Instituto Nacional de Estadística

OPOSICIONES AL CUERPO SUPERIOR DE ESTADÍSTICOS DEL ESTADO

BOE NÚM. 270, DE 12 DE OCTUBRE DE 2020, PÁG. 87165

Econometría

Grupo de Materias Específicas: Economía

INE
28 de septiembre de 2021

Índice general

Tema 1

Modelos causales y no causales. Datos

Este tema está elaborado como una adaptación de los capítulos 2 y 3:

Cameron, A.C. y Trivedi, P.K., 2005. Microeconometrics. Methods and applications. Cambridge University Press. Así como de la bibliografía complementaria.

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al Órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

- Modelos estructurales.
- Exogeneidad.
- Modelo de ecuaciones lineales simultáneas.
- Conceptos de identificación.
- Modelos de una sola ecuación.
- Modelo de resultados potenciales.
- Modelización causal y estrategias de estimación.
- Datos observacionales. Datos de experimentos sociales.
- Datos de experimentos naturales.

1.1 Modelos estructurales

Un **modelo** es la especificación de la distribución de probabilidad para un **conjunto de observaciones** (**datos**). Una **estructura** es la especificación de los parámetros de dicha distribución. Por lo tanto, una estructura es un modelo en el que a todos los parámetros se les asignan valores numéricos.

Un modelo tiene por objetivo explicar los valores de una variable y, o más generalmente, de un vector de variables $\mathbf{y}$, donde $\mathbf{y}' = (y_1, ..., y_G)$. Cada valor del vector $\mathbf{y}$ es una función de:

- algunos otros elementos de $\mathbf{y}$, por lo tanto puede haber interdenpendencia entre las variables incluidas en el vector $\mathbf{y}$

- de un conjunto de variables explicativas $\mathbf{z}$, cuyas posibles interdependencias no forman parte de la modelización
- de un término puramente aleatorio $\mathbf{u}$ denominado error, que en la literatura econométrica también suele denotarse por ε, o por ϵ.

A un elemento del conjunto de observaciones lo referenciamos mediante i, y nos referimos a el mismo como observación o elemento $i-$ésimo del conjunto de datos. El conjunto de datos referidos a las variables del modelo lo particionamos en dos bloques

$$[\mathbf{Y}, \mathbf{Z}]$$

que responden a una ordenación a priori y conforme a unas relaciones causa-efecto sobre el modelo.

La i-ésima observación satisface el conjunto de ecuaciones implícitas siguiente

$$\mathbf{g}(\mathbf{y}_i, \mathbf{z}_i, \mathbf{u}_i|\boldsymbol{\theta}) \tag{1.1}$$

donde g es una función conocida y el vector $\boldsymbol{\theta}$ es el vector de parámetros (estructurales). Precisamente a esto le denominamos **modelo estructural**. La especificación de la forma funcional g y de las restricciones sobre los parámetros $\boldsymbol{\theta}$ puede ser paramétrica, semiparamétrica o no-paramétrica.

Goldberger define un modelo estructural como aquel que representa una relación causal, en oposición a a una relación que simplemente captura asociaciones estadísticas. Una ecuación estructural puede obtenerse de un modelo económico, o puede obtenerse a través de razonamientos informales. A veces, el modelo estructural se puede estimar directamente. Otras veces debemos combinar supuestos auxiliares sobre otras variables con manipulaciones algebraicas para llegar a un modelo estimable, como vemos a continuación.

En caso de haber una solución única para $\mathbf{y}_i$ para cada $(\mathbf{z}_i, \mathbf{u}_i)$, entonces podríamos escribir las ecuaciones explícitas para $\mathbf{y}$ como función de ($\mathbf{z}$,$\mathbf{u}$)

$$\mathbf{y}_i = \mathbf{f}(\mathbf{z}_i, \mathbf{u}_i|\boldsymbol{\pi}) \tag{1.2}$$

siendo $\boldsymbol{\pi}$ un vector de parámetros que son función de $\boldsymbol{\theta}$. La expresión (1.2) se denomina **forma reducida** del modelo estructural. En efecto, la forma reducida se obtiene resolviendo el modelo estructural para las variables (interdependientes, que denominamos **endógenas**) $\mathbf{y}_i$ dados $(\mathbf{z}_i, \mathbf{u}_i)$. Por estos motivos, $\boldsymbol{\pi}$ un vector de parámetros de la forma reducida y son función del vector de parámetros de la forma estructural.

Un objetivo prioritario en econometría es realizar inferencia precisamente sobre el vector de parámetros estructurales, y las relaciones implícitas de la expresión (1.1) nos darían una vía directa si pudieremos estimar el modelo estructural. Sin embargo, dado que el vector $\boldsymbol{\pi}$ es una función de $\boldsymbol{\theta}$, entonces (1.2) proporciona una vía alternativa de lograr el objetivo de hacer inferencia sobre $\boldsymbol{\theta}$.

Si conociéramos la forma funcional (1.2) y fuera aditivamente separable en $\mathbf{z}_i$ y en $\mathbf{u}_i$, de tal modo que pudiéramos escribir

$$\mathbf{y}_i = \mathbf{g}(\mathbf{z}_i|\boldsymbol{\pi}) + \mathbf{u}_i = \mathbb{E}(\mathbf{y}_i|\mathbf{z}_i) + \mathbf{u}_i$$

entonces la regresión de $\mathbf{y}$ sobre $\mathbf{z}$ sería una función de predicción natural para el vector $\mathbf{y}$ dado el vector $\mathbf{z}$. O dicho en otros términos, la forma reducida tiene, en este contexto, la funcionalidad de realizar predicciones condinadas de de $\mathbf{y}_i$ dados $(\mathbf{z}_i, \mathbf{u}_i)$ siempre que podamos estimar el vector de parámetros de la forma reducida, que es a priori más sencillo que el vector de parámetros de la forma estructural.

Hemos considerado con naturalidad que del conjunto de variables del modelo podíamos distinguir entre aquellas variables cuya variación es explicada por el modelo (endógenas) y aquellas que cuya variación está determinada "fuera" del modelo (**exógenas**). De hecho unas las denotamos por $\mathbf{y}$ y a las segundas por $\mathbf{z}$. Es evidente que este punto discriminación entre variables es crítico en la modelización y por ello se requiere introducir un nuevo elemento en el modelo: la función de distribución de probabilidad conjunta de las variables $[\mathbf{Y}, \mathbf{Z}]$

$$F(\mathbf{W}), \mathbf{W} = [\mathbf{Y}, \mathbf{Z}]$$

a partir del cual se podrá dar una definición de exogeneidad formal, que posteriormente irá materializándose y quedando más clara a lo largo de este tema y de los siguientes temas del bloque econométrico.

1.2 Exogeneidad

Hemos considerado con naturalidad que del conjunto de variables del modelo podíamos distinguir entre aquellas variables cuya variación es explicada por el modelo (endógenas) y aquellas que cuya variación está determinada "fuera" del modelo (**exógenas**). De hecho unas las denotamos por $\mathbf{y}$ y a las segundas por $\mathbf{z}$. Es evidente que este punto discriminación entre variables es crítico en la modelización y por ello se requiere introducir un nuevo elemento en el modelo: la función de distribución de probabilidad conjunta de las variables $[\mathbf{Y}, \mathbf{Z}]$

$$F(\mathbf{W}), \mathbf{W} = [\mathbf{Y}, \mathbf{Z}].$$

Considermos la distribución conjunta de **W,** con parámetros ϑ divididos como $(\boldsymbol{\theta}_1\boldsymbol{\theta}_2)$. Dichas distribución se factoriza en la densidad condicional (f_C) de **Y** dado **Z,** y la distribución marginal de **Z** (f_M):

$$f_J(\mathbf{W}|\boldsymbol{\theta}) = f_C(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta}) \times f_M(\mathbf{Z}|\boldsymbol{\theta})$$

Un caso especial de este resultado ocurre si

$$f_J(\mathbf{W}|\boldsymbol{\theta}) = f_C(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta}_1) \times f_M(\mathbf{Z}|\boldsymbol{\theta}_2)$$

siendo $(\boldsymbol{\theta}_1\boldsymbol{\theta}_2)$ funcionalmente independientes. En tal caso, decimos que Z es exógeno respecto de $\boldsymbol{\theta}_1$, en el sentido de que no se precisa del conocimiento de $f_M(\mathbf{Z}|\boldsymbol{\theta}_2)$ para hacer inferencia sobre $\boldsymbol{\theta}_1$, por lo que es perfectamente correcto condicionar la distribución de **Y** a **Z**.

Existen en econometría dos formas de exogeneidad que se irán explicando en los siguientes temas: exogeneidad débil y exogeneidad fuerte, el nexo de unión es el concepto de indepencia condicionada:

Decimos que x_1 e y son condicionalmente independientes dado x_2 si

$$f(y|x_1, x_2) = f(y|x_2)$$

es decir, conocido x_2 (condicionado a) las variables y y x_1 son independientes. Una versión menos restrictiva sería la independencia en media:

$$\mathbb{E}(y|x_1, x_2) = \mathbb{E}(y|x_2)$$

que indica la ausencia de capacidad predictiva de x_1 en y después de condionar sobre x_2. Situación conocida como ausencia de causalidad tipo Granger.

1.3 Modelo de ecuaciones lineales simultáneas

Un caso importante del modelo (1.1) es el modelo lineal de ecuaciones simultáneas, que se describe a partir de las siguientes ecuaciones

$$\begin{aligned} y_{1i}\beta_{11} + ... + y_{Gi}\beta_{1G} + z_{1i}\gamma_{11} + ... + z_{Ki}\gamma_{1K} &= u_{1i} \\ ..... &= ... \\ y_{1i}\beta_{G1} + ... + y_{Gi}\beta_{GG} + z_{1i}\gamma_{G1} + ... + z_{Ki}\gamma_{GK} &= u_{Gi} \end{aligned}$$

que matricialmente escribimos como

$$(y_{1i}, ..., y_{Gi}) \begin{pmatrix} \beta_{11} & & \beta_{1G} \\ \vdots & \ddots & \vdots \\ \beta_{G1} & & \beta_{GG} \end{pmatrix} + (z_{1i}, ..., z_{Ki}) \begin{pmatrix} \gamma_{11} & & \gamma_{1K} \\ \vdots & \ddots & \vdots \\ \gamma_{G1} & & \gamma_{GK} \end{pmatrix} = (u_{1i}, ..., u_{Gi})$$

y más compactamente

$$\mathbf{y}_i'\mathbf{B} + \mathbf{z}_i'\mathbf{\Gamma} = \mathbf{u}_i'$$

donde dados unos valores para los parámetros $(\mathbf{B}, \mathbf{\Gamma})$ y para $(\mathbf{z}_i, \mathbf{u}_i)$, permiten resolver simultáneamente el sistema de ecuaciones lineales para $\mathbf{y}_i$.

En esta especificación apreciamos los elementos de la estructura de los modelos. Tenemos un conjunto de variables que diferenciamos entre endógenas y exógenas. Estas últimas no guardan correlación alguna con las variables de término error $(u_{1i}, ..., u_{Gi})$. También es necesario completar el modelo con una serie de restricciones sobre los parámetros que son las siguientes:

1. **B** es una matriz no singular de rango G
2. El rango de $\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1' \\ : \\ \mathbf{z}_G' \end{bmatrix}$ es K

3. El plim $(\mathbf{Z}'\mathbf{Z}) = \mathbf{\Sigma_{ZZ}}$ es una matriz simétrica cuadrada K que es definida positiva
4. Unos elementos que caractericen la función de densidad de los errores del modelo. Por ejemplo, aunque como veremos ya no es lo más habitual, una parametrización de dicha función es la siguiente: $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$: $\mathbb{E}(\mathbf{u}_i) = 0, \mathbb{E}(\mathbf{u}_i\mathbf{u}_i')$=$\mathbf{\Sigma}$que una matriz simétrica cuadrada G que es definida positiva
5. Los errores en cada ecuación son serialmente independientes

Los parámetros estructurales son en este caso $(\mathbf{B}, \mathbf{\Gamma}, \mathbf{\Sigma})$. Obsérvese que estas restricciones o supuestos pueden variar para hacerlos más generales (menos restrictivos), y lo haremos más adelante, pero por ahora al objeto de la exposión no se hace necesario contemplar otros. Dado que este sistema se cumple para cada elemento i, podemos escribir más compactamente el modelo estructural

$$\mathbf{YB} + \mathbf{Z\Gamma} = \mathbf{U}$$

donde

$$\underset{N\times G}{\mathbf{Y}} = \begin{bmatrix} \mathbf{y}_{1i}' \\ : \\ \mathbf{y}_{Ni}' \end{bmatrix}, \underset{N\times K}{\mathbf{Z}} = \begin{bmatrix} \mathbf{z}_{1i}' \\ : \\ \mathbf{z}_{Ni}' \end{bmatrix}, \underset{N\times G}{\mathbf{U}} = \begin{bmatrix} \mathbf{u}_{1i}' \\ : \\ \mathbf{u}_{Ni}' \end{bmatrix}$$

y las matrices paramétricas $\mathbf{B}, \mathbf{\Gamma}$ son de dimensiones $G \times G, K \times G$, respectivamente. En este caso, el modelo de ecuaciones simultáneas junto con las restricciones permite obtener la **forma reducida**

$$\mathbf{YBB}^{-1} + \mathbf{Z\Gamma B}^{-1} = \mathbf{UB}^{-1};$$

$$\mathbf{Y} = \mathbf{Z\Pi} + \mathbf{V}, \tag{1.3}$$

donde $\mathbf{\Pi} = -\mathbf{\Gamma B}^{-1}$ y $\mathbf{V} = \mathbf{UB}^{-1}$.

El modelo estructural especialmente importante en econometría cuando estamos intentado investigar o conocer los efectos parciales de las variables exógenas sobre las endógenas. En efecto, las ecuaciones contienen interpretaciones en términos establecidos por las relaciones económicas (oferta-demanda, producción-costes, salario-formación, etc.), y por tanto son susceptibles de que se impongan de forma natural restricciones procedentes de la teoría económica. En consecuencia, **B** y Γ son parámetros que describen el comportamiento económico. Por tanto, se puede invocar la teoría a priori para formar expectativas sobre el signo y el tamaño de los coeficientes individuales. Por el contrario, los parámetros (no restringidos) de la denominada forma reducida son funciones (potencialmente complicadas) de los parámetros estructurales, y, como tales, puede ser difícil evaluarlos después tras la estimación.

La forma reducida expresa cada variable endógena como una función lineal de todas las variables exógenas y todas los errores estructurales. Los errores de la forma reducida son combinaciones lineales de erroes estructurales. La forma reducida para la i-ésima observación es

$$\mathbb{E}[\mathbf{y}_i|\mathbf{z}_i] = \mathbf{z}_i'\mathbf{\Pi}; Var[\mathbf{y}_i|\mathbf{z}_i] = \mathbf{B}^{-1'}\mathbf{\Sigma}\mathbf{B}^{-1}$$

Los parámetros de forma reducida, Π, son parámetros derivados a partir de funciones de los parámetros estructurales. Si Π se puede estimar de manera consistente, entonces

la forma reducida se puede usar para hacer predicciones sobre variaciones en **Y** debido a cambios exógenos en **Z**. Obsérvese que estas predicciones condicionadas (condicionales a **Z**) son posible incluso si no se conocen **B** y Γ. Dada la exogeneidad de **Z**, el conjunto completo de regresiones de la forma reducida es un modelo de regresión multivariante que puede estimarse consistentemente por mínimos cuadrados. Por tanto si estamos interesados solo en la **predicción** la forma reducida es extramadamente útil.

Si estamos interesados en las **relaciones causales** (establecidas en las ecuaciones estructurales), entonces ya hemos visto que es posible recuperar los parámetros estructurales bajo ciertas restricciones. Los parámetros estructurales desconocidos, los elementos distintos de cero de **B**, Γ y Σ, juegan un papel clave porque reflejan la estructura causal del modelo. **B** describe la interdependencia entre variables endógenas, mientra que las respuestas de las variables endógenas a impactos exógenos en **Z** se reflejan en la matriz de parámetros Γ.

En esta configuración, los parámetros causales de interés son aquellos que miden el impacto marginal directo de un cambio en una variable explicativa, y_j o z_k sobre la variable de interés $y_l, l \neq j$. Los elementos de Σ describen las propiedades de dispersión y dependencia de los errores aleatorios y, por lo tanto, miden algunas propiedades de la forma en que se generan los datos.

Las justificaciones habituales para centrarse en los parámetros estructurales son las siguientes: (1) nos interesan las estimaciones de "parámetros económicos" (piense el lector por ejemplo en las elasticidades de la oferta de trabajo); (2) las estimaciones de los parámetros estructurales nos permiten obtener los efectos de una variedad de intervenciones políticas (por ejemplo, cambios en las tasas impositivas).

1.4 Conceptos de identificación

En términos generales, la identificación se refiere a la determinación de un parámetro con suficientes observaciones. En este sentido, es un concepto asintótico. La incertidumbre estadística afecta necesariamente a cualquier inferencia basada en un número finito de observaciones. Al considerar hipotéticamente la posibilidad de que haya un número suficiente de observaciones disponibles, es posible considerar si es lógicamente factible determinar un parámetro de interés, entendiendo por deteminar el establecer su valor puntual o bien determinar el conjunto al que pertenece dicho parámetro.

Dos modelos son observacionalmente equivalentes si, dados los datos, los dos modelos estructurales implican distribuciones de probabilidad conjunta idénticas de las variables. En este sentido, la existencia de múltiples estructuras observacionalmente equivalentes implica el fracaso de la identificación. Un ejemplo simple de no identificación ocurre cuando hay una colinealidad perfecta entre regresores en la regresión lineal $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$. Entonces podemos identificar la combinación lineal $\mathbf{C}\beta$, donde rango [**C**] <rango [β], pero no podemos identificar β .

La identificación requiere por tanto que no haya una estructura observacionalmente equivalente. En el contexto de los modelos de ecuaciones simultáneas, la identificación requiere que haya una tripleta única (**B**, Γ, Σ) consistente con los datos observados.

Implica por tanto poder obtener estimaciones únicas de parámetros estructurales dados los momentos muestrales de los datos.

En el caso de la forma reducida (1.3), bajo los supuestos establecidos, el estimador de mínimos cuadrados proporciona estimaciones únicas de Π, es decir,

$$\widehat{\mathbf{\Pi}} = [\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{Y};$$

y la identificación de $\mathbf{B}$, Γ requiere que exista solución para los elementos de dichas matrices a través de las relaciones establecidas en las ecuaciones

$$\mathbf{\Pi} + \mathbf{\Gamma}\mathbf{B}^{-1} = \mathbf{0}$$

y las restricciones a priori (que generalmente provienen de la teoría económica) sobre el modelo. Si estas restricciones son suficientes, entonces implicaría que el modelo estaría identificado, y por tanto todos los parámetros del modelo lo estarían.

En la práctica, se pueden imponer una variedad de restricciones que incluyen

1. normalizaciones, como establecer elementos diagonales de $\mathbf{B}$ igual a la unidad,
2. restricciones de exclusión, es decir, indicar que algunas variables tiene nulo impacto en alguna(s) de las variables exógenas; esto implica que ciertas direciones de la causalidad están excluidas a priori, lo que posibilita identificar otras direcciones de la causalidad.
3. restricciones sobra la covarianza que pueden ser de desigualdad.

Si no hay restricciones sobre la varianza Σ, y los elementos de la diagonal principal de B se normalizan a 1, entonces una **condición necesaria** para la identificación es la **condición de orden**, que establece que el número de variables exógenas excluidas debe ser al menos igual al número de variables endógenas incluidas.

Una **condición suficiente** es la **condición de rango:** los parámetros de la j-ésima ecuación $\mathbf{\Pi\Gamma}_j = -\mathbf{B}_j$ generen una solución única para $(\mathbf{\Gamma}_j, \mathbf{B}_j)$ dado Π.

El modelo puede estar **exactamente identificado**, la condición del pedido se satisface exactamente; o bien puede estar sobreidentificado, cuando el número de restricciones en el sistema excede el requerido para una identificación exacta.

1.5 Modelos de una sola ecuación

Consideremos la primera ecuación de un sistema lineal sujeto a la normalización $\beta_{11} = 1$. Sea $y = y_1$, sea $\mathbf{y}_1$ las componentes endógenas de $\mathbf{y}$ distintas de y_1, y $\mathbf{z}_1$ las componentes exógenas de $\mathbf{z}$ con

$$y = \mathbf{y}_1'\alpha + \mathbf{z}_1'\gamma + u$$

que también puede reescribirse como

$$y = \mathbf{x}'\beta + u$$

donde algunos componentes de $\mathbf{x}$ son endógenos (implícitamente $\mathbf{y}_1$) y otros son exógenos (implícitamente $\mathbf{z}_1$). La atención se centra entonces en estimar el impacto de los cambios en los regresores clave que pueden ser endógenos o exógenos, según los supuestos. La técnica de estimación natural se basa en el uso de variables instrumentales que se comentará más adelante en los temas.

Es habitual especificar al menos algunas de las ecuaciones restantes en el modelo, incluso si no son el foco de investigación. Suponga que $\mathbf{y}_1$ tiene dimensión unitaria. Entonces la primera posibilidad es especificar la ecuación estructural para y_1 y para las otras variables endógenas que pueden aparecer en esta ecuación estructural para y_1. Una segunda posibilidad es especificar la ecuación en forma reducida para y_1. Esto mostrará las variables exógenas que afectan a y_1, pero que no afectan directamente a y. Una ventaja es que en tal escenario las variables instrumentales emergen de forma natural.

1.6 Modelo de resultados potenciales

La motivación para la inferencia causal en los modelos econométricos es especialmente fuerte cuando la atención se centra en el impacto de las políticas públicas y / o las variables de decisión privada sobre algunos resultados específicos. Por ejemplo, el impacto del tamaño de la clase en el aprendizaje de los estudiantes o el impacto de tener un seguro médico (privado) en la utilización del servicio de atención sanitaria. En muchos casos, las propias variables causales reflejan decisiones individuales y, por tanto, son potencialmente endógenas. Cuando, como suele ser el caso, la estimación econométrica y la inferencia se basan en datos observacionales, la identificación y la inferencia de los parámetros causales plantean muchos desafíos. Estos desafíos pueden volverse potencialmente menos serios si los problemas causales se abordan utilizando datos procedentes de un **experimento social controlado** con un diseño estadístico adecuado. Aunque tales experimentos se han implementado, generalmente son costosos de organizar y ejecutar. Realmente lo más atractivo sería desarrollar modelos causales utilizando datos generados por un experimento natural o en un entorno cuasi-experimental. Un entorno de este tipo es como un escenario en el que alguna variable causal cambia exógena e independientemente de otras variables explicativas, lo que hace relativamente más fácil identificar los parámetros causales.

Un obstáculo importante para el modelado de causalidad surge del problema fundamental de la inferencia causal. Sea X la causa hipotética e Y el resultado. Al manipular el valor de X podemos cambiar el valor de Y. Suponga que el valor de X se cambia de x_{11} a x_2. Luego, se forma una medida del impacto causal del cambio en Y comparando los dos valores de Y: y_2, que resulta del cambio, e y_1, que habría resultado si no hubiera ocurrido ningún cambio en X. Sin embargo, si X cambiara, entonces, en ausencial del cambio, no se observaría el valor de Y. Por lo tanto, no se puede decir nada más sobre el impacto causal sin alguna hipótesis sobre qué valor habría asumido Y en ausencia del cambio en X. Este último se conoce como un hecho contrafactual, que significa un valor hipotético no observado. En pocas palabras, toda inferencia causal implica la comparación de un resultado fáctico con un resultado contrafactual. En el modelo econométrico

convencional no es necesario establecer explícitamente un contrafactual.

Un aspecto de cierto atractivo en la actualidad y estrechamente relacionado con esto es la evaluación de programas o la evaluación de tratamientos, que proporciona un marco estadístico para la estimación de parámetros causales. En la literatura estadística, este marco también se conoce como el modelo causal de Rubin (RCM). Los parámetros causales basados en contrafactuales proporcionan definiciones operacionales y estadísticamente significativas de la causalidad que en algunos aspectos difieren de la definición tradicional de la econometría. En primer lugar, en entornos ideales, este marco conduce a una considerable simplicidad de los métodos econométricos. Segundo, este marco se centra un menor número de parámetros causales los cuales se cree que son más relevantes para las cuestiones de política que se examinan. Esto contrasta con el enfoque econométrico tradicional que se centra simultáneamente en todos los parámetros estructurales, como hemos visto más arriba. En tercer lugar, el enfoque proporciona información adicional sobre las propiedades de los parámetros causales estimados por los métodos estructurales estándar. A continuación se desarrolla este modelo causal.

El objeto último de los experimentos es aprender o tener información sobre el efecto que tiene sobre una unidad de análisis el estar expuesto a un tratamiento. Supongamos que estamos interesados en asesorar la toma de una decisión sobre inscribirse en un programa de formación laboral (tratamiento), o bien sobre ir a la universidad o no. Es razonable preguntarse sobre cuáles son los beneficios de apuntarse (de recibir el tratamiento). Un marco analítico con muchas ventajas para asesorar dicha decisión es imaginar qué pasaría en el hipotético caso de recibir el tratamiento y qué resultado tendría en caso de no recibirlo (es decir, de no inscribirse en el curso de formación laboral, por ejemplo). En estos términos, la diferencia entre ambos resultados sería el efecto causal individual del tratamiento (de apuntarse al programa).

Necesitamos identificar una variable respuesta que indique el resultado que obtendría una unidad (individuo, en este caso) al recibir un tratamiento. Por simplicidad podemos considerar un tratamiento binario, es decir o se expone al tratamiento o no. La variable binaria $X_i = 1$, si el individuo i-ésimo recibe el tratamiento, y $X_i = 0$, en caso de que no lo recibiera. En tal caso, definimos una variable respuesta que recoja los dos potenciales resultados:

$$\text{resultado potencial} = \begin{cases} Y_i(1) & \text{si } X_i = 1 \\ Y_i(0) & \text{si } X_i = 0 \end{cases}.$$

Nos interesa la diferencia de resultados potenciales, es decir, nos interesa la diferencia entre $Y_i(1)$ y $Y_i(0)$ en la medida en que es el efecto causal de estar expuesto a tratamiento (ir a la universidad, o bien recibir formación de inserción laboral).

El principal problema es que no es posible medir el efecto causal para una sola persona, es decir, solo uno de los dos resultados potenciales puede ser realizado por el individuo (y por tanto un solo resultado es observado). Nótese que antes de que se tome una decisión ambos son potencialmente observables, de ahí que se le donomine resultado potencial. Este resultado potencial no hay que confundirlo con el resultado observado o realizado, que denominamos Y_i. Ambos conceptos se relacionan fácilmente a partir de la expresión siguiente, que realmente nos permite definir el resultado observado a

partir de los resultados potenciales:

$$Y_i = Y_i(1)X_i + Y_i(0)\,(1 - X_i)\,. \tag{1.4}$$

Debido a la imposibilidad de medir el efecto causal de un tratamiento X para un individuo (es decir, $Y_i(1) - Y_i(0)$), es suficiente con considerar el **efecto causal promedio**. Cuando algunos individuos reciben el tratamiento y otros no, la diferencia esperada en los resultados entre los dos grupos es

$$\mathbb{E}\,(Y_i\,|X_i = 1\,) - \mathbb{E}\,(Y_i\,|X_i = 0\,) = \mathbb{E}\,(Y_i(1)\,|X_i = 1\,) - \mathbb{E}\,(Y_i(0)\,|X_i = 0\,)\,, \tag{1.5}$$

donde la igualdad se obtiene simplemente usando la expresión (1.4). Esta expresión pone en relación la diferencia observada entre las medias de los resultados experimentales, y las medias de los resultados potenciales, sin embargo no es exactamente en lo que estamos interesados. Nuestro interés es saber cuándo a partir de las diferencias observadas podemos extraer el efecto causal promedio de la población de la que extrajeron los sujetos. Dicho en otros términos, si nuestro interés es la diferencia salarial entre los que van a la universidad y los que no, comparar las medias salariales de los individuos que fueron a la universidad y la de los que no fueron no nos proporciona necesariamente una medida del efecto causal de ir a la universidad. De hecho, es posible que la diferencia de una y otra exagere por exceso el efecto causal, toda vez que es bastante posible que en media aquellos que han ido a la universidad hubieran ganado más (que los que no fueron) incluso en caso de no haber ido. Es decir, la simple diferencia de medias no considera que hay un *sesgo de selección* que distorsiona las conclusiones.

Para verlo formalmente tenemos que introducir la expresión

$$\mathbb{E}\,(Y_i(0)\,|X_i = 1\,)\,,$$

que refleja el resultado potencial esperado que habría obtenido el individuo que ha sido expuesto al tratamiento, en caso de no haber sido expuesto. En el ejemplo en el que el tratamiento es ir a la universidad, la expresión considera cuál hubiera sido el salario de una persona que ha ido a la universidad, con sus característica propias, en caso de que no hubiera ido. En el ejemplo en el que el tratamiento es atender a un programa de formación laboral, el término recoge cuál hubiera sido el salario medio en caso de que el sujeto que atendió al programa no hubiera atendido. Podemos introducir esta expresión en (1.5) haciendo lo siguiente

$$\underbrace{\mathbb{E}\,(Y_i\,|X_i = 1\,) - \mathbb{E}\,(Y_i\,|X_i = 0\,)}_{\text{Diferencias de promedios observados}} = \underbrace{\mathbb{E}\,(Y_i(1)\,|X_i = 1\,) - \mathbb{E}\,(Y_i(0)\,|X_i = 1\,)}_{\text{Efecto promedio del tratamiento en tratados}} + \underbrace{\mathbb{E}\,(Y_i(0)\,|X_i = 1\,) - \mathbb{E}\,(Y_i(0)\,|X_i = 0\,)}_{\text{Sesgo de selección}}.$$

Justamente esta expresión nos permite visualizar dos cosas importantes: (a) La utilidad del concepto de resultado potencial. El primer sumando recoge la diferencia de las medias de resultados potenciales que los sujetos tratados obtendrían si en lugar de haber sido tratados, no lo hubieran sido. (b) El papel potencialmente distorsionador del

término «sesgo de selección». Debido a que es posible que aun así los que han ido a la universidad obtuvieran mayores salarios que los que tendrían los que no han ido, las diferencias de promedios observados sobrestimarían el efecto causal promedio, es decir, el sesgo de selección en este caso sería positivo. Es incluso posible que en algunos casos el sesgo sea de tal magnitud que vele los efectos de un tratamiento determinado. Por este motivo es fundamental afrontar la cuestión de cómo cancelar el sesgo de selección.

El objetivo por tanto es estimar el efecto causal promedio para un población dada, para lo cual es importante eliminar el sesgo de selección. Esto es teóricamente posible en el caso de los experimentos aleatorizados controlados. Veamos por qué. En general, el efecto causal individual de un tratamiento puede variar de un individuo a otro porque su efecto puede depender de otras características del sujeto, lo que implica que las distribuciones de $Y_i(1)$ y $Y_i(0)$ serían distintas. Sin embargo, si hacemos una selección aleatoria de los individuos a partir de una población, las variables respuesta (y por tanto sus efectos causales) se pueden considerar extraídas de una misma distribución, por lo que el valor esperado (promedio) del efecto muestral coincidiría con el valor esperado del efecto poblacional. Por otro lado, si los sujetos pudieran ser asignados aleatoriamente a los grupos de tratamiento y control, entonces el estado de un sujeto (X_i, tratado o no tratado) se distribuiría independientemente de todos los atributos personales del individuo, así como de las potenciales respuestas, $Y_i(1)$ y $Y_i(0)$. Formalmente, la independencia implica que los promedios en tal caso satisfacen $\mathbb{E}\left(Y_i(0)\,|X_i=1\right) = \mathbb{E}\left(Y_i(0)\,|X_i=0\right)$, por lo que sustituyendo en la expresión (1.5) se tiene

$$\begin{aligned}\mathbb{E}\left(Y_i\,|X_i=1\right)-\mathbb{E}\left(Y_i\,|X_i=0\right) &= \mathbb{E}\left(Y_i(1)\,|X_i=1\right)-\mathbb{E}\left(Y_i(0)\,|X_i=1\right)\\ &= \mathbb{E}\left(Y_i(1)-Y_i(0)\,|X_i=1\right)\\ &= \mathbb{E}\left(Y_i(1)-Y_i(0)\right),\end{aligned}$$

donde la última igualdad se obtiene de la independencia inducida por la asignación aleatoria del tratamiento[1].

Esta última expresión indica que si a partir de una selección aleatoria de sujetos, asignamos aleatoriamente el tratamiento, entonces la diferencia de promedios de los resultados

[1]En general, sin embargo, esto no será así. Para comprobarlo supongamos que no hay efecto causal, de modo que $Y_i(0) = Y_i(1)$ para todos los individuos, y por tanto el efecto causal promedio será nulo, $\mathbb{E}\left(Y_i(1)-Y_i(0)\right)=0$. Consideremos igualmente que el tratamiento X_i está, por ejemplo, positivamente correlacionado con el *resultado potencial*. En el ejemplo en el que el tratamiento es ir a la universidad, es bastante probable que los estudiantes que van a la universidad (tratados) sean los más motivados o los que tienen mayores habilidades. En esta situación de correlación positiva entre estar tratado $X_i=1$ y el resultado potencial, resultaría que

$$\mathbb{E}\left(Y_i(1)\,|X_i=1\right) > \mathbb{E}\left(Y_i(1)\right),$$

$$\mathbb{E}\left(Y_i(0)\,|X_i=0\right) < \mathbb{E}\left(Y_i(0)\right).$$

Es decir, que en tal caso, el salario promedio potencial por ir a la universidad de la población que efectivamente ha ido a la universidad es mayor que el salario promedio potencial por ir a la universidad de la población con independencia de si efectivamente ha ido o no a la universidad. Esto significa que entonces

$$\mathbb{E}\left(Y_i(1)\,|X_i=1\right)-\mathbb{E}\left(Y_i(0)\,|X_i=0\right) > \mathbb{E}\left(Y_i(1)\right)-\mathbb{E}\left(Y_i(0)\right)=0,$$

donde la igualdad se debe al supuesto que hemos hecho de efecto causal nulo.

observados en el experimento entre los tratados y lo no tratados coincide con el efecto promedio causal del tratamiento en la población. Dicho en otros términos, si denominamos experimento aleatorizado controlado a aquel expermiento que satisface el diseño que hemos indicado, entonces el sesgo de selección se anula, de modo que las diferencias de medias observadas nos permiten capturar el efecto causal promedio, que es lo que buscábamos.

Obviamente, este experimento sería ideal, y la realidad nos devuelve a situaciones que no garantizan la implementación de las condiciones del experimento aleatorizado controlado ideal. Es decir condiciones en las que la asignación del tratamiento es independiente de los resultados potenciales

$$X_i \perp [Y_i(0), Y_i(1)] .$$

De hecho en el ejemplo del efecto causal sobre el salario por motivo de ir a la universidad, no se dan las condiciones de experimento aleatorizado controlado. La asignación del tratamiento, en ese ejemplo, no es aleatoria. La mayoría de los trabajos aplicados con experimentos ideales se han hecho en áreas relacionadas con la bioestadística. No obstante, es posible que se den situaciones en ramas socio-económicas (de hecho se han dado) en las que es posible llevar a cabo un experimento ideal. Normalmente cuando esto ocurre las conclusiones de esta investigación tienen mucha relevancia a la hora de asesorar en la toma de decisiones sobre el desarrollo de programas.

1.7 Modelización causal y estrategias de estimación

Tanto en el contexto análitico o marco referencial previsto en los modelos de ecuaciones simultáneas, como en el previsto en los modelos de respuesta potencial, la modelización de la causalidad -dada su inherente complejidad y limitaciones empíricas debido a la naturaleza del tipo de datos utilizados en la ciencias sociales- un(a) económetra tiene varias alternativas para alcanzar dicha modelización.

Distinguimos, primeramente entre modelos de información completa e información limitada. La modelizción en un contexto de **información completa**, desde un punto de vista estadístico, consiste en considerar la distribución de probabilidad conjunta de las variables endógenas, dadas las variables exógenas, y esto será la base de la inferencia sobre la causalidad. Las relaciones no se derivan necesariamente de un modelo de comportamiento optimizador. Se colocan restricciones paramétricas para asegurar la identificación de los parámetros del modelo que son el objetivo de la inferencia estadística. El modelo completo se estima simultáneamente utilizando la estimación de máxima verosimilitud o basada en momentos. Para modelos bien especificados, este es un enfoque atractivo pero, en general, su limitación potencial es que puede contener algunas ecuaciones que están mal especificadas. La articulación de la distribución conjunta puede derivar de la interdependencia contemporánea o dinámica entre variables endógenas y / o los errores de las ecuaciones.

Por el contrario, cuando el objeto central de la inferencia estadística es la estimación de uno o dos parámetros clave, se puede utilizar un enfoque de **información limitada**. Una

característica de este enfoque es que, aunque una ecuación es el centro de la inferencia, se explota la dependencia conjunta entre ella y otras variables endógenas. Esto requiere que se hagan suposiciones explícitas sobre algunas características del modelo que no son el objeto principal de inferencia. Los métodos de variables instrumentales, los métodos secuenciales de varios pasos y los métodos de máxima verosimilitud de información limitada son ejemplos específicos de este enfoque. Para implementar el enfoque, uno normalmente trabaja con una (o más) ecuaciones estructurales y algunas ecuaciones de forma reducida implícita o explícitamente establecidas. Esto contrasta con el enfoque de información completa donde todas las ecuaciones son estructurales. El enfoque de información limitada a menudo es computacionalmente más manejable que el de información completa. Estadísticamente, podemos interpretar el enfoque de información limitada como uno en el que la distribución conjunta se factoriza en el producto de un modelo condicional para las variables endógenas de interés, digamos y1, y un modelo marginal para otras variables endógenas, digamos y2, que están en el conjunto de las variables condicionantes, como en

$$f(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}) = g(\mathbf{y}_1|\mathbf{x},\mathbf{y}_2,\boldsymbol{\theta}_1)h(\mathbf{y}_2|\mathbf{x},\boldsymbol{\theta}_2), \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

La modelización se basa en $g(\mathbf{y}_1|\mathbf{x},\mathbf{y}_2,\boldsymbol{\theta}_1)$ con escasa atención al factor $h(\mathbf{y}_2|\mathbf{x},\boldsymbol{\theta}_2)$ si es que $\boldsymbol{\theta}_2$ se pueden considerar parámetros molestos (nuisance parameters).

Otra alternativa en el marco de las ecuaciones simultáneas es partir de la forma reducida identificada, como hemos visto anterioremente.

Necesitamos, independientemente de la alternativa de modelización (limitada o no), formas de identificar los parámetros fundamentales del modelo. Sin embargo hay varios riesgos que iremos viendo a lo largo de los siguientes temas. Riesgos tales como: posibles variables omitidas, especificaciones incorrectas de la forma funcional, errores de medición en las variables explicativas, el uso de datos no representativos de la población, y el ignorar la endogeneidad de las variables explicativas, entre los más relevantes.

A veces, los datos se generan mediante experimentos naturales y cuasiexperimentos (lo veremos más adelante). La idea aquí es simplemente que una variable de política puede cambiar exógenamente para alguna subpoblación mientras que permanece igual para otras subpoblaciones. Por ejemplo, las leyes de salario mínimo en un estado pueden cambiar mientras permanecen sin cambios en un estado vecino. Tales eventos crean de "forma natural" grupos de tratamiento y control. Si el experimento natural se aproxima a una asignación de tratamiento aleatoria, entonces aprovechar esos datos para estimar parámetros estructurales puede ser más simple que la estimación de un modelo de ecuaciones simultáneas más grande con variables de tratamiento endógenas. También es posible que la variable de tratamiento en un experimento natural pueda considerarse exógena, pero el tratamiento en sí no se asigna al azar.

La identificación puede verse amenazada por la presencia de un gran número de parámetros molestos. Por ejemplo, en un modelo de regresión de sección transversal, la función media condicional E [yi | xi] puede involucrar un efecto fijo específico individual αi, que se supone que está correlacionado con el error de regresión. Este efecto no puede identificarse sin muchas observaciones sobre cada individuo (es decir, datos de

panel). Sin embargo, con solo un panel corto, podría eliminarse mediante una transformación del modelo. Otro ejemplo es la presencia de variables exógenas no observadas invariantes en el tiempo que pueden ser comunes a grupos de individuos.

Cuando se omiten las variables de una regresión, y cuando los factores omitidos se correlacionan con las variables incluidas, se produce un sesgo. Por ejemplo, en una regresión con los ingresos como variable dependiente y la escolaridad como variable explicativa, la capacidad individual puede considerarse una variable omitida porque normalmente sólo se dispone de aproximaciones imperfectas. Esto significa que posiblemente no se pueda identificar el coeficiente de la variable escolaridad. Una posible estrategia es introducir variables de control en el modelo; el enfoque general se denomina enfoque de función de control. Estas variables son un intento de aproximar la influencia de las variables omitidas. Por ejemplo, varios tipos de puntuaciones sobre el rendimiento escolar podrían servir como controles de capacidad individual.

Si la identificación está en peligro porque la variable de tratamiento es endógena, entonces una solución estándar es utilizar variables instrumentales válidas. Esto es más fácil de decir que de hacer, como veremos en el tema dedicado a este tipo de estimación. La elección de la variable instrumental así como la interpretación de los resultados obtenidos debe hacerse con cuidado porque los resultados pueden ser sensibles a la elección de instrumentos.

La inferencia basada en muestras sobre la población solo son válidas si los datos de la muestra son representativos de la población. El problema de la selección muestral o el denominado sesgo muestral surge cuando los datos de la muestra no son representativos, en cuyo caso no se identifican los parámetros de la población. Existen técnicas de reponderación de la muestra que corrijan los sesgos, tal y como se ha visto en los temas de muestreo de poblaciones.

1.8 Datos observacionales

Tanto en macroeconometría como en microeconometría, los datos por lo general son observacionales (es decir, no proceden de un experimento controlado, si no de la observación). Los datos observaciones y los experimentales son distintos porque, en principio, dentro de un entorno experimental puede supervisarse y controlarse de cerca. Esto hace posible variar una variable causal de interés, manteniendo otras covariables en entornos controlados. Por el contrario, los datos observacionales se generan en un entorno no controlado, dejando abierta la posibilidad de que la presencia de factores de confusión dificulte la identificación de la relación causal de interés. Este hecho diferencial es fundamental dado que condiciona (y diferencia) la gran mayoría de las técnicas de estimación, así como su alcance causal.

Una fuente de datos observacionales son las encuestas de hogares, empresas y datos administrativos del gobierno. Los datos del censo también se pueden utilizar para generar muestras. Muchas otras muestras se generan a menudo en los puntos de contacto entre las partes que realizan la transacción. Por ejemplo, los datos de marketing se pueden generar en el punto de venta y / o encuestas entre compradores (reales

o potenciales). Internet (por ejemplo, subastas en línea) también es una fuente de datos.

El término dato observacional generalmente se refiere a datos de encuestas recopilados mediante el muestreo de la población relevante de sujetos sin ningún intento de controlar las características de los datos muestreados. Sea t el subíndice de tiempo, y **w** un conjunto de variables de interés. En este contexto, t puede ser un punto en el tiempo o un intervalo de tiempo. Sea S_t una muestra de la distribución de probabilidad poblacional $F\left(w_t \mid \boldsymbol{\theta}_t\right)$; S_t es un dibujo de $F\left(w_t \mid \boldsymbol{\theta}_t\right)$, donde ϑ es un vector de parámetros.

La población debe considerarse como un conjunto de puntos con características de interés y, por simplicidad, asumimos que se conoce la forma de la distribución de probabilidad F. Un esquema de muestreo aleatorio simple permite que todos los elementos de la población tengan la misma probabilidad de ser incluidos en la muestra. Se pueden considerar no obstante esquemas de muestreo más complejos.

El abstracto concepto de **población estacionaria** proporciona un punto de referencia útil. Si los momentos de las características de la población son constantes, entonces podemos escribir $\boldsymbol{\theta}_t = \vartheta$, para todo t. Este es un supuesto fuerte porque implica que los momentos de las características de la población son invariantes en el tiempo. Por ejemplo, la distribución por edad y sexo debe ser constante. De manera más realista, algunas características de la población no serían constantes. Para manejar tal posibilidad, (los parámetros de) cada población pueden considerarse como una extracción de una **superpoblación** con características constantes. Específicamente, pensemos en cada $\boldsymbol{\theta}_t$ como una extracción de una distribución de probabilidad con un (hiper)parámetro constante ϑ.

Como punto de referencia para los posteriores temas, consideraremos el **muestreo aleatorio simple** en el que la probabilidad de la unidad de muestreo i de una población de tamaño N, con N grande, es $1/N$ para todo i.

Particionemos **w** en [y: **x**] y supongamos que nuestro interés está en modelar y, una variable de resultado condicionado al vector exógeno de covariables **x**, cuya distribución conjunta se denota $f_J(y, \mathbf{x})$. Esto siempre se puede factorizar como el producto de la distribución condicional $f_C(y|\mathbf{x}, \boldsymbol{\theta})$ y la distribución marginal $f_M(\mathbf{x})$

$$f_J(y, \mathbf{x}) = f_C(y|\mathbf{x}, \boldsymbol{\theta}) f_M(\mathbf{x})$$

El muestreo aleatorio simple implica extraer las combinaciones (**y**, **x**) uniformemente de toda la población.

Lo interesante del muestreo es que si se extrae una muestra aleatoria, entonces la distribución de probabilidad de los datos es la misma que la distribución de la población. Sin embargo, hay ciertas desviaciones del muestreo aleatorio que provocan una divergencia entre los dos; esto se conoce como **sesgo muestral** o **sesgo de muestreo**. La distribución de datos difiere de la distribución de la población de una manera que depende de la naturaleza de la desviación del muestreo aleatorio. La desviación del muestreo aleatorio se produce porque a veces es más conveniente o económicamente rentable obtener los datos de una subpoblación aunque no sean representativos de toda la población. Ahora

consideramos varios ejemplos de tales desviaciones, comenzando con un caso en el que no hay desviación de la aleatoriedad.

En efecto, el caso 'ideal' sería el (sub)muestreo exógeno. Sucede cuando el analista segmenta la muestra disponible en submuestras basándose únicamente en un conjunto de variables exógenas **x**, pero no en la variable de respuesta y. Por ejemplo, la clasificación por categorías de ingresos, género, salud o nivel socioeconómico.

Bajo los supuestos de muestreo exógeno, la distribución de probabilidad de las variables exógenas es independiente de y, y no contiene información sobre los parámetros poblacionales de interés, ϑ. Por lo tanto, se puede ignorar la distribución marginal de las variables exógenas y simplemente basar la estimación en la distribución condicional $f(y|\mathbf{x}, \boldsymbol{\theta})$.

Si la suposición es incorrecta y la distribución observada de la variable de resultado puede depender de la variable de segmentación seleccionada, que puede estar correlacionada con el resultado, entonces nos alejaríamos del muestreo exógeno. Veamos algunos casos especialmente habituales:

Considere, por ejemplo, que en un estudio de los determinantes del número de visitas a un sitio recreativo solo se incluyen aquellos con al menos una visita. En este caso la probabilidad de que un individuo sea incluido en la muestra depende de las respuestas o elecciones hechas por ese individuo. De modo que las elecciones están relacionadas con la propia variable endógena y se produce un sesgo muestras causado por la respuesta (**response-based biased**). Se necesitaría una muestra aleatoria muy grande para generar suficientes observaciones (información) sobre un resultado o elección relativamente poco frecuente y, por lo tanto, es más barato recolectar una muestra de aquellos que realmente han hecho la elección. La importancia práctica de esto es que la estimación coherente de los parámetros de población ϑ ya no se puede realizar utilizando únicamente la densidad de población condicional $f(y|\mathbf{x})$. El efecto del esquema de muestreo también debería tenerse en cuenta.

Otra situación es el generado por el muestreo con sesgo por extensión o longitud (**lenght-biased**) muestral. Aparece cuando en el muestreo sobre una población es usado para hacer inferencias sobre una población diferente. Por ejemplo, supongamos que deseamos conocer el tiempo medio de transición del estado "desempleada" al estado "empleada". La clave es observar de dónde procede la muestra. Un posiblidad es que provenga de individuos que están desempleados en una fecha particular, otra por ejemplo es que provenga de la fuerza laboral (independientemente de su estado actual), y otra es que se muestree del grupo de personas que están en transición. Así pues, podríamos sesgar los resultados del tiempo medio fácilmente en función de dónde proceda la muestra.

Veamos una tercera fuente relacionado con el muestreo, y conocido como **sesgo de selección muestral**. Considere que una investigadora está interesada en medir el efecto de la formación o capacitación para alguna labor, denotado por z (tratamiento), sobre los salarios posteriores a la formación, denotado por y (resultado), dadas las características del trabajador, denotado por x.

La variable z toma el valor 1 si el trabajador ha recibido capacitación y es 0 en caso

contrario. Las observaciones están disponibles en (x, D) para todos los trabajadores, pero para y solo están disponible para aquellos que recibieron capacitación (D = 1). La investigadora está interesada en hacer inferencia sobre el impacto promedio de la capacitación en el salario posterior al proceso de formacion o capacitación de un trabajador elegido al azar con características conocidas y que actualmente no ha sido formado o capacitado (D = 0). El problema de la selección muestral, y por tanto su sesgo, se refiere a la dificultad de hacer tal inferencia. El objetivo es por tanto establecer la siguiente probabilidad:

$$Pr[y|x] = Pr[y|x, D = 1]Pr[D = 1|x] + Pr[y|x, D = 0]Pr[D = 0|x]$$

El proceso de muestreo puede identificar tres de los cuatro términos del lado derecho, pero no proporciona información sobre el término $Pr[y|x, D = 0]$, y por tanto será necesario establecer algún tipo de supuesto sobre $Pr[y|x]$ para poder inferir algo sobre el comportamiento medio, es decir, sobre $E[y|x]$:

$$E[y|x] = E[y|x, D = 1] \cdot Pr[D = 1|x] + E[y|x, D = 0] \cdot Pr[D = 0|x].$$

Esta lista de sesgos se puede completar con los problemas derivados de la calidad de las respuestas. Los datos observacionales a menudo provienen de encuestas. Las encuestas son normalmente voluntarias y el incentivo para participar puede variar sistemáticamente según las características de la entidad (por ejemplo, un hogar) y del tipo de pregunta que se haga. Las personas pueden negarse a responder algunas preguntas. Si existe una relación sistemática entre la negativa a responder una pregunta y las características del individuo, entonces surge el problema de la representatividad de una encuesta si el cuestionario permite la ausencia de respuesta. Si se ignora la falta de respuesta, y si el análisis se lleva a cabo utilizando únicamente los datos de los encuestados, ¿cómo se verá afectada la estimación de los parámetros de interés? La falta de respuesta a la encuesta es un caso especial del problema de selección mencionado en la sección anterior. Ambos involucran muestras sesgadas.

Otro problema de las encuestas es que los encuestados que se enfrentan a un cuestionario extenso no necesariamente responderán todas las preguntas e incluso si lo hacen, las respuestas pueden ser deliberada o fortuitamente falsas. Suponga que la encuesta por muestreo intenta obtener un vector de respuestas denotado como $\mathbf{x}_i = (x_{i1}, ..., x_{iK})$ de N individuos, $i = 1, ..., N$. Consideremos que si un individuo no proporciona información sobre uno o más elementos de $\mathbf{x}_i$, se descarta el vector completo.

Entonces, el primer problema resultante de la falta de datos es que se reduce el tamaño de la muestra. El segundo problema potencialmente más serio es que los datos faltantes pueden conducir a sesgos similares al sesgo de selección. Si faltan datos de manera sistemática, entonces la muestra que queda por analizar puede no ser representativa de la población. Una forma de sesgo de selección puede ser inducida por cualquier patrón sistemático de falta de respuesta. Por ejemplo, es posible que los encuestados de altos ingresos no respondan sistemáticamente a las preguntas sobre ingresos. Por el contrario, si los datos faltan completamente al azar, descartar las observaciones incompletas reducirá la precisión pero no generará sesgos.

Estos dos problemas conviven con el propio error de medida derivadas del descuido, de la información deliberadamente errónea, del recuerdo defectuoso de eventos pasados, de la interpretación incorrecta de preguntas y de los propios errores en el procesamiento de datos. Una fuente más profunda de error de medición se debe a que la variable medida es, en el mejor de los casos, una proxy imperfecto del concepto teórico relevante.

Terminamos este apartado indicando los tipos más habituales de datos observacionales en la actualidad.

Los **datos de corte transversal** se obtienen observando **w**, para la muestra S_t para algún t. Aunque generalmente no es práctico muestrear todos los hogares en el mismo momento, los datos de corte transversal siguen siendo una instantánea de las características de cada elemento de un subconjunto de la población que se utilizará para hacer inferencias sobre la población. Si la población es estacionaria, las inferencias hechas sobre $\boldsymbol{\theta}_t$ utilizando S_t pueden ser válidas también para t.

Los **datos de cortes repetidos** se obtienen mediante una secuencia de muestras independientes S_t tomadas de $F(\mathbf{w}_t|\boldsymbol{\theta}_t)$, $t = 1, ..., T$. Debido a que el diseño de la muestra no intenta retener las mismas unidades en la muestra, se pierde información sobre la dependencia dinámica en el comportamiento. Si la población es estacionaria, los datos de cortes transversales repetidos se obtienen mediante un proceso de muestreo similar al muestreo con reemplazo de la población constante. Si la población no es estacionaria, las secciones transversales repetidas se relacionan de una manera que depende de cómo cambia la población a lo largo del tiempo. En tal caso, el objetivo sería hacer inferenciaa sobre los parámetros (hiper) constantes subyacentes.

Si existe una dependencia significativa entre el comportamiento pasado y actual, entonces se requieren datos longitudinales o de panel para identificar la relación de interés. Por ejemplo, las decisiones pasadas pueden afectar los resultados actuales; la inercia o la persistencia del hábito pueden explicar las compras actuales, pero dicha dependencia no puede modelarse si el historial de compras no está disponible. Ésta es una de las limitaciones impuestas por los datos de corte transversal.

Los **datos de panel o longitudinales** se obtienen seleccionando inicialmente una muestra S y luego ir recolectando observaciones para una secuencia de períodos de tiempo, $t = 1, ..., T$. Esto se puede lograr entrevistando a los sujetos y recolectando datos presentes y pasados al mismo tiempo, o rastreando a los sujetos una vez que hayan sido incluidos en la encuesta. Esto produce una secuencia de vectores de datos $\{\mathbf{w}_1, \ldots, \mathbf{w}_T\}$ que se utilizan para hacer inferencias sobre el comportamiento de la población o el de la muestra particular de individuos. La metodología adecuada en cada caso puede no ser la misma. Si los datos se extraen de una población no estacionaria, el objetivo apropiado debería ser la inferencia de (hiper) parámetros de la superpoblación.

Los datos observacionales recogido a través de un panel tienen muy buenas propiedades, sin embargo hay que ser consciente de los riesgos y limitaciones. La primera cuestión es la representatividad del panel. Los problemas de inferencia con respecto al comportamiento de la población utilizando datos longitudinales se vuelven más difíciles si la población no es estacionaria. Para analizar la dinámica del comportamiento,

retener los hogares originales en el panel durante el mayor tiempo posible es una opción atractiva. En la práctica, los conjuntos de datos longitudinales sufren el problema del "desgaste de la muestra", quizás debido a la "fatiga de la muestra". Esto simplemente significa que los encuestados no continúan respondiendo a los cuestionarios. Esto crea dos problemas: (1) El panel se desequilibra y (2) existe el peligro de que el hogar retenido no sea "típico" y que la muestra no sea representativa de la población. Tanto (1) como (2) compromenten la validez de la inferencia.

1.9 Datos de experimentos sociales

Decíamos anteriormente que en ciencias sociales abundan los datos observacionales en detrimento de los experimientales. En economía, los datos análogos a los datos experimentales provienen de experimentos sociales o de experimentos de "laboratorio" en pequeños grupos de participantes voluntarios que imitan el comportamiento de los agentes económicos. Los experimentos sociales son relativamente poco comunes y, sin embargo, los conceptos, métodos y datos experimentales sirven como punto de referencia para evaluar estudios econométricos basados en datos observacionales.

La característica central de la metodología experimental implica una comparación entre los resultados del grupo experimental seleccionado al azar que se somete a un "tratamiento" con los de un grupo de control (comparación). En un buen experimento, se ejerce un cuidado considerable en emparejar los grupos de control y experimentales ("tratados") y en evitar posibles sesgos en los resultados. Considere, por ejemplo, dos regiones o estados contiguos, uno de los cuales pone en marcha una política de salario mínimo diferente de la otra, creando las condiciones de un experimento natural en el que las observaciones del estado "tratado" se pueden comparar con las del estado "control".

Un experimento social implica variaciones exógenas en el entorno económico al que se enfrenta el conjunto de sujetos experimentales, que se divide en un subconjunto que recibe el tratamiento experimental y otro que sirve como grupo de control. El modelo de resultados potenciales es muy adecuado para este tipo de datos experimentales.

Los experimentos sociales están motivados por cuestiones de política sobre cómo reaccionarían los sujetos ante un tipo de política que nunca se ha probado y, por tanto, para la que no existen datos de respuesta observada. La idea de un experimento social es reclutar a un grupo de participantes dispuestos, algunos de los cuales se asignan al azar a un grupo de tratamiento y el resto a un grupo de control. La diferencia entre las respuestas de aquellos en el grupo de tratamiento, sujetos al cambio de política, y aquellos en el grupo de control, que no lo están, es el efecto estimado de la política. El elemento clave es la asignación aleatoria a ser parte del grupo de control o del de tratamiento.

Los experimentos aleatorizados (**randomized trials**) también permiten una mayor variación en las variables y parámetros de las políticas que los que están presentes en los datos de observación, lo que facilita la identificación y el estudio de las respuestas a los cambios en las políticas. Un experimento puede producir datos de corte transversal o

longitudinales, aunque las consideraciones asociadas a sus respectivos costes generalmente limitarán la dimensión de tiempo muy por debajo de lo que es habitual en los datos observacionales.

La ventaja clave proviene de los experimentos aleatorizados es que eliminan cualquier correlación entre las características observadas y no observadas de los participantes del programa. La contribución del tratamiento a la diferencia entre el resultado entre los grupos "tratamiento" y "control" puede estimarse sin sesgo de confusión, incluso si no se pueden controlar las variables de confusión. La presencia de correlación entre el tratamiento y las variables de confusión a menudo afecta los estudios observacionales y complica la inferencia causal. Por el contrario, un estudio experimental realizado en circunstancias ideales puede producir una estimación consistente de la diferencia promedio en los resultados de los grupos tratados y no tratados sin mucha complejidad computacional. Sin embargo, si un resultado depende del tratamiento, así como de otros factores observables, entonces deberíamos controlar por este último para mejorar en general la precisión de la estimación del impacto causal.

Las principales limitaciones de lo experimentos sociales son las siguientes. En la gran mayoría de ocasiones estos experimentos tienen un coste monetario y no monetario alto. El primero de ellos es evidente, dado que compromente muchos recursos, más aún cuando el experimento comportara decisiones que pudieran generar ganancias potenciales en los participantes. Los costes no monetarios son los derivados de aplicar sobre seres humanos metodologías desarrolladas para entidades no-humanas. No siempre es posible, no siempre es ético y no siempre los resultados se pueden entender en el mismo sentido que en el mundo experimental sobre no-humanos.

Un segundo problema es el de la selección muestral, que es relevante porque la participación es voluntaria. Por razones éticas, hay muchos experimentos que simplemente no se pueden realizar (por ejemplo, asignación aleatoria de estudiantes a años de educación). A diferencia de los experimentos médicos que pueden alcanzar el estándar de oro de un protocolo "doble ciego", en los experimentos sociales, los experimentadores y los sujetos saben si están en grupos de tratamiento o de control. Además, aquellos en los grupos de control pueden obtener tratamiento (por ejemplo, entrenamiento) de fuentes alternativas. Si la decisión de participar no está correlacionada con $\mathbf{x}$ o $\mathbf{u}$, el análisis de los datos experimentales se simplifica.

Un tercer problema es el desgaste de la muestra causado por los sujetos que abandonan el experimento después de que ha comenzado. Incluso si la muestra inicial fuera aleatoria, el efecto de la deserción no aleatoria bien puede conducir a un problema similar al sesgo de deserción propio de los datos de panel. Finalmente, está el problema del efecto Hawthorne. Los sujetos humanos, a diferencia de los objetos inanimados, pueden cambiar o adaptar su comportamiento mientras participaba en el experimento. En este caso, la variación en la respuesta observada en condiciones experimentales no puede atribuirse únicamente al tratamiento.

1.10 Datos de experimentos naturales

Es posible que los datos disponibles provengan de un "experimento natural". Un experimento natural ocurre cuando un subconjunto de la población está sujeto a una variación exógena en una variable, quizás como resultado de un cambio de política, que normalmente estaría sujeta a una variación de la variable endógena.

Supongamos que existe una intervención exógena (externa) que cambia x en un modelo simple $y = \beta_1 + \beta_2 x + u$. Ejemplos de tal intervención externa son las normas administrativas, la legislación generada y no anticipada por los agentes o instituciones, los eventos naturales como los nacimientos de gemelos, los cambios relacionados con el clima ... . La intervención exógena crea una oportunidad para evaluar su impacto comparando el comportamiento del grupo afectado tanto "antes" como "después" de la intervención, o con el de un grupo no afectado después de la intervención. Es decir, los grupos de comparación "naturales" son generados por el evento que facilita la estimación del β_2. La estimación se simplifica porque x puede tratarse como exógena.

Consideremos algunos ejemplos ilustrativos. Supongamos que estamos interesados en estudiar los efectos de la inmigración sobre el mercado laboral. Una de las preguntas que más interesan a los economistas, y también a la sociedad, es saber si la inmigración reduce los salarios. La teoría económica sugiere que al desplazarse a la derecha la curva de oferta de trabajo, ceteris paribus todo lo demás, se llegaría a una situación de equilibrio estable en la que los salarios (precios de trabajo) serían más bajos que antes. ¿Qué haríamos si pudiéramos realizar un experimento aleatorizado controlado? Un experimento para estimar el efecto sobre los salarios de la inmigración asignaría aleatoriamente un número diferente de inmigrantes (diferentes tratamientos del experimento) a distintos mercados de trabajo (sujetos del experimento), y luego mediría el efecto sobre los salarios (respuesta observada en el experimento) y compararía. Sin embargo, es evidente que por muchos motivos de distinta naturaleza esto no lo podemos hacer. Podemos por tanto pensar en un cuasiexperimento. De hecho el trabajo del economista David Card (1990) es un ejemplo de cuasiexperimento.

El levantamiento temporal de las restricciones sobre emigración desde Cuba en 1980, supuso un éxodo de cubanos hacia Miami, inmigrantes que pasaron a formar parte del mercado laboral de Miami. Este hecho institucional puntual, que constituye una fuente exógena (natural) de variación, fue utilizado por Card para estimar el efecto causal sobre los salarios. Para ello comparó la variación de los salarios de trabajadores poco cualificados en Miami con la variación de salarios de trabajadores similares en otras ciudades comparables (con Miami) de EE.UU. durante el mismo periodo. La conclusión fue que no hubo efecto estadísticamente significativo sobre los salarios de los trabajadores menos cualificados.

Otros temas que han sido analizados por economistas por medio de experimentos naturales son: efectos de legislación sobre salarios mínimos por medio de cambios en las normas estatales o federales en los EE.UU.; los efectos del tamaño familiar sobre las elecciones de las familias utilizando los partos gemelares como fuente exógena de variación; los efectos de los impuestos sobre la oferta de trabajo y la inversión examinando reformas impositivas; los efectos del seguro por enfermedad sobre la salud, la oferta

de trabajo y las ayudas con hijos dependientes, utilizando para ello ampliaciones de programas que han permitido ampliar la posibilidad de selección de nuevos sujetos; los efectos de las restricciones de liquidez sobre la inversión utilizando los cambios en los precios del crudo como shocks para el cashflow de filiales no dependendientes del petróleo; por citar algunos ejemplos clásicos.

El tratamiento econométrico habitual se denomina estimación de diferencias en diferencias, y se ve en otra parte del temario.

Bibliografía complementaria

Matilla-García, M et al. 2017. Econometría y Predicción. McGraw Hill

Stock J. and Watson J. Introducción a la econometría. Pearson.

Tema 2

Analisis de regresion con datos de seccion cruzada I

Este tema está elaborado como una adaptación de los capítulos 2 y 3:

Wooldridge. J. 4th Ed., Introductory Econometrics.

Así como de la bibliografía complementaria.

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al Órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

- Análisis de regresión con datos de sección cruzada.
- El estimador de Mínimos Cuadrados Ordinarios (MCO).
- Valor esperado y varianza el estimador MCO.
- Eficiencia.

2.1 Análisis de regresión con datos de sección cruzada

En la Economía nos interesa la relación entre dos o más variables económicas. Esto motiva que en Econometría nos concentramos en poblaciones con una dimensión, al menos, bivariante.

La herramienta econométrica más comúnmente aplicada es la estimación por mínimos cuadrados, también conocida como **regresión**. La técnica de mínimos cuadrados es una herramienta para estimar la media condicional de una variable (la **variable dependiente**) dado otro conjunto de variables (denominadas generalmente como **regresores**).

Así, en el caso bivariante, si Y y X son dos variables que representan alguna población, entonces estamos interesados en explicar Y en términos de X, o puesto en otros términos, nos interesa estudiar cómo varía Y ante variaciones en X. Un ejemplo elemental es que Y sea la tasa de ahorro de una familia y X la renta de dicha familia. Otro es que Y represente las ventas de un servicio/producto y X los gastos en publicidad de dicho servicio/producto.

Al tratarse de datos no-experimentales, consideraremos que nuestras variables son aleatorias. La variable «Y», que la denotaremos, generalmente, por Y, es una variable aleatoria que tendrá una distribución poblacional desconocida, y lo mismo sucederá para la variable «X». Desconocemos el valor de las mismas antes de que sean observadas.

Así pues, podemos utilizar las herramientas propias de la teoría de la probabilidad. Las variables aleatorias tienen una función de distribución conjunta, $F(X, Y)$, y también una función de densidad de probabilidad conjunta, $f(X, Y)$.

El objeto de interés es, no tanto el comportamiento aislado de cada una de las variables aleatoria, sino las relaciones que se establecen entre las dos variables. Esta relación será de naturaleza estocástica y por tanto queremos analizar conjuntamente el vector de variables aleatorias (X, Y). En particular nuestro interés será la (generalmente desconocida) **distribución poblacional conjunta** de X y de Y. Utilizaremos dicha distribución conjunta para conocer la posible relación entre ambas variables. En caso de que no hubiera relación entre cada una de las variables, diríamos que son estadísticamente independientes, y lógicamente vendría recogido en la distribución conjunta adecuadamente.

Una forma muy conveniente e informativa para capturar la potencial relación entre «Y» y «X» es a través de la **función de esperanza condicionada (FEC)**. Es importante condicionar puesto que se trata de variables aleatorias. En caso de variables deterministas hubieramos trabajado con la función de esperaza marginal.

La esperanza condicionada de Y a un nivel de X determinado recoge la media de la distribución de Y. La función de esperanza condicionada informa sobre la evolución de las medias a medida que cambia el nivel de X. Esta idea se puede expresar como sigue

$$\mathbb{E}(Y \mid X) = m(X).$$

Se observa que la $\mathbb{E}(Y \mid X)$ es aleatoria en la medida en que cambia con X. Como variable aleatoria será suceptible de tener esperanza matemática no condicionada, es decir,

$$\mathbb{E}\left[\mathbb{E}(Y \mid X)\right],$$

que es una media que se va calculando a partir de las esperanzas condicionadas que a su vez consideran los distintos valores que va tomando la variable X. Un resultado interesante e intuitivo a este respecto es la Ley de las Esperanzas Totales:

$$\mathbb{E}(Y) = \mathbb{E}\left[\mathbb{E}(Y \mid X)\right],$$

que nos indica que una media (no condicionada) puede escribirse como la media no condicionada del promedio de las funciones de esperanzas condicionadas.

Por otra parte, otro resultado operativo interesante es la propiedad de que cuando condicionas una función de X a X, puedes tratar la esperanza condicionada como una constante

$$\mathbb{E}(g(X) \mid X) = X.$$

El objetivo principal es estudiar la relación que se establece a nivel poblacional entre la variable Y y la variable X. Normalmente nos encontraremos que la variable objeto de estudio Y está relacionada no solo con X, sino con otras variables X_1, X_2,..., X_k, y entonces nuestro objetivo será explicar cómo varía «Y» ante cambios en alguna(s) de

las «k» variables explicativas. La lista de las «k» variables, con toda seguridad, no será una relación exhaustiva de las variables que expliquen el comportamiento de «Y», de manera que la relación entre «Y» y las «k» variables no será exacta o determinada, sino solo aproximada. La FEC es una forma excelente de resumir la relación entre la «Y» y todas las «k» variables. Son varios los motivos por los que la relación queda bien resumida.

La primera razón es que la FEC permite que cualquier variable aleatoria Y se pueda descomponer como la suma de dos términos

$$Y = \mathbb{E}[Y\,|X_1, X_2, ..., X_k] + e$$

donde el primer sumando es la FEC y recoge la parte de Y explicada por las «k» variables del vector $\mathbf{x} = (X_1, X_2, ..., X_k)$ y el segundo sumando que denotamos por e el cual, por un lado no está correlacionado (no covaría) por construcción con el vector $\mathbf{x}$, $\mathbb{E}[e\,|\mathbf{x}] = 0$; ni con ninguna función de $\mathbf{x}$, $\mathbb{E}[e \cdot h(\mathbf{x})] = 0$.

La segunda razón es que la FEC es la función que mejor sirve para predecir la variable «Y». Y la tercera razón por la que conviene usar la FEC es porque la esperanza de una variable aleatoria es un buen resumen de la información de dicha variable. Por estos motivos, la FEC es el eje central de construcción de los modelos econométricos.

La FEC será no lineal y también será desconocida. Es decir en general $\mathbb{E}[Y\,|X_1, X_2, ..., X_k]$ es no lineal, sin embargo, vamos a considerar que la podemos aproximar mediante una función lineal.

La relación promedio entre las variables a la derecha del signo igual (las $X_j, j = 1, ..., k$) y la variable Y se expresa adecuadamente mediante la FEC, y también se la denomina función de regresión poblacional (**FRP**):

$$\mathbb{E}[Y\,|X_1, X_2, ..., X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k, \tag{2.1}$$

que indica que el valor esperado de la variable Y condicionado a los valores que toman las variables explicativas $X_j, j = 1, ..., k$. Esta expresión nos indica que estamos considerando como forma de trabajo (de aproximación) que la la FEC es lineal, aunque no tiene porqué ser lineal. Esto es bastante razonable, ya que

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

es la mejor aproximación lineal que podemos hacer de la FEC, es decir, de $\mathbb{E}[Y\,|X_1, X_2, ..., X_k]$.

Esta importante propiedad es una de las principales razones por la que utilizamos la especificación lineal y que se denomina **modelo de regresión lineal**. Así pues, el modelo de regresión lineal lo interpretarmos como una aproximación lineal a la FEC. Esto enfatiza que el objetivo principal que perseguimos en el modelo de regresión lineal es estudiar hechos esenciales de la relación estadística de «Y» con las «X». Esto es, estamos centrados en la distribución de «Y» más que en su valor concreto.

Otro motivo importante por el que usamos la caracterización lineal de la FEC es que la forma o expresión lineal que le damos a la FEC nos asegura que captura el ***efecto parcial***

(efecto *ceteris paribus*) de cada una de las «k» variables sobre «Y»: el efecto esperado sobre Y de la variación de una variable (digamos, X_1) manteniendo constantes el resto de factores incluidos $(X_2, X_3, ..., X_k)$. De hecho, el coeficiente de la pendiente de X_1 o parámetro β_1 captura el efecto que X_1 tiene sobre Y *teniendo en cuenta* (controlando por) los otros factores explicitados en la relación. Esta interpretación se obtiene fácilmente si a partir de la FRP imaginamos una variación de X_1 por una cuantía ΔX_1, mientras que el resto de variables no varían (se mantienen constantes). El cambio de X_1 hará que cambie $\mathbb{E}[Y\,|X_1, X_2, ..., X_k]$ en una cierta cantidad $\Delta\mathbb{E}[Y\,|X_1, X_2, ..., X_k]$. El nuevo valor resultante será

$$\mathbb{E}[Y\,|X_1, X_2, ..., X_k] + \Delta\mathbb{E}[Y\,|X_1, X_2, ..., X_k] = \beta_0 + \beta_1\,(X_1 + \Delta X_1) + \beta_2 X_2 + \ldots + \beta_k X_k. \tag{2.2}$$

Si a esta expresión le restamos el valor esperado de Y cuando no hay cambios, entonces se obtiene

$$\Delta\mathbb{E}[Y\,|X_1, X_2, ..., X_k] = \beta_1\,(\Delta X_1)\,,$$

por lo que la expresión

$$\beta_1 = \frac{\Delta\mathbb{E}[Y\,|X_1, X_2, ..., X_k]}{\Delta X_1}$$

indica que el coeficiente poblacional β_1 es el efecto (cambio esperado) sobre $\mathbb{E}(Y|X)$ ante un cambio en X_1, manteniendo fijas X_j,$j = 2, 3, ..., k$.

El término constante «β_0» frecuentemente no es relevante en el análisis empírico, si bien hay algunas aplicaciones en las que sí lo es. Su interpretación es sencilla: es el valor esperado de Y, cuando $X_1 = X_2 = ... = X_k = 0$.

Dados estos elementos estamos en condiciones de formar el modelo, es decir, la forma particular con la que vamos a relacionar las «k» variables con «Y» es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon. \tag{2.3}$$

La ecuación anterior define lo que llamaremos **modelo de regresión lineal múltiple**

De singular importancia es la variable «ε» se denomina **término error** y representa todos los otros factores que además de X_1, X_2,..., X_k determinan el valor de la variable dependiente Y para una observación concreta que llamamos observación i, por lo que para cada observación i habrá un error ε_i. Es decir ε_i representa los diferentes factores, distintos de las variables explicativas X_{1i}, X_{2i},... de la Ecuación (2.3) que afectan a la variable dependiente Y_i para cada i.

Así pues tenemos todos los elementos esenciales del modelo que nos permite hacer análisis de regresión: la variable dependiente, Y, las variables exógeneas o independientes o regresosres, $X_j, j = 1, ..., k$, el término error ε, los parámetros de pendiente o de efectos parciales.

2.2 El estimador de Mínimos Cuadrados Ordinarios (MCO)

El objetivo de esta sección estimar los parámetros que aparecen en la relación poblacional indicada en la FEC:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k.$$

Deseamos por tanto estimar los coeficientes $\beta_j, j = 0, 1, ..., k$, que son los que nos permiten, como hemos visto, interpretar los efectos parciales de cada variable explicativa en el comportamiento esperado de la variable dependiente.

Podemos estimar $\beta_j, j = 0, 1, ..., k$ a partir de datos observados que incluyan los valores conjuntos de las variables $(Y, X_1, ..., X_k)$. Estas variables son aleatorias, y hemos de distinguirlas de las observaciones realizadas que serán un conjunto finito de n observaciones

$$\{(Y_i, X_{1i}, ..., X_{ki}); i = 1, ..., n\}$$

que denominamos **muestra**.

Desde el punto de vista del análisis econométrico trataremos esta muestra como realizaciones de un proceso aleatorio. En particular asumimos que las observaciones son muestras de una población subyacente idéntica común para todas ellas, y que podemos denotar como F. La muestra está formada por n observaciones, cada una de ellas de la forma $\{(Y_i, X_{1i}, ..., X_{ki}); i = 1, ..., n\}$. Obsérvese que el orden de los elementos de la muestra no tiene ninguna importancia.

Esta distribución común F se le suele denominar **población** o también **proceso generador de datos**. Una forma útil de pensar sobre el proceso generador de datos es como una población infinítamente grande de la que hemos extraído una muestra finita.

El modelo de regresión simple y múltiple se aplica perfectamente a partir de las observaciones o muestra de la forma $\{(Y_i, X_{1i}, ..., X_{ki}); i = 1, ..., n\}$.

Supongamos, inicialmente, que queremos estimar la función de regresión poblacional del modelo de regresión simple de la Ecuación:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i, \tag{2.4}$$

De cada una de ellas (de cada i) tenemos un valor observado para la variable Y, que denotamos Y_i, y otro para la variable X_1, que indicamos por X_{1i}. Queremos *estimar* los parámetros de la FEC, es decir, de $\beta_0 + \beta_1 X_1$. Estos coeficientes o parámetros poblacionales son desconocidos, y tendremos que utilizar los datos disponibles de ambas variables para estimarlos. Estimados los coeficientes por alguna técnica estadística, nos referiremos a ellos por «$\hat{\beta}_0, \hat{\beta}_1$». Esta notación resulta muy conveniente y cómoda puesto que simplemente observando si lleva o no acento «^» sabremos si nos referimos al coeficiente muestral o al poblacional. Diremos que estos coeficientes estimados serán los homólogos muestrales de los coeficientes poblacionales, β_0, β_1. Con ellos podremos explicitar la estimación de la (FMR) función de regresión muestral $\hat{\beta}_0 + \hat{\beta}_1 X_1$, que así mismo es homóloga a la función de regresión poblacional $\beta_0 + \beta_1 X_1$.

A partir de la FRM podemos obtener $\hat{Y}_i$, que denominaremos valor estimado de «Y_i» dado el valor que toma X_{1i}. También puede considerarse que $\hat{Y}_i$ es el valor de predicción de Y_i a partir de la recta de regresión estimada (es decir a partir de la FRM). La diferencia entre el valor observado y el valor estimado o previsto se denomina **residuo de la regresión**

$$\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i = Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_{1i}\right), \tag{2.5}$$

que es el homólogo muestral del término (poblacional) error ε_i. Obsérvese que la Ecuación (2.5) nos permite descomponer el valor observado como la suma del valor estimado (valor de predicción) y el residuo:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\varepsilon}_i = \hat{Y}_i + \hat{\varepsilon}_i. \tag{2.6}$$

Lo mismo sucede si consideramos k variables explicativas de la variable dependiente,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon. \tag{2.7}$$

En este caso la variable a explicar depende, y por tanto, varía en función del valor que tomen varias variables. Algo que, por otra parte, es perfectamente esperable para las variables de naturaleza económico-empresarial.

Entendemos que la aproximación que hacemos a la función de esperanza condicionada es una especificación del modelo que se hace mediante una función lineal que recoge k variables explicativas, además de una constante o intercepto:

$$\mathbb{E}[Y\,|X_1, X_2, ..., X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k.$$

Re-escribimos esta relación lineal de forma más compacta. Sea el vector

$$\mathbf{x} = \begin{pmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$$

y el vector coeficientes poblacionales

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

de modo que si efectuamos el producto $\mathbf{x}'\boldsymbol{\beta}$, donde la «′» indica la transpuesta del vector o matriz afectados por «′», se obtiene una forma más compacta de escribir la ecuación anterior

$$\mathbb{E}[Y\,|X_1, X_2, ..., X_k] = \mathbf{x}'\boldsymbol{\beta}.$$

El método de los mínimos cuadrados localiza el vector

$$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)'$$

que minimiza

$$\min_{\boldsymbol{\beta}} : \mathbb{E}\left[(Y - \mathbf{x}'\boldsymbol{\beta})^2\right] \tag{2.8}$$

El vector de coeficientes que resuelve el problema será aquel vector que satisfaga las condiciones de primer orden de derivadas parciales nulas.

Desarrollamos el cuadrado y calculamos la esperanza

$$\mathbb{E}\left[(Y - \mathbf{x}'\boldsymbol{\beta})^2\right] = \mathbb{E}\left(Y^2\right) - 2\boldsymbol{\beta}'\mathbb{E}\left(\mathbf{x}Y\right) + \boldsymbol{\beta}'\mathbb{E}\left(\mathbf{x}\mathbf{x}'\right)\boldsymbol{\beta}.$$

A continuación derivamos respecto de las variables que están en el vector de coeficientes e igualamos todas a cero:

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\beta}}\mathbb{E}\left[(Y - \mathbf{x}'\boldsymbol{\beta})^2\right]$$

Para hacer esta derivada vectorial conviene desarrollar los productos $\boldsymbol{\beta}'\mathbb{E}\left(\mathbf{x}Y\right)$ y $\boldsymbol{\beta}'\mathbb{E}\left(\mathbf{x}\mathbf{x}'\right)\boldsymbol{\beta}$, y luego derivar respecto de los elementos del vector columna $\boldsymbol{\beta}$, o alternativamente utilizar las técnicas de derivación matricial. El resultado será un vector columna, que como hemos dicho, igualamos al vector columna $\mathbf{0}$, para obtener la condición necesaria de mínimo.

$$\mathbf{0} = -2\mathbb{E}\left(\mathbf{x}Y\right) + 2\mathbb{E}\left(\mathbf{x}\mathbf{x}'\right)\boldsymbol{\beta}$$

que resolviendo arroja el vector poblacional que minimiza $\mathbb{E}\left[(Y - \mathbf{x}'\boldsymbol{\beta})^2\right]$

$$\boldsymbol{\beta} = \left[\mathbb{E}\left(\mathbf{x}\mathbf{x}'\right)\right]^{-1}\mathbb{E}\left(\mathbf{x}Y\right).$$

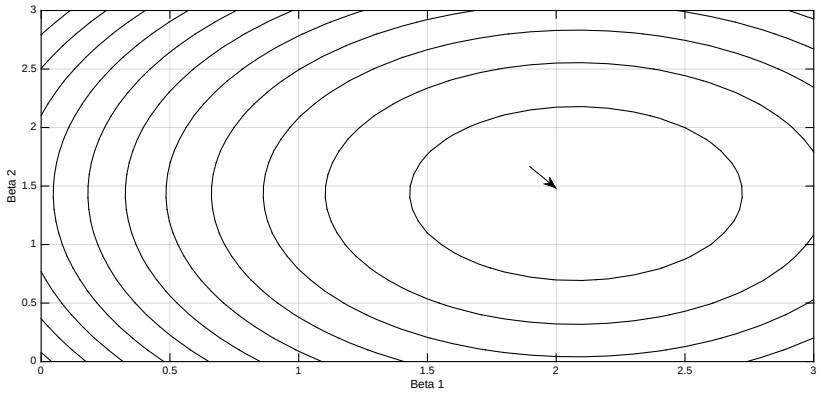
Varias observaciones son importantes:

- Esta expresión que resuelve el problema de minimización es poblacional, y no muestral (por ello no lleva ningún tipo de acento superior).
- El término $\mathbb{E}\left(\mathbf{x}\mathbf{x}'\right)$ es una matriz de orden $K \times K$ y el término $\mathbb{E}\left(\mathbf{x}Y\right)$ es un vector de orden $K \times 1$, por lo que $\boldsymbol{\beta}$ es un vector columna bien definido como producto. K es el número de filas del vector $\boldsymbol{\beta}$.
- Para que exista un único vector solución $\boldsymbol{\beta}$, es necesario que exista la inversa de la matriz $\mathbb{E}\left(\mathbf{x}\mathbf{x}'\right)$.

El modelo (2.7) no es observable directamente puesto que solo tenemos acceso a una muestra y no a la población. Siempre podemos definir el modelo estimable a partir de $\hat{\varepsilon}_i \equiv Y_i - \hat{Y}_i$, luego

$$Y_i = \hat{Y}_i + \hat{\varepsilon}_i. \tag{2.9}$$

Figura 2.1: Gráfico de contorno de $SCR(\beta_1, \beta_2))$



El objetivo es localizar los parámetros que permiten minimizar la suma de los cuadrados de los residuos, es decir

$$\min_{\beta_0,\beta_1,...,\beta_k} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - .... - \hat{\beta}_k X_{ki}\right)^2. \tag{2.10}$$

Conviene observar la similitud entre esta expresión muestral del proceso de minimización y la minimización de la expresión poblacional equivalente (2.8).

Se trataría de buscar entre todos los posibles vectores de $\boldsymbol{\beta}$, aquel vector que haga mínima la suma

$$SCR(\boldsymbol{\beta}) := \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$$

donde observamos que la suma cuadrática residual (SCR) es función de $\boldsymbol{\beta}$, puesto que lo demás son datos observados. Dada una colección formada por b vectores $\boldsymbol{\beta}_{(1)}, ..., \boldsymbol{\beta}_{(b)}$, tendremos una colección de sumas cuadráticas para cada correspondientes $SCR_{(1)}, ..., SCR_{(b)}$. Optaremos por aquella que sea mínima. En particular la Figura (2.1) muestra una gráfico de curvas de nivel donde cada curva de nivel SCR toma un valor constante a lo largo de la curva. En este caso, a medida que nos acercamos al punto indicado con la flecha los valores de la curva de nivel son cada vez más bajos. En el límite, el mínimo de ellos se obtiene para la combinación de valores $(\hat{\beta}_1, \hat{\beta}_2)$ correspondiente al estimador de mínimos cuadrados ordinarios. En este caso parece dar un vector próximo a $(\hat{\beta}_1 = 2, \hat{\beta}_2 = 1{,}5)$.

El tratamiento gráfico no es, lógicamente, el óptimo ni adecuado para resolver el problema (pese a ser instructivo). Utilizaremos en el cálculo diferencial de varias variables para resolverlo. Formalmente, el vector candidato a mínimo se consigue derivando respecto a cada parámetro e igualando a cero. Operando se llega a $k + 1$ **ecuaciones normales**:

$$\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - .... - \hat{\beta}_k X_{ki}\right) = \sum_{i=1}^{n} \hat{\varepsilon}_i = 0, \tag{2.11}$$

$$\sum_{i=1}^{n} X_{1i}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - .... - \hat{\beta}_k X_{ki}\right) = \sum_{i=1}^{n} X_{1i}\hat{\varepsilon}_i = 0, \tag{2.12}$$

$$.....$$

$$\sum_{i=1}^{n} X_{ki}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - .... - \hat{\beta}_k X_{ki}\right) = \sum_{i=1}^{n} X_{ki}\hat{\varepsilon}_i = 0, \tag{2.13}$$

que permiten deducir los $k+1$ coeficientes o parámetros de la regresión. La Ecuación (2.11) nos indica que la suma de los residuos es nula, por consiguiente su media también lo es ($\bar{\hat{\varepsilon}} = 0$). Puesto que $Y_i = \hat{Y}_i + \hat{\varepsilon}_i$ y la media de los errores es nula, se deduce que la media de la variable dependiente observada y la estimada son iguales ($\bar{Y} = \bar{\hat{Y}}$).

A partir de la Ecuación (2.11) dividiendo por n en ambas partes y realizando operaciones sencillas se llega a

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 + ... + \hat{\beta}_k \bar{X}_k. \tag{2.14}$$

Se observa que cuando la regresión pasa por las medias de las variables independientes, los errores se anulan (la relación es exacta en las medias).

En ocasiones resulta operativo considerar el mismo modelo pero centrado resepecto de sus medias. Para ello, si centramos en torno a su media a la variable dependiente Y_i:

$$Y_i - \bar{Y} = \hat{\beta}_1\left(X_{1i} - \bar{X}_1\right) + \hat{\beta}_2\left(X_{2i} - \bar{X}_2\right) + ... + \hat{\beta}_k\left(X_{ki} - \bar{X}_k\right) + \hat{\varepsilon}_i. \tag{2.15}$$

Si realizamos los cambios, $y_i = Y_i - \bar{Y}$ y $x_{ki} = X_{ki} - \bar{X}_i$ entonces

$$y_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki} + \hat{\varepsilon}_i, \tag{2.16}$$

donde se ha cancelado el término constante $\hat{\beta}_0$. La estimación mínimo cuadrática en desviaciones respecto de las medias es

$$\hat{y}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki}. \tag{2.17}$$

A partir de las ecuaciones normales (2.11) a (2.13) se despejan los parámetros $\hat{\beta}_j$.

A estos mismos estimadores también llegamos si derivamos vectorialmente

$$SCR(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 = \sum_{i=1}^{n} Y_i^2 - 2\boldsymbol{\beta}'\sum_{i=1}^{n}\left(\mathbf{x}_i Y_i\right) + \boldsymbol{\beta}'\sum_{i=1}^{n}\left(\mathbf{x}_i\mathbf{x}_i'\right)\boldsymbol{\beta}$$

y posteriormente igualamos a cero, es decir

$$\mathbf{0} = \frac{\partial}{\partial\boldsymbol{\beta}} SCR(\boldsymbol{\beta}) = -2\sum_{i=1}^{n}\left(\mathbf{x}_i Y_i\right) + 2\sum_{i=1}^{n}\left(\mathbf{x}_i\mathbf{x}_i'\right)\hat{\boldsymbol{\beta}}$$

ya que esta ecuación vectorial define justamente al sistema de **ecuaciones normales**, como podemos fácilmente comprobar

$$\sum_{i=1}^{n}\left(\mathbf{x}_i Y_i\right) = \sum_{i=1}^{n}\left(\mathbf{x}_i\mathbf{x}_i'\right)\hat{\boldsymbol{\beta}}$$

Podemos resolver el sistema por Cramer simplemente usando la inversa de la matriz $\sum_{i=1}^{n}(\mathbf{x}_i\mathbf{x}_i')$, es decir

$$\hat{\boldsymbol{\beta}}_{MCO} = \left[\sum_{i=1}^{n}(\mathbf{x}_i\mathbf{x}_i')\right]^{-1}\left[\sum_{i=1}^{n}(\mathbf{x}_i Y_i)\right]$$

Esta expresión es equivalente a la siguiente:

$$\hat{\boldsymbol{\beta}}_{MCO} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}. \tag{2.18}$$

donde $\hat{\boldsymbol{\beta}}_{MCO}$ es el vector columna de los parámetros estimados $\left\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k\right\}$.

Resulta ilustrativo fijarse en la analogía entre el vector solución poblacional y el muestral. Observemos que la solución poblacional es

$$\boldsymbol{\beta} = \left[\mathbb{E}\left(\mathbf{x}\mathbf{x}'\right)\right]^{-1}\mathbb{E}\left(\mathbf{x}Y\right)$$

que hemos obtenido por un proceso similar de minimización de $\mathbb{E}\left[(Y - \mathbf{x}'\boldsymbol{\beta})^2\right]$. Mientras que la solución muestral es

$$\hat{\boldsymbol{\beta}}_{MCO} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}$$

que se ha localizado minimizando la $SCR(\boldsymbol{\beta}) = \sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$. En la muestra se plantea que para cada n:

$$\begin{aligned} Y_1 &= \mathbf{x}_1'\boldsymbol{\beta} + \varepsilon_1 \\ Y_2 &= \mathbf{x}_2'\boldsymbol{\beta} + \varepsilon_2 \\ &... \\ Y_n &= \mathbf{x}_n'\boldsymbol{\beta} + \varepsilon_n. \end{aligned}$$

Si definimos

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

entonces

$$\mathbf{X}' = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & ... & \mathbf{x}_n \end{pmatrix}$$

y serán unas matrices de dimensiones $n \times (k+1)$ y $(k+1) \times n$, respectivamente. Podemos comprobar:

- (a) que el producto $\mathbf{x}_i\mathbf{x}_i'$ es una matriz de dimensiones $(k+1) \times (k+1)$,
- (b) que cada matriz $\mathbf{x}_i\mathbf{x}_i'$ está formada por los $(k+1) \times (k+1)$ productos de la forma $(x_{ji}x_{si})$, $j, s = 0, 1, .., k$,

- (c) que la expresión $\sum_{i=1}^{n}(\mathbf{x}_i\mathbf{x}_i')$ es por tanto la suma en i de los productos cruzados $(x_{ji}x_{si}), j, s = 0, 1, .., k$,
- (d) que dichas sumas están precisamente recogidas en cada una de las posiciones de la matriz $\mathbf{X}'\mathbf{X}$, la cual es también de dimensiones $(k+1)\times(k+1)$.
- (e) que la expresión $\sum_{i=1}^{n}(\mathbf{x}_i Y_i)$ es por tanto la suma en i de los productos cruzados $(x_{ji}Y_i), j = 0, 1, .., k$, y que están recogidos en el producto $\mathbf{X}'\mathbf{y}$, donde $\mathbf{y} = \begin{pmatrix} Y_1 & Y_2 & ... & Y_n \end{pmatrix}'$

En general las características algebraicas del estimador MCO son las siguientes:

- $\left(\sum_{i=1}^{n}\hat{y}_i\hat{\varepsilon}_i = 0\right)$, la estimación de la variable regresada «$\hat{Y}_i$» y los residuos «$\hat{\varepsilon}_i$» no están correlacionados, lo que implica que su covarianza es nula [cov $\left(\hat{Y},\hat{\varepsilon}\right) = 0$]:

$$\sum_{i=1}^{n}\hat{y}_i\hat{\varepsilon}_i = \hat{\mathbf{y}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{b}'\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{b}) =$$

$$= \boldsymbol{b}'\mathbf{X}'\mathbf{y} - \boldsymbol{b}'\mathbf{X}'\mathbf{X}\boldsymbol{b} = \boldsymbol{b}'\mathbf{X}'\mathbf{y} - \boldsymbol{b}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = 0$$

- Las variables independientes «X_{ji}» y los residuos «$\hat{\varepsilon}_i$» también están incorrelacionados [cov $(X_1,\hat{\varepsilon}) =$ cov $(X_2,\hat{\varepsilon}) = ... =$ cov $(X_k,\hat{\varepsilon}) = 0$]:

$$\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{X}'(\mathbf{y}-\mathbf{X}\mathbf{b}) = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = 0$$

- var (Y_i) = var $\left(\hat{Y}_i + \hat{\varepsilon}_i\right)$ = var $\left(\hat{Y}_i\right)$ + var $(\hat{\varepsilon}_i)$

$$var(\mathbf{y}) = var(\hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}) = var(\hat{\mathbf{y}}) + var(\hat{\boldsymbol{\varepsilon}}) + cov(\hat{\mathbf{y}},\hat{\boldsymbol{\varepsilon}}) = var(\hat{\mathbf{y}}) + var(\hat{\boldsymbol{\varepsilon}})$$

  donde la última igualdad se debe a que como hemos visto la covarianza es nula.

Esta última descomposición se utilizar para la elaboración del denominado R-cuadrado. El coeficiente de determinación o R^2 es una mera medida bondad del ajuste mínimo cuadrático. Existen otras mejores que requieren análisis estadístico (y no algebraico como en este caso):

$$R^2 = \frac{\text{var}\left(\hat{Y}_i\right)}{\text{var}\,(Y_i)} = 1 - \frac{\text{var}\,(\hat{\varepsilon}_i)}{\text{var}\,(Y_i)} = 1 - \frac{SCR}{SCT}. \tag{2.19}$$

donde *SCT* es la suma cuadrática de la variable dependiente en desviaciones a las medias, *SCE* es la suma cuadrática de la variable estimada en desviaciones a las medias y *SCR* es la suma cuadrática de los residuos estimados

En efecto, a partir de la definición de varianza podemos escribir

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n}SCE + \frac{1}{n}SCR = \frac{1}{n}SCT, \tag{2.20}$$

La suma cuadrática total es

$$\begin{aligned} SCT &= \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(Y_i^2 + \bar{Y}^2 - 2\bar{Y}Y_i\right) \\ &= \sum_{i=1}^{n} Y_i^2 + n\bar{Y}^2 - 2\bar{Y}\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} Y_i^2 + n\bar{Y}^2 - 2\bar{Y}n\left(n^{-1}\sum_{i=1}^{n} Y_i\right) \\ &= \sum_{i=1}^{n} Y_i^2 + n\bar{Y}^2 - 2n\bar{Y}^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = \mathbf{y}'\mathbf{y} - n\bar{Y}^2. \end{aligned} \tag{2.21}$$

A partir de la suma cuadrática de los residuos y, teniendo en cuenta la forma matricial de las ecuaciones normales, tenemos que

$$\begin{aligned} SCR &= \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \\ &= \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \text{ puesto que } \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \\ &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}. \end{aligned} \tag{2.22}$$

Sustituyendo (2.22) y (2.21), obtenemos la expresión matricial del coeficiente de determinación

$$\begin{aligned} R^2 &= \frac{SCT - SCR}{SCT} = \frac{\left(\mathbf{y}'\mathbf{y} - n\bar{Y}^2\right) - \left(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}\right)}{\mathbf{y}'\mathbf{y} - n\bar{Y}^2} \\ &= \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{Y}^2}{\mathbf{y}'\mathbf{y} - n\bar{Y}^2}. \end{aligned} \tag{2.23}$$

que es la expresión matricial del R-cuadrado.

Una desventaja del R-cuadrado en el contexto del modelo de regresión múltiple es que a medida que aumentamos el número de regresores X_j el coeficiente de determinación «R^2» necesariamente aumenta salvo que el coeficiente estimado sea *exactamente* nulo. Debido a esto, un incremento del R^2 no significa necesariamente que añadir una nueva variable realmente haya mejorado la calidad del ajuste de nuestro modelo. En realidad incluso si la nueva variable incluida en el modelo mejora nuestro ajuste, sabemos que necesariamente el R^2 de la nueva regresión estará artificialmente «inflado» por el mero hecho de incorporar un nuevo regresor. Por este motivo se utiliza el $\bar{R}^2$ **corregido**, que ajusta por el número de coeficientes estimados y cuya definición es

$$\overline{R^2} = 1 - \frac{SCR/_{n-k-1}}{SCT/_{n-1}} = 1 - \frac{\hat{\sigma}^2}{S_Y^2}, \tag{2.24}$$

donde se divide la suma cuadrática de los residuos por «n» menos el número de parámetros estimados «k+1», es decir, por «n –k – 1»; y la suma cuadrática total se divide por «n – 1». «$\hat{\sigma}^2$» es un estimador insesgado de la verdadera varianza de los residuos, «σ^2 » y «S_Y^2» es la varianza muestral de «Y».

En las ciencias sociales y particularmente con datos de sección cruzada, los R-cuadrado bajos en los modelos ajustados no son infrecuentes. Un R cuadrado aparentemente bajo no significa necesariamente que una ecuación de regresión estimada por MCO sea inútil. La capacidad de estimación de los efectos parciales de las variables explicativas sobre la variable dependiente no depende directamente del tamaño de R-cuadrado. Recordemos que el R-cuadrado es simplemente una estimación de cuánta variación en Y se explica por x1, x2, ..., xk en la población.

2.3 Valor esperado y varianza del estimador MCO

En el apartado anterior se ha visto una técnica de estimación del vector de parámetros del modelo de regresión lineal múltiple. Naturalmente esta no es la única técnica estadística disponible, sin embargo hay motivos que nos hacen pensar que es bastante buena si se cumplen ciertos requisitos que se exploran en este apartado. A partir de una muestra podremos, sin requerir especiales requisitos técnicos, lograr un vector $\hat{\boldsymbol{\beta}}$. Lo verdaderamente importante será conocer si dicho vector estimado es útil para saber los efectos parciales (poblacionales) que tienen cada una de las variables sobre la variable estimada. Recordemos que estos efectos parciales recogen la noción ceteris paribus (control por varios factores). En efecto, los parámetros β_j son los efectos parciales o efectos ceteris paribus de un cambio en la variable asociada al parámetro, es decir, X_j. ¿En qué condiciones esta interpretación de los parámetros es correcta? Es decir, en un contexto no experimental (donde no controlamos los demás factores distintos del j-ésimo), ¿cómo es factible actuar como si de un experimento controlado se tratase? Pues bien, bajo ciertos supuestos este objetivo es realizable.

Al conjunto de supuestos que nos conducen a una serie de propiedades deseables como la que hemos indicado, completan lo que se denomina el **modelo de regresión lineal múltiple**. Y se convierte por este motivo en la piedra angular de gran parte de la teoría econométrica, ya que plantea los supuestos poblacionales necesarios para que los estimadores muestrales (función de regresión muestral) cumplan una serie de propiedades deseables respecto de los verdaderos valores poblacionales (dados en la función de regresión poblacional).

***Supuesto 1. Linealidad en los parámetros**. El modelo poblacional puede escribirse como*

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i,$$

donde ε_i, $i = 1, ..., n$ es la sucesión de términos error de cada una de las observaciones.

Este supuesto indica que la población es compatible con una explicación lineal en la que el modelo estimado podrá diferir del modelo poblacional. La linealidad es respecto a los parámetros $\beta_j, j = 0, ..., k$, y por tanto no en las variables. Este hecho da bastante flexibilidad puesto que permite que las variables del modelo puedan ser funciones no lineales de variables subyacentes económicamente interesantes.

***Supuesto 2. Muestra aleatoria**. Tenemos una muestra aleatoria de n observaciones obtenidas del modelo poblacional del Supuesto 1:*

$$(X_{1i}, X_{2i}, ..., X_{ki}, Y_i), i = 1, ..., n \text{ son } i.i.d.$$

La muestra de datos que tenemos está formada por observaciones tomadas de modo que sean *independientes* unas de las otras y estén distribuidas *idénticamente* (provengan de la misma distribución conjunta). Este supuesto indica que las observaciones están en condiciones de asemejarse a una muestra aleatoria simple.

Supuesto 3. No multicolinealidad perfecta .

$$\mathbb{E}(\mathbf{x}_i\mathbf{x}_i') > 0, \text{ es decir, es una matriz definida positiva}$$

donde $\mathbf{x}_i = (X_{1i}, ..., X_{ki})'$.

El supuesto de no multicolinealidad perfecta permite que las variables independientes estén correlacionadas, pero no admite que estén exacta o perfectamente correlacionadas. Se trata de un requisito técnico que nos permitirá hacer la estimación MCO. Intuitivamente nos indica si una variable explicativa es una función lineal de otros regresores, en cuyo caso no se podrían calcular los coeficientes por MCO. Recordemos que una matriz $\mathbf{A}$ es definida positiva si para cualquier vector no nulo $\mathbf{a}$, se tiene que el escalar $\mathbf{a}'A\mathbf{a} > 0$. En este caso el supuesto implica que no existe ningún vector $\mathbf{a}$, tal que $\mathbf{a}'\mathbf{x}_i = 0$, ya que $\mathbf{a}'\mathbb{E}(\mathbf{x}_i\mathbf{x}_i')\mathbf{a} = \mathbb{E}(\mathbf{a}'\mathbf{x}_i\mathbf{x}_i'\mathbf{a})$, y dado que $(\mathbf{a}'\mathbf{x}_i\mathbf{x}_i'\mathbf{a}) = \sum (\mathbf{a}'\mathbf{x}_i)^2$ solo puede ser nulo o positivo; con lo que el supuesto excluye la nulidad. Otra forma de enunciar este supuesto es indicando que la matriz $\mathbf{X}$ es de rango completo.

En la práctica, la multicolinealidad perfecta aparece por problemas con el conjunto de datos que estamos manejando. Algunos de los motivos más habituales son los siguientes:

- Incluir el mismo regresor dos veces.
- Incluir un regresor que por confección de la base de datos está definido como una combinación lineal de otros regresores incluidos.
- Incluir una variable dummy (veremos más adelante que se trata de variables que valen o «cero» o «uno») y su cuadrado.
- Estimar una regresión sobre una submuestra en la cual una variable es o bien ceros o bien unos
- Incluir una variable dummy de interacción que arroja todo ceros.
- Incluir más regresores que observaciones
- ...

Obsérvese que la multicolinealidad no-perfecta o quasimulticolinealidad no viola ninguno de los supuestos, pero cuando la dependencia entre dos regresores ses aproxima a uno, entonces la varianza puede hacerse realmente grande, lo que implica una mayor varianza en el estimador de parámetro. Por tanto, cuando los regresores son altamente dependientes es estadísticamente difícil distinguir el impacto de β_j del de cualquier otro β_s. Decimos que la precisión del estimador individual se ve reducida.

No está definido cuándo la multicolinealidad es realmente un problema, es decir; no hay una regla fija o comúnmente aceptada sobre la importancia del problema. En todo caso y a efectos prácticos cabe decir que lo mejor es que la relación entre las variables regresoras sea pequeña, y que cuantas más observaciones tengamos mejor. Las soluciones no son fáciles. Aumentar el tamaño muestral recogiendo más datos sin duda ayudará, pero en ocasiones esto no está en la mano del analista de datos. Sustituir variables colineales

por variables proxies puede ser una solución en algunos casos. Eliminar variables del modelo entraña bastantes riesgos pues como veremos más adelante puede invalidar directamente el modelo.

Supuesto 4. Exogeneidad *. El valor esperado del término error es nulo para cualesquiera valores de las variables independientes*

$$\mathbb{E}\left(\varepsilon_i|\,\mathbf{x}_i\right)=0,\ i=1,2,...,n.$$

Este supuesto central indica que el valor medio de las variables incluidas en el error es el mismo independientemente los valores que tome el vector $\mathbf{x}_i$, es decir no varía con los niveles que tomen las variables explicativas. Por este motivo cuando el supuesto de exogeneidad se satisface, decimos que el error, $\boldsymbol{\varepsilon}$, es independiente en media de $\mathbf{x}$. Y además, siempre que el modelo incluya una variable constante, el valor medio será nulo.

Puede resultar útil pensar que en un experimiento controlado, la asignación de una determinada medida (por ejemplo, dar un tratamiento) a un sujeto se hace aleatoriamente de modo que el tratamiento esté administrado se distribuya independientemente de las características singulares del sujeto. En el análisis de regresión nuestro objetivo es modelizar la media condicionada, las variables en $\mathbf{x}$ no necesitan estar distribuidas independientemente de todos los factores que dejamos en el error $\boldsymbol{\varepsilon}$. Sin embargo sí es necesario que la media $\boldsymbol{\varepsilon}$ no esté relacionada con $\mathbf{x}$, es decir, $\mathbb{E}\left(\varepsilon_i|\,\mathbf{x}_i\right)=0$.

Si la media condicionada de una variable dada otra es nula, entonces ambas variables están tienen covarianza nula y están por tanto no correlacionadas, es decir $corr(\varepsilon_i,\mathbf{x}_i)=0$. Esto indica que no hay asociación lineal, y que en caso de que la hubiera, entonces el supuesto de exogeneidad sería violado.

Este supuesto no se satisface, por ejemplo, cuando omitimos en el error (por los motivos que fuere) alguna variable que esté correlacionada con alguna(s) de las variables incluidas en el vector $\mathbf{x}$. Hay otros motivos también importantes que producen la violación del supuesto de exogeneidad, y que serán tratados en otros temas. En todo caso, cuando se satisface este supuesto diremos que las variables explicativas contenidas en el vector $\mathbf{x}$ son exógenas.

Es importante observar que cuando se cumple el supuesto se tiene

$$\mathbb{E}\left(Y|\,\mathbf{x}_i\right)=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k. \tag{2.25}$$

ya que

$$\begin{aligned}\mathbb{E}\left(Y|\,\mathbf{x}_i\right) &= \mathbb{E}\left(\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_kX_{ki}+\varepsilon_i\mid\mathbf{x}_i\right)\\ &= \beta_0+\beta_1\mathbb{E}\left(X_{1i}\mid\mathbf{x}_i\right)+\beta_2\mathbb{E}\left(X_{2i}\mid\mathbf{x}_i\right)+...+\beta_k\mathbb{E}\left(X_{ki}\mid\mathbf{x}_i\right)+\mathbb{E}\left(\varepsilon_i\mid\mathbf{x}_i\right)\\ &= \beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k.\end{aligned}$$

Y por lo tanto si se cumple el supuesto (junto con los anteriores) los coeficientes $\beta_j, j=0,...,k$ recogen el efecto causal y efecto ceteris paribus de la variable X_j sobre el valor esperado de Y.

Otra consecuencia directa del supuesto de exogeneidad es que

$$\mathbb{E}\left(\varepsilon_i\right) = 0$$

para comprobarlo basta aplicar la Ley de las Esperanzas Totales

$$\mathbb{E}(\varepsilon_i) = \mathbb{E}\left[\mathbb{E}(\varepsilon_i \mid \mathbf{x}_i)\right] = 0.$$

Estos cuatro supuestos nos facilitan comprobar que el estimador MCO del vector de parámetro del modelo es un estimador insesgado.

TEOREMA. INSESGADEZ DE LOS PARÁMETROS MUESTRALES.
Bajo el supuesto de esperanza condicionada nula de los errores, los estimadores muestrales MCO son insesgados:

$$\mathbb{E}\left(\hat{\beta}_j\right) = \beta_j,\ j = 1, 2, ..., k, \tag{2.26}$$

$$\mathbb{E}\left(\hat{\beta}_j \,|\mathbf{X}\right) = \beta_j. \tag{2.27}$$

La primera ecuación del teorema indica que el estimador MCO es insesgado, es decir, que está centrado en torno al verdadero valor β_j. La segunda ecuación indica que el estimador es insesgado para cualquier realización de la matriz de regresores $\mathbf{X}$.

La demostración es la siguiente:

$$\begin{aligned}\mathbf{b} := \hat{\boldsymbol{\beta}}_{MCO} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon};\end{aligned}$$

haciendo la esperanza condicionada por $\mathbf{X}$, se tiene

$$\begin{aligned}\mathbb{E}(\mathbf{b}|\ \mathbf{X}) = \quad & \mathbb{E}[\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \mid \mathbf{X}] \\ = \ & \mathbb{E}[\boldsymbol{\beta} \mid \mathbf{X}]+\mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \mid \mathbf{X}] \text{ (recuérdese que } \beta \text{ no es aleatorio, por tanto)} \\ = \ & \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\boldsymbol{\varepsilon} \mid \mathbf{X}] \\ = \ & \boldsymbol{\beta} + \mathbf{0} = \boldsymbol{\beta} \text{ (por el supuesto de exogeneidad).}\end{aligned}$$

Por otra parte, la Ley de las esperanzas totales indica que $\mathbb{E}[\mathbb{E}(\mathbf{b} \mid \mathbf{X})] = \mathbb{E}(\mathbf{b})$, por lo que

$$\mathbb{E}(\mathbf{b}) = \boldsymbol{\beta}.$$

El estimador MCO, como vector aleatorio, del mismo modo que podíamos calcular su esperanza condicionada, también será susceptible de tener una varianza condicionada. Consideramos ahora dicha varianza.

Observamos inicialmente que la matriz de varianzas del vector error de regresión $\boldsymbol{\varepsilon}$ es la matriz $n \times n$ siguiente:

$$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'} = \mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \,|\mathbf{X}\right),$$

donde el elemento i-ésimo de la diagonal principal es

$$\mathbb{E}(\varepsilon_i^2 \,|\mathbf{x}_i) = \sigma_i^2,$$

que es la varianza de ε_i, mientras que los elementos fuera de la diagonal de la matriz $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'}$son

$$\mathbb{E}(\varepsilon_i \varepsilon_j \,|\mathbf{X}) = \mathbb{E}(\varepsilon_i \,|\mathbf{x}_i)\mathbb{E}(\varepsilon_j \,|\mathbf{x}_j) = 0,$$

al ser independientes (por el supuesto de muestra aleatoria) las observaciones j e i-ésimas.

La varianza condicionada del estimador MCO, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{A}'\mathbf{y}$, donde definimos $\mathbf{A}(\mathbf{X}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ será entonces

$$\begin{aligned}
\text{var}\left(\hat{\boldsymbol{\beta}} \,|\mathbf{X}\right) &= \text{var}\left(\mathbf{A}'\mathbf{y} \,|\mathbf{X}\right) \\
&= \text{var}\left(\mathbf{A}'\boldsymbol{\varepsilon} \,|\mathbf{X}\right) \\
&= \mathbf{A}'\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'}\mathbf{A} \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'}\mathbf{X}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}.
\end{aligned}$$

Donde apreciamos la dependencia de la varianza condicionada respecto de la matriz de varianzas y covarianzas del término error. En este caso, será una matriz de la forma

$$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'} = \mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \,|\mathbf{X}\right) = \begin{pmatrix} \sigma_1^2 & 0\ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0\ldots & \sigma_n^2 \end{pmatrix}$$

que nos indica que los **errores** son **heterocedásticos**. Alternativamente, si los errores fueran tales que $\sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2 = \sigma^2$, diríamos que los **errores** son **homocedásticos**, algo verdaderamente infrecuente (y por tanto solo teóricamente interesante) en los datos y modelos económicos. En tal caso,

$$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'} = \begin{pmatrix} \sigma^2 & 0\ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0\ldots & \sigma^2 \end{pmatrix} = \sigma^2\mathbf{I},$$

y la varianza condicionada del estimador MCO bajo homocedasticidad sería

$$\text{var}\left(\hat{\boldsymbol{\beta}} \,|\mathbf{X}\right)_{Homo} = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{2.28}$$

Tanto del caso heterocedástico como del homocedástico apreciamos que la varianza del estimador del parámetro β_j, digamos $\sigma^2_{\hat{\beta}_j}$, será menor cuanto mayor sea la varianza

de la variable X_j y cuanto menor sea la varianza del error σ_i^2. Por tanto, de cara a la precisión de las estimaciones realizadas por MCO respecto de las pendientes, preferiremos variables explicativas que tengan bastante variabilidad. Una forma de aumentar la variabilidad en las variables independientes es incrementando el tamaño muestral cuando esto es posible.

En todo caso será necesario estimar adecuadamente σ_i^2. Al fin y al cabo este parámetro mide la variación de la parte no explicada del modelo. El método de los momentos nos conduce a un estimador muestral obvio a partir de los residuos del modelo $\hat{\varepsilon}_i$

$$\widehat{\sigma^2} = \frac{1}{n}\sum \hat{\varepsilon}_i^2.$$

En el modelo de regresión esta suma de cuadrados es, donde definimos $M := \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ y aplicamos las propiedades de la traza:

$$\frac{1}{n}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \frac{1}{n}\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} = \frac{1}{n}tr(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}) = \frac{1}{n}tr(\mathbf{M}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}).$$

La esperanza condicionada entonces será

$$\begin{aligned}\mathbb{E}(\widehat{\sigma^2}|\mathbf{X}) &= \frac{1}{n}tr(\mathbb{E}\left(\mathbf{M}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}|\mathbf{X}\right)) = \frac{1}{n}tr(\mathbf{M}\mathbb{E}\left(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}|\mathbf{X}\right)) \\ &= \frac{1}{n}tr(\mathbf{M}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'})\end{aligned}$$

que en el caso de ser un error homocedástico se reduce a

$$\mathbb{E}(\widehat{\sigma^2}|\mathbf{X}) = \frac{1}{n}tr(\mathbf{M}\sigma^2\mathbf{I}) = \frac{1}{n}\sigma^2 tr(\mathbf{M}) = \frac{1}{n}\sigma^2(n-K)$$

donde usamos $tr(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = n - K$. Por tanto el estimador insesgado de la varianza de los errores en caso de homocedasticidad sería

$$s^2 = \frac{1}{n-K}\sum \hat{\varepsilon}_i^2.$$

En el caso heterocedástico lo veremos en otro apartado del temario de la oposición.

2.4 Eficiciencia

Aunque hemos dicho que la homocedasticidad es la excepción, hay un resultado que es interesante en sí mismo y se conoce como Teorema de Gauss-Markov:

TEOREMA DE GAUSS-MARKOV
Bajo los supuestos 1, 2, 3, 4 y considerando que los errores son homocedásticos, el estimador MCO es eficiente respecto de la clase de estimadores lineales insesgados. Esto es, para cualquier estimador insesgado $\hat{\boldsymbol{\beta}}$ lineal

$$\text{var}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) \geq \text{var}(\mathbf{b} \mid \mathbf{X}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}, .$$

donde $\mathbf{b} := \hat{\boldsymbol{\beta}}_{MCO}$ y $\hat{\boldsymbol{\beta}}$ es un estimador diferente del MCO

Demostración: Podemos escribir $\hat{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y}$ dado que es lineal en $\mathbf{y}$, donde $\mathbf{C}$ es una matriz que posiblemente es función de $\mathbf{X}$. Sea la matriz diferencia $\mathbf{D} \equiv \mathbf{C} - \mathbf{A}$, donde $\mathbf{A} \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, entonces

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{D}+\mathbf{A})\mathbf{y} &&= \mathbf{D}\mathbf{y} + \mathbf{A}\mathbf{y} \\
&= && \mathbf{D}(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) + \mathbf{A}(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \\
&= && \mathbf{D}\mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\varepsilon} + \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon} \quad [\text{porque } \mathbf{A}\mathbf{X} = \mathbf{0}] \\
&= && \boldsymbol{\beta} + \mathbf{D}\mathbf{X}\boldsymbol{\beta} + (\mathbf{D}+\mathbf{A})\,\boldsymbol{\varepsilon} \quad [\text{dado que hemos reordenado}].
\end{aligned}$$

Por tanto, la esperanza de $\hat{\boldsymbol{\beta}}$ condicionada a $\mathbf{X}$ será

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) &= \mathbb{E}(\boldsymbol{\beta} + \mathbf{D}\mathbf{X}\boldsymbol{\beta} + (\mathbf{D}+\mathbf{A})\,\boldsymbol{\varepsilon} \mid \mathbf{X}) \\
&= \boldsymbol{\beta} + \mathbb{E}(\mathbf{D}\mathbf{X}\boldsymbol{\beta} \mid \mathbf{X}) + \mathbb{E}(\mathbf{C}\boldsymbol{\varepsilon} \mid \mathbf{X}) \\
&= \boldsymbol{\beta} + \mathbf{D}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}\mathbb{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) \quad [\mathbf{C} \text{ y } \mathbf{D} \text{ son funciones de } \mathbf{X}] \\
&= \boldsymbol{\beta} + \mathbf{D}\mathbf{X}\boldsymbol{\beta} \quad [\text{por el supuesto de exogeneidad}].
\end{aligned}$$

Ahora bien, dado que el estimador $\hat{\boldsymbol{\beta}}$ debe ser, por el enunciado del teorema, insesgado, entonces ha de suceder que la matriz $\mathbf{D}$ sea tal que $\mathbf{D}\mathbf{X} = \mathbf{0}$. Así el estimador puede expresarse

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{C}\boldsymbol{\varepsilon},$$

por lo que $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{C}\boldsymbol{\varepsilon}$, y por tanto su matriz de varianzas y covarianzas condicionada será

$$\begin{aligned}
\text{var}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) &= \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mid \mathbf{X}] \\
&= \mathbb{E}[(\mathbf{C}\boldsymbol{\varepsilon})\,(\mathbf{C}\boldsymbol{\varepsilon})' \mid \mathbf{X}] \\
&= \mathbf{C}\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}]\mathbf{C}' \quad [\text{porque } \mathbf{C} \text{ es función de } \mathbf{X}] \\
&= (\mathbf{D}+\mathbf{A})\,\sigma^2\mathbf{I}\,(\mathbf{D}+\mathbf{A})' \quad [\text{por definición de C y por homocedasticidad}] \\
&= \sigma^2\left(\mathbf{D}\mathbf{D}' + \mathbf{A}\mathbf{A}' + \mathbf{A}\mathbf{D}' + \mathbf{D}\mathbf{A}'\right).
\end{aligned}$$

La matriz cuadrada producto $\mathbf{D}\mathbf{A}' = \mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}$ dado que, como se ha comprobado, $\mathbf{D}$ es tal que $\mathbf{D}\mathbf{X} = \mathbf{0}$; así, $(\mathbf{D}\mathbf{A}')' = \mathbf{A}\mathbf{D}' = \mathbf{0}$. Por otra parte es inmediato obtener que el producto $\mathbf{A}\mathbf{A}' = (\mathbf{X}'\mathbf{X})^{-1}$, por lo que

$$\begin{aligned}
\text{var}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) &= \sigma^2\left(\mathbf{D}\mathbf{D}' + (\mathbf{X}'\mathbf{X})^{-1}\right) \\
&\geq \sigma^2\left((\mathbf{X}'\mathbf{X})^{-1}\right) \quad [\text{dado que } \mathbf{D}\mathbf{D}' \text{ es semidefinida positiva}].
\end{aligned}$$

El teorema de Gauss-Markov ofrece claramente una justificación adicional para el uso de MCO. Sin embargo, el teorema tiene dos limitaciones severas. La primera es que los

supuestos bajo los que es cierto pueden fácilmente no satisfacerse en la práctica. Si el término error es condicionalmente heterocedástico, como ocurre en la mayoría de las aplicaciones en economía, entonces deja de ser el eficiente entre los lineales e insesgados. Como hemos dicho anteriormente, en el caso de tener errores heterocedásticos, si utilizamos errores estándar que fueran robustos a la heterocedasticidad, algo que se trata en otro tema, podremos realizar con garantías inferencias, pero entonces MCO ya no es el estimador óptimo (más eficiente). La segunda limitación es que incluso si las condiciones del teorema se cumplen, existen potencialmente otros estimadores que son no lineales e insesgados que podrían ser más eficientes que los MCO. Así pues la clase de modelos en los que es aplicable está restringida a regresiones homocedásticas y la clase de estimadores potenciales está restringida a estimadores lineales insesgados. Esta última restricción es particularmente insatisfactoria ya que no existe una motivación clara para centrarse en estimadores lineales.

De hecho es posible demostrar (si bien está fuera del alcance técnico de estas notas)

TEOREMA DE GAUSS-MARKOV GENERAL
Bajo los supuestos 1, 2, 3, 4 y considerando que los errores son homocedásticos, si $\mathbb{E}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \boldsymbol{\beta}$, entonces

$$\text{var}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) \geq \text{var}(\mathbf{b} \mid \mathbf{X}) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}, .$$

donde $\mathbf{b} := \hat{\boldsymbol{\beta}}_{MCO}$ y $\hat{\boldsymbol{\beta}}$ es un estimador diferente del MCO

Obsérvese que no se requiere que el estimador sea lineal.

Bibliografía complementaria

Matilla-García, M et al. 2017. Econometría y Predicción. McGraw Hill

Stock J. and Watson J. Introducción a la econometría. Pearson.

Tema 3

Analisis de regresion con datos de seccion cruzada II

Este tema está elaborado como una adaptación de los capítulos 4 y 5:

Wooldridge. J. 4th Ed., Introductory Econometrics.

Así como de la bibliografía complementaria.

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al Órgano convocante ni al Tribunal actuante.

- Distribución en el muestreo de los estimadores MCO.
- Intervalos de confianza y contrastes de hipótesis.
- Comportamiento asintótico del estimador mínimo cuadrático.
- Consitencia.
- Normalidad e inferencia asintótica MCO.
- Eficiencia.

El objeto ahora es hacer inferencia sobre los parámetros poblaciones. Sabemos que bajo ciertas condiciones (supuestos del tema anterior) relativamente bastante generales, el estimador MCO es insesgado y sabemos cuál es también su varianza condicionada, la cual tiene ciertas propiedades de optimalidad. Sin embargo, esto no es suficiente para conocer la distribución muestral de dicho estimador.

A los efectos de obtener la distribución muestral tenemos varias alternativas. Vamos a trabajar con dos posibles soluciones. La primera consiste en estudiar la distribución muestral del estimador MCO para cualquier tamaño muestral. Veremos que esto supone aumentar el número de supuestos del modelo, y por tanto perder se pierde generalidad. Alternativamente, la segunda solución es trabajar con resultados asintóticos, es decir, resultados que se alcanzan a medida que aumenta el tamaño muestral. Como veremos esta segunda alternativa requiere menos supuestos y por tanto es más general. En los primeros dos apartados trataremos la distribución muestral, mientras que los restantes apartados estarán dedicados a la distribución asintótica.

3.1 Distribución en el muestreo (muestral) de los estimadores MCO

SUPUESTOS DEL **MODELO DE REGRESIÓN LINEAL NORMAL**. Las observaciones $(Y_i, \mathbf{x}_i)$, $i = 1, 2, ..., n$, satisfacen la ecuación lineal de regresión

LINEALIDAD

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i. \tag{3.1}$$

MUESTRA ALEATORIA

$$(X_{1i}, X_{2i}, ..., X_{ki}, Y_i), i = 1, ..., n \text{ son } i.i.d.$$

EXOGENEIDAD

$$\mathbb{E}(\varepsilon_i \,|\mathbf{x}_i\,) = 0. \tag{3.2}$$

NO MULTICOLINEALIDAD PERFECTA

$$\mathbb{E}(\mathbf{x}_i\mathbf{x}_i') > 0, \text{ donde } \mathbf{x}_i = (X_{0i}, X_{1i}, ..., X_{ki})'. \tag{3.3}$$

NORMALIDAD Y HOMOCEDASTICIDAD CONDICIONADA

$$\varepsilon_i \,\big|\mathbf{x}_i \sim N(0, \sigma^2), i = 1, ..., n\ . \tag{3.4}$$

Estos supuestos implican que

$$Y_i \,|\mathbf{x}_i = \mathbf{x}_i'\boldsymbol{\beta} + N(0, \sigma^2).$$

Desafortunadamente cuando trabamajos con variables económicas y sus relaciones, tenemos pocos argumentos para considerar que la variable económica Y se distribuya como normal. Tampoco es realista considerar, como dijimos en el tema anterior, que los errores del modelo lineal sean homocedásticos. Desde este punto de vista, este apartado debe entenderse como un ejercicio instrutivo y didáctico para llegar a resultados más realistas que veremos en los dos últimos apartados de este tema.

Desde el punto de vista técnico, sin embargo, nos resulta muy cómodo trabajar con la normalidad. Los resultados se simplifican bastante. Así por ejemplo, se tiene el siguiente resultado:

Proposición. Bajo los supuestos del Modelo Regresión Lineal Normal, resulta que:

$$\begin{aligned} \mathbf{b}|\ \mathbf{X} &\sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}) \\ \mathbf{Xb}|\ \mathbf{X} &\sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{P}) \\ \mathbf{e}|\ \mathbf{X} &\sim N(\mathbf{0}, \sigma^2\mathbf{M}) \end{aligned} \tag{3.5}$$

donde $\mathbf{b} := \boldsymbol{\beta}_{MCO}; \mathbf{e} := \boldsymbol{\varepsilon}; \mathbf{P} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'; \mathbf{M} := \mathbf{I} - \mathbf{P}$

Demostración: La primera expresión se obtiene al aplicar el Teorema 2.3 y la Ecuación (2.28). La segunda es evidente a partir de que $\mathbb{E}(\mathbf{Xb}|\ \mathbf{X}) = \mathbf{X}\mathbb{E}(\mathbf{b}|\ \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$. La expresión de la varianza

se obtiene

$$\begin{aligned}
\text{var}(\mathbf{Xb}|\ \mathbf{X}) &= \text{var}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \mid \mathbf{X}) \\
&= \text{var}(\mathbf{Py} \mid \mathbf{X}) \quad \text{por definición } \mathbf{P} \\
&= \mathbf{P}\text{var}(\mathbf{y} \mid \mathbf{X})\mathbf{P}' \quad \text{al ser } \mathbf{P} \text{ dada} \\
&= \mathbf{P}\sigma^2\mathbf{I}_n\mathbf{P}' \\
&= \sigma^2\mathbf{P} \quad (\mathbf{P} = \mathbf{P}', \text{ y } \mathbf{P} \text{ es idempotente}) \\
&= \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.
\end{aligned} \tag{3.6}$$

Finalmente, dadas las propiedades de la matriz $\mathbf{M}$ y su relación con la matriz $\mathbf{P}$, se tiene que $\mathbf{e} = \mathbf{M}\boldsymbol{\varepsilon}$, por lo que es inmediato comprobar que $\mathbb{E}(\mathbf{e} \mid \mathbf{X}) = \mathbb{E}(\mathbf{M}\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{M}\mathbb{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$. La varianza la deducimos a partir de

$$\begin{aligned}
\text{var}(\mathbf{e}|\ \mathbf{X}) &= \text{var}(\mathbf{M}\boldsymbol{\varepsilon} \mid \mathbf{X}) \\
&= \mathbb{E}(\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{M} \mid \mathbf{X}) \ \text{(por } \mathbb{E}(\mathbf{M}\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0} \text{ y las propiedades de } \mathbf{M}) \\
&= \mathbf{M}\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X})\mathbf{M} \quad \text{(al ser } \mathbf{M} \text{ es función de } \mathbf{X}) \\
&= \mathbf{M}\sigma^2\mathbf{I}\mathbf{M} \quad \text{(por el supuesto de homocedasticidad)} \\
&= \sigma^2\mathbf{M}\mathbf{M} \quad \text{(al ser } \sigma^2 \text{ un escalar y por la propiedad de la matriz } \mathbf{I}) \\
&= \sigma^2\mathbf{M} \quad \text{(por la propiedad de idempotencia de } \mathbf{M}),
\end{aligned} \tag{3.7}$$

.

Podemos incluso derivar la distribución de s^2 a partir del supuesto de normalidad de los errores. Para ello recuérdese que $\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$. De acuerdo con el supuesto de normalidad de $\boldsymbol{\varepsilon}$, se tiene que $\boldsymbol{\varepsilon}/\sigma \mid \mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_n)$. Consideremos por tanto el producto $\frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}'}{\sigma}\mathbf{M}\frac{\boldsymbol{\varepsilon}}{\sigma}$. La matriz proyección $\mathbf{M}$ es idempotente. Por tanto el último producto representa una forma cuadrática que está sumando el cuadradado de variables normales independientes, siendo la suma ponderada por una matriz idempotente. Sabemos que una suma, en este caso ponderada por la matriz $\mathbf{M}$, de variables normales independientes es una distribución χ^2 con grados de libertad igual al rango de la matriz $\mathbf{M}$. También sabemos que $traza(\mathbf{M}) = rango(\mathbf{M})$ siempre que $\mathbf{M}$ sea idempotente. En consecuencia, y dado que $traza(\mathbf{P}) = K$, se tiene que $\frac{\mathbf{e}'\mathbf{e}}{\sigma^2} \sim \chi^2(n-K)$, es decir, una chi-cuadrado con $n-K$ grados de libertad. A partir de aquí se tiene:

Proposición. Bajo los supuestos del Modelo Regresión Lineal Normal, resulta que

$$s^2 \sim \frac{1}{n-K}\sigma^2 \sim \chi^2(n-K). \tag{3.8}$$

Por un lado observamos que los grados de libertad están en sintonía con el hecho de que utilicemos los residuos, y no los errores, para estimar la varianza. Si pudiéramos observar los errores del modelo, y dado el supuesto de normalidad de estos, se tendría que $\frac{\boldsymbol{\varepsilon}}{\sigma} \sim N(\mathbf{0}, \mathbf{I}_n)$ y en consecuencia $\frac{\boldsymbol{\varepsilon}'}{\sigma}\frac{\boldsymbol{\varepsilon}}{\sigma}$ se distribuiría como una chi-cuadrado con n grados de libertad. Por otro lado, esta proposición establece que las distribuciones marginal y condicionada de s^2 son idénticas dado que la distribución de s^2 dado $\mathbf{X}$ no depende de $\mathbf{X}$. Sin embargo, esto no ocurre con las distribuciones de $\mathbf{b}, \mathbf{Xb}$ y $\mathbf{e}$.

3.2 Intervalos de confianza y contrastes de hipótesis (en el Modelo Regresión Lineal Normal)

Consideremos inicialmente que estamos interesados en uno de los coeficientes de la regresión poblacional, por ejemplo y sin pérdida de generalidad, β_k. Es muy posible que dicho interés esté motivado en que la propia teoría económica indique la relevancia o el interés de que dicho coeficiente tome un valor en concreto. Por ejemplo, la teoría podría indicar una restricción de la forma $\beta_k = 1$. La estimación MCO de dicho coeficiente será b_k. La probabilidad de que $b_k = 1$ es cero, si bien la insesgadez del estimador nos garantiza que, en media y bajo los supuestos establecidos, será 1 si el parámetro poblacional lo es. Parece entonces razonable la decisión de no rechazar que la restricción sea cierta por el hecho de que la estimación no es idéntica a la unidad. Para tomar una decisión en ese sentido será necesario establecer cuándo la discrepacia o error muestral (esto es $b_K - 1$) es «tan grande» como para no dar por cierta a la restricción. Para saber si es «muy grande» o no, bajo ciertas circunstancias, es posible construir un intervalo de confianza o un test estadístico cuya distribución sea conocida cuando la restricción (o hipótesis sobre la población) es cierta. Este test o contraste nos permitirá decidir sobre si la estimación para nuestra muestra, esto es, b_k, está cerca (en términos estadísticos) del valor hipotético previsto por la teoría, es decir en este caso, 1. La restricción a ser contrastada se denomina *hipótesis nula* y se denota habitualmente por H_0. Bajo la H_0 junto con el conjunto de supuestos mantenidos en lo que denominamos modelo de regresión lineal normal, es posible obtener un contraste o test estadístico de distribución conocida.

El test estadístico es también una variable aleatoria que se distribuye según una distribución conocida cuando la hipótesis nula es cierta. Si el valor empírico que toma el test para una muestra concreta es un valor que aparece frecuentemente de acuerdo a la distribución del estadístico bajo la H_0, entonces el test o contraste decimos que no da muestras de ir contra la hipótesis nula, y por tanto no rechazaríamos dicha H_0. Lo contrario sucedería, esto es rechazaríamos H_0, si el valor que tomara el contraste fuera un valor extremo, es decir un valor que en raras ocasiones aparece en la distribución prevista bajo la hipótesis nula.

En el caso que nos ocupa del Modelo de Regresión Lineal Normal, para desarrollar la distribución del error muestral observemos inicialmente cuál es su expresión, que por conveniencia reescribimos a continuación

$$\mathbf{b} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon},$$

donde observamos que es una función de $(\mathbf{X}, \boldsymbol{\varepsilon})$ y además es lineal en $\boldsymbol{\varepsilon}$. Como ya hemos indicado en el epígrafe anterior, bajo el supuesto de normalidad sobre el término $\boldsymbol{\varepsilon}$, y dado que la combinación lineal de distribuciones normales es también una normal, el error muestral (errores de muestreo) también se distribuirá como una normal. Por tanto

$$(\mathbf{b} - \boldsymbol{\beta}) \mid \mathbf{X} \sim N(\mathbf{0}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}), \tag{3.9}$$

que, como vemos, hemos podido obtenerla sin especificar cuál es la distribución conjunta de $(\mathbf{X}, \boldsymbol{\varepsilon})$.

3.2.1 El test o contraste exacto de la t

Es habitual que estemos interesados en contrastar una hipótesis determinada sobre un coeficiente, digamos el coeficiente k, como en el ejemplo anterior. En ese caso la hipótesis nula sería del tipo

$$H_0 : \beta_k = \beta_k^0,$$

donde β_k^0 es cualquier valor que deseemos contrastar, por ejemplo, $\beta_k^0 = 0$ constituye en este caso la hipótesis nula H_0. Una forma de entender esta hipótesis nula es considerarla como una restricción lineal sobre el modelo. La hipótesis alternativa es $H_1 : \beta_k \neq \beta_k^0$, y se hará a un nivel de significación α determinado por el usuario.

A partir de la Ecuación (3.9) podemos obtener la distribución en caso de imponer la restricción que queremos contrastar. Por simplificar la notación nuevamente denotamos al estimador MCO del parámetro poblacional β_k por b_k, obteniéndose

$$\left(b_k - \beta_k^0\right) \mid \mathbf{X} \sim N(0, \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}]_{k,k}),$$

y entonces simplemente dividiendo por la desviación estándar se tiene la variable

$$z_k \equiv \frac{(b_k - \beta_k^0)}{\sqrt{\sigma^2[(\mathbf{X}'\mathbf{X})^{-1}]_{k,k}}} \sim N(0,1),$$

cuya distribución, por la forma en que la hemos construido, es la normal estándar.

Por tanto, z_k podría ser utilizado como test estadístico para contrastar H_0. Es decir, podría ser utilizado para saber si el error muestral $(b_k - \beta_k^0)$ es demasiado grande: esto sucede si el valor de z_k para la realización que tenemos del modelo resulta incompatible (por ser un valor extraño de acuerdo a la distribución prevista bajo la hipótesis nula) para un nivel de significación decidido anteriormente.

En cuanto al test conviene observar que su distribución [$N(0,1)$] no depende de **X**, por lo que la distribución marginal (es decir, la distribución no condicionada) y la distribución condicionada por **X** son la misma, pese a que z_k sí dependa de **X**. Por lo tanto, z_k y **X** se distribuyen de manera independiente y, con independencia del valor de **X**, la distribución de z_k es la misma y coincide por tanto con la no condicionada o marginal.

Otro hecho a favor del test o contrastes (o ratio) tipo t es que su distribución es conocida. En la práctica el cálculo del test estadístico depende de un parámetro desconocido σ^2, por lo que será necesario estimarlo previamente. Parece lógico utilizar a tal efecto el estimador $s^2 = \frac{SCR}{n-K} = \frac{\mathbf{e}'\mathbf{e}}{n-K}$.

En este último caso deberíamos sustituir, en la expresión de z_t, σ^2 por su versión estimable s^2. El denominador será ahora $[\widehat{\text{var}}(b_k)]^{1/2}$ que ya definimos como «error estándar de b_k». Sin embargo esta sustitución va a cambiar la distribución del test dado que s^2 es función de la muestra y por tanto es una variable aleatoria (a diferencia de σ^2 que es constante desconocida, y por tanto no aleatoria). Afortunadamente la distribución del nuevo contraste, que llamaremos t_k, es conocida, tal y como muestra la siguiente proposición.

Proposición. Bajo los supuestos del Modelo Regresión Lineal Normal, el estadístico tipo-t siguiente

$$t_k \equiv \frac{(b_k - \beta_k^0)}{\sqrt{s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{k,k}}} \tag{3.10}$$

se distribuye como una $t - student$ con $(n - K)$ grados de libertad.

Demostración. Reescribimos t_k del siguiente modo

$$\begin{aligned} t_k &= \frac{\left(b_k - \beta_k^0\right)}{\sqrt{\sigma^2[(\mathbf{X}'\mathbf{X})^{-1}]_{k,k}}}\sqrt{\frac{\sigma^2}{s^2}} \\ &= \frac{z_k}{\sqrt{s^2/\sigma^2}} \\ &= \frac{z_k}{\sqrt{\left(\frac{\mathbf{e}'\mathbf{e}}{n-K}\right)/\sigma^2}} = \frac{z_k}{\sqrt{\left(\frac{\mathbf{e}'\mathbf{e}}{\sigma^2}\right)/(n-K)}}. \end{aligned}$$

El cociente entre una variable $N(0,1)$ y la raíz de una variable chi-cuadrado dividida entre sus correspondientes grados de libertad tiene, por definición, una distribución t con dichos grados de libertad, siempre que las variables del numerador y del denominador sean independientes. Respecto del numerador de la última igualdad, z_k, hemos mostrado que es una N(0,1). Por otra parte, el denominador contiene a $\left(\frac{\mathbf{e}'\mathbf{e}}{\sigma^2}\right)$, por lo que a partir del resultado 3.8, se tiene que $\left(\frac{\mathbf{e}'\mathbf{e}}{\sigma^2}\right) \sim \chi^2(n-K)$. Solo resta comprobar que el numerador y el denominador son variables aleatorias independientes, dado **X**. Para ello obsérvese que z_k depende de $\mathbf{b}$, mientras que $\left(\frac{\mathbf{e}'\mathbf{e}}{\sigma^2}\right)$ dependen de $\mathbf{e}$. $\mathbf{b}$ y $\mathbf{e}$ se distribuyen condicionados en **X** como una normal conjunta dado que $\mathbf{b}$ y $\mathbf{e}$ son funciones lineales de $\boldsymbol{\varepsilon}$.

Bajo el supuesto de normalidad, dos variables no correlacionadas son independientes. En realidad este es el caso, ya que $\text{cov}(\mathbf{b}, \mathbf{e} \mid \mathbf{X}) = \mathbf{0}$ como mostramos a continuación:

$$\begin{aligned} \text{cov}(\mathbf{b}, \mathbf{e}|\ \mathbf{X}\ ) &= \mathbb{E}[(\mathbf{b} - \mathbb{E}(\mathbf{b} \mid \mathbf{X}))(\mathbf{e} - \mathbb{E}(\mathbf{e} \mid \mathbf{X})')|\ \mathbf{X}] \\ &= \mathbb{E}[\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right)(\mathbf{M}\boldsymbol{\varepsilon} - \mathbb{E}(\mathbf{M}\boldsymbol{\varepsilon} \mid \mathbf{X}))' \mid \mathbf{X}] \\ &= \mathbb{E}[\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right)(\mathbf{M}\varepsilon)' \mid \mathbf{X}] (\text{dado } \mathbb{E}(\mathbf{M}\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{M}\mathbb{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}) \\ &= \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{M} \mid \mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] \\ &= \mathbf{0}\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \mathbf{0}, \\ &\quad (\text{dado que } \mathbf{X}'\mathbf{M} = \mathbf{X}'(\mathbf{I}_n - \mathbf{P}) = \mathbf{X}' - \mathbf{X}'\mathbf{P} = \mathbf{X}' - \mathbf{X}'(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \mathbf{0}). \end{aligned}$$

Algo que intuitivamente en el resultado 3.8 pudimos comprobar: que los residuos MCO no son informativos sobre los parámetros de regresión $\boldsymbol{\beta}$. Por tanto, dado que z_k es función de $\mathbf{b}$, y $\left(\frac{\mathbf{e}'\mathbf{e}}{\sigma^2}\right)$ es función de $\mathbf{e}$, siendo $\mathbf{b}$ y $\mathbf{e}$ independientes entre sí, entonces también lo son el numerador y denominador de $t_k = \frac{z_k}{\sqrt{\left(\frac{\mathbf{e}'\mathbf{e}}{\sigma^2}\right)/(n-K)}}$.

Al test o contraste basado en este ratio lo denominaremos *test o contraste de la* t. En este caso lo utilizaremos para realizar inferencia con la intención de contrastar una hipótesis nula (H_0) sobre un coeficiente del modelo de regresión poblacional. Para ello es preciso

establecer el nivel de significación α, que indica la probabilidad de rechazar la hipótesis nula cuando esta es cierta (es decir, obtener un falso negativo para H_0).

La distribución t-student está centrada en 0 y es simétrica. La regla de decisión del test de la t consiste, en términos generales, en verificar si el valor observado del estadístico (3.10), para la hipótesis y muestra concreta, está muy alejado de 0. Bajo la H_0 la distribución es una t-student con $(n-K)$ grados de libertad, por tanto podemos localizar en las tablas correspondientes aquellos valores (valores críticos) que delimitan el área establecida en el nivel de significación α. Estos valores críticos, por ser una distribución simétrica, serán simétricos y por tanto los podemos denotar sin generar confusión por $\pm t_{\alpha/2}(n-K)$, de modo que a la derecha de $t_{\alpha/2}(n-K)$ se delimite, por ejemplo, un área 0.025 (2.5 %) y a la izquierda de $-t_{\alpha/2}(n-K)$ un área simétrica de 0.025 (2.5 %), de modo que en este ejemplos el nivel sería del 5 %. De esta manera podemos indicar que si la H_0 es verdadera, entonces

$$\Pr(-t_{\alpha/2}(n-K) < t < t_{\alpha/2}(n-K)) = 1-\alpha.$$

Esto nos sirve para establecer el significado de *«estar alejado de 0»* y poder establecer la regla de decisión del test: no rechazar («aceptar») H_0 si $|t_k| < t_{\alpha/2}(n-K)$ ya que indica que el valor obtenido t_k para la muestra concreta es compatible con la distribución prevista bajo H_0. Rechazar la hipótesis nula en caso contrario.

Otra forma alternativa para realizar el contraste de la t es elaborando un **intervalo de confianza** para β_k^0. Cuando la H_0 se «acepta» (es más correcto decir «no se rechaza») estamos en la «región de aceptación», es decir, en

$$\left[-t_{\alpha/2}(n-K) < \frac{(b_k-\beta_k^0)}{\sqrt{\widehat{\text{var}}(b_k)}} < t_{\alpha/2}(n-K)),\right]$$

lo que es equivalente a

$$\left[b_k - t_{\alpha/2}(n-K)\cdot\sqrt{\widehat{\text{var}}(b_k)} < \beta_k^0 < b_k + t_{\alpha/2}(n-K))\cdot\sqrt{\widehat{\text{var}}(b_k)},\right]$$

que pone de manifiesto que el intervalo será más estrecho, cuanto menor sea el error estándar de b_k. El intervalo de confianza, que es aleatorio al ser función de los datos, se construye de modo que nos dé información sobre el rango de valores de β_k^0 que son consistentes, es decir aquellos para los cuales el test no rechaza la nula.

Finalmente podemos realizar el contraste de la t utilizando el conocido y ampliamente utilizado $p-valor$. Recuérdese que este valor indica precisamente el nivel más pequeño para el cual el test rechaza la H_0, es decir, el test rechaza para todos los niveles por encima del $p-valor$. Dicho de otra manera, si t_k tiene asociado un $p-valor$ determinado y denotado por $p-valor(t_k)$, entonces estamos soportando una probabilidad de cometer un Error Tipo I de $p-valor(t_k)$ cuando optamos por rechazar la hipótesis nula. En el caso de tests de dos colas tendremos

$$p-valor = 2\cdot\Pr(t > |t_k|),$$

al ser una distribución simétrica.

Es posible relacionar los tres métodos para contrastar la hipótesis nula. Por ejemplo, utilizando un nivel de significación determinado α, rechazaríamos cuando $p-valor(t_k) < \alpha$, y esto ocurre si y solo si $|t_k| > t_{\alpha/2}(n-K)$, por lo que la equivalencia es evidente. Por otro lado se observa que el hecho de indicar el $p-valor$ es más informativo para el usuario ya que permite intuir la fuerza del rechazo de la hipótesis nula: cuanto más próximo a cero esté el p-valor, con mayor claridad se estará rechazando la hipótesis nula.

3.2.2 Test o contraste de la F para varias restricciones

En muchas ocasiones resulta útil o puede interesar contrastar más de una restricción lineal. Supongamos que deseamos contrastar $r \leq K$ restricciones sobre los coeficientes de los regresores del modelo. Estas restricciones, que configurarán la hipótesis nula, H_0, podemos escribirlas a través de un simple sistema de ecuaciones lineales:

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}, \tag{3.11}$$

donde **R** y **r** toman valores previamente especificados de acuerdo con la hipótesis nula. Por ejemplo, si queremos contrastar que dos parámetros son iguales, digamos los dos últimos $\beta_K = \beta_{K-1}$, y que un tercer parámetro (β_{K-2}) toma valor cero, tendríamos que

$$\begin{aligned} \mathbf{R} &= \begin{bmatrix} 0 & \cdots & 0 & 1 & -1 \\ 0 & \cdots & 1 & 0 & 0 \end{bmatrix} \\ \mathbf{r} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \end{aligned}$$

por lo que, en general, **R** será una matriz de dimensiones (número de restricciones)$\times$(número de parámetros del modelo), que denotamos por $(r)\times(K)$. Es evidente que r coincide con el rango de la matriz **R**, ya que de lo contrario habría restricciones redundantes.

Una vez que hemos establecido cómo son las restricciones, construimos un test estadístico que tenga una distribución *exacta* bajo la hipótesis nula descrita en (3.11).

A partir del resultado 3.8 se tiene que, bajo $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$,

$$\begin{aligned} \mathbf{Rb} \quad &| \quad \mathbf{X} \sim N(\mathbf{R}\boldsymbol{\beta}, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}') \\ (\mathbf{Rb}-\mathbf{r}) \quad &| \quad \mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}') \text{ (dado que bajo } H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}). \end{aligned} \tag{3.12}$$

por lo que sería posible, a priori, construir un test si reemplazáramos σ^2 por la varianza estimada s^2. La siguiente proposición recoge dicho resultado.

Proposición. Bajo los supuestos del Modelo Regresión Lineal Normal, y bajo la hipótesis nula $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, el cociente

$$F \equiv \frac{(\mathbf{Rb}-\mathbf{r})' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{Rb}-\mathbf{r}) /r}{s^2}, \tag{3.13}$$

conocido por test de la F, se distribuye como una F(r,$n-K$).

Como sucedía en el caso del test de la t, dado que la distribución de la F no depende de $\mathbf{X}$, la distribución condicionada y no condicionada coinciden, y por lo tanto basta con probar que la distribución condicionada por $\mathbf{X}$ se distribuye como indica la proposición anterior. Dividimos y multiplicamos por σ^2, y posteriormente usamos que $s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$, de modo que escribimos

$$\begin{aligned} F &= \frac{\sigma^2 (\mathbf{Rb}-\mathbf{r})' [\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{Rb}-\mathbf{r})/r}{s^2} \\ &= \frac{\sigma^2 w/r}{\mathbf{e}'\mathbf{e}/(n-K)} = \frac{w/r}{\left(\frac{\mathbf{e}'\mathbf{e}}{\sigma^2}\right)/(n-K)}, \end{aligned}$$

donde por simplificar la notación hacemos que $w \equiv (\mathbf{Rb}-\mathbf{r})' [\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{Rb}-\mathbf{r})$. Por el resultado (3.8) sabemos que $\left(\frac{\mathbf{e}'\mathbf{e}}{\sigma^2}\right) \mid \mathbf{X} \sim \chi^2(n-K)$. Falta por comprobar (a) que $w \mid \mathbf{X} \sim \chi^2(r)$ y (b) que $\left(\frac{\mathbf{e}'\mathbf{e}}{\sigma^2}\right)$ y w se distribuyen independientemente condicionados por $\mathbf{X}$.

Resultado (a): bajo H_0: $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, resulta por la expresión (3.12) que $\mathbf{Rb}-\mathbf{r}$ se distribuye como una normal con media $\mathbf{0}$, y

$$\text{var}(\mathbf{Rb}-\mathbf{r}|\ \mathbf{X}) = \sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}',$$

por lo que podemos reescribir $w = (\mathbf{Rb}-\mathbf{r})' [\text{var}(\mathbf{Rb}-\mathbf{r} \mid \mathbf{X})]^{-1} (\mathbf{Rb}-\mathbf{r})$ que es una expresión que suma r normales al cuadrado, al ser $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ una matriz cuadrada de orden r. Por tanto $w \mid \mathbf{X} \sim \chi^2(r)$.

Resultado (b): ya probamos para la Ecuación (3.10) que utilizando el supuesto de normalidad, $\mathbf{b}$ y $\mathbf{e}$ se distribuyen condicionados por $\mathbf{X}$ de forma independiente. Dado que w es una función de $\mathbf{b}$ y $\left(\frac{\mathbf{e}'\mathbf{e}}{\sigma^2}\right)$ lo es de $\mathbf{e}$, queda probada la independencia entre los dos.

La definición de una distribución F como cociente de dos variables aleatorias distribuidas como chi-cuadrado, divididas cada una de ellas por sus respectivos grados de libertad, concluye la demostración.

En este caso el test o contraste es de *una sola cola*. Si la hipótesis nula es verdadera, entonces $\mathbf{Rb}-\mathbf{r} = \mathbf{R}(\mathbf{b}-\boldsymbol{\beta})$ tenderá a tomar valores pequeños haciendo que el numerador de (3.13) sea también pequeño, y por tanto un valor alto del test F sería indicativo de un rechazo de la H_0. La regla de decisión es por tanto rechazar la hipótesis nula si el valor que toma el estadístico F es superior al valor crítico asociado al nivel de significación determinado de antemano.

El test (3.13) puede interpretarse a partir de la distinción entre el concepto de *regresión restringida* y *regresión no restringida*.

Recordemos que la técnica MCO consiste en minimizar SCR, y ahora el mínimo estará sujeto a un conjunto de restricciones lineales expresadas precisamente por la H_0 : $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. El problema de estimación MCO se transforma en este

$$\min_{\tilde{\beta}} SCR(\tilde{\boldsymbol{\beta}}) \text{ sujeto a } \mathbf{R}\tilde{\boldsymbol{\beta}} = \mathbf{r}. \tag{3.14}$$

La obtención del $\tilde{\boldsymbol{\beta}}$ que satisface el problema anterior se denomina *mínimos cuadrados restringidos* o *regresión restringida*. Denotaremos por $\hat{\boldsymbol{\beta}}$ al estimador restringido del parámetro resultado de resolver la Ecuación (3.14). A los efectos de solventar este problema

formaremos el Lagrangiano correspondiente a la optimización restringida

$$L(\tilde{\boldsymbol{\beta}}, \boldsymbol{\lambda}) = \frac{1}{2}\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right) + \boldsymbol{\lambda}'\left(\mathbf{R}\tilde{\boldsymbol{\beta}} - \mathbf{r}\right),$$

donde el vector λ de orden $(r \times 1)$ está formado por los *multiplicadores de Lagrange* del problema. Las condiciones de primer orden se obtienen a partir de desarrollar $L(\tilde{\boldsymbol{\beta}}, \lambda)$ y de igualar a cero sus derivadas parciales respecto de $\tilde{\boldsymbol{\beta}}, \lambda$:

$$L(\tilde{\boldsymbol{\beta}}, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\tilde{\boldsymbol{\beta}} + \frac{1}{2}\tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{\lambda}'\mathbf{R}\tilde{\boldsymbol{\beta}} - \boldsymbol{\lambda}'\mathbf{r},$$

derivando e igualando a cero se obtiene que los estimadores restringidos serán los $\hat{\boldsymbol{\beta}}$ que satisfagan las ecuaciones

$$\begin{aligned}
\frac{\partial L(\tilde{\boldsymbol{\beta}}, \boldsymbol{\lambda})}{\partial \tilde{\boldsymbol{\beta}}} &= \mathbf{0} \Leftrightarrow -\mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{R}'\boldsymbol{\lambda} = \mathbf{0} \\
&\Leftrightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\lambda} \qquad (3.15) \\
\frac{\partial L(\tilde{\boldsymbol{\beta}}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} &= \mathbf{0} \Leftrightarrow \mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r} = \mathbf{0}
\end{aligned}$$

de modo que premultiplicando la expresión (3.15) por $\mathbf{R}$ y usando la segunda ecuación (la restricción en sí) se tiene que

$$\begin{aligned}
\mathbf{r} &= \mathbf{R}\hat{\boldsymbol{\beta}} = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\lambda} \\
&\Leftrightarrow \mathbf{r} = \mathbf{R}\mathbf{b} - \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\lambda} \\
&\Leftrightarrow (\mathbf{R}\mathbf{b} - \mathbf{r}) = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\lambda} \\
&\Leftrightarrow \boldsymbol{\lambda} = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}),
\end{aligned}$$

lo que nos permite expresar (3.15) del siguiente modo:

$$\hat{\boldsymbol{\beta}} = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}).$$

Esta regresión restringida tendrá unos residuos diferentes de la regresión no restringida. De hecho, la suma del cuadrado de los residuos restringidos, SCR_R, será ahora

$$\begin{aligned}
SCR_R &= \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \\
&= \left[\mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X}\left(\mathbf{b} - \hat{\boldsymbol{\beta}}\right)\right]'\left[\mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X}\left(\mathbf{b} - \hat{\boldsymbol{\beta}}\right)\right] \\
&= \left[\mathbf{e} + \mathbf{X}\left(\mathbf{b} - \hat{\boldsymbol{\beta}}\right)\right]'\left[\mathbf{e} + \mathbf{X}\left(\mathbf{b} - \hat{\boldsymbol{\beta}}\right)\right] \\
&= \mathbf{e}'\mathbf{e} + \left(\mathbf{b} - \hat{\boldsymbol{\beta}}\right)'(\mathbf{X}'\mathbf{X})\left(\mathbf{b} - \hat{\boldsymbol{\beta}}\right) \quad (\text{pues } \mathbf{e}'\mathbf{X} = \mathbf{0})
\end{aligned}$$

y por tanto la diferencia entre la suma del cuadrado de los residuos restringidos, SCR_R,

y la suma de cuadrados no restringidos, SCR_{NR}, será

$$\begin{aligned} SCR_R - SCR_{NR} &= \left(\mathbf{b} - \hat{\boldsymbol{\beta}}\right)' \mathbf{X}'\mathbf{X} \left(\mathbf{b} - \hat{\boldsymbol{\beta}}\right) \\ &= \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})\right]' \times \\ &\quad (\mathbf{X}'\mathbf{X})\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})\right] \\ &= (\mathbf{R}\mathbf{b} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) \times \\ &\quad (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}) \\ &= (\mathbf{R}\mathbf{b} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}), \end{aligned} \tag{3.16}$$

donde hemos utilizado fundamentalmente el hecho de que las matrices $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ y $(\mathbf{X}'\mathbf{X})^{-1}$ son simétricas, junto con la propiedad de la inversa que indica que $[\mathbf{A}^{-1}]' = [\mathbf{A}']^{-1}$ siendo $\mathbf{A}$ una matriz invertible.

Observamos que a partir de la Ecuación (3.16) y de la definición de s^2, podemos expresar (3.13) del siguiente modo:

$$F = \frac{(SCR_R - SCR_{NR})/r}{SCR_{NR}/(n-K)} \tag{3.17}$$

que como ya probamos se distribuye como una $F(r,n-K)$. Luego tanto la expresión (3.17) como la expresión (3.13) arrojan el mismo resultado. Utilizar el test de *F* según (3.17) implica realizar dos regresiones (una con las restricciones activas y otras sin ellas), guardar los residuos y calcular el ratio descrito por (3.17). En cambio, en el caso del contraste (3.13) solo es necesario la regresión no restringida.

3.2.3 Un contraste de significación global

A menudo estamos interesados en contrastar la significatividad general del modelo, esto es, si las variables explicativas resultan en su *conjunto* estadísticamente significativas. Es posible mejorar la evaluación estadística de la bondad del ajuste al poder relacionarla con un contraste estadístico.

Dado que la forma más general de contrastar un conjunto de hipótesis sobre un modelo es a partir del test de la F presentado anteriormente, el primer paso será expresar la hipótesis nula en los términos que venimos usando ($\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$). El modelo no restringido será el modelo con una constante habitual, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$

$$\underset{n\times 1}{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}} = \underset{n\times K}{\begin{bmatrix} 1 & X_{12} & \cdots & X_{1K} \\ 1 & X_{22} & \cdots & X_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & \cdots & X_{nK} \end{bmatrix}} \cdot \underset{K\times 1}{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}} + \underset{n\times 1}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}},$$

mientras que el restringido será exactamente el mismo, esto es $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, solo que sujeto a la restricción descrita a continuación:

$$\overbrace{\underset{(K-1)\times(K)}{\left[\begin{array}{cc} \mathbf{0}_{(K-1)\times 1} & \mathbf{I}_{K-1} \end{array}\right]}}^{\mathbf{R}} \underset{K\times 1}{\left[\begin{array}{c} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{array}\right]} = \mathbf{r} = \mathbf{0}_{(K-1)\times 1}. \tag{3.18}$$

El estimador MCO no restringido es $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Por otra parte, el estimador MCO restringido por la condición de la Ecuación (3.18) será aquel $\hat{\boldsymbol{\beta}}$ que cumpliendo la restricción (es decir, $\mathbf{R}\hat{\boldsymbol{\beta}} = \mathbf{0}$) minimice la suma cuadrática de los residuos. Cumplir la restricción implica que $\hat{\beta}_2 = \hat{\beta}_3 = \cdots = \hat{\beta}_K = 0$, por lo que quedaría únicamente estimar por MCO el parámetro β_1 que como sabemos es $\bar{y}$ para un modelo con constante. Así resulta que $\hat{\beta} = [\ \bar{y} \quad 0 \quad \cdots \quad 0\ ]'$ y por tanto para esta restricción se tiene que $\mathbf{X}\hat{\boldsymbol{\beta}} = [\ \bar{y} \quad \bar{y} \quad \cdots \quad \bar{y}\ ]'$ y en consecuencia

$$SCR_R = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)' \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = (\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}}). \tag{3.19}$$

A partir de la Ecuación (3.16) podemos utilizar esta última expresión de la suma cuadrática de los residuos restringidos por la condición (3.18) para mostrar que

$$(\hat{\mathbf{y}} - \bar{\mathbf{y}})'(\hat{\mathbf{y}} - \bar{\mathbf{y}}) = SCR_R - SCR_{NR} = \left(\mathbf{b} - \hat{\boldsymbol{\beta}}\right)' \mathbf{X}'\mathbf{X} \left(\mathbf{b} - \hat{\boldsymbol{\beta}}\right).$$

Estos resultados específicos de la restricción (3.18) junto con la definición de R^2, nos permiten expresar el contraste de F en función de la bondad del ajuste. A tal efecto, reescribimos $F = \frac{(SCR_R - SCR_{NR})}{SCR_{NR}} \frac{(n-K)}{r}$, de modo que usando los resultados anteriores, también podemos escribir

$$\begin{aligned} R^2 &= \frac{(\hat{\mathbf{y}} - \bar{\mathbf{y}})'(\hat{\mathbf{y}} - \bar{\mathbf{y}})}{(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})} = 1 - \frac{\mathbf{e}'\mathbf{e}}{(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})} \\ &= \frac{SCR_R - SCR_{NR}}{SCR_R} = 1 - \frac{SCR_{NR}}{SCR_R}. \end{aligned} \tag{3.20}$$

Usando esta nueva expresión del R^2 podemos desarrollar F del siguiente modo

$$F = \frac{(SCR_R - SCR_{NR})}{SCR_{NR}} \frac{(n-K)}{r} \tag{3.21}$$

$$= \frac{(SCR_R - SCR_{NR})/SCR_R}{SCR_{NR}/SCR_R} \frac{(n-K)}{r} \tag{3.22}$$

$$= \frac{\frac{(\hat{\mathbf{y}} - \bar{\mathbf{y}})'(\hat{\mathbf{y}} - \bar{\mathbf{y}})}{(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})}}{\frac{\mathbf{e}'\mathbf{e}}{(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})}} \frac{(n-K)}{r} \tag{3.23}$$

$$= \frac{R^2}{1 - R^2} \frac{(n-K)}{r}, \tag{3.24}$$

que en este caso se ditribuirá como una $F(r = K - 1, n - K)$.

De este modo si el valor numérico del estadístico supera al de la tabla de la $F(K - 1, n - K)$ rechazaríamos H_0, esto es, rechazaríamos la hipótesis de que «todos los parámetros (excepto el de la constante) son nulos», luego el modelo sería globalmente válido. Lógicamente esto último encaja perfectamente con la formulación del test en términos del R^2 ya que F tomará valores numéricos altos cuando R^2 sea elevado (para un valor fijo de $\frac{(n-K)}{r}$), es decir, cuando el modelo no restringido ajuste relativamente bastante bien. No obstante, nótese que numerador y denominador están ponderados por la relación que exista entre el número de observaciones y el número de parámetros independientes del modelo en cuestión. En la práctica es posible que bajos R^2 puedan ser compatibles con un modelo globalmente significativo siempre que el número de observaciones sea muy elevado en relación con el número de parámetros.

Esta observación pone de manifiesto la relevancia de contar con un test para la significación global del modelo, ya que este permite no dejarnos guiar exclusivamente por el valor arrojado por el R^2, sino completar nuestra valoración del modelo de una forma más sólida utilizando un test estadístico sobre la validez del modelo completo.

Resulta más cómodo a efectos prácticos contrastar hipótesis utilizando la expresión (3.17) que la (3.13), del mismo modo que es más rápido usar (3.24) para contrastar la significatividad global del modelo que el estadístico equivalente basado en las sumas cuadráticas residuales. Existe una expresión equivalente para el test general de hipótesis lineales (3.17) también en términos de R^2. Para ello, distinguimos entre el R^2 de la regresión restringida y de la no restringida, $R^2_R \equiv [1 - (SCR_R)/(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})]$ y $R^2_{NR} \equiv [1 - (SCR_{NR})/(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})]$, respectivamente. Utilizando estas definiciones expresamos el estadístico F como

$$F = \frac{(SCR_R - SCR_{NR})}{SCR_{NR}} \frac{(n-K)}{r} \tag{3.25}$$

$$= \frac{(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})[(1 - R^2_R) - (1 - R^2_{NR})]}{(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})(1 - R^2_{NR})} \frac{(n-K)}{r} \tag{3.26}$$

$$= \frac{R^2_{NR} - R^2_R}{1 - R^2_{NR}} \frac{(n-K)}{r} \tag{3.27}$$

que expresa otra forma equivalente de realizar el contraste de la F para cualquier conjunto de restricciones lineales. De hecho, en el caso de la restricción de significación global (3.18) será un caso particular de este último resultado. Así, bajo la hipótesis nula del modelo restringido se tiene (3.19) y por tanto $R^2_R \equiv [1 - (SCR_R)/(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})] = 0$, por lo que (3.27) queda reducido a (3.24).

3.3 Comportamiento asintótico del estimador mínimo cuadrático.

Como decíamos a comienzo de tema, una forma muy atractiva y más aplicable de realizar inferencia es trabajar con la distribución asintótica del estimador MCO. Decimos que es más aplicable en la medida en que los supuestos no son tan restrictivos como lo eran en el modelo lineal de regresión normal.

Los supuestos que ahora necestimos están recogidos a continuación

SUPUESTOS DEL **MODELO DE REGRESIÓN LINEAL**. Las observaciones $(Y_i, \mathbf{x}_i)$, $i = 1, 2, ..., n$, satisfacen la ecuación lineal de regresión

LINEALIDAD

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i. \tag{3.28}$$

MUESTRA ALEATORIA

$$(X_{1i}, X_{2i}, ..., X_{ki}, Y_i), i = 1, ..., n \text{ son } i.i.d.$$

EXOGENEIDAD

$$\mathbb{E}\left(\mathbf{x}_i \cdot \varepsilon_i\right) = \mathbf{0}. \tag{3.29}$$

NO MULTICOLINEALIDAD PERFECTA

$$\mathbb{E}(\mathbf{x}_i\mathbf{x}_i') > 0, \text{ donde } \mathbf{x}_i = (X_{0i}, X_{1i}, ..., X_{ki})'. \tag{3.30}$$

MOMENTOS ORDEN CUATRO: ATÍPICOS POCO PROBABLES

$$\mathbb{E}Y_i^4 < \infty \text{y la} \mathbb{E}\left\|X_i^4\right\| < \infty \tag{3.31}$$

Seguidamente explicaremos el motivo de estos supuestos. En todo caso se observa que algunos supuestos son los mismos (o están relacionados), mientras que el cambio importante está en el hecho de no requerir normalidad de los residuos ni tampoco homocedasticidad.

3.3.1 Consitencia.

La consistencia es una propiedad de los estimadores, de hecho que un estimador sea consistente es una buena propiedad para el estimador. Significa que para cualquier distribución de datos, existe un tamaño muestral n lo suficientemente grande como para que el estimador **b** esté, con una alta probabilidad, tan cercano como deseemos al verdadero valor $\boldsymbol{\beta}$.

Para comprobar que el estimador MCO **b** es consistente para el vector de parámetros o coeficientes $\boldsymbol{\beta}$ daremos estos tres pasos:

1. Mostrar que el estimador MCO puede escribirse como una función continua de un conjunto de momentos muestrales.
2. Usar una Ley de grandes números (LGN) que nos permita verificar que los momentos muestrales convergen a los poblacionales.
3. Utilizar un resultado técnico que nos garantice que las funciones continuas preservan la convergencia.

El primer paso consiste en reescribir el estimador MCO, $\mathbf{b}$ ($\mathbf{b}_n$), del siguiente modo

$$\mathbf{b}_n = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i Y_i\right) \tag{3.32}$$

o bien

$$\mathbf{b}_n = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\mathbf{y}\right)$$

o bien

$$\mathbf{b}_n - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}.$$

En cualquier caso enfatizamos la dependencia del tamaño muestral y el hecho de que el estimador MCO es una función g que depende de

$$\mathbf{b}_n = g\left(\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right), \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right)\right)$$

Observemos que cada uno de los términos del producto está calculando promedios y estos son el objeto principal de las Leyes de Grandes Números (LGN), que se estudian en otros apartados de este temario y por tanto no explicaremos de nuevo.

Uno de los supuestos que se mantienen es el hecho de que $(Y_i, \mathbf{x}_i)$ sea una muestra aleatoria, y este supuesto es central para aplicar LGNs. Ahora bien ¿cualquier función $(Y_i, \mathbf{x}_i)$ será también iid? En particular, ¿lo serán $(\mathbf{x}_i\mathbf{x}_i')$ y $(\mathbf{x}_i\, Y_i)$ o en su caso $(\mathbf{x}_i\varepsilon_i)$? La respuesta es afirmativa siempre que las transformaciones sean continuas, y en este caso lo son. La demostración se puede encontrar en la bibliografía recomendada.

Por otro lado las LGN requieren que $(\mathbf{x}_i\mathbf{x}_i')$ y $(\mathbf{x}_i\, Y_i)$ tengan medias finitas. Para ello requeriremos unas **condiciones de regularidad**

$$\mathbb{E}Y^2 < \infty,$$

$$\mathbb{E}\left\|\mathbf{x}\right\|^2 < \infty.$$

Bajo estas condiciones podemos aplicar una LGN de modo que cuando $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' \xrightarrow{p} \mathbb{E}\left(\mathbf{x}_i\mathbf{x}_i'\right), \tag{3.33}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i Y_i \xrightarrow{p} \mathbb{E}\left(\mathbf{x}_i Y_i\right)$$

y

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i \xrightarrow{p} \mathbb{E}\left(\mathbf{x}_i\varepsilon_i\right).$$

En las expresiones (3.32) del estimador $\mathbf{b}$ tenemos una función de estos últimos promedios. La pregunta ahora es saber si esta función preserva la convergencia en probabilidad. De nuevo la respuesta es positiva, si la función g es continua, como es el caso.

El estimador MCO consiste es una función de dos argumentos

$$\mathbf{b} - \boldsymbol{\beta} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\boldsymbol{\varepsilon}},$$

donde $\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right), \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\boldsymbol{\varepsilon}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right)$. La función será continua en aquellos puntos en los que exista la inversa $\widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}$. Lo que nos lleva a introducir un supuesto que garantice tal existencia: $\boldsymbol{\Sigma}_{\mathbf{xx}} \equiv \mathbb{E}(\mathbf{x}\mathbf{x}')$ debe ser una matriz definida positiva. Este supuesto ya lo teníamos anteriormente en el modelo de regresión lineal normal. Observamos también que para que

$$\mathbf{b}_n - \boldsymbol{\beta} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}\boldsymbol{\varepsilon}} \xrightarrow{p} \mathbf{0},$$

será necesario que la relación entre las explicativas y el error sea

$$\mathbb{E}\left(\mathbf{x}_i \cdot \varepsilon_i\right) = \mathbf{0}$$

que es una condición de ortogonalidad entre los mismos menos exigente que la exogeneidad requerida en muestras finitas. De hecho, si los regresores son exógenos, $\mathbb{E}(\varepsilon_i \,|\mathbf{x}_i) = 0$, entonces la condición de ortogonalidad se cumple inmediatamente, mientras el reverso no es cierto. Por otra parte, obsérvese que es estimador MCO, bajo el supuesto de ortogonalidad del error, no necesariamente será insesgado, pese a que será consistente. Al fin y al cabo la distribución para muestras finitas es exacta por lo que es normal que el requisito sea más restrictivo que en el caso asintótico.

Así pues llegamos al resultado esperado de que el estimador MCO es consistente

TEOREMA DE CONSISTENCIA DEL ESTIMADOR MCO
Bajo el supuesto de que $(Y_i, \mathbf{x}_i)$ sea iid, $\mathbb{E}Y^2 < \infty$, $\mathbb{E}\left\|\mathbf{x}\right\|^2 < \infty$, $\mathbb{E}\left(\mathbf{x}_i \cdot \varepsilon_i\right) = \mathbf{0}$, y si $\boldsymbol{\Sigma}_{\mathbf{xx}} \equiv \mathbb{E}(\mathbf{x}\mathbf{x}')$ es definida positiva, entonces se tiene que $\mathbf{b}$ es consistente, es decir:

$$\mathbf{b} \xrightarrow{p} \boldsymbol{\beta},$$

o bien

$$plim(\mathbf{b}) = \boldsymbol{\beta},$$

o bien

$$\mathbf{b} = \boldsymbol{\beta} + o_p(1).$$

Estas tres expresiones indican lo mismo, que el estimador MCO ($\mathbf{b}$, o de modo equivalente $\mathbf{b}_n$) converge en probabilidad hacia $\boldsymbol{\beta}$ a medida que el tamaño muestral crece, y por lo tanto el estimador MCO es consistente.

También es especialmente interesante observar que el supuesto de media condicionada nula (exogeneidad condicionada) implica que hemos modelizado correctamente la esperanza condicionada de la variable objetivo, es decir, la FEC. Lo cual implicaba, como dijimos, que los efectos parciales o efectos ceteris paribus sobre el valor esperado de la variable dependiente podían ser estimados.

Sin embargo, el supuesto de ortogonalidad (correlación nula entre errores y explicativas) nos permite ampliar la aplicación del estimador MCO a una aproximación lineal de la FEC, algo por otra parte bastante natural y habitual puesto que no es difícil considerar que cuando modelizamos, lo que hacemos es aproximar linealmente la FEC, máxime sabiendo que dicha aproximación lineal es la mejor lineal en términos predicitivos. Así pues esta aproximación lineal nos indica que, en caso de que algunos de los supuestos del Modelo de Regresión Lineal Normal no se satisficieran, y aún así estimaros por MCO, entonces estimaríamos precisamente la mejor combinación lineal de las variables que hemos llamado «explicativas» para «predecir» la variable dependiente.

3.3.2 Normalidad e inferencia asintótica

El último teorema nos permite saber que el estimador converge al verdadero vector parámetros, pero esto en sí mismo no es suficiente para poder realizar inferencia estadística. Es decir, necesitamos la distribución del estimador MCO. Este apartado muestra cómo es posible llegar a la ***distribución asintótica*** del estimador MCO.

Del mismo modo que la consistencia se fundamenta en la LGN, la normalidad asintótica lo hace en los teoremas centrales del límite (TCL) aplicados sobre la expresión

$$\mathbf{b} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon},$$

o alternativamente sobre

$$\mathbf{b}_n - \boldsymbol{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i\mathbf{x}_i')\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right). \tag{3.34}$$

El lector habrá de repasar en este punto el TCL de Lindeberg-Levy:

TCL Linderberg-Levy: Sea $\{z_n\}$ una sucesión de variables aleatorias independientemente e idénticamente distribuidas (iid), tal que $\mu \equiv \mathbb{E}(z_n) < \infty$ y $\sigma^2 \equiv \text{var}(z_n) < \infty$. Si $\sigma^2 \neq 0$, entonces

$$\sqrt{n}\left(\bar{z}_n - \mu\right)/\sigma = \frac{1}{\sqrt{n}}\sum_{n=1}^{N}\left(z_n - \mu\right)/\sigma \xrightarrow{d} N(0,1),$$

o alternativamente

$$\sqrt{n}\left(\bar{z}_n - \mu\right) \xrightarrow{d} N(0,\sigma^2). \tag{3.35}$$

Esta ecuación directamente muestra que para poder aplicar el TCL expuesto en (3.35), necesitamos escalar la expresión por $\sqrt{n}$, con lo que obtenemos

$$\sqrt{n}\left(\mathbf{b}_n - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i\mathbf{x}_i')\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right). \tag{3.36}$$

Así pues el estimador escalado $\sqrt{n}\,(\mathbf{b}_n - \boldsymbol{\beta})$ es una función de la media muestral $\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i\mathbf{x}_i')$ y del promedio $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i$, que tiene media cero, por lo que estamos en condiciones de intentar aplicar el TCL (ver TCL (3.35)). Para verificar que se satisfacen las condiciones del TCL, necesitamos, en primer lugar, observar que el supuesto de muestreo aleatorio, nos aseguran que $(\mathbf{x}_i\mathbf{x}_i')$ y $(\mathbf{x}_i\,Y_i)$ son variables iid, y dado que ε_i es una combinación lineal de Y_i con $\mathbf{x}_i$, también será iid la variable $(\mathbf{x}_i\varepsilon_i)$. Estas variables aleatorias deben tener (para aplicar el TCL) momentos de primer y segundo orden finitos (deben existir sus medias y varianzas-covarianzas). La matriz de varianzas-covarianzas $\text{var}(\mathbf{x}_i\varepsilon_i)$ la denotamos por

$$\boldsymbol{\Omega} \equiv \mathbb{E}(\mathbf{x}_i\mathbf{x}_i'\varepsilon_i^2)(= \text{var}(\mathbf{x}_i\varepsilon_i)). \tag{3.37}$$

La existencia de varianza y covarianzas de $(\mathbf{x}_i\mathbf{x}_i')$ y de $(\mathbf{x}_i\varepsilon_i)$ requiere que contemplemos la existencia de los momentos de orden cuatro de las variables x_i y ε_i.

La $\mathbb{E}Y_i^4 < \infty$ y la $\mathbb{E}\,\|X_i^4\| < \infty$.

Recordemos que la existencia de estos momentos bajo las condiciones del MPL garantiza la existencia de $\mathbb{E}\varepsilon_i^4$. Este supuesto es el que introdujimos anteriormente, y entonces le dábamos una interpretación en términos de atípicos.

Bajo las condiciones establecidas en Teorema de la consistencia y añadiendo las condiciones de los momentos de orden cuatro podemos aplicar el TCL

Bajo el supuesto de que $(Y_i, \mathbf{x}_i)$ sea iid, $\mathbb{E}\,(\mathbf{x}_i \cdot \varepsilon_i) = \mathbf{0}$, $\boldsymbol{\Sigma}_{\mathbf{xx}} \equiv \mathbb{E}(\mathbf{x}\mathbf{x}')$ es definida positiva, y momentos de orden cuatros finitos para las variables del modelo, se tiene

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega})$$

cuando $n \to \infty$.

Si ahora utilizamos este resultado y las expresiones (3.33) y (3.36),

$$\sqrt{n}\,(\mathbf{b}_n - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}N(\mathbf{0}, \boldsymbol{\Omega}) = N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Omega}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1})$$

cuando $n \to \infty$.

Por tanto, hemos demostrado el Teorema siguiente

TEOREMA NORMALIDAD ASINTÓTICA DEL ESTIMADOR MCO
Bajo el supuesto de que $(Y_i, \mathbf{x}_i)$ sea iid, $\mathbb{E}(\mathbf{x}_i \cdot \varepsilon_i) = \mathbf{0}$, $\boldsymbol{\Sigma}_{\mathbf{xx}} \equiv \mathbb{E}(\mathbf{x}\mathbf{x}')$ es definida positiva, y momentos de orden cuatros finitos

$$\sqrt{n}(\mathbf{b}_n - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}),$$

donde

$$\mathbf{V} \equiv \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\Omega} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}, \tag{3.38}$$

$\boldsymbol{\Sigma}_{\mathbf{xx}} \equiv \mathbb{E}(\mathbf{x}_i \mathbf{x}_i')$ y $\boldsymbol{\Omega}$ está definida en (3.37)
O bien, alternativamente,

$$\mathbf{b} = \boldsymbol{\beta} + O_p(n^{-1/2}).$$

Este resultado indica que la distribución de $\sqrt{n}$ veces el error muestral $(\mathbf{b}_n - \boldsymbol{\beta})$ se aproxima a una distribución normal cuando n es suficientemente grande. A la matriz $\mathbf{V}$ se le suele denominar matriz asintótica de varianzas y covarianzas de $\mathbf{b}$.

Para que este resultado sea operativo es necesario estimar consistentemente $\mathbf{V}$, y así poder luego hacer inferencia. La estimación de dicha matriz pertenece a otro tema, y trabajaremos por ahora como si ya tuviéramos una estimación consistente de la misma.

En particular si quisiéramos **contrastar una hipótesis lineal simple** sobre el coeficiente k-ésimo del modelo, el teorema de normalidad asintótica implica que bajo la hipótesis nula

$$H_0 : \beta_k = c$$

entonces

$$\sqrt{n}(b_k - c) \xrightarrow{d} N(0, \mathbf{V}(b_k)),$$

donde $\mathbf{V}(b_k)$ es el elemento (k,k) de la matriz $K \times K$ que hemos denominado $\mathbf{V}$. Por lo tanto,

$$t_k \equiv \frac{\sqrt{n}(b_k - c)}{\sqrt{\widehat{\mathbf{V}}(b_k)}} = \frac{(b_k - c)}{ee(b_k)} \xrightarrow{d} N(0,1),$$

donde ee se refiere al error estándar del parámetro, es decir,

$$ee(b_k) \equiv \sqrt{\frac{1}{n}\widehat{\mathbf{V}}(b_k)} = \sqrt{\frac{1}{n}\left(\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \widehat{\boldsymbol{\Omega}} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\right)_{kk}}$$

Obsérvese que este error estándar admite que los errores sean condicionalmente heterocedásticos. De hecho, no hemos hecho en esta sección supuesto alguno al respecto. Este tipo de ratio se distingue del ratio tipo t en muestras finitas en que los errores estándar están calculados de forma robusta a la heterocedasticad cuando $\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\Omega} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$ se estima robustamente frente a la misma. Este aspecto se desarrolla en otro tema.

A partir de este momento el contraste de hipótesis se conduce de forma similar al realizado para muestras finitas pero considerando que la distribución sobre la contrastar la hipótesis nula es ahora la N(0,1).

No obstante conviene resaltar que el estadístico para realizar el contraste de la t en muestras finitas es un estadístico con distribución *exacta*, mientras que t asintótico es un estadístico con distribución *asintótica*. Esto último implica que el tamaño exacto del test o contraste (la probabilidad del Error Tipo I dado un tamaño muestral) es *aproximadamente* igual al tamaño nominal del test (es decir, el nivel-α deseado de significatividad). La diferencia entre uno y otro es asintóticamente nula cuando el tamaño muestral n crece hasta infinito. Igualmente la forma de cómputo es distinta ya que, por un lado, en el caso exacto se utiliza la distribución de una $t - student$, y en el asintótico la de una normal estándar. Por otro lado, los errores estándar se calculan de forma diferente, y además se calculan bajo supuestos distintos también.

Por último, si necesitamos contrastar varias restricciones lineales simultáneamente, podremos utilizar un contraste de Wald o LM asintótico. El siguiente resultado recoge los resultados para una restricción y varias

Bajo la hipótesis nula $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, siendo $\mathbf{R}$ una matriz $r \times K$ de rango completo, suponiendo que se cumplen las condiciones del Teorema de Normalidad Asintótica, y suponiendo que $\hat{\mathbf{V}}$ es un estimador consistente de $\mathbf{V}$, entonces

$$n\left(\mathbf{R}\mathbf{b}_n - \mathbf{r}\right)' \left(\mathbf{R}\hat{\mathbf{V}}\mathbf{R}'\right)^{-1} \left(\mathbf{R}\mathbf{b}_n - \mathbf{r}\right) \xrightarrow{d} \chi^2(r).$$

En el caso particular de que $H_0 : \beta_k = c$, entonces

$$t_k \equiv \frac{\sqrt{n}\left(b_k - c\right)}{\sqrt{\widehat{\mathbf{V}}(b_k)}} = \frac{\left(b_k - c\right)}{ee(b_k)} \xrightarrow{d} N(0,1).$$

Bibliografía complementaria

Matilla-García, M et al. 2017. Econometría y Predicción. McGraw Hill

Stock J. and Watson J. Introducción a la econometría. Pearson.

Tema 4

Analisis de regresion con datos de seccion cruzada III

Este tema está elaborado como una adaptación del capítulo 6:

Wooldridge. J. 4th Ed., Introductory Econometrics.

Así como de la bibliografía complementaria.

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al Órgano convocante ni al Tribunal actuante.

- Temas adicionales en el análisis de Regresión Múltiple.
- Efectos del cambio de escala sobre los estimadores MCO.
- Formas funcionales, selección de modelos, predicción y análisis residual.

4.1 Temas adicionales en el análisis de Regresión Múltiple

A lo largo de este tema se presentan varios resultados que están relacionados con el uso del modelo de regresión múltiple. La aplicabilidad de estos resultados está garantizada tanto para los supuestos del modelo de regresión lineal como para el modelo teórico normal.

El tema trata por tanto aspectos prácticos de bastante utilidad para un económetra cuando plantea y estima modelos que expliquen el comportamiento esperado de la variable dependiente condicinado a unos valores de las variables explicativas. Se estudiará qué sucede con las estimaciones MCO cuando las variables sufren cambios de escala. Muchas veces estos cambios de escala son imprescindibles o muy convenientes para el tipo de problema que tenemos entre manos. En otras ocasiones, la económetra tiene que afinar también las formas funcionales que admite el modelo pese a ser lineal. Veremos que esto es fundamental para la correcta interpretación de los parámetros estimados y de sus efectos marginales sobre la variable dependiente estimada.

Dado el carácter condicionado de los modelos de regresión, suele ser habitual que estemos interesados en controlar o condicionar las relaciones económicas por ciertas características determinantes de la muestra analizada. Para ello utilizamos variables dicotómicas. Los modelos de regresión presentados admiten este tipo de variables, y resultan ser muy útiles cuando interactuan con otras variables de naturaleza continua. También estos aspectos serán objeto del interés a lo largo de las siguiente páginas.

Por último, el tema termina con unas secciones dedicadas al propio proceso de mo-

delización, en particular a la selección de modelos y en su utilidad para evaluar los impactos sobre la variable dependiente de escenarios futuros utilizando así el modelo para hacer predicciones, si bien es cierto que en los modelos de sección cruzada, como es los que estamos estudiando, el objetivo principal suele ser el estimar correctamente los efectos parciales de las variables explicativas.

4.2 Efectos del cambio de escala sobre los estimadores MCO

Modificar la escala en la que los datos son introducidos en un modelo es algo relativamente habitual. Se utiliza con fines prácticos o incluso estéticos, como puede ser el reducir el número de ceros después de un punto decimal en un coeficiente estimado. Al elegir con criterio las unidades de medida, podemos mejorar la apariencia de una ecuación estimada sin cambiar nada que sea esencial.

Los cambios en las unidades de medida técnicamente se denominan cambios de escala y se pueden representar e ilustrar de un modo bastante general a partir de la siguiente expresión, que fácilmente es extensible a un mayor número de variables:

$$w_1 Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 w_2 X_{1i} + \tilde{\beta}_2 X_{2i} + \tilde{\varepsilon}_i, \tag{4.1}$$

donde w_1 es el cambio de escala de la variable dependiente y w_2 el cambio de escala de la variable independiente. Utilizamos $\sim$ para distinguir los coeficientes cuyas variables tienen cambios de escala respecto de los originales o sin cambio de escala. La cuestión que nos plantemos es cómo varían los valores estimados respecto del modelo general

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\varepsilon}_i.$$

A partir del modelo estimado anterior, si incorporamos los cambios de escala, entonces para que siga siendo cierto se tendrá

$$w_1 Y_i = \left(w_1 \hat{\beta}_0\right) + \left(\frac{w_1}{w_2}\hat{\beta}_1\right) w_2 X_{1i} + \left(w_1 \hat{\beta}_2\right) X_{2i} + \left(w_1 \hat{\varepsilon}_i\right).$$

Se puede comprobar que en estas condiciones

$$\tilde{\beta}_1 = \frac{w_1}{w_2}\hat{\beta}_1, \tilde{\beta}_2 = w_1 \hat{\beta}_2, \tag{4.2}$$

$$\tilde{\beta}_0 = w_1 \hat{\beta}_0, \tag{4.3}$$

$$\tilde{\varepsilon}_i = w_1 \hat{\varepsilon}_i. \tag{4.4}$$

Por consiguiente la pendiente de la variable independiente que ha sido escalada se ve afectada por los cambios de escala de ambas variables, expresión (4.2). El término constante, la variable independiente X_2, y los errores estimados (residuos) sin embargo solo son afectados por el cambio de escala de la variable explicada [(4.2), (4.3) y (4.4)].

Los estadísticos tipo t para realizar contrastes de significatividad de una variable explicativa tampoco se verían modificados. Sin embargo los errores estándar sí. Ambas

cosas suceden tanto para cambios sobre la variable dependiente como para cambios en la variable independiente. Los intervalos de confianza, por otra parte, sí se verían alterados por el factor w_1/w_2. Sin embargo, el R-cuadrado tampoco se vería afectado por estos cambios de escala.

Cuando alguna de las variables tiene una escala de valores de difícil interpretación puede ser interesante medirla en términos tipificados o estandarizados. Tipificar no es más que restar la media a todos los valores de la variable y dividirla por su desviación típica o error estándar

$$Z_j = \frac{X_j - \bar{X}_j}{S_{X_j}}. \tag{4.5}$$

Cuando tipificamos obtenemos variables con media nula y varianza unitaria. La unidad de medida en este caso es la desviación típica (o error estándar). Si la variable se distribuye normalmente entonces un incremento de una desviación típica equivale a un incremento aproximado del 34 % sobre su valor medio y un incremento de 0,25 desviaciones a un incremento del 10 %.

En otras ocasiones puede resultar adecuado expresar todo el modelo estandarizado. En este caso el modelo se denomina habitualmente *modelo de coeficientes beta*. Si en el modelo de regresión múltiple restamos a todas las variables su media y las dividimos por sus respectivos errores estándar obtenemos el siguiente modelo

$$\begin{array}{l} \frac{Y_i - \bar{Y}}{S_y} = \left(\frac{S_{X_1}}{S_y}\right)\hat{\beta}_1\left(\frac{X_{1i}-\bar{X}_1}{S_{X_1}}\right) + \left(\frac{S_{X_2}}{S_y}\right)\hat{\beta}_2\left(\frac{X_{2i}-\bar{X}_2}{S_{X_2}}\right) \\ + ... + \left(\frac{S_{X_k}}{S_y}\right)\hat{\beta}_k\left(\frac{X_{ki}-\bar{X}_k}{S_{X_k}}\right) + \frac{\hat{\varepsilon}_i}{S_y}. \end{array} \tag{4.6}$$

donde desaparece el término constante, pues estamos utilizando una regresión en desviaciones a las medias y los coeficientes de la regresión del modelo en niveles aparecen multiplicados por el cociente de las desviaciones típicas en aplicación de las expresiones (4.2), (4.3), y (4.4) podemos expresar (4.6) en términos de variables tipificadas Z

$$Z_y = \tilde{\beta}_1 Z_1 + \tilde{\beta}_2 Z_2 + ... + \tilde{\beta}_k Z_k + \tilde{\varepsilon}, \tag{4.7}$$

donde utilizamos $\widetilde{\beta}$ para distinguir los coeficientes beta respecto de los mínimo cuadráticos «$\hat{\beta}$».

Una de las ventajas de los **coeficientes beta** es que no dependen de las unidades de medida utilizadas y permiten determinar la influencia de las variables explicativas sobre la explicada a partir de la magnitud del coeficiente, lo que normalmente no ocurre en los otros casos en que los coeficientes pueden modificarse cambiando las unidades de medida de las variables.

4.3 Formas funcionales, selección de modelos, predicción y análisis residual

4.3.1 Formas funcionales

El modelo de regresión es lo suficientemente flexible como para contemplar relaciones no lineales. Los modelos de regresión no lineales *en las variables* los podemos linealizar mediante cambios de variable, y es habitual realizar transformaciones en las variables en los estudios aplicados. Algunas de las transformaciones más comunes son: los modelos logarítmicos o de elasticidad constante (log-log), los semilogarítmicos [logarítmicos lineales (log-nivel) y lineales logarítmicos (nivel-log)] y los recíprocos.

Cuando la relación entre las variables es exponencial del tipo

$$Y = \beta_0 X^{\beta_1} e^{\varepsilon}, \tag{4.8}$$

si tomamos logaritmos y operamos, la Ecuación (4.8) se puede expresar como

$$\ln Y = \ln \beta_0 + \beta_1 \ln X + \varepsilon = \alpha_0 + \beta_1 \ln X + \varepsilon, \tag{4.9}$$

puesto que ln β_0 es una constante podemos hacer el cambio (ln $\beta_0 = \alpha_0$). Por consiguiente el modelo (4.8) lo hemos transformado en otro, expresión (4.9), en el que las variables están en logaritmos. A este tipo de modelo se le conoce por el nombre de modelo **log-log** o modelo de elasticidad constante.

En el modelo logarítmico el coeficiente $\hat{\beta}_1$ (0,97 para el caso de la demanda de tabaco) estima la elasticidad de Y respecto de X. En este modelo, por tanto, una variación de un 1 % en la variable explicativa (que está en logaritmos) está asociada con una variación en la variable dependiente (también en logaritmos) de un β_1 %.

Resulta útil repasar la relación entre el logaritmo y el porcentaje para entender el porqué de las interpretaciones que hacemos cuando aparecen logaritmos. Consideremos una variación «pequeña» de cualquier variable x que denotamos como Δx. La diferencia entre el logaritmo de $x + \Delta x$ y el logaritmo de x es «aproximadamente» $\Delta x/x$. Por ejemplo, si $x = 100$ y $\Delta x = 1$, entonces $\Delta x/x = 1/100 = 0{,}01$, mientras que $ln\,(x + \Delta x) - ln(x) = ln(101) - ln(100)$ que arroja un valor de $0{,}00995$, que es aproximadamente igual (indistinguible en la práctica) de $0{,}01$. Por tanto, siempre que $\Delta x/x$ sea pequeño, la diferencia[1] de los logaritmos captura la variación porcentual en x dividida entre 100. Es decir, $\Delta x/x = 0{,}01$ implica que la variación *porcentual* en x ha sido del $0{,}01 \times 100 = 1\,\%$.

Consideremos ahora la variación en lnY ante un cambio en la variable en $ln(X)$, esto es

$$ln\,(Y + \Delta Y) - ln\,(Y) = [\beta_0 + \beta_1 ln\,(X + \Delta X)] - [\beta_0 + \beta_1 ln(X)] = \beta_1\,(ln\,(X + \Delta X) - ln(X))\,,$$

y aplicamos en ambos la relación comentada anteriormente:

$$ln(x + \Delta x) - ln(x) \cong \frac{\triangle x}{x},$$

[1] En términos de cálculo matemático esta interpretación se basa en que la diferencial de la función ln(x), $d(\ln x) = dx/x$.

entonces se tiene

$$\frac{\triangle Y}{Y} \cong \beta_1 \frac{\triangle X}{X},$$

o lo que es lo mismo

$$\beta_1 = \frac{\triangle Y/Y}{\triangle X/X},$$

que es el ratio de variación de proporciones, y por tanto si multiplicamos por 100, obtenemos el ratio de cambio porcentual, que es la elasticidad.

Si la variable endógena Y está en logaritmos y la variable explicativa X en niveles entonces el modelo se denomina logarítmico lineal (**log-lin** o **log-nivel**), su forma general es

$$\ln Y = \beta_0 + \beta_1 X + \varepsilon, \tag{4.10}$$

donde la pendiente β_1 multiplicada por 100 es aproximadamente la tasa porcentual de cambio de la variable dependiente $100 \cdot \beta_1 \Delta X = \Delta Y$ %, y se suele denominar semielasticidad. Lo que se interpreta fácilmente ya que si X cambia en una unidad (cambio unitario), este cambio está asociado a un cambio de $100 \times \beta_1$ % en Y. Esto es así[2] porque si comparamos los valores de ln$Y$ antes y después de que se haya producido una variación discreta ΔX en X, tenemos

$$ln\,(Y + \Delta Y) - ln\,(Y) = [\beta_0 + \beta_1\,(X + \Delta X)] - [\beta_0 + \beta_1 X] = \beta_1\,(\Delta X)\,.$$

Si aplicamos a la diferencia que está a la izquierda del igual, el resultado, visto anteriormente, de la diferencia de logaritmos se aproxima a $\Delta Y/Y$, entonces

$$\Delta Y/Y \simeq \beta_1 \Delta X,$$

luego un cambio unitario en X genera un cambio en $\Delta Y/Y$ de β_1, que implica una variación porcentual en Y de $100 \times \beta_1$ %.

En el modelo lineal logarítmico **(lin-log)** la variable dependiente está en niveles mientras que la independiente aparece en logaritmos, es decir que ahora el modelo poblacional es

$$Y = \beta_0 + \beta_1\,(\ln X) + \varepsilon, \tag{4.11}$$

donde la pendiente β_1 dividida por 100 es aproximadamente el cambio de la variable explicada $\Delta Y = (\beta_1/100)\Delta X$[3]. Esta interpretación es así por lo siguiente. Consideremos la diferencia en la función de regresión poblacional entre los valores de X que se diferencian en la cantidad ΔX: es decir

$$[\beta_0 + \beta_1 ln\,(X + \Delta X)] - [\beta_0 + \beta_1 ln\,(X)] = \beta_1\,(ln\,(X + \Delta X) - ln\,(X)) \simeq \beta_1\,(\Delta X/X)\,.$$

Por tanto si cambia X en un 1 %, es decir si $\Delta X/X = 0{,}01$, entonces dicho cambio tiene asociado en este modelo una variación en Y de $0,01 \times \beta_1$.

[2] Diferenciando a ambos lados (4.10) tenemos $dY/Y = \beta_1 dx$. Si multiplicamos por 100 en ambos lados, y sustituimos el diferencial por un pequeño incremento discreto (Δx), resulta: $\beta_1 \Delta X \cdot 100 = (\Delta Y/Y) \cdot 100 = \Delta Y$ %.

[3] Diferenciando a ambos lados de la ecuación lin-log, se tiene $dY = \beta_1 dX/X$. Sustituyendo diferenciales por incrementos pequeños, tenemos $\beta_1(\Delta X/X) = \Delta Y$, multiplicando y dividiendo en el lado izquierdo por 100 obtenemos el cambio $(\beta_1/100)(\Delta X/X) \cdot 100 = \Delta Y$.

Tabla 4.1: Formas funcionales habituales

Modelo	Variable Dependiente	Variable Independiente	Interpretación del Cambio	Elasticidad
Nivel-nivel	Y	X	$\Delta E(Y) = \beta \Delta X$	$\beta \left({}^{X}/_{Y}\right)$
Nivel-log	Y	$\ln X$	$\Delta E(Y) = \left(\frac{\beta}{100}\right) \Delta X\,\%$	$\beta \left({}^{1}/_{Y}\right)$
Log-nivel	$\ln Y$	X	$\Delta E(Y)\,\% = 100\beta \Delta X$	βX
Log-log	$\ln Y$	$\ln X$	$\Delta E(Y)\,\% = \beta \Delta X\,\%$	β

Se conoce como modelo recíproco a aquel en que la variable independiente aparece en su forma inversa, es decir

$$Y = \beta_0 + \beta_1 \left(1/X\right) + \varepsilon. \tag{4.12}$$

A medida que X aumenta la variable independiente disminuye $1/X$, en el límite se va acercando a cero, momento en que la variable explicada Y se hace igual al término constante $Y = \beta_0$, por tanto este tipo de modelos tiene sentido cuando la variable dependiente tiene límite asintótico β_0.

La elección de la forma funcional en los modelos de regresión simple puede ser *a priori* relativamente fácil de determinar puesto que podemos realizar el gráfico de las variables y hacernos una idea de cómo podría ser la forma funcional. El problema se agrava cuando introducimos más de una variable independiente (regresión múltiple), entonces la elección de la forma funcional de las distintas variables puede ser todo lo complicada que queramos. En ocasiones la teoría económica (o el sentido económico) nos sugieren una forma funcional determinada. También puede resultar útil el cálculo de la tasa de cambio y la elasticidad de los parámetros. La Tabla 4.1 muestra la interpretación del cambio en las variables y el cálculo de las elasticidades de los modelos en niveles y en logaritmos.

4.3.1.1 Formas funcionales cuadráticas

La regresión múltiple permite establecer relaciones funcionales de una variable que no se pueden tratar o modelizar mediante la regresión simple. En esta sección y las dos siguientes vamos a tratar este tipo de consideraciones relativas a la forma funcional.

Supongamos una relación cuadrática del siguiente tipo

$$Y = \beta_0 - \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon, \tag{4.13}$$

donde la variable explicativa X_1 aparece en niveles y al cuadrado. Esta aproximación se denomina *modelo de regresión cuadrático*[4] porque la función de regresión poblacional, es decir $\mathbb{E}\left(Y_i \,|X_i\right) = \beta_0 - \beta_1 X_{1,i} + \beta_2 X_{1,i}^2$, define una función cuadrática respecto de una *sola* variable independiente, en este caso, X_1. El modelo por tanto relaciona la variable dependiente Y con una variable independiente X_1 de un modo no lineal y, pese a que

[4]Con independencia de los signos de parámetros poblacionales.

esto podría parecer *a priori* de complejo tratamiento, la regresión *múltiple* nos permite tratarla adecuadamente al considerar como variables distintas a X_1^2, y a X_1.

La interpretación del efecto en la variable Y de un cambio en la variable X_1 será diferente. Para ver la relación entre ambas variables observemos que aproximadamente

$$\frac{\Delta Y}{\Delta X_1} \approx (-\beta_1 + 2\beta_2 X_1)\,. \tag{4.14}$$

Lo primero que advertimos es que la variación esperada en la variable dependiente Y ahora depende del nivel inicial en el que se encuentre la variable explicativa X_1. Lo segundo es que existirá un nivel determinado para el cual la variación esperada en la variable dependiente ante un cambio en la variable explicativa sea nula. Si igualamos a cero la Ecuación (4.14) obtenemos

$$-\beta_1 + 2\beta_2 X_1 = 0;\ X_1 = \frac{\beta_1}{2\beta_2}. \tag{4.15}$$

Luego, en este caso, a partir del nivel umbral encontrado, el efecto sobre la variación en la variable Y será distinto si la variable independiente está por encima o por debajo del mismo. Al ser la segunda derivada positiva, el efecto de X_1 sobre Y será decreciente hasta llegar al valor $\beta_1/2\beta_2$ y creciente a partir de ese momento. Si invertimos los signos, $Y = \beta_0 + \beta_1 X_1 - \beta_2 {X_1}^2 + \varepsilon$ estaremos ante un máximo, de manera que la relación será creciente hasta $\beta_1/2\beta_2$ y decreciente a partir de ese momento.

4.3.1.2 Formas funcionales con términos que interactúan

En ocasiones es adecuado para dotar de mayor realismo o afinación al modelo previsto hacer que una variable explicativa dependa de la magnitud o nivel que alcanza otra variable independiente. Es como si ambas variables explicativas tuvieran un efecto parcial no solo aisladamente, sino también conjuntamente. Este tipo de interacción se puede considerar introduciendo en el modelo un término nuevo que actúe como **término de interacción**. El caso para dos variables con término de interacción es

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon, \tag{4.16}$$

donde la variable producto $X_1 X_2$ es el término de interacción. El efecto parcial de X_1 es ahora

$$\frac{\Delta E(Y|\mathbf{x})}{\Delta X_1} = \beta_1 + \beta_3 X_2. \tag{4.17}$$

En ocasiones se reparametriza el modelo para interpretar de forma más clara el término de interacción.

4.3.1.3 Formas funcionales con variables explicativas discontinuas

En numerosas ocasiones nos encontraremos con que algunos de los factores que afectan a la variable dependiente tienen carácter cualitativo, es decir, son variables que solo

admiten escala nominal, como por ejemplo género, raza, religión, nacionalidad, región geográfica, acciones de política económica o empresarial, etcétera. En general, se trata de acontecimientos que solo admiten una valoración cualitativa. En estos casos se suelen utilizar *variables dicotómicas* (también la literatura se refiere a este tipo de variables con el anglicismo de variable *dummy* o *dummies*) para incluir su influencia en el modelo de regresión.

Cuando el modelo incorpore variables binarias podremos realizar entonces interpretaciones similares de los coeficientes estimados, interpretaciones relacionadas con el efecto parcial o marginal de la variable en cuestión sobre la variable dependiente. Será posible hacer interactuar la variable binaria con otra variable no binaria del modelo, y dotar así al modelo de regresión múltiple de nuevas capacidades explicativas sobre la variable de interés y su relación con las variables explicativas.

Podremos igualmente llevar a cabo contrastes de hipótesis con técnicas *robustas* a la heterocedasticidad y/o autocorrelación sobre los coeficientes del modelo poblacional, de acuerdo a lo presentado en el tema anterior. Igualmente podremos realizar predicciones de la variable dependiente para distintos escenarios configurados por determinados valores de las variables explicativas.

Modelo ANOVA

El modelo ANOVA general tiene la forma siguiente

$$Y_i = \beta_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + ...\alpha_m D_{mi} + \varepsilon_i. \tag{4.18}$$

Hay, por tanto, m variables dummies. La interpretación es la misma que en el caso más simple si las variables dicotómicas son excluyentes, es decir si se trata del análisis de la misma característica que tiene m + 1 categorías, en este caso el modelo se denomina de categorías múltiples.

Modelo ANCOVA

Las variables dicotómicas se pueden utilizar, lógicamente, si la estructura de los datos es una serie temporal. La variable binaria tomaría valores 1 o 0 en el tiempo en función de si para ese momento temporal se da o no un hecho determinado y de interés para el modelo. La interpretarción básicamente es la misma.

Supongamos el modelo más sencillo en el que tenemos una regresión simple a la que añadimos una variable binaria

$$Y_t = \beta_0 + \beta_1 X_{1t} + \alpha_1 D_{1t} + \varepsilon_t. \tag{4.19}$$

A la variable explicada solo le afectan dos factores, la variable cuantitativa X_{1t} y la variable dicotómica o binaria D_{1t}. La interpretación del modelo (4.19) es la siguiente: cuando se cumple la característica o acontecimiento al que hace referencia la variable binaria, entonces el término constante se descompone en la suma del término β_0 y el parámetro de la variable dummy $\alpha_1 D_{1t}$, mientras que cuando no se cumple, el término constante es solo β_0. La pendiente no se ve afectada, puesto que está determinada por el parámetro de la variable cuantitativa β_1. Cuando la característica o el acontecimiento

se cumple, D_{1t} = 1, el término constante aumenta. Para el mismo valor de la variable independiente X_{1t} la variable explicada Y_t aumenta en la cantidad α_1.

Modelos con variables que interactuan con dicotómicas

El caso más sencillo es considerar una regresión simple en la que incluimos una variable dicotómica que modifica el término constante y que también interactúa con la variable no binaria o cuantitativa. Consideremos el siguiente modelo:

$$\begin{aligned} Y_i &= \beta_0 + \alpha_0 D_{1i} + \beta_1 X_{1i} + \alpha_1 D_{1i} X_{1i} + \varepsilon_i \\ &= (\beta_0 + \alpha_0 D_{1i}) + (\beta_1 + \alpha_1 D_{1i}) X_{1i} + \varepsilon_i. \end{aligned}$$

El primer paréntesis determina el término constante: cuando la dummy tiene valor unitario, el término constante es β_0 + α_0, y cuando tiene valor nulo β_0, en términos geométricos, la predicción se desplaza paralelamente manteniendo la pendiente constante.

El segundo paréntesis modifica la pendiente. Cuando la variable binaria tiene valor unitario, la pendiente es β_1 + α_1; en caso contrario, la pendiente es β_1. Las distintas posibilidades las podemos visualizar también en la Figura 4.1, de manera que en función de los valores que tomen los parámetros de la variable dicotómica $\alpha_0 + \alpha_1$ las estimaciones pueden converger, divergir o cruzarse.

Regresión por tramos

Cuando analizamos las interacciones con variables binarias, consideramos el modelo

$$Y_i = \beta_0 + \alpha_0 D_{1i} + \beta_1 X_{1i} + \alpha_1 D_{1i} X_{1i} + \varepsilon_i. \tag{4.20}$$

Resulta también fácil ver que la expresión (4.20) equivale a calcular dos regresiones separadas.

Cuando la variable binaria tiene valor nulo, entonces el modelo es

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \tag{4.21}$$

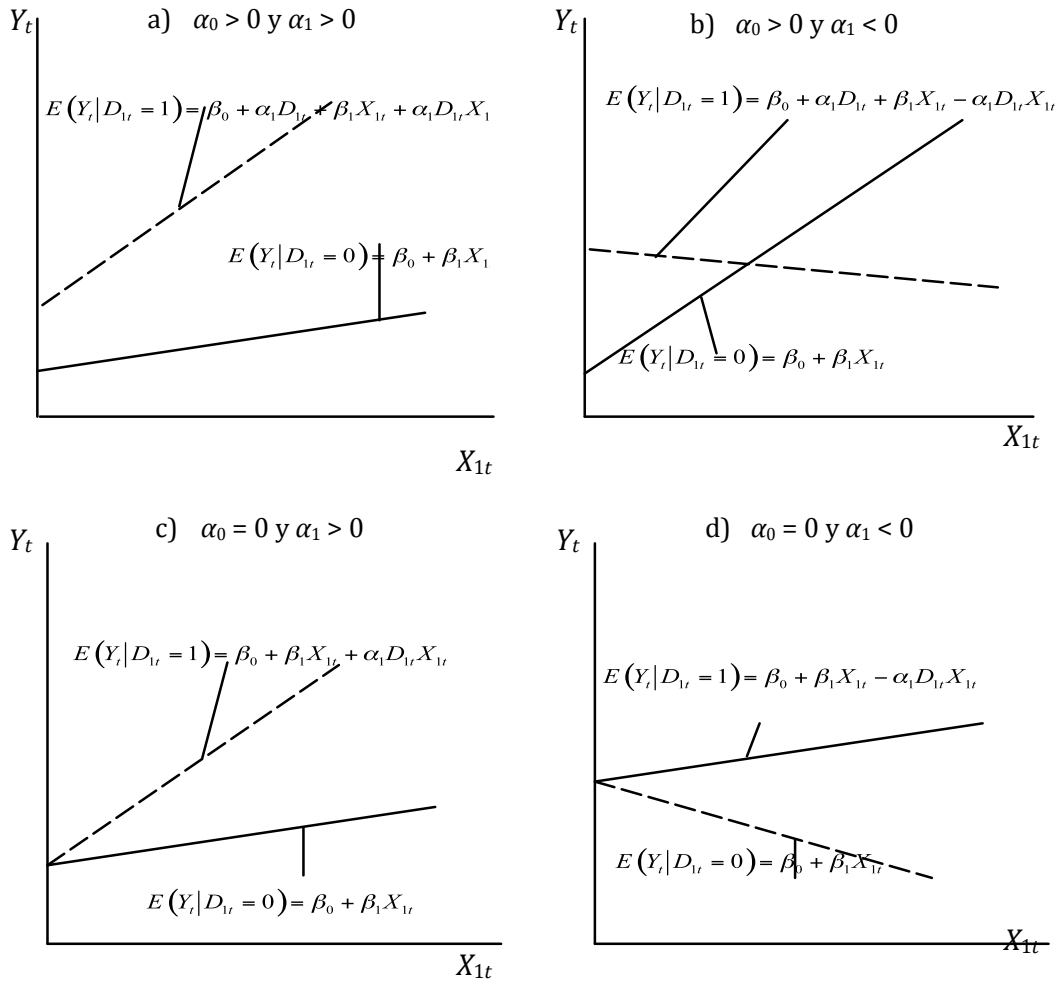
y cuando la dummy tiene valor unitario su expresión es

$$Y_i = (\beta_0 + \alpha_0) + (\beta_1 + \alpha_1) X_{1i} + \varepsilon_i. \tag{4.22}$$

Supongamos ahora que la variable ficticia D_{1i} lo que hace es dividir la variable independiente X_{1i} en dos tramos diferentes, es decir que D_{1i} tiene valor nulo si X_{1i} tiene un valor menor que un valor determinado j* y unitario si es mayor o igual que ese valor (D_{1i} = 0 si $X_{1i}<j$* y D_{1i} = 1 si $X_{1i}\geq j$*). El valor umbral j* se conoce como nudo y para transformar la función en continua para todo el recorrido (que es en lo que consiste la estimación de un modelo de regresión por tramos) tenemos que garantizar que en ese punto ambos tramos coincidan en j*, es decir, se tiene que cumplir que para $X_{1i} = j$*, las expresiones (4.21) y (4.22) coinciden

$$\begin{aligned} &\beta_o + \beta_1 j* = \beta_0 + \alpha_0 + (\beta_1 + \alpha_1) j*; \\ &0 = \alpha_0 + \alpha_1 j*; \\ &\alpha_0 = -\alpha_1 j* \end{aligned} \tag{4.23}$$

Figura 4.1: Modelos con cambio de pendiente



de manera que la regresión por tramos consiste en estimar (4.20) por mínimos cuadrados restringidos imponiendo la restricción (4.23). Sustituyendo (4.23) en (4.20) y operando tenemos que

$$\begin{aligned} Y_i &= \beta_0 + \alpha_0 D_{1i} + \beta_1 X_{1i} + \alpha_1 D_{1i} X_{1i} + \varepsilon_i \\ &= \beta_0 - \alpha_1 j * D_{1i} + \beta_1 X_{1i} + \alpha_1 D_{1i} X_{1i} + \varepsilon_i \\ &= \beta_0 + \beta_1 X_{1i} + \alpha_1 D_{1i} \left(X_{1i} - j*\right) + \varepsilon_i, \end{aligned}$$

que es el denominado modelo de regresión por tramos.

4.3.2 Selección de modelos

De cara a la práctica de la modelización hemos ofrecido varias alternativas o aspectos que vamos a sintetizar en esta sección.

En ocasiones podemos observar que los datos que tenemos para llevar a término un estudio o responder una pregunta de interés están en una escala que no nos resulta conveniente. Hemos comprobado que en esta situación habitual podemos modificar la escala sin cambiar ninguna de las relaciones económicas que subyacen entre las variables. Hemos de usar unas unidades de medida que nos sean útiles en la práctica y que nos permitan dar sentido y facilitar la comprensión de los coeficientes estimados.

El punto de partida de prácticamente todo análisis econométrico es la teoría económica. ¿Qué dice la teoría económica sobre una relación determinada de interés?¿Qué dice el sentido económico? En pocas ocasiones nos vamos a encontrar que la respuesta explícitamente diga que la relación es lineal. A veces nos encontraremos que el análisis económico puede llegar a sugerir una relación no lineal. Si bien en muchos casos la teoría no entrará directamente en esta cuestión explícitamente.

Como quiera que sea, el «económetra» o la economista debe elegir una forma algebraica para establecer la relación económica. Esto, como hemos visto, requiere elegir la «transformación» adecuada de las variables originales. Cómo hacerlo es algo que se adquiere sin duda desde la práctica, e inicialmente no es fácil. Para facilitar este proceso hemos considerado en este tema algunas transformaciones simples como potencias y logaritmos naturales. Usando estas transformaciones se abre un sorprendente abanico de posibilidades y de formas.

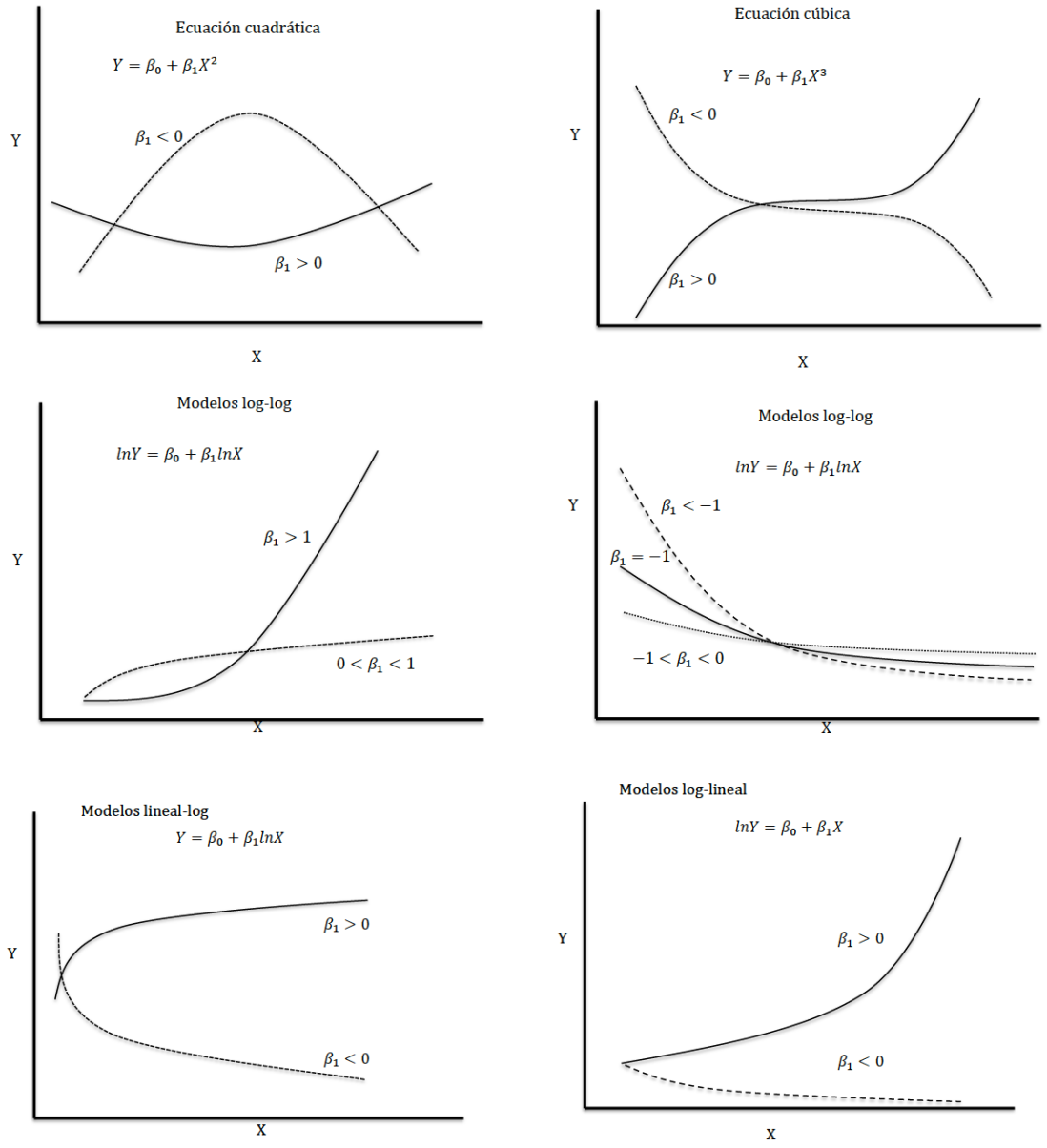
En la Figura 4.2 hemos representado varias alternativas que nos ofrece el conjugar las potencias y las relaciones con logaritmos que anticipamos en epígrafes anteriores. Como vimos entonces, cuando transformamos las variables entonces la interpretación de los resultados cambia, pues las variables ahora están relacionadas de forma no lineal.

En la Tabla 4.1 incorporamos las interpretaciones de los modelos con algunos datos en logaritmos; también introdujimos el caso cuando algún regresor está en forma de potencia. Así pues, tenemos modelos alternativos que contienen diferentes transformaciones tanto de la variable dependiente, como de las independientes. Más aún, algunas de la formas tienen ciertas semejanzas.

En términos generales, la guía más natural para elegir la forma funcional, si bien no es la única y podría matizarse en función del problema a tratar, consistiría en:

- (i) optar por una forma que sea consistente con lo que indica la teoría económica sobre la relación,
- (ii) elegir una forma que sea suficientemente flexible para «ajustar» los datos, y
- (iii) elegir una forma funcional que (mejor) asegure que los supuestos son satisfechos, de modo que los estimadores –en este caso MCO– tengan igualmente las propiedades deseadas para un estimador. El análisis de los residuos del (de los) modelo(s) estudiado(s) será revelador al respecto de la calidad del modelo seleccionado finalmente.

Figura 4.2: Formas funcionales



Resulta enormemente útil no olvidar que nunca sabemos el «verdadero» modelo, es decir, la «verdadera» relación funcional entre las variables socio-económicas. Nuestro modelo seleccionado, tras haber realizado suficientes pruebas y comprobaciones, siempre será una aproximación (y esperemos que útil).

Recuerde el lector a estos efectos lo comentados en el tema precedente a este respecto: Cuando decimos que el modelo es una aproximación nos referimos al hecho innegable de la excesiva complejidad del comportamiento económico debido entre otros a la dificultad de medir con precisión (incluso de definir con precisión aspectos determinantes del comportamiento económico) y debido a que el economista tiene poco o ningún control sobre el fenómeno bajo estudio.

En estas circunstancias resulta demasiado optimista considerar que los modelos econométricos (modelos de probabilidad) propuestos son suficientemente adecuados para capturar esta complejidad inherente. Por este motivo, resulta más ajustado considerar que un modelo econométrico (o la modelización econométrica) es una cruda aproximación a la relación (verdadera) que existe entre los datos observados.

Por último, suele ser tentador usar el R-cuadrado para seleccionar modelos. Sin embargo, conviene recordar que de los supuestos del modelo lineal ninguno requiere que el R-cuadrado esté por encima de un valor particular. Así pues la calidad de la estimación de los efectos parciales en la relación lineal de las explicativas respecto a la esperanza de la explicada no reside en un valor del R-cuadrado determinado. De hecho es posible obtener modelos con R-cuadrados bastante pequeños (lo que significa que no hemos tenido en cuenta varios factores que afectan a Y) y sin embargo esto no significa que los factores en el error estén correlacionados con las variables explicativas. El supuesto de media condicional nula es el que realmente determina si obtenemos estimadores insesgados de los efectos parciales (ceteris paribus) de las variables independientes, y el tamaño del R-cuadrado no tiene relación directa con esto. Por tanto, el R-cuadrado por sí solo no es un buen criterio para seleccionar un modelo, pese a la atracción que a priori puede tener el disponer un número acotado, como el R-cuadrado.

A estos efectos el estadístico F nos puede resultar útil para probar la significación conjunta de un grupo de variables; esto nos permite decidir, a un nivel de significación particular, si al menos una variable del grupo afecta a la variable dependiente. Sin embargo, este contraste no nos permite decidir cuál de las variables tiene efecto. Cuando tenemos que seleccionar entre modelos que están anidados (uno es un caso particular de otro), el contraste de la F sí puede ser de utilidad. En cambio, si los modelos que estamos comparando no están anidados (uno no es un caso particular del otro), entonces el contraste F no es operativo para seleccionar entre modelos. En tal caso, el R-cuadrado ajustado, que vimos en su momento, podría ser útil si se trata de la elección de la forma funcional con la que entran en el modelo las variables explicativas.

4.3.3 Predicción con datos de sección cruzada

Después de la estimación de los parámetros o coeficientes del modelo por MCO es habitual utilizar el modelo estimado para hacer una previsión de la variable dependiente.

La predicción o pronóstico consiste en valorar el modelo estimado para un escenario dado por valores particulares (de interés para el usuario) de las variables explicativas. Es decir, deseamos saber qué valor tomaría la variable dependiente para un vector de variables explicativas determinado y que denotamos por $(X_1^0, X_2^0, ..., X_k^0)$, y por tanto este vector puede ser entendido como un *escenario económico*. Supongamos que hemos estimado el modelo general siguiente

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_k X_k, \tag{4.24}$$

y que queremos realizar una predicción para los valores particulares de las variables independientes $X_1^0, X_2^0, ..., X_k^0$ donde el superíndice 0 indica valores particulares de las variables explicativas. La predicción entonces es

$$\mathbb{E}\left(Y|\, X_1^0, ..., X_k^0\right) = \hat{Y}^0 = \hat{\beta}_0 + \hat{\beta}_1 X_1^0 + ... + \hat{\beta}_k X_k^0. \tag{4.25}$$

El estimador de la predicción o predictor, expresión (4.25), es un estimador puntual, y puesto que lo hemos elaborado a partir de las estimaciones mínimo cuadráticas, expresión (4.24), está sujeto a variación muestral, es decir, el predictor está sujeto a la variabilidad de los estimadores MCO. En consecuencia debemos obtener alguna medida de la incertidumbre asociada al pronóstico realizado.

La *varianza del predictor* para la *regresión simple* y bajo el supuesto de homocedasticidad es

$$\begin{aligned}
\text{var}\left(\hat{Y}^0\,|\mathbf{X}\right) &= \text{var}\left(\hat{\beta}_0 + \hat{\beta}_1 X_1^0\,|\mathbf{X}\right) && (4.26)\\
&= \left(\ 1,\quad X_1^0\ \right)\text{var}\left(\boldsymbol{\beta}\,|\mathbf{X}\right)\left(\ 1,\quad X_1^0\ \right)' && (4.27)\\
&= \sigma^2\left(\ 1,\quad X_1^0\ \right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\ 1,\quad X_1^0\ \right)' && (4.28)\\
&= \sigma^2\left[\frac{1}{n} + \frac{\left(X_1^0 - \bar{X}_1\right)^2}{\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_1\right)^2}\right], && (4.29)
\end{aligned}$$

donde la última igualdad se obtiene operando algebraicamente y se deja como ejercicio para el lector interesado. Esta expresión para el modelo de regresión simple se puede generalizar para la regresión múltiple en términos matriciales

$$\begin{aligned}
\text{var}\left(\hat{Y}^0\right) &= \sigma^2\left[\mathbf{X^{0\prime}}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X^0}\right] && (4.30)\\
&= \sigma^2\left[n^{-1} + \left(\mathbf{X^0} - \bar{\mathbf{X}}\right)'\left(\mathbf{x}'\mathbf{x}\right)^{-1}\left(\mathbf{X^0} - \bar{\mathbf{X}}\right)\right], && (4.31)
\end{aligned}$$

donde la expresión x′x indica en esta ocasión que las variables están tomadas en diferencias respecto de sus medias. Las expresiones (4.26) y (4.30) son ilustrativas al mostrar que la varianza del predictor y, en consecuencia, también sus errores estándar aumentan a medida que las variables explicativas se alejan de sus respectivos valores medios.

Una forma alternativa de cálculo del error estándar consiste en restar las expresiones (4.24) y (4.25), de donde operando mínimamente se tiene

$$\hat{Y} = \hat{Y}^0 + \hat{\beta}_1 \left(X_1 - X_1^0\right) + ... + \hat{\beta}_k \left(X_k - X_k^0\right). \tag{4.32}$$

Esta expresión sugiere que el error estándar asociado a la constante en la expresión (4.32) coincide con el error estándar del predictor de la expresión (4.25), cuya forma de cálculo es la habitual.

Como hemos dicho, al hacer la predicción cometemos un error que denominamos ***error de predicción***; teniendo en cuenta el modelo poblacional, este error será

$$\hat{\varepsilon}^0 = Y^0 - \hat{Y}^0 = \left(\beta_0 + \beta_1 X_1^0 + ... + \beta_k X_k^0 + \varepsilon^0\right) - \hat{Y}^0. \tag{4.33}$$

La varianza del error de predicción es

$$\text{var}\left(\hat{\varepsilon}^0 \,|\mathbf{X}\right) = \text{var}\left(\varepsilon^0\right) + \text{var}\left(\hat{Y}^0\right), \tag{4.34}$$

dado que ε^0 y $\hat{Y}^0$ son independientes y el resto de términos de convarianzas entre $X_j^0 \hat{Y}^0, j = 0, 1, ..., k$ se anulan. Si utilizamos ahora la expresión (4.26) se tiene

$$\text{var}\left(\hat{\varepsilon}^0 \,|\mathbf{X}\right) = \sigma^2 \left[1 + \frac{1}{n} + \frac{\left(X_1^0 - \bar{X}_1\right)^2}{\sum_{i=1}^{n} \left(X_{1i} - \bar{X}_1\right)^2}\right].$$

Sustituyendo la varianza de los errores var(ε^0) = σ^2 por su estimador insesgado podemos estimar

$$\text{var}\left(\hat{\varepsilon}^0\right) = \text{var}\left(\hat{Y}^0\right) + \hat{\sigma}^2. \tag{4.35}$$

Una práctica extendida consiste en establecer un intervalo al 95 % de confianza. Siguiendo la regla que ya utilizamos anteriormente, podemos considerar que el valor en tablas es aproximadamente 2 y, entonces, el intervalo de confianza del predictor sería

$$\hat{Y}^0 \pm 2ee(\hat{\varepsilon}^0) = \hat{Y}^0 \pm 2\left\{\left[ee\left(\hat{Y}^0\right)\right]^2 + \hat{\sigma}^2\right\}^{\frac{1}{2}}. \tag{4.36}$$

Como hemos visto anteriormente, con frecuencia la variable dependiente es el logaritmo de la variable objetivo. Se puede demostrar que cuando la variable explicada está en logaritmos, la esperanza de la predicción en niveles, si se cumplen los supuestos del modelo lineal clásico y el supuesto de normalidad, es

$$\mathbb{E}\left(Y \,|\, \mathbf{X}\right) = \hat{Y} = \exp\left(\frac{\sigma^2}{2}\right) \exp\left(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k\right), \tag{4.37}$$

de manera que utilizar el procedimiento de estimar la predicción en niveles a partir de la expresión $\hat{Y}$= exp[$\widehat{ln(Y)}$] es por tanto incorrecto al subestimar el valor esperado, y debe ser ajustado (multiplicando por $\exp\left(\frac{\sigma^2}{2}\right)$).

No obstante, el método previsto por la expresión (4.37) es sesgado (pese a ser consistente) y además depende crucialmente de que los errores de la regresión se distribuyan normalmente. El problema del sesgo del estimador no se puede tratar fácilmente, sin embargo el de la normalidad sí es tratable. Sabemos que cuando los errores no se distribuyen normalmente, los estimadores MCO tienen buenas propiedades. Basta con suponer que el error es independiente de las variables explicativas para que podamos realizar la predicción consistente a partir de la siguiente regresión mínimo cuadrática

$$\hat{Y} = \hat{\gamma} \exp\left[\widehat{lnY}\right], \tag{4.38}$$

donde $\hat{\gamma}$ es un estimador de $\exp(\varepsilon)$. Este estimador se obtiene en tres pasos:

- Obtener los valores ajustados $\widehat{lnY_i}$ de la regresión de $ln(Y)$ sobre $X_k, k = 1, 2, ..., K$.
- Para cada i, calcular $\hat{c}_i = \exp(\widehat{lnY_i})$.
- Hacer una regresión simple sin constante de Y sobre $\hat{c}$. El coeficiente estimado constituye la estimación de γ.

A veces, es útil examinar observaciones individuales para ver si el valor real de la variable dependiente está por encima o por debajo del valor predicho; es decir, examinar los residuos de las observaciones individuales. Este proceso se denomina análisis de residuos.

Los residuos miden la desviación de los valores ajustados de los valores reales de la variable dependiente. Se pueden utilizar para:

- detectar errores de especificación del modelo;
- detectar valores atípicos u observaciones con un ajuste deficiente; y
- detectar observaciones influyentes u observaciones con un gran impacto en el modelo ajustado.

El análisis de residuos, en particular el análisis visual, puede indicar potencialmente la naturaleza de la especificación incorrecta y las formas en que se puede corregir, así como proporcionar una idea de la magnitud del efecto de la especificación incorrecta. Por el contrario, los contrastes estadísticos más formales de especificación incorrecta del modelo pueden ser cajas negras, que producen solo un número que luego se compara con un valor crítico. Además, si se contrasta al mismo nivel de significatividad (generalmente 5 %) sin tener en cuenta el tamaño de la muestra, cualquier modelo que utilice datos reales será rechazado con una muestra suficientemente grande incluso si se ajusta bien a los datos.

Para modelos lineales, un residuo se define fácilmente como la diferencia entre el valor real y ajustado. Para los modelos no lineales, la definición de un residuo no es única.

Quizás la forma más fructífera de utilizar los residuos es dibujar los residuos frente a otras variables de interés. Dichos gráficos incluyen los residuos representados:

- frente a los valores predichos de la variable dependiente, por ejemplo, para ver si el ajuste es deficiente para valores pequeños o grandes de la variable dependiente;

- frente a regresores omitidos, para ver si existe alguna relación, en cuyo caso se deben incluir los residuos;
- y frente a regresores incluidos, para ver si los regresores deben ingresar a través de una forma funcional diferente a la especificada.

Una alternativa relativamente frecuente es contrastar la normalidad de los residuos. Una primera aproximación al estudio de la normalidad de los residuos puede ser (pero no solo) hacer una inspección gráfica del histograma de los residuos del modelo estimado. Los software habituales incorporan esta utilidad.

El histograma no es más que la representación gráfica de una variable. El eje de abscisas se divide en intervalos, y en ordenadas el número de observaciones registradas dentro de cada uno de ellos {o su proporción respecto del total [(n.º obs. del intervalo)/n]}.

Algunos programas nos permiten además introducir en el mismo gráfico del histograma la distribución teórica de referencia, en este caso sería la distribución normal, para hacernos una idea sobre lo adecuado de la aproximaciónUna variación es graficar el valor real de y contra el valor predicho. Sin embargo, esta gráfica es difícil de interpretar si la variable dependiente toma solo unos pocos valores.

Como sabemos, la distribución normal se caracteriza por ser simétrica respecto a su media (lo que podemos medir mediante el coeficiente de asimetría S: si es igual a cero entonces es simétrica) y también por el apuntamiento de la distribución, es decir, si es más alta o menos que la distribución teórica normal (lo que también podemos medir mediante el coeficiente de curtosis K: si tiene el mismo apuntamiento que la distribución teórica normal entonces este coeficiente vale tres).

El estadístico Jarque-Bera, *JB*, es válido asintóticamente o para muestras grandes y es el siguiente

$$JB = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right], \tag{4.39}$$

donde S es el coeficiente de asimetría y K el de curtosis.

El estadístico sigue una chi cuadrado χ^2 con 2 grados de libertad y sirve para contrastar la hipótesis nula H_0 de que los residuos siguen una distribución normal H_0: los residuos estimados se distribuyen normalmente. Si el valor estimado es mayor que el de tablas para un determinado nivel de confianza, entonces rechazamos la hipótesis nula.

Al 95 % de confianza [o al 5 % de significatividad (α)] el valor de tablas es 5,99 ($\chi^2_{2.\alpha} = \chi^2_{2,0,05} = 5,99$), de manera que si el valor empírico de (4.39) es mayor que 5,99 entonces rechazamos la hipótesis nula y los residuos estimados no se distribuyen normalmente con el 95 % de confianza.

En el caso de los usuarios de internet el coeficiente de asimetría es 0,0978 y el de curtosis 3,838. El valor empírico del estadístico *JB* es 5,15 {167[(0,0978^2/6)+(3,838-3)^2/24] = 5,15}, luego no podemos rechazar la hipótesis de normalidad de los residuos al 95 % de confianza.

Bibliografía complementaria

Matilla-García, M et al. 2017. Econometría y Predicción. McGraw Hill

Stock J. and Watson J. Introducción a la econometría. Pearson.

Tema 5

Analisis de regresion con datos de seccion cruzada IV

Este tema está elaborado como una adaptación del capítulo 8:

Wooldridge. J. 4th Ed., Introductory Econometrics.

Así como de la bibliografía complementaria.

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al Órgano convocante ni al Tribunal actuante.

- Heterocedasticidad.
- Consecuencias para los MCO.
- Inferencia robusta a la heterocedasticidad en la estimación MCO.
- Tests de heterocedasticidad.
- Estimación por Mínimos Cuadrados Ponderados.

5.1 Heterocedasticidad.

En los temas anteriores hemos incidido en la relevancia práctica de considerar que, en los datos de naturaleza económica, la heterocedasticidad es la norma, y no la excepción. Pese a ello, esto no supone en la actualidad un problema de difícil solución. De hecho, el modelo de regresión lineal, cuyos supuestos expusimos en detalle anteriormente, y que seguidamente recopilamos, nos permite estimar y realizar inferencia estadística sobre los parámetros estimados.

Hay varios motivos para pensar que los errores son heterocedásticos. En los modelos de aprendizaje, por ejemplo, los agentes aprenden por la experiencia y lo normal es que la variabilidad de los errores se reduzca con el paso del tiempo.

En ocasiones, no pocas variables explicativas (ingresos, beneficios, educación, renta, etc.) acentúan la probabilidad de la existencia de una mayor variabilidad en el comportamiento de los agentes económicos (generalmente porque tienen más grados de libertad en su comportamiento). En estos casos lo normal es que la variabilidad residual aumente a medida que lo hacen las variables explicativas.

La mejora en las técnicas de recolección de datos provenientes de los agentes económicos también podría significar la potencial comisión de menores errores, lo que reduciría la varianza de los errores.

La presencia en la muestra de datos atípicos severos (en el sentido de ser datos muy diferentes del resto) propicia la aparicción de heterocedasticidad, especialmente cuando la muestra es pequeña.

La fuente de heterocedasticidad más preocupante se produce como consecuencia de un modelo mal especificado (por ejemplo la no inclusión de variables relevantes), o por una transformación incorrecta de los datos (estimar en niveles cuando lo correcto sería en logaritmos o en diferencias). Esta fuente de heterocedasticidad vulnera inicialmente el supuesto de exogeneidad causando que la esperanza condicionada de los errores ya no sea nula.

Normalmente el problema de heterocedasticidad es más frecuente con información de corte transversal, donde las observaciones suelen ser más heterogéneas, que con datos de series temporales.

Los supuestos bajo los que vamos a desarrollar este tema son los más generales posibles. Los recordamos a continuación

SUPUESTOS DEL **MODELO DE REGRESIÓN LINEAL**. Las observaciones $(Y_i, \mathbf{x}_i)$, $i = 1, 2, ..., n$, satisfacen la ecuación lineal de regresión

LINEALIDAD

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i. \tag{5.1}$$

MUESTRA ALEATORIA

$$(X_{1i}, X_{2i}, ..., X_{ki}, Y_i), i = 1, ..., n \text{ son } i.i.d.$$

EXOGENEIDAD

$$\mathbb{E}\left(\mathbf{x}_i \cdot \varepsilon_i\right) = \mathbf{0}. \tag{5.2}$$

NO MULTICOLINEALIDAD PERFECTA

$$\mathbb{E}(\mathbf{x}_i\mathbf{x}_i') > 0, \text{ donde } \mathbf{x}_i = (X_{0i}, X_{1i}, ..., X_{ki})'. \tag{5.3}$$

MOMENTOS ORDEN CUATRO: ATÍPICOS POCO PROBABLES

$$\mathbb{E}Y_i^4 < \infty \text{y la} \mathbb{E}\left\|X_i^4\right\| < \infty \tag{5.4}$$

5.2 Consecuencias para los MCO.

Las propiedades que establecimos en temas anteriores respecto a la consistencia e insesgadez de los coeficientes de las variables explicativas siguen siendo ciertas puesto que dichas propiedades fueron deducidas bajo los supuestos del modelo de regresión lineal y dichos supuestos no afectan al comportamiento de la varianza condicionada de los errores, $var(\varepsilon_i|\mathbf{x}_i)$. Así pues la heterocedasticidad condicionada no causa en modo alguno problemas de sesgo ni de falta de consistencia.

Tampoco tiene efectos sobre el R-cuadrado convencional ni sobre su versión corregida.

Ambos pueden interpretarse como estimadores de un R-cuadrado poblacional de la forma

$$R^2_{pob} = 1 - \frac{\sigma^2_\varepsilon}{\sigma^2_Y}$$

donde las varianzas son de la poblacion del error y de la variable dependiente. Ambas varianzas son incondicionadas, y por tanto no se ven afectadas por el hecho de que la varianza condicionada del error, $var(\varepsilon_i|\mathbf{x}_i) = \mathbb{E}(\varepsilon_i^2|\mathbf{x}_i)$, sea o no constante a lo largo de i. Si $(Y_i, \mathbf{x}_i)$ son iid, una vez que condicionamos por $\mathbf{x}_i$, la media y la varianza condicionadas pueden varíar $\mathbf{x}_i$. Lo mismo le sucederá al error

$$\varepsilon_i = Y_i - \mathbf{x}_i'\boldsymbol{\beta}$$

es decir, será iid pero una vez que condicionamos por $\mathbf{x}_i$, y consideramos la distribución de ε_i condicionada a $\mathbf{x}_i$, la varianza condicional de esta distribución puede variar con $\mathbf{x}_i$, pese a la varianza incondicinal sea constante. Por otra parte, sabemos que los estimadores que incorpora el R-cuadrado (ajustado o no) para estimar σ^2_ε y σ^2_Y son consistentes, independientemente de si la varianza condicionada es constante o no.

El problema de tener errores condicionalmente heterocedásticos, como sucede habitualmente en los datos de sección cruzada, es que la inferencia se realiza de forma incorrecta si utilizamos los estimadores de la varianza de los coeficientes deducidos bajo el supuesto de homocedasticidad condicionada. En particular, dado que esta varianza es fundamental para calcular el error estándar del parámetro estimado, si el error estandar $ee(\hat{\beta}_j)$ es sesgado, también lo será la inferencia que realicemos a partir estos errores en presencia de heterocedasticidad, como sucede cuando realizamos un contraste de significatividad del parámetro. Lo mismo sucede con los contrastes tipo F y tipo LM que vimos en el tema dedicado a los contrastes de hipótesis.

Lo interesante es que en el marco del modelo de regresión que hemos planteado, es posible obtener errores estándar de los coeficientes estimados de tal menera que sean válidos si hay heterocedasticidad condicionada aunque dicha heterocedasticidad sea desconocida o incluso si hay homocedasticidad. Este es el objeto del siguiente epígrafe.

5.3 Inferencia robusta a la heterocedasticidad en la estimación MCO.

Recordemos, como vimos en la expresión (3.38), que bajo los supuestos de que $(Y_i, \mathbf{x}_i)$ sea iid, $\mathbb{E}(\mathbf{x}_i \cdot \varepsilon_i) = \mathbf{0}$, $\boldsymbol{\Sigma}_{\mathbf{xx}} \equiv \mathbb{E}(\mathbf{x}\mathbf{x}')$ es definida positiva, y momentos de orden cuatros finitos

$$\sqrt{n}\,(\mathbf{b}_n - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}),$$

donde

$$\mathbf{V} \equiv \boldsymbol{\Sigma}^{-1}_{\mathbf{xx}} \boldsymbol{\Omega} \boldsymbol{\Sigma}^{-1}_{\mathbf{xx}}, \tag{5.5}$$

En esta sección tratamos la estimación consistente de $\mathbf{V}$ en condiciones no homocedásticas, es decir, se trata de una estimación general, que tiene como caso particular la homocedasticidad.

En el caso homocedástico y también en el heterocedástico es fundamental estimar consistentemente la varianza del término error $\boldsymbol{\varepsilon}$. El siguiente resultado garantiza que los estimadores habituales como $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n e_i^2$ o $s^2 = \frac{1}{n-K}\sum_{i=1}^n e_i^2$ son consistentes.

Bajo los supuestos de que $(Y_i, \mathbf{x}_i)$ sea iid, $\mathbb{E}(\mathbf{x}_i \cdot \varepsilon_i) = \mathbf{0}$, $\boldsymbol{\Sigma}_{\mathbf{xx}} \equiv \mathbb{E}(\mathbf{x}\mathbf{x}')$ es definida positiva, y momentos de orden cuatros finitos, resulta que

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K} \xrightarrow{p} \sigma^2 \text{ donde } \sigma^2 = \mathbb{E}\left(\varepsilon_i^2\right).$$

Demostración: $s^2 = \frac{1}{n-K}\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} = \frac{n}{n-K}\left(\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} - \frac{\boldsymbol{\varepsilon}'\mathbf{X}}{n}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right)$. Por un lado, la demostración del teorema precedente permite establecer, por Ley de los Grandes Números, que $(\mathbf{X}'\mathbf{X})/n = n^{-1}\sum_{i=1}^n (\mathbf{x}_i\mathbf{x}_i') \xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{xx}}$ y $(\mathbf{X}'\varepsilon)/n = n^{-1}\sum_{i=1}^n \mathbf{x}_i\varepsilon_i \xrightarrow{p} \mathbf{0}$. Si usamos las propiedades del operador límite en probabilidad ($plim$), se tiene que

$$plim\left(\frac{\varepsilon'\mathbf{X}}{n}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\frac{\mathbf{X}'\varepsilon}{n}\right) = plim\frac{\varepsilon'\mathbf{X}}{n}plim\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}plim\frac{\mathbf{X}'\varepsilon}{n}$$

y por tanto converge a cero. Por otra parte, asintóticamente el término $\frac{n}{n-K}$ converge a 1, y como resultado $plim\, s^2 = plim\frac{\varepsilon'\varepsilon}{n} = plim\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2$, es decir, la media de una variable aleatoria. Como tal es posible aplicar la Ley de los Grandes Números de nuevo, ya que los supuestos garantizan que ε_i son iid y que el momento de segundo orden de ε_i^2 (esto es, el momento de orden cuarto) también existe, y por tanto $plim\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2\right) (= plim\left(s^2\right)) = \sigma^2$.

Una vez que tenemos un estimador consistente de la varianza de los incondicionada de los errores, H. White propuso una forma para estimar consistentemente $\boldsymbol{\Omega} (= \mathbb{E}(\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i') = \text{var}(\mathbf{x}_i\varepsilon_i))$ y la recogemos en el siguiente teorema.

TEOREMA: ESTIMADOR CONSISTENTE DE $\mathbf{V}$ Bajo los supuestos de que $(Y_i, \mathbf{x}_i)$ sea iid, $\mathbb{E}(\mathbf{x}_i \cdot \varepsilon_i) = \mathbf{0}$, $\boldsymbol{\Sigma}_{\mathbf{xx}} \equiv \mathbb{E}(\mathbf{x}\mathbf{x}')$ es definida positiva, y momentos de orden cuatros finitos, resulta que

$$\hat{\mathbf{V}} = \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Omega}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} = \mathbf{V}$$

donde $\hat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^n (e_i^2\mathbf{x}_i\mathbf{x}_i') = \mathbf{e}'\mathbf{e}\,(\mathbf{X}'\mathbf{X})/n$, y $\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}} = (\mathbf{X}'\mathbf{X})/n = n^{-1}\sum_{i=1}^n (\mathbf{x}_i\mathbf{x}_i')$.

Por lo dicho anteriormente basta con demostrar $\hat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$. Es decir, mostraremos que

$$\frac{1}{n}\sum_{i=1}^n (e_i^2\mathbf{x}_i\mathbf{x}_i') \xrightarrow{p} \mathbb{E}(\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i').$$

Para ello partimos de

$$\mathbf{e}'\mathbf{e}\left(\mathbf{X}'\mathbf{X}\right)/n = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}\left(\mathbf{X}'\mathbf{X}\right)/n = \left(\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n} - \frac{\boldsymbol{\varepsilon}'\mathbf{X}}{n}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right)\left(\mathbf{X}'\mathbf{X}\right).$$

Sabemos que el segundo término del paréntesis converge en probabilidad a cero, $plim\frac{\boldsymbol{\varepsilon}'\mathbf{X}}{n}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n} = 0$. Por tanto $plim\left(\mathbf{e}'\mathbf{e}\left(\mathbf{X}'\mathbf{X}\right)/n\right) = plim\left(\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n}(\mathbf{X}'\mathbf{X})\right) = plim\frac{1}{n}\sum_{i=1}^n (\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i')$. (a) La Suposición

3.3.2 garantiza que la variable aleatoria $(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i')$ tenga definida su media $\mathbb{E}(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i') < \infty$ y su varianza. (b) Igualmente, el supuesto de iid garantiza que la variable aleatoria $(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i')$ también es iid. Por (a) y (b) se cumplen las condiciones para aplicar la Ley de los Grandes Números, y por lo tanto $plim \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i') = \mathbb{E}(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i')$.

El estimador de la matriz de covarianzas, $\hat{\mathbf{V}}$, permite obtener, para el caso en el que $\boldsymbol{\beta}$ es un vector, los errores estándar de los parámetros del vector,

$$ee(\beta_j), j = 1, 2, ..., k : n^{-1/2} \sqrt{\hat{\mathbf{V}}_{jj}}$$

donde el subíndice (j, j) indica el elemento j-ésimo de la diagonal principal de la matriz de varianzas y covarianzas. Cuando los errores estándar son calculados por este procedimiento es habitual decir que los errores estándar son *robustos* a la heterocedasticidad, precisamente porque son asintóticamente válidos para cualquier tipo de heterocedasticidad.

Errores estándar asintóticos: homocedasticidad y heterocedasticidad

Reconsideremos inicialmente la expresión de la varianza asintótica de $\sqrt{n}\,(\mathbf{b}_n - \boldsymbol{\beta})$, esto es en $\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\Omega} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$, bajo los supuestos del modelo de regresión lineal normal expuestos en otro tema. El supuesto de homocedasticidad quedaba reformulado para muestras aleatorias simples como $\mathbb{E}(\varepsilon_i^2 \mid \mathbf{x}_i) = \sigma^2 > 0 \quad (i = 1, 2, ..., n)$. En ese caso

$$\begin{aligned} \boldsymbol{\Omega} &\equiv \mathbb{E}(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i') \\ &= \mathbb{E}(\mathbb{E}(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i' \mid \mathbf{x}_i)) \\ &= \mathbb{E}(\mathbb{E}(\varepsilon_i^2 \mid \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i') \\ &= \sigma^2 \mathbb{E}(\mathbf{x}_i \mathbf{x}_i') = \sigma^2 \boldsymbol{\Sigma}_{\mathbf{xx}}, \end{aligned}$$

y por tanto bajo estos supuestos tendríamos que la varianza asintótica de $\sqrt{n}\,(\mathbf{b}_n - \beta)$ sería

$$\mathbf{V}_0 = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\Omega} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} = \sigma^2 \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$$

El estimador más obvio que podemos utilizar para estimar $\mathbf{V}_0$ será

$$\hat{\mathbf{V}}_0 = s^2 \left(\mathbf{X}'\mathbf{X}/n\right)^{-1}$$

toda vez que $s^2 \xrightarrow{p} \sigma^2$ y $\mathbf{X}'\mathbf{X}/n \xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{xx}}$, ya que en ese caso $\hat{\mathbf{V}}_0 \xrightarrow{p} \mathbf{V}_0$.

La cuestión interesante desde la óptica del modelo de regresión propuesto en este tema, es decir, el compatible con la heterocedasticidad, es que este modelo sugiere que la regresión se interprete como una aproximación a la función de esperanza condicionada. Bajo este punto de vista, vamos a ver que la heterocedasticidad surge de forma natural.

Si la función de esperanza condicionada es no lineal y utilizamos el estimador MCO para aproximarla, entonces la calidad del ajuste entre la línea de regresión y la función de esperanza condicionada variará con $\mathbf{x}_i$. En promedio los residuos serán mayores

para aquellos valores de $\mathbf{x}_i$ donde el ajuste sea más pobre. La siguiente expresión nos permite ver el motivo:

$$\begin{aligned}\mathbb{E}\left[(Y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \,|\mathbf{x}_i\right] = \\ &= \mathbb{E}\left\{(Y_i - \mathbb{E}\,(Y_i\,|\mathbf{x}_i\,) + \mathbb{E}\,(Y_i\,|\mathbf{x}_i\,) - \mathbf{x}_i'\boldsymbol{\beta})^2\,|\mathbf{x}_i\right\} \\ &= \text{var}\,(Y_i\,|\mathbf{x}_i\,) + (\mathbb{E}\,(Y_i\,|\mathbf{x}_i\,) - \mathbf{x}_i'\boldsymbol{\beta})^2\,.\end{aligned}$$

El segundo término es distinto de cero al ser $\mathbb{E}\,(Y_i\,|\mathbf{x}_i\,)$ no lineal. Por tanto, incluso si var $(Y_i\,|\mathbf{x}_i\,)$ fuera constante, la varianza de los residuos aumentaría con el cuadrado de la discrepancia entre la recta de regresión y la función de esperanza condicionada. Por este motivo, la utilidad práctica nos conduce a optar por usar los errores estándar robustos. Generalmente se dice robusto porque, en muestras grandes, los errores estándar robustos proporcionan contrastes de hipótesis precisos a partir de mínimos supuestos sobre los datos y el modelo.

Los estimadores consistentes para el supuesto de homocedasticidad y para el caso robusto a la heterocedasticidad son, respectivamente,

$$\hat{\mathbf{V}}_0 = s^2\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \tag{5.6}$$

y

$$\hat{\mathbf{V}} = \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-\mathbf{1}}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-\mathbf{1}} = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(e_i^2\mathbf{x}_i\mathbf{x}_i')\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{5.7}$$

Es necesario aclarar siempre qué tipo de estimador de la varianza se está utilizando, y esto nos indicará bajo qué supuestos (modelo) se está trabajando. Estos dos tipos de estimadores nos conducen a los errores estándar que generalmente son los más utilizados. El primero por razones históricas en la evolución de la Econometría y de la del propio software econométrico. El segundo porque es el que se ha establecido como estimador robusto, si bien hay otras alternativas que a continuación comentaremos. Antes, sin embargo, queremos llamar la atención sobre una cuestión práctica en el uso habitual del estimador robusto (5.7), nos referimos al estimador de la varianza de $\sqrt{n}\,(\mathbf{b}_n - \boldsymbol{\beta})$, de donde podemos deducir la varianza de $\mathbf{b}$,

$$\widehat{\text{var}}\,(\mathbf{b}) = n^{-1}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\hat{\boldsymbol{\Omega}}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}.$$

La segunda observación es que la matriz $\hat{\boldsymbol{\Omega}}$ en ocasiones es ligeramente distinta dado que estamos ajustando el potencial sesgo a la baja debido a la estimación de $K = k + 1$ coeficientes de regresión, al dividir entre $n - K$ en lugar de entre n. No obstante, los resultados asintóticos son equivalentes.

Errores estándar asintóticos: alternativas robustas

Si retomamos la expresión de la varianza teórica del vector de discrepancias entre los parámetros estimados y verdaderos, que dimos en la expresión (3.38), dicha expresión podemos reescribirla de esta manera

$$\begin{aligned}\mathbf{V} &= \mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Omega}\mathbf{\Sigma}_{\mathbf{xx}}^{-1} \\ &= \mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbb{E}(\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i')\mathbf{\Sigma}_{\mathbf{xx}}^{-1} \\ &= \mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{X}'\mathbb{E}(\varepsilon_i^2)\mathbf{X}\mathbf{\Sigma}_{\mathbf{xx}}^{-1},\ i=1,..,n.\end{aligned}$$

La matriz de varianzas y covarianzas del error, $\mathbb{E}(\varepsilon_i^2),\ i=1,2,...n$, recoge las varianzas de los errores para cada elemento de la muestra, y las potenciales covarianzas entre los distintos errores individuales. Bajo el supuesto de muestreo aleatorio, estas covarianzas son nulas. Por tanto, la matriz $\mathbb{E}(\varepsilon_i^2)$ no es más que la matriz diagonal que definimos como

$$\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2),$$

que en el caso especial de errores homocedásticos se reduce a la matriz $\mathbf{I}_n\sigma^2$.

En el caso heterocedástico, la varianza condicionada de $\sqrt{n}\left(\mathbf{b}_n - \boldsymbol{\beta}\right)$ es

$$\text{var}\left(\sqrt{n}\left(\mathbf{b}-\boldsymbol{\beta}\right)|\mathbf{X}\right) = \text{var}\left(\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{X}'\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\mathbf{X}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}|\mathbf{X}\right),$$

donde $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} = \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X})$; de manera que entonces la varianza condicionada del estimador MCO, $\mathbf{b}$, será, utilizando la definición de $\mathbf{\Sigma}_{\mathbf{xx}} = \mathbb{E}(\mathbf{X}'\mathbf{X})$,

$$\text{var}\left(\mathbf{b}|\mathbf{X}\right) = \frac{1}{n}\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\mathbf{X}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right) = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\left(\mathbf{X}'\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}\mathbf{X}\right)\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}.$$

Un estimador de la varianza condicionada del estimador MCO que fuera robusto a la heterocedasticidad consistiría en localizar estimadores de $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} = \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$. El estimador consistente de White del Teorema 5.3 se basa en utilizar los residuos estimados, es decir, $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} = \mathbf{e}'\mathbf{e} = diag(e_1^2, e_2^2, ..., e_n^2)$, que define exactamente a $\hat{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^n(e_i^2\mathbf{x}_i\mathbf{x}_i') = \mathbf{e}'\mathbf{e}\left(\mathbf{X}'\mathbf{X}\right)/n$.

Una segunda alternativa es utilizar los residuos MCO, pero estandarizados. Para obtener la expresión matricial de los residuos estandarizados recurrimos a la expresión del proyector $\mathbf{M}$, que recordemos era

$$\underset{n\times n}{\mathbf{M}} = \mathbf{I}_n - \mathbf{P} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$

cuyos elementos de la diagonal principal los denotamos por $(1-h_{ii})$ para $i=1,...,n$. El elemento $h_{ii} = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$, que es el elemento i-ésimo de la diagonal principal de la matriz de proyección $\mathbf{P}$. Llamamos

$$\mathbf{M}^* = diag\left\{(1-h_{11})^{-1}, (1-h_{22})^{-1}, ..., (1-h_{nn})^{-1}\right\}.$$

Los residuos MCO los podemos expresar

$$\begin{aligned}\mathbf{My} &= \mathbf{e}\\ \mathbf{M}(\boldsymbol{\varepsilon} + \mathbf{X}\boldsymbol{\beta}) &= \mathbf{e}\\ \mathbf{M}\boldsymbol{\varepsilon} &= \mathbf{e}.\end{aligned}$$

Si queremos estandarizar los residuos $\mathbf{e}$ a fin de que tengan una varianza condicionada constante, entonces primero vemos cómo es la varianza condicionada, y posteriormente re-escalamos. La varianza es

$$\begin{aligned}\text{var}\left(\mathbf{e}\,|\mathbf{X}\right) &= \text{var}\left(\mathbf{M}\boldsymbol{\varepsilon}\,|\mathbf{X}\right)\\ &= \mathbf{M}\text{var}\left(\boldsymbol{\varepsilon}\,|\mathbf{X}\right),\end{aligned}$$

por lo que el factor de escala consiste en dividir cada e_i por la raíz cuadrada del elemento i-ésimo de la diagonal principal de la matriz $\mathbf{M}$. Por tanto el residuo estándar, e_i^*, sería

$$e_i^* = (1 - h_{ii})^{-1/2} e_i,$$

o matricialmente

$$\mathbf{e}^* = \mathbf{M}^{*1/2}\mathbf{e}.$$

A partir de los errores estandarizados $\mathbf{e}^*$, estimamos $\boldsymbol{\Sigma}_{\varepsilon\varepsilon} = \mathbb{E}(\varepsilon\varepsilon')$ del siguiente modo

$$\bar{\boldsymbol{\Sigma}}_{\varepsilon\varepsilon} = \mathbf{e}^{*\prime}\mathbf{e}^* = diag(e_1^{*2}, e_2^{*2}, ..., e_n^{*2}).$$

En este caso tendríamos que $\overline{\boldsymbol{\Omega}} = \frac{1}{n}\mathbf{e}^{*\prime}\mathbf{e}^*(\mathbf{X}'\mathbf{X})$. Y por tanto, la matriz de varianzas robusta a la heterocedasticidad sería

$$\begin{aligned}\overline{\text{var}\left(\mathbf{b}\,|\mathbf{X}\right)} &= \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\left(\mathbf{X}'\bar{\boldsymbol{\Sigma}}_{\varepsilon\varepsilon}\mathbf{X}\right)\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\\ &= \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(e_i^{*^2}\mathbf{x}_i\mathbf{x}_i')\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}.\end{aligned}$$

Otra tercera alternativa para obtener un estimador robusto a la heterocedasticidad es utilizar el residuo de la predicción MCO, más conocido como error de predicción. Ahora vamos a dar una formulación matricial que complementa lo tratado en dicha sección.

La estimación de residuos MCO, e_i, no son los verdaderos errores que cometeríamos al hacer una predicción, dado que su construcción está basada en la muestra completa incluyendo, por tanto, Y_i. Este término de la variable a explicar Y_i no está disponible cuando haces su predicción. Una predicción adecuada de Y_i debería basarse en las estimaciones utilizando solo las observaciones distintas de la i-ésima. Esto se puede hacer fácilmente definiendo el estimador MCO del vector $\boldsymbol{\beta}$ que deja dicha observación fuera del proceso de cálculo, es decir, se obtiene en realidad a partir de la muestra con

$n-1$ observaciones, al excluir la observación i-ésima:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{(-i)} &= \left(\frac{1}{n-1}\sum_{j\neq i}^{n}\left(\mathbf{x}_j\mathbf{x}_j'\right)\right)^{-1}\left(\frac{1}{n-1}\sum_{j\neq i}^{n}\mathbf{x}_j y_j\right)\\ &= \left(\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}\right)^{-1}\mathbf{X}_{(-i)}'\mathbf{y}_{(-i)}.\end{aligned}$$

Una expresión útil alternativa a estas dos últimas es

$$\widehat{\boldsymbol{\beta}}_{(-i)} = \mathbf{b} - (1-h_{ii})^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i, \tag{5.8}$$

cuya obtención se encuentra en el Apéndice técnico de este tema.

La predicción para Y_i con el estimador MCO que excluye (deja una fuera) del vector $\boldsymbol{\beta}$ es

$$Y_i^{**} = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)},$$

y el error de predicción o residuo del estimador MCO «excluyente» es la ecuación

$$e_i^{**} = Y_i - Y_i^{**}.$$

A partir de esta expresión del error de predicción, y utilizando (5.8) tenemos

$$\begin{aligned}e_i^{**} &= Y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}\\ &= Y_i - \mathbf{x}_i'\mathbf{b} + (1-h_{ii})^{-1}\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i\\ &= e_i + (1-h_{ii})^{-1}h_{ii}e_i\\ &= (1-h_{ii})^{-1}e_i.\end{aligned}$$

Esta última expresión nos indica que el cómputo del error de predicción solo requiere un ajuste lineal en el residuo MCO.

Utilizando este residuo o error de predicción, como decíamos antes, podemos estimar la matriz $\boldsymbol{\Sigma}_{\varepsilon\varepsilon} = \mathbb{E}(\varepsilon\varepsilon')$ del siguiente modo:

$$\tilde{\boldsymbol{\Sigma}}_{\varepsilon\varepsilon} = \mathbf{e}^{**\prime}\mathbf{e}^{**} = diag(e_1^{**2}, e_2^{**2}, ..., e_n^{**2}).$$

En este caso la matriz tendríamos que $\widetilde{\boldsymbol{\Omega}} = \frac{1}{n}\mathbf{e}^{**\prime}\mathbf{e}^{**}(\mathbf{X}'\mathbf{X})$. Y por tanto, la matriz de varianzas robusta a la heterocedasticidad sería

$$\begin{aligned}\widetilde{\text{var}\,(\mathbf{b}\,|\mathbf{X})} &= \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\left(\mathbf{X}'\tilde{\boldsymbol{\Sigma}}_{\varepsilon\varepsilon}\mathbf{X}\right)\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\\ &= \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(e_i^{**^2}\mathbf{x}_i\mathbf{x}_i')\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}.\end{aligned}$$

Estos tres estimadores robustos de la matriz de varianzas y covarianzas difieren entre ellos por los distintos estimadores utilizados para estimar la matriz $\boldsymbol{\Omega}$. En la demostración del Teorema 5.3 hemos usado un estimador consistente de dicha matriz, es decir

$\widehat{\mathbf{\Omega}} \xrightarrow{p} \mathbf{\Omega}$. Para verificar que sus homólogos $\bar{\mathbf{\Omega}}$ y $\tilde{\mathbf{\Omega}}$ también son consistentes basta con comprobar que cuando $n \to \infty$ sus respectivas diferencias $\bar{\mathbf{\Omega}} - \widehat{\mathbf{\Omega}}$ y $\tilde{\mathbf{\Omega}} - \widehat{\mathbf{\Omega}}$ convergen a cero en probabilidad. La demostración consiste en demostrar que asintóticamente la influencia de cualquier individuo de una muestra grande es despreciable, esto es

$$\max_{1 \leq i \leq n} h_{ii} = o_p(1).$$

Los estimadores robustos alternativos que hemos propuesto no aparecen en todos los paquetes informáticos. Cuando lo hacen para localizarlos, habitualmente, tenemos que señalar la opción de estimadores robustos, y posteriormente optar por los que están disponibles, que suelen denotarse mediante los acrónimos hc1, hc2,...

5.4 Contrastes de heterocedasticidad.

Los errores estándar robustos a la heterocedasticidad vistos anteriormente son en la actualidad la forma más común de enfrentarse a la heterocedasticidad en los análisis de datos de sección cruzada. No en vano, los errores estandar robustos proporcionan una forma sencilla de calcular los estadísticos tipo t con una distribución asintótica conocidad con independencia de que esté presente la heterocedasticidad. Lo mismo sucede con los contrastes tipo F y LM robustos a la heterocedasticidad.

Por otra parte, en ciertas situaciones es interesante desde un punto de vista económico saber si la varianza condicionada es una función de los regresores. En estos casos, la literatura ofrece varios constrastes estadísticos de homocedasticidad (heterocedasticidad). Es importante considerar este marco para saber para qué y para qué no sirven los contrastes de heterocedasticidad. Utilizar un contraste de este tipo para determinar si utilizar MCO u otra técnica de estimación (ver siguiente apartado), o para saber si usar errores estándar habituales o los robustos, constituye un uso relativamente poco adecuado de un contraste de hipótesis sobre heterocedasticidad. Un contraste de heterocedasticidad debería utilizarse para contestar a la pregunta de si la varianza condicionada es una función de las variables explicativas siempre se tenga interés realmente en la forma de la varianza condicionada.

Se han propuesto muchos contrastes de heterocedasticidad y generalmente los programas especializados los realizan de forma rutinaria. Algunos de ellos son capaces de detectar de forma directa la heterocedasticidad, pero no contrastan de forma directa el supuesto de que la varianza de los errores no depende de las variables independientes. Nos limitaremos en esta sección a indicar el contraste de Breusch-Pagan y el contraste de White.

Partimos del modelo lineal general, esto es

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i,$$

junto con el resto de supuestos que hemos determinado al comienzo del tema y que configuran el modelo de regresión. Recordemos que este conjunto de supuestos no incorpora el de homocedasticidad

$$H_0 : \text{var}\left(\varepsilon_i | \mathbf{X}\right) = \text{var}\left(\varepsilon_i\right) = \mathbb{E}\left(\varepsilon_i^2\right) = \sigma^2. \tag{5.9}$$

El objetivo es contrastar si ${\varepsilon_i}^2$ se relaciona, en valor esperado, con una o más variables explicativas. Una forma simple es suponer una función lineal del tipo siguiente

$$\varepsilon_i^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + ... + \alpha_k X_{ki} + e_i, \tag{5.10}$$

que en el caso de homocedasticidad de la expresión (5.9), se cumple (5.10) si

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0.$$

Para contrastar la homocedasticidad de los errores de la expresión (5.1), podemos utilizar un estadístico tipo F de significatividad global de las variables explicativas de la expresión (5.10), que tiene una justificación asintótica.

Evidentemente nunca conoceremos los verdaderos errores ε_i pero sí su estimación $\hat{\varepsilon}_i$. El uso de los residuos MCO en lugar de los errores no afecta a la distribución asintótica de estadísticos tipo F o LM. Así pues, podemos estimar

$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + ... + \alpha_k X_{ki} + u_i. \tag{5.11}$$

Desde esta regresión del cuadrado de los residuos del modelo original, se construyen los tests. Como hemos visto anteriormente, los estadísticos F dependen del R^2 de la regresión (5.11), para no generar equívocos nos referiremos al R-cuadrado de la regresión secundaria como $R^2_{\hat{\varepsilon}^2}$. El estadístico F en consecuencia es

$$F = \frac{R^2_{\hat{\varepsilon}^2}/k}{\left(1 - R^2_{\hat{\varepsilon}^2}\right)/(n-k-1)},$$

que se distribuye, bajo la hipótesis nula, como una F de Snedecor con k y n - k - 1 grados de libertad ($F_{k,n-k-1}$).

También podemos construir y utilizar el estadístico de tipo *LM*, que se calcula:

$$LM = n \cdot R^2_{\hat{\varepsilon}^2}, \tag{5.12}$$

que se distribuye, bajo la hipótesis nula, como una chi cuadrado con k grados de libertad ${\chi^2}_k$.

A esta última forma de efectuar el contraste se la conoce como contraste de heterocedasticidad de Breusch-Pagan (BP-test). Si el valor empírico del contraste BP es mayor que el valor crítico para un determinado nivel de significatividad entonces rechazamos la hipótesis nula de homocedasticidad y en consecuencia concluimos que los residuos son heterocedásticos.

Halbert White propuso un contraste parecido al de BP en el que de un modo muy intuitivo añade los cuadrados y productos cruzados de todas las variables independientes (distintas de la constante) de la expresión (5.11).

La lógica subyacente al contraste es que el supuesto de homocedasticidad puede ser reemplazado por uno menos exigente. En concreto, bastaría que el cuadrado de los errores estuviera no correlacionado con ninguna de las variables explicativas, ni con sus cuadrado y tampoco con los productos cruzados entre variables.

En el caso de dos variables explicativas, el test consistiría primero en estimar

$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{1i} X_{2i} + \alpha_4 X_{1i}^2 + \alpha_5 X_{2i}^2 + u_i$$

Seguidamente, en construir el estadístico de contraste tipo *LM* ya mostrado en la expresión (5.12).

Si el número de variables explicativas crece, el número de parámetros a estimar crece de forma no lineal, y se pierden grados de libertad. En efecto, con tres regresores, tendríamos un modelo con nueve variables independientes y una constante; pero cuatro regresores, tendríamos catorce más la constante, y así sucesivamente. Un recurso para evitar la pérdida de grados de libertad es usando los valores estimados de la variable dependiente, $\hat{Y}_i$. Esto es así porque estos valores estimados son una combinación lineal de los parámetros, de modo que haciendo el cuadrado de esta combinación lineal recogeríamos una combinación de cuadrados y producto cruzados. Esta idea conduce a utilizar la siguiente regresión para reconducir el test de White

$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 \hat{Y}_i + \alpha_2 \hat{Y}_i^2 + v$$

En tal caso los contrastes tipo F y tipo LM utilizarían la hipótesis nula

$$H_0 : \alpha_1 = \alpha_2 = 0$$

y bajo esta nula de homocedasticidad, los contrastes tendrían las ditribuciones comentadas anteriormente pero con muchos menos grados de libertad.

5.5 Estimación por Mínimos Cuadrados Ponderados.

Una de las consecuencias de tener heterocedasticidad es que el estimador MCO ya no es el mejor estimador lineal insesgado. En esta sección veremos por qué y cómo es posible obtener un mejor estimador que MCO cuando se conoce la forma de heterocedasticidad.

La cuestión relevante es la forma que presenta el estimador MCO cuando consideramos la varianza de este estimador. Para ello comprobemos cómo es la varianza del estimador de coeficientes MCO en el modelo de regresión anterior, que no excluye la heterocedasticidad condicionada:

La matriz de varianzas del vector error de regresión $\boldsymbol{\varepsilon}$ es la matriz $n \times n$ siguiente:

$$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'} = \mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \,|\mathbf{X}\right),$$

donde el elemento i-ésimo de la diagonal principal es

$$\mathbb{E}(\varepsilon_i^2 \,|\mathbf{x}_i) = \sigma_i^2,$$

mientras que los elementos fuera de la diagonal de la matriz $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'}$son

$$\mathbb{E}(\varepsilon_i \varepsilon_j \,|\mathbf{X}) = \mathbb{E}(\varepsilon_i \,|\mathbf{x}_i)\mathbb{E}(\varepsilon_j \,|\mathbf{x}_j) = 0,$$

al ser independientes las observaciones j e i-ésimas (por el supuesto de muestra aleatoria) .

La varianza del estimador MCO, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{A}'\mathbf{y}$, donde definimos $\mathbf{A}(\mathbf{X}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ será entonces

$$\begin{aligned} \text{var}\left(\hat{\boldsymbol{\beta}}\,|\mathbf{X}\right) &= \text{var}\left(\mathbf{A}'\mathbf{y}\,|\mathbf{X}\right) \\ &= \text{var}\left(\mathbf{A}'\boldsymbol{\varepsilon}\,|\mathbf{X}\right) \\ &= \mathbf{A}'\boldsymbol{\Sigma}_{\varepsilon\varepsilon'}\mathbf{A} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\varepsilon\varepsilon'}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \end{aligned}$$

que no es más que una versión ponderada de la matriz $\mathbf{X}'\mathbf{X}$ al ser el término

$$\mathbf{X}'\boldsymbol{\Sigma}_{\varepsilon\varepsilon'}\mathbf{X} = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\sigma_i^2.$$

Observamos pues que la var $\left(\hat{\boldsymbol{\beta}}\,|\mathbf{X}\right)$ o, mejor, su versión convenientemente escalada,

$$\begin{aligned} \text{var}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)|\mathbf{X}\right) &= n\text{var}\left(\hat{\boldsymbol{\beta}}\,|\mathbf{X}\right) \\ &= n\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\varepsilon\varepsilon'}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\Sigma}_{\varepsilon\varepsilon'}\mathbf{X}\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}, \end{aligned} \tag{5.13}$$

no es tan fácilmente accesible dado que desconocemos los n elementos de la matriz $\boldsymbol{\Sigma}_{\varepsilon\varepsilon'}$. Nótese que en el caso homocedástico, esta matriz se reduce a una matriz diagonal en la que todos los elementos de la misma son iguales a $\sigma_i^2 = \sigma_j^2 = \sigma^2$.

Paralelamente, también sabemos por el teorema de Gauss-Markov que el estimador MCO de los coeficientes del modelo lineal de regresión homocedástico es el de menor varianza de entre todos los lineales e insesgados, si bien es cierto que esto solo es correcto en el caso teórico de la homocedasticidad. Por el contrario, en el modelo de regresión lineal planteado en el recuadro, el estimador lineal e insesgado de menor varianza es diferente. Para verlo con claridad consideremos que, por los motivos que fuera, la varianza condicionada del error var$(\varepsilon_i\,|\mathbf{x}_i) = \sigma_i^2$ es conocida.

Para verlo con claridad consideremos que, por los motivos que fuera, la varianza condicionada del error var$(\varepsilon_i\,|\mathbf{x}_i) = \sigma_i^2$ es conocida.

El objetivo es cómo podemos utilizar esta información para transformar la expresión (5.1) de forma que podamos estimar los parámetros con errores homocedásticos.

Dividiendo la Ecuación (5.1) por su desviación típica σ_i conocida conseguimos que los errores sean homoscedásticos,

$$\frac{Y_i}{\sigma_i} = \frac{\beta_0}{\sigma_i} + \beta_1\frac{X_{1i}}{\sigma_i} + \beta_2\frac{X_{2i}}{\sigma_i} + ... + \beta_k\frac{X_{ki}}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}.$$

Haciendo un cambio de variable podemos escribir:

$$Y_i^* = \beta_0 X_{0i}^* + \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + ... + \beta_k X_{ki}^* + \varepsilon_i^*. \tag{5.14}$$

Es fácil observar que ahora la varianza de los errores es constante

$$\text{var}\left(\varepsilon_i^* \left|\mathbf{x}_i\right.\right) = \mathbb{E}\left[\left(\varepsilon_i^* \left|\mathbf{x}_i\right.\right)^2\right] = \mathbb{E}\left[\left(\frac{\varepsilon_i}{\sigma_i}\left|\mathbf{x}_i\right.\right)^2\right] = \mathbb{E}\left(\frac{\varepsilon_i^2}{\sigma_i^2}\left|\mathbf{x}_i\right.\right) = \frac{\mathbb{E}\left(\varepsilon_i^2 \left|\mathbf{x}_i\right.\right)}{\sigma_i^2} = \frac{\sigma_i^2}{\sigma_i^2} = 1.$$

Por tanto, la expresión (5.14) tendría errores homoscedásticos. A estas expresiones se las conoce con el nombre de estimador de **mínimos cuadrados ponderados** (**MCP**) puesto que todas las variables están ponderadas por $1/\sigma_i$. Si al ponderar adecuadamente logramos que el modelo sea homocedástico, entonces estaríamos bajo las condiciones de aplicabilidad del teorema de Gauss-Markov, y la expresión del estimador MCO ya no será la de menor varianza. Ahora el estimador lineal insesgado óptimo (de mínima varianza) sería una versión adecuadamente ponderada por la inversa de la desviación típica del error de cada observación. La expresión matricial es la siguiente

$$\hat{\boldsymbol{\beta}}_{MCP} = \left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*'}\mathbf{y}^* = \left(\mathbf{X}'\boldsymbol{\Sigma}_{\varepsilon\varepsilon'}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\varepsilon\varepsilon'}^{-1}\mathbf{y}. \tag{5.15}$$

Aunque pueda resultar evidente, conviene observar cómo se obtienen las matrices y vectores transformados

$$\mathbf{X}^* := \begin{pmatrix} 1/\sigma_1 & 0 & ... & 0 \\ 0 & 1/\sigma_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & 1/\sigma_n \end{pmatrix}\begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}; \mathbf{y}^* := \begin{pmatrix} 1/\sigma_1 & 0 & ... & 0 \\ 0 & 1/\sigma_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & 1/\sigma_n \end{pmatrix}\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

y de este modo comprobamos que el lado derecho de la expresión (5.15) se obtiene fácilmente tras observar qué forma tiene la inversa de la varianza heterocedástica

$$\boldsymbol{\Sigma}_{\varepsilon\varepsilon'}^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 0 & ... & 0 \\ 0 & 1/\sigma_2^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & 1/\sigma_n^2 \end{pmatrix} = \begin{pmatrix} 1/\sigma_1 & 0 & ... & 0 \\ 0 & 1/\sigma_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & 1/\sigma_n \end{pmatrix}^2$$

es decir hemos descompuesto $\boldsymbol{\Sigma}_{\varepsilon\varepsilon'}^{-1}$ en el producto de dos matrices, que podemos denotar por $\mathbf{V}$, es decir $\boldsymbol{\Sigma}_{\varepsilon\varepsilon'}^{-1} = \mathbf{V}'\mathbf{V}$.

La Ecuación (5.15), al ser ELIO, implica que el estimador $\hat{\boldsymbol{\beta}}_{MCO}$ sería ineficiente en un contexto tan general y habitual como es el heterocedástico. No obstante, para poder utilizar un estimador lineal e insesgado más eficiente sería preciso salirnos de los supuestos del modelo de regresión que hemos indicado al comienzo del tema, dado que para poder utilizarlo precisamos suponer que contamos con cierta información sobre la función de varianza condicionada $\text{var}(\varepsilon_i \left|\mathbf{x}_i\right.) = \sigma_i^2$. En este sentido se puede observar que la matriz de covarianzas condicionadas tiene n parámetros desconocidos, que aumentan con el tamaño muestral. Por tanto se necesita parametrizar esta matriz a fin de hacerla operativa. A continuación tratamos cómo estimaríamos si dispusiéramos de este tipo de información.

Mínimos cuadrados ponderados cuando conocemos la forma funcional de la heterocedasticidad

Como hemos de realizar al menos un supuesto adicional, consideremos que la varianza condicionada del error es conocida salvo por un factor de proporcionalidad; es decir

$$\text{var}(\varepsilon_i \,|X_i\,) = \lambda h(X_i),$$

donde h es una función que suponemos conocida y λ es una constante. El estimador MCP, como hemos visto anteriormente, se obtiene siempre dividiendo la variable dependiente e independiente por la raíz cuadrada de h y luego haciendo la regresión por MCO de la variable dependiente transformada y el regresor también transformado.

Como hemos visto anteriormente en este procedimiento, conocer parcialmente la forma de la varianza del error, nos permite transformar el término error heterocedástico en un término error transformado de modo que ahora ya es homocedástico. Por tanto, aplicar MCO a dicho modelo transformado nos conduce a estimadores ELIO, ya que se cumplen los supuestos del teorema de Gauss-Markov.

La cuestión obvia es que en la práctica desconocemos la función h, y por tanto la propuesta no es factible o realizable, al no poder llevarse a cabo. No obstante se suelen indicar algunos supuestos tentativos sobre el patrón de heterocedasticidad (especialmente útiles en el caso de regresión simple), a fin de hacer factible el método de los MCP. Veamos algunos casos.

Caso I. La varianza de los errores es proporcional, digamos, a una de las variables explicativas, por ejemplo a X_{1i}^2, es decir que

$$\text{var}(\varepsilon_i \,|X_i\,) = \sigma^2 X_{1i}^2.$$

Entonces estimamos la regresión (5.1) ponderada por X_{1i}

$$\frac{Y_i}{X_{1i}} = \beta_0 \frac{1}{X_{1i}} + \beta_1 + \beta_2 \frac{X_{2i}}{X_{1i}} + ... + \beta_k \frac{X_{ki}}{X_{1i}} + \frac{\varepsilon_i}{X_{1i}}.$$

Esta expresión nos conduce a (5.14), y por tanto a una situación homocedástica. Es decir,

$$\mathbb{E}\left[(\varepsilon_i^* \,|\mathbf{x}_i\,)^2\right] = \mathbb{E}\left[\left(\frac{\varepsilon_i}{X_i}\,|\mathbf{x}_i\right)^2\right] = \mathbb{E}\left(\frac{\varepsilon_i^2}{X_i^2}\,|\mathbf{x}_i\right) = \frac{\mathbb{E}\,(\varepsilon_i^2\,|\mathbf{x}_i\,)}{X_i^2} = \frac{\sigma^2 X_{1i}^2}{X_{1i}^2} = \sigma^2.$$

Podemos comprobar que este Caso I es fácilmente aplicable a situaciones similares como pueden ser que consideremos que la varianza condicionada del error sea proporcional a X_i o incluso a una combinación lineal de las variables explicativas, como es el caso de utilizar una varianza condicionada proporcional al valor medio de la variable $\hat{Y}_i$. En uno y otro caso tendríamos modelos transformados del tipo siguiente:

$$\frac{Y_i}{\sqrt{X_{1i}}} = \beta_0 \frac{1}{\sqrt{X_{1i}}} + \beta_1 \sqrt{X_{1i}} + \beta_2 \frac{X_{2i}}{\sqrt{X_{1i}}} + ... + \beta_k \frac{X_{ki}}{\sqrt{X_{1i}}} + \frac{\varepsilon_i}{\sqrt{X_{1i}}},$$

$$\frac{Y_i}{\sqrt{\hat{Y}_i}} = \beta_0 \frac{1}{\sqrt{\hat{Y}_i}} + \beta_1 \frac{X_{1i}}{\sqrt{\hat{Y}_i}} + \beta_2 \frac{X_{2i}}{\sqrt{\hat{Y}_i}} + ... + \beta_k \frac{X_{ki}}{\sqrt{\hat{Y}_i}} + \frac{\varepsilon_i}{\sqrt{\hat{Y}_i}},$$

en ambos casos los modelos ahora serían homocedásticos[1]. El considerar uno u otro caso dependerá críticamente del tipo de relación económica que se esté estudiando.

Otro tipo de relación que se puede dar dentro de este tipo de situaciones (Caso I) es que se diera la situación en que la muestra pudiera ser dividida en dos (o más grupos), teniendo cada grupo una varianza distinta. Podríamos estimar dichas varianzas a partir de los residuos de cada uno de los grupos, y transformar el modelo para hacerlo homocedástico, para lo cual seguir la mísma estrategia de dividir el modelo entre la desviación típica estimada en cada uno de los grupos (supongamos por simplificar que hubiera solo dos, $\hat{\sigma}_I, \hat{\sigma}_{II}$):

$$\frac{Y_{i(I)}}{\hat{\sigma}_I} = \beta_{0,(I)} \frac{1}{\hat{\sigma}_I} + \beta_1 \frac{X_{2i(I)}}{\hat{\sigma}_I} + ... + \beta_k \frac{X_{ki(I)}}{\hat{\sigma}_I} + \frac{\varepsilon_{i(I)}}{\hat{\sigma}_I}, i = 1, ..., N_I$$

$$\frac{Y_{i(II)}}{\hat{\sigma}_{II}} = \beta_{0(II)} \frac{1}{\hat{\sigma}_{II}} + \beta_1 \frac{X_{2i(II)}}{\hat{\sigma}_{II}} + ... + \beta_k \frac{X_{ki(II)}}{\hat{\sigma}_{II}} + \frac{\varepsilon_{i(II)}}{\hat{\sigma}_{II}}, i = 1, ..., N_{II},$$

$$N = N_I + N_{II}.$$

Obsérvese que el modelo prevé que los coeficientes que recogen los efectos parciales de las variables explicativas son los mismos para las dos submuestras, mientras que los interceptos (términos independientes) no lo son. Por otra parte, estos modelos ponderados son ahora homocedásticos, y por tanto su estimación de $\beta_j, j = 1, ..., k$ es más eficiente que MCO.

Con carácter general, ha de tenerse en cuenta que en ocasiones puede resultar útil reducir la heterocedasticidad considerando las variables en logaritmos. De hecho transformar a logaritmos comprime las escalas en las que las variables (dependientes e independientes) son medidas, por tanto se produce una reducción en la diferencia entre valores.

Caso II. En este caso, a diferencia del anterior, consideramos que es necesario estimar la varianza condicionada. En la mayoría de las situaciones la forma de la heterocedasticidad no es conocida de manera que es difícil encontrar la función de las variables independientes $h(\mathbf{X})$ que determina la forma de heterocedasticidad. Pero podemos estimarla $\hat{h}(\mathbf{X})$, y su utilización, en vez de la verdadera función $h(\mathbf{X})$, se suele denominar MCP-*factibles*.

Hay distintas formas de modelizar la heterocedasticidad, una posibilidad es

$$\text{var}\left(\varepsilon_i | \mathbf{X}\right) = \sigma^2 \exp\left(\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + ... + \alpha_k X_{ki}\right).$$

Es decir, incluimos el supuesto de que

$$h\left(\mathbf{X}\right) = \exp\left(\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + ... + \alpha_k X_{ki}\right). \tag{5.16}$$

En este caso utilizamos la función exponencial para garantizar que $\hat{h}(\mathbf{X})$ tenga valor positivo. No conocemos sin embargo los coeficientes de la ecuación anterior

[1] En la última ecuación, también podríamos incluir la eventualidad de que la varianza del error fuera proporcional al cuadrado del valor esperado de Y_i, en tal caso, habríamos de ponderar por $1/\hat{Y}_i$.

(si los conociéramos, entonces aplicaríamos MCP tal y como en el Caso I). Por tanto es preciso estimarlos. Bajo el supuesto de que se cumple (5.16) podemos escribir

$$\varepsilon_i^2 = \sigma^2 \exp\left(\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + ... + \alpha_k X_{ki}\right) u_i.$$

Suponiendo que u_i tiene media unitaria y que es independiente de las variables explicativas. podemos escribir

$$\ln\left(\varepsilon_i^2\right) = \delta_0 + \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + ... + \alpha_k X_{ki} + e_i,$$

donde e_i tiene media nula y es independiente de las variables explicativas. Estamos aún en una situación no implementable en la práctica dado que los errores de la expresión (5.1) los desconocemos. No obstante, sí conocemos los errores estimados, $\hat{\varepsilon}_i$ en la regresión inicial MCO, y estos los podemos utilizar para estimar consistentemente los parámeteros de la ecuación logarítimica, es decir que estimamos

$$\ln\left(\hat{\varepsilon}_i^2\right) = \delta_0 + \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + ... + \alpha_k X_{ki} + e_i.$$

Haciendo $g_i \equiv \ln\left(\hat{\varepsilon}_i^2\right)$, la estimación de $h(\mathbf{X})$ es

$$\hat{h}\left(\mathbf{X}\right) = \exp\left(\hat{g}_i\right).$$

Finalmente utilizamos $1/\hat{h}^{1/2}\left(\mathbf{X}\right)$ como ponderación en la expresión (5.1) como hicimos en los casos precedentes:

$$\frac{Y_i}{\sqrt{\hat{h}\left(\mathbf{X}\right)}} = \beta_0 \frac{1}{\sqrt{\hat{h}\left(\mathbf{X}\right)}} + \beta_1 \frac{X_{1i}}{\sqrt{\hat{h}\left(\mathbf{X}\right)}} + \beta_2 \frac{X_{2i}}{\sqrt{\hat{h}\left(\mathbf{X}\right)}} + ... + \beta_k \frac{X_{ki}}{\sqrt{\hat{h}\left(\mathbf{X}\right)}} + \frac{\varepsilon_i}{\sqrt{\hat{h}\left(\mathbf{X}\right)}}.$$

Naturalmente $h(\mathbf{X})$ puede tomar otras formas dependiendo del problema. Por ejemplo, podríamos considerar que

$$\text{var}(\varepsilon_i \,|X_i\,) = \theta_0 + \theta_1 X_{1i}^2, \theta_0 > 0, \theta_1 \geq 0$$

lo que nos llevaría a tener que estimar $\text{var}(\varepsilon_i \,|X_i\,)$. Esto lo haríamos a partir de los residuos MCO haciendo la regresión de $\hat{\varepsilon}_i^2$ sobre X_{1i}^2, sujetos a la restricción $\theta_0 > 0, \theta_1 \geq 0$. Esto nos llevaría la expresion factible de

$$\widehat{\text{var}}(\varepsilon_i \,|X_i\,) = \hat{\theta}_0 + \hat{\theta}_1 X_{1i}^2$$

que utilizaríamos para formar la nuevas variables explicativas transformadas (ponderadas)

$$\frac{Y_i}{\sqrt{\widehat{\text{var}}(\varepsilon_i \,|X_i\,)}} = \beta_0 \frac{1}{\sqrt{\widehat{\text{var}}(\varepsilon_i \,|X_i\,)}} + \beta_1 \frac{X_{1i}}{\sqrt{\widehat{\text{var}}(\varepsilon_i \,|X_i\,)}} + ... + \beta_k \frac{X_{ki}}{\sqrt{\widehat{\text{var}}(\varepsilon_i \,|X_i\,)}} + \frac{\varepsilon_i}{\sqrt{\widehat{\text{var}}(\varepsilon_i \,|X_i\,)}}.$$

En general, podríamos contemplar casos en los que podemos intentar modelizar, basados en el tipo de problema que estamos analizando, la $h(X)$ y por tanto la varianza condicionada $\text{var}(\varepsilon_i \,|\mathbf{X}_i\,)$ haciendo que esta no solo dependa de una de las variables explicativas, si no también de otras variables del modelo especificado o incluso de otras que no estén en la función de regresión poblacional. Siempre que tengamos datos correspondientes, podremos hacer una estimación factible con sus correpondientes restricciones a fin de caracterizar a una varianza, y posteriormente transformar el modelo a estimar.

Tanto en un caso como en otro, conviene recordar que hemos añadido supuestos que nos permitan transformar el modelo heterocedástico en un homocedástico, y luego procedemos con la estimación MCO, que en caso de haber modelizado adecuadamente la varianza condicionada de los errores, nos conduciría asintóticamente a estimadores más eficientes que los MCO.

Obsérvese también que a partir de la estimación por MCP, que nos permite estimar los valores de los coeficientes, podemos construir intervalos de confianza para los coeficientes estimados y utilizaremos para ello los errores estándar típicos de los casos teóricos de homocedasticidad.

Cuando NO conocemos la forma funcional de la heterocedasticidad

En realidad, como vimos anteriormente, hay otra alternativa para solucionar el efecto de la heterocedasticidad: utilizar los estimadores de los errores estándar robustos. Así pues se presentan dos opciones: o bien estimamos los coeficientes $\boldsymbol{\beta}$ por mínimos cuadrados ponderados $\hat{\boldsymbol{\beta}}_{MCP}$, o bien los estimamos por MCO $\hat{\boldsymbol{\beta}}_{MCO}$, y luego utilizamos los errores estándar robustos a la heterocedasticidad. Para decidir qué usar en la práctica veamos las ventajas y desventajas de uno y de otro método.

Una ventaja de MCP es su mayor eficiente respecto del estimador MCO de los coeficientes del modelo de regresión original, al menos asintóticamente. La desventaja es que necesariamente requiere conocer la función de la varianza condicionada y estimar, adecuadamente, sus correspondientes parámetros. En la práctica habitual no se conoce dicha función, y tenemos que postular cuál es. De hecho si hemos especificado incorrectamente la forma funcional de la varianza condicionada, entonces los errores estándar calculados por MCP no son válidos y nos conducirían a conclusiones erróneas.

La ventaja de usar errores estándar robustos a la heterocedasticidad es que asintóticamente proporcionan valores válidos para llevar a cabo inferencias incluso si se desconoce la forma de la función de varianza condicionada. Afortunadamente en la actualidad el software econométrico incorpora esta opción lo que facilita un uso a bajo coste para el usuario.

Considerando pros y contras, junto con el hecho de que en la práctica raramente conocemos la expresión de la varianza condicionada del error, parece oportuno y más sencillo utilizar errores estándar robustos sin necesidad de hacer elucubraciones sobre la varianza condicionada.

Bibliografía complementaria

Matilla-García, M et al. 2017. Econometría y Predicción. McGraw Hill

Stock J. and Watson J. Introducción a la econometría. Pearson.

Tema 6

Otras técnicas de estimación

Este tema está elaborado como una adaptación de

Wooldridge. J. Econometric Analysis of Cross Section and Panel Data. Capitulo 8 apartados 8.1, 8.2, 8.3 y 8.5; Capítulo 12 apartado 12.10; Capítulo 21 apartados 21.1-21.3. Así como de la bibliografía complementaria.

- El estimador de momentos.
- Estimación por el método generalizado de los momentos (GMM),.
- Regresión cuantílica.
- Estimador diferencias en diferencias.

6.1 El estimador de momentos

Revisamos inicial y someramente lo que ya hemos visto para así contextualizar alguna técnica de estimación alternativa a las ya expuestas.

Los economistas suelen utilizar la regresión lineal para cuantificar una relación entre variables económicas. Una regresión lineal entre Y y X es una relación de la forma

$$Y = \mathbf{x}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

donde β y ε se eligen de tal manera que ε no está correlacionado con **x** (que normalmente incluye un término constante). Por lo tanto, el vector de parámetros β satisface la condición

$$\mathbb{E}\left[\mathbf{x}\left(Y - \mathbf{x}'\boldsymbol{\beta}\right)\right] = \mathbf{0} \tag{6.1}$$

Si $\mathbb{E}(\mathrm{xx}\prime)$ tiene rango completo, entonces el sistema de esperanzas tiene una solución única

$$\boldsymbol{\beta} = \left[\mathbb{E}\left(\mathbf{x}\mathbf{x}'\right)\right]^{-1}\mathbb{E}\left(\mathbf{x}Y\right) \tag{6.2}$$

Una propiedad importante de la regresión lineal es que es un predictor óptimo de Y dado **x** en el siguiente sentido

$$\boldsymbol{\beta} = \underset{\mathbf{b}}{argmin}\mathbb{E}\left[\left(Y - \mathbf{x}'\boldsymbol{b}\right)^2\right]$$

Es decir, minimiza el cuadrado del error de predicción lineal esperado. Es por eso que $\mathbf{x}'\beta$ se denomina "mejor predictor lineal" o "proyección lineal".

Además, si la expectativa condicional de Y dado **x** es lineal, resulta que coincide con el predictor lineal. Si altenativamente, $\mathbb{E}(Y|\mathbf{x})$ no es lineal, entonces la proyección lineal es una aproximación óptima a ella en el sentido de que

$$\boldsymbol{\beta} = \underset{\mathbf{b}}{argmin}\mathbb{E}\left\{\left[\mathbb{E}(Y|\mathbf{x}) - \mathbf{x}'\boldsymbol{b}\right]^2\right\}$$

Es por eso que a veces la notación $\mathbb{E}(Y|\mathbf{x}) = \mathbf{x}\prime\boldsymbol{\beta}$ es utilizada ya que enfatiza la proximidad de los conceptos de proyección lineal y expectativa condicional. Por lo tanto, β es un vector con información útil si estamos interesados en una predicción lineal de Y dada **x**, o si estamos interesados en estudiar cómo cambia la media de Y para diferentes valores de **x**, y pensamos que $\mathbb{E}(Y|\mathbf{x})$ es lineal o aproximadamente lineal. La regresión lineal también puede ser de interés como relación estructural o causal entre Y y **x** si tenemos razones a priori para creer que los determinantes no observables de y no están correlacionados con **x**.

Si estamos interesados en una relación estructural entre Y, **x**, y unas variables no observables **u**

$$Y = \mathbf{x}'\boldsymbol{\delta} + \boldsymbol{u}$$

tal que **u** está correlacionada con al menos algunos de los componentes de **x**, entonces claramente $\delta \neq \beta$, en general. En muchas situaciones de interés en econometría, δ puede considerarse como la solución de ecuaciones de momento de la forma

$$\mathbb{E}\left[\mathbf{z}\left(Y - \mathbf{x}'\boldsymbol{\delta}\right)\right] = \mathbf{0} \tag{6.3}$$

donde **z** es un vector de variables que a priori se puede suponer que no está correlacionado con **u** a estas variables las llamaremos variables instrumentales más adelante en este temario[1]. Si $\mathbb{E}(\mathbf{z}\mathbf{x}\prime)$ tiene rango completo, entonces el sistema de esperanzas (6.3) tiene una solución única. Además, si **z** y **x** son de la misma dimensión, se tiene

$$\boldsymbol{\delta} = \left[\mathbb{E}\left(\mathbf{z}\mathbf{x}'\right)\right]^{-1}\mathbb{E}\left(\mathbf{z}Y\right) \tag{6.4}$$

Pues bien, obsérvese que las ecuaciones (6.2) y (6.4) pueden describirse como "problemas de momentos" porque los parámetros de interés resuelven ecuaciones sobre los momentos.

De acuerdo con el **principio de analogía**, dada una muestra representativa $\{Y_i, \mathbf{x}_i, \mathbf{z}_i\}, i = 1, ..., N$, elegimos como estimador candidato para una característica poblacional, la misma característica definida en la muestra. De esta forma, los coeficientes de regresión lineal muestrales resuelven

$$\frac{1}{N}\sum_{i=1}^{N}\left[\mathbf{x}_i\left(Y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)\right] = \mathbf{0}$$

[1]En este punto puede considerarse oportuno trabajar el Tema 8 de este bloque para facilitar la lectura, si bien puede continuarse la lectura sin ninguna dificultad si el lector está familiarizado con el concepto y técnicas basadas en variables instrumentales. En caso contrario, es posible que sea conveniente leerse el Tema 8.

arrojando el estimador MCO

$$\widehat{\boldsymbol{\beta}}_{MM} = \left[\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'\right]^{-1} \sum_{i=1}^{N} \mathbf{x}_i Y_i.$$

Del mismo modo para la estimación con variables instrumentales, se tiene la contrapartida muestral de las restricción poblacional, es decir,

$$\frac{1}{N} \sum_{i=1}^{N} \left[\mathbf{z}_i \left(Y_i - \mathbf{x}_i' \boldsymbol{\delta}\right)\right] = \mathbf{0}$$

con solución

$$\widehat{\boldsymbol{\delta}}_{MM} = \left[\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i'\right]^{-1} \sum_{i=1}^{N} \mathbf{z}_i Y_i.$$

Ambos estimadores están basados en las restricciones sobre los momentos, por ello los indicamos con el subíndice MM (método de los momentos). Obsérvese igualmente que en estas situaciones los estimadores MCO y los estimadores MM coinciden es este caso.

Esta es la idea esencial de los denominados estimadores por el método de los momentos que ahora formalizamos considerando, tal y como hicimos en el Tema 1, que es posible tener más de una variable endógena.

El punto de partida es asumir la existencia de r condiciones sobre los momentos relativos q parámetros,

$$\mathbb{E}[\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}_0)] = \mathbf{0} \tag{6.5}$$

donde ϑ es un vector $q \times 1$, $\mathbf{h}(\cdot)$ es una función vectorial $r \times 1$ con $r \geq q$, y $\boldsymbol{\theta}_0$ denota el valor de ϑ en un supuesto proceso generador de datos. El vector $\mathbf{w}$ incluye todos los observables incluyendo, cuando sea relevante, una variable dependiente $\mathbf{y}$, regresores potencialmente endógenos $\mathbf{x}$ y variables instrumentales $\mathbf{z}$. La variable dependiente $\mathbf{y}$ puede ser un vector, por lo que se incluyen las aplicaciones con sistemas de ecuaciones o con datos de panel. La esperanza es con respecto a todos los componentes estocásticos de $\mathbf{w}$ y, por tanto respecto, $\mathbf{y}$, $\mathbf{x}$ y $\mathbf{z}$. La elección de la forma funcional para $\mathbf{h}(\cdot)$ es cualitativamente similar a la elección del modelo y variará con la aplicación. Dos ejemplos son (6.3) y (6.1), donde $\mathbf{h}(\cdot)$ es respectivamente la función sobre la que se toman esperanzas.

Si $r = q$, entonces se puede aplicar el método de momentos. La igualdad a cero del momento poblacional se reemplaza por la igualdad a cero del momento muestral correspondiente, y el estimador del método de momentos $\widehat{\boldsymbol{\theta}}_{MM}$ se define como aquel que soluciona

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{h}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{6.6}$$

Este estimador equivalentemente minimiza la forma

$$Q_N(\boldsymbol{\theta}) = \left[\frac{1}{N} \sum_{i=1}^{N} \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})\right]' \left[\frac{1}{N} \sum_{i=1}^{N} \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})\right]$$

6.2 Estimación por el método generalizado de los momentos (GMM)

El estimador GMM se basa en r condiciones independientes, (6.5), sobre los momentos mientras que se estiman q parámetros. Si r = q, se dice que el modelo está exactamente identificado y se puede utilizar el estimador MM anteriormente descrito. Si $r > q$ se dice que el modelo está sobreidentificado y (6.6) no tiene solución para $\widehat{\boldsymbol{\theta}}$ ya que hay mas ecuaciónes r que incógnitas q. La idea de los estimadores GMM, $\widehat{\boldsymbol{\theta}}_{GMM}$, es seleccionar $\widehat{\boldsymbol{\theta}}$ de modo que la forma $N^{-1}\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i,\widehat{\boldsymbol{\theta}})$ sea lo más cercana posible a cero. Es decir, se trata de minimizar la función objetivo

$$Q_N(\boldsymbol{\theta}) = \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})\right]' \mathbf{W}_N \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})\right]$$

donde la matriz $\mathbf{W}_N$, es una matriz de pesos cuadrada simétrica $r \times r$, definida positiva, posiblemente estocástica con límite en probabilidad finito. Esta matriz de pesos no depende de $\boldsymbol{\theta}$, pero problemente dependa del tamaño muestral N. Diferentes matrices de pesos conducen a distintos estimadores. Si por ejemplo $\mathbf{W}_N = \mathrm{I}$, entonces $Q_N(\boldsymbol{\theta}) = \bar{h_1}^2 + \bar{h_2}^2 + .. + \bar{h_r}^2$, $\bar{h_j} = N^{-1}\sum_{i=1}^{N} h_j(\mathbf{w}_i,\boldsymbol{\theta})$, siendo h_j el componente $j-$ésimo de $\mathbf{h}$. Se trata localizar matrices de pesos óptimas, y hay técnicas para ello.

Para alcanzar el mínimo de la forma cuadrática diferenciamos respecto de $\boldsymbol{\theta}$ y obtenemos las condiciones de primer orden de GMM

$$\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})'}{\partial\boldsymbol{\theta}}\bigg|_{\widehat{\boldsymbol{\theta}}}\right]' \mathbf{W}_N \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})\right] = \mathbf{0}, \tag{6.7}$$

que en general serán ecuaciones no lineales que deben ser resueltas numéricamente.

En el caso particular de las restricciones (6.3), la forma cuadrática objetivo será, utilizando la notación matricial $\mathbf{y} = \mathbf{X}\boldsymbol{\delta} + \mathbf{u}$ y considerando la matriz de instrumentos $\mathbf{Z}$ de dimensión $N \times r$:

$$Q_N(\boldsymbol{\delta}) = \left[\frac{1}{N}(\mathbf{y}-\mathbf{X}\boldsymbol{\delta})'\mathbf{Z}\right]' \mathbf{W}_N \left[\frac{1}{N}\mathbf{Z}'(\mathbf{y}-\mathbf{X}\boldsymbol{\delta})\right].$$

Las condiociones de primer orden

$$\frac{\partial Q_N(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}} = -2\left[\frac{1}{N}\mathbf{X}'\mathbf{Z}\right]' \mathbf{W}_N \left[\frac{1}{N}\mathbf{Z}'(\mathbf{y}-\mathbf{X}\boldsymbol{\delta})\right] = \mathbf{0}$$

cuya resolución conduce al estimador GMM de variables instrumentales:

$$\hat{\boldsymbol{\delta}}_{\mathbf{W}}^{GMM} = (\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{y} \tag{6.8}$$

donde deberemos seleccionar una matriz $\mathbf{W}_N$ adecuada.

Es interesante observar que el estimador (6.8) utiliza combinaciones ponderadas de las variables en incluidas en $\mathbf{Z}$ (exógeneas e instrumentos). Por este motivo el estimador que

en su momento veremos de mínimos cuadrados en dos etapas (MC2E) de la expresión será un caso particular de (6.8). En particular será una combinación ponderada de los instrumentos, en el que el problema de minimización se resuelve para la matriz de ponderaciones particular $\mathbf{W}_N = (N^{-1}\mathbf{Z}'\mathbf{Z})^{-1}$. Igualmente, otros métodos de estimación son reconciliables con GMM siempre que determinemos una $\mathbf{Z}$ y una $\mathbf{W}_N$ adecuadas. Por ejemplo, si seleccionamos la matriz de ponderaciones $\mathbf{W}_N = (N^{-1}\mathbf{X}'\mathbf{X})^{-1}$ y consideramos que los instrumentos son todos variables exógenas, es decir, si consideramos que no hay problemas de endogeneidad, $\mathbf{Z} = \mathbf{X}$, entonces (6.8) coincide[2] con el estimador MCO.

La distribución asintótica del estimador $\hat{\boldsymbol{\beta}}_{\mathbf{W}}^{GMM}$ (Ecuación (6.8)) se deriva igual que haremos (en el tema dedicado a las variables instrumentales). Lo mismo sucederá con su varianza. El resultado general es

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{\mathbf{W}}^{GMM} - \boldsymbol{\beta}\right) \xrightarrow{d} Normal\left(\mathbf{0}, \mathbf{V}_{\mathbf{W}}^{GMM}\right),$$

$$\mathbf{V}_{\mathbf{W}}^{GMM} = \left(\mathbf{Q_{XZ}W Q_{ZX}}\right)^{-1}\mathbf{Q_{XZ}W\Omega W Q_{ZX}}\left(\mathbf{Q_{XZ}W Q_{ZX}}\right)^{-1}. \tag{6.9}$$

Es de interés saber si hay matrices de ponderación asintóticamente más eficientes que otras. La eficiencia dependerá de la varianza, es decir de (6.9). De nuevo las propiedades de los errores del modelo jugarán, como en el caso MCO, un papel determinante.

Vamos a considerar el caso en el que los **errores son homocedásticos**. Recordemos que en MCO, bajo este supuesto, el teorema de Gauss-Markov ofrece un resultado en términos de eficiencia de los estimadores. En el caso de trabajar con variables instrumentales (VI) hay un resultado análogo que indica que la estimación MC2E es asintóticamente eficiente en la clase de estimadores VI en los que los instrumentos son combinaciones lineales de las filas de $\mathbf{Z}$.

Con homocedasticidad, $\mathbb{E}\left(u_i^2 \,|\mathbf{Z}_i\right) = \sigma_u^2$, se tiene que

$$\mathbf{\Omega} = \mathbb{E}\left(\mathbf{Z}_i\mathbf{Z}_i'u_i^2\right) = \mathbb{E}\left[\mathbb{E}\left(\mathbf{Z}_i\mathbf{Z}_i'u_i^2 \,|\mathbf{Z}_i\right)\right] = \mathbb{E}\left[\mathbf{Z}_i\mathbf{Z}_i'\mathbb{E}\left(u_i^2 \,|\mathbf{Z}_i\right)\right] = \sigma_u^2\mathbf{Q_{ZZ}}.$$

Esta expresión hace que ahora (6.9) se convierta en

$$\mathbf{V}_{homo}^{GMM} = \sigma_u^2\left(\mathbf{Q_{XZ}W Q_{ZX}}\right)^{-1}\mathbf{Q_{XZ}W Q_{ZZ}W Q_{ZX}}\left(\mathbf{Q_{XZ}W Q_{ZX}}\right)^{-1}. \tag{6.10}$$

Igualmente, con homocedasticidad, la configuración de la matriz $\mathbf{\Omega}$ generará, tras una simplificación sencilla, la siguiente expresión de la varianza del estimador MC2E

$$\mathbf{V}_{homo}^{MC2E} = \sigma_u^2\left(\mathbf{Q_{XZ}Q_{ZZ}}^{-1}\mathbf{Q_{ZX}}\right)^{-1}. \tag{6.11}$$

Demostrar que MC2E es asintóticamente eficiente entre la clase de estimadores que son combinaciones lineales de $\mathbf{Z}$ consiste en probar que

$$\mathbf{c}'\mathbf{V}_{homo}^{GMM}\mathbf{c} \geq \mathbf{c}'\mathbf{V}_{homo}^{MC2E}\mathbf{c} \tag{6.12}$$

[2] En tal caso, observe el lector que $\boldsymbol{\beta} = \boldsymbol{\delta}$.

para todas las matrices $\mathbf{W}$ semidefinidas positivas y todos los vectores $\mathbf{c}$ de orden $(k+r+1)\times 1$.

Por tanto, en el caso homocedástico, la eficiencia del estimador VI se encuentra haciendo que la matriz de ponderaciones $\mathbf{W}=(N^{-1}\mathbf{Z}'\mathbf{Z})^{-1}$, que es la que, como hemos visto, da lugar a la estimación MC2E. Podemos además observar la cercanía entre la expresión eficiente bajo homocedasticidad de $\mathbf{W}$ y $\mathbf{\Omega}^{-1}=(1/\sigma_u^2)\mathbf{Q}_{\mathbf{ZZ}}^{-\mathbf{1}}$.

En el caso de errores heterocedásticos, el estimador MC2E no es eficiente entre la clase de estimadores VI que utilizan combinaciones lineales de $\mathbf{Z}$ como instrumentos. En este caso el estimador eficiente se encuentra a partir del estimador GMM, expresión (6.8). Por analogía al caso homocedástico, donde la expresión de la varianza que nos conduce a un estimador eficiente es aquella correspondiente a una selección de la matriz de ponderaciones que lleva a (6.11), en el caso heterocedástico la matriz de ponderaciones que nos conduce a una expresión similar (6.11), y por analogía eficiente, es cuando $\mathbf{W}=\mathbf{\Omega}^{-1}$. En este caso la expresión (6.9) se reduce, tras simplificar,

$$\mathbf{V}^{GMM}=\left(\mathbf{Q}_{\mathbf{XZ}}\mathbf{\Omega}^{-1}\mathbf{Q}_{\mathbf{ZX}}\right)^{-1}.$$

Se puede demostrar también que

$$\mathbf{c}'\mathbf{V}_{homo}^{GMM}\mathbf{c}\geq\mathbf{c}'\mathbf{V}^{GMM}\mathbf{c}.$$

Por lo que el estimador eficiente bajo heterocedasticidad se alcanza cuando $\mathbf{W}=\mathbf{\Omega}^{-1}$, y si sustituimos esto en 6.8, obtemos

$$\tilde{\boldsymbol{\beta}}^{GMM}=\left(\mathbf{X}'\mathbf{Z}\mathbf{\Omega}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\mathbf{\Omega}\mathbf{Z}'\mathbf{y}.$$

Para lograr que este estimador sea factible necesitamos que la matriz $\mathbf{\Omega}$ sea estimada consistentemente. Este estimador se calcula en dos etapas. La primera consiste en estimar consistentemente el vector de coeficientes $\boldsymbol{\beta}$ de la Ecuación habitual

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon},\tag{6.13}$$

donde ahora $\mathbf{X}$ es una matriz de orden $n\times(k+r+1)$ que contiene a los regresores exógenos incluidos y a los regresores endógenos, de modo que la fila i-ésima es

$$\mathbf{X}_i=\left(1,X_{1i},X_{2i},...,X_{ri},Y_{1i},...,Y_{ki}\right);$$

el vector de errores, $\boldsymbol{\varepsilon}$, es de orden $n\times 1$; y finalmente el vector de orden $n\times 1$, $\mathbf{Y}$, está formado por la variable dependiente, $Y_{0i}, i=1,...,n$.. Esto nos permite obtener los residuos de la ecuación de interés, y por tanto, podemos formar $\hat{\mathbf{\Omega}}=(1/N)\sum_{i=1}^{N}\mathbf{Z}_i\mathbf{Z}_i'\hat{u}_i^2$. En la segunda etapa se calcula la matriz de ponderaciones óptima $\hat{\mathbf{\Omega}}^{-1}$ y se calcula el estimador GMM eficiente:

$$\hat{\boldsymbol{\beta}}^{GMM}=\left(\mathbf{X}'\mathbf{Z}\hat{\mathbf{\Omega}}^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\hat{\mathbf{\Omega}}^{-1}\mathbf{Z}'\mathbf{y}.$$

6.3 Regresión cuantílica

En un la estadística descriptiva, los estadísticos de resumen para la distribución de la muestra incluyen cuantiles, como la mediana, los cuartiles inferior y superior y los percentiles, además de la media de la muestra.

En el contexto de la regresión, también podríamos estar interesados en los cuantiles condicionales. Por ejemplo, el interés puede radicar en cómo los percentiles de la distribución de ingresos para los trabajadores con bajo nivel educativo están mucho más comprimidos que los de los trabajadores con alto nivel educativo. En este ejemplo simple, uno puede simplemente hacer cálculos separados para trabajadores con bajo nivel educativo y para trabajadores con alto nivel educativo. Sin embargo, este enfoque se vuelve inviable si hay varios regresores que toman varios valores. En cambio, se necesitan métodos de regresión por cuantiles (regresión cuantílica) para estimar los cuantiles de la distribución condicional de y dado **x**.

La regresión cuantílica corresponde al uso de una función de perdida absoluta y asimétrica, mientras que el caso especial de regresión mediana usa función de pérdida absoluta del error, digamos u:

La función de pérdida, $L(u)$, en valor absoluto y asimétrica es

$$L(u) = \begin{cases} (1-\alpha)|u| & \text{si } u < 0 \\ \alpha|u| & \text{si } u \geq 0 \end{cases} \tag{6.14}$$

mientras la función de pérdida para la medina sería

$$L(u) = |u| \tag{6.15}$$

Los métodos basados en estas funciones de pérdida (penalización) proporcionan una alternativa a MCO, que utiliza como función de pérdida el error al cuadrado, $L(u) = u^2$.

Los métodos de regresión cuantílica tienen ventajas más allá de proporcionar una caracterización más rica de los datos. La regresión mediana es más robusta a los valores atípicos que la regresión por mínimos cuadrados. Además, los estimadores de regresión cuantílica pueden ser consistentes bajo supuestos estocásticos más débiles de lo que es posible con la estimación por mínimos cuadrados.

En la exposición primero nos centraremos sobre los cuatiles poblacionales para posteriormente pasar a la estimación de los cuantiles muestrales.

Para una variable aleatoria continua y, el cuantil poblacional q es ese valor μ_q tal que y es menor o igual a μ_q con probabilidad q. Por tanto,

$$q = Pr[y \leq \mu_q] = F_y(\mu_q)$$

donde F_y es la función de distribución acumulativa (CDF) de y. Por ejemplo, si $\mu_{0,75}$ = 3, la probabilidad de que $y \leq 3$ es igual a 0,75. De ello se deduce que

$$\mu_q = F_y^{-1}(q).$$

Los ejemplos principales son la mediana, $q = 0,5$, el cuartil superior, $q = 0,75$ y el cuartil inferior, $q = 0,25$. Para la distribución normal estándar $\mu_{0,5} = 0{,}0$, $\mu_{0,95} = 1,645$ y $\mu_{0,975} = 1,960$.

Para el modelo de regresión, el cuantil poblacional q de y condicional a **x** es la función $\mu_q(\mathbf{x})$ tal que y condicional (condionada) a **x** es menor o igual a $\mu_q(\mathbf{x})$ con probabilidad **q**, donde la probabilidad se evalúa usando la distribución condicional (condicionada de y dado **x**. De ello se deduce que

$$\mu_q(\mathbf{x}) = F^{-1}_{y|\mathbf{x}}(q)$$

donde $F_{y|\mathbf{x}}$ es la CDF condicional de y dada[3] **x**.

Resulta ilustrativo para familiarizarse con la regresión cuantílica derivar la función cuantílica $\mu_q(x)$ si se supone que el proceso generador de datos es el modelo lineal con heterocedasticidad multiplicativa:

$$y = \mathbf{x}'\boldsymbol{\beta} + u$$

$$u = \mathbf{x}'\boldsymbol{\alpha} \times \varepsilon, \mathbf{x}'\boldsymbol{\alpha} > 0$$

$$\varepsilon \sim iid(0, \sigma^2).$$

El cuantil poblacional q de y condicional a **x** es aquella función $\mu_q(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ tal que

$$\begin{aligned} q &= Pr[y \leq \mu_q(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})] \\ &= Pr[\mathbf{x}'\boldsymbol{\beta} + u \leq \mu_q(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})] \\ &= Pr[u \leq \mu_q(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) - \mathbf{x}'\boldsymbol{\beta}] \\ &= Pr[\varepsilon \times \mathbf{x}'\boldsymbol{\alpha} \leq (\mu_q(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) - \mathbf{x}'\boldsymbol{\beta})] \\ &= Pr[\varepsilon \leq (\mu_q(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) - \mathbf{x}'\boldsymbol{\beta}) / \mathbf{x}'\boldsymbol{\alpha}] \\ &= F_\varepsilon [(\mu_q(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) - \mathbf{x}'\boldsymbol{\beta}) / \mathbf{x}'\boldsymbol{\alpha}] \end{aligned}$$

siendo F_ε la CDF de ε. De esto se sigue por tanto que

$$F^{-1}_\varepsilon(q) = (\mu_q(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) - \mathbf{x}'\boldsymbol{\beta}) / \mathbf{x}'\boldsymbol{\alpha}$$

de modo que

$$\begin{aligned} \mu_q(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \mathbf{x}'\boldsymbol{\beta} + \mathbf{x}'\boldsymbol{\alpha} \times F^{-1}_\varepsilon(q) \\ &= \mathbf{x}' \left(\boldsymbol{\beta} + \boldsymbol{\alpha} \times F^{-1}_\varepsilon(q)\right) \end{aligned}$$

por lo que los cuantiles condicionales son lineales en **x**. En el caso particular de homocedasticidad, al ser $\mathbf{x}'\boldsymbol{\alpha}$ una constante en tal caso, se tendría que todos los cuantiles condicionales tendrían la misma pendiente y diferirían únicamente en el término constante (intercepto), que crecería a medida que lo hiciera q.

Si pasamos ahora de los cuantiles poblacionales a los muestrales, lo primero que hemos de recordar es cómo se obtienen los cuantiles a partir de una realización de tamaño

[3] Obsérvese que hemos suprimido el papel de los parámetros de esta distribución

muestral N de una variable aleatoria. El proceso consiste en ordenar los datos observados de menor a mayor, y entonces μ_q será el valor $N \times q$-ésimo valor más pequeño (redondeándose al alza el valor $N \times q$) de la muestra. Este mecanismo de obtención del $q-$ésimo cuantil es equivalente a resolver el siguiente problema de minimización con respecto a β

$$\sum_{i:y_i \geq \beta}^{N} q|y_i - \beta| + \sum_{i:y_i < \beta}^{N} (1-q)|y_i - \beta|$$

Esta función objetivo es fácilmente extensibel al caso de la regresión lineal, de modo que el estimador de la regresión cuantílica q-ésima, $\widehat{\boldsymbol{\beta}}_q$, minimiza $\boldsymbol{\beta}_q$ en la expresión

$$Q_N(\boldsymbol{\beta}_q) = \sum_{i:y_i \geq \mathbf{x}_i'\boldsymbol{\beta}}^{N} q|y_i - \mathbf{x}_i'\boldsymbol{\beta}_q| + \sum_{i:y_i < \mathbf{x}_i'\boldsymbol{\beta}}^{N} (1-q)|y_i - \mathbf{x}_i'\boldsymbol{\beta}_q| \tag{6.16}$$

donde se observa que coincide con un problema de minimización de una función de pérdida como la indicada en (6.14) para el caso particular de $u = y - \mathbf{x}'\boldsymbol{\beta}_q$. Obsérvese también que para $q = 0,5$ se obtiene el estimador de la regresión de la mediana, y en ese caso la función de pérdida es (6.15).

De cara a la estimación, la expresión (6.16) no es diferenciable y por tanto hay que utilizar métodos númericos para su resolución; algo que en la actualidad no supone mayores dificultades para el software habitual econométrico. Sin embargo esta ausencia de diferenciabilidad si ha sido un reto para obtener la distribución asintótica. En este sentido se puede demostrar que

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q\right) \underset{d}{\rightarrow} Normal\left[0, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right]$$

donde

$$\mathbf{A} = plim\frac{1}{N}\sum_{i=1}^{N} f_{u_q}(0|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'$$

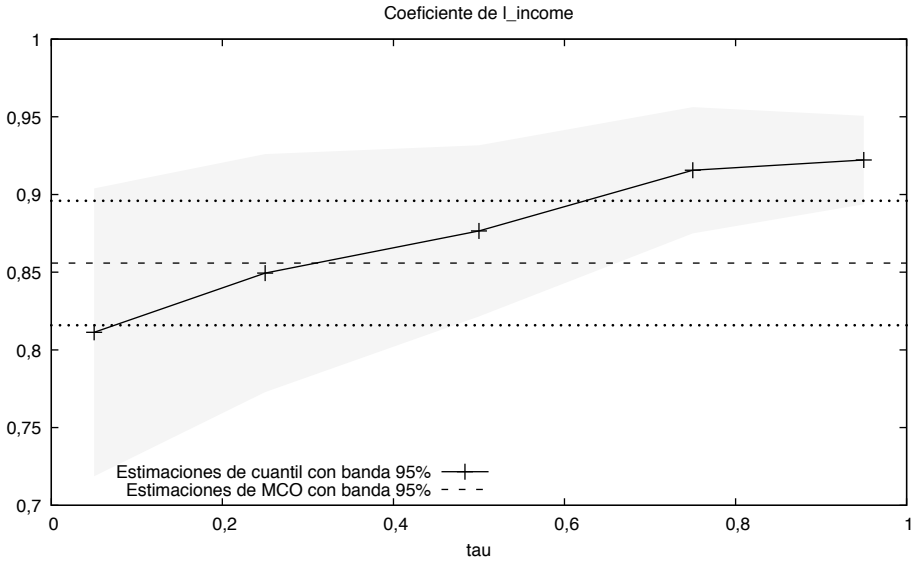
$$\mathbf{B} = plim\frac{1}{N}\sum_{i=1}^{N} q(1-q)\mathbf{x}_i\mathbf{x}_i',$$

y $f_{u_q}(0|\mathbf{x})$ es la densidad condicional de término error $u_q = y - \mathbf{x}'\boldsymbol{\beta}_q$ evaluada en $u_q = 0$.

Para ilustrar este tipo de estimación, vamos a utilizar datos de gasto anual en comida e ingresos anuales de hogares de clase trabajadora. El conjunto de datos está formado por 235 hogares. Los datos de ingresos y gastos están en escala logarítmica y por tanto la estimación de la pendiente se interpreta como la elasticidad del gasto en comida respecto de los ingresos.

En la regresión cuantílica hemos considerado cinco cuantiles $q = \{0,05; 0,25; 0,50; 0,75; 0,95\}$. La siguiente figura muestra los coeficientes de pendiente para diferentes valores de q, junto con un intervalo de confianza al 95 %. La elasticidad estimada aumenta con el nivel de ingreso del hogar. La estimación clásica MCO es de 0,8505, que se muestra

Figura 6.1: Regresión cuantílica para los cuantiles (tau=q)



con la línea constante discontinua, mientras que la regresión por cuantiles varía de $0,81 - 0,92$.

Desde otro punto de vista, la regresión cuantílica puede ser informativa respecto a una incorrecta especificación del modelo para la media condicional. En efecto una variedad de pendientes y puntos de corte (interceptos) que significativamente varían con q se interpreta como evidencia de heterocedasticidad. Otra interpretación es que la media condicional es no lineal en **x** con pendiente creciente y esto conduce a coeficientes de pendiente de cuantiles que aumentan con el cuantil q.

6.4 Estimador diferencias en diferencias

En economía abundan, como dijimos en el Tema 1 del temario, los datos observacionales, esto es, datos que generalmente son de naturaleza no experimental. Lo interesante, y que en buena medida justifica el tratamiento ofrecido en la sección anterior, es que los métodos e ideas de los experimentos aleatorizados controlados pueden, en ciertas circunstancias, trasladarse y en su caso aplicarse a datos no experimentales. Podríamos por tanto a partir de esos datos analizar los resultados para observaciones de un grupo de tratamiento y otro de control en los que el tratamiento no hubiera sido asignado aleatoriamente.

En estos casos ya no estamos en el marco de los experimentos aleatorizados, y por tanto la literatura se refiere a ellos como ***cuasiexperimentos*** o ***experimentos naturales***. El primer término, heredado de la psicología, enfatiza el hecho más sustantivo de que no se trata de experimentos. El segundo término incide en el hecho singular de que para poder realizar un estudio de este tipo es necesario que existan variaciones en circunstancias individuales (externas) que hagan que parezca «como si» la asignación del tratamiento hubiera sido aleatoria. Estas variaciones en las circunstancias individuales pueden

surgir como consecuencia de factores no relacionados con el efecto causal de estudio (por tanto exógenos). Estos factores en ocasiones provienen de fuentes de aleatoriedad natural como son las fechas de nacimiento, la lluvia o, en general, cuestiones genéticas. También se pueden encontrar en factores institucionales como la ubicación, el calendario de aplicación de un programa o acción, la entrada en vigor de una norma, etcétera. Un buen cuasiexperimento es aquel en el que hay una transparente fuente de variación exógena en las variables explicativas que determine la asignación del tratamiento.

Lo que caracteriza a un cuasiexperimento es que el tratamiento no está asignado al azar, es «como si» estuviera asignado al azar cuando condicionamos algunas variables observadas, W. Debido a que el investigador no tiene control sobre la aleatoridad en la asignación del tratamiento, es probable que la correcta comparabilidad entre grupos no esté garantizada, incluso después de haber controlado W. Existe aún la posibilidad de que haya variables omitidas, por ejemplo, que sean permanentes en los dos grupos y explicaran también los distintos resultados potenciales. Esto supondría que las conclusiones obtenidas con los estimadores de las diferencias que hemos visto anteriormente no serían veraces.

Una forma atractiva de tratar esta situación es analizando la variación experimentada antes y después del tratamiento por la variable resultado Y en cada uno de los dos grupos, tratados y no-tratados (controles). Esto supone que consideramos que hay un «antes» y hay un «después». Por ejemplo, en el caso del estudio sobre el efecto de la inmigración sobre el salario de los trabajadores, se compara la variación de los salarios en Miami antes y después del éxodo, con la variación en otras ciudades similares en EE.UU. antes y después de cuando se produjo la entrada de inmigrantes. Esto nos permite ver que lo que en último término analizamos es la diferencia entre las variaciones (que son diferencias), lo cual explica el motivo por el que a esta técnica se denomina **estimador de diferencias en diferencias**. Veamos en qué consiste.

Llamamos $\bar{Y}_{antes}^{tratamiento}$ a la media muestral de Y para los sujetos dentro del grupo de tratamiento antes de que sean expuestos al tratamiento, y sea $\bar{Y}_{después}^{tratamiento}$ media muestral de Y para los sujetos dentro del grupo de tratamiento después de que sean expuestos al tratamiento. Para los sujetos (unidades de análisis) que integran el grupo de control definimos de manera análoga las variables $\bar{Y}_{antes}^{control}$, $\bar{Y}_{después}^{control}$. Como hemos dicho, el estimador de diferencias en diferencias es la diferencia entre la variación promedio en Y de aquellos en el grupo de tratamiento y la variación promedio de aquellos en el grupo de control,

$$\hat{\beta}^{DID} = \left(\bar{Y}_{después}^{tratamiento} - \bar{Y}_{antes}^{tramiento}\right) - \left(\bar{Y}_{después}^{control} - \bar{Y}_{antes}^{control}\right) = \Delta\bar{Y}^{tratamiento} - \Delta\bar{Y}^{control}, \tag{6.17}$$

siendo las variaciones promedio postexperimentales y preexperimentales $\Delta\bar{Y}^{tratamiento}$ y $\Delta\bar{Y}^{control}$, respectivamente.

Esta doble diferencia elimina los posibles sesgos asociados a diferencias permanentes entre los dos grupos que no están relacionados con el tratamiento. Imaginemos que en el ejemplo de la inmigración y el efecto sobre los sueldos, estos eran más bajos en Miami antes del éxodo cubano que en otra de las ciudades con las que se hace el cuasiexperimento. Ambos niveles de salarios se explican posiblemente por motivos permanentes de sus mercados laborales. Consideremos, por ejemplo, que después

del éxodo a Miami se registra un descenso de los salarios en Miami, y en ese mismo lapso de tiempo en la ciudad que hace de control los salarios se mantienen iguales debido a factores de su propio mercado de trabajo. En esta situación, si comparamos la diferencia entre los salarios (promedio) en Miami y los salarios de otra ciudad después de la inmigración a Miami, observaremos una diferencia exagerada y no enteramente imputable a la entrada de inmigrantes en el mercado laboral de Miami. Esto es así porque antes del tratamiento ya existía una diferencia en contra del nivel de salarios de Miami, por lo que el efecto del tratamiento no debe incorporar tal diferencia permanente. Eso es precisamente lo que hace el estimador DID al calcular la diferencia entre la variación producida en el grupo de tratamiento (que en este caso es negativa) y la variación en los salarios promedio del grupo de control (que en este ejemplo es nula), por lo que la variación de las diferencias es ahora solo imputable al efecto del tratamiento, evitando así el sesgo inicial.

El estimador DID se puede expresar con la notación habitual de la regresión. Sea ΔY_i la diferencia de Y para el individuo i-ésimo registrada antes y después del experimento. El estimador DID es el estimador MCO de la regresión

$$\Delta Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \tag{6.18}$$

En caso de que no hubiera tratamiento $X_i = 0$, la diferencia experimentada sería imputable a los factores permanentes propios que quedarían recogidos en β_0, cuyo estimador es, en este caso, la media aritmética de las diferencias de Y_i entre individuos. En caso de existir tratamiento, $X_i = 1$, las diferencias individuales se explican en media por el componente permanente y el efecto propio del tratamiento, β_1, que consideramos constante entre los individuos. El estimador MCO del coeficiente β_1 en el modelo anterior es igual a (6.17).

El estimador DID se puede ampliar para incluir regresores adicionales que midan características individuales que estuvieran presentes antes de la realización del experimento. Estos regresores adicionales W_i transforman el modelo (6.18) en un modelo de regresión múltiple

$$\Delta Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + ... + \beta_{1+r} W_{ri} + \varepsilon_i. \tag{6.19}$$

El estimador MCO de β_1 de (6.19) será insesgado siempre que X_i esté asignado «como si fuera aleatorio», condicionado a $W_{1i}, ..., W_{ri}$. Esto es así porque recordemos que en tal caso el error ε_i satisfaría la condición de independencia en media condicionada, y podría tener por tanto un significado causal.

Tanto para el caso de los modelos simple y múltiple (ecuaciones (6.18) y (6.19)) en realidad tenemos un panel de dos periodos (antes y después del tratamiento)[4], por lo que el estimador se puede ampliar a casos en que el número de periodos del panel sea superior a dos.

Un caso diferente al panel es cuando el conjunto de datos procede de una sección cruzada repetida. Este tipo de conjuntos se caracteriza por el hecho de que cada conjunto de datos de sección cruzada corresponde a un periodo de tiempo diferente. Por ejemplo,

[4]Pese a tener un panel de $T = 2$, el estimador DID no tiene en cuenta el hecho singular del panel, esto es, estima ignorando que las observaciones proceden de la misma unidad en ambos periodos.

el conjunto de datos podría estar formado por observaciones de 300 sujetos en el periodo t, y por 350 sujetos diferentes en el periodo $t+1$, lo que configuraría un total de 650 sujetos observados.

Para poder utilizar este conjunto de datos configurado a partir de secciones en dos momentos diferentes es necesario considerar que si los individuos de la sección en t son extraídos aleatoriamente de una población, entonces estos individuos se pueden utilizar como sustitutos de los individuos (sujetos) de los grupos de tratamiento y control en la sección cruzada $t+1$.

En este caso de dos periodos, el modelo de regresión para este tipo de datos sección cruzada repetida es

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 G_i + \beta_3 D_t + \beta_4 W_{1it} + ... + \beta_{3+r} W_{rit} + \varepsilon_{it}, \tag{6.20}$$

donde X_{it} se refiere al tratamiento del $i-$ésimo sujeto en la sección cruzada de tiempo $t, t = 1, 2$; G_i es una variable indicador (variable binaria) de si el sujeto está en el grupo de tratamiento (ya sea antes, tratamiento sustituto, o después del tratamiento); y D_t es otro indicador del periodo en el que está el sujeto (pretratamiento, $t = 0$, o postratamiento, $t = 1$). A patir de estas definiciones resulta fácil comprobar que un sujeto recibe el tratamiento si está en el grupo de tratamiento ($G_i = 1$) y además está en el segundo periodo ($D_t = 1$), es decir un sujeto tratado se caracteriza por $X_{it} = G_i \times D_t$.

Si el cuasiexperimento hace que el tratamiento X_{it} fuera «como si» estuviera asignado al azar, condicionado a los controles W, entonces el efecto causal del tratamiento puede ser estimado por el estimador MCO de $\hat{\beta}_1$. Podemos observar que el modelo de la Ecuación (6.20) nos conduce al mismo estimador de la Ecuación (6.17), por lo que los modelos (6.20) y (6.19) son equivalentes. Para verlo consideremos el caso más simple de (6.19), es decir, cuando no hay W. En tal caso, se puede comprobar fácilmente que

$$\left[\mathbb{E}\left(Y^{tratados}_{después}\right) - \mathbb{E}\left(Y^{tratados}_{antes}\right)\right] - \left[\mathbb{E}\left(Y^{control}_{después}\right) - \mathbb{E}\left(Y^{control}_{antes}\right)\right] = \beta_1,$$

cuyo estimador consistente es (6.17).

En algunos cuasiexperimentos, es posible que tengamos disponibilidad de otra variable adicional, que llamaremos Z, de la que sabemos que influye en la recepción o exposición al tratamiento, X, y que está administrada «como si» fuera al azar entre los sujetos. Por ejemplo, consideremos de nuevo el efecto del tratamiento «ir a la universidad» sobre los salarios. Supongamos que a algunos individuos se le asignó aleatoriamente una ayuda económica para cubrir gastos de formación universitaria. Sea Z la variable binaria que indica si un individuo recibe la ayuda, y que podemos denominar *instrumento*. En este caso es esperable que el instrumento Z_i pueda afectar a la decisión de un individuo sobre ir a la universidad (tratamiento).

En este escenario podemos comprobar que, dado que la variable tratamiento es binaria, entonces el estimador siguiente (conocido por estimador de Wald)

$$\begin{aligned} \hat{\beta}^{wald} &= \frac{\mathbb{E}\left(\widehat{Y_i \,|Z_i = 1}\right) - \mathbb{E}\left(\widehat{Y_i \,|Z_i = 0}\right)}{\mathbb{E}\left(\widehat{X_i \,|Z_i = 1}\right) - \mathbb{E}\left(\widehat{X_i \,|Z_i = 0}\right)} \\ &= \frac{\sum Y_i Z_i / \sum Z_i - \sum Y_i \left(1 - Z_i\right) / \sum \left(1 - Z_i\right)}{\sum X_i Z_i / \sum Z_i - \sum X_i \left(1 - Z_i\right) / \sum \left(1 - Z_i\right)} \end{aligned} \tag{6.21}$$

es consistente. Al tratarse Z_i de una variable binaria, el denominador captura el efecto medio de recibir la ayuda económica sobre la decisión de ir a la universidad, y dado que las ayudas motivarán que ciertos estudiantes, que en otro caso no irían, vayan a la universidad, se espera que sea un número entre 0 y 1. En cambio el numerador es el efecto de la ayuda sobre los salarios, dado que las ayudas aumentan el número de universitarios, lo que incrementa sus salarios. Por tanto, el estimador está ponderando los efectos de los salarios (numerador) por la proporción de la población afectada por la ayuda económica.

La consistencia del estimador quedaría comprobada si realmente $\hat{\beta}^{wald} \xrightarrow{p} \mathbb{E}\left(Y_i(1) - Y_i(0)\right)$, donde, como antes, $Y_i(1)$ y $Y_i(0)$ son resultados potenciales que se obtendrían en caso de ser o no tratados. Recordemos que estos resultados no son observables simultáneamente a nivel individual

$$Y_i = Y_i(1)X_i + Y_i(0)\left(1 - X_i\right).$$

Ahora además contamos con el instrumento binario que puede afectar al tratamiento recibido. Por tanto, el estado del tratamiento dependerá de los valores que tome el instrumento Z_i, por lo que potencialmente tendremos el estado $X_i(1)$ en el caso de que $Z_i = 1$, y alternativamente el estado será potencialmente $X_i(0)$ si $Z_i = 0$. Ahora también cabe decir que para un individuo solo podremos observar uno de los posibles tratamientos

$$X_i = Z_i X_i(1) + \left(1 - Z_i\right) X_i(0). \tag{6.22}$$

Conviene observar que, por un lado, suponemos que el instrumento afecta al tratamiento observado (recibido), esto es, la probabilidad de recibir tratamiento en caso de que $Z_i = 1$ es diferente de la probabilidad de recibir tratamiento si $Z_i = 0$,

$$Pr\left(X_i(1) = 1\right) \neq Pr\left(X_i(0) = 1\right).$$

Por otro lado, asumimos que el instrumento Z_i está asignado aleatoriamente, lo que implica que es independiente también de los tratamientos potenciales $X_i(1), X_i(0)$,

$$Z_i \perp Y_i(0), Y_i(1), X_i(0), X_i(1). \tag{6.23}$$

Como hemos dicho, nuestro interés está en estimar el efecto potencial del tratamiento, que vamos a considerar de nuevo constante para los individuos

$$\mathbb{E}\left(Y_i(1) - Y_i(0)\right) = \beta,$$

por lo que el efecto medio del tratamiento también será β.

En estas circunstancias la Ley de los grandes números nos garantiza que

$$\hat{\beta}^{wald} \xrightarrow{p} \frac{\mathbb{E}\left(Y_i \,|Z_i = 1\right) - \mathbb{E}\left(Y_i \,|Z_i = 0\right)}{\mathbb{E}\left(X_i \,|Z_i = 1\right) - \mathbb{E}\left(X_i \,|Z_i = 0\right)},$$

que está expresado en resultados no-potenciales (observados), y que podemos relacionar con los potenciales, que son en los que están expresados los efectos causales promedio.

Así, el numerador puede expresarse

$$\begin{aligned}\mathbb{E}\left(Y_i\left|Z_i=1\right.\right)-\mathbb{E}\left(Y_i\left|Z_i=0\right.\right) &= \mathbb{E}\left(Y_i(1)X_i(1)+Y_i(0)\left(1-X_i(1)\right)\left|Z_i=1\right.\right)\\ &\quad -\mathbb{E}\left(Y_i(1)X_i(0)+Y_i(0)\left(1-X_i(0)\right)\left|Z_i=0\right.\right)\\ &= \mathbb{E}\left(Y_i(1)X_i(1)+Y_i(0)\left(1-X_i(1)\right)\right)\\ &\quad -\mathbb{E}\left(Y_i(1)X_i(0)+Y_i(0)\left(1-X_i(0)\right)\right)\\ &= \mathbb{E}\left[\left(Y_i(1)-Y_i(0)\right)\left(X_i(1)-X_i(0)\right)\right]\\ &= \beta\mathbb{E}\left[\left(X_i(1)-X_i(0)\right)\right],\end{aligned}$$

donde la segunda igualdad proviene de la independencia de Z_i (expresión (6.23)) y la tercera del supuesto de efectos constantes.

El denominador se simplifica utilizando (6.22) a $\mathbb{E}\left(X_i(1)\right)-\mathbb{E}\left(X_i(0)\right)$, por lo que el cociente indica que

$$\hat{\beta}^{wald}\xrightarrow{p}\frac{\mathbb{E}\left(Y_i\left|Z_i=1\right.\right)-\mathbb{E}\left(Y_i\left|Z_i=0\right.\right)}{\mathbb{E}\left(X_i\left|Z_i=1\right.\right)-\mathbb{E}\left(X_i\left|Z_i=0\right.\right)}=\beta.$$

A este estimador consistente es al que llegamos utilizando la técnica de las variables instrumentales de los temas anteriores. Para comprobarlo observemos que podemos escribir

$$Y_i(1)=Y_i(0)+\beta,$$

por lo que el resultado observado será

$$Y_i=Y_i(0)+\beta X_i=\mathbb{E}\left(Y_i(0)\right)+\beta X_i+\varepsilon_i,\varepsilon_i\equiv Y_i(0)-\mathbb{E}\left(Y_i(0)\right),$$

que podemos reescribir del modo habitual

$$Y_i=\beta_0+\beta X_i+\varepsilon_i,\beta_0\equiv\mathbb{E}\left(Y_i(0)\right). \tag{6.24}$$

Dado que estamos considerando que los potenciales resultados puedan estar correlacionados con el tratamiento, es decir, dado que fácilmente podemos tener un problema de endogeneidad, $\mathbb{E}\left(\varepsilon_i\left|X_i\right.\right)\neq 0$, entonces existe riesgo derivado de la falta de consistencia en los estimadores. Este problema lo podremos paliar, como hemos visto en el tema dedicado a las variables instrumentales, en la medida en la que exista un instrumento, Z_i, que sea independiente de los resultados potenciales. Precisamente este es el supuesto que hacemos en (6.23), y por tanto sabemos con certeza que el estimador consistente existe y es el de la expresión (8.19), es decir, el estimador MC2E. Dada la naturaleza binaria de las variables instrumento Z_i y tratamiento X_i.

Bibliografía complementaria

Matilla-García, M et al. 2017. Econometría y Predicción. McGraw Hill

Stock J. and Watson J. Introducción a la econometría. Pearson.

Tema 7

Tests de especificación y selección de modelos.

Este tema está elaborado como una adaptación de

Wooldridge. J. 4th Ed. , Introductory Econometrics. Capítulo 9. Wooldridge. J. Econometric Analysis of Cross Section and Panel Data. Capítulo 6. A.C. Cameron and P.K. Trivedi. Microeconometrics: methods and applications. Capítulo 8. Secciones 8.1, 8.2, 8.3, 8.4 8.5, 8.6 y 8.7. Así como de la bibliografía complementaria.

- Introducción
- Tests tipo m
- Tests de Hausman
- Tests para algunos errores comunes
- Discriminación entre modelos no anidados
- Consecuencias de los tests
- Diagnosis de modelos

7.1 Introducción

Dos aspectos prácticos importantes en la modelización econométrica son determinar si un modelo se especifica correctamente y seleccionar modelos alternativos. Para estos propósitos, a menudo es posible utilizar contrastes de hipótesis conocidos y ya presentado para el caso en el que los modelos estén anidados.

La estimación de un modelo de regresión mal especificado generalmente produce estimaciones de parámetros sesgadas e inconsistentes. Esto es cierto para los modelos de regresión siempre que omitimos incorrectamente uno o más regresores que están correlacionados con los regresores incluidos en el modelo. Excepto en ciertos casos especiales, también es cierto para tipos más generales de errores de especificación. Esto sugiere que la especificación de cada modelo econométrico debe ser probada a fondo antes de que aceptemos siquiera provisionalmente sus resultados.

Los procedimientos que pueden usarse como pruebas de especificación son muy variados, algunos de ellos ya están tratados (o se verán también en otros temas). Pensemos en

pruebas tipo-t tipo-F para variables omitidas, o contrastes de constancia de parámetros, o los contraste de heterocedasticidad, o los de correlación serial (Sección 7.7).

Este tema presentamos pruebas estadísticas conocidas como tests tipo-m (m-tests) basados en contrastes sobre si se satisfacen las restricciones impuestas a los momentos condicionales. En estos contrastes no hay una declaración explícita de un modelo con hipótesis alternativas.

Otros contrastes especialmente abundantes son los contrastes tipo Hausman que contrastan la diferencia entre dos estimadores consistentes si el modelo se especifica correctamente pero divergen si el modelo se especifica incorrectamente.

Finalmente, es interesante contar con contrastes para selección de modelos no anidados requieren métodos especiales porque el enfoque habitual de prueba de hipótesis solo se puede aplicar cuando un modelo está anidado dentro de otro.

Estos métodos se utilizan en un ciclo de especificación, estimación, prueba y evaluación del modelo. Este ciclo puede pasar de un modelo general a un modelo específico, o de un modelo específico a uno más general que se considera que captura las características más importantes de los datos.

7.2 m-Tests

En estadística y econometría, los **estimadores de extremos** son una clase de estimadores para modelos paramétricos que se calculan mediante la maximización (o minimización) de una determinada función objetivo, que depende de los datos.

Precisamente los **estimadores M** son a su vez una clase amplia de estimadores extremos para los cuales la función objetivo es un promedio muestral, es decir, una media. El método de mínimos cuadrados es un estimador M prototípico, ya que el estimador se define como el mínimo de la suma de cuadrados de los residuos. Otro estimador M popular es la estimación por máxima verosimilitud (ML o MV). Para una familia de funciones de densidad de probabilidad f parametrizadas por ϑ, se calcula un estimador de máxima verosimilitud de ϑ para cada conjunto de datos maximizando la función de verosimilitud sobre el espacio paramátrico ϑ. Lógicamente, tanto los mínimos cuadrados no lineales como la estimación de máxima verosimilitud son casos especiales de estimadores M. El procedimiento estadístico de evaluar un estimador M en un conjunto de datos se llama estimación M o tipo m.

De manera más general, un estimador M puede definirse como un cero de una función de estimación. Esta función de estimación es a menudo la derivada de otra función estadística. Por ejemplo, una estimación de máxima verosimilitud es el punto donde la derivada de la función de verosimilitud con respecto al parámetro es cero; por lo tanto, un estimador de máxima verosimilitud es un punto crítico de la función de puntuación.

Los tests tipo-m provienen de estimadores M. Se trata de contrastes (tests) de momentos condicionales, y son un procedimiento para contrastes de especificación general que abarca muchos contrastes de especificación comunes.

Suponga que un modelo implica la condición poblacional sobre el momento

$$H_0 : \mathbb{E}[\mathbf{m}_i(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}$$

donde **w** es un vector de observables, generalmente la variable dependiente **y** y regresores **x** y, a veces, variables adicionales **z**, ϑ es un vector $q \times 1$ de parámetros, y $\mathbf{m}_i(\cdot)$ es un vector h × 1. Un ejemplo simple es

$$\mathbf{m}(\mathbf{w}, \boldsymbol{\beta}) = (y - \mathbf{x}'\boldsymbol{\beta})\mathbf{z}$$

de modo que

$$\mathbb{E}[(y - \mathbf{x}'\boldsymbol{\beta})\mathbf{z}] = \mathbf{0}$$

si el vector z puede ser omitido en el modelo lineal $y = \mathbf{x}'\boldsymbol{\beta} + \mathbf{u}$.

Un test-m es un contraste de la proximidad a cero del momento muestral correspondiente, es decir, de

$$\widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}}) = N^{-1} \sum_{i=1}^{N} [\mathbf{m}_i(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})].$$

Si H_0 es cierta, entonces se puede demostrar que

$$\sqrt{N}\widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}}) \underset{d}{\rightarrow} Normal[\mathbf{0}, \mathbf{V}_m]$$

donde

$$\mathbf{V}_m = \mathbf{H}_0 \mathbf{J}_0 \mathbf{H}_0'$$

con

$$\mathbf{H}_0 = [\mathbf{I} - \mathbf{C}_0 \mathbf{A}_0^{-1}]$$

donde $\mathbf{C}_0 = plim(N^{-1} \sum_i \partial \mathbf{m}_{i0} / \partial \boldsymbol{\theta}')$, $\mathbf{A}_0 = plim(N^{-1} \sum_i \partial \mathbf{s}_{i0} / \partial \boldsymbol{\theta}')$ y

$$\mathbf{J}_0 = plim[N^{-1} \begin{bmatrix} \sum_i \mathbf{m}_{i0}\mathbf{m}_{i0}' & \sum_i \mathbf{m}_{i0}\mathbf{s}_{i0}' \\ \sum_i \mathbf{s}_{i0}\mathbf{m}_{i0}' & \sum_i \mathbf{s}_{i0}\mathbf{s}_{i0}' \end{bmatrix}$$

siendo $\mathbf{m}_{i0} = \mathbf{m}_i(\mathbf{w}_i, \boldsymbol{\theta}_0)$ y $\mathbf{s}_{i0} = \mathbf{s}_i(\mathbf{w}_i, \boldsymbol{\theta}_0)$. La complejidad de esta varianza procede fundamentalmente debido a que $\mathbf{m}_{i0} = \mathbf{m}_i(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})$ tiene dos fuentes de variación estocástica:$(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})$.

Naturalmente es posible derivar un test tipo chi-cuadrado a partir de la forma cuadrática con distribución tabulada baja la hipótesis nula:

$$M = N\widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}})' \widehat{\mathbf{V}}_m^{-1} \widehat{\mathbf{m}}_N(\widehat{\boldsymbol{\theta}}) \underset{d}{\rightarrow} \chi^2(rango[\mathbf{V}_m]).$$

La hipótesis nula se rechaza si el estadístico M está alejado estadísticamente de cero.

El proceso de construcción de este tipo de contrastes tipo-m es aparentemente muy simple. Su dificicultad está en calcular $\widehat{\mathbf{V}}_m$ y en seleccionar los momentos a contrastar $\mathbf{m}(\cdot)$.

7.3 Tests de Hausman

Las tests o contrastes basados en comparaciones entre dos estimadores diferentes se denominan tests de Hausman.

Considere un contraste de endogeneidad de un regresor en una sola ecuación. Dos estimadores alternativos son los estimadores MCO y MC2E, donde el estimador MC2E utiliza instrumentos para controlar la posible endogeneidad del regresor. Si hay endogeneidad, MCO es inconsistente, por lo que los dos estimadores tendrán un límite de probabilidad diferente. Por el contrario, si no hay endogeneidad, ambos estimadores son consistentes, por lo que los dos estimadores tienen el mismo límite de probabilidad. Esto sugiere probar la endogeneidad probando la diferencia entre los estimadores MCO y MC2E.

En particular consideremos dos estimadores $\widehat{\boldsymbol{\theta}}$ y $\widetilde{\boldsymbol{\theta}}$. Los tests tipo Hausman contrastan

$$H_0 : plim(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}) = \mathbf{0}$$
$$H_a : plim(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}) \neq \mathbf{0}$$

Suponga que la diferencia, convenientemente escalada por raíz de N, entre los dos estimadores consistentes es también consistente bajo H_0 con media $\mathbf{0}$ y una distribución asintótica normal, de modo que

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}) \underset{d}{\rightarrow} Normal[\mathbf{0}, \mathbf{V}_H]$$

donde $\mathbf{V}_H$ denota la matriz de varianza en la distribución límite. Entonces el estadístico de la prueba (o test) de Hausman es

$$H = (\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})'(N^{-1}\widehat{\mathbf{V}}_H)^{-1}(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}) \tag{7.1}$$

cuya distribución asintótica, si la hipótesis nula es cierta, es $\chi^2(q)$. Así pues, rechazamos H_0 al nivel α si $H > \chi^2_\alpha(q)$.

Los tests de Hausman se pueden aplicar solo a un subconjunto de los parámetros. Por ejemplo, el interés puede residir únicamente en el coeficiente del regresor posiblemente endógeno y si cambia al pasar de MCO a MC2E. En tal caso, solo se usa un componente de ϑ y el estadístico de prueba se distribuye $\chi^2(1)$. Como en otras configuraciones, este test sobre un subconjunto de parámetros puede llevar a una conclusión diferente a la del test sobre todos los parámetros.

7.4 Tests para algunos errores comunes de especificación

Tests de endogeneidad

Los estimadores de variables instrumentales solo deben usarse cuando sea necesario (muchas veces lo son como veremos en otros temas). Esto es así ya que, si todos los

regresores son exógenos, entonces los estimadores de mínimos cuadrados son más eficientes. Por lo tanto, puede ser útil probar si se necesitan métodos alternativos como los denominados de variables instrumentales (IV). Un test de endogeneidad de regresores compara las estimaciones por IV con las estimaciones de MCO. Si los regresores son endógenos, entonces en el límite estas estimaciones diferirán, mientras que si los regresores son exógenos, los dos estimadores no diferirán. Por lo tanto, las grandes diferencias entre las estimaciones de MCO y IV pueden interpretarse como evidencia de endogeneidad. Este marco nos indica que estamos en condiciones de aplicar un test tipo Hausman.

Considere el modelo de regresión lineal

$$y = \mathbf{x}_1'\boldsymbol{\beta}_1 + \mathbf{x}_2'\boldsymbol{\beta}_2 + u$$

donde $\mathbf{x}_1$ es potencialmente endógeno, mientras que las variables en $\mathbf{x}_2$ no lo son, es decir, son exógenas. Sean $\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}$ los estimadores MCO y MC2E del modelo anterior. Sabemos que si los errores son homocedásticos, MCO es eficiente bajo la hipótesis nula de no endogeneidad. Un test tipo Hausman de endogeneidad de $\mathbf{x}_1$ desarrollando la prueba H indicada en (7.1). En efecto, si el estimado MCO es eficiente bajo la hipótesis nula, entonces $cov(\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}) = Var(\widehat{\boldsymbol{\beta}})$, esto implica que

$$Var(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}) = Var(\widetilde{\boldsymbol{\beta}}) - Var(\widehat{\boldsymbol{\beta}})$$

por lo que, (7.1) se convierte en

$$H = (\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})'(Var(\widetilde{\boldsymbol{\beta}}) - Var(\widehat{\boldsymbol{\beta}}))^{-1}(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}) \tag{7.2}$$

dado que

$$N^{-1}\widehat{\mathbf{V}}_H = Var(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}) = Var(\widehat{\boldsymbol{\beta}}) + Var(\widetilde{\boldsymbol{\beta}}) - 2cov(\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}).$$

Es posible mostrar que el test se puede realizar más fácilmente contrastando $\boldsymbol{\gamma} = \mathbf{0}$ en la regresión MCO aumentada siguiente

$$y = \mathbf{x}_1'\boldsymbol{\beta}_1 + \mathbf{x}_2'\boldsymbol{\beta}_2 + \widehat{\mathbf{x}}_1'\boldsymbol{\gamma} + u$$

donde $\widehat{\mathbf{x}}_1$ es la predicción de los regresores exógenos de la regresión de $\mathbf{x}_1$ sobre los instrumentos $\mathbf{z}$. El procedimiento es equivanlente a contrastar $\boldsymbol{\gamma} = \mathbf{0}$ en la regresión MCO aumentada siguiente alternativa

$$y = \mathbf{x}_1'\boldsymbol{\beta}_1 + \mathbf{x}_2'\boldsymbol{\beta}_2 + \widehat{\mathbf{v}}_1'\boldsymbol{\gamma} + u \tag{7.3}$$

donde $\widehat{\mathbf{v}}_1$ son los residuos de la regresión e $\mathbf{x}_1$ sobre los instrumentos $\mathbf{z}$. La intuición para estos tests es que si u no está correlacionado con $\mathbf{x}_1$ y $\mathbf{x}_2$ en el modelo

$$y = \mathbf{x}_1'\boldsymbol{\beta}_1 + \mathbf{x}_2'\boldsymbol{\beta}_2 + u \tag{7.4}$$

entonces $\gamma = \mathbf{0}$. Si en cambio u está correlacionado con $\mathbf{x}_1$, entonces esto será recogido significativamente en transformaciones sobre $\mathbf{x}_1$ como son $\widehat{\mathbf{v}}_1$ y $\widehat{\mathbf{x}}_1$.

En el caso en el que solo un componente del vector de parámetros ($\boldsymbol{\theta} = [\mathbf{x}_1, \mathbf{x}_2]$) del modelo fuera objeto de contraste, entonces el test (7.2) sería tan sencillo como

$$H = \frac{(\widehat{\theta} - \tilde{\theta})}{(\widetilde{s}^2 - \hat{s}^2)} \sim \chi^2(1)$$

siendo $\hat{s}, \widetilde{s}$ los errores estándar de $\widehat{\theta}, \tilde{\theta}$, respectivamente.

En el caso más general de heterocedasticidad en los errores, las estimaciones MCO serían ineficientes y por tanto no podríamos usarlas. Sin embarago sería correcto usar las regresiones (7.3) y (7.4) utilizando estimadores de la varianza robustos a la heterocedasticidad.

Por completar la exposición, suponga que

$$y = \mathbf{x}_1'\boldsymbol{\beta}_1 + \mathbf{x}_2'\boldsymbol{\beta}_2 + \mathbf{x}_3'\boldsymbol{\beta}_3 + u$$

donde $\mathbf{x}_1$ es potencialmente endógeno, las variables en $\mathbf{x}_2$ asumimos que son endógenas, y se asume que $\mathbf{x}_3$ es exógeno. Entonces, la endogeneidad de $\mathbf{x}_1$ se puede contrastar comparando el estimador MC2E usando instrumentos únicamente para $\mathbf{x}_2$ con el estimador MC2E con instrumentos para $\mathbf{x}_1$ y para $\mathbf{x}_2$.

Tests de exogeneidad de instrumentos

Si se utiliza un estimador de IV, los instrumentos deben ser exógenos para que el estimador de IV sea consistente. Para modelos exactamente identificados, no es posible contrastar la exogeneidad del instrumento. En tal caso precisamos utilizar argumentos a priori para justificar la validez del instrumento. Sin embargo en modelos sobreidentificados, es posible construir una prueba o contraste de exogeneidad de instrumentos.

En la notación del tema anterior, el estimador GMM se basa en el supuesto de que $\mathbb{E}[\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta}_0)] = \mathbf{0}$. Si el modelo está sobreidentificado, entonces solo q de estas restricciones sobre los momentos se utilizan en la estimación, lo que lleva a $(r - q)$ condiciones de ortogonalidad linealmente dependientes, siendo $r = dim[\mathbf{h}(\cdot)]$, que se pueden usar para formar una contraste tipo-m de los vistos a comienzo de este tema.

En el caso particular del modelo habitual $y = \mathbf{x}'\boldsymbol{\beta} + u$ con instrumentos $\mathbf{z}$, son efectivamente instrumentos válidos si

$$\mathbb{E}[u|\mathbf{z}] = \mathbf{0} \text{ o si} \mathbb{E}[\mathbf{z}u] = \mathbf{0}$$

Un contraste obvio de

$$H_0 : E[\mathbf{z}u] = \mathbf{0}$$

es que $N^{-1}\sum_i \mathbf{z}_i\hat{u}_i$ se aleje del cero. En el caso sobreidentificado, la prueba de restricciones de sobreidentificación se puede elaborar a partir de (6.7), llegando a la siguiente expresión

$$(\hat{\mathbf{u}}'\mathbf{Z})\,\hat{\mathbf{S}}^{-1}\,(\mathbf{Z}'\hat{\mathbf{u}}) \sim \chi^2(r - K)$$

donde

$$\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$$

y $\widehat{\boldsymbol{\beta}}$ es -en este caso- el estimador óptimo GMM (en el sentido de que minimiza $(\mathbf{u}'\mathbf{Z})\,\hat{\mathbf{S}}^{-1}\,(\mathbf{Z}'\mathbf{u})$), y $\hat{\mathbf{S}}$ es un estimador consistente del $plim[N^{-1}\sum_i u_i^2\mathbf{z}_i\mathbf{z}_i']$.

El rechazo de la hipótesis nula generalmente se interpreta como evidencia de que los instrumentos **z** son endógenos, pero también podría ser evidencia de una especificación incorrecta del modelo.

Test RESET

Una especificación errónea de forma funcional bastante habitual es la que proviene de no considerar alguna(s) relación no-lineal en los regresores. Por ejemplo, dentro del esquema de regresión $\mathbf{y} = \mathbf{x}'\boldsymbol{\beta} + \mathbf{u}$, asumimos que los regresores entran linealmente y no están correlacionados asintóticamente con el error **u**. Para contrastar la no linealidad, un enfoque sencillo es ingresar funciones de potencia de variables exógenas (generalmente cuadrados de las exégoneas) como regresores independientes adicionales y probar la significancia estadística de estas variables adicionales usando un test de Wald o un contraste tipo-F. Proceder de este modo requiere que el investigador tenga razones específicas para considerar la no linealidad

El denominado contraste RESET es un test de variables omitidas de la regresión que puede formularse como un test sobre la forma funcional. La propuesta es ajustar la regresión inicial y generar nuevos regresores que sean funciones de valores ajustados $\hat{\mathbf{y}} = \mathbf{x}'\hat{\boldsymbol{\beta}}$, como son

$$\mathbf{w} = [\hat{\mathbf{y}}^2, \hat{\mathbf{y}}^3, ..., \hat{\mathbf{y}}^p]$$

Posteriomente, se estima el modelo el modelo

$$\mathbf{y} = \mathbf{x}'\boldsymbol{\beta} + \mathbf{w}'\boldsymbol{\gamma} + \mathbf{u}$$

y el contraste de no linealidad es un contraste tipo Wald de p restricciones

$$H_0 : \boldsymbol{\gamma} = \mathbf{0}$$

$$H_a : \boldsymbol{\gamma} \neq \mathbf{0}$$

utilizando generalmente valores de p no superiores a 3.

La lógica del contraste es la siguiente. Bajo el supuesto de exogeneidad habitual, sabemos que cualquier función de **x** no está correlacionada con **u** (motivo por el que podemos incluir cuadrados y/o productos cruzados de **x** como regresores adicionales). En tal caso $\hat{\mathbf{y}}^p = \left(\mathbf{x}'\hat{\boldsymbol{\beta}}\right)^p$ no está correlacionado con **u** para cualquier entero p. Como no observamos ß lo reemplazamos con el estimador MCO, $\hat{\boldsymbol{\beta}}$. Por definición de MCO, la covarianza muestral entre $\hat{u}_i$ y $\hat{y}_i$ es cero. Podríamos entonces contrastar si los $\hat{u}_i$ están suficientemente correlacionados con polinomios de orden bajo en $\hat{y}_i$ como prueba de no linealidad. Una forma de hacerlo es agregar estos términos a la ecuación tal y

como hemos indicado más arriba, lo cual nos llevaría una F estándar, $F_{p-1,N-K-p-1}$. Otra posibilidad es usar un test LM: regresamos $\hat{u}_i$ sobre $\hat{\mathbf{y}}^2, \hat{\mathbf{y}}^3, ..., \hat{\mathbf{y}}^p$ para formar

$$NR^2 \sim \chi^2_{p-1}$$

donde el R-cuadrado se refiere a la última regresión auxiliar indicada.

7.5 Discriminación entre modelos no anidados

Dos modelos están (son) anidados si uno es un caso especial del otro; por el contrario no están anidados si ninguno puede representarse como un caso especial del otro. Es posible discriminar entre modelos anidados utilizando un contraste de hipótesis estándar de restricciones paramétricas tales que que reduzcan un modelo al otro. En el caso no anidado, sin embargo, es necesario desarrollar métodos alternativos.

Consideremos dos modelos, por ahora, que hacemos explícitos

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i. \tag{7.5}$$

$$Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + ... + \beta_k Z_{ki} + \varepsilon_i \tag{7.6}$$

Así los modelos (8.1) y (7.5) son modelos no anidados puesto que las variables independientes de ambos modelos son diferentes (no es necesario que todas las variables independientes sean distintas, basta con que alguna de las variables en ambos modelos sean distintas para que el modelo sea no anidado). Cuando los modelos son no anidados, no podemos utilizar los contrastes tipo-*F*.

Así pues de modo más general, supongamos dos modelos hipotéticos para explicar una misma variable dependiente

$$H_1 : \mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}_1$$
$$H_2 : \mathbf{y} = \mathbf{z}(\boldsymbol{\gamma}) + \mathbf{u}_2$$

donde los vectores paramétricos son de longitud k_1, k_2. Decimos que estos modelos no están anidados si en general es imposible localizar retricciones en el vector $\boldsymbol{\beta}$ tales que, para un arbitrario $\boldsymbol{\gamma}$, $\mathbf{x}(\boldsymbol{\beta})$ sea igual $\mathbf{z}(\boldsymbol{\gamma})$, y que sea imposible restringir $\boldsymbol{\gamma}$ de tal manera que para cualquier $\boldsymbol{\beta}$, $\mathbf{z}(\boldsymbol{\gamma})$ sea igual a $\mathbf{x}(\boldsymbol{\beta})$. Así pues, no debe de existir ninguna aplicación, digamos $\mathbf{g}$, definida sobre todo el espacio paramétrico sobre el cual $\boldsymbol{\gamma}$ está definida, tal que $\mathbf{z}(\boldsymbol{\gamma}) = \mathbf{x}(g(\boldsymbol{\gamma}))$, e igualmente no debe de existir una $\mathbf{h}$ tal que $\mathbf{x}(\boldsymbol{\beta}) = \mathbf{z}(\mathbf{h}(\boldsymbol{\beta}))$.

Si particularizamos para dos funciones lineales, tendríamos por ejemplo

$$x(\boldsymbol{\beta}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \tag{7.7}$$

$$z(\boldsymbol{\gamma}) = \gamma_0 + \gamma_1 X_{1i} + \gamma_3 X_{3i} \tag{7.8}$$

donde apreciamos que cada función de regresión contiene un regresor que no está en la otra, y por tanto no están anidados. Sin embargo, si a la expresión (7.8) le añadimos el término X_{2i}, tenemos una nueva función de regresión

$$z^*(\boldsymbol{\gamma}) = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{3i} \tag{7.9}$$

que en este caso estaría anidada en (7.7) puesto que la retricción $\gamma_3 = 0$ haría que (7.9) fuera equivalente a (7.7).

Una solución inicial, dentro del marco de la estimación por mínimos cuadrados, es estimar un **modelo artificial** general que contenga las variables explicativas de ambos modelos, es decir, estimar

$$Y_i = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{2i} + ... + \delta_k X_{ki} + \delta_{k+1} Z_{(k+1)i} + \delta_{k+2} Z_{(k+2)i} + ... + \delta_{2k} Z_{2ki} + \varepsilon_i. \tag{7.10}$$

Y contrastar mediante la «*F*» habitual, la hipótesis nula de que el modelo correcto es (8.1) «$H_0 : \delta_{k+1} = \delta_{k+2} = ... = \delta_{2k} = 0$», y posteriormente contrastar la hipótesis nula de que el otro modelo (7.5) es correcto «$H_0 : \delta_1 = \delta_2 = ... = \delta_k = 0$».

Alternativamente podemos utilizar la prueba «***J***», que proviente de «***J***oint», basada en el estadístico «*t*» usual, en la siguiente expresión:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \phi_1 \widehat{Y}_i^Z + e_i, \tag{7.11}$$

donde la variable «$\widehat{Y}_i^Z$» es la estimación MCO del modelo (7.6). Si «ϕ_1» es significativo, rechazamos el modelo (8.1).

Lo mismo hacemos a partir de la expresión (7.6), estimando el modelo ampliado siguiente:

$$Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + ... + \beta_k Z_{ki} + \phi_2 \widehat{Y}_i^X + e_i, \tag{7.12}$$

donde la variable «$\widehat{Y}_t^X$» es la estimación mínimo cuadrática del modelo (8.1). Si «ϕ_2» es significativo, rechazamos el modelo (7.5).

Los contrastes de modelos no anidados pueden llevarnos a soluciones en las que no prevalece un modelo sobre otro, es decir al rechazo o «aceptación» de ambos modelos. En el caso de «aceptación» de ambos modelos (ϕ_1 y ϕ_2 no significativos) podemos utilizar el coeficiente de determinación, corregido o sin corregir, o los criterios de Akaike o Schwarz para decidirnos por uno de ellos. En el caso de rechazo de ambos (ϕ_1 y ϕ_2 significativos) tendremos que seguir trabajando la especificación del modelo.

Probar cada uno de dos modelos no anidados contra el otro puede permitirnos o no elegir un modelo sobre el otro. De manera más general, si tenemos m modelos y realizamos $m(m-1)$ pruebas por pares, no podemos esperar razonablemente encontrar que uno y solo uno de los modelos nunca sea rechazado. Por tanto, si nuestro objetivo es elegir el mejor modelo entre m modelos en competencia, y no nos importa si incluso el mejor modelo es falso, no necesariamente deberíamos utilizar pruebas de hipótesis

no anidadas vistas anteriormente. En su lugar, deberíamos utilizar un procedimiento diseñado explícitamente para la selección del modelo. Dicho procedimiento generalmente implica calcular algún tipo de función de criterio para cada uno de los modelos y seleccionar el modelo para el cual esa función se maximiza o minimiza.

En el marco de estimación es máxima verosimilitud, tenemos los criterios basados en información, que tienen una lógica de usabilidad parecida a las decisiones basadas en el estadístico R-cuadrado.

Los criterios de información son criterios de (log)verosimilitud con grados de ajuste de libertad. Se prefiere el modelo con el criterio de información más pequeño. La intuición es que existe una tensión entre el ajuste del modelo, medido por el valor que maximiza la (log)verosimilitud, y el principio de parsimonia, que favorece un modelo simple. El ajuste del modelo se puede mejorar aumentando la complejidad del modelo. Sin embargo, los parámetros solo se agregan si la mejora resultante en el ajuste compensa suficientemente la pérdida de parsimonia. Tenga en cuenta que desde este punto de vista no es necesario que el conjunto de modelos en consideración deba incluir el "verdadero proceso generador de datos". Los diferentes criterios de información varían en la medida en que penalizan la complejidad del modelo.

Akaike propuso originalmente el criterio de información de Akaike

$$AIC = -2lnL + 2q$$

donde q es el número de parámetros, prefiriéndose el modelo con el AIC más bajo.

Se han propuesto un número considerable de modificaciones a AIC, todas de la forma

$$-2lnL + g(q, N)$$

para la función de penalización especificada g $(\cdot)$ que excede $2q$.

El más popular es el criterio de información bayesiano

$$BIC = -2lnL + (lnN)q$$

de Schwarz.

Un refinamiento de AIC de espíritu similar a BIC es

$$CAIC = -2lnL + (1 + lnN)$$

Si la parsimonia del modelo es lo importante, entonces BIC se usa más ampliamente ya que la penalización del tamaño del modelo para AIC es relativamente baja.

Un tratamiento más general es el basado en la contrucción de un test de selección de modelos que además permita que las funciones de densidad condicional implicadas en la comparación de modelos sean desconocidas.

La idea es asignar significación estadística entre la diferencia en las funciones (log) verosimilitud de dos modelos no anidados competitivos.

Sea $f_1(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}_1)$ y $f_2(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}_2)$ dos modelos competitivos para explicar la densidad $D(\mathbf{y}_i|\mathbf{x}_i)$, donde posiblemente ambos modelos estén mal especificados. A partir de los estimadores cuasi-verosímiles $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2$ (vistos en otra parte de este temario) que covergen a unos vectores pseudo-verdaderos $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*$, evaluamos la función de cuasi-verosimilitud $\mathcal{L}_m = \sum_{i=1}^N \ell_{im}(\hat{\boldsymbol{\theta}}_m)$ en $m = 1, 2$, y entonces se puede demostrar que

$$\begin{aligned} N^{-1/2}\left(\mathcal{L}_1 - \mathcal{L}_2\right) &= N^{-1/2}\sum_{i=1}^{N}\left[\ell_{i1}(\hat{\boldsymbol{\theta}}_1) - \ell_{i2}(\hat{\boldsymbol{\theta}}_2)\right] \\ &= N^{-1/2}\sum_{i=1}^{N}\left[\ell_{i1}(\boldsymbol{\theta}_1^*) - \ell_{i2}(\boldsymbol{\theta}_2^*)\right] + o_P(1) \end{aligned}$$

es decir, esta última expresión indica que los estimadores cuasi-verosímiles $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2$ no afectan a la distribución asintótica de $N^{-1/2}\left(\mathcal{L}_1 - \mathcal{L}_2\right)$. Por tanto, podemos obtener una distribución normal asintótica para $N^{-1/2}\left(\mathcal{L}_1 - \mathcal{L}_2\right)$ bajo la hipótesis nula

$$H_0 : \mathbb{E}[\ell_{i1}(\boldsymbol{\theta}_1^*)] = \mathbb{E}[\ell_{i2}(\boldsymbol{\theta}_2^*)]$$

en particular bajo la hipótesis nula se tiene

$$N^{-1/2}\sum_{i=1}^{N}\left[\ell_{i1}(\boldsymbol{\theta}_1^*) - \ell_{i2}(\boldsymbol{\theta}_2^*)\right] \underset{d}{\to} Normal(0, \eta^2)$$

donde

$$\eta^2 = N^{-1}\sum_{i=1}^{N}\left[\ell_{i1}(\hat{\boldsymbol{\theta}}_1) - \ell_{i2}(\hat{\boldsymbol{\theta}}_2)\right]^2$$

cuyo estimador consistente es

$$\hat{\eta}^2 = N^{-1}\sum_{i=1}^{N}\left[\ell_{i1}(\hat{\boldsymbol{\theta}}_1) - \ell_{i2}(\hat{\boldsymbol{\theta}}_2)\right]^2.$$

El estadístico de selección de modelos conocido como estadístico de Voung

$$\begin{aligned} VMS &= N^{-1/2}\left(\mathcal{L}_1 - \mathcal{L}_2\right)/\hat{\eta} \\ &= \frac{N^{-1}\sum_{i=1}^{N}\left[\ell_{i1}(\hat{\boldsymbol{\theta}}_1) - \ell_{i2}(\hat{\boldsymbol{\theta}}_2)\right]}{\left\{N^{-1}\sum_{i=1}^{N}\left[\ell_{i1}(\hat{\boldsymbol{\theta}}_1) - \ell_{i2}(\hat{\boldsymbol{\theta}}_2)\right]^2\right\}^{1/2}/\sqrt{N}} \underset{d}{\to} Normal(0,1) \end{aligned}$$

No debemos utilizar la estadística VMS y su distribución normal estándar limitante para probar modelos anidados con la especificación correcta. Recuerde que el estadístico de razón de verosimilitud LR es simplemente $LR = 2\left(\mathcal{L}_{no-restringido} - \mathcal{L}_{restringido}\right)$, y, bajo la nula, LR tiene una distribución chi-cuadrado con Q igual al número de restricciones. El punto importante es que, si los modelos están anidados y especificados correctamente, "" entonces $\ell_{i1}(\boldsymbol{\theta}_1^*) - \ell_{i2}(\boldsymbol{\theta}_2^*) = \ell_{i1}(\boldsymbol{\theta}_0) - \ell_{i2}(\boldsymbol{\theta}_0) = 0$. Es decir, la diferencia de las verosimilitudes logarítmicas es evaluada en los $plims$ de los estimadores es idénticamente

cero bajo la nula. Esto hace que sea inútil para derivar un estadístico contraste porque la varianza η^2 sería idénticamente cero. Para los modelos anidados, no dividimos la diferencia en las verosimilitudes logarítmicas por $\sqrt{N}$ porque el resultado sería un estadístico que converge en probabilidad a cero.

El test VMS se aplica a modelos no anidados en los que la hipótesis nula es que ambos modelos están mal especificados pero ajustan igualmente bien (en el sentido de que tienen las mismas verosimilitudes logarítmicas esperadas). Es decir, si rechazamos el modelo 2 en favor del modelo 1 porque VMS es estadísticamente mayor que cero, entonces solo podemos concluir que el modelo 1 encaja mejor en el sentido de que $\mathbb{E}[\ell_{i1}(\boldsymbol{\theta}_1^*)] > \mathbb{E}[\ell_{i2}(\boldsymbol{\theta}_2^*)]$. No significa que el modelo 1 esté correctamente especificado (aunque podría estarlo). Hay muchos modelos que pueden encajar mejor que otro modelo dado, y claramente no todos pueden ser correctos.

7.6 Consecuencias de los tests

El uso de tests de especificación para elegir un modelo complica la distribución de un estimador. Por ejemplo, supongamos que elegimos entre dos estimadores $\hat{\theta}$ y $\tilde{\theta}$ sobre la base de un contraste estadístico al 5 %. Por ejemplo, $\hat{\theta}$ y $\tilde{\theta}$ pueden ser estimadores en modelos restringidos y sin restricciones. Entonces, el estimador real es

$$\theta^+ = w\hat{\theta} + (1 - w)\tilde{\theta}$$

donde la variable aleatoria w toma el valor 1 si la prueba favorece a $\hat{\theta}$ y 0 si la prueba favorece a $\tilde{\theta}$. Es decir, el estimador θ^+ depende de los estimadores restringidos y no restringidos y también depende de una variable aleatoria w, que a su vez depende del nivel de significatividad del contraste. Por tanto, θ^+ es un estimador con propiedades no triviales. A este estimador se le denomina estimador de prueba previa (pretest estimator), ya que el estimador se basa en una prueba inicial previa. La distribución de θ^+ no es estándar incluso bajo los supuestos de linealidad y normalidad. Dada esta complejidad, en la práctica, la inferencia se basa en la distribución de $\hat{\theta}$ si $w = 1$ o de $\tilde{\theta}$ si $w = 0$, ignorando por lo tanto la aleatoriedad en w. Si bien en teoría la inferencia debería basarse en θ^+.

Otra observación importante cuando usamos varios contrastes es que se pueden sacar diferentes conclusiones según el orden en que se realicen los tests. Una posible ordenación en los tests es la denominada modelización **de lo general a lo específico**. Por ejemplo, se puede estimar un modelo general para la demanda y posteriormente ser específico y probar las restricciones de la teoría de la demanda del consumidor, como la homogeneidad y la simetría. Alternativamente el proceso puede ser revertido, es decir, podemos usar la modelización **de lo específico a lo general**, añadiendo regresores progresivamente según sea necesario. Tales ordenaciones son naturales al elegir qué regresores incluir en un modelo, pero cuando también se realizan pruebas de especificación, no es raro usar ordenamientos tanto de **general a específico** como de **específico a general** en el mismo estudio.

Estos procedimientos, en función del tipo de modelización que llevemos a cabo, puede conducirnos a un uso extensivo de tests para seleccionar, este tipo de uso extensivo para un modelo se le denomina **minería de datos**. Por ejemplo, uno puede buscar entre varios cientos de posibles predictores de y y elegir sólo aquellos predictores que sean significativos al 5 % en una prueba de dos colas. Fácilmente podemos generar códigos que automatizan estas búsquedas. Desafortunadamente, búsquedas tan amplias conducirán al descubrimiento de relaciones falsas, ya que una prueba con un tamaño de 0.05 conduce a resultados erróneos de significatividad estadística el 5 % de las veces. Esta forma de proceder tiende a sobrestimar las medidas de bondad del ajuste (por ejemplo, R-cuadrado) y subestima las varianzas muestrales de los coeficientes de regresión, incluso cuando logra descubrir las variables que aparecen en el proceso generador de datos. Usar pruebas estándar y reportar p-valores sin tener en cuenta el procedimiento de selección del modelo es engañoso porque los p-valores nominales y los reales no son los mismo. Los métodos bootstrap pueden servir para calcular la verdadera significatividad estadística de los regresores.

La motivación para la minería de datos es a veces conservar grados de libertad o evitar la sobre-parametrización. Más importante aún como motivación para la minería de datos (data mining), es que muchos aspectos de la especificación, tales como la forma funcional de las covariables, no están resueltos por la teoría económica subyacente. Dada esta incertidumbre en la especificación, existe justificación para los mecanismos de búsqueda de especificación. Sin embargo, debe tenerse cuidado especialmente si se analizan muestras pequeñas y el número de búsquedas de especificación es grande en relación con el tamaño de la muestra. Cuando la búsqueda de especificación es secuencial, con un gran número de pasos, y con cada paso determinado por un resultado de prueba anterior, las propiedades estadísticas del procedimiento en su conjunto son complejas y analíticamente intratables, como hemos indicado anteriormente.

Las formas de reducir estos problemas de "pre-testing" esbozados anteriormente son varias. Una consiste en usar la teoría económica para conducir o guiar la selección de regresores, reduciendo en gran medida el número de regresores potenciales. Ahora abien si el tamaño de la muestra es grande, no sirve de nada eliminar variables "insignificantes". De hecho, los resultados finales suelen utilizar regresiones que incluyen regresores estadísticamente no significativos para las variables de control. La sobre-parametrización se puede evitar no informando de los coeficientes sin importancia en una especificación completa del modelo, pero haciendo constar este hecho en un lugar apropiado dentro del estudio. Es cierto, que esto puede llevar a una cierta pérdida de precisión al estimar los regresores clave de interés, pero protege contra el sesgo causado por la eliminación errónea de las variables que deberían incluirse.

Otra alternativa muy interesante, es usar solo parte de la muestra ("muestra de entrenamiento") para búsquedas de especificación y selección de modelo, y luego reportar los resultados usando el modelo seleccionado estimado usando una parte completamente separada de la muestra ("muestra de estimación"). En tales circunstancias, las pruebas preliminares (pre-test) no afectan a la distribución del estimador, si las submuestras son independientes. Este procedimiento generalmente solo se implementa cuando los tamaños de muestra son muy grandes, porque el uso de una muestra menor en la estimación final conduce a una pérdida en la precisión del estimador.

7.7 Diagnosis de modelos

La bondad de ajuste se interpreta como la cercanía de los valores ajustados a los valores muestrales de la variable dependiente. Para los modelos lineales con K regresores, la medida más directa es el **error estándar de la regresión**, que es la estimación de la desviación estándar del error

$$s = \left[(N-K)^{-1} \sum_{i}^{N} (y_i - \hat{y}_i)^2 \right]^{1/2}$$

Otra medida del ajuste se puede construir usando el error absoluto. El error absoluto medio será

$$\left[(N-K)^{-1} \sum_{i}^{N} |y_i - \hat{y}_i| \right]$$

Una medida relacionada en los modelos lineales es el R-cuadrado, el coeficiente de determinación múltiple, que hemos visto en temas precedentes y que explica la fracción de variación de la variable dependiente explicada por los regresores. El estadístico R-cuadrado suele aparecer más habitualmente que el s, aunque s puede ser más informativo para evaluar la bondad del ajuste.

Si el modelo de regresión no es lineal, es posible definir un pseudo-R-cuadrado. De hecho hay varias medidas pseudo-R-cuadrado posibles que en los modelos no lineales difieren y no necesariamente tienen las propiedades de estar entre cero y uno y aumentar a medida que se agregan regresores. En efecto, si partimos de la descomposición la suma cuadrática total (SCT) como siguie

$$\sum (y_i - \bar{y})^2 = \underbrace{\sum (y_i - \hat{y}_i)^2}_{SCR} + \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{SCE} + 2 \sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}),$$

se pueden establecer estas medidas

$$R^2_{RES} = 1 - \frac{SCR}{SCT},$$

$$R^2_{EXP} = \frac{SCE}{SCT}.$$

En general, estas medidas son diferentes, sin embargo en el modelo lineal estimado por MCO resulta que el tercer sumando de la descomposición anterior es nulo, y por tanto en dicho caso sucede que $R^2_{RES} = R^2_{EXP}$. Como estas medidas tienen sentido para marcos no-lineales, destacamos que en tal situación R^2_{RES} podría ser menor a cero, R^2_{EXP} podría ser superior a la unidad, y ambos podrían disminuira a medida que incluimos regresores.

Una medida relacionada es el cuadrado de la correlación entre los valores observados y ajustados

$$R^2_{CORR} = \widehat{Corr^2}[y_i, \hat{y}_i]$$

que tiene la característica de estar acotado entre 0 y 1, y que es igual al R-cuadrado convencional en un modelo lineal con intercepto estimado por MCO. Al igual que los anteriores es susceptible de disminuir al incorporar más regresores al modelo.

Cuando la heterocedasticidad intrínseca del modelo quiere ser explícitamente tenida en consideración a la hora de valorar la bondad del ajuste, entonces es posible usar el siguiente pseudo-R-cuadrado ponderado

$$R^2_W = 1 - \frac{SCR_W}{SCT_W}$$

donde $SCR_W = \sum (y_i - \hat{y}_i)^2 / \hat{\sigma}_i^2$, $SCT_W = \sum (y_i - \hat{\mu})^2 / \hat{\sigma}^2$; por otra parte $\hat{\sigma}_i^2$ es la varianza condicional de y_i, $\hat{\sigma}^2, \hat{\mu}$ son la varianza y la media estimadas del modelo simple que solo tiene como regresor un intercepto (término constante).

Por último cabe destacar una generalización de las medida tipo $R-$cuadrado para medidas objetivo diseñadas por el usuario. Sea $Q_N(\theta)$ la función objetivo a maximizar, Q_0 el valor de dicha función en el modelo con únicamente un término constante, Q_{fit} el valor de la función en el modelo ajustado, y Q_{max} el valor máximo posible alcanzable por $Q_N(\theta)$. Así pues la potencial ganancia sobre la función objetivo resultante de incluir regresores será $Q_{max} - Q_0$, y por tanto una medida sería la ganancia relativa (RG), es decir,

$$R^2_{RG} = \frac{Q_{fit} - Q_0}{Q_{max} - Q_0} = 1 - \frac{Q_{max} - Q_{fit}}{Q_{max} - Q_0}$$

Esta medida tiene las ventajas de que estár acotada entre 0 y 1, y que aumenta a medida que aumentamos los regresores incluidos. Conviene que observar que en el caso MCO, la función a maximizar es el valor negativo de la suma cuadrática de los residuos, de manera que $Q_0 = -SCT, Q_{fit} = -SCR, Q_{max} = 0$, y por tanto $R^2_{RG} = SCE/SCT$. Todas la pseudo-medidas presentadas pueden ser ajustadas por los grados de libertad.

Para la diagnosis del modelo suele ser común analizar los residuos. Una forma habitual es mediante gráficos de los residuos por sí mismos, o frente a otras variables de interés. Así, los gráficos de los residuos frente a los valores ajustados pueden revelar un ajuste deficiente del modelo; los gráficos de los residuos frente a las variables omitidas pueden sugerir que se incluyan más regresores en el modelo; y los gráficos de residuos contra regresores incluidos pueden sugerir la necesidad de una forma funcional diferente. También puede ser útil aplicar alguna técnica de representación no paramétrica.

Algunos modelos paramétricos implican que un residuo adecuadamente definido debe distribuirse normalmente. Esto puede comprobarse mediante un gráfico de que ordena los residuos e_i de menor a mayor y los contrapone frente a los valores pronosticados si los residuos estuvieran exactamente distribuidos normalmente. Por tanto se dibuja e_i frente a $\bar{e} + \hat{\sigma}_e \Phi^{-1}((i - 0{,}5)/N)$, siendo $\bar{e}$ la media muestral de los residuos, $\hat{\sigma}_e$ su desviación estándar y $\Phi^{-1}(\cdot)$ la inversa de la función de distribución normal.

En otras partes del temario se desarrollan aspectos complementarios del análisis de residuos. El lector puede encontrar interesante completar el apartado con el análisis de residuos propio de las series temporales.

Bibliografía complementaria

Matilla-García, M et al. 2017. Econometría y Predicción. McGraw Hill

Stock J. and Watson J. Introducción a la econometría. Pearson.

Tema 8

Endogeneidad y estimación con variables instrumentales

Este tema está elaborado como una adaptación de los capítulo 15 y 16 de:

Wooldridge. J. 4th Ed., Introductory Econometrics,

y del capítulo 4 (secciones 4.6, 4.7, 4.8 y 4.9)

Cameron and Trivedi. Microeconometrics: methods and applications. Así como de la bibliografía complementaria

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al Órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

- Fuentes de endogeneidad.
- Variables instrumentales.
- Estimación con variables instrumentales.
- Estimación de mínimos cuadráticos bietápicos.
- Contrastes de endogeneidad.
- Modelos de ecuaciones simultáneas

8.1 Fuentes de endogeneidad

El modelo de regresión lineal presentado con anterioridad es hasta cierto punto bastante general, sin embargo aunque la aproximación lineal sea factible para las variables consideradas a analizar, la realidad de las relaciones económicas y de los datos económicos nos conducen fácilmente a situaciones en las que alguno(s) de los supuestos que caracterizan al modelo de regresión lineal no son satisfechas, y por tanto la validez del modelo es limitada.

Hay varias situaciones muy recurrentes en el contexto de los datos observacionales de sección cruzada que claramente invalidan la estimación por MCO. Destacamos las siguientes fuentes de endogeneidad: el *sesgo por omisión de variables*, los *errores en las variables*, y la *causalidad simultánea* Estas tres situaciones comparten el hecho de que

$$\mathbb{E}\left(\varepsilon_i|\,\mathbf{X}\right) \neq 0,\ i = 1, 2, ..., n,$$

es decir, son casos en los que se viola el *supuesto de exogeneidad*, y por tanto decimos que el modelo tiene o sufre problemas de **endogeneidad**. La violación de este supuesto

se genera directamente porque existe correlación entre alguna(s) de la(s) variable(s) explicativa(s), X, y el término error, ε,

$$\mathbb{E}\left(\varepsilon_i \mathbf{x}_i\right) \neq \mathbf{0},\ i = 1, 2, ..., n.$$

Cuando el supuesto de exogeneidad no se cumple, es decir cuando algunas de las variables son **endógenas**, el estimador MCO es incosistente en el sentido de que el sesgo no desaparece al crecer el tamaño de la muestra[1]. Sucede que las correlaciones entre las variables observables y los errores (donde se incluyen al conjunto de variables determinantes o influyentes pero no tenidas en consideración en el modelo) contaminan persistentemente a nuestros estimadores, haciendo que sea prácticamente imposible obtener información «limpia» de los coeficientes poblacionales $\beta_j, j = 1, 2, ..., k$.

Por tanto, este es un problema grave en tanto que invalida el método y las conclusiones que pudieran derivarse de estudios con variables endógenas. Se hace necesario contar con métodos de estimación alternativos que permitan controlar la endogeneidad. En este tema veremos que uno de ellos consiste en la utilización de variables intrumentales, pero antes veremos en cierto detalle las fuentes más comunes de endogeneidad.

Sesgo por omisión de variables relevantes

Supongamos que partimos de la siguiente especificación general, que consideramos correcta,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i. \tag{8.1}$$

sin embargo estimamos

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_{k-1} X_{(k-1)i} + e_i.$$

Por lo tanto hemos omitido la variable X_k y formará parte del término error nuevo. El principal inconveniente que puede aparecer al omitir una variable relevante es el denominado *sesgo de variable omitida*. Para que el problema sea tal no solo consiste en omitir una variable determinante de la variable dependiente, sino que es necesario que la variable omitida cumpla otra condición en relación al resto de variables especificadas en el modelo. En concreto, el sesgo de variable omitida se produce cuando se satisfacen dos condiciones:

1. La variable omitida está correlacionada con los regresores incluidos en la regresión, X_k.
2. La variable omitida es un factor determinante de la variable dependiente, Y.

[1] Esto parcialmente justifica que históricamente esta inconsistencia sea etiquetada como *sesgo de estimación* o *sesgo de endogeneidad*, cuando realmente se trata de inconsistencia.

Este "sesgo de variable omitida" significa que el supuesto de exogeneidad no se cumple, es decir, $\mathbb{E}(\varepsilon_i \,|X_i\,) \neq 0$. Consideremos el modelo de regresión simple, en el que el término error ε_i representa todos los factores, distintos de X_i, que son determinantes de Y_i. Si uno de esos factores está correlacionado con X_i, esto significa necesariamente que el término error (que contiene este factor) está correlacionado con X_i. Debido a que entonces X_i y ε_i están correlacionados, la media condicionada de ε_i dado X_i ya no será constante, y por lo tanto el supuesto central de exogeneidad no se satisface.

la consecuencia de que el supuesto de exogeneidad no se cumpla. Para ello consideremos formalmente el estimador MCO del coeficiente de la variable explicativa

$$\hat{\beta}_1 = \beta_1 + \frac{(1/n)\sum(X_i - \bar{X})\varepsilon_i}{(1/n)\sum(X_i - \bar{X})^2}.$$

Bajo el supuesto de muestra aleatoria y el supuesto sobre atípicos, el numerador y el denominador del segundo sumando de la expresión anterior pueden reemplazarse por sus contrapartidas poblacionales, $\text{cov}(\varepsilon_i, X_i) = \rho_{X\varepsilon}\sigma_\varepsilon\sigma_X$ y σ_X^2, respectivamente, donde el término $\rho_{X\varepsilon} = \text{corr}(X, \varepsilon)$. Si sustituimos estas expresiones obtendremos

$$\hat{\beta}_1 = \beta_1 + \frac{(1/n)\sum(X_i - \bar{X})\varepsilon_i}{(1/n)\sum(X_i - \bar{X})^2} \xrightarrow{p} \beta_1 + \rho_{X\varepsilon}\frac{\sigma_\varepsilon\sigma_X}{\sigma_X^2}. \tag{8.2}$$

El sesgo precisamente se produce porque al estar correlacionado el error con la variable explicativa entonces el término $\rho_{X\varepsilon}$ es distinto de cero, lo que hace que el estimador $\hat{\beta}_1$ no converja en probabilidad al verdadero valor β_1, incluso si el tamaño muestral es grande, por lo que también tendremos que el estimador no será consistente. El sesgo será grande o pequeño en función de la correlación $\rho_{X\varepsilon}$: cuanto mayor sea en términos absolutos, mayor será el sesgo. La dirección del denomiando sesgo depende de si X y ε están positiva o negativamente correlacionadas. El estimador $\hat{\beta}_1$ en un modelo que omite una variable relevante, X_2, no recoge el efecto parcial sobre Y de un cambio en X_1, pues al correlacionar con X_2 cuando varía X_1 también lo hace X_2. De hecho lo que captura es el efecto directo sobre Y de un cambio en X_1, más el efecto indirecto de X_1 sobreX_2, que termina afectando a Y.

Error de medida en la variable explicativa

Supongamos que el modelo es de la siguiente forma

$$Y_i = \beta_0 + \beta_1 X_{1i}^* + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i, \tag{8.3}$$

cuyo error de medida es

$$w_1 = X_1 - X_1^*. \tag{8.4}$$

Sustituyendo (8.4) en (8.3) tenemos

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + (\varepsilon_i - \beta_1 w_{1i}) = \beta_0 + \beta_1 X_{1i} + ... + \underbrace{[\varepsilon_i - \beta_1 (X_{1i} - X_{1i}^*)]}_{v_i}. \quad (8.5)$$

Las propiedades de los estimadores del modelo (8.5) dependen de cómo consideremos o caractericemos los errores de medida, es decir, dependen de qué supuestos hagamos sobre error de medida de la expresión (8.4). La expresión (8.5) sugiere que los sesgos dependerán de la correlación entre el error v_i, que incluye el error de medición, y el regresor X_{1i}. De modo que si w_i estuviera correlacionado con X_{1i}, también lo estaría v_i y habría sesgo e incosistencia en $\hat{\beta}_1$.

Es posible que el marco en el que se obtienen los datos nos lleve a suponer que el error de medida no está correlacionado con la variable observable, digamos «X_1», es decir que

$$\text{corr}\,(X_1, w_1) = 0. \quad (8.6)$$

Este marco puede producirse, por ejemplo, cuando los datos provienen de una encuesta en la que nos parece razonable considerar que el encuestado hace su mejor aproximación, dada toda su información, acerca del verdadero valor de la variable sobre la que es preguntado. El error de aproximación (esto es, el error de medida) no está entonces correlacionado con la respuesta de cada individuo, si ha utilizado toda su información.

Por los supuestos del modelo de regresión lineal sabemos que «X_1» y «ε_i» están incorrelacionados en (8.5); además, por el supuesto (8.6), «X_1» y «w_{1i}» también están incorrelacionados. Por tanto «$\varepsilon_i - \beta_1 w_{1i}$» de la expresión (8.5) tiene media cero y está incorrelacionado con «X_1». En definitiva si se cumple el supuesto (8.6), el modelo con errores de medida (8.5) tendrá estimadores consistentes. En tal caso, como hemos visto tendremos estimadores consistentes, pero no olvidemos que la varianza del coeficiente $\hat{\beta}_1$ será mayor que en el caso de ausencia del error, puesto que var $(\varepsilon_i - \beta_1 w_{1i}) >$ var (ε_i).

Alternativamente el marco en el que se recolectan los datos nos puede hacer pensar que el error de medición es puramente aleatorio, lo que se denomina *modelo clásico de error de medición*. En tal caso podríamos considerar que el error de medida está incorrelacionado con la variable no observable:

$$\text{corr}\,(X_1^*, w_1) = 0, \quad (8.7)$$

y los errores de medida son de la forma

$$X_1 = X_1^* + w_1, \quad (8.8)$$

donde el componente aleatorio w_1 es tal que tiene media cero y varianza constante, y además corr $(w_i, \varepsilon_i) = 0$. Teniendo en cuenta la expresión (8.8), entonces la variable observada «X_1» y el error de medida «w_1» estarán correlacionados:

$$\text{cov}\left(X_1, w_1\right) = \text{E}\left(X_1 w_1\right) = \text{E}\left(X_1^* w_1\right) + \text{E}\left(w_1^2\right) = \sigma_w^2. \tag{8.9}$$

El error de medida «ε_i» y la variable observable «X_1» están correlacionados en la expresión (8.5), lo que incumple el supuesto de esperanza condicionada nula, y por tanto los estimadores de (8.5) son sesgados e inconsistentes.

El caso del modelo de regresión simple utilizado en esta sección nos permite comprobar la expresión del sesgo para el modelo clásico de error de medición. Si desarrollamos el límite en probabilidad del estimador MCO se tiene

$$\begin{aligned} plim\hat{\beta}_1 &= \frac{plim\left(\sum_i x_i y_i\right)^{-n}}{plim\left(\sum_i x_i^2\right)^{-n}} = \frac{cov(X,Y)}{var(X)} \\ &= \frac{cov(X^* + w, Y^*)}{var(X^* + w)} = \frac{cov(X^*, Y) + cov(Y, w)}{var(X^*) + var(w)} \end{aligned}$$

si multiplicamos por $var(X*)/var(X^*)$, al ser la unidad, tendremos dado que $cov(Y, w) = 0$

$$plim\hat{\beta}_1 = \frac{cov(X^*, Y)/var(X^*)}{\left[var(X^*) + var(w)\right]/var(X^*)} = \frac{\beta_1}{1 + var(w)/var(X^*)} = \underbrace{\frac{var(X^*)}{var(X^*) + var(w)}}_{\lambda}\beta_1.$$

El término λ es un ratio de varianzas. En el numerador la varianza de la señal y en el denominador la varianza total (de la señal y del error de medida de la misma). Por tanto este ratio está entre cero y la unidad. También esta última expresión nos permite ver (operando mínimamente en el último igual) cuál es el sesgo (asintótico) de $\hat{\beta}_1$:

$$plim(\hat{\beta}_1 - \beta_1) = \lambda\beta_1 - \beta_1 = -(1-\lambda)\beta_1 = -\left(\frac{var(w)}{var(X^*) + var(w)}\right)\beta_1.$$

Y por tanto: (i) el estimador $\hat{\beta}_1$ estará sesgado hacia cero; es decir, si $\beta_1 > 0$, entonces $\left(\hat{\beta}_1 - \beta_1\right) < 0$, mientras que si $\beta_1 < 0$, entonces $\left(\hat{\beta}_1 - \beta_1\right) > 0$; por lo que podemos decir que $\hat{\beta}_1$ infraestima el coeficiente poblacional de la variable medida con error; (ii) la inconsistencia puede ser despreciable si la variabilidad del error de medida en relación a la variablidad de la variable explicativa original, es decir, si $var(X^*)$ es alta en relación a la $var(w)$.

A modo de resumen hemos comprobado que dependiendo del supuesto que hagamos, (8.6) u (8.7), los estimadores de los modelos con errores de medida serán consistentes o inconsistentes, y esta inconsistencia podría resultar insignificante, pero no nula. Resulta difícil determinar en la práctica cuál de los dos supuestos es más realista.

Un elemento a considerar es el sesgo que el error de medida puede inducir sobre otras variables del modelo distintas de la medida con error. Para comprobarlo consideremos

el efecto sobre la variable constante del modelo de regresión simple. Calculamos el límite en probabilidad para el mismo

$$\begin{aligned} plim\left(\hat{\beta}_0\right) &= plim\left(\bar{Y} - \hat{\beta}_1 X^*\right) = \mathbb{E}(Y) - plim\left(\hat{\beta}_1 X^*\right) \\ &= \mathbb{E}(Y) - plim\hat{\beta}_1 \mathbb{E}(X + w) \\ &= \mathbb{E}(Y) - \lambda\beta_1 \mathbb{E}(X), \end{aligned}$$

que no es consistente, es decir no colapsa asintóticamente con β_0, incluso si $\mathbb{E}(w) = 0$.

$$\beta_0 = \mathbb{E}(Y) - \hat{\beta}_1 \mathbb{E}(X).$$

Obsérvese que el error w_i no está correlacionado con el valor de la variable medida X_i, ni lógicamente tampoco la constante del modelo, y sin embargo se genera inconsistencia en su coeficiente asociado. Pues bien, esto sucede también cuando consideramos un modelo de regresión múltiple. En general (salvo casos muy particulares, casi solo contemplados en la teoría), el error de medida en una variable produce inconsistencia en todos los coeficientes estimados $\hat{\beta}_j$.

Causalidad simultánea

Cuando tenemos un modelo, suponemos que las variables explicativas, X, son las que «causan» o generan cambios en la variable Y. La causalidad simultánea se produce cuando la variable Y genera o «causa» cambios en alguna(s) de las variables X. Una regresión estimada por MCO capturará ambos efectos, por lo que el estimador MCO será necesariamente inconsistente.

Para verlo con mayor detenimiento vamos a comprobar cómo la causalidad simultánea induce a la existencia de correlación entre el regresor X y el término error en la regresión poblacional de interés. Por comodidad y a efectos ilustrativos consideremos que existen solo dos variables X e Y, y que existen dos ecuaciones que indican las relaciones de causalidad entre ambas:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{8.10}$$

$$X_i = \gamma_0 + \gamma_1 Y_i + v_i. \tag{8.11}$$

La Ecuación (8.10) es la ecuación poblacional habitual en la que el coeficiente β_1 es el efecto sobre Y de una variación en X, y donde el término ε representa otros factores influyentes distintos de X. La Ecuación (8.11) representa el efecto causal inverso de Y sobre X.

En esta situación es fácil ver que la causalidad simultánea conduce a la correlación entre X_i y ε_i. Para verlo, imaginemos, por ejemplo, que ε_i es negativo en la Ecuación (8.10), lo que hace disminuir Y_i. Este menor valor de Y_i afecta al valor de X_i a través de la Ecuación (8.11), de modo que si el coeficiente γ_1 es, por ejemplo, positivo, un valor bajo de Y_i conducirá a un valor bajo en X_i, y en tal caso, existirá correlación (positiva, en la misma dirección) entre X_i y ε_i.

Matemáticamente también es fácil comprobar la existencia de correlación entre X_i y ε_i. La Ecuación (8.11) implica

$$\begin{aligned}\text{cov}(X_i, \varepsilon_i) &= \text{cov}(\gamma_0 + \gamma_1 Y_i + v_i, \varepsilon_i) \\ &= \gamma_1 \text{cov}(Y_i, \varepsilon_i) + \text{cov}(v_i, \varepsilon_i) \\ &= \gamma_1 \text{cov}(Y_i, \varepsilon_i) \\ &= \gamma_1 \text{cov}(\beta_0 + \beta_1 X_i + \varepsilon_i, \varepsilon_i) \\ &= \gamma_1 \beta_1 \text{cov}(X_i, \varepsilon_i) + \gamma_1 \sigma_\varepsilon^2.\end{aligned}$$

Si despejamos $\text{cov}(X_i, \varepsilon_i)$, obtenemos

$$\text{cov}(X_i, \varepsilon_i) = \frac{\gamma_1 \sigma_\varepsilon^2}{1 - \gamma_1 \beta_1}.$$

Este modelo sencillo nos permite entonces comprender la naturaleza del sesgo inducido por la existencia de un proceso con causalidad en doble dirección entre dos variables, cuando la que nos importa es una de ellas.

Afortunadamente, hay una alternativa a la técnica MCO que se comporta mejor, al menos para muestras grandes. Esta alternativa que vamos a exponer en este tema aprovecha el hecho de que, incluso cuando $\mathbb{E}(\varepsilon_i \mathbf{x}_i) \neq 0$, es posible (a menudo) utilizar el propio análisis económico (esto es, la teoría económica que subyace a la relación de las variables económicas) para localizar otras variables que no estén correlacionadas con el término error ε_i. Estas variables que hemos detectado por el razonamiento económico pueden ser consideradas (cuando se cumplen ciertas condiciones) como un instrumento que nos facilite estimar $\beta_j, j = 1, 2, ..., k$, y por este motivo se denominan ***variables instrumentales*** **(VI)**.

8.2 Variables instrumentales (VI).

La inconsistencia de MCO causada por cualquiera de las fuentes de endogeneidad anteriores implica que los cambios en la variable explicativa X están asociados no solo con cambiós en la variable dependiente Y, sino también a cambios en las variables que hay en el término error. Para paliar este problema necesitamos de un método que genere variación solo variación exógena en la X. Idealmente podríamos conseguir dicha fuente de varación de un expermiento aleatorizado, pero esto raramente es factible.

Una forma alternativa de conseguir esta variación exógena es a través de una variable que hará de instrumento, y que denominaremos **variable instrumental** o **instrumento**. Intuitivamente, podemos comprender cómo funciona la técnica de VI si consideramos que la variable dependiente, X, consta de dos partes. Una parte, que por algún motivo, está correlacionada con ε, y por tanto es la parte que genera disfunciones por endogeneidad; y otra parte que no está correlacionada con ε. Si fuera posible obtener información que permitiera aislar la primera parte de X, podríamos estudiar solo las variaciones de X que no están correlacionadas con el error, y obviar las variaciones de X que sesgan la estimación MCO. Esta información procedería de otra u otras variables que llamamos variables instrumentales.

Así pues una variable, Z, que es instrumento tiene la interesante propiedad de que cambios en Z están asociados con cambios en X, pero no lo están a cambios en Y (a parte de la vía de relación **indirecta** que se produce a través de X). La variable dependiente Y y el instrumento Z están correlacionadas meramente por la relación indirecta a través de X, que hemos comentado. En todo caso, Z no será una variable explicativa para la modelización de la variable Y.

Imaginemos que queremos estimar la respuesta de la demanda de mercado de un producto agrícola a cambios exógenos en el precio de mercado. La cantidad demandada depende claramente del precio, pero los precios no son exógenos puesto que están determinados en parte por la demanda del mercado. Un instrumento adecuado para la variable $X = precio$ sería una variable que correlacionara con $precio$ pero que no afectara directamente a la cantidad demandada. Un candidato obvio es una variable que afecte a la oferta del mercado, ya que esta también afectaría a los precios, pero que simultáneamente no fuera un determinante directo de la demanda. Un ejemplo de variable instrumental válida para este tipo de mercado agrícula sería un variable que capturara las condiciones favorables para el crecimiento del bien agrícuola en cuestión. En efecto, las condiciones de crecimiento favorables, no afectan directamente a la demanda, y por tanto esta variación externa sirve como potencial instrumento de la variable $precio$.

Una vez identificado cómo debería de ser un instrumento, el siguiente paso es estimar los parámetros del modelo.

8.3 Estimación con variables instrumentales.

Vamos a explicar esta técnica de estimación a partir de un ejemplo que ilustra bien este tipo de estimación y que se utiliza con frecuencia en la literatura de variables instrumentales. Consideremos un modelo de economía laboral que explica el salario por hora de los trabajadores. Supongamos que el modelo *bien especificado* es el siguiente:

$$\ln\left(salario\right)_i = \beta_0 + \beta_1 \cdot estudios_i + \beta_2 \cdot habilidad_i + \varepsilon_i, \tag{8.12}$$

donde el término error representa los factores omitidos que determinan la variable dependiente, en este caso, ln(salario). Indudablemente la habilidad o capacidad intrínseca o natural del trabajador debe influir en el salario, al igual que lo hace el nivel de estudios alcanzado (nivel de formación) por el trabajador. Por lo general, el nivel de estudios alcanzado por el individuo y su capacidad o habilidad están también correlacionados: los niveles de formación suelen ser más altos para aquellos con mayores habilidades o capacidades. La variable «*habilidad*» es difícil de definir, y en términos prácticos, muy difícil de medir, de manera que nos encontramos con una variable importante que es inobservable.

Si consideramos la posibilidad de estimar el modelo sin la variable inobservable «$habilidad$», es decir, estimaríamos el modelo de regresión simple

$$\ln\left(salario\right)_i = \beta_0 + \beta_1 estudios_i + v_i, \tag{8.13}$$

en el que la variable «$habilidad$» pasa a formar parte de los errores «v», que ahora necesariamente son distintos de los errores de la función de esperanza condicionada (FEC) (8.12), ε.

Por otra parte, dado el modelo de la FEC, sabemos que

$$\text{cov}\left(estudios_i, v_i\right) \neq 0, \tag{8.14}$$

es decir, la variable que habitualmente denotamos por X está correlacionada con el término error «v» de la ecuación estimada. Cuando sucede esto decimos que la **variable explicativa es endógena**. Solo cuando X no está correlacionada con los errores poblacionales, decimos que la **variable explicativa es exógena**.

Cuando la ecuación *mal especificada* por omisión de variable relevante, como en la Ecuación (8.13), sabemos que la variable explicativa (estudios) será endógena al estar correlacionada con el error v, ya que hemos considerado que el modelo poblacional es (8.12), y por tanto el error de la i-ésima observación incorpora la variable $habilidad_i$. Por este motivo, MCO generará estimaciones no consistentes de los coeficientes de interés en el modelo (8.13).

Para estimar consistentemente «β_0» y «β_1» en estas condiciones, tenemos que utilizar información externa a la proporcionada en el modelo (8.13). Más concretamente debemos encontrar una variable «Z» (variable instrumental o instrumento) que satisfaga dos condiciones necesarias para que el «instrumento» tenga el efecto deseado de permitirnos estimar consistentemente los coeficientes de interés, es decir, dos condiciones para que el instrumento sea válido. A estas condiciones se las conoce como ***condición de exogeneidad del instrumento*** y la ***condición de relevancia***

$$\text{cov}\left(Z, v\right) = 0, \tag{8.15}$$

y

$$\text{cov}\left(Z, X\right) \neq 0, \tag{8.16}$$

es decir, tenemos que encontrar una variable «Z» que no covaríe con los errores «v» de la expresión (8.13) y covaríe con la variable explicativa endógena «X». Bajo estos supuestos decimos que «Z» es una variable instrumental de «X». También podemos decir que el instrumento «Z» es exógeno en el modelo (8.13) y está correlacionado con la variable explicativa endógena «X». Si un instrumento es relevante, la variación en el instrumento está relacionada con la variación en X_i. Si, además, es exógeno, entonces la parte de variación de X_i capturada por el instrumento Z_i es exógena. Por tanto, un instrumento relevante capta los movimientos de X_i que son exógenos. Esta variación ahora exógena puede ser utilizada para estimar «sin contaminación» el coeficiente β_1. El requisito de exogeneidad del instrumento excluye la posiblilidad de que Z sea un regresor en el modelo Y, dado que si por el contrario Y dependiera tanto de X como de

Z, e Y fuera regresada solo sobre X, entonces Z quedaría absorvida dentro del error de tal manera que entonces Z estaría correlacionda con el error.

Con la información adicional que proporciona el instrumento Z, si partimos de un modelo de regresión simple tal que

$$Y = \beta_0 + \beta_1 X + v; cov(X, v) \neq 0; Z \text{ es un instrumento válido}$$

se tiene que

$$cov(Z, Y) = \beta_1 cov(Z, X) + cov(Z, v)$$

la condición de exogeneidad implica que los parámetros poblacionales son

$$\beta_1 = \frac{cov(Z, Y)}{cov(Z, X)}$$

$$\beta_0 = \mathbb{E}(Y) - \beta_1 \mathbb{E}(X) = \mathbb{E}(Y) - \frac{cov(Z, Y)}{cov(Z, X)} \mathbb{E}(X).$$

Reemplazando estos momentos poblacionales por sus expresiones muestrales, obtenemos el estimador de variables instrumentales

$$\hat{\beta}_1^{VI} = \frac{\widehat{cov(Z, Y)}}{\widehat{cov(Z, X)}} = \frac{\sum z_i y_i}{\sum z_i x_i}$$

donde las variables están en desviaciones respecto a sus medias; y

$$\hat{\beta}_0^{VI} = \bar{Y} - \hat{\beta}_1^{VI} \bar{X}.$$

Ahora este estimador será consistente al usar el instrumento

$$\begin{aligned} \hat{\beta}_1^{VI} &= \frac{(1/n)\sum z_i y_i}{(1/n)\sum z_i x_i} = \frac{(1/n)\sum z_i(\beta_1 x_i + v_i)}{(1/n)\sum z_i x_i} = \beta_1 + \frac{(1/n)\sum z_i v_i}{(1/n)\sum z_i x_i} \\ \hat{\beta}_1^{VI} &\xrightarrow{p} \beta_1 + \frac{cov(z, v)}{cov(z, x)} = \beta_1 + \frac{\rho_{Zv}}{\rho_{ZX}} \frac{\sigma_v}{\sigma_X} = 0 \end{aligned}$$

puesto que el instrumento garantiza $cov(z, v) = 0$ y que por tanto el estimador de VI es consistente. Resulta ilustrativo observar que el estimador de VI coincide con el estimador de MCO en caso de que la variable X tenga las propieadades de instrumento de sí misma, $Z = X$. En tal caso la condición de relevancia será inmediatamente satisfecha puesto que la correlación entre X y Z será perfecta. Si la condición de exogeneidad se cumpliera, es decir, si $cov(Z = X, v) = 0$, entonces sustituyendo en la expresión de $\hat{\beta}_1^{VI}$ se comprueba inmediatamente que $\hat{\beta}_1^{VI} = \hat{\beta}_1^{MCO}$ y lo mismo sucerá para el estimador del término independiente.

La expresión matricial del estimador es

$$\hat{\boldsymbol{\beta}}_{VI} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y}$$

dada una matriz $\mathbf{Z}$ de dimensiones $n \times (k+1)$ donde hay el mismo número de instrumentos que variables explicativas. Naturalmente la matriz de instrumentos incluye el término independiente, que hace de instrumento de si mismo.

El estimador de la varianza de $\hat{\boldsymbol{\beta}}_{VI}$ se obtiene siguiendo exactamente los mismos pasos que para obtener el de la varianza del estimador de MCO :

$$\widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}_{\text{VI}}\right) = \left(\mathbf{Z}'\mathbf{X}\right)^{-1} \hat{\boldsymbol{\Omega}} \left(\mathbf{Z}'\mathbf{X}\right)^{-1}, \text{ donde } \hat{\boldsymbol{\Omega}} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \hat{\varepsilon}_i^2.$$

Por tanto, con errores heterocedásticos el estimador de VI es asintóticamente normal, con media $\boldsymbol{\beta}$, y la varianza indicada anteriormente.

Esta varianza será mayor que la de estimador de MCO. Sin embargo no tiene mucho sentido la comparación entre una y otra puesto que el estimador MCO no es consistente si hay endogeneidad. Lo que sí es interesante es que la varianza tenderá a mayor cuanto menor sea la correlación entre instrumentos y regresores.

Por tanto, con errores heterocedásticos el estimador de VI es asintóticamente normal, con media $\boldsymbol{\beta}$, y la varianza indicada anteriormente.

Para ilustrar el problema y la solución que ofrece el estimador de VI hemos generado mediante simulación el siguiente modelo

$$\begin{cases} Y = \beta_0 + \beta_1 X + U; & \beta_0 = 0, \beta_1 = 0{,}5 \\ X = \alpha_0 + \alpha_1 Z + V; & \alpha_0 = 0, \alpha_1 = 1 \\ Z \sim N(2,1); & \\ U \sim N(0,1), V \sim N(0,1) & corr(U,V) = 0{,}8 \end{cases}$$

La variable explicativa X está por construcción correlacionada con V y V a su vez está correlacionada con U, por tanto $corr(X,U) \neq 0$ y podemos decir que la variable X es endógena. El estimador MCO de β_0, β_1 será por tanto inconsistente. El modelo dispone de una variable Z que está correlacionada por construcción con X (mediante el parámetro α_1). Podemos decir entonces que cumple el criterio de relevancia. Además, Z es una variable que se distribuye como una normal que es independiente de U, por lo que la variable Z es exógena respecto de U, $corr(Z,U) = 0$. Así pues cumple el requisito de exogeneidad. En resumen, podemos afirmar que Z es un instrumento de X.

En esta simulación hemos generado dos muestras, una de tamaño 100 y otra de tamaño $10{,}000$. Para cada muestra hemos estimado los parámetros de interés β_0, β_1 tanto por MCO como por VI. La siguiente tabla muestra los resultados.

Observamos varias cuestiones relevantes. El estimador MCO de la pendiente es inconsistente: para un tamaño muestral de 100 es de 0,91 (cuando debería estar próximo a 0,5), y este problema no desaparece cuando aumentamos el tamaño muestra mil veces (para 10.000 observaciones el estimador MCO sigue anclado en el 0,9). Sin embargo, al utilizar un instrumento y estimar por VI los resultados cambian drásticamente. En efecto, el estimador de VI de la pendiente del modelo arroja un valor muy próximo a 0,5, independientemente del tamaño muestral. Algo similar sucede con el estimador del término independiente.

Tabla 8.1: Estimación MCO y VI

	$n = 100$		$n = 10{,}000$	
	MCO	VI	MCO	VI
Constante	-0.98	-0.16	-0.81	-0.01
	(0.1151)	(0.1912)	(0.013)	(0.020)
X	0.91	0.49	0.90	0.51
	(0.0500)	(0.0912)	(0.0061)	(0.0101)

El ejemplo de simulación anterior ilustra claramente la conveniencia de usar VI. Dicha utilidad naturalmente está supeditada a la existencia de suficientes instrumentos válidos. Si los instrumentos no son válidos, las conclusiones carecerán de sentido. Es por tanto importante saber evaluar cuándo los instrumentos son o no válidos. Sobre este aspecto volveremos más adelante en el tema, por el momento consideraremos que los instrumentos son válidos.

En el caso del salario, expresión (8.12), los expertos en el mercado laboral han utilizado como variable instrumental el «nivel educativo de la madre». Este instrumento cumple claramente la condición de relevancia al estar correlacionada con la variable «$estudios$» del hijo, y puede resultar que también cumpla la condición de exogeneidad del instrumento, si consideramos que la habilidad del hijo no está correlacionada con el nivel de estudios alcanzado por la progenitora, cuestión que ha sido discutida en la literatura especializada. Un instrumento menos controvertido que se ha utilizado es la proximidad a la universidad o centro de formación superior. A priori, cuanto más alejado se esté del centro, menor es la probabilidad de recibir educación superior. Y por otra parte, es bastante factible que no haya relación directa entre la variable habilidad y la proximidad geográfica a la universidad.

8.4 Estimación de mínimos cuadráticos bietápicos.

Introducción

Consideremos que disponemos de un instrumento que cumple las condiciones (8.15) y (8.16) y que necesitamos que instrumentalice a la variable explicativa endógena X, entonces podemos estimar consistentemente la ecuación

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, cov(X, \varepsilon) \neq 0 \tag{8.17}$$

mediante un estimador de VI denominado *mínimos cuadrados en dos etapas* (MC2E), aunque en el modelo poblacional exista correlación entre X_i y ε_i. El estimador consta de dos fases.

La primera etapa consiste en una regresión poblacional que relaciona a X con Z

$$X_i = \pi_0 + \pi_1 Z_i + u_i,$$

donde los parámetros π_0, π_1 son el intercepto y la pendiente, respectivamente, y donde u_i es el término error de esta regresión auxiliar. Esta regresión define las dos partes que necesitamos. A partir de las condiciones de validez del instrumento, la parte no problemática de X_i es $\pi_0 + \pi_1 Z_i$, que es la parte de X_i que captura o explica Z_i. Dado que Z_i es exógena, esta componente está incorrelacionada con el término error de (8.17), ε_i. La otra parte restante, es decir, u_i será la parte de X_i problemática por estar correlacionada con ε_i. Los MC2E utilizan la parte no problemática, pero para ello es necesario estimar por MCO los coeficientes π_0, π_1 y formar la variable $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$. La segunda etapa consiste en la estimación por MCO de la regresión de la variable a explicar Y_i sobre $\hat{X}_i$. Los estimadores resultantes de la segunda regresión son los estimadores MC2E que denotamos por $\hat{\beta}_0^{MC2E}$ y $\hat{\beta}_1^{MC2E}$.

En una regresión por VI simple, esto es, con una sola variable explicativa endógena y un solo instrumento, la estimación MC2E nos conduce a estimaciones consistentes de los coeficientes de la Ecuación (8.17). Para ello expresamos β_1 en función de las covarianzas poblaciones que induce la Ecuación (8.17):

$$\begin{aligned} \text{cov}(Z_i, Y_i) &= \text{cov}\left[Z, (\beta_0 + \beta_1 X_i + \varepsilon_i)\right] \\ &= \beta_1 \text{cov}\left(Z_i, X_i\right) + \text{cov}\left(Z_i, \varepsilon_i\right). \end{aligned}$$

A partir del requisito de exogeneidad del instrumento, $\text{cov}(Z_i, \varepsilon_i) = 0$, y dado el cumplimiento del requisito de relevancia, $\text{cov}(Z_i, X_i) \neq 0$, podemos encontrar (identificar) la expresión poblacional del parámetro β_1:

$$\beta_1 = \frac{\text{cov}\left(Z_i, Y_i\right)}{\text{cov}\left(Z_i, X_i\right)}. \tag{8.18}$$

Es decir, el coeficiente poblacional es el cociente de la covarianza poblacional entre Z e Y y la covarianza poblacional entre Z y X.

Para estimar consistentemente estas covarianzas poblacionales, podemos utilizar sus análogos muestrales:

$$\hat{\beta}_1^{MC2E} = \frac{\widehat{\text{cov}}\left(Z_i, Y_i\right)}{\widehat{\text{cov}}\left(Z_i, X_i\right)} = \frac{\sum_{i=1}^n \left(Z_i - \bar{Z}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^n \left(Z_i - \bar{Z}\right)\left(X_i - \bar{X}\right)}, \tag{8.19}$$

y

$$\hat{\beta}_0^{MC2E} = \bar{Y} - \widehat{\beta}_1^{MC2E} \bar{X}, \tag{8.20}$$

si el instrumento «Z» y la variable endógena explicativa «X» coinciden, entonces los estimadores por VI y MCO $\left(\widehat{\beta}_1 = \frac{\widehat{\text{cov}}(X_i, Y_i)}{\widehat{\text{var}}(X_i)}\right)$ coinciden. De hecho si recordamos las ecuaciones normales de la estimación MCO

$$\sum_{i=1}^n \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = \sum_{i=1}^n \hat{\varepsilon}_i = 0$$

y

$$\sum_{i=1}^n X_i \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = \sum_{i=1}^n X_i \hat{\varepsilon}_i = 0,$$

que nos conducían al estimador MCO, y utilizamos ahora la variable Z para «instrumentalizar» la variable explicativa endógena X en dichas ecuaciones obtendríamos:

$$\sum_{i=1}^{n}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i\right)=\sum_{i=1}^{n}\hat{\varepsilon}_i=0 \tag{8.21}$$

y

$$\sum_{i=1}^{n}Z_i\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i\right)=\sum_{i=1}^{n}Z_i\hat{\varepsilon}_i=0, \tag{8.22}$$

que resolviendo nos permitiría recuperar la expresión del estimador MC2E (8.19), y por tanto también el estimador MC2E del término independiente.

Dado que las covarianzas muestrales en dicha expresión (8.19) son estimadores consistentes de sus respectivas poblacionales, es decir, $\widehat{\text{cov}}\left(Z_i,Y_i\right)\xrightarrow{p}\text{cov}\left(Z_i,Y_i\right)$, y $\widehat{\text{cov}}\left(Z_i,X_i\right)\xrightarrow{p}\text{cov}\left(Z_i,X_i\right)$, tendremos que

$$\hat{\beta}_1^{MC2E}\xrightarrow{p}\beta_1,$$

por lo que el estimador de VI es también consistente.

De nuevo el uso de teorema central del límite, al tratarse de promedios de variables aleatorias, nos conduce a la normalidad. Por tanto, para muestras grandes resulta que el estimador de MC2E nos conduce a una distribución normal

$$\hat{\beta}_1^{MC2E}\overset{as}{\sim}N\left(\beta_1,\sigma^2_{\hat{\beta}_1^{MC2E}}\right),$$

donde

$$\sigma^2_{\hat{\beta}_1^{MC2E}}=\frac{\text{var}\left(\left(Z_i-\mu_Z\right)\varepsilon_i\right)}{n\left[\text{cov}\left(X_i,Z_i\right)\right]^2}. \tag{8.23}$$

La expresión (8.23) se puede estimar a partir de la estimación de la varianza y covarianza que aparecen en la misma. La raíz cuadrada de la estimación de (8.23) es el error estándar del estimador VI. Dado que el error podría ser heterocedástico hemos de asegurarnos de utilizar las versiones robustas a la heterocedasticidad por los mismos motivos que lo hacíamos con el estimador MCO en regresión múltiple.

Para contrastar hipótesis sobre β_1 utilizamos un estadístico tipo t, y si queremos construir un intervalo de confianza al 95 %, siempre que la muestra sea grande, lo haremos de este modo

$$\hat{\beta}_1^{MC2E}\pm1,96\times ee\left(\hat{\beta}_1^{MC2E}\right).$$

Generalización a varias variables

En un modelo general de regresión con VI hay que considerar cuatro tipos de variables. Vamos a introducir una notación ligeramente distinta a la que hemos mantenido hasta ahora en este tema para referirnos a las variables explicativas endógenas: la *variable dependiente* que es endógena, Y_0; las *variables explicativas (regresores) endógenas*, que están correlacionadas con el término error, y que por ser endógenas, pero distintas de la

dependiente, las denotamos por Y_k siendo el subíndice $k > 0$; los regresores que son *variables exógenas* incluidas, X; y por último las *variables instrumentales, Z*.

El caso del modelo simple de regresión por VI [Ecuación (8.17)] quedaría con esta notación de la siguiente manera

$$Y_{0,i} = \beta_0 + \beta_1 Y_{1,i} + \varepsilon_i. \tag{8.24}$$

En este caso la regresión VI simple de la sección anterior es factible que se practique porque hay el mismo número de regresores endógenos, $k = 1$, que instrumentos, Z. Si hubiera menos, es decir, en caso de que no hubiera instrumentos no podríamos realizar la regresión de la primera etapa. Ahora bien, en caso de que hubiera más instrumentos que regresores endógenos sí podríamos hacer la regresión de la primera etapa. Por tanto, es especialmente relevante la relación entre el número de instrumentos (m) y el número de regresores endógenos (k). Decimos que los coeficientes de regresión están ***exactamente identificados*** si el número de instrumentos es igual al número de regresores endógenos, es decir, $m = k$. Los coeficientes están ***sobreidentificados*** si el número de instrumentos es mayor que el número de regresores endógenos, $m > k$. Si el número de instrumentos es menor, diríamos que los coeficientes están ***subidentificados***. Para estimar los coeficientes mediante cualquier técnica de variables instrumentales, estos deben estar identificados o sobreidentificados.

Modelo con un único regresor endógeno

Supongamos una ecuación general con «r» variables explicativas exógenas y una explicativa endógena

$$Y_{0i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_r X_{ri} + \beta_{r+1} Y_{1i} + \varepsilon_i \tag{8.25}$$

donde la variable «Y_{1i}» es una variable explicativa endógena (correlacionada con los errores «ε_i»), el resto de variables explicativas son exógenas (no correlacionadas con el término de error «ε_i»), y por tanto el número de regresores endógenos es 1, $k = 1$; el número de regresores total es $r + 1 (= r + k)$, al que habrá que incorporar la constante. La Ecuación (8.25) a veces se denomina **ecuación estructural**.

Supongamos por ahora que tenemos solo una variable instrumental ($m = 1$) y que por tanto cumple las condiciones de exogeneidad, «Z_{1i}» exógena respecto a (8.25), esto es, no correlacionada con los errores «cov $(Z_{1i}, \varepsilon_i) = 0$»; y de relevancia «$Z_{1i}$», está correlacionada con la variable endógena explicativa «cov $(Z_{1i}, Y_{1i}) \neq 0$».

Esta última condición de relevancia la podemos intentar contrastar directamente mediante la primera etapa cuando formamos la regresión:

$$Y_{1i} = \pi_0 + \pi_1 X_{1i} + \pi_2 X_{2i} + ... + \pi_k X_{ki} + \pi_{k+1} Z_{1i} + u_i, \tag{8.26}$$

donde regresamos la variable explicativa endógena «Y_{1i}» con todas las variables exógenas de la ecuación estructural (8.25) y el instrumento. Esta ecuación se denomina **forma**

reducida del modelo estructural (8.25) para la variable $Y_{k=1}$. El requisito de correlación entre la variable explicativa endógena «Y_{1i}» y el instrumento «Z_{1i}» se confirma si el estimador del coeficiente asociado al único instrumento «π_{k+1}» de (8.26) es significativamente distinto de cero, para lo que podemos hacer un test tipo t. De manera que si «$\pi_{k+1} \neq 0$» entonces «Z_{1i}» cumpliría el requisito de relevancia respecto de la variable explicativa endógena «Y_{1i}». En tal caso sabemos que entonces la ecuación estructural (8.25) está *identificada*, y podría estimarse[2].

En este caso particular, en la primera etapa del método de MC2E estimamos la variable explicativa endógena «$\hat{Y}_{1i}$» por MCO utilizando la forma reducida (8.26), y en la segunda etapa estimamos el modelo estructural (8.25), también por MCO, pero sustituyendo la variable explicativa endógena «Y_{1i}» por la estimación realizada en la primera etapa «$\hat{Y}_{1i}$».

Puesto que la forma reducida (8.26) está constituida por variables exógenas [no correlacionadas con los errores del modelo estructural (8.25)], la estimación de la variable explicativa endógena «$\hat{Y}_{1i}$» tampoco está correlacionada con los errores del modelo estructural. La ecuación en forma reducida también se puede escribir como «$Y_{1i} = \hat{Y}_{1i} + u_i$» y sustituyendo esta expresión en (8.25) tenemos que

$$Y_{0i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \beta_{k+1} \hat{Y}_{1i} + (\varepsilon_i + \beta_{k+1} u_i)\,, \tag{8.27}$$

donde comprobamos que los errores «$\varepsilon_i + \beta_{k+1} u_i$» tienen media cero y están incorrelacionados con todas las variables explicativas, y ahora además los estimadores son consistentes.

En ocasiones contamos con más de una variable instrumental para «Y_{1i}», por ejemplo si «Z_{1i}» y «Z_{2i}» son variables excluidas de modelo (8.25) y exógenas [en el sentido de estar incorreladas con los residuos «ε_i» del modelo estructural (8.25)], entonces incluiremos ambos instrumentos «Z_{1i}» y «Z_{2i}» en la forma reducida (8.26) junto con el resto de exógenas incluidas en el modelo.

$$Y_{1i} = \pi_0 + \pi_1 X_{1i} + \pi_2 X_{2i} + ... + \pi_k X_{ki} + \pi_{k+1} Z_{1i} + \pi_{k+2} Z_{2i} + u_i.$$

Estimaríamos esta ecuación de la primera etapa por MCO

$$\widehat{Y_{1i}} = \hat{\pi}_0 + \hat{\pi}_1 X_{1i} + \hat{\pi}_2 X_{2i} + ... + \hat{\pi}_k X_{ki} + \hat{\pi}_{k+1} Z_{1i} + \hat{\pi}_{k+2} Z_{2i}. \tag{8.28}$$

Si son en conjunto significativamente distintas de cero, es decir si podemos rechazar la hipótesis nula «$H_0 : \pi_{k+1} = \pi_{k+2} = 0$», entonces ambas variables cumplen el requisito de relevancia para «Y_{1i}» o, dicho de otra forma, el modelo (8.25) está identificado y lo podemos estimar, como vimos en la sección anterior, consistentemente por el método de **mínimos cuadrados en dos etapas (MC2E)**. La ecuación de la segunda etapa consistirá en usar de nuevo la expresión (8.27) pero usando la variable estimada en la primera etapa indicada en (8.28).

[2]Podríamos utilizar cada una de las variables instrumentales para estimar el modelo estructural (8.25) por MC2E, pero entonces tendríamos dos estimadores diferentes y normalmente ninguno de los dos sería eficiente. Una buena combinación de ambos será más eficiente.

Los programas especializados suelen estimar de forma rutinaria por MC2E y por tanto no es necesario realizar las dos etapas manualmente. Esto es especialmente importante porque los errores estándar que calcularíamos a partir de la segunda etapa se calcularían (como vemos) con estimadores del término error inapropiados pues no solo incorporarían (en tal caso) a los ε_i. El software econométrico especializado evita el realizar las dos etapas, y solventa esta fuente de confusión o error. Generalmente estos programas piden que se especifique la ecuación estructural (8.25) y otro conjunto de variables que incorpore todas las variables exógenas del modelo estructural y las variables instrumentales propiamente dichas. En todo caso, salvo que expresamente se indique lo contrario, presentaremos la regresión con Y_{1i} en lugar de con $\hat{Y}_{1i}$ indicando siempre cuáles han sido los instrumentos.

Para comprender aún más la lógica de los MC2E, consideremos el modelo de regresión simple con más de un instrumento disponible. En realidad para estimar por MC2E, según nos indican las ecuaciones (8.21) y (8.22), bastaría un solo instrumento por lo que podríamos desechar los restantes. Sin embargo desechar instrumentos es desperdiciar información (si los instrumentos son buenos). Como hemos descrito, el método de los mínimos cuadrados en dos etapas (MC2E) nos conduce a considerar toda la información a través de la variable $\hat{Y}_1$, constituida a partir de los instrumentos disponibles, y en ese caso las ecuaciones normales serían

$$\sum_{i=1}^{n}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i\right)=\sum_{i=1}^{n}\hat{\varepsilon}_i=0$$

$$\sum_{i=1}^{n}\hat{Y}_{1i}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i\right)=\sum_{i=1}^{n}\hat{Y}_{1i}\hat{\varepsilon}_i=0.$$

Una forma alternativa de estimación es posible. Consideremos, por simplificar, que tenemos dos instrumentos (Z_1, Z_2) para el modelo de regresión simple. Ahora, además de las restricicciones sobre los momentos (8.21) y (8.22), habrá otra condición o restricción nueva, por lo que tendremos un total de tres restricciones

$$\sum_{i=1}^{n}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i\right)=\sum_{i=1}^{n}\hat{\varepsilon}_i=0,$$

$$\sum_{i=1}^{n}Z_{1i}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i\right)=\sum_{i=1}^{n}Z_{1i}\hat{\varepsilon}_i=0,$$

$$\sum_{i=1}^{n}Z_{2i}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i\right)=\sum_{i=1}^{n}Z_{2i}\hat{\varepsilon}_i=0$$

es decir, ahora tenemos tres ecuaciones con solo dos incógnitas, por lo que en principio podríamos obviar una de las ecuaciones y usar las dos restantes para resolver y despejar las incógnitas. Sin embargo para evitar desperdiciar información, podemos seleccionar los $\hat{\beta}_1, \hat{\beta}_0$ que más se aproximen a satisfacer simultáneamente las tres *restricciones muestrales*.

Esta vía de estimación conduce a la denominada estimación por el *Método Generalizado de los Momentos* (MGM o *GMM*, por sus siglás en inglés), y que desarrollamos en otro tema. De hecho, como entonces se verá, el estimador GMM es más eficiente que el de MC2E, estimador (este último) que bajo ciertos supuestos es un caso particular (de dichos supuestos) de estimación GMM.

Extensión al caso de múltiples regresores endógenos

Es perfectamente factible que el modelo tenga más de una variable explicativa endógena. Consideremos un modelo estructural general con «r» variables exógenas y «k» variables explicativas endógenas

$$Y_{0i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_r X_{ri} + \beta_{(r+1)} Y_{(1)i} + ... + \beta_{(r+k)} Y_{(k)i} + \varepsilon_i. \tag{8.29}$$

Para estimar (8.29) necesitamos un número «m», igual o mayor a «k», de variables instrumentales exógenas al modelo estructural (8.29) que estén correlacionadas con las «k» variables endógenas del modelo estructural; esto se denomina **condición de orden** (número de instrumentos al menos igual al número de variables explicativas endógenas).

Desafortunadamente la condición de orden es necesaria pero no suficiente para identificar y por tanto para poder estimar el modelo estructural (8.29). La condición suficiente para identificar el modelo estructural, denominada **condición de rango**, requiere estimar todas las ecuaciones reducidas del modelo estructural. Para la Ecuación (8.29), tenemos «k» ecuaciones reducidas

$$\begin{aligned}
Y_{(1)i} &= \pi_{01} + \pi_{11} X_{1i} + \pi_{21} X_{2i} + ... + \pi_{r1} X_{ri} + \pi_{(r+1)1} Z_{(1)i} + ... + \pi_{(r+m)1} Z_{(m)i} + u_{1i} \\
Y_{(2)i} &= \pi_{02} + \pi_{12} X_{1i} + \pi_{22} X_{2i} + ... + \pi_{r2} X_{ri} + \pi_{(r+1)2} Z_{(1)i} + ... + \pi_{(r+m)2} Z_{(m)i} + u_{2i} \\
&\cdots \\
Y_{(k)i} &= \pi_{0k} + \pi_{1k} X_{1i} + \pi_{2k} X_{2i} + ... + \pi_{rk} X_{ri} + \pi_{(r+1)k} Z_{(1)i} + ... + \pi_{(r+m)k} Z_{(m)i} + u_{ki}.
\end{aligned} \tag{8.30}$$

Si utilizamos notación matricial para mostrar los estimadores de los instrumentos, obtenemos la siguiente matriz:

$$\begin{pmatrix}
\pi_{(r+1)1} & \pi_{(r+2)1} & \cdots & \pi_{(r+m)1} \\
\pi_{(r+1)2} & \pi_{(r+2)2} & \cdots & \pi_{(r+m)2} \\
\vdots & \vdots & \ddots & \vdots \\
\pi_{(r+1)k} & \pi_{(r+2)k} & \cdots & \pi_{(r+m)k}
\end{pmatrix}. \tag{8.31}$$

En el caso de que el número de instrumentos sea igual al número de variables endógenas, la matriz anterior sería una matriz cuadrada, al ser el número de filas igual al número de columnas. Lo que interesa no es tanto el número de filas y columnas, sino el rango de la matriz para que el modelo sea estimable. Para que el modelo estructural (8.29) esté

identificado y sea estimable, el rango de la matriz de orden $(k \times m)$ debe ser igual al número de variables endógenas explicativas «k» (igual al número de filas).

En caso de que $m = k$, la condición se satisface si su determinante es distinto de cero. Si el rango es menor (determinante nulo) entonces el modelo estructural no está identificado y no es estimable.

Si el número de instrumentos, m, es mayor que el número de endógenas explicativas entonces la matriz de (8.31) tendrá un número de columnas mayor que «k», en tal caso el modelo estructural está identificado si podemos construir una matriz de «k» columnas y «k» filas cuyo determinante sea distinto de cero (matriz de rango «k»).

La condición de orden determina si el modelo está **sobreidentificado**, situación que se produce cuando el número de instrumentos excluidos de la ecuación estructural 8.29 es mayor que el número de variables endógenas explicativas. Cuando el número de instrumentos es igual al número de variables endógenas, entonces sabemos que el modelo estructural (8.29) está **exactamente identificado** y podemos estimarlo.

La estimación por MC2E del caso general de la Ecuación (8.29) comprende las siguientes dos etapas:

1. Regresiones en la primera etapa: regresar por MCO Y_{1i} sobre una lista de variables formada por las variables instrumentales $(Z_{1i}, ..., Z_{mi})$ y por las variables exógenas incluidas $(X_{1i}, ..., X_{ri})$, incluyendo el intercepto. Esto nos permite calcular los valores estimados de Y_{1i}, que hemos denominado $\hat{Y}_{1i}, i = 1, ..., n$. Esto se repite para todos los regresores endógenos, $Y_{2i}, ..., Y_{ki}$, calculando por tanto sus valores estimados respectivos, $\hat{Y}_{2i}, ..., \hat{Y}_{ki}, i = 1, ..., n$.
2. Regresión en la segunda etapa: regresar por MCO Y_i sobre una lista de variables formada por los valores estimados de las variables endógenas $\left(\hat{Y}_{1i}, ..., \hat{Y}_{ki}\right)$ y sobre las variables exógenas incluidas $(X_{1i}, ..., X_{ri})$, incorporando el intercepto. El estimador MC2E, $\hat{\boldsymbol{\beta}}^{MC2E} = \left(\hat{\beta}_0^{MC2E}, \hat{\beta}_1^{MC2E}, ..., \hat{\beta}_{r+k}^{MC2E}\right)$, son los coeficientes estimados en esta segunda etapa.

El estimador en forma matricial y su desarrollo se encuentra en la última sección de este tema y lo reproducimos a continuación del siguiente modo

$$\hat{\boldsymbol{\beta}}^{MC2E} = \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1}\hat{\mathbf{X}}'\mathbf{Y}$$

donde

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}.$$

En el modelo simple de la sección anterior, establecimos dos condiciones [(8.16) y (8.15)] para la validez de un instrumento. Las condiciones o requisitos para la validez de los instrumentos en el modelo general de VI lógicamente han de incorporar a estas como caso particular. En el caso de que existan varias variables endógenas incluidas en el modelo estructural, debemos garantizar que no hay multicolinealidad perfecta en la regresión poblacional de la segunda etapa. Esto es así porque si hubiera multicolinealidad perfecta en el modelo poblacional no podríamos estimar, dado que los instrumentos

no proporcionarían información suficiente sobre los movimientos exógenos de las endógenas, y por tanto no podríamos «aislar» sus efectos sobre la variable dependiente Y.

Resumimos a continuación las dos condiciones para la validez de un conjunto de m instrumentos:

1. **Relevancia del instrumento**: Los vectores $\left(1, X_{1i}, ..., X_{ri}, \hat{Y}^*_{1i}, ..., \hat{Y}^*_{ki}\right)$ no deben ser perfectamente multicolineales, donde $\hat{Y}^*_{1i}$ es el valor de predicción de Y_{1i} a partir de la regresión poblacional de Y_{1i} sobre los instrumentos (Z) y los regresores exógenos incluidos (X), y «1» es el valor que toma el término constante a todas las observaciones $i = 1, ..., n$. Si solo hay una variable endógena incluida, Y_1, esto se cumple si al menos el coeficiente de un instrumento Z en la regresión poblacional de Y_1 sobre los m instrumentos Z y las r exógenas incluidas X es distinto de cero.
2. **Exogeneidad del instrumento**: Los instrumentos no están correlacionados con el término error: $corr(Z_{ji}, \varepsilon_i) = 0, j = 1, 2, ..., m$.

8.5 Contrastes de endogeneidad.

En esta sección vamos a considerar dos tipos de tests, uno sobre la endogeneidad de una variable, y otro sobre los instrumentos.

La cuestión inicial es saber si una variable es o no endógena debido a todos los problemas que de invalidez que supone. Consideremos un modelo general con una sola variable explicativa endógena

$$Y_0 = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \beta_{(k+1)} Y_1 + \varepsilon, \tag{8.32}$$

donde sospechamos que «Y_1» es endógena. Además contamos con dos instrumentos «Z_1» y «Z_2» (la validez de los instrumentos depende de si son exógenos al modelo o no correlacionados con «ε»).

Para contrastar si «Y_1» es verdaderamente una variable explicativa endógena, Hausman propuso comparar las estimaciones MCO y MC2E y si las diferencias entre ambas estimaciones son significativas concluimos que «Y_1» es endógena, pues de lo contrario (en caso de exogeneidad de la variable) ambos estimadores serían consistentes y no deberían existir diferencias entre una y otra.

Para realizarlo partimos de la forma reducida de «Y_1»

$$Y_1 = \pi_0 + \pi_1 X_1 + ... + \pi_k X_k + \pi_{(k+1)} Z_1 + \pi_{(k+2)} Z_2 + u, \tag{8.33}$$

ecuación que estimamos por MCO. A partir de la misma obtenemos

$$cov(Y_1, \varepsilon) = cov(u, \varepsilon)$$

por tanto

$$cov(Y_1, \varepsilon) = 0 \iff cov(u, \varepsilon) = 0.$$

Vemos entonces que contrastar $cov(Y_1, \varepsilon) = 0$ es equivalente a contrastar $cov(u, \varepsilon) = 0$. Bajo la hipótesis nula $H_0 : cov(Y_1, \varepsilon) = 0$, se verificaría que el coeficiente δ en la regresión

$$\varepsilon = \delta u + error$$

sería nulo ($\delta = 0$). Es decir, contrastar $\delta = 0$ equivaldría a contrastar $H_0 : cov(Y_1, \varepsilon) = 0$. En la práctica dado que no observamos u utilizaríamos el residuo MCO de la primera etapa, $\hat{u}$.

Precisamente, el contraste de endogeneidad, también denominado **contraste de especificación de Hausman**, consiste en introducir los residuos estimados en la forma reducida «$\hat{u}$» como una variable explicativa más, es decir, estimamos la siguiente ecuación por MCO

$$Y_0 = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \beta_{(k+1)} Y_1 + \delta \hat{u} + \varepsilon, \tag{8.34}$$

y contrastamos «δ» de la forma usual (mediante el contraste de la «t»); si «δ» es significativa entonces concluimos que «Y_1» es endógena porque la «ε» de la forma estructural y la «u» de la forma reducida están correlacionadas (además los estimadores «β_i» de esta ecuación coinciden con los estimados por MC2E).

En el caso de r variables potencialmente endógenas, el contraste de Hausman requiriría (i) estimar las r formas reducidas con sus correspondientes residuos (de la forma reducida); (ii) incluir en el modelo de interés como r regresores adicionales cada uno de los residuos obtenido en la fase (i); (iii) hacer un contraste de significación conjunta de dichos residuos mediante el estadístico tipo F siguiente

$$W_0 = (n - k - 1) \frac{SCR_R^* - SCR_{NR}^*}{SCR_{NR}^*} \xrightarrow{d} \chi_r^2$$

donde SCR_R^* es la suma de los cuadrados de los residuos del modelo original (es decir sin incluir los residuos de las formas reducidas), y SCR_{NR}^* la suma de los cuadrados de los residuos del modelo (no reducido), es decir que sí incorpora como regresores los correspondientes residuos de las r formas reducidas. Si los residuos son conjuntamente significativos (esto es, si se rechaza la hipótesis nula), entonces al menos una de las variables explicativas potencialmente endógena lo es en la realidad.

Por otra parte, la exogeneidad también es un requisito para que un instrumento sea válido. En caso de que el instrumento no sea exógeno (sea endógeno), entonces el estimador MC2E convergía necesariamente hacia algo diferente del coeficiente de regresión del modelo poblacional, que es el que nos interesa. Para establecer si un instrumento es exógeno, un primer paso es pensar en los argumentos acerca de por qué puede o no serlo. Para ello es conveniente preguntarse cuáles son los factores que forman parte del término error en la especificación que hayamos hecho del modelo, y entonces pensar si esa lista de factores puede estar (y en qué grado de verosimilitud) relacionada con los instrumentos. Cuando el modelo está sobreidentificado podríamos realizar un contraste sobre la validez de alguno(s) de los instrumentos.

Supongamos que disponemos de un solo regresor endógeno y de dos instrumentos (modelo sobreidentificado). Bajo estas condiciones podríamos estimar el modelo utilizando solo un instrumento, en cuyo caso tendríamos dos estimaciones, una que utiliza

el primer instrumento y otra con el segundo. Si ambos instrumentos son exógenos a la ecuación estructural y están correlacionados con la variable explicativa endógena, es decir si se cumple (8.15) y (8.16), entonces ambas estimaciones deberían ser cercanas (no iguales debido a la variación muestral), y si no es así, entonces parece razonable concluir que uno de los instrumentos, o bien los dos no son exógenos a la ecuación estructural, en el sentido de que no se verifica la expresión (8.15).

Esto es lo que hacemos de forma implícita cuando estimamos por MCO la siguiente expresión:

$$\hat{\varepsilon}_i^{MC2E} = \delta_0 + \delta_1 \cdot Z_{1i} + \delta_2 \cdot Z_{2i} + \delta_{2+1} \cdot X_{1i} + ... + \delta_{2+r} \cdot X_{ri} + \eta_i, \tag{8.35}$$

donde «$\hat{\varepsilon}_i^{MC2E}$» son los residuos estimados por MC2E usando todos los instrumentos, y son por tanto las versiones muestrales de ε_i. Contrastamos mediante la *F* la hipótesis nula de que los instrumentos no son significativos «$H_0 : \delta_1 = \delta_2 = 0$». El estadístico para el contraste de sobreidentificación, también denominado estadístico «J», se construye de la siguiente forma: «$J = mF$», donde «m» es el número de instrumentos, y cuya distribución para muestras grandes sigue una «χ_q^2» donde «q» o grado de sobreidentificación es el número de instrumentos, en nuestro caso «$m = 2$», menos el número de variables explicativas endógenas, en este caso con valor unitario «$q = 2 - 1 = 1$». Nada impide aplicar este contraste de forma general a modelos con más variables explicativas endógenas siempre que el número de instrumentos sea mayor que el número de regresores endógenos. De manera que el contraste «J» nos permite contrastar la exogeneidad de los instrumentos siempre que el modelo esté sobreidentificado.

8.6 Modelos de ecuaciones simultáneas

La simultaneidad en las relaciones teóricas entre las variables económicas es algo extraordinariamente frecuente en economía. El propio concepto o mecanismo de equilibrio induce de forma natural a un determinación conjunta entre variables. En términos generales, la simultaneidad aparece cuando una o más de las variables explicativas se determinan conjuntamente con la variable dependiente.

Como vimos en el primer epígrafe de este tema, la simultaneidad conduce a un sesgo de endogeneidad derivado de la bidireccionalidad causal. La solución a este tipo de sesgos también se produce mediante el uso de VI.

El ejemplo económico ilustrativo que vamos a exponer se circunscribe al marco del equilibrio. Consideramos dos ecuaciones, una de oferta y otra de demanda. Precisamente la técnica de regresión (y estimación) por VI tiene su origen en los trabajos de Phillip y Sewall Wright que pretendían estimar las curvas de oferta y demanda de bienes de naturaleza agrícola.

Supongamos que estamos interesados en estimar la *elasticidad precio de la demanda* de un bien agrícola concreto. Para ello especificamos la siguiente ecuación:

$$Q^d = \beta_0 - \beta_1 \cdot p + \beta_2 \cdot yd + \varepsilon_1, \tag{8.36}$$

donde suponemos que las variables están en logaritmos, la variable «Q^d» es la cantidad demandada, «p» el precio, «yd» la renta disponible y «ε_1» los errores en los que se incluyen el resto de variables independientes no incluidas específicamente. Consideramos además que el bien es normal, en el sentido de que la elasticidad precio «β_1» es negativa y la elasticidad renta «β_2» positiva. En estas condiciones la demanda usual tiene pendiente negativa, los incrementos de los precios reducen las cantidades a lo largo de la curva; y los incrementos de la renta desplazan la curva de demanda.

Sabemos que la determinación del precio y de la cantidad finalmente consumida depende también de la oferta del bien, podemos estimar la siguiente ecuación de oferta:

$$Q^o = \gamma_0 + \gamma_1 \cdot p + \gamma_2 \cdot lluvia + \varepsilon_2, \tag{8.37}$$

donde suponemos que las variables están también en logaritmos, la variable «Q^o» es la cantidad ofrecida, «p» el precio, «$lluvia$» es la lluvia caída y «ε_2» los errores en los que se incluyen el resto de variables independientes no incluidas específicamente. Consideramos además que la elasticidad precio de la oferta «γ_1» es positiva tal y como indica la teoría y que el parámetro «γ_2» de la lluvia caída es positiva, en el sentido de que el aumento de la lluvia en los terrenos donde se cultiva el producto agrícola analizado provoca mayores cosechas y en consecuencia aumentos de oferta. Así pues la oferta tiene pendiente positiva, los incrementos de los precios incrementan la cantidades ofrecidas a lo largo de la curva y los incrementos de lluvia desplazan la curva de oferta.

Hasta ahora hemos establecido una ecuación de oferta (8.37) y otra de demanda (8.36). Cada una de estas dos ecuaciones refleja o resuelve un problema de comportamiento, ya sea del consumidor (en el caso de la demanda), ya sea de la empresa (en el caso de la oferta). Son por tanto dos ecuaciones estructurales, que en este caso están relacionadas internamente.

En efecto, sabemos por la teoría económica que el mercado está en equilibrio (es decir, las transacciones se efectúan) cuando la cantidad ofrecida y demandada coinciden, es decir cuando

$$Q^o = Q^d. \tag{8.38}$$

Cuando el/la econométra tiene datos de precios y cantidades efectivamente observados (intercambiados), está registrando pares de cantidad y precio en periodos diferentes, donde en cada periodo las curvas de oferta y demanda están sujetas a los cambios asociados (desplazamientos) a factores distintos del precio, pero que afectan a la oferta y a la demanda de este mercado. En el modelo expuesto estos factores distintos de precio son la renta y la lluvia. Estimar por MCO una recta de ajuste no permitiría saber si se trata de una curva de demanda, o de una de oferta, ya que dichos puntos han sido determinados tanto por cambios de oferta, como por cambios de demanda. Es decir, no podemos estimar la ecuación de demanda sin tener en cuenta la influencia de la oferta y viceversa, porque ambas se establecen simultáneamente (conjuntamente).

La única forma de estimar (o identificar) la ecuación de demanda es considerar las ecuaciones de oferta y demanda conjuntamente y permitir que la ecuación de oferta

se desplace (modificando los valores de la lluvia caída) de manera que los sucesivos valores de equilibrio se correspondan con la ecuación de demanda.

Podemos expresar ambas ecuaciones de la siguiente forma:

$$\begin{aligned} Q &= \beta_{10} + \alpha_{11}{\cdot}p + \beta_{11}{\cdot}yd + \varepsilon_1 \\ p &= \beta_{20} + \alpha_{22}{\cdot}Q + \beta_{22}{\cdot}lluvia + u, \end{aligned} \tag{8.39}$$

ambas ecuaciones se denominan **modelo de ecuaciones simultáneas**.

La segunda ecuación, la de oferta, la expresamos en su forma inversa [$p = f(Q)$]. Al hacerlo así vemos con mayor claridad que la relación de causalidad entre precios y cantidades no es unidireccional (sino bidireccional), en este caso las cantidades se explican por los precios pero también los precios se explican por las cantidades. Es decir, precios y cantidades son conjuntamente dependientes entre sí, hay una relación de causalidad en ambos sentidos o simultánea o, dicho de otra forma, **precios y cantidades son ambas variables endógenas**.

Si como hemos indicado nuestro interés consiste en estimar la elasticidad de *demanda*, utilizando la terminología de VI, diremos que la variable p es una variable explicativa endógena, $\mathbb{E}(\varepsilon_{1i}|p_i) \neq 0$, la variable exógena incluida es «yd», $\mathbb{E}(\varepsilon_{1i}|yd_i) = 0$. Nos falta por tanto ver qué rol desempeña la variable «$lluvia$» cuando estimamos la función de demanda.

La variable «$lluvia$» nos permite identificar la demanda es que (*i*) la «$lluvia$» está correlacionada con la variable «precio» (porque desplaza la curva y por tanto varía el precio, tal y como indica la ecuación inversa de demanda (8.39)), y (*ii*) porque la «$lluvia$» no está correlacionada con otros factores distintos del precio que determinan la demanda, factores que están recogidos en el término ε_1, es decir, las lluvias no deben tener un efecto directo sobre la demanda del bien agrícola. Podemos decir entonces que la variable «$lluvia$» es justamente el instrumento (la variable instrumental) porque satisface por (*i*) la condición de relevancia, y por (*ii*) la condición de exogeneidad.

Es ilustrativo tratar esta cuestión también analíticamente. Sustituyendo las cantidades de la primera ecuación (demanda) en la segunda (oferta) tenemos

$$(1 - \alpha_{22}\alpha_{11})\, p = (\beta_{20} + \alpha_{22}\beta_{10}) + \alpha_{22}\beta_{11}{\cdot}yd + \beta_{22}{\cdot}lluvia + (\alpha_{22}{\cdot}\varepsilon_1 + u), \tag{8.40}$$

y si «$\alpha_{22}\alpha_{11} \neq 1$» [lo que es muy probable puesto que hemos supuesto que «α_{11}» es negativo (demanda) y «α_{22}» positivo (oferta)] podemos dividir ambas partes de la expresión (8.40) por «$1 - \alpha_{22}\alpha_{11}$» lo que nos lleva a

$$\begin{aligned} p &= \tfrac{\beta_{20}+\alpha_{22}\beta_{10}}{1-\alpha_{22}\alpha_{11}} + \tfrac{\alpha_{22}\beta_{11}}{1-\alpha_{22}\alpha_{11}} yd + \tfrac{\beta_{22}}{1-\alpha_{22}\alpha_{11}} lluvia + \tfrac{\alpha_{22}{\cdot}\varepsilon_1+u}{1-\alpha_{22}\alpha_{11}} \\ &= \pi_{20} + \pi_{21}{\cdot}yd + \pi_{22}{\cdot}lluvia + \varepsilon, \end{aligned} \tag{8.41}$$

que es lo que hemos denominado ecuación en la *forma reducida* para los precios. Además como los errores de la forma reducida «ε» son una combinación lineal de los errores de las ecuaciones estructurales de demanda «ε_1» y oferta «u», los precios «p» y los errores «ε_1» están correlacionados «$\text{cov}(p, \varepsilon_1) \neq 0$», que es precisamente el motivo

por el que la variable precios «p» es una variable explicativa endógena en la ecuación estructural de demanda, y por tanto es lo que genera el sesgo en la estimación MCO del coeficiente de p en la ecuación estructural de demanda. Este sesgo además no puede desaparecer al aumentar el número de observaciones, por lo que el estimador además es inconsistente.

La solución que plantea VI es precisamente usar la variable «$lluvia$» como instrumento, justamente lo que hacemos en la forma reducida (8.41) y que constituye la base de estimación de *primera etapa* en MC2E. Así luego, una vez aislado en $\hat{p}$ la parte incorrelacionada con el error estructural de demanda, podamos en la segunda etapa estimar sin sesgo el coeficiente de la variable explicativa endógena.

Los modelos de ecuaciones simultáneas con datos de naturaleza económica y de sección cruzada (como es el caso) pueden incorporar, lógicamente, más de dos ecuaciones. La identificabilidad de cada ecuación dependerá de las condiociones de orden y rango anterioremente expuestas. Lo central en los verdaderos modelos de ecuaciones simultáneas es que cada ecuación en el sistema debe tener una interpretación causal en términos ceteris paribus, tal y como sucede en el caso de la demanda y oferta, donde cada ecuación por separado responde, en este caso, a un comportamiento diferenciado por parte o bien de los demandantes o bien de los oferentes.

Bibliografía complementaria

Matilla-García, M et al. 2017. Econometría y Predicción. McGraw Hill

Stock J. and Watson J. Introducción a la econometría. Pearson.

Tema 9

Modelos de panel lineales

Este tema está elaborado como una adaptación de los capítulo 13 y 14 de:

Wooldridge. J. 4th Ed., Introductory Econometrics,

y del capítulo 21 (secciones 21.1, 21.2 y 21.4. 21.5, 21.6, 21.7 y 21.8)

Cameron and Trivedi. Microeconometrics: methods and applications.

Así como de la bibliografía complementaria

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al Órgano convocante ni al Tribunal actuante.

- Modelo de datos apilados.
- Modelo de efectos fijos y su estimación.
- Modelo de efectos aleatorios y su estimación.
- Modelos de efectos fijos vs. modelos de efectos aleatorios

Los economistas utilizan tradicionalmente el término datos de panel para referirse a estructuras de datos que consisten en observaciones sobre individuos durante múltiples períodos de tiempo. Otros campos, como por ejemplo el de la estadística, suelen referirse a esta estructura como datos longitudinales. Los "individuos" observados pueden ser, por ejemplo, personas, hogares , trabajadores, empresas, escuelas, plantas de producción, industrias, regiones, estados o países. La característica distintiva respecto de los conjuntos de datos de sección cruzada, que es los que hasta el momento hemos considerado, es la presencia de múltiples observaciones para cada individuo. Los métodos de datos de panel se pueden aplicar a cualquier contexto que tenga dependencia del tipo de conglomerado.

Hay varias ventajas de los datos de panel en relación con los datos de sección cruzada. Una es la posibilidad de controlar la endogeneidad invariante en el tiempo no observada sin el uso de variables instrumentales. Otra ventaja es que permite el tratamiento de formas más amplias de heterogeneidad.

La forma general del panel para una de las variables sería:

	1	2		..	..	T	
1	y_{11}	y_{12}	...			y_{1T}	$y_{1.}$
2	y_{21}	y_{22}	...			y_{2T}	$y_{2.}$
:	...	...	...	y_{it}	...	...	$y_{i.}$
N	y_{N1}	y_{N2}				y_{NT}	$y_{N.}$
	$y_{,1}$	$y_{,2}$	...	$y_{.t}$	...	$y_{.T}$	

Este panel define una variable y_{it} en dos dimensiones, la individual de la sección cruzada, i, y la dimensión temporal, t. Ambas configuran el ancho y el largo del panel, y por tanto no son dimensiones intercambiables. El índice temporal marca una ordenación (en el tiempo cronológico: días, semanas, meses, trimestres, años,...) y dota de una interpretación común a muchos paneles. Sin embargo, el índice individual, i, no tiene ningún orden, y además su interpretación o contenido varía según la aplicación en cuestión. Se puede referir a personas, empresas, municipios, países, árboles, etcétera.

En función de la forma del panel podríamos distinguir entre paneles de series temporales $(T > N)$ que son comunes en macroeconomía, y paneles de secciones cruzadas $(N > T)$ que dominan en microeconomía (especialmente en economía laboral). También se hace referencia a paneles largos cuando el número de periodos es mayor que el número de observaciones transversales $(T > N)$ o cortos cuando ocurre lo contrario $(T < N)$.

Hay dos categorías amplias de conjuntos de datos de panel en aplicaciones económicas: micro paneles y macro paneles. Los micro paneles son típicamente encuestas o registros administrativos de individuos y se caracterizan por un gran número de personas (a menudo 1000 o más) y un número de tiempo relativamente pequeño períodos (a menudo de 2 a 20 años). Los paneles macro son típicamente variables macroeconómicas nacionales o regionales, y se caracterizan por un número moderado de individuos (por ejemplo, 7-20) y un número moderado de tiempo períodos (20-60 años).

9.1 Modelos de datos apilados

Nos referimos a **datos fusionados o datos apilados** cuando utilizamos datos obtenidos mediante muestreo aleatorio en diferentes momentos de tiempo. La característica fundamental de este conjunto de datos es que provienen de observaciones muestrales independientes aunque probablemente las observaciones referidas a distintos momentos de tiempo puedan no estar idénticamente distribuidas. Veremos que esta cuestión se puede incorporar al análisis de regresión permitiendo que el término constante (y a veces también la pendiente) varíen con el tiempo.

Una de las razones para utilizar estos datos es que al fusionar las secciones de distintas encuestas incrementamos el tamaño de la muestra. Siempre que la relación entre la variable dependiente y al menos alguna de las variables explicativas permanezca constante a lo largo del tiempo resultará beneficioso fusionar los datos de las secciones independientes, puesto que se consiguen estimadores más precisos. Estadísticamente el tratamiento es similar al que hacemos en una sección. Ahora el número de elementos muestrales es NT, por lo que tomamos muestras de tamaño N en diferentes T momentos

del tiempo, lo que invita a considerar que las observaciones no necesariamente han de estar idénticamente distribuidas. Por ejemplo, la distribución de la renta o de los salarios ha cambiado a lo largo del tiempo.

Estas características hacen que este tipo de análisis de datos fusionados resulte útil para evaluar los efectos de política económica o los cambios provocados como consecuencia de distintos escenarios. De hecho se pueden relacionar fácilmente estas técnicas con la literatura sobre experimentos naturales donde hay grupos de control y de experimentación.

Matricialmente el modelo de datos fusionados es el siguiente

$$Y_{it} = \alpha + \mathbf{x}_{it}^{'}\boldsymbol{\beta} + \varepsilon_{it} \tag{9.1}$$

donde $\mathbf{x}_{it}$ es un vector $k \times 1$ de variables independientes, y el resto son escalares con los significados habituales en el modelo de regresión, pero con los subíndices de tiempo y de individuos que anteriormente indicamos. Apreciamos que todos los coeficientes son constantes a lo largo del tiempo, lo que introduce *a priori* una fuerte restricción, y que parcialmente podemos relajar introduciendo en el vector $\mathbf{x}_{it}$ alguna variable que no cambie en el tiempo, para lo que usaríamos variables binarias de género, industria, estado o región en función del tipo de entidad considerada en i.

Podemos dar una expresión matricial aún más compacta para la ecuación (9.1) de modo que *para cada entidad* (individuo, empresa, región,...) definimos

$$\mathbf{y}_i = \mathbf{W}_i\boldsymbol{\delta} + \boldsymbol{\varepsilon}_i$$

donde $\boldsymbol{\delta} = \begin{bmatrix} \alpha & \boldsymbol{\beta}' \end{bmatrix}'$ de dimensiones $(k+1) \times 1$ es el vector de parámetros, los vectores $\mathbf{y}_i$ y $\boldsymbol{\varepsilon}_i$ son $T \times 1$ formados por la respectiva entrada *t*-ésima de y_{it} y ε_{it}, y la matriz $\mathbf{W}_i$ es de dimensiones $T \times (k+1)$ donde la fila *t*-ésima es $\mathbf{w}_{it}' = [1\ \mathbf{x}_{it}]'$.

Dado que tenemos N entidades, si las apilamos o fusionamos una a continuación de otra tendremos

$$\mathbf{y} = \mathbf{W}\boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

donde ahora $\mathbf{y}$ y $\boldsymbol{\varepsilon}$ son vectores $NT \times 1$, y $\mathbf{W}$ es una matriz de regresores de dimensiones $NT \times (k+1)$ con la primera columna de unos.

Para conseguir estimadores MCO de los parámetros

$$\hat{\boldsymbol{\delta}}_{FUSIONADOS} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}$$

es preciso requerir que la matriz de regresores no sea perfectamente multicolineal; para que sean consistentes y asintóticamente normales se requiere que los regresores no estén correlacionados con los errores del modelo, $\mathbb{E}(\boldsymbol{\varepsilon}\,|\mathbf{W}) = \mathbf{0}$; y para realizar inferencia, lo hacemos con errores estándar robustos a la autocorrelación y a la heterocedasticidad estimados a partir de

$$\widehat{Var(\hat{\boldsymbol{\delta}}_{FUSIONADOS})} = \left[\sum_{i=1}^{N}\mathbf{W}_i'\mathbf{W}_i\right]^{-1}\sum_{i=1}^{N}\mathbf{W}_i'\hat{\boldsymbol{\varepsilon}}_i\hat{\boldsymbol{\varepsilon}}_i{}'\mathbf{W}_i\left[\sum_{i=1}^{N}\mathbf{W}_i'\mathbf{W}_i\right]^{-1}$$

donde $\hat{\boldsymbol{\varepsilon}}_i = \mathbf{y}_i - \mathbf{W}_i\hat{\boldsymbol{\delta}}$. De forma más compacta y equivalente lo podemos expresar así:

$$\widehat{Var(\hat{\boldsymbol{\delta}}_{FUSIONADOS})} = [\mathbf{W}'\mathbf{W}]^{-1}\,\mathbf{W}'\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}'\mathbf{W}\,[\mathbf{W}'\mathbf{W}]^{-1}\,.$$

Este tipo de modelos con datos apilados también podría utilizar la técnica de estimación por variables instrumentales (VI) y los contrastes o test de especificación desarrollados en temas precedentes.

9.2 Modelo de efectos fijos y su estimación

Estos modelos se aplican propiamente a los **datos de panel** (también denominados *datos longitudinales*). Recordemos que son datos que también tienen conjuntamente dimensión transversal y temporal, pero que se diferencian de los datos fusionados en que las entidades individuales observadas (familias, empresas, ciudades, estados, etc.) son las mismas a lo largo del tiempo. Lógicamente, no podemos suponer que las observaciones estén distribuidas de forma independiente en el tiempo, pues se trata de las mismas unidades y por lo tanto es factible que los factores no observados afecten a lo largo del tiempo considerado en la estructura del panel.

Una suposición mantenida típica para los micropaneles (que seguimos en este capítulo) es que los individuos son mutuamente independientes, mientras que las observaciones de un individuo dado están correlacionadas a través de los periodos de tiempo. Esto significa que las observaciones siguen una estructura de dependencia agrupada. Debido a esto, la práctica econométrica actual es utilizar estimadores de la matriz de covarianza robustos por conglomerados cuando sea posible. Con frecuencia se utilizan supuestos similares para los paneles macro, aunque el supuesto de independencia entre individuos (por ejemplo, países) es mucho menos convincente.

El caso más sencillo es aquel panel formado por dos periodos y una variable observable explicativa. Supongamos que tenemos datos para dos periodos temporales $(t = 1, 2)$ y N valores de corte transversal para entidades individuales $(i = 1, 2, ..., N)$ relativos a dos variables, «Y_{it}» y «X_{it}», donde el subíndice «i» indica la entidad individual y el subíndice «t» el periodo de tiempo. Además, consideremos que el modelo que relaciona ambas variables es

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \varepsilon_{it}, \tag{9.2}$$

donde la variable «Z_i» es una variable que influye en «Y_{it}» pero que no varía con el tiempo, es decir tiene carácter idiosincrásico para las entidades individuales, como los hábitos culturales, personales, rasgos o atributos característicos, etc. Si se tratara de una ecuación salarial, dicha variable podría referirse a la habilidad natural de cada trabajador; si se tratara de una ecuación de producción, podría referirse a los conocimientos organizativos de la entidad i considerada. De este modo parte de la heterogeneidad o singularidad de cada entidad sería contemplada. Debido a que esta variable no varía con el tiempo (temporalmente invariantes), la influencia será igual en

ambos periodos y por ello solo incluimos el subíndice «i» en la expresión. Consideramos además que la variable «Z_i» es inobservable.

En estas condiciones, si realizamos la regresión de corte transversal entre «Y_i» y «X_i» y dejamos fuera de la misma a la variable Z_i, puesto que no es observable, corremos el riesgo de que el estimador del efecto parcial sea sesgado, y lo será sin duda si X_i y Z_i están correlacionados.

Alternativamente, y debido a que Z_i no cambia con el tiempo, puede eliminarse (sin correr el riesgo de obtener estimadores inconsistentes) mediante el análisis de las diferencias temporales entre ambos periodos. Formalmente tenemos que la estimación del primer periodo es

$$Y_{i1} = \beta_0 + \beta_1 X_{i1} + \beta_2 Z_i + \varepsilon_{i1}, \tag{9.3}$$

y la del segundo periodo

$$Y_{i2} = \beta_0 + \beta_1 X_{i2} + \beta_2 Z_i + \varepsilon_{i2}, \tag{9.4}$$

de manera que la diferencia entre ambas es

$$\begin{gathered}(Y_{i2} - Y_{i1}) = \beta_1 (X_{i2} - X_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1}) \\ \triangle Y_i = \beta_1 \triangle X_i + \triangle \varepsilon_i,\end{gathered} \tag{9.5}$$

y por tanto el **estimador de la diferencia** es una forma de calcular «β_1» sin incurrir en el problema de variables omitidas del modelo (9.2). La intuición es clara: El estimador de la diferencia da cuenta del cambio de la variable Y_{it} producido para una unidad individual i entre un periodo (antes) y el siguiente considerado (después). Si la variable Z_i, pese tener un efecto diferente sobre las distintas unidades individuales, no experimentó ningún cambio de un periodo a otro, entonces no pudo ejercer ningún efecto sobre el cambio de Y_{it} (es decir, sobre $\triangle Y_i$). Los cambios de Y_{it} provienen de los cambios en la variable explicativa X_{it} al pasar de $t = antes$ a $t = después$, y de los cambios en otros factores que determinan la variable Y_{it}, pero que no hemos hecho explícitos y por tanto están en la variación (cambio) de los errores, $\triangle \varepsilon_i$.

Así pues, el estimador de la diferencia es el estimador MCO en la Ecuación (9.5), que como hemos comprobado explota la singularidad de los datos de panel: medir la asociación entre regresores específicos de cada entidad que cambian de un periodo a otro y los cambios de un periodo a otro en la variable dependiente también específicos de la entidad correspondiente. Se observa fácilmente que este método de estimación no permite identificar a los coeficientes de los regresores invariantes en el tiempo. Sin embargo, este estimador de la diferencia está limitado a dos periodos.

El **método de efectos fijos** que explicamos seguidamente facilita estimar directamente el modelo (9.2) con dos o más periodos.

Debido a que la variable inobservable «Z_i» de (9.2) no varía entre periodos para cada entidad individual, también podemos escribir la Ecuación (9.2) con «n» términos cons-

tantes[1] (tantos como entidades individuales), y por ello la expresión más habitual de (9.2) es

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \varepsilon_{it}. \tag{9.6}$$

En este modelo los α_i, o efectos fijos individuales, se tratan como términos independientes a estimar en la ecuación (para cada entidad individual). Hay por tanto n efectos fijos individuales, efectos que son distintos como resultado de las variables *omitidas* invariantes en el tiempo. Se observa que el coeficiente poblacional de la pendiente, β_1, es el mismo para todas las entidades, siendo el intercepto o término constante lo que varía entre las mismas. Se comprueba también que, al igual que sucede con el *estimador de la diferencia*, este modelo es menos restrictivo que el modelo de datos fusionados toda vez que permite que el intercepto o constante varíe a lo largo de los individuos, y así se captura cierto grado de heterogeneidad individual no observada.

La estimación por MCO no es adecuada si se aplica directamente sobre la ecuación (9.6) porque produce estimadores sesgados e inconsistentes toda vez que $\mathbb{E}(\varepsilon_{it}) = \alpha_i$. Una alternativa es considerar a α_i como un coeficiente de una variable dummy (binaria), de esta manera podemos caracterizar el modelo de efectos fijos utilizando variables binarias para cada entidad individual, es decir, podemos considerar el modelo de regresión siguiente:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \gamma_3 D3_i + ... + \gamma_n Dn_i + \varepsilon_{it}. \tag{9.7}$$

Las variables dicotómicas del modelo (9.7) tienen valor unitario para la entidad individual a la que hacen referencia y valor nulo para el resto. Así «$D2_i$» tiene valor unitario para la segunda entidad individual (segunda observación de corte transversal) y valor nulo para el resto. La interpretación es clara: «β_0» es el efecto fijo individual de la primera entidad de corte transversal «α_1»; el segundo, $\beta_0 + \gamma_2 = \alpha_2$, y en general $\beta_0 + \gamma_i = \alpha_i$.

Así pues el estimador de efectos fijos, a diferencia del estimador por MCO fusionado, explota la singularidad propia de los datos de panel. Mide la asociación *entre* las desviaciones específicamente individuales de los regresores respecto de sus correspondientes promedios temporales y las desviaciones específicas individuales de la variable dependiente respecto de su promedio temporal. Un inconveniente del estimador de efectos fijos, compartido también con el estimador de las diferencias, es que la estimación de parámetros asociados a variables que no cambian en el tiempo no es factible[2]. Lo cual lógicamente impide que podamos estimar el efecto parcial de una de este tipo de variables (pensemos por ejemplo en la condición de genéro en una ecuación de salarios) sobre la variable dependiente.

[1] Para facilitar la notación de algunas partes utilizaremos en ocasiones n para referirnos al tamaño muestral de la sección cruzada.

[2] En general, este tipo de modelos no permiten identificar coeficientes de regresores invariantes en el tiempo.

Tanto el estimador de efectos fijos como el estimador de la diferencia producen estimaciones consistentes de los k parámetros o coeficientes de los regresores $X_{it,j}, j = 1, 2, ..., k$, esto es de los efectos parciales sobre la variable dependiente de los cambios en los mismos. Mientras que los N parámetros $\alpha_i, i = 1, ..., N$ tienen un interés escaso o incidental, si bien su presencia es necesaria para la calidad de la estimación de los k. Sobre estos aspectos volveremos más adelante en la exposición.

Normalmente los programas especializados calculan el estimador de efectos fijos sin utilizar variables binarias. Lo hacen en dos fases: En la primera se le resta a cada variable observable la media específica de cada entidad individual y en la segunda se estima la regresión en desviaciones a las medias por MCO. Veamos cómo se calcula para el caso de una sola variable explicativa observable. Por un lado tenemos la ecuación de efectos fijos,

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \varepsilon_{it}, \tag{9.8}$$

calculamos las medias de cada entidad individual de la forma usual: $\overline{Y}_{i.} = T^{-1} \sum_{t=1}^{T} Y_{it}$, $\overline{X}_{i.} = T^{-1} \sum_{t=1}^{T} X_{it}$, y $\overline{\varepsilon}_{i.} = T^{-1} \sum_{t=1}^{T} \varepsilon_{it}$, de manera que la ecuación de efectos fijos para los valores medios es

$$\bar{Y}_{i.} = \beta_1 \bar{X}_{i.} + \alpha_i + \bar{\varepsilon}_{i.}, \tag{9.9}$$

y debido a que el efecto fijo α_i es constante también aparece en la ecuación de valores medios. Se puede considerar que esta ecuación es una ecuación de sección cruzada.

Restando ambas ecuaciones para cada t, obtenemos la ecuación de efectos fijos en diferencias a las medias temporales, en la que los efectos individuales específicos han desaparecido:

$$Y_{it} - \bar{Y}_{i.} = \beta_1 \left(X_{it} - \bar{X}_{i.} \right) + \left(\varepsilon_{it} - \bar{\varepsilon}_{i.} \right), t = 1, 2, ...T. \tag{9.10}$$

Estimar por MCO esta última ecuación para $t = 1, 2, .., T; i = 1, 2, ..., N$ conduce exáctamente al mismo estimador de efectos fijos obtenido anteriormente.

Su extensión a k variables explicativas observables es

$$Y_{it} - \bar{Y}_{i.} = \beta_1 \left(X_{it1} - \bar{X}_{i.,1} \right) + \beta_2 \left(X_{it2} - \bar{X}_{i.,2} \right) + ... + \beta_k \left(X_{itk} - \bar{X}_{i.,k} \right) + \left(\varepsilon_{it} - \bar{\varepsilon}_{i.} \right). \tag{9.11}$$

A partir de esta estimación se calculan los efectos fijos α_i, de la siguiente forma:

$$\widehat{\alpha}_i = \bar{Y}_{i.} - \hat{\beta}_1 \bar{X}_{i.,1} - \hat{\beta}_2 \bar{X}_{i.,2} - ... - \hat{\beta}_k \bar{X}_{i.,k}. \tag{9.12}$$

Se observa que en este estimador la medias aritméticas son calculadas a partir de la variación temporal observada *dentro de* cada observación, i, de sección cruzada (tanto para la variable dependiente, como para las independientes o explicativas), y por ello se denomina **estimador intragrupos** o **estimador** «**within**», siendo un estimador que por

diseño tiene en cuenta información importante sobre cómo las variables consideradas (explicativas y explicada) varían en el tiempo.

Esto contrasta con otro estimador, que no vamos a estudiar porque hay otros claramente mejores, pero que al menos vamos a enunciar. Nos referimos a un estimador que solo usa la variación entre secciones cruzadas (estimador «**between**») y consiste en estimar los coeficientes por MCO desde la ecuación

$$\bar{Y}_{i.} = \alpha + \beta_1 \bar{X}_{i.,1} + ... + \beta_k \bar{X}_{i.,k} + (\alpha_i - \alpha + \bar{\varepsilon}_{i.}) \,. \tag{9.13}$$

A modo de completar esta sección, merece la pena hacer notar que cuando tenemos paneles con dos periodos temporales llegamos a los mismos estimadores utilizando cualquiera de los tres procedimientos: el que estima el modelo de las diferencias, el que estima el modelo con variables binarias y el que estima el modelo en diferencias a las medias (intragrupos). Cuando el panel tiene más de dos periodos entonces podemos estimar el modelo de efectos fijos mediante la utilización de variables binarias o mediante el estimador en diferencias a las medias, y también podemos hacerlo con el estimador de la diferencia, si bien este caso es claramente menos eficiente. Como el uso de programas especializados está generalizado y su estimación por el procedimiento de diferencias a las medias es menos tediosa (y obtenemos los mismos estimadores de «β_i»), cuando nos referimos a estimaciones de efectos fijos, en general nos estamos refiriendo a la estimación por el procedimiento de diferencias a las medias.

El siguiente recuadro recoge las características definitorias del modelos de efectos fijos.

Modelo de regresión de efectos fijos

Para cada i el modelo es

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \varepsilon_{it},$$

1. $\mathbb{E}\left(\varepsilon_{it} \left| X_{i1}, X_{i2}, ..., X_{iT}, \alpha_i \right.\right) = 0$, para $t = 1, 2, ..., T$
2. $\left(X_{i1}, X_{i2}, ..., X_{iT}, \varepsilon_{i1}, ..., \varepsilon_{iT}\right), i = 1, 2, ..., N$ son extracciones iid de su distribución conjunta
3. $(X_{it}, \varepsilon_{it})$ tienen momentos de orden cuatro finitos
4. No hay multicolinealidad perfecta

Bajo estos supuestos, los estimadores de efectos fijos son insesgados y consistentes, la estimación adecuada es MCO utilizando errores robustos a la autocorrelación y a heterocedasticidad (HAC).

Este modelo también puede incorporar efectos fijos, no solo transversales, sino también temporales. Esto facilita descomponer aún más la heterogeneidad. Esta descomposición de la heterogeneidad no observada u omitida contemplando variables que son iguales para los individuos de la sección en un periodo, pero varían a lo largo de tiempo (en los distintos periodos) tales como tipos de interés, precios, nivel de confianza en la economía, etcétera; nos referiremos a este tipo de variables por μ_t.

Esto nos permite entender la heterogeneidad no observada y omitida que reside en un término error $\varepsilon_{i,t}$ de un modelo, como si la hubiéramos descompuesto en variables

omitidas con efectos individuales, variables con efectos temporales y el resto de variables con efectos, es decir, variables no observadas con efectos temporales e individuales. De nuevo el reto consiste en controlar el efecto de las variables omitidas para estimar y realizar una correcta inferencia sobre los parámetros estructurales β_j.

Ahora introducimos una variable inobservable que varía con el tiempo pero que es constante para todas las entidades individuales. Es decir, consideramos ahora que el modelo tiene la forma siguiente

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + \varepsilon_{it}, \tag{9.14}$$

donde S_t no es observable y el subíndice t indica que todas las entidades individuales se ven afectadas por igual en cada periodo, y donde solo hay una variable explicativa, X. Si la variable S_t está correlacionada con X_{it} y no la introducimos en la ecuación obtenemos estimadores sesgados.

Podemos escribir esta ecuación en términos de efectos fijos de forma semejante a como hicimos en la expresión (9.6) añadiendo efectos fijos temporales,

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \mu_t + \varepsilon_{it}, \tag{9.15}$$

donde añadimos la variable μ_t que se mantiene constante para todas las entidades individuales y solo cambia con el tiempo, es decir, se añade un término independiente para cada periodo temporal.

También puede expresarse de una forma más compacta usando sumatorios

$$Y_{it} = \sum_{j=1}^{n} \alpha_j D_{j,it} + \sum_{s=2}^{T} \delta_s B_{s,it} + \beta_1 X_{it1} + \beta_2 X_{it2} + ... + \beta_k X_{itk} + \varepsilon_{it}$$

donde hay n variables binarias para los efectos fijos individuales iguales a la unidad si $i = j$, $(T-1)$ binarias para los efectos fijos temporales iguales a la unidad si $s = t$, y en este caso no podríamos incluir la constante pues hemos considerado directamente los n efectos fijos individuales. Recuérdese que los estimadores son consistentes para los parámetros que varían en el tiempo, y por tanto podremos estimar consistentemente los β_j y los δ_s. Por este motivo en las expresiones matriciales que incluimos en el apéndice técnico de este tema el vector $\mathbf{x}_{it}$ incorpora las $(T-1)$ variables binarias relativas a los coeficientes δ_s.

La estimación se realiza por el procedimiento en diferencias a las medias de un panel equilibrado. En primer lugar se calcula la Y_{it} y las X_{it}, en desviaciones a las medias individuales y temporales, y posteriormente estimamos la ecuación en desviaciones a las medias por MCO. El estimador en diferencias a las medias es

$$\begin{aligned}\left(Y_{it} - \overline{Y}_{i.} - \overline{Y}_{.t} + \overline{Y}_{..}\right) = \beta_1 \left(X_{it1} - \overline{X}_{i.,1} - \overline{X}_{.t1} + \overline{X}_{..,1}\right) + \beta_2 \left(X_{it2} - \overline{X}_{i.,2} - \overline{X}_{.t2} + \overline{X}_{..,2}\right) \\ + ... + \beta_k \left(X_{itk} - \overline{X}_{i.,k} - \overline{X}_{.tk} + \overline{X}_{..,k}\right) + \left(\varepsilon_{it} - \overline{\varepsilon}_{i.} - \overline{\varepsilon}_{.t} + \overline{\varepsilon}_{..}\right)\end{aligned} \tag{9.16}$$

donde $\overline{Y}_{..} = (nT)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} y_{it}$ y $\overline{\varepsilon}_{..}, \overline{X}_{..,j}$ se definen de forma equivalente. El motivo por el que es necesario hacer estas transformaciones es para asegurar que desaparecen los términos de los efectos temporales e individuales

En términos generales, con independecia del dipo de efectos fijos considerados, la expresión matricial del modelo (9.11), en términos semejantes a los que utilizamos en datos fusionados es la siguiente

$$\tilde{Y}_{it} = \tilde{\mathbf{w}}'_{it}\boldsymbol{\beta} + \tilde{\varepsilon}_{it}.$$

Podemos a partir de esta expresión colocar las observaciones temporales de cada agente (como hicimos en el modelo apilado)

$$\tilde{\mathbf{y}}_i = \tilde{\mathbf{W}}_i\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}_i$$

donde $\tilde{\mathbf{y}}_i$ es un vector $T \times 1$, al igual que $\tilde{\boldsymbol{\varepsilon}}_i$, y $\tilde{\mathbf{W}}_i$ será la matriz con T filas y el número de columnas indicativo de los regresores que varían en el tiempo, digamos k. Es posible compactar más aún la expresión matricial simplemente apilando los N individuos o agentes

$$\tilde{\mathbf{y}} = \tilde{\mathbf{W}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$$

con sus respectivas dimensiones $NT \times 1$,$NT \times k$,$k \times 1$,$NT \times 1$.

El estimador del modelo de efectos fijos (EF), tal y como explicamos en el texto principal, es el estimador MCO de este último modelo

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{EF} &= \left(\tilde{\mathbf{W}}'\tilde{\mathbf{W}}\right)^{-1}\tilde{\mathbf{W}}'\tilde{\mathbf{y}} \\ &= \left(\sum_{i=1}^{N}\tilde{\mathbf{W}}'_i\tilde{\mathbf{W}}_i\right)^{-1}\sum_{i=1}^{N}\tilde{\mathbf{W}}_i\tilde{\boldsymbol{y}}_i.\end{aligned}$$

Desde esta expresión podemos comprobar las condiciones para la consistencia simplemente mediante el álgebra habitual

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{EF} &= \boldsymbol{\beta} + \left(\tilde{\mathbf{W}}'\tilde{\mathbf{W}}\right)^{-1}\tilde{\mathbf{W}}'\tilde{\boldsymbol{\varepsilon}} \\ &= \boldsymbol{\beta} + \left(\sum_{i=1}^{N}\tilde{\mathbf{W}}'_i\tilde{\mathbf{W}}_i\right)^{-1}\sum_{i=1}^{N}\tilde{\mathbf{W}}_i\tilde{\boldsymbol{\varepsilon}}_i.\end{aligned}$$

Dada la independencia a lo largo de los individuos i, la condición fundamental para que la consistencia es que $\mathbb{E}\left(\tilde{\mathbf{W}}_i\tilde{\boldsymbol{\varepsilon}}_i\right) = 0$. Una condición suficiente para ello es precisamente $\mathbb{E}\left(\varepsilon_{it}\,|X_{i1}, X_{i2}, ..., X_{iT}, \alpha_i\right) = 0$.

La varianza asintótica es entonces

$$\widehat{Var(\hat{\boldsymbol{\beta}}_{EF})} = \left[\sum_{i=1}^{N}\tilde{\mathbf{W}}'_i\tilde{\mathbf{W}}_i\right]^{-1}\sum_{i=1}^{N}\tilde{\mathbf{W}}'_i\hat{\tilde{\boldsymbol{\varepsilon}}}_i\hat{\tilde{\boldsymbol{\varepsilon}}}_i{}'\tilde{\mathbf{W}}_i\left[\sum_{i=1}^{N}\tilde{\mathbf{W}}'_i\tilde{\mathbf{W}}_i\right]^{-1}$$

donde $\hat{\tilde{\boldsymbol{\varepsilon}}}_i = \tilde{\mathbf{y}}_i - \tilde{\mathbf{W}}_i\hat{\boldsymbol{\beta}}_{EF}$, por lo que es un estimador que solo requiere independencia entre las entidades, pero acepta que tanto $Var(\varepsilon_{it})$ como $cov(\varepsilon_{it}, \varepsilon_{is})$ varíe con i, t, s.

9.3 Modelo de efectos aleatorios y su estimación

La gran ventaja de la estimación por efectos fijos es que las variables no observables individuales α_i pueden estar correlacionadas con las variables explicativas X_{itj}, es decir, el modelo de efectos fijos permite que la heterogeneidad individual no observada pueda estar correlacionada con los regresores.

Pues bien, si estamos dispuestos a sostener (porque el tipo de análisis o estudio que estamos realizando lo permite) la restricción de que estas variables, α_i, no están correlacionadas con el resto de variables explicativas X_{itj}, entonces los estimadores de efectos fijos, que son (y seguirían siendo) consistentes, pueden mejorar en su eficiencia. Este supuesto generalmente no es siempre posible. Por ejemplo, si un panel está conformado por observaciones individuales de trabajadores, una variable observable y de interés habitual es el salario por hora del trabajador. Este salario puede estar correlacionado fácilmente con una variable no observable como, por ejemplo, las habilidades del trabajador en cuestión, que implícitamente está incorporada en el error específico individual α_i, por lo que entonces el error podrá estar correlacionado con otras variables explicativas con las que correlacione la habilidad, como puede ser el nivel educativo alcanzado, entre otras.

En todo caso, si estamos en condiciones de asumir dicha restricción, entonces la forma de conseguir estimadores eficientes en estas condiciones es recurrir al modelo de efectos aleatorios. Si bien, en caso de que realmente hubiera correlación entre α_i y X_{itj}, el modelo de efectos aleatorios dejaría de producir estimadores consistentes. Así pues, si se cumple que

$$\text{cov}\,(X_{itj}, \alpha_i) = 0,\ t = 1, 2, ..., T,\ j = 1, 2, ..., k \tag{9.17}$$

junto con los supuestos ya aludidos para el modelo de efectos fijos, podemos estimar eficientemente los coeficientes β_{itj} mediante el estimador de efectos aleatorios que exponemos a continuación.

El modelo de efectos aleatorios considera, además de la incorrelación indicada en (9.17), que el término error está compuesto de la siguiente forma, $v_{it} = \alpha_i + \varepsilon_{it}$, donde α_i y ε_{it} son variables aleatorias *iid* con media y varianza definidas:

$$\alpha_i \sim [\alpha, \sigma_\alpha^2], \varepsilon_{it} \sim [0, \sigma_\varepsilon^2].$$

De esta manera podemos escribir el modelo como

$$\begin{aligned} Y_{it} &= \beta_0 + \beta_1 X_{it1} + \beta_2 X_{it2} + ... + \beta_k X_{itk} + (\alpha_i + \varepsilon_{it}) \\ &= \beta_0 + \beta_1 X_{it1} + \beta_2 X_{it2} + ... + \beta_k X_{itk} + v_{it}. \end{aligned} \tag{9.18}$$

Modelo de regresión de efectos aleatorios

$$Y_{it} = \beta_0 + \beta_1 X_{it1} + \beta_2 X_{it2} + ... + \beta_k X_{itk} + v_{it},$$

1. $\mathbb{E}\left(\varepsilon_{it} \left| X_{i1}, X_{i2}, ..., X_{iT}, \alpha_i \right.\right) = 0, t = 1, 2, ..., T$
2. $\mathbb{E}\left(\alpha_i \left| X_{i1}, X_{i2}, ..., X_{iT} \right.\right) = \mathbb{E}(\alpha_i) = 0$
3. $\left(X_{i1}, X_{i2}, ..., X_{iT}, \varepsilon_{i1}, ..., \varepsilon_{iT}\right), i = 1, 2, ..., N$ son extracciones iid de la distribución conjunta
4. $(X_{it}, \varepsilon_{it})$ tienen momentos de orden cuatro finitos
5. No hay multicolinealidad perfecta

El supuesto o característica 2 del modelo de efectos aleatorios es nuevo respecto del de efectos fijos. Este supuesto evita la existencia de correlación entre el efecto no observado invariante en el tiempo, α_i, y las variables explicativas. Debido a que hemos incluido los efectos individuales invariantes en el tiempo en el término de error v_{it}, este ahora presenta autocorrelación:

$$\mathbb{E}(v_{it}^2) = \mathbb{E}\left(\alpha_i + \varepsilon_{it}\right)^2 = \mathbb{E}\alpha_i^2 + \mathbb{E}\varepsilon_{it}^2 + 2\mathbb{E}\left(\alpha_i \varepsilon_{it}\right)$$

que por el supuesto 1 del cuadro de referencia se puede reducir a

$$\mathbb{E}(v_{it}^2) = \mathbb{E}\alpha_i^2 + \mathbb{E}\varepsilon_{it}^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2.$$

Por otra parte, la aplicación de los suspuestos nos facilita el desarrollo de la covarianza ($t \neq s$)

$$cov(v_{it}, v_{is}) = E(v_{it} v_{is}) - E(v_{it})E(v_{is}) = E(v_{it} v_{is})$$

que desarrollando conduce

$$cov(v_{it}, v_{is}) = E(v_{it} v_{is}) = E[(\alpha_i + \varepsilon_{it})(\alpha_i + \varepsilon_{is})] = \sigma_\alpha^2.$$

Y por tanto la autocorrelación es

$$\text{Corr}\left(v_{it}, v_{is}\right) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2},\ t \neq s \tag{9.19}$$

donde, como hemos dicho, $\sigma_\alpha^2 = \text{Var}\left(\alpha_i\right)$ y $\sigma_\varepsilon^2 = \text{Var}\left(\varepsilon_{it}\right)$. Se aprecia que esta correlación es el ratio de la varianza de α_i sobre la del error compuesto v_{it}, por lo que mide la importancia relativa de los efectos invariantes α_i.

En este nuevo modelo la estimación MCO claramente producirá errores estándar incorrectos y para corregir la autocorrelación debemos recurrir el estimador por el método de mínimos cuadrados generalizados (MCG). La transformación utilizada parte de

$$\lambda = 1 - \frac{\sigma_\varepsilon^2}{\left(\sigma_\varepsilon^2 + T\sigma_\alpha^2\right)^{1/2}}, \tag{9.20}$$

cuyo valor está entre cero y uno. De tal manera que la ecuación transformada (MCG) para calcular los estimadores de efectos aleatorios es

$$\begin{aligned} Y_{it} - \lambda \bar{Y}_{i.} = \beta_0 (1-\lambda) + \beta_1 \left(X_{it1} - \lambda \bar{X}_{i.,1} \right) + \beta_2 \left(X_{it2} - \lambda \bar{X}_{i.,2} \right) + ... \\ + \beta_k \left(X_{itk} - \lambda \bar{X}_{i.,k} \right) + \left(v_{it} - \lambda \bar{v}_{i.} \right), \end{aligned} \tag{9.21}$$

que es una estimación en cuasidiferencias a las medias temporales.

Una de las ventajas del estimador de efectos aleatorios es que permite incluir variables explicativas binarias (variables que no se pueden incluir en el estimador de efectos fijos).

El parámetro λ no se conoce en la práctica por lo que recurrimos al estimador mínimos cuadrados generalizados factibles. En general los programas especializados calculan rutinariamente el estimador de efectos aleatorios.

Cuando en la expresión (9.21) $\lambda = 0$ la estimación de efectos aleatorios y MCO (sin incluir efectos fijos) coinciden (es decir, el modelo de datos fusionados). En cambio cuando $\lambda = 1$ los estimadores de efectos aleatorios y de efectos fijos coinciden. Además y puesto que los estimadores de efectos fijos son consistentes incluso bajo los supuestos de efectos aleatorios a medida que el número de observaciones aumenta, λ tiende asintóticamente a 1 $(NT \to \infty)$.

El modelo de efectos aleatorios individuales y temporales es similar. Ahora el error compuesto es $v_{it} = \alpha_i + \eta_t + \varepsilon_{it}$, donde hemos añadido los efectos fijos temporales. La estimación de efectos aleatorios individuales y temporales requieren que tanto los errores individuales como los temporales no estén correlacionados con las variables explicativas y de igual forma tenemos que recurrir para su estimación a MCGF, la estrategia para calcularlo es similar a la de efectos fijos individuales pero más compleja[3]. Por suerte los programas especializados calculan este estimador de forma rutinaria.

Es posible estimar efectos aleatorios individuales y efectos fijos temporales y viceversa. Para ello seguimos los mismos pasos que en el caso del modelo de efectos fijos, llegamos a expresiones matriciales similares a las obtenidas anteriormente, siempre que realicemos las transformaciones adecuadas. En este caso tendremos un modelo

$$\check{Y}_{it} = \check{\mathbf{w}}'_{it} \boldsymbol{\beta} + \check{\varepsilon}_{it}$$

donde $\check{Y}_{it} = Y_{it} - \lambda \bar{Y}_i$,$\check{\mathbf{w}}_{it} = \mathbf{w}_{it} - \lambda \mathbf{w}_{it}$; y donde λ se estimará a partir de los estimadores muestrales de las varianzas correspondientes, como indicaremos más adelante. Con el modelo así formulado, podemos colocar las observaciones temporales de cada agente (como hicimos en el modelo apilado)

$$\check{\mathbf{y}}_i = \check{\mathbf{W}}_i \boldsymbol{\beta} + \check{\boldsymbol{\varepsilon}}_i$$

donde $\check{\mathbf{y}}_i$ es un vector $T \times 1$, al igual que $\tilde{\varepsilon}_i$, y $\check{\mathbf{W}}_i$ será la matriz con T filas y el número de columnas indicativo de todos los regresores, digamos q. Es posible compactar más aún la expresión matricial simplemente apilando los N individuos o agentes

$$\check{\mathbf{y}} = \check{\mathbf{W}} \boldsymbol{\beta} + \check{\boldsymbol{\varepsilon}}$$

[3] La transformación es, para la variable explicada: $Y^*_{it} = \left(Y_{it} - \theta_1 \bar{Y}_{i.} - \theta_2 \bar{Y}_{.t} - \theta_3 \bar{Y}_{..} \right)$, con $\theta_1 = 1 - \frac{\sigma_v}{\sqrt{T\sigma^2_\alpha + \sigma^2_v}}$; $\theta_2 = 1 - \frac{\sigma_v}{\sqrt{N\sigma^2_\eta + \sigma^2_v}}$; $\theta_1 = 1 - \frac{\sigma_v}{\sqrt{T\sigma^2_\alpha + N\sigma^2_\eta + \sigma^2_v}}$. Las transformaciones para las variables explicativas y el error son similares.

con sus respectivas dimensiones $NT \times 1$,$NT \times q$,$q \times 1$,$NT \times 1$.

El estimador del modelo de efectos aleatorios (RE), tal y como explicamos en el texto principal, es el estimador MCO de este último modelo

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{RE} &= \left(\check{\mathbf{W}}'\check{\mathbf{W}}\right)^{-1}\check{\mathbf{W}}'\check{\mathbf{y}} \\ &= \left(\sum_{i=1}^{N}\check{\mathbf{W}}_i'\check{\mathbf{W}}_i\right)^{-1}\sum_{i=1}^{N}\check{\mathbf{W}}_i\check{\mathbf{y}}_i.\end{aligned}$$

Desde esta expresión podemos comprobar las condiciones para la consistencia simplemente mediante el álgebra habitual

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{RE} &= \boldsymbol{\beta} + \left(\check{\mathbf{W}}'\check{\mathbf{W}}\right)^{-1}\check{\mathbf{W}}'\check{\boldsymbol{\varepsilon}} \\ &= \boldsymbol{\beta} + \left(\sum_{i=1}^{N}\check{\mathbf{W}}_i'\check{\mathbf{W}}_i\right)^{-1}\sum_{i=1}^{N}\check{\mathbf{W}}_i\check{\boldsymbol{\varepsilon}}_i.\end{aligned}$$

Dada la independencia a lo largo de los individuos i, la condición fundamental para la consistencia es que $\mathbb{E}\left(\tilde{\mathbf{W}}_i\tilde{\boldsymbol{\varepsilon}}_i\right) = 0$, que está garantizada si el modelo es de efectos aleatorios.

La varianza asintótica es entonces

$$\widehat{Var(\hat{\boldsymbol{\beta}}_{RE})} = \left[\sum_{i=1}^{N}\check{\mathbf{W}}_i'\check{\mathbf{W}}_i\right]^{-1}\sum_{i=1}^{N}\check{\mathbf{W}}_i'\widehat{\check{\boldsymbol{\varepsilon}}}_i\widehat{\check{\boldsymbol{\varepsilon}}}_i'\check{\mathbf{W}}_i\left[\sum_{i=1}^{N}\check{\mathbf{W}}_i'\check{\mathbf{W}}_i\right]^{-1}$$

donde $\widehat{\check{\boldsymbol{\varepsilon}}}_i = \check{\mathbf{y}}_i - \check{\mathbf{W}}_i\hat{\boldsymbol{\beta}}_{RE}$, por lo que es un estimador que solo requiere independencia entre las entidades, pero acepta que tanto $Var(\varepsilon_{it})$ como $cov(\varepsilon_{it}, \varepsilon_{is})$ varíe con i, t, s.

Cualquiera de estas estimaciones requiere que se estimen consistentemente las varianzas $\sigma_\alpha^2 = \text{Var}(\alpha_i)$ y $\sigma_\varepsilon^2 = \text{Var}(\varepsilon_{it})$, y así poder estimar λ. Los programas informáticos especializados en econometría y que ofrezcan la estimación con datos en forma de panel obtienen dichas estimaciones consistentes de

$$\hat{\sigma}_\varepsilon^2 = (N(T-1)-k)^{-1}\sum_i\sum_t\left[(Y_{it}-\bar{Y}_{i.}) - (\mathbf{x}_{it}-\mathbf{x}_i)'\hat{\boldsymbol{\beta}}_{EF}\right]^2.$$

Este estimador se utiliza para estimar la varianza σ_α^2. La podemos obtener a partir del vector estimado $\hat{\boldsymbol{\beta}}_B$ de la regresión del modelo que hemos denominado «between» en la ecuación (9.13), cuyo término error tiene una varianza de $\sigma_\alpha^2 + \sigma_\varepsilon^2/T$. Así pues un estimador consistente será

$$\hat{\sigma}_\alpha^2 = (N-(k+1))^{-1}\sum_i\left(\bar{Y}_i - \hat{\alpha}_B - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_B\right)^2 - (1/T)\hat{\sigma}_\varepsilon^2.$$

9.4 Modelos de efectos fijos vs. modelos de efectos aleatorios

No hay una regla sencilla que nos ayude a decidir entre el estimador de efectos aleatorios y efectos fijos y ante la duda lo más sencillo es utilizar el estimador de efectos fijos, puesto que estos son también consistentes bajo los supuestos de efectos aleatorios, lo que no ocurre a la inversa, es decir si los efectos fijos están correlacionados con las variables explicativas entonces el estimador de efectos aleatorios es sesgado e inconsistente.

En los trabajos aplicados en muchas ocasiones se decide utilizar efectos fijos o aleatorios en función de si los efectos fijos son considerados como parámetros a estimar o como resultados de una variable aleatoria. Cuando los datos no pueden considerarse como una muestra aleatoria de una población grande es usual decantarse también por el estimador de efectos fijos.

El test tipo Hausman establece la siguiente estrategia para contrastar la hipótesis nula de efectos aleatorios individuales $[\mathrm{H}_0 : \mathbb{E}\left(\alpha_i \left| X_{i1}, X_{i2}, ..., X_{iT}\right.\right) = \mathbb{E}(\alpha_i) = 0]$ comparando los estimadores de efectos fijos (FE) y efectos aleatorios (RE) a partir del siguiente estadístico

$$Q_{FE,RE} = \left(\widehat{\boldsymbol{\beta}}_{FE} - \widehat{\boldsymbol{\beta}}_{RE}\right)' \left(\widehat{\sigma}^2_{\widehat{\boldsymbol{\beta}}_{FE}} - \widehat{\sigma}^2_{\widehat{\boldsymbol{\beta}}_{RE}}\right)^{-1} \left(\widehat{\boldsymbol{\beta}}_{FE} - \widehat{\boldsymbol{\beta}}_{RE}\right), \tag{9.22}$$

que no es más que el cociente del cuadrado de las diferencias de los estimadores y las diferencias entre la matriz de varianzas y covarianzas. El test de Hausman converge a una distribución χ^2_k.

La idea con la que se construye el test consiste en aprovechar que tanto el estimador de efectos aleatorios como el de efectos fijos son consistentes si no hay correlación entre las variables explicativas $X_{it,j}$ y α_i. Si ambos son consistentes entonces deberían converger a verdadero valor del parámetro β_j. Es decir, para muestras grandes las estimaciones deberían ser similares, por lo que la diferencia entre ambos valores estimados debe ser pequeña (al menos asintóticamente). Por otra parte, en caso de correlación entre $X_{it,j}$ y α_i, el estimador de efectos aleatorios sabemos que es inconsistente, mientras que el de efectos fijos sigue siendo consistente, por lo que este último estimador convergerá a los verdaderos valores de los parámetros, mientras que el de efectos aleatorios no lo hará. En tal caso, esperamos apreciar diferencias estadísticas significativas entre ambas estimaciones, constituyendo esta diferencia evidencia en contra de la hipótesis nula anteriormente señalada. Así, al haber evidencia en contra del supuesto $\mathbb{E}\left(\alpha_i \left| X_{i1}, X_{i2}, ..., X_{iT}\right.\right) = \mathbb{E}(\alpha_i) = 0$, sería preferible que el modelo se estimara con el estimador de efectos fijos.

El test así planteado contrasta una hipótesis nula conjunta al comparar todos los coeficientes estimables. En ocasiones estamos interesados en un solo coeficiente del modelo, en tal caso es posible plantear una versión del test de Hausman a través de un ratio de la t para dicho parámetro o coeficiente. El estadístico tipo t también compara la diferencia entre las estimaciones de cada estimación respectiva de un solo coeficiente, digamos el

coeficiente k-ésimo

$$t = \frac{\hat{\beta}_{k,(FE)} - \hat{\beta}_{k,(RE)}}{\left[\widehat{var}(\hat{\beta}_{k,(FE)}) - \widehat{var}(\hat{\beta}_{k,(RE)})\right]^{1/2}} = \frac{\hat{\beta}_{k,(FE)} - \hat{\beta}_{k,(RE)}}{\left[\widehat{\sigma}^2_{\widehat{\boldsymbol{\beta}}_{FE}} - \widehat{\sigma}^2_{\widehat{\boldsymbol{\beta}}_{RE}}\right]^{1/2}}$$

cuya distribución asintótica es la normal estándar.

El test de Hausman se aplica de forma similar también para efectos individuales y temporales o solo de efectos temporales. Normalmente los programas especializados realizan el test de Hausman de forma rutinaria.

El modelo de efectos fijos tiene el atractivo de que permite estudiar los efectos parciales e incluso causales de las variables explicativas sobre la variable dependiente con supuestos más flexibles (menos restrictivos) que los que se necesitan para establecer una relación causal con datos de sección cruzada o con modelos de paneles sin efectos fijos, como es el caso del modelo de efectos aleatorios y, lógicamente, también el de datos fusionados. Salvo que el esquema de causas esté muy claro y los datos se hayan obtenido de un experimento controlado (o las circunstancias del mismo estén cercanas a ser un experimento controlado), es preferible utilizar el modelo de efectos fijos en la medida en que estemos interesados en medir relaciones causales.

Lógicamente la disyuntiva entre efectos fijos y aleatorios se presenta porque los efectos fijos tienen algunas desventajas. La más relevante es que la estimación de los coeficientes de regresores que sean invariantes en el tiempo no es posible y quedará absorbida dentro del coeficiente del efecto fijo individual. Esto provoca que únicamente podamos hacer previsiones (a partir del modelo) sobre la variación la media condicionada a partir de cambios en los regresores que varían a lo largo del tiempo. Por estos motivos, incluso al coste de que el análisis causal no quede garantizado, también utilizamos el modelo de efectos aleatorios.

Bibliografía complementaria

Matilla-García, M et al. 2017. Econometría y Predicción. McGraw Hill

Stock J. and Watson J. Introducción a la econometría. Pearson.

Tema 10

Procesos estocásticos estacionarios

Este tema está elaborado como una adaptación del capítulo 1 (apartado 1) y capítulo 2 de:

Enders, W. Applied Econometric Time Series. 4 ed. Wiley, y del capítulo 2 de:

Box, Jenkins, Reinsel y Ljung Time Series, Analysis, Forecasting and Control. 5th. ed. Wiley. Capítulo 2. Así como de la bibliografía complementaria

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al Órgano convocante ni al Tribunal actuante.

- Series temporales.
- Procesos estocásticos estacionarios en sentido estricto y débil.
- Medias, varianzas y autocovarianzas, ergodicidad.
- Ruido blanco.
- El espectro y su estimación.
- Procesos AR y MA.
- Procesos ARIMA y SARIMA

10.1 Series temporales

Los modelos que vamos a estudiar en este tema son modelos *univariantes* porque estudian el comportamiento de una serie temporal basándose únicamente en el análisis de su propia historia. Los pronósticos se fundamentan en la hipótesis de que las condiciones futuras serán análogas a las pasadas por lo que en buena medida son especialmente adecuados para la predicción a corto o muy corto plazo, y más cuestionables en escenarios temporales de mayor plazo.

Los modelos ARIMA históricamente se han establecido como una herramienta potente para la predicción de series temporales y también son un camino convergente con los modelos dinámicos (de naturaleza económica) que se ven en otras partes de este temario.

El análisis de series temporales que se va a presentar a continuación es el propio o el característico de la economía, en el sentido que está orientado hacia el tipo de preguntas

y datos temporales propios de las series económicas. En efecto, el análisis de series temporales es fundamental para el estudio del comportamiento de la economía (en su conjunto) tanto a nivel nacional como internacional. Así por ejemplo si necesitamos hacer una predicción del crecimiento PIB o el de la inflación, miramos el comportamiento de algunos indicadores económicos y consideramos su comportamiento en el pasado reciente. De manera similar, podemos analizar la evolución reciente de una industria determinada para pronosticar el potencial de ventas de una empresa perteneciente a dicha industria. Ni que decir tiene que muchas de las decisiones de inversión financiera (ya sea corto, medio o largo plazo) se toman considerando la evolución de la cotización de uno o varios valores bursátiles, así como de la evolución de tipos de cambio, o de la senda trazada por los tipos de interés, entre otros. En cada uno de estos casos, necesitamos analizar series temporales.

Algunas características de las series económicas es que suelen ser relativamente cortas (en comparación con las disponibles en otros dominios científicos), por lo que las herramientas de análisis deben estar adecuados a esta realidad. Suelen presentar tendencias, y el tratamiento de las mismas es fundamental tanto desde el punto de vista analítico, como desde el punto de vista económico. Igualmente pasa con la existencia de ciclos estacionales, consustanciales a la realidad estudiada a través de la serie.

Un elemento singular de las series a analizar es que la teoría económica juega un papel central, en el sentido de que es realmente interesante integrar las técnicas y procedimientos de análisis basados (solo) en su propia historia con los análisis basados en una estructura teórica, lo cual es razonable por varios motivos, entre ellos porque el agente estudiado tiene acceso al análisis de la propia serie lo cual le puede reportar beneficio económico.

La mayoría de las series temporales económicas son grabadas o registradas, por organismos, en intervalos discretos de forma anual, trimestral, mensual, semanal, o diaria. El número de observaciones por año se denomina **frecuencia**.

10.2 Procesos estocásticos estacionarios en sentido estricto y débil: Medias, varianzas y autocovarianzas, ergodicidad

A diferencia del tratamiento dado a las observaciones transversales, que eran consideradas muestras aleatorias de una población subyacente, dicho tratamiento no es apropiado para las series temporales debido a la dependencia existente entre observaciones. En este contexto, tratamos la muestra observada $\{Z_1, ..., Z_T\}$ como la realización de un proceso estocástico dependiente, que definimos seguidamente. A menudo es útil ver $\{Z_1, ..., Z_T\}$ como un subconjunto de una secuencia doblemente infinita $\{..., Z_{t-1}, Z_t, Z_{t+1}, ...\}$.

Cuando observamos una serie temporal vamos a entender que esta serie es una realización de un *proceso estocástico*. Podemos definir un proceso estocástico «Z» como un conjunto de «T» variables aleatorias «Z_t» en momentos de tiempo sucesivos. Cada una de estas «T» variables se comporta como lo hacen las variables aleatorias usuales. Por tanto, la variable aleatoria Z_t se caracteriza por su función de distribución, y el conjunto $\{Z_t, Z_{t+1}, ..., Z_{t+u}\}$ se caracteriza por su distribución conjunta. Analíticamente se puede

expresar como

$$Z = \{Z_1,\ Z_2, ...,\ Z_T\}\ ;\ Z \in \{Z\,(s,t)\,; s \in S, t \in T\}\,, \tag{10.1}$$

donde «s» representa el comportamiento en el estado de los sucesos aleatorios y «t» el comportamiento en la dimensión temporal.

Es en este contexto en el que cabe interpretar las series de tiempo como realizaciones de un proceso estocástico, es decir, dado (realizado) un suceso determinado, s_0 (del espacio de sucesos), observamos la serie $Z(s_0, t)$ a lo largo del tiempo (ordenada cronológicamente)[1]. Bajo ciertas condiciones de estabilidad temporal, los datos que observamos (fruto de una realización) pueden permitir caracterizar al proceso generador de datos. En otros términos, dado que solo hay una muestra o realización del proceso, si queremos aprender algo sobre estas distribuciones necesitaremos cierto grado de constancia o regularidad, como veremos.

Lógicamente un proceso estocástico tiene que tener una función de distribución conjunta, del tipo habitual,

$$F\,(Z) = F\,\{Z_1, Z_2, ..., Z_T\} = \Pr\,\{Z_1 < z_1, Z_2 < z_2, ..., Z_T < z_T\}\,, \tag{10.2}$$

pero normalmente en un proceso estocástico solo conocemos un valor de cada una de las «T» variables que componen el proceso (o punto muestral), y en consecuencia no podemos conocer su función de distribución conjunta, que puede ser muy compleja.

Kolgomorov demostró que si se cumplen las condiciones de **simetría** (cuando la permutación temporal de las variables del proceso no afecta a su distribución conjunta) y **compatibilidad** (cuando el proceso estocástico se puede reducir mediante marginalización al análisis de un conjunto finito de elementos), entonces no es necesario conocer la función de distribución conjunta para poder hacer inferencia estadística. Ambas condiciones se cumplen si el proceso estocástico es *estacionario*, un concepto que definiremos inmediatamente.

Para analizar teóricamente los procesos estocásticos definimos los siguientes momentos de las distribuciones marginales:

$$\mathbb{E}\,(Z_t) = \mu_t,\ \text{para}\ t = 1, 2, ..., T, \tag{10.3}$$

que no es más que la esperanza no condicionada de las variables aleatorias (o **esperanza marginal**) de que consta el proceso.

La **varianza marginal** es

$$\text{var}\,(Z_t) = \sigma_t^2,\ para\ t = 1, 2, ..., T. \tag{10.4}$$

[1] La función real $Z(s,t)$ es un proceso estocástico que depende del tiempo y del suceso. Si fijamos el tiempo, $Z(s, t_0)$ define una variable aleatoria, si fijamos tiempo y suceso, $Z(s_0, t_0)$ define un número real.

La dependencia entre las variables aleatorias del proceso estocástico se representa por las funciones de covarianza y correlación entre dos variables del proceso en dos instantes cualesquiera:

$$\gamma(t, t+u) = \text{cov}(Z_t, Z_{t+u}) = \mathbb{E}\left[(Z_t - \mu_t)(Z_{t+u} - \mu_{t+u})\right] \tag{10.5}$$

es la covarianza, que denominaremos **función de autocovarianza**, puesto que se refiere a la covarianza de dos variables cualesquiera del proceso en distintos momentos de tiempo. u representa el retardo (o adelanto en el tiempo) respecto de Z_t. Si el desfase es nulo, $u = 0$, y entonces obtenemos de nuevo la varianza del proceso.

El coeficiente de correlación, que denominaremos **función de autocorrelación**, mide la correlación entre dos variables del proceso en distintos momentos de tiempo,

$$\rho(t, t+u) = \frac{\text{cov}(Z_t, Z_{t+u})}{\sigma_t \sigma_{t+u}} = \frac{\gamma(t, t+u)}{\gamma^{1/2}(t,t)\,\gamma^{1/2}(t+u, t+u)}, \tag{10.6}$$

y como cualquier correlación su valor está acotado entre 1 y -1.

La forma más común de asumir cierta regularidad o constancia es la de estacionariedad. Un proceso estocástico es estacionario, en **sentido estricto**, si las funciones de distribución de las variables aleatorias que lo componen son idénticas, es decir, si

$$F(Z_t) = F(Z_{t+u}), \tag{10.7}$$

donde las funciones de distribución marginal (o de cada una de las variables del proceso) son iguales, lo que permite considerarlo de hecho, como si fuera una única variable aleatoria con «T» repeticiones. La constancia o regularidad anteriores se refieren en este caso a que la distribución de Z_t no cambia en el tiempo. Tampoco cambiarían la distribuciones bivariantes (Z_t, Z_{t+1}), ni las multivariantes $(Z_t, ..., Z_{t+u})$.

El conocimiento de las funciones de distribución de las variables que componen el proceso resulta inalcanzable si, como es habitual, solo tenemos una realización de cada una de las variables. Para solventar este problema suele recurrirse al concepto de **proceso estocástico estacionario en sentido débil**.

Un proceso es **estacionario** en sentido **débil** si

$$\begin{aligned} &1.^{\text{a}}\ \mu_t = \mu \text{ para todo } t, \\ &2.^{\text{a}}\ \sigma_t^2 = \sigma^2 \text{ para todo } t, \\ &3.^{\text{a}}\ \gamma(t, t+u) = \gamma(t, t-u) = \gamma_u = \gamma_{-u}. \end{aligned} \tag{10.8}$$

La constancia o regularidad, en este caso, significa que el proceso estocástico es estable a lo largo del tiempo en media y varianza constantes para todo t y la función de autocovarianza solo depende del desfase temporal u. Esta última condición también se puede escribir como $\rho_u = \rho_{-u}$. A estos procesos también se les conoce como procesos *estacionarios en covarianza*.

De las definiciones y de los ejemplos de series podemos ver que la estacionariedad significa que la distribucion es constante a lo largo del tiempo. No significa, sin embargo, que el proceso tenga limitaciones en cuanto a la dependencia, y tampoco que haya serias limitaciones en cuanto a la ausencia de patrones periódicos.

Una propiedad importante por su utilidad de los procesos estacionarios es que la estacionariedad se preserva tras realizar transformaciones que incluyan la historia entera de Z_t:

Si Z_t es un proceso estrictamente estacionario y $x_t = f(z_t, z_{t-1}, ...)$ es una variable aleatoria, entonces x_t es estrictamente estacionario.

Por ejemplo una transformación como esta

$$x_t = \sum_{j=0}^{\infty} a_j Z_{t-j}$$

donde a_j son coeficientes y Z_t es estacionario, genera un proceso x_t que también es estacionario.

Hemos visto en los apartados anteriores que los procesos estacionarios en covarianza se caracterizan por los primeros momentos de su distribución. Veamos bajo qué condiciones podemos estimarlos y hacer inferencia a partir de las correspondientes estimaciones.

Supongamos, como es habitual, que contamos con una única realización de un proceso estocástico estacionario, con media $\mu = \mathbb{E}(Z_t)$, varianza $\sigma_z^2 = \gamma_0 = \text{var}(Z_t)$ y autocovarianzas $\gamma_u = \text{cov}(Z_t, Z_{t-u})$.

En estas condiciones un estimador del primer momento, es decir, de la media poblacional, es la media muestral, que será una media temporal,

$$\bar{Z} = \frac{\sum_{t=1}^{T} Z_t}{T}. \tag{10.9}$$

Una forma alternativa de estimar la media (esperanza) del proceso sería obtener diferentes realizaciones del proceso y ensamblar entonces la media a partir de las mismas. Esto, sin embargo, supondría repetir la historia más de una vez, lo cual no es posible con los datos económicos. En consecuencia, aspiramos a poder estimar la media poblacional a partir de la media muestral *temporal*.

Consideremos que tenemos datos independientes, en ese caso la varianza de la media muestral es $\text{var}\left(\bar{Z}\right) = \sigma^2/T$, y por tanto al aumentar el tamaño de la muestra, el *error cuadrático medio* $\mathbb{E}\left(\bar{Z} - \mu\right)^2$ de la estimación tiende a cero, lo cual es deseable. Desafortunadamente, esta convergencia no está garantizada para todo proceso estacionario. Por ejemplo, un proceso con media cero y varianza σ^2 tan simple como $Z_1 = Z_2 = Z_3 = ...$, es estacionario sin embargo, pese a que la esperanza $\mathbb{E}(Z) = 0$, cuando tengamos una única realización $Z_1 = Z_2 = Z_3 = ... = Z_T$, la varianza de la media muestral (temporal),

es decir,

$$\begin{aligned}\text{var}\left(\bar{Z}\right) &= \frac{1}{T^2}\text{var}\left(\sum_{t=1}^{T} Z_t\right) \\ &= \frac{1}{T^2}\left[\text{var}\left(Z_1\right) + ... + \text{var}\left(Z_T\right) + 2\text{cov}\left(Z_1, Z_2\right) + 2\text{cov}\left(Z_1, Z_3\right)\right. \\ &\quad +... + 2\text{cov}\left(Z_1, Z_T\right) + 2\text{cov}\left(Z_2, Z_3\right) + ... + 2\text{cov}\left(Z_2, Z_T\right) + \\ &\quad \left.+... + 2\text{cov}\left(Z_{T-1}, Z_T\right)\right],\end{aligned}$$

no será asintóticamente nula. La correlación entre una observación y la siguiente, que es idéntica por la definición del proceso, es unitaria y por tanto las covarianzas de la expresión serán constantes a lo largo del tiempo. Lo fundamental es percatarse de que en este proceso en particular cada nueva observación, al ser idéntica a la anterior, no proporciona nueva información, es decir, tiene una dependencia muy fuerte respecto de la anterior, y esta dependencia no se atenúa con el paso del tiempo. Este tipo de dependencias entre las observaciones no permite que la media muestral colapse asintóticamente con la esperanza o media poblacional, μ, incluso si el proceso es estacionario, cuya media poblacional es por definición constante. Estas situaciones han de ser descartadas porque nos conducen a estimaciones erróneas del primer momento (media poblacional).

Para garantizar la convergencia, es necesaria una condición más: que el proceso sea **ergódico**. Si el proceso es ergódico la media muestral temporal nos conduce asintóticamente a la media poblacional.

Decimos que un **proceso es ergódico** para la estimación de la media cuando se cumple que

$$\lim_{T\to\infty} \mathbb{E}\left(\bar{Z} - \mu\right)^2 \to 0. \tag{10.10}$$

Esta es una propiedad importante porque garantiza que la diferencia entre el estimador de la media $\bar{Z}$ y su verdadero valor μ tiende a cero cuando T aumenta, es decir, $\text{var}\left(\bar{Z}\right) \to 0$. Nos interesa por tanto conocer bajo qué condiciones el proceso estacionario es ergódico para la media.

Ya hemos visto que cuando las observaciones son independientes en un proceso estacionario, el error cuadrático medio tiende a cero cuando T aumenta, sin embargo cuando las observaciones del proceso estacionario no son independientes, para calcular el error cuadrático medio debemos tener en cuenta las funciones de covarianza, es decir que

$$\begin{aligned}\text{var}\left(\bar{Z}\right) = \mathbb{E}\left(\bar{Z} - \mu\right)^2 &= \tfrac{1}{T^2}\mathbb{E}\left[\textstyle\sum_{t=1}^{T}\left(Z_t - u\right)\right]^2 \\ &= \tfrac{1}{T^2}\mathbb{E}\left[\textstyle\sum_{t=1}^{T}\mathbb{E}\left(Z_t - u\right)^2 + 2\sum_{i=1}^{T}\sum_{j=i+1}^{T}\mathbb{E}\left(\left(z_i - \mu\right)\left(z_j - \mu\right)\right)\right] \\ &= \tfrac{1}{T}\mathbb{E}\left[\sigma^2 + 2\textstyle\sum_{i=1}^{T-1}\left(1 - \frac{i}{T}\right)\gamma_i\right],\end{aligned} \tag{10.11}$$

de manera que la condición para que la varianza tienda a cero al aumentar la muestra es que el sumatorio de la última expresión converja hacia una constante. La condición suficiente (no necesaria) para que esto suceda es que

$$\lim_{u\to\infty} \gamma_u \to 0, \tag{10.12}$$

lo que se denomina **dependencia débil en covarianza** e implica que cuando aumenta el desfase la covarianza tiende a cero. En consecuencia, para que el proceso sea ergódico es condición suficiente que exista dependencia débil en covarianza.

Por tanto podemos decir que la dependencia serial en el proceso es admisible siempre que tienda a desaparecer con el tiempo. Esta observación está en cercana sintonía con lo previsto para el comportamiento de la covarianza en la versión de la Ley de los Grandes Números. Adicionalmente, conviene tener presente que para cualquier función medible f, la sucesión $\{f(Z_i)\}$ es ergódica siempre que lo sea el proceso Z_i, por lo que si un proceso es estacionario y ergódico, entonces cualquiera de sus momentos (si existen) se podrá estimar consistentemente a partir del correspondiente momento muestral.

En términos más generales, un proceso es ergódico respecto a un parámetro ξ cuando el estimador $\hat{\xi_T}$ calculado sobre una serie temporal *converge en media cuadrática* a un estimador ξ análogo definido sobre una muestra de réplicas independientes del proceso. La ergodicidad es una restricción sobre la memoria del proceso necesaria para poder estimar consistentemente las características del mismo a partir de una única realización.

La estacionariedad no garantiza la ergodicidad. Hemos visto que la memoria de un proceso se mide por la covarianza entre dos variables distanciadas u periodos. En cambio la condición de estacionariedad no implica una restricción de memoria, obsérvese que solamente afecta a la homogeneidad temporal: todas las variables distanciadas u periodos tienen una misma covarianza, γ_u, para cualquier u.

A modo de resumen, y para cerrar la cuestión de la ergodicidad, y evitar así una exposición más técnica, que excede el nivel diseñado para este manual, cabe indicar lo siguiente respecto a los requisitos de ergodicidad y estacionariedad. La estacionariedad la pensamos en términos de restricción sobre la heterogeneidad temporal del proceso, mientras que la ergodicidad limita su memoria. Se pueden concebir procesos estacionarios no ergódicos y procesos ergódicos no estacionarios, aunque, en general, la ergodicidad no suele definirse para procesos no estacionarios. El requerimiento conjunto de estacionariedad y ergodicidad asegura que con una única serie temporal se pueden obtener estimadores consistentes de los momentos poblacionales. De ambos requisitos el más fuerte es el relativo a la estacionariedad. Sería posible técnicamente relajar la estacionariedad requiriendo condiciones más fuertes que la ergodicidad. Todos los procesos presentados en este tema serán lineales y estocásticos, lo cual garantiza la ergodicidad, y por tanto no nos preocuparemos por ella.

Un elemento de enorme utilidad en un proceso estocástico estacionario y ergódico es que la *función de densidad conjunta* de un subconjunto de T variables de dicho proceso, condicional en unos valores iniciales dados, coincide con el producto de las funciones de densidad condicionales escalares (individuales) con un número finito de parámetros

constantes. Justamente esta propiedad es la que permite realizar inferencias sin la necesidad de conocer la función de distribución conjunta. De lo contrario sería casi imposible.

10.3 Ruido blanco

Un proceso estocástico estacionario de gran importancia es el denominado proceso de ruido blanco. Responde a la siguiente expresión analítica

$$Z_t = \bar{Z} + \varepsilon_t, \tag{10.13}$$

y puede ser escrito en desviaciones a las medias sin pérdida de generalidad, es decir que expresado de esta forma, el proceso **ruido blanco** sería:

Ruido blanco

$$z_t = \varepsilon_t, \tag{10.14}$$

y satisface

$$\begin{aligned} &1.^{\text{a}}\ \mathbb{E}(z_t) = 0,\ t = 1, 2, ... \\ &2.^{\text{a}}\ \text{var}(z_t) = \sigma_\varepsilon^2,\ t = 1, 2, ... \\ &3.^{\text{a}}\ \text{cov}(z_t, z_{t+u}) = \gamma_u = 0 \text{ para todo } u \neq 0. \end{aligned} \tag{10.15}$$

Consiste, por tanto, en una secuencia de variables aleatorias con media nula y varianza constante. La tercera condición también se puede escribir como: $\rho_u = 0$, para todo $u \neq 0$. Intuitivamente, y de forma menos precisa, podemos decir que en un proceso de ruido blanco, conocer los valores pasados no proporciona ninguna información sobre los valores futuros. El proceso no tiene memoria. Por tanto es evidente que el ruido blanco es un proceso estacionario débil.

A modo ilustrativo consideremos, por ejemplo, un proceso que consiste en empezar en un z_0 cualquiera. Si observamos que $z_t > 0$, entonces z_{t+1} lo extraemos de una distribución normal $N(0, 1)$; pero si observamos que $z_t < 0$, entonces z_{t+1} lo extraemos de una distribución uniforme $U\left(-1/\sqrt{3}, 1/\sqrt{3}\right)$. Este proceso estocástico es estacionario, aunque temporalmente no es independiente. También se puede comprobar fácilmente que es un proceso de ruido blanco: su media es nula, la varianza es constante, y las covarianzas son nulas. Así pues, el ruido blanco no es necesario que sea serialmente independiente. La independencia serial requeriría reemplazar la tercera condición por la siguiente

$$z_t, z_{t+u} \text{ son independientes para } u \neq 0$$

y entonces diríamos que es un proceso de *ruido blanco independiente* o *ruido blanco estricto*. Esto es así puesto que se trata de ruido blanco, la distribución es la misma a lo largo del tiempo y el proceso es temporalmente independiente, y se corresponde con las siglas i.i.d, que hemos utilizado en otros temas. Requerir que un proceso sea i.i.d. es más restrictivo que requerir que sea ruido blanco.

En los casos en los que ε_t se acomoda a la distribución normal $N(0, \sigma^2)$ decimos que se trata de *ruido blanco gausiano*. En este caso, lógicamente también sería i.i.d.

10.4 Procesos AR y MA

En general un proceso estocástico estacionario se denomina proceso autorregresivo de orden p [también denominado $AR(p)$] si el valor actual de la serie (Z_t) depende de la propia variable en retardos sucesivos desde 1 hasta p. Analíticamente un $AR(p)$ presenta la forma siguiente:

$$Z_t = \phi_0 + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + ... + \phi_p Z_{t-p} + \varepsilon_t, \tag{10.16}$$

donde ε_t es ruido blanco [$\mathbb{E}(\varepsilon_t) = 0,\ \text{var}(\varepsilon_t) = \sigma_\varepsilon^2$ y $\gamma_u = 0,\ u \neq 0$]. Los parámetros ϕ_i se estiman por MCO.

Proceso autorregresivo de primer orden AR(1)

El proceso autorregresivo más simple es el proceso $AR(1)$. Diremos que un proceso es un proceso autorregresivo de primer orden si ha sido generado a partir de la siguiente expresión:

$$Z_t = \phi_0 + \phi_1 Z_{t-1} + \varepsilon_t, \tag{10.17}$$

donde ϕ_0 y ϕ_1 son valores a determinar y ε_t ruido blanco.

Supongamos que el proceso así establecido comienza a partir de un valor cualquiera, que denominaremos por ejemplo I, de manera que el primer valor es $Z_0 = I$, el segundo valor será $Z_1 = \phi_0 + \phi_1 I + \varepsilon_1$, el siguiente, $Z_2 = \phi_0 + \phi_1 Z_1 + \varepsilon_2 = \phi_0 + \phi_1(\phi_0 + \phi_1 I + \varepsilon_1) + \varepsilon_2$, y sustituyendo así sucesivamente tenemos que

$$\begin{aligned}
&Z_1 = \phi_0 + \phi_1 I + \varepsilon_1 \\
&Z_2 = \phi_0(1 + \phi_1) + \phi_1^2 I + \phi_1 \varepsilon_1 + \varepsilon_2 \\
&Z_2 = \phi_0(1 + \phi_1 + \phi_1^2) + \phi_1^3 I + \phi_1^2 \varepsilon_1 + \phi_1 \varepsilon_2 + \varepsilon_3 \\
&\vdots \\
&Z_t = \phi_0 \textstyle\sum_{i=0}^{t-1}(\phi_1^i) + \phi_1^t I + \sum_{i=0}^{t-1} \phi_1^i \varepsilon_{t-i}.
\end{aligned} \tag{10.18}$$

Y como la $\mathbb{E}(\varepsilon_t) = 0$, la esperanza del proceso es

$$\mathbb{E}(Z_t) = \phi_0 \sum_{i=0}^{t-1} \left(\phi_1^i\right) + \phi_1^t I. \tag{10.19}$$

Para que el proceso sea estacionario en media necesitamos que el primer término converja a una constante y que el segundo se anule, lo que se consigue solo si $|\phi_1| < 1$.

En efecto, el primer término $\phi_0 \sum_{i=0}^{t-1} (\phi_1^i) = \phi_0 \left(1 + \phi_1 + \phi_1^2 + ... + \phi_1^{t-1}\right)$ es la suma de una progresión geométrica de razón ϕ_1 cuya suma converge a $\phi_0/(1-\phi_1)$, y el segundo ϕ_1^t tiende a cero a medida que t aumenta. En definitiva si $|\phi_1| < 1$, la $\mathbb{E}(Z_t)$ converge en media rápidamente a $\phi_0/(1-\phi_1)$ con independencia de las condiciones iniciales.

Al mismo resultado llegamos si partimos de la expresión (10.17). Para que la media sea constante en todo el proceso se tiene que cumplir que $\mathbb{E}(Z_t) = \mathbb{E}(Z_{t-1}) = ... = \mu$ y como la $\mathbb{E}(\varepsilon_t) = 0$, y aplicando esperanzas a la expresión (10.17) obtenemos que

$$\begin{array}{c} \mu = \phi_0 + \phi_1 \mu; \\ \mu = \frac{\phi_0}{1-\phi_1}, \end{array} \tag{10.20}$$

es decir que la media marginal es constante para todo el periodo si se cumple la expresión anterior.

Igual que hicimos con el proceso ruido blanco, normalmente los procesos $AR(1)$ se expresan, sin pérdida de generalidad, en desviaciones a las medias

$$z_t = \phi_1 z_{t-1} + \varepsilon_t, \tag{10.21}$$

donde perdemos el término constante y las variables, z_t y z_{t-1}, aparecen en minúsculas indicando que son variables centradas o en diferencias a las medias.

También es usual utilizar el operador de retardos cuya definición[2] es

$$B^p z_t = z_{t-p}, \tag{10.22}$$

por lo que el proceso $AR(1)$ utilizando el operador de retardos es

$$\begin{array}{c} z_t = \phi_1 B z_t + \varepsilon_t; \\ (1 - \phi_1 B) z_t = \varepsilon_t, \end{array} \tag{10.23}$$

es decir, una serie centrada (o en desviaciones a las medias) sigue un proceso $AR(1)$ con parámetro ϕ_1 si al aplicarle el operador $(1 - \phi_1 B)$ se obtiene un proceso ruido blanco. Si consideramos el operador como una ecuación en B, el coeficiente ϕ_1 se denomina factor de la ecuación, y también podemos llegar a la condición de estacionaridad utilizando la raíz de la ecuación. Es decir, igualando el operador a cero y resolviendo la ecuación con B como incógnita tenemos la ecuación y la solución (raíz)

$$\begin{array}{c} 1 - \phi_1 B = 0; \\ B = \frac{1}{\phi_1}, \end{array} \tag{10.24}$$

y el proceso será estacionario si B está fuera del círculo unidad, es decir, si $|B| > 1$

$$\begin{array}{l} |B| = \left|\frac{1}{\phi_1}\right| > 1; \\ |\phi_1| < 1. \end{array} \tag{10.25}$$

[2]En este punto se recomienda leer el apéndice correspondiente en el que se describen algunas propiedades de este operador.

Un proceso $AR\,(1)$ y en general los procesos AR se pueden expresar también como la suma ponderada de procesos ruido blanco (denominado procesos de medias móviles como veremos más adelante). En efecto, retardando un periodo la expresión (10.21) tenemos

$$z_{t-1} = \phi_1 z_{t-2} + \varepsilon_{t-1}, \tag{10.26}$$

y sustituyendo en (10.21)

$$z_t = \phi_1 z_{t-1} + \varepsilon_t = \phi_1 \left(\phi_1 z_{t-2} + \varepsilon_{t-1}\right) + \varepsilon_t = \phi_1^2 z_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t. \tag{10.27}$$

Aplicando este procedimiento sucesivamente llegamos a

$$z_t = \phi_1^t z_1 + \phi_1^{t-1} \varepsilon_1 + \phi_1^{t-2} \varepsilon_2 + ... + \phi_1^2 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t, \tag{10.28}$$

y bajo el supuesto de que t es grande, ϕ^t será a efectos prácticos cero, es decir que un proceso $AR\,(1)$ (y en general cualquier proceso AR) se puede representar como la suma de ruido blanco ponderada por una constante que decrece geométricamente, y suponiendo que la serie comienza en el pasado lejano $(-\infty)$, podemos expresar los procesos AR como la suma infinita siguiente:

$$z_t = \sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i}. \tag{10.29}$$

Aunque la esperanza del proceso (10.17) es $\phi_0 / \left(1 - \phi_1\right)$, si utilizamos la expresión (10.21), entonces el proceso $AR\,(1)$ centrado (o en desviaciones a la media, $\phi_0 = 0$) tiene esperanza nula

$$\mathbb{E}\left(z_t\right) = \mathbb{E}\left(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i}\right) = \sum_{i=0}^{\infty} \phi_1^i \mathbb{E}\left(\varepsilon_{t-i}\right) = 0. \tag{10.30}$$

Sabiendo que el proceso centrado tiene media nula, tras elevar al cuadrado la expresión (10.21), si aplicamos esperanzas obtenemos la varianza incondicional o incondicionada

$$\mathbb{E}\left(z_t^2\right) = \mathbb{E}\left(\phi_1^2 z_{t-1}^2 + 2 z_{t-1} \varepsilon_t + \varepsilon_t^2\right). \tag{10.31}$$

Si denominamos a la varianza del proceso por σ_z^2, tenemos que

$$\sigma_z^2 = \phi_1^2 \sigma_z^2 + \sigma_\varepsilon^2, \tag{10.32}$$

de manera que

$$\sigma_z^2 = \frac{\sigma_\varepsilon^2}{1 - \phi_1^2}. \tag{10.33}$$

Lo cual nuevamente nos permite comprobar que para que la varianza sea positiva se hace necesario el cumplimiento de la condición de estacionaridad, $|\phi_1| < 1$.

Para calcular las funciones de autocorrelación partimos de la expresión (10.29). Multiplicando por z_{t+u} y aplicando esperanzas,

$$\begin{aligned}\gamma_u &= \mathbb{E}(z_t z_{t+u}) = \mathbb{E}\left(\sum_{i=0}^{\infty}\phi_1^i\varepsilon_{t-i}\sum_{j=0}^{\infty}\phi_1^j\varepsilon_{t+u-j}\right)\\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\phi_1^i\phi_1^j\mathbb{E}(\varepsilon_{t-i}\varepsilon_{t+u-j}).\end{aligned} \tag{10.34}$$

Encontramos que las esperanzas solo serán distintas de cero cuando los subíndices entre paréntesis coincidan (puesto que el proceso ruido blanco está incorrelacionado), es decir, cuando $t-i = t-j+u$, lo que ocurre cuando $j = u+i$, y sustituyendo j por $u+i$ en la expresión anterior tenemos que

$$\begin{aligned}\gamma_u &= \mathbb{E}(z_t z_{t+u}) = \sum_{i=0}^{\infty}\phi_1^i\phi_1^{u+i}\mathbb{E}\left(\varepsilon_{t-i}^2\right) =\\ &= \sigma_\varepsilon^2\sum_{i=0}^{\infty}\phi_1^{2i+u} = \sigma_\varepsilon^2\phi_1^u\sum_{i=0}^{\infty}\phi_1^{2i} =\\ &= \sigma_\varepsilon^2\phi_1^u(1+\phi_1^2+\phi_1^4+...) = \frac{\sigma_\varepsilon^2\phi_1^u}{1-\phi_1^2}\end{aligned} \tag{10.35}$$

donde el único valor que cambia con el desfase temporal es ϕ_1^u, que decrece geométricamente si se cumple la condición de estacionaridad $|\phi_1| < 1$.

La función de autocorrelación es

$$\rho_u = \frac{\gamma_u}{\gamma_0} = \frac{\frac{\sigma_\varepsilon^2\phi_1^u}{1-\phi_1^2}}{\frac{\sigma_\varepsilon^2}{1-\phi_1^2}} = \phi_1^u, \tag{10.36}$$

de manera que la **función de autocorrelación (FAT)** de un proceso $AR(1)$ decrece de forma geométrica en valor absoluto. Si el valor del parámetro ϕ_1 es negativo lo hará de igual forma, pero cambiando de signo en desfases sucesivos.

La función de autocorrelación ayuda a caracterizar a los procesos AR: en general, si decrece rápidamente nos encontraremos, como veremos posteriormente, ante un proceso AR.

Sin embargo la función de autocorrelación no nos informa del *orden del proceso* autorregresivo. Para determinarlo, es decir para identificar el orden p de un proceso $AR(p)$ debemos recurrir a la **función de autocorrelación parcial** (**FAP**).

La autocorrelación parcial con k desfases, que denominaremos ϕ_{kk}, mide la influencia de z_{t-k} sobre z_t descontada la influencia de los $k-1$ valores anteriores de z: $z_{t-1}, z_{t-2}, ...z_{t-k-1}$. Es decir, la autocorrelación parcial de orden k se refiere a la correlación entre z_t, z_{t-k} condicionada a $z_{t-1}, z_{t-2}, ...z_{t-k-1}$.

Esta correlación vendrá dada por un coeficiente, ϕ_{kk}, que irá definiendo la función de autocorrelación parcial para los desfases temporales 1 a k de cualquier proceso estacionario. El coeficiente (poblacional) será el último coeficiente de los siguientes modelos que crecen secuencialmente en orden ($k = 1, 2, ...$), y que pueden ser fácilmente estimados por MCO:

$$\begin{aligned} &z_t = \phi_1 z_{t-1};\ \text{donde } \phi_1 = \phi_{11} \\ &z_t = \varphi_1 z_{t-1} + \varphi_2 z_{t-2};\ \text{donde } \varphi_2 = \phi_{22} \\ &... \\ &z_t = \iota_1 z_{t-1} + \iota_2 z_{t-2} + ... + \iota_k z_{t-k};\ \text{donde } \iota_k = \phi_{kk}. \end{aligned} \tag{10.37}$$

De esta manera la función de autocorrelacción *parcial* (FAP) vendrá dada (poblacionalmente) por la sucesión $\{\phi_{11}, \phi_{22}, ..., \phi_{uu}, ..., \phi_{kk}\}$. Hay que distinguir esta función de la función de correlación $\{\rho_1, \rho_2, ..., \rho_u, ..., \rho_k\}$, y para ello denominaremos a esta última función de autocorrelación *total* (FAT).

Para estimar estos coeficientes FAP usaríamos la técnica MCO sobre modelos secuenciales

$$\begin{aligned} &z_t = \phi_1 z_{t-1} + \varepsilon_t^{(1)};\ \text{donde } \phi_1 = \phi_{11} \\ &z_t = \varphi_1 z_{t-1} + \varphi_2 z_{t-2} + \varepsilon_t^{(2)};\ \text{donde } \varphi_2 = \phi_{22} \\ &... \\ &z_t = \iota_1 z_{t-1} + \iota_2 z_{t-2} + ... + \iota_k z_{t-k} + \varepsilon_t^{(k)};\ \text{donde } \iota_k = \phi_{kk}. \end{aligned} \tag{10.38}$$

Donde $\varepsilon_t^{(u)}$ hace referencia a los errores para cada uno de los modelos secuenciales. En caso de que el proceso fuera un AR(1), el error $\varepsilon_t^{(1)}$ sería ruido blanco (por definición de AR(1)), y el coeficiente poblacional $\phi_{11} \neq 0$, de lo contrario no sería un AR(1), y el coeficiente muestral estimado $\hat{\phi}_{11}$ convergería asintóticamente al verdadero (a un valor distinto de cero). Pensemos ahora cómo sería el último coeficiente de una de las ecuaciones siguientes, es decir $u > 1$. En estos casos, dado que el proceso es AR(1) y por tanto solo es relevante el valor que toma la serie en el momento justamente anterior, una vez que dicho valor ya ha sido considerado, pues la regresión u-ésima contiene el efecto en cuestión, tendríamos que $\phi_{uu} = 0$, y por tanto su contrapartida muestral $\hat{\phi}_{uu}$ convergería a cero. Hemos comprobado que los procesos autorregresivos de primer orden se caracterizan por tener una FAP en la que solo el primer retardo es estadísticamente distinto de cero, mientras que los siguientes desfases son estadísticamente nulos.

Para determinar si un valor concreto ϕ_{uu} de los k estimados en la FAP es significativamente distinto de cero consideramos el proceso ruido blanco, es decir, consideramos que los valores sucesivos de ϕ_{uu} para $uu = 1, 2, ...$ son independientes y se distribuyen como una normal de media cero y varianza unitaria. En estas condiciones, cada ϕ_{uu} se distribuye de la siguiente forma:

$$\begin{aligned} &\mathbb{E}\left(\phi_{uu}\right) = 0 \\ &\text{Var}\left(\phi_{uu}\right) = \tfrac{1}{T} \\ &\text{Cov}\left(\phi_{uu}, \phi_{uu+h}\right) = 0 \text{ para } h \geq 1, \end{aligned} \tag{10.39}$$

de manera que para contrastar la hipótesis nula de que un ϕ_{uu} concreto es nulo $(\text{H}_0 : \phi_{uu} = 0)$ realizamos el contraste de hipótesis usual, y rechazamos la hipótesis nula con el 95 % de confianza si

$$\left|\frac{\phi_{uu}-0}{1/\sqrt{T}}\right| > 1,96;\ |\phi_{uu}| > \frac{1,96}{\sqrt{T}}, \tag{10.40}$$

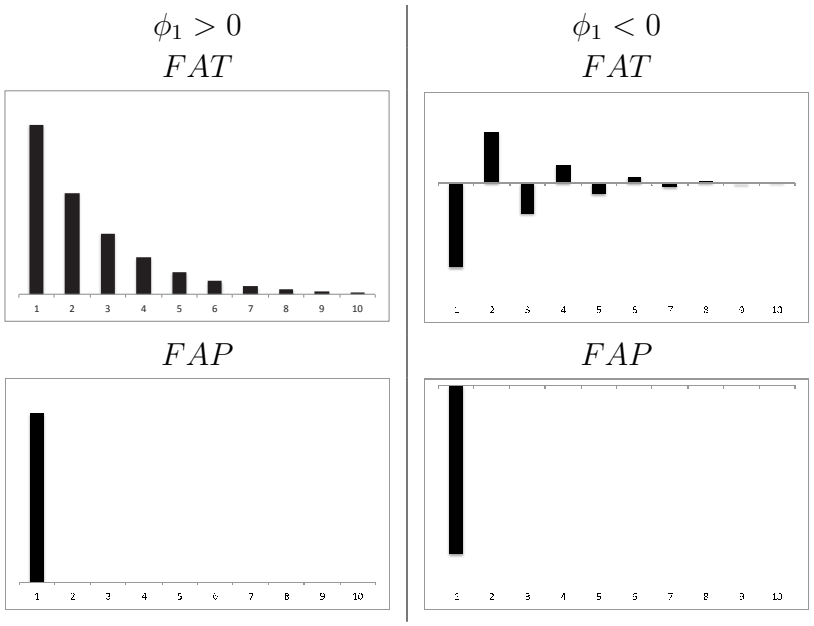
es decir que si un valor concreto de ϕ_{uu} está fuera del intervalo $\pm 1,96/\sqrt{T}$ entonces podemos afirmar que ϕ_{uu} es distinto de cero con el 95 % de confianza.

Un proceso $AR(1)$ presentará una FAP en la que solo el valor del primer desfase temporal ϕ_{11} será distinto de cero y todos los demás serán nulos ($\phi_{uu} = 0,$ para $uu > 1$). Un $AR(2)$ presentará los dos primeros desfases de su FAP distintos de cero ϕ_{11} y ϕ_{22} y el resto nulos y así sucesivamente. De manera que el orden del proceso AR lo determina la FAP.

Llamamos correlograma completo a la representación gráfica de las funciones de autocorrelación total (FAT) y parcial (FAP) de desfases sucesivos de un proceso estocástico estacionario.

El correlograma completo de un proceso $AR(1)$ en consecuencia mostrará una función de autocorrelación total que decrece de forma geométrica a medida que se incrementa el desfase temporal y una función de autocorrelación parcial con un solo valor significativo, el de orden uno.

Figura 10.1: Correlogramas de un $AR(1)$



La Figura 10.1 muestra el correlograma completo de dos procesos $AR(1)$. Los dos gráficos de la izquierda muestran la FAT y la FAP de un proceso $AR(1)$ con parámetro positivo $\phi_1 > 0$, la FAT (gráfico superior izquierdo) decrece geométricamente mientras que la FAP (gráfico inferior izquierdo) presenta un solo valor distinto de cero en el primer desfase. Los gráficos de la derecha muestran el correlograma de un proceso $AR(1)$ con parámetro negativo $\phi_1 < 0$, la FAT (gráfico superior derecho) decrece rápidamente pero cambiando de signo en desfases sucesivos, la FAP (gráfico inferior

derecho) muestra un solo desfase distinto de cero, el del primer desfase, con signo negativo.

Una forma alternativa y teóricamente muy atractiva de llegar a los valores de la autocorrelación parcial (FAP) es a partir de los valores de la función de autocorrelación total (FAT) ρ_u, utilizando para ello las denominadas *ecuaciones de Yule-Walker*. En efecto, partiendo de (10.21), multiplicando ambos lados de la ecuación por z_{t-1} y aplicando esperanzas, tenemos

$$\begin{aligned} &\mathbb{E}\left(z_t z_{t-1}\right) = \phi_1 \mathbb{E}\left(z_{t-1}^2\right) + \mathbb{E}\left(z_{t-1}\varepsilon_t\right); \\ &\gamma_1 = \phi_1 \gamma_0, \end{aligned} \tag{10.41}$$

puesto que $\mathbb{E}\left(z_{t-1}\varepsilon_t\right) = 0$ y dividiendo ambas partes por la varianza γ_0, llegamos a

$$\rho_1 = \phi_1, \tag{10.42}$$

es decir que la función de autocorrelación total con un desfase ρ_1 es igual a la función de autocorrelación parcial con un desfase ϕ_{11} en cualquier proceso estacionario.

En general podemos calcular FAP para un desfase cualquiera u. Partiendo de un modelo $AR\left(p\right)$ en desviaciones a las medias

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + ... + \phi_p z_{t-p} + \varepsilon_t, \tag{10.43}$$

y multiplicando ambas partes de la ecuación por z_{t-u}, aplicando esperanzas y dividiendo por la varianza γ_0 llegamos a la expresión de la *ecuación de Yule-Walker* para u desfases

$$\rho_u = \phi_1 \rho_{u-1} + \phi_2 \rho_{u-2} + ... + \phi_p \rho_{u-p}. \tag{10.44}$$

Teniendo en cuenta el carácter par de la FAP $\rho_u = \rho_{-u}$, dando valores a u, y recordando que estamos interesados solo en el último coeficiente denominado $\phi_{uu} = \phi_p$, se obtiene

$$\begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{uu} \end{pmatrix} = \begin{pmatrix} \rho_0 & \rho_1 & \cdots & \rho_{u-1} \\ \rho_1 & \rho_0 & \cdots & \rho_{u-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{u-1} & \rho_{u-2} & \cdots & \rho_0 \end{pmatrix}^{-1} \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_u \end{pmatrix}, \tag{10.45}$$

que permite calcular las funciones de autocorrelación parcial de orden u de forma *sucesiva* a partir de la función de autocorrelación total de cualquier proceso estacionario.

Por ejemplo, en el caso de un AR(1) la expresión sería

$$\phi_{11} = \rho_1/\rho_0 = \rho_1$$

$$\begin{pmatrix} \phi_1 \\ \phi_{22} \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix}$$

que resolviendo y teniendo en cuenta que es un proceso AR(1), como hemos visto, $\rho_2 = \rho_1^2$, se llega fácilmente a que $\phi_{22} = 0/(1-\rho_1^2) = 0$. Lo mismo sucede si operamos para calcular ϕ_{33}, que será nulo, y así sucesivamente para cualquier retardo distinto de primero.

Procesos de media móviles

Un proceso estacionario de media móvil de orden q, denominado usualmente $MA(q)$, analíticamente obedece a la siguiente expresión

$$Z_t = c + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - ... - \theta_q\varepsilon_{t-q}, \tag{10.46}$$

donde ε_t es un proceso ruido blanco centrado o en desviaciones a la media[3]. El signo negativo de los parámetros θ_i es una mera convención y, de hecho, estos pueden ser tanto negativos como positivos. La media del proceso es el término constante.

Los procesos AR se pueden escribir como procesos $MA(\infty)$, mientras que los procesos MA dependen solo de un número finito de retardos p, en este sentido los procesos MA tienen memoria más corta que los procesos AR.

Los modelos $MA(q)$ se pueden escribir, sin pérdida de generalidad, en diferencias a las medias

$$z_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - ... - \theta_q\varepsilon_{t-q}, \tag{10.47}$$

donde desaparece el término constante c y la variable z_t la mostramos en minúscula para indicar que es una variable centrada o en diferencias a las medias.

Proceso de media móvil de orden uno, MA(1)

Podemos entender el modelo MA(1) como la construcción (más simple) de un modelo con dependencia temporal tomando como punto de partida el ruido blanco. El proceso generador obedece a la siguiente expresión en desviaciones a las medias

$$z_t = \varepsilon_t - \theta_1\varepsilon_{t-1}. \tag{10.48}$$

Utilizando el operador de retardos B podemos escribir

$$z_t = \varepsilon_t - \theta_1 B\varepsilon_t = \varepsilon_t(1 - \theta_1 B), \tag{10.49}$$

de manera que el proceso es la suma de dos procesos estacionarios ε_t y $\theta_1\varepsilon_t$, y en consecuencia el proceso $MA(1)$ es por construcción estacionario.

Retardando un periodo la expresión (10.48), tenemos,

[3] $\mathbb{E}(\varepsilon_t) = 0$, $\text{Var}(\varepsilon_t) = \sigma_\varepsilon^2$ y $\rho_u = 0$ para todo $u \geq 1$.

$$\begin{aligned} z_{t-1} &= \varepsilon_{t-1} - \theta_1 \varepsilon_{t-2}; \\ \varepsilon_{t-1} &= z_{t-1} + \theta_1 \varepsilon_{t-2}, \end{aligned} \tag{10.50}$$

y sustituyendo en (10.48) obtenemos

$$\begin{aligned} z_t &= \varepsilon_t - \theta_1 \varepsilon_{t-1} = \varepsilon_t - \theta_1 \left(z_{t-1} - \theta_1 \varepsilon_{t-2} \right) \\ &= \varepsilon_t - \theta_1 z_{t-1} + \theta_1^2 \varepsilon_{t-2}. \end{aligned} \tag{10.51}$$

Realizando el proceso iterativamente llegamos a

$$z_t = \varepsilon_t - \sum_{i=1}^{t-1} \theta_1^i z_{t-i} - \theta_1^t \varepsilon_0, \tag{10.52}$$

que es un proceso autorregresivo. Esta expresión nos permite divisar el escaso sentido que tendría que el parámetro θ_1 fuera superior a la unidad, pues en tal circunstancia el efecto del pasado de z_t tiene mayor incidencia para explicar el valor actual de z_t cuanto más lejano está en el tiempo. Es decir, contemplamos casos en los que el efecto (los efectos) va(n) disminuyendo a medida que aumentan los retardos, para lo cual debemos imponer la condición $|\theta_1| < 1$, en cuyo caso decimos que el proceso MA es **invertible** en un proceso autorregresivo AR. Además esta restricción es perfectamente compatible con el hecho de ser un proceso débilmente dependiente. Por otra parte, también observamos que θ_1^t converge a cero a medida que t aumenta. En estas condiciones si el proceso comienza en el pasado lejano, tan lejano como queramos $(-\infty)$, lo podemos escribir como

$$z_t = \varepsilon_t - \sum_{i=1}^{\infty} \theta_1^i z_{t-i}. \tag{10.53}$$

Por tanto un proceso $MA\,(1)$ se puede escribir como un $AR\,(\infty)$. En general todos los procesos MA invertibles son los que se pueden escribir como un proceso $AR\,(\infty)$.

Aplicando esperanzas a la expresión (10.48) llegamos a la conclusión de que el proceso $MA\,(1)$ centrado tiene media nula

$$\mathbb{E}\left(z_t\right) = \mathbb{E}\left(\varepsilon_t - \theta_1 \varepsilon_{t-1}\right) = \mathbb{E}\left(\varepsilon_t\right) - \theta_1 \mathbb{E}\left(\varepsilon_{t-1}\right) = 0. \tag{10.54}$$

La varianza es

$$\mathbb{E}\left(z_t^2\right) = \mathbb{E}\left(\varepsilon_t - \theta_1 \varepsilon_{t-1}\right)^2 = \mathbb{E}\left(\varepsilon_t^2 + \theta_1^2 \varepsilon_{t-1}^2 - 2\theta_1 \varepsilon_t \varepsilon_{t-1}\right) = \sigma_\varepsilon^2 \left(1 + \theta_1^2\right). \tag{10.55}$$

La función de autocovarianza de orden u la obtenemos multiplicando a ambos lados de la ecuación (10.48) por z_{t-u} y aplicando esperanzas,

$$\gamma_u = \mathbb{E}\left(z_t z_{t-u}\right) = \mathbb{E}\left(\varepsilon_t z_{t-u}\right) - \mathbb{E}\left(\theta_1 \varepsilon_{t-1} z_{t-u}\right). \tag{10.56}$$

Para u = 1 tenemos

$$\gamma_1 = \mathbb{E}(z_t z_{t-1}) = \mathbb{E}(\varepsilon_t z_{t-1}) - \mathbb{E}(\theta_1 \varepsilon_{t-1} z_{t-1}) = -\theta_1 \sigma_\varepsilon^2, \tag{10.57}$$

puesto que $\mathbb{E}(\varepsilon_t z_{t-1}) = \mathbb{E}[\varepsilon_t(\varepsilon_{t-1} - \theta_1 \varepsilon_{t-2})] = 0$ y $\theta_1 \mathbb{E}[\varepsilon_{t-1}(\varepsilon_{t-1} - \theta_1 \varepsilon_{t-2})] = \theta_1 \sigma_\varepsilon^2$.

Para $u = 2$,

$$\gamma_2 = \mathbb{E}(z_t z_{t-2}) = \mathbb{E}(\varepsilon_t z_{t-2}) - \mathbb{E}(\theta_1 \varepsilon_{t-1} z_{t-2}) = 0, \tag{10.58}$$

y para $u > 2$ obtenemos también funciones de autocovarianza nulas ($\gamma_u = 0,$ para $u > 1$).

La función de autocorrelación con un retardo es

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{-\theta_1 \sigma_\varepsilon^2}{\sigma_\varepsilon^2 (1 + \theta_1^2)} = \frac{-\theta_1}{(1 + \theta_1^2)}, \tag{10.59}$$

y para $u > 1$ las funciones de autocorrelación son cero ($\rho_u = 0$).

Por consiguiente, la función de autocorrelación de un $MA(1)$ presentará un solo valor distinto de cero, en el primer desfase. Es decir que el orden del proceso lo determina la función de autocorrelación total (FAT). Por tanto la FAT de un $MA(1)$ tiene la misma interpretación, determinar el orden del proceso que tenía la FAP para un proceso $AR(1)$. Esta misma dualidad se presenta también en la FAP de un $MA(1)$ puesto que este proceso se puede escribir como un $AR(\infty)$, que tiene una FAP que registra el efecto directo de z_{t-u} sobre z_t de magnitud θ_1^u, por lo que la FAP de un $MA(1)$ decrecerá rápidamente en u, siendo todos poblacionalmente no nulos. Por tanto esta característica nos servirá para determinar el orden del proceso MA. La Figura 10.2 muestra el correlograma de un proceso $MA(1)$.

Proceso de medias móviles de orden q, MA(q)

Como ya sabemos, un proceso $MA(q)$ analíticamente presenta la siguiente forma:

$$z_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} = \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}. \tag{10.60}$$

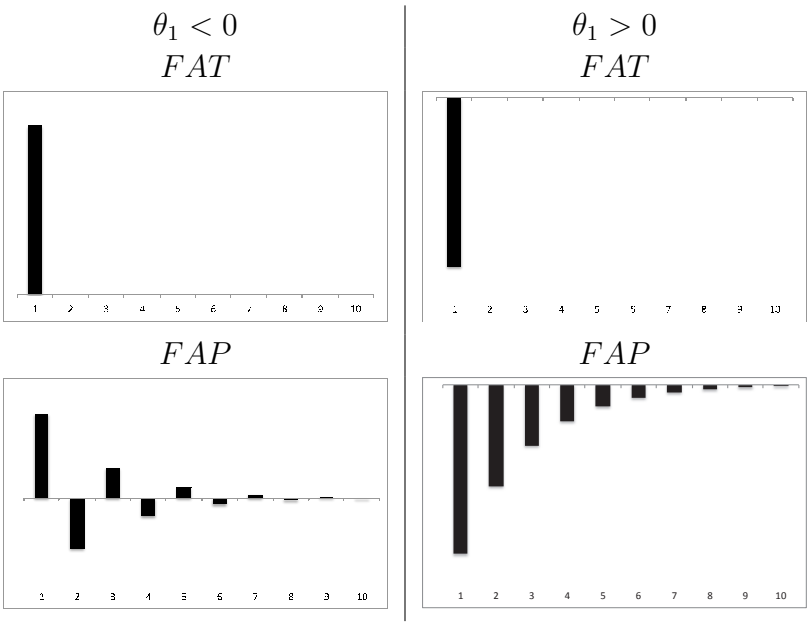
Utilizando el operador de retardos B podemos escribirlo también como

$$z_t = \varepsilon_t - \theta_1 B \varepsilon_t - \theta_2 B^2 \varepsilon_t - ... - \theta_q B^q \varepsilon_t = \varepsilon_t \left(1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q\right), \tag{10.61}$$

donde el operador de retardos $MA(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q$ opera sobre ε_t, y nos permite llegar a la notación general compacta de un modelo $MA(q)$

$$\begin{aligned} z_t &= MA(B)\, \varepsilon_t; \\ MA(B)^{-1} z_t &= \varepsilon_t. \end{aligned} \tag{10.62}$$

Figura 10.2: Correlogramas de un $MA(1)$



Esta última expresión permite expresar el modelo $MA(q)$ como un $AR(\infty)$

$$MA(B)^{-1} = 1 - \eta_1 B - \eta_2 B^2 - ... \tag{10.63}$$

Los coeficientes η_i se obtienen imponiendo la condición $MA(B)^{-1} MA(B) = 1$; los procesos MA deben ser invertibles, lo que se cumplirá si las raíces de la ecuación $MA(B) = 0$ caen fuera del círculo unidad y la serie $MA(B)^{-1}$ será convergente y podremos escribir el proceso MA como

$$z_t = \sum_{i=1}^{\infty} \eta_i z_{t-i} + \varepsilon_t, \tag{10.64}$$

que es un proceso $AR(\infty)$ y, por consiguiente, la FAP de un proceso $MA(q)$ tiene la misma estructura que la FAT de un proceso AR del mismo orden. Es decir, la FAP de un proceso $MA(q)$ decrece rápidamente de forma geométrica o sinusoidal y determina la naturaleza del proceso.

Multiplicando (10.60) por z_{t+u} para $u \geq 0$ y tomando esperanzas, obtenemos la autocovarianza del proceso

$$\begin{aligned} \mathbb{E}(z_t z_{t-u}) &= \mathbb{E}\left[\sum_{i=0}^{q-u}(\theta_i \varepsilon_{t-i}) \sum_{j=0}^{q-u}(\theta_j \varepsilon_{t+u-j})\right] \\ &= \sum_{i=0}^{q-u}\sum_{j=0}^{q-u} \theta_i \theta_j \mathbb{E}(\varepsilon_{t-i}\varepsilon_{t+u-j}). \end{aligned} \tag{10.65}$$

Teniendo en cuenta que $\mathbb{E}(\varepsilon_{t-i}\varepsilon_{t+u-j}) = \sigma_\varepsilon^2$ solo cuando los subíndices coinciden y cero en caso contrario, podemos igualar ambos subíndices $t - i = t + u - j$, lo que implica que cuando $j = i + u$, ambos subíndices son iguales y la esperanza es distinta de cero. Por tanto, la expresión anterior se puede escribir también de la siguiente forma:

$$\gamma_u = \mathbb{E}\left(z_t z_{t-u}\right) = \sum_{i=0}^{q-u} \theta_i \theta_{i+u} \mathbb{E}\left(\varepsilon_{t-i}^2\right). \tag{10.66}$$

Para $u = 0$ tenemos que la varianza del proceso es

$$\sigma_z^2 = \gamma_0 = \sigma_\varepsilon^2 \left(1 + \theta_1^2 + ... + \theta_q^2\right). \tag{10.67}$$

Para $u = 1, 2, ..., q$ obtenemos las autocovarianzas del proceso distintas de cero

$$\gamma_u = \sigma_\varepsilon^2 \sum_{i=0}^{q} \theta_i \theta_{i+u} \text{ para } u = 1, 2, .., q. \tag{10.68}$$

Para $u > q$, los subíndices no coinciden en ningún momento y, por tanto

$$\gamma_u = 0 \text{ para } u > q. \tag{10.69}$$

La funciones de autocorrelación son

$$\begin{array}{l} \rho_u = \frac{\sigma_\varepsilon^2 \sum_{i=0}^{q} \theta_i \theta_{i+u}}{\sigma_\varepsilon^2 \sum_{i=0}^{q} \theta_i^2} = \frac{\sum_{i=0}^{q} \theta_i \theta_{i+u}}{\sum_{i=0}^{q} \theta_i^2} \text{ para } u = 1, 2, .., q \\ \rho_u = 0 \text{ para } u > q, \end{array} \tag{10.70}$$

de manera que la función de autocorrelación total determina el orden del proceso. Un proceso $MA\left(q\right)$ presenta los q primeros desfases distintos de cero y el resto nulos. La Figura 10.3 muestra los correlogramas de un proceso $MA\left(2\right)$.

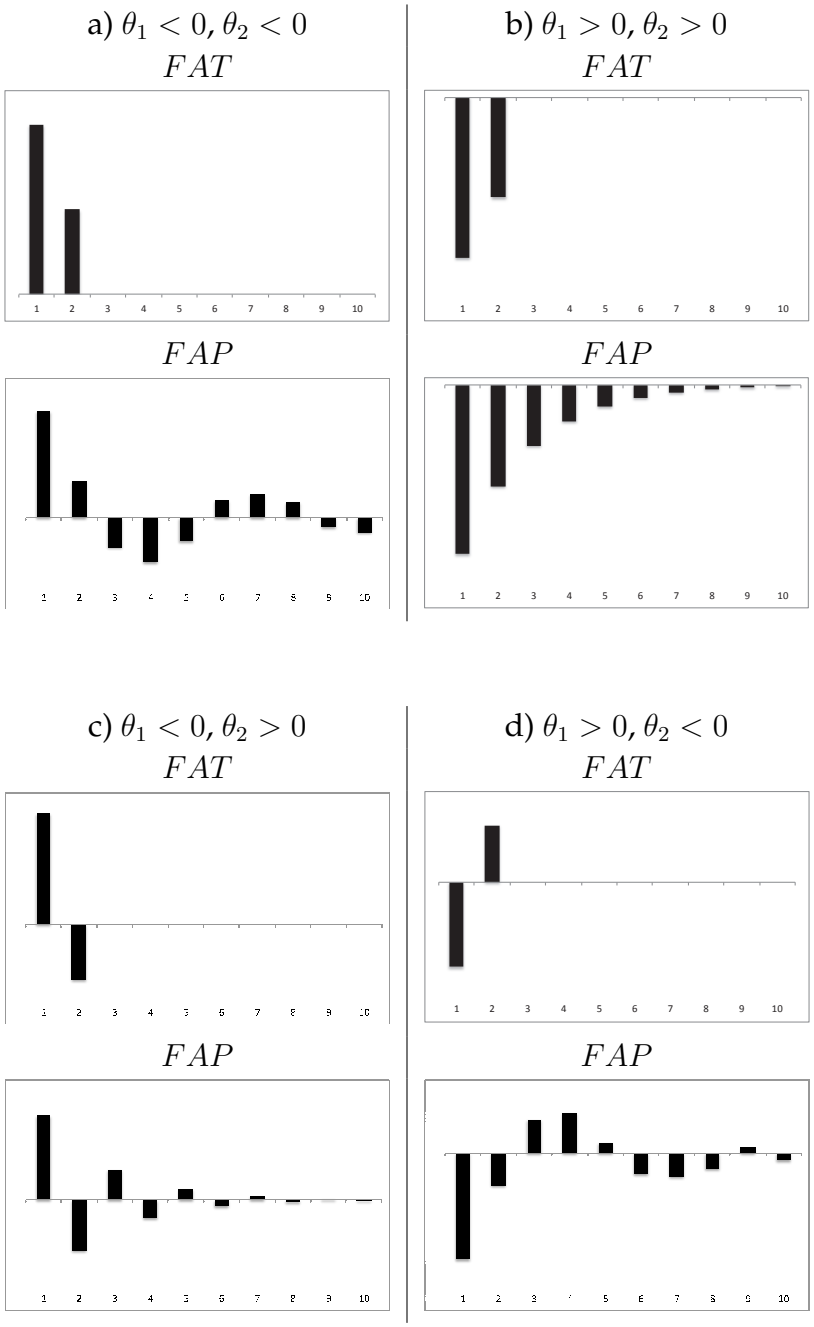
En buena medida la importancia de los procesos MA reside en el denominado Teorema de Descomposición de Wold que establece que todo proceso estacionario débil (ya sea lineal o no) puede ser aproximado tanto como deseemos (en términos de precisión) por un proceso MA, con un orden q largo, más una parte determinista (por ejemplo, una constante o una fución trigonométrica con dependencia temporal)

$$z_t = \delta_t + \sum_{j=0}^{\infty} \psi_j a_{t-j}$$

donde la parte determinista está en el primer sumando, mientras que el segundo es una suma ponderada de ruido blanco formado por combinaciones lineales de $z_s, s < t$. Este último sumando contiene o representa una suma ponderada de *errores de predicción* generados al intentar predecir z_t a partir de combinaciones lineales de z_s.

Los parámetros de los procesos MA no se pueden estimar por MCO puesto que la suma cuadrática de las discrepancias no son una función lineal de los parámetros a estimar y se suelen utilizar procedimientos como el de máxima verosimilitud condicional o exacta. Afortunadamente los programas especializados incorporan estos algoritmos y calculan los parámetros de los modelos MA de forma rutinaria.

Figura 10.3: Correlogramas de un $MA(2)$



PROCESOS ARMA

Los procesos que combinan los modelos AR y MA conjuntamente se denominan procesos $ARMA$. El correlograma de un proceso ARMA es bastante más complejo que los que hemos visto hasta ahora. La parte AR se puede escribir como un $MA(\infty)$ pero con pautas de decrecimiento geométrico; la parte MA tiene pocos parámetros pero irrestrictos. Por consiguiente, los procesos $ARMA$ se pueden aproximar a un modelo $MA(\infty)$ en el que los primeros desfases no tienen restricciones pero, a partir de un determinado desfase, decrecen de forma geométrica.

Precisamente el mencionado Teorema de Descomposición de Wold prevé la existencia de un polinomio infinito

$$\sum_{j=0}^{\infty} \psi_j a_{t-j} = \Psi(B) a_t$$

que podría ser obtenido por el cociente de dos polinomios de retardos finitos (digamos, p, q). Es decir

$$z_t = \Psi(B) a_t = \frac{MA(B)}{AR(B)} a_t, \text{ y } AR(B) z_t = MA(B) a_t,$$

que observamos combinan polinomios de retardos p, q.

Proceso ARMA $(1,1)$

El proceso más sencillo es el proceso $ARMA(1,1)$,

$$z_t = \phi_1 z_{t-1} + \varepsilon_t - \theta_1 \varepsilon_{t-1}, \tag{10.71}$$

donde las variables están en diferencias a las medias y ε_t es ruido blanco. Utilizando el operador de retardos B tenemos

$$(1 - \phi_1 B) z_t = (1 - \theta_1 B) \varepsilon_t. \tag{10.72}$$

Para que el proceso sea invertible se debe cumplir que $|\theta_1| < 1$, y para que sea estacionario que $|\phi_1| < 1$. En tal caso podemos expresar un ARMA(1,1) tanto como un AR, como un MA, ambos de orden infinito y serán útiles para caracterizar los correlogramas de este tipo de procesos.

En primer lugar vamos a invertir[4] la parte AR

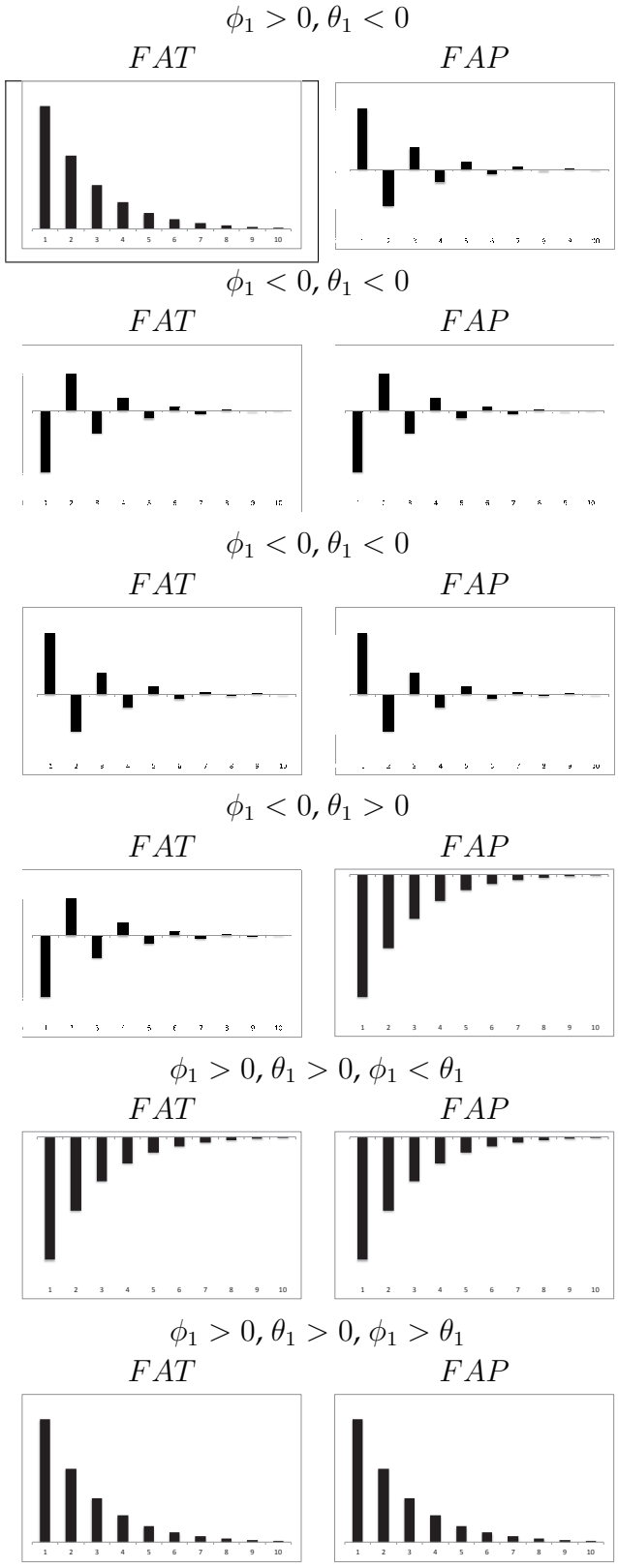
$$z_t = (1 - \phi_1 B)^{-1} (1 - \theta_1 B) \varepsilon_t = \left(1 + \phi_1 B + \phi_1^2 B^2 + ...\right) (1 - \theta_1 B) \varepsilon_t$$

$$z_t = \left(\varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + ...\right) - \left(\theta_1 \varepsilon_{t-1} + \phi_1 \theta_1 \varepsilon_{t-2} + \phi_1^2 \theta_1 \varepsilon_{t-3} + ...\right)$$

$$z_t = \varepsilon_t + (\phi_1 - \theta_1) \varepsilon_{t-1} + \phi_1 (\phi_1 - \theta_1) \varepsilon_{t-2} + \phi_1^2 (\phi_1 - \theta_1) \varepsilon_{t-3} + ...$$

[4] Es aconsejable que el lector que experimente alguna dificultad técnica en la inversión revise el Apéndice de este tema.

Figura 10.4: Correlogramas de un $ARMA\,(1,1)$



Esta representación en términos de un MA (∞) es consecuencia de la inversión de la parte AR (1), y nos facilita observar que los parámetros decrecen de forma geométrica en potencias sucesivas de ϕ_1. Así pues la parte *AR*(1) del *ARMA*(1,1) facilitará una FAT de similar trayectoria a un AR(1).

En segundo lugar, el proceso ARMA $(1,1)$ también se puede escribir como una AR (∞) puesto que la parte del proceso MA es invertible

$$\varepsilon_t = (1-\phi_1 B)(1-\theta_1 B)^{-1} z_t = \left(1+\theta_1 B+\theta_1^2 B^2+...\right)(1-\phi_1 B) z_t$$

que tras operar y despejar, podemos expresar del siguiente modo

$$z_t = (\phi_1-\theta_1) z_{t-1} + \theta_1(\phi_1-\theta_1) z_{t-2} + \theta_1^2(\phi_1-\theta_1) z_{t-3} + ... + \varepsilon_t.$$

Esta última expresión nos permite ver que el *efecto directo* de z_{t-u} sobre z_t decrece geométricamente en potencias de θ_1, es decir, θ_1^u. Es de esperar por lo tanto que la FAP presente un decrecimiento geométrico como consecuencia de la influencia de la parte MA en el proceso *ARMA* $(1,1)$.

A modo de resumen, podemos decir que la FAP y FAT de este tipo de procesos tendrán una descripción estructural muy parecida: El primer valor depende de la diferencia paramétrica $(\phi_1-\theta_1)$; los siguientes valores de la FAP y FAT irán decreciendo a una tasa determinada por θ_1 y ϕ_1, respectivamente. La Figura 10.4 contempla distintos escenarios en función del valor y signos de dichos parámetros.

A los efectos de determinar el valor preciso de los coeficientes relevantes, procedemos inicialmente elevando al cuadrado la expresión (10.71), aplicando ahora esperanzas, obtenemos la varianza del proceso

$$\begin{aligned} \sigma_z^2 &= \mathbb{E}(z_t^2) = \mathbb{E}(\phi_1 z_{t-1} + \varepsilon_t - \theta_1\varepsilon_{t-1})^2 = \phi_1^2\sigma_z^2 - 2\phi_1\theta_1\sigma_\varepsilon^2 + \theta_1^2\sigma_\varepsilon^2 + \sigma_\varepsilon^2, \\ \sigma_z^2 &= \gamma_0 = \sigma_\varepsilon^2 \frac{\left(1-2\phi_1\theta_1+\theta_1^2\right)}{1-\phi_1^2}. \end{aligned} \tag{10.73}$$

Multiplicando también (10.71) por z_{t-u} y aplicando esperanzas obtenemos las funciones de autocovarianza del proceso

$$\gamma_u = \phi_1\gamma_{u-1} + \mathbb{E}(\varepsilon_t z_{t-u}) - \theta_1\mathbb{E}(\varepsilon_{t-1} z_{t-u}). \tag{10.74}$$

Esta expresión nos facilita comprobar que para $u=1$ se tiene que

$$\gamma_1 = \phi_1\gamma_0 - \theta_1\sigma_\varepsilon^2. \tag{10.75}$$

Así pues podemos obtener el primer coeficiente de autocorrelación

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{1}{\sigma_\varepsilon^2}\frac{1-\phi_1^2}{(1-2\phi_1\theta_1+\theta_1^2)}\left(\phi_1\gamma_0 - \theta_1\sigma_\varepsilon^2\right) = \frac{(\phi_1-\theta_1)(1-\phi_1\theta_1)}{(1-2\phi_1\theta_1+\theta_1^2)}$$

donde la última igualdad se obtiene tras sustituir y simplicar.

De modo similar, para $u > 1$ la función de autocovarianza es

$$\gamma_u = \phi_1 \gamma_{u-1}, \tag{10.76}$$

que podemos calcular recursivamente. Por tanto,

$$\rho_u = \phi_1 \rho_{u-1}$$

y la FAT decrecerá también de forma geométrica como consecuencia de la influencia de la parte $AR\,(1)$ del proceso $ARMA\,(1,1)$.

En todo caso los correlogramas reales de los procesos $ARMA\,(1,1)$ pueden diferir de los teóricos representados en la Figura 10.4. Cuanto mayor sea el peso de la parte AR respecto de la parte MA, el correlograma del proceso $ARMA$ será más parecido al correlograma teórico de un modelo AR. Por el contrario, si la parte MA pesa más, su correlograma se acercará al teórico de un modelo MA.

Por último, la expresión de ρ_1 nos permite considerar el caso particular de $\theta_1 = \phi_1$, ya que en tal situación $\rho_1 = 0$, y por tanto también serán nulos los siguientes retardos, es decir, $\rho_u = 0, u = 1, 2, 3, ...$, que es justamente el correlograma del ruido blanco. El motivo por el que sucede esto es porque el polinomio de la parte *MA* y el de la parte *AR* comparten, en ese caso, una raíz común, por lo que podríamos reducir la expresión $(1 - \phi_1 B)\, z_t = (1 - \theta_1 B)\, \varepsilon_t$ simplemente multiplicando ambos miembros por el factor $(1 - \phi_1 B)^{-1} = (1 - \theta_1 B)^{-1}$, lo que significaría que $z_t = \varepsilon_t$, o lo que es lo mismo que el proceso sería ruido blanco.

Proceso ARMA (p, q)

Un proceso $ARMA\,(p,q)$ combina los procesos $AR\,(p)$ y $MA\,(q)$, y analíticamente tiene la forma siguiente:

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + ... + \phi_p z_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} + \varepsilon_t. \tag{10.77}$$

Utilizando el operador de retardos B podemos escribir el proceso como sigue

$$\left(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p\right) z_t = \left(1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q\right) \varepsilon_t, \tag{10.78}$$

o, en notación compacta[5]

$$AR\,(B)\, z_t = MA\,(B)\, \varepsilon_t. \tag{10.79}$$

El proceso será invertible si las raíces de $MA(B) = 0$ están fuera del círculo unidad y estacionario si ocurre lo mismo con las raíces de los retardos autorregresivos $AR\,(B) = 0$. Además hemos de suponer que no hay raíces comunes entre ambas partes. Al igual que

[5]Es evidente que las expresiones AR(B) o MA(B) son referentes a los polinomios en B relativos a cada una de las partes AR y MA. En el tratamiento que a estos efectos damos en el Apéndice Técnico a este tema, dichos polinomios se denotan de forma diferente.

con el modelo *ARMA*(1,1), podemos expresar cualquier modelo *ARMA*(p,q) como un *MA* de orden infinito y como un *AR* de orden infinito.

De igual modo a como hemos procedido con los modelos anteriores, podemos obtener las autocovarianzas simplemente multiplicando por z_{t-u}

$$\left(1-\phi_1 B-\phi_2 B^2-...-\phi_p B^p\right) z_t z_{t-u}=\left(1-\theta_1 B-\theta_2 B^2-...-\theta_q B^q\right)\varepsilon_t z_{t-u}$$

y tomando a continuación esperanzas

$$\gamma_u-\phi_1\gamma_{u-1}-\phi_2\gamma_{u-2}-...-\phi_p\gamma_{u-p}=\mathbb{E}\left(\varepsilon_t z_{t-u}\right)-\theta_1\mathbb{E}\left(\varepsilon_t z_{t-u-1}\right)-\theta_2\mathbb{E}\left(\varepsilon_t z_{t-u-2}\right)-...-\theta_q\mathbb{E}\left(\varepsilon_t z_{t-u-q}\right).$$

Considerando que para $u>q$ la parte de la derecha se anula, tendremos, tras dividir entre γ_0 :

$$\rho_u-\rho_1\gamma_{u-1}-\rho_2\gamma_{u-2}-...-\rho_p\gamma_{u-p}=0.$$

La forma típica de la FAT de un modelo $ARMA\left(p,q\right)$ es geométricamente decreciente, y esto es así debido a la parte AR del proceso; sin embargo este decrecimiento puede estar atenuado por el componente MA. Lo mismo puede decirse de la FAP pero a la inversa. En consecuencia puede resultar complejo identificar el orden del proceso $ARMA\left(p,q\right)$ en la práctica

10.5 Procesos ARIMA y SARIMA

Hemos visto anteriormente que un proceso es integrado de orden d, $I\left(d\right)$ si obtenemos un proceso estacionario al aplicar d diferencias sucesivas.

Supongamos que denominamos W_t a la serie original $I\left(1\right)$ y Z_t a la serie estacionaria consecuencia de aplicar una diferencia. En estas condiciones podemos escribir

$$W_t=W_{t-1}+Z_t, \tag{10.80}$$

que es un proceso autorregresivo de parámetro unitario y cuya primera diferencia es estacionaria. Utilizando el operador de retardos B tenemos que

$$dW_t=\triangle W_t=\left(1-B\right)W_t=W_t-W_{t-1}=Z_t. \tag{10.81}$$

Los modelos $ARIMA$ incorporan esta posibilidad, es decir, permiten incorporar un proceso autorregresivo de parámetro unitario previo a la aplicación de los procesos $ARMA$.

Un proceso $ARIMA\left(p,d,q\right)$ es un proceso integrado de orden d $\left[I\left(d\right)\right]$ que combina además una parte autorregresiva de orden p $\left[AR\left(p\right)\right]$ y una parte de medias móviles de orden q $\left[MA\left(q\right)\right]$.

Así en el proceso $ARIMA\left(1,1,1\right)$, el primer dígito indica el orden del componente AR, el segundo el grado de integración y el tercero el orden del componente MA. Utilizando el operador de retardos, un *ARIMA*(1,1,1) puede escribirse como

$$(1-\phi_1 B)(1-B)W_t = c + (1-\theta_1 B)\varepsilon_t. \tag{10.82}$$

Análogamente un proceso es integrado de orden dos, $I(2)$, si aplicando dos diferencias sucesivas obtenemos un proceso estacionario,

$$d^2 W_t = \triangle^2 W_t = (1-B)^2 W_t = Z_t \tag{10.83}$$

desarrollando el paréntesis al cuadrado obtenemos

$$(1-B)^2 W_t = \left(1 + B^2 - 2B\right) W_t = W_t + W_{t-2} - 2W_{t-1} = Z_t.$$

En general, un modelo $ARIMA(p,d,q)$ se escribe como

$$\left(1-\phi_1 B - \phi_2 B^2 - ... - \phi_p B^p\right)(1-B)^d W_t = c + \left(1-\theta_1 B - \theta_2 B^2 - ... - \theta_q B^q\right)\varepsilon_t. \tag{10.84}$$

Cuando las observaciones de una serie tienen una periodicidad inferior al año, el componente estacional s puede ser muy importante[6]. En general, los procesos $ARIMA$ se pueden generalizar a los procesos $SARIMA$ (o simplemente modelos $ARIMA$ estacionales) sin más que sustituir los desfases regulares $i = 1, 2, 3, ...$ por los estacionales $s = s, 2s, 3s, ...$ Además los modelos estacionales presentan los mismos correlogramas que los modelos $ARIMA$ no estacionales pero considerando solo los retardos o desfases estacionales.

Así un modelo $SAR(1)$ o un modelo $AR(1)$ estacional tiene la forma $z_t = \Phi_1 z_{t-s} + \varepsilon_t$. Un $SAR(2)$se escribe como $z_t = \Phi_1 z_{t-s} + \Phi_2 z_{t-2s} + \varepsilon_t$. Por lo tanto estos modelos tendrán los mismos momentos y correlogramas que los $AR(1)$ y $AR(2)$ regulares, pero considerando solo los retardos estacionales. Los procesos SAR presentan una FAT que decrece rápidamente en los retardos estacionales de forma exponencial o sinusoidal indicando la naturaleza del proceso. La FAP indica el orden del proceso SAR: un $SAR(1)$ tiene un solo valor distinto de cero en el primer retardo estacional s. Un $SAR(2)$ tiene solo dos valores distintos de cero, s y $2s$, el resto serán nulos.

Lo mismo podemos decir del modelo $SMA(1)$ o un modelo $MA(1)$ estacional, que tiene la forma $z_t = \Theta_1 \varepsilon_{t-s} + \varepsilon_t$. Un $SMA(2)$ se escribe como $z_t = \Theta_1 \varepsilon_{t-s} + \Theta_2 \varepsilon_{t-2s} + \varepsilon_t$. Tienen también los mismos momentos y correlogramas que los $MA(1)$ y $MA(2)$ pero considerando solo los retardos estacionales. Los procesos SMA presentan una FAP que decrece rápidamente en los retardos estacionales de forma exponencial o sinusoidal indicando la naturaleza del proceso. La FAT indica el orden del proceso SMA: un $SMA(1)$ tiene un solo valor distinto de cero en el primer retardo estacional s. Un $SMA(2)$ tiene solo dos valores distintos de cero s y $2s$, el resto serán nulos. Los procesos $SARMA$ presentan también la misma forma funcional y los mismos correlogramas que los procesos $ARMA$ pero considerando solo los desfases estacionales.

[6]Su valor es $s = 2$ cuando la serie es semestral; $s = 4$ si es trimestral; $s = 12$ si es mensual; $s = 52$ si es semanal; y $s = 365$ si la serie tiene observaciones diarias.

El modelo $SARIMA\,(P,D,Q)$ es un proceso integrado de orden D estacional que se combina con un proceso $SAR\,(P)$ y un proceso $SMA\,(Q)$.

Un modelo $SARIMA\,(1,1,1)$ analíticamente es

$$DW_t = \triangle_s W_t = (1-B^s)\,W_t = Z_t = c + \Phi Z_{t-s} - \Theta\varepsilon_{t-s} + \varepsilon_t. \tag{10.85}$$

Los modelos $ARIMA$ regulares y los modelos $SARIMA$ estacionales se pueden combinar en modelos generales del tipo $SARIMA\,(p,d,q)\,(P,D,Q)_s$ donde el componente d y D indica el orden de integración regular y estacional $I\,(d,D)$, incluyendo también el componente autorregresivo regular $AR\,(p)$ y estacional $SAR\,(P)$ y el componente medias móviles regular $MA\,(q)$ y estacional $SMA\,(Q)$.

Por ejemplo, un modelo $SARIMA\,(1,1,1)\,(1,1,1)_s$ analíticamente es

$$(1-B)\,(1-B^s)\,W_t = Z_t = c + \phi_1 Z_{t-1} + \Phi_1 Z_{t-s} - \theta_1\varepsilon_{t-1} - \Theta_1\varepsilon_{t-s} + \varepsilon_t, \tag{10.86}$$

que presentará el correlograma típico de un proceso $ARMA\,(1,1)$ en el orden regular y un correlograma similar en el orden estacional $SARMA\,(1,1)_s$.

El modelo anterior presenta una variable W_t integrada de orden uno regular y estacional $I\,(1,1)$, es decir que realizando una diferencia regular y una estacional obtenemos una serie estacionaria Z_t; esto lo podemos escribir como

$$\triangle\triangle_s W_t = dDW_t = (1-B)\,(1-B^s)\,W_t = Z_t; \tag{10.87}$$

desarrollando el paréntesis tenemos que

$$\begin{aligned}(1-B)\,(1-B^s)\,W_t &= (1-B^s-B+B^{s+1})\,W_t \\ &= W_t - W_{t-s} - W_{t-1} + W_{t-s-1} = Z_t.\end{aligned} \tag{10.88}$$

10.6 El espectro y su estimación

El análisis de series temporales puede llevarse a cabo indistintamente en el *dominio del tiempo* utilizando los modelos ARIMA que vimos en un tema anterior, o en el *dominio de las frecuencias* en cuyo caso emplearemos el análisis espectral. El análisis en el dominio de las frecuencias centra su atención en el estudio de los movimientos cíclicos de una serie temporal. Como en el caso de los modelos ARIMA, trataremos de explicar estos movimientos con la información exclusiva de la propia serie en el pasado, es decir sin relacionarla con otra u otras variables, siendo asimismo la predicción uno de los objetivos de esta aproximación.

Consideremos una serie de tiempo estacionaria y_t cuyo movimiento está causado por distintas oscilaciones o variaciones en distintas frecuencias, $1.\ldots,j$. Un modelo natural

para explicar su variación sería:

$$y_t = \sum_{j=1}^{k} Z_j \cos(w_j t + p_j) + e_t \tag{10.89}$$

donde t es el tiempo, w es la frecuencia, Z la amplitud, p la fase y e_t es una perturbación aleatoria con las características que hemos venido considerando habituales en los temas anteriores. Diferentes oscilaciones implican que hay diversas frecuencias relevantes a la hora de explicar el movimiento de y_t.

Si Z_j y p_j son constantes, el movimiento no sería estacionario puesto que E(y_t) dependería del tiempo, lo que impediría la aplicación de este método, pero este problema puede evitarse suponiendo que alguno de esos elementos son variables aleatorias con las características apropiadas.

Por otra parte, dadas las propiedades de las relaciones trigonométricas, la expresión anterior puede ser escrita de forma equivalente como:

$$y_t = \sum_{j=1}^{k} (a_j \cos w_j t + b_j \text{sen}\, w_j t) + e_t \tag{10.90}$$

Puesto que debemos contemplar todas las frecuencias, no tiene sentido restringir el sumatorio anterior entre los límites 1 y k. Si en la expresión (2) hacemos tender k a infinito, puede demostrarse que cualquier proceso estacionario discreto puede representarse por:

$$y_t = \int_0^{\pi} \cos wt du(w) + \int_0^{\pi} \text{sen}\, wt dv(w) \tag{10.91}$$

La expresión anterior es la representación espectral del proceso y_t siendo $u(w)$ y $v(w)$ sendos procesos estocásticos estacionarios, incorrelados y con incrementos ortogonales. Aunque el límite superior en las integrales anteriores debería ser ∞, en el caso de procesos discretos medidos a intervalos unitarios de tiempo como los que solemos manejar en economía, no hay pérdida de generalidad en sustituirlo por π[7].

De acuerdo con (10.91) cada frecuencia comprendida en el rango $(0, \pi)$ puede contribuir a explicar la variación del proceso. Sin embargo, las integrales anteriores son matemáticamente complejas y difíciles de manejar, lo que unido al escaso interés práctico de los procesos u y v,resta atractivo a esta expresión. En su lugar se emplea el resultado de un teorema según el cual para todo proceso estocástico estacionario con función de autocovarianza γ_k, existe una función monótona creciente, $F(w)$ tal que[8]:

$$\gamma_k = \int_0^{\pi} \cos wk dF(w) \tag{10.92}$$

[7] Ver Chatfield (1996), p. 94.
[8] Teorema de Wiener-Khintchine.

La expresión anterior es la representación espectral de la función de autocovarianza que ya conocemos de los temas anteriores. La función $F(w)$ es la función de distribución espectral y representa la contribución a la varianza de todas las frecuencias comprendidas en $(0, \pi)$, es decir tiene la interpretación de una típica función de distribución estadística. Puesto que no hay variación para frecuencias negativas:

$$F(w) = 0 \text{ para } w < 0$$

Y dado que la máxima frecuencia es π, se deduce que:

$$F(w) = \text{var}(y_t), \text{ para } w = \pi,$$

resultado este último que también puede derivarse directamente de (10.92):

$$\gamma_0 = \int_0^\pi \cos wk dF(w) = \int_0^\pi dF(w) = F(\pi)$$

A veces, en lugar de $F(w)$, se emplea la función de distribución espectral normalizada, que viene dada por:

$$F^*(w) = \frac{F(w)}{\sigma_y^2} \tag{10.93}$$

La derivada de la función de distribución espectral[9], es decir:

$$f(w) = \frac{dF(w)}{dw} \tag{10.94}$$

es la función de densidad espectral o simplemente el *espectro*. Su interpretación es la propia de una función de densidad: $f(w)$ representa la contribución a la varianza de y_t de las frecuencias comprendidas en el rango (w, w+dw).

Combinando (10.93) y (10.94) se deduce que

$$\gamma_k = \int_0^\pi \cos wk f(w) d(w) \tag{10.95}$$

La expresión anterior expresa la relación entre la función de autocovarianza y la función de densidad espectral. Puede demostrarse que la relación inversa viene dada por[10]:

$$f(w) = \frac{1}{\pi}\left[\gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k \cos wk\right] \tag{10.96}$$

Ambas, es decir (10.95) y (10.96) ponen de relieve el hecho ya mencionado de que el espectro y la función de autocovarianza son formas equivalentes de analizar una serie temporal estacionaria.

[9] En los términos en los que está definida $F(w)$ y para todos los casos relevantes desde el punto de vista práctico, $F(w)$ es diferenciable.

[10] Utilizando el teorema de Moivre el espectro puede escribirse también como $f(w) = \frac{1}{\pi}\sum_{j=-\infty}^{\infty}\gamma_j e^{-iwj}$.

Igual que sucedía con la función de distribución espectral, en ocasiones la función de densidad espectral se emplea también en términos normalizados:

$$f*(w) = \frac{1}{\pi}\left[1 + 2\sum_{k=1}^{\infty}\rho_k \cos wk\right] \tag{10.97}$$

Para terminar conviene señalar que en la literatura pueden encontrarse otras definiciones diferentes del espectro. La mayoría de ellas difieren de (10.96) por una constante multiplicativa y/o por el rango de definición de w. Por ejemplo, es muy frecuente encontrar:

$$f(w) = \frac{1}{2\pi}\left[\gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k \cos wk\right]$$

No obstante a lo largo de este tema y mientras no se diga lo contrario, nos basaremos en la ecuación (10.96).

En la práctica, como sucedía en el caso de los modelos ARIMA, hemos de estimar el proceso generador de datos a partir de una muestra concreta. Un instrumento natural para estimar el espectro poblacional es el *peridograma muestral*, $I(w)$. Puede demostrarse que el periodograma está directamente relacionado con la función (muestral) de autocovarianza. Para una serie de tamaño T:

$$I(w_i) = \frac{1}{\pi}\left(c_0 + 2\sum_{k=1}^{T-1} c_k \cos w_i k\right) \tag{10.98}$$

Si disponemos de una muestra de T observaciones para la serie y_t, es posible ajustar una función que pase por todos sus puntos. Describiremos el proceso para el supuesto de que el número de observaciones T es par, aunque puede emplearse un procedimiento similar para el caso de que dispongamos de un número impar de observaciones.

Comenzamos definiendo el siguiente conjunto de frecuencias w_j, denominadas frecuencias de Fourier[11]:

$$w_j = \frac{2\pi j}{T}, \qquad j = 1, \ldots, T/2 \tag{10.99}$$

Por tanto, la frecuencia más alta que consideramos es $w = \pi$, también denominada frecuencia de Nyquist y corresponde a un periodo[12] de dos unidades de tiempo (el periodo mínimo en el análisis de ciclos).

El siguiente paso consiste en definir el par de términos trigonométricos cos($w_j t$) y sen($w_j t$) para cada una de estas frecuencias, siendo t el tiempo, es decir $t = 1, 2, \ldots, T$. Entonces puede demostrarse que:

$$y_t = \alpha_0 + \sum_{j=1}^{T/2}(\alpha_j \cos w_j t + \delta_j \sin w_j t) \tag{10.100}$$

[11] Si T es impar la diferencia consiste en que j varía entre 1 y (T-1)/2. En este caso no se anula la serie de seno correspondiente a último armónico.

[12] El periodo es el inverso de la frecuencia.

donde α_0 es la media de y_t. La ecuación anterior puede interpretarse como una ecuación de regresión múltiple en la que los términos cos($w_j t$) y sen($w_j t$) juegan el papel de variables explicativas. La parte derecha de la ecuación explica completamente la variación de la serie y_t, razón por la cual (10.100) no incluye término de error. No obstante, lo habitual es considerar, no el conjunto completo de las frecuencias de Fourier, sino un subconjunto más reducido, en cuyo caso se añadiría un término de error que supondremos tiene las características habituales.

Para la última frecuencia, la variable sen($w_j t$) es nula para todos los valores de t, dado que sen(π) = 0 y siendo t entero, también lo será sen($\pi \cdot t$). Teniendo en cuenta el resto de los términos y la constante, α_0, resulta que disponemos de T variables para explicar el movimiento de y_t, cuyo tamaño es asimismo T. No tiene pues ningún mérito que con tantas variables podamos explicar toda la variación.

Si escribimos el estimador MCO de la ecuación de regresión (16), en lenguaje matricial, tenemos:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} T & 0 & \cdots & 0 \\ 0 & T/2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & T \end{pmatrix}^{-1} \mathbf{X}'\mathbf{Y} \tag{10.101}$$

Es decir que **X'X** es una matriz diagonal, lo que significa que los elementos cos($w_j t$) y sen($w_j t$) son independientes unos de otros. Por lo tanto podemos eliminar cualquier conjunto de variables de la expresión (10.100) sin que ello afecte al resto de los coeficientes de las variables que permanecen. Es decir, podemos estimar de forma aislada el subconjunto de coeficientes que nos interesen.

Es relativamente fácil comprobar (ver ejercicios) que:

$$\hat{\alpha}_j = \frac{2}{T}\sum_{t=1}^{T} y_t \cos(w_j t) \tag{10.102}$$

$$\hat{\delta}_j = \frac{2}{T}\sum_{t=1}^{T} y_t \text{sen}(w_j t) \tag{10.103}$$

Por otra parte, la varianza de la serie y_t puede descomponerse en:

$$T^{-1}\sum_{t=1}^{T}(y_t - \bar{y})^2 = \frac{1}{2}\sum_{j=1}^{T/2}(\hat{\alpha}_j^2 + \hat{\delta}_j^2) \tag{10.104}$$

Este resultado es el conocido *teorema de Parseval* y permite afirmar que la contribución a la explicación de la varianza total de y_t del ciclo correspondiente a la frecuencia w_j, viene dada por $\frac{1}{2}(\hat{\alpha}_j^2 + \hat{\delta}_j^2)$.

La representación gráfica de las frecuencias (en abscisas) junto con su correspondiente contribución a la varianza (en ordenadas) recibe el nombre de *periodograma*. En este gráfico la contribución a la varianza suele expresarse en términos proporcionales. La existencia de *picos* en el periodograma, es decir valores de $I(w_j)$ mayores que los

adyacentes $I(w_{j-1})$ e $I(w_{j+1})$, se interpreta como evidencia de que los ciclos en esa periodicidad son relevantes a la hora de explicar la variación de la serie que estamos analizando. De manera que mediante este instrumento podemos encontrar los componentes cíclicos de una serie de tiempo estacionaria. Basta para ello observar en el periodograma, cuáles son las frecuencias que más contribuyen a explicar la variación de nuestra serie.

Las propiedades estadísticas de los coeficientes de Fourier pueden derivarse de la teoría estándar de mínimos cuadrados. Los elementos de $\hat{\beta}$ se distribuyen de forma normal (dado el supuesto de normalidad de ε_t) con media nula (se deduce de (10.102) y (10.103), dado que todos los elementos de esos sumatorios tienen media nula) y varianza dada por[13]

$$var(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{10.105}$$

El hecho de emplear el conjunto de frecuencias de Fourier es conveniente pero arbitrario. En efecto, en este modelo atribuimos toda la variación muestral de la serie a alguna de las $T/2$ frecuencias de Fourier. Si parte de la variación fuese debida a ciclos con frecuencias distintas de las consideradas en (10.99), dicha variación es automáticamente atribuida a alguna de esas $T/2$ frecuencias.

queda claro que el periodograma debe ser modificado si queremos disponer de un estimador consistente del espectro poblacional. Una posibilidad es suavizar el periodograma.

Podemos obtener una mejor aproximación al espectro mediante un suavizado que consistiese en promediar cada valor de $I(w_j)$ con las frecuencias adyacentes. Por ejemplo, podríamos emplear:

$$\hat{f}(w_j) = \frac{1}{m}\sum_{i=-m*}^{m*} I(w_{j-i}) \tag{10.106}$$

con $m = 2\ m* + 1$. Siguiendo un razonamiento similar al empleado anteriormente, encontramos que[14]:

$$\hat{f}(w_j) = \frac{\sigma^2}{4\pi m}\chi^2_{2m}$$

Siendo la varianza de este estimador:

$$var\left[\hat{f}(w_j)\right] = \frac{\sigma^4}{4\pi^2 m} \tag{10.107}$$

y aunque es también constante, una forma de conseguir que disminuya con T es hacer que m dependa del tamaño muestral. Si $m = \lambda T$, (obviamente $\lambda < 1$), entonces:

$$var\left[\hat{f}(w_j)\right] = \frac{\sigma^4}{4\pi^2 \lambda T} \tag{10.108}$$

[13] La matriz inversa, es decir $(\mathbf{X'X})^{-1}$ será también una matriz diagonal. Además, todos sus elementos excepto el primero, correspondiente a la constante y el último, correspondiente a la frecuencia $j = T/2$, tienen el mismo valor, $2/T$, de manera que la varianza de todos estos coeficientes será también la misma, en cada ecuación de regresión concreta. Para la constante y el coeficiente correspondiente a cos(πt) suponiendo que este último estuviera incluido, el valor de la varianza será exactamente la mitad que la del resto de los coeficientes, es decir $1/T$.

[14] La suma de m variables aleatorias independientes distribuidas como una $\chi^2(2)$, es una χ^2 con $2m$ g.l.

que tiende a cero a medida que T crece, es decir, proporcionando un estimador consistente. Aunque nos hemos basado en el proceso de ruido blanco para describir la idea, ésta puede generalizarse inmediatamente.

En el proceso de suavizado descrito por la expresión (10.106) solo se tiene en cuenta un subconjunto de valores del periodograma. Este grupo define lo que se conoce como *ventana*. La anchura de la ventana o rango[15], es precisamente el número de valores considerados, m.

La ponderación implícita en (10.106) otorga la misma importancia a cada uno de los m valores contemplados, pero es natural considerar otros sistemas en los que el peso disminuya con la distancia, es decir donde se otorgue más importancia a los valores de I próximos a la frecuencia de interés. En general:

$$\hat{f}(w_j) = \sum_{i=-m*}^{m*} h_i I(w_{j-i}) \tag{10.109}$$

donde la suma de los pesos h_i es la unidad y m =2$m*$ +1. La primera y última frecuencias, solo tienen elementos adyacentes por uno de los lados, por lo que supondremos que el periodograma es totalmente simétrico en esos puntos, asignando a los valores adyacentes exactamente el doble del peso considerado en el resto de los casos.

En consecuencia hay que tomar dos decisiones: elegir el valor de m y además, el sistema de ponderaciones. En cuanto a este último, en lugar de elegirse lo que se denomina una ventana rectangular (que otorga la misma ponderación a todos los valores considerados), se opta por una triangular (la ponderación disminuye con la distancia). Respecto al valor de m no hay muchos consejos prácticos en la literatura, aunque algunos autores recomiendan probar con diversos valores en el entorno de $T/40$.

Esta elección se ve comprometida por el hecho de que procedimientos como los descritos pueden proporcionar estimadores consistentes, pero al precio de introducir sesgo. En efecto, se deduce el sesgo de (10.106):

$$E[\hat{f}(w)] \approx m^{-1} \sum_i f(w_i) \tag{10.110}$$

que solo coincidirá con $f(w)$ si el espectro es lineal en el intervalo considerado, algo que solo garantiza el proceso de ruido blanco[16]. En general, cuanto mayor sea el valor de m, menor será la varianza del estimador, pero mayor el sesgo introducido (y viceversa).

Un sistema de ponderaciones utilizable en este contexto consiste en elegir los pesos de acuerdo con:

$$h_i = \sum_{i=-m*}^{m*} \left(\frac{m* + 1 - |i|}{(m* + 1)^2} \right) \tag{10.111}$$

[15] También se emplea el concepto de ancho de banda (*bandwidth*) que es la anchura expresada en radianes.

[16] No obstante, el sesgo puede carecer de importancia en la medida en que $f(w)$ sea una función razonablemente suavizada y m pequeño en relación a T.

Por ejemplo, para el supuesto más simple, $m*$ = 1 (h = 3), i toma los valores -1, 0 y 1 y los pesos son:

$$h_1 = \frac{1+1-1}{(1+1)^2} = \frac{1}{4}, \quad h_2 = \frac{1+1-0}{4} = \frac{1}{2} \text{ y } h_3 = \frac{1+1-1}{4} = \frac{1}{4}$$

Es decir, para cada w_j, $\hat{f}(w_j) = 0{,}25I(w_{j-1}) + 0{,}5I(w_j) + 0{,}25I(w_{j+1})$. Para la primera frecuencia no existe w_{j-1} de manera que $\hat{f}(w_1) = 0{,}5I(w_j) + 0{,}5I(w_{j+1})$. Análogamente, para la última $\hat{f}(w_\pi) = 0{,}5I(w_{j-1}) + 0{,}5I(w_j)$.

La alternativa a la suavización del periodograma consiste en ponderar la función de autocovarianza. Tomando la expresión (10.98) como punto de partida y teniendo en cuenta que la precisión de las c_k disminuye a medida que k aumenta, es razonable tener en cuenta este hecho al elegir el procedimiento de ponderación. Podemos considerar entonces un estimador como:

$$I(w_i) = \frac{1}{\pi}\left(h_0 c_0 + 2\sum_{k=1}^{M} h_k c_k \cos w_i k\right) \tag{10.112}$$

donde el *punto de truncamiento* $M < T$ y h_k es un conjunto de ponderaciones que se conoce con el nombre de *lagwindow*. El primero determina hasta qué valor de k vamos a considerar las autocovarianzas (las autocovarianzas para $k > M$ simplemente no se tienen en cuenta), y el segundo el peso que se otorga a las cada una de las M autocovarianzas consideradas. De manera que también aquí hay que tomar dos decisiones y también hay un *trade off* entre el tamaño de M y el valor del sesgo y la varianza.

Dos de los sistemas de ponderación más utilizados son la ventana de Tukey

$$h_k = \frac{1}{2}\left(1 + \cos\frac{\pi k}{M}\right), \quad k = 0, 1, ..., M \tag{10.113}$$

y la ventana de Parzen:

$$\begin{array}{ll} h_k = & 1 - 6\left(\frac{k}{M}\right)^2 + 6\left(\frac{k}{M}\right)^3, \quad 0 \le k \le M/2 \\ & 2(1-k/M)^3, \qquad\qquad M/2 < k \le M \end{array} \tag{10.114}$$

En cuanto al punto de truncamiento no hay tampoco muchas recomendaciones sugiriéndose $2\sqrt{T}$ como valor aproximado.

Bibliografía complementaria

Matilla-García, M et al. 2017. Econometría y Predicción. McGraw Hill

Tema 11

Modelos con tendencias

Este tema está elaborado como una adaptación del capítulo 4 de:

Enders, W. Applied Econometric Time Series. 4 ed. Wiley, y del capítulo 3-8 de:

Box, Jenkins, Reinsel y Ljung Time Series, Analysis, Forecasting and Control. 5th. ed. Wiley. Así como de la bibliografía complementaria

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al Órgano convocante ni al Tribunal actuante.

- Modelos con tendencias.
- Raíces Unitarias.
- Eliminación de tendencias.
- Contrastes de raíces unitarias.
- Cambio estructural.
- Tendencias y Descomposición Univariante

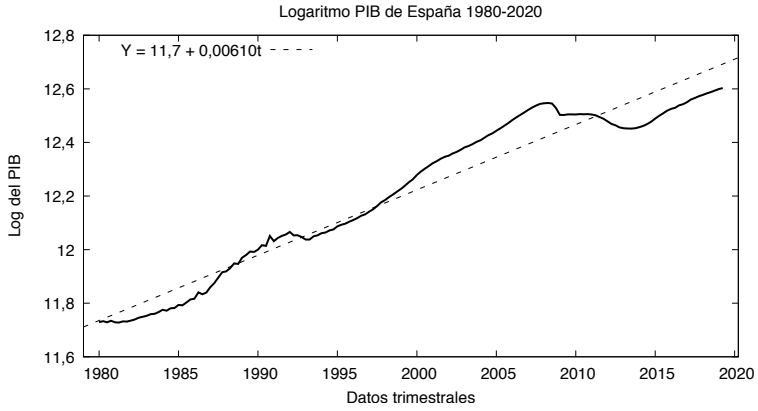
11.1 Modelos con tendencias

Cuando planteamos un modelo de regresión múltiple con datos en forma de serie temporal estamos imponiendo criterios que de alguna manera se acomoden en su conjunto a cierto tipo de estacionariedad: los parámetros deben ser estables en el tiempo, la varianza de los errores ha de ser constante, los errores deben estar no autocorrelacionados,..., por mencionar los más relevantes.

Muchas series macroeconómicas son no estacionarias y por tanto estos supuestos de estabilidad no siempre se satisfacen. Incluso algunas series siguen siendo no estacionarias después de eliminar una tendencia determinista temporal.

Cuando las variables utilizadas en el análisis de regresión son *no estacionarias*, sucede entonces que no podemos recurrir al Teorema Central del Límite y por tanto los resultados asintóticos que entonces derivábamos se encuentran seriamente comprometidos. En estas condiciones (incumplimiento de los supuestos clásicos), el utilizar en el análisis de regresión con series temporales no estacionarias puede llegar a ser crítico y las conclusiones de sus resultados lo más probable es que sean necesariamente erróneas. Es más, los estadísticos habituales tipo t y Durbin Watson, así como medidas como el

Figura 11.1: Tendencia del PIB



R-cuadrado, dejan de tener las características y el comportamiento esperado cuando trabajamos con datos no estacionarios. Hacer regresiones con este tipo de datos puede fácilmente conducirnos a regresiones que «informan» de una relación cuando realmente tal supuesta relación es inexistente.

La Figura 11.1 representa el PIB de España desde 1980 hasta principios de 2020, en escala logarítmica. Se aprecia que la media no es constante (el proceso no es estacionario) y que la serie tiene un movimiento persistente creciente a lo largo del tiempo. Esa persistencia podría estar recogida en la línea recta que aparece representada en trazo discontinuo, mientras el (log) PIB, fluctúa en torno a la misma. Dicho movimiento persistente de la variable a largo plazo es a lo que nos referimos por tendencia.

La persistencia posiblemente se deba a que las tasas de crecimiento de algunas variables económico-empresariales (o sociales) provienen de un proceso con ciertas características que son intrínsecamente estables. Dichas características están reflejadas en el tipo de tendencia. Por ejemplo, la serie del PIB en niveles podría decirse que es compatible con un crecimiento a una tasa constante manteniendo algunas desviaciones a lo largo de su historia. Por este motivo el logaritmo de la misma crece linealmente. El modelo de crecimiento exponencial $Y_t = e^{\beta_0 t}$ en el que la variable Y_t es en este caso el PIB crece a una tasa $d(Y_t)/dt = \beta_0 Y_t$, determinada por la constante β_0. De hecho, tomando logaritmos se tiene $ln(Y_t) = \beta_0 t$, que permite verificar el crecimiento lineal de la serie en logaritmos. Por este motivo es muy conveniente trabajar con la serie en escala logarítmica.

En el caso del PIB español la tendencia podría modelizarse mediante una función no aleatoria del tiempo, esto es mediante una tendencia que se denomina «determinista». De hecho es lo que hemos hecho en la Figura 11.1 a través de la recta dibujada en trazo discontinuo. Si el logaritmo del PIB tuviera una **tendencia determinista**, indicaría que el PIB tendría una tendencia exponencial del tipo $Y_t = Y_0 e^{\beta_0 t}$, como hemos visto. El término $0{,}0061t$ indica que cada trimestre el PIB crecería un 0,61 %. En general se tendrá que la diferencia $\Delta \ln Y_t = \beta_0$. El comportamiento en torno a la tendencia determinista lo incorporamos con un término error contemporáneo, es decir,

$$Y_t = Y_0 e^{\beta_0 t} \varepsilon_t \Rightarrow \ln Y_t = Y_0 + \beta_0 t + \ln \varepsilon_t. \tag{11.1}$$

En esta expresión, el logaritmo del PIB puede diferir del valor marcado por su tendencia en una cantidad indeterminada recogida en el error ($\ln\varepsilon_t$). Dado que el error es estacionario, la serie del logaritmo del PIB se alejará puntualmente de su tendencia, pero a medida que transcurra el tiempo convergerá a la línea (función) de tendencia $Y_0 + \beta_0 t$. Puesto en otros términos, en los modelos de tendencia determinista, el término error afecta a lo que está pasando en el periodo actual, pero deja de tener efectos en los periodos siguientes.

Veamos ahora algunos tipos de tendencias deterministas. Una tendencia determinista, por tanto, consiste en modelizar el componente persistente mediante una función no aleatoria del tiempo. También se suele denominar a estos procesos no estacionarios como *procesos estacionarios en tendencia*. Están compuestos por una parte estocástica que es estacionaria y por una parte no estacionaria, que en este caso es determinista. En general tendremos

$$Y_t = f(t, \boldsymbol{\beta}) + \text{estacionario}$$

Las formas que toma la parte no estacionaria y determinista suelen ser lineales, exponenciales o cuadráticas.

Obtenemos una **tendencia lineal** si calculamos por MCO la siguiente expresión:

$$\hat{Y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 t, \text{ para } t = 1, 2, ...;$$

obtenemos una **tendencia exponencial** mediante

$$\widehat{\ln Y_t} = \widehat{\beta}_0 + \widehat{\beta}_1 t, \text{ para } t = 1, 2, ...,$$

y una **tendencia cuadrática** calculando la siguiente expresión:

$$\hat{Y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 t + \widehat{\beta}_2 t^2, \text{ para } t = 1, 2, ... \tag{11.2}$$

Alternativamente, podríamos considerar otro proceso de modelización para el mismo PIB en el cual el crecimiento (y por tanto la tendencia) pudiera cambiar en el tiempo de modo que el 0,61 % fuera una tendencia «en promedio», cambiando de trimestre en trimestre. Esto es equivalente a considerar que el cambio del 0,61 % sería de naturaleza estocástica y se cumpliría en media, es decir

$$\Delta \ln Y_t = \beta_0 + \ln\varepsilon_t.$$

A este tipo de **tendencia** se le denomina **estocástica** donde el cambio en la variable dependiente en ocasiones estará por encima de β_0 y en otras por debajo, y dado que el error tiene esperanza nula, el modelo prevé que el cambio esperado será de un $100 \times \beta_0\,\%$. En el caso del PIB tendríamos que el modelo sería

$$Y_t = Y_{t-1} e^{\beta_0} \varepsilon_t \Rightarrow \ln Y_t = \ln Y_{t-1} + \beta_0 + \ln\varepsilon_t.$$

En este modelo de tendencia estocástica, el PIB crecería un 0,61 % ($100 \times \beta_0\,\%$) respecto del trimestre anterior.

Para poder comparar mejor ambos modelos de tendencia (estocástico y determinista) sustituimos recursivamente

$$\ln Y_t = \ln Y_{t-1} + \beta_0 + \ln \varepsilon_t = \ln Y_{t-2} + 2\beta_0 + \ln \varepsilon_t + \ln \varepsilon_{t-1} = ... =$$

$$\ln Y_t = \ln Y_0 + \beta_0 t + \sum_{i=1}^{t} \ln \varepsilon_i. \tag{11.3}$$

Esta expresión del modelo de tendencia estocástica se puede comparar ahora fácilmente con el modelo de tendencia determinista de la expresión (11.1). El modelo (11.3) está formado por un componente determinista $\beta_0 t$ y por un componente estocástico $\ln Y_0 + \sum_{i=1}^{t} \ln \varepsilon_i$, que puede interpretarse como un intercepto estocástico. En ausencia de shocks el intercepto sería $\ln Y_0$. En caso contrario, cada shock representa un cambio estocástico de la constante. Es importante observar que dado que todos los valores de los errores tienen asignado un coeficiente unitario, entonces el efecto sobre el intercepto es permanente, y por tanto la esperanza condicionada del proceso es un efecto de naturaleza permanente aunque estocástica. Este efecto permanente contrasta con el modelo de tendencia determinista donde el efecto era puntual y desaparecía con el paso del tiempo.

En el caso del PIB, en un modelo de tendencia estocástica, un shock generará efectos económicos permanentes, mientras que en el caso determinista sucederá que el shock será transitorio. Naturalmente esto tiene profundas implicaciones para la teoría y política macroeconómica ya que el diseño de una política dependerá crucialmente de saber si los efectos perdurarán o por el contrario se diluirán con el paso del tiempo.

11.2 Raíces Unitarias

La tendencia estocástica no se caracteriza por presentar una media que cambia a una tasa constante en el tiempo. Podemos definir la tendencia estocástica a partir de una función aleatoria (estocástica) del tiempo: la media no se mantiene constante y, en contraste con la tendencia determinista, su cambio es impredecible.

La tendencia estocástica considera que el componente tendencial de las series económicas es aleatorio y por tanto la variación (incremento o decremento) en la media del proceso es de naturaleza estocástica. Recordemos que la tendencia determinista considera sin embargo que tal variación es constante cada periodo.

La forma técnica de introducir una tendencia con un comportamiento estocástico como el que describimos en lo párrafos anteriores es mediante una **raíz unitaria**. Un proceso estocástico temporal tiene una raíz unitaria cuando una de las raíces del polinomio de retardos es unitaria. El caso más sencillo es este

$$X_t = \phi_1 X_{t-1} + \varepsilon_t, \phi_1 = 1, \varepsilon_t \text{ es } iid.$$

El lector podrá recordar que consideramos que las series podían presentar tendencias de este tipo en la media, y desaparecían tras tomar diferencias (*procesos estacionarios en*

diferencias), tal es el caso:

$$(1-B)X_t = Z_t, \text{ donde } Z_t \text{ es un proceso estacionario.}$$

Esta representación es indicativa de que *la variación o el cambio de* X en el tiempo se representa mediante un proceso estacionario Z. De ahí que dijéramos entonces que la serie X era «integrada» puesto que si Z es el cambio o variación de X, entonces X es la «suma», a lo largo de t, de Z. De hecho una serie no estacionaria se considera integrada de orden uno, $I\,(1)$, si para obtener una serie estacionaria (desestacionarizar la serie) se requiere aplicar una primera diferencia sobre la serie original. Dado que Z_t es una serie estacionaria y X_t a la serie original es $I\,(1)$, entonces

$$X_t = X_{t-1} + Z_t. \tag{11.4}$$

Comparando esta expresión con la formulación de la tendencia determinista, se observa que hay una parte estacionaria, Z_t, y ahora el término equivalente a la parte determinista de la tendencia viene recogido en X_{t-1}, que ya no es determinista.

Un caso ilustrativo es aquel donde, por ejemplo, la parte estacionaria es $Z_t = \beta_0 + v_t$ con v_t una variable con esperanza nula, y X_t tiene una tendencia estocástica en forma de proceso $I(1)$:

$$\begin{aligned} X_t &= X_{t-1} + \beta_0 + v_t \\ Y_t &= X_t + \varepsilon_t,\ \varepsilon_t \text{ es } iid. \end{aligned} \tag{11.5}$$

Este caso nos permite preguntarnos sobre cuál es la contribución de la tendencia temporal estocástica sobre la variable Y. La respuesta es que la contribución de la tendencia es $X_t - X_{t-1} = \beta_0 + v_t$, por lo que claramente apreciamos que ya no es una constante β_1, sino que la contribución de la tendencia es de naturaleza estocástica. Esto supone que la tendencia será una variable aleatoria y tendrá una media y una varianza, que en su momento consideraremos.

El papel que desempeñan las perturbaciones aleatorias en los modelos con tendencia derterminista y estocástica es claramente diferente y tiene implicaciones importantes. Para comprobarlo consideramos nuevamente el modelo con tendencia determinista lineal (11.11) y el modelo con tendencia estocástica (11.5). En el caso determinista el cambio de Y de un periodo a otro consecutivo es

$$Y_t - Y_{t-1} = \beta_0 + \beta_1 t + \varepsilon_t - \beta_0 - \beta_1(t-1) - \varepsilon_{t-1} = \beta_1 + \varepsilon_t - \varepsilon_{t-1},$$

es decir, la perturbación producida en $t-1$ que nos alejó de la línea o senda $(\beta_0 + \beta_1 t)$, esto es ε_{t-1}, desaparece en el perido t, revertiendo de este modo Y a su senda, y haciendo que el efecto sea transitorio. Por el contrario, el cambio en Y en el modelo con tendencia estocástica (11.5) sería

$$Y_t - Y_{t-1} = X_t - X_{t-1} + \varepsilon_t - \varepsilon_{t-1} = \beta_0 + v_t + \varepsilon_t - \varepsilon_{t-1},$$

al igual que en el caso determinista, el efecto de ε_{t-1} sobre Y desaparece cuando llega la perturbación ε_t, es decir en el periodo t. Sin embargo en el periodo t, cuando se produce

el efecto de v_t, no desaparece el correspondiente de v_{t-1}, y por tanto Y no revierte y no regresa a su senda.

Retomemos ahora, para completar y contrastar, una variante no estacionaria del modelo de la ecuación (11.12), particular cuando $\rho = 1$. Ahora la parte estocástica ya no es estacionaria al presentar una raíz unitaria

$$Y_t = \beta_0 + \beta_1 t + Z_t$$

$$Z_t = Z_{t-1} + \varepsilon_t.$$

La esperanza matemática condicionada (la predicción s periodos adelante) se obtiene haciendo $\rho = 1$ en (11.13), con lo que se tiene

$$\begin{aligned} \mathbb{E}(Y_{t+s} \,|Y_t, Y_{t-1}, \ldots) &= \beta_0 + \beta_1 (t+s) + Z_t \\ &= \beta_0 + \beta_1 (t+s) + Y_t - \beta_0 - \beta_1 t \\ &= \beta_1 s + Y_t. \end{aligned}$$

A diferencia del modelo con tendencia determinista, ahora el valor actual de Y tiene un efecto permanente en la predicción futura para todos los horizontes temporales.

Esta exposición ha dejado clara la relevancia de distinguir entre tendencias deterministas y estocásticas. En términos gráficos, en un proceso con tendencia determinista, las desviaciones con respecto a la tendencia son puramente aleatorias y se corrigen rápidamente. El movimiento a largo plazo de la serie está completamente determinado por el componente determinista, es decir, por la tendencia. Por el contrario, en el caso de una tendencia estocástica, el componente aleatorio es mucho más persistente y sí afecta al movimiento a largo plazo. Para empeorar más las cosas, es posible que un proceso presente a la vez los dos tipos de tendencia. Más adelante se presenta un contraste estadístico para distinguir entre estas posibilidades.

A continuación presentamos algunos procesos con tendencias estocásticas relevantes y útiles. Como hemos visto en los casos contemplados anteriormente, el introducir términos autorregresivos es una forma sencilla de representación de tendencias estocásticas.

Paseos aleatorios

Un proceso estocástico importante, y que hemos utilizado ya en este tema, es el conocido por *paseo aleatorio* cuya estructura es la siguiente:

$$Y_t = Y_{t-1} + \varepsilon_t, \tag{11.6}$$

donde suponemos que ε_t es ruido blanco independiente, es decir que $\mathbb{E}(\varepsilon_t) = 0$, $\text{var}(\varepsilon_t) = \sigma_\varepsilon^2$ y $\rho(\varepsilon_t, \varepsilon_{t+u}) = 0$ para $u > 0$.

Intuitivamente, un paseo aleatorio se caracteriza porque el valor de la serie «mañana» es el valor que toma «hoy» y se altera por una variable impredecible.

Realizando sustituciones sucesivas, podemos obtener el modelo de paseo aleatorio como la suma de variables puramente aleatorias,

$$Y_t = \varepsilon_t + \varepsilon_{t-1} + ... + \varepsilon_1 + Y_0 = Y_0 + \sum_{i=0}^{T-1} \varepsilon_{t-i}, \tag{11.7}$$

y aplicando esperanzas no condicionadas tenemos que:

$$\mathbb{E}(Y_t) = \mathbb{E}(Y_0), \tag{11.8}$$

por lo que la esperanza no depende del tiempo t sino de las condiciones iniciales del proceso, y bajo el supuesto usual de que el proceso comienza con el valor cero, $Y_0 = 0$, la esperanza del proceso sería cero para todo t, $\mathbb{E}(Y_t) = 0$. Lo principal es que el valor está presente en el proceso, y no desaparece a lo largo de los distintos periodos. También observamos que las innovaciones del proceso ε_t se acumulan en el componente $\sum_{i=0}^{T-1} \varepsilon_{t-i}$, por lo que un shock o innovación en t tendrá el esperado efecto permanente.

La varianza no condicionada es

$$\text{var}(Y_t) = \text{var}(\varepsilon_t) + \text{var}(\varepsilon_{t-1}) + ... + \text{var}(\varepsilon_1) + \text{var}(Y_0) = \sigma_\varepsilon^2 t, \tag{11.9}$$

de manera que la varianza depende del tiempo t, aumentando a medida que transcurre, por consiguiente, el proceso paseo aleatorio no es estacionario en varianza, lo que es indicativo de que la incertidumbre sobre la situación del proceso crece con t. Comprobamos entonces que el ***paseo aleatorio no es estacionario***. Sin embargo, como hemos visto, la diferenciación del proceso nos devolvería un proceso estacionario.

Además el comportamiento del proceso paseo aleatorio es persistente en covarianza, esto lo podemos comprobar calculando la predicción para h periodos en el futuro a partir del valor del momento actual Y_t,

$$Y_{t+h} = \varepsilon_{t+h} + \varepsilon_{t+h-1} + .. + \varepsilon_{t+1} + Y_t,$$

donde incluimos el término Y_t por ser el último valor conocido.

Su valor esperado condicionado es

$$\mathbb{E}(Y_{t+h} \mid Y_t) = Y_t, \text{ para } h \geq 1,$$

de manera que con independencia de lo lejano que sea el periodo de predicción h, la mejor predicción es su valor actual Y_t.

La función de autocovarianza con u desfases γ_u del proceso paseo aleatorio es

$$cov(t, t+u) = \mathbb{E}\left((Y_t - Y_0)(Y_{t+u} - Y_0)\right) = \mathbb{E}\left(\sum_{i=0}^{T-1} \varepsilon_{t-i} \sum_{j=0}^{T+u-1} \varepsilon_{t-j}\right) = \sigma_\varepsilon^2 t,$$

por lo que la autocovarianza varía a lo largo de t.

La función de autocorrelación con u retardos es, en consecuencia,

$$\rho_u = \frac{\mathbb{E}\left(Y_t Y_{t+u}\right)}{\sqrt{\text{var}\left(Y_t\right)}\sqrt{\text{var}\left(Y_{t+u}\right)}} = \frac{\sigma_\varepsilon^2 t}{\sqrt{\sigma_\varepsilon^2 t}\sqrt{\sigma_\varepsilon^2\left(t+u\right)}} = \frac{t}{\sqrt{t\left(t+u\right)}} = \left(\frac{t}{t+u}\right)^{1/2},$$

para valores de t grandes $t/\left(t+u\right)$ será cercano a uno y ρ_u decrecerá aproximadamente de forma lineal; por consiguiente la función de autocorrelación (FAT) de un proceso paseo aleatorio decrece de forma lineal y no de forma geométrica como requieren los procesos estacionarios (por tanto el proceso no es débilmente dependiente en covarianza, sino persistente).

Además las autocovarianzas no dependen solo del desfase, como ocurre en los procesos estacionarios, con u retardos la función de autocovarianza γ_{-u} es

$$\gamma_{-u} = \mathbb{E}\left(Y_t Y_{t-u}\right) = \mathbb{E}\left(\sum_{i=0}^{T-1}\varepsilon_{t-i}\sum_{j=0}^{T-u-1}\varepsilon_{t-j}\right) = \sigma_\varepsilon^2\left(t-u\right),$$

de manera que la función de autocovarianza no depende solo del desfase,

$$\gamma_{-u} = \sigma_\varepsilon^2\left(t-u\right) \neq \gamma_u = 2\sigma_\varepsilon^2 t.$$

La persistencia en covarianza es una cuestión importante desde el punto de vista económico. Si el PIB es fuertemente persistente en covarianza, el PIB de los próximos años puede estar muy correlacionado con el PIB de no pocos años atrás. En consecuencia debemos tener siempre en cuenta que las políticas económicas que causan una variación del PIB actual puede tener efectos durante muchos años.

Esta persistencia es independiente de si el proceso tiene o no tendencia. De hecho es posible tener series altamente persistentes (como el PIB, la tasa de inflación, la tasa de desempleo, o incluso tipos de interés de los bancos centrales) que también puedan tener tendencia (el PIB suele tener una tendencia creciente, pero las series de inflación o el desempleo las tendencias no son tan evidentes). Por este motivo es interesante introducir otro tipo de tendencia estocástica que construimos a partir del paseo aleatorio que hemos presentado.

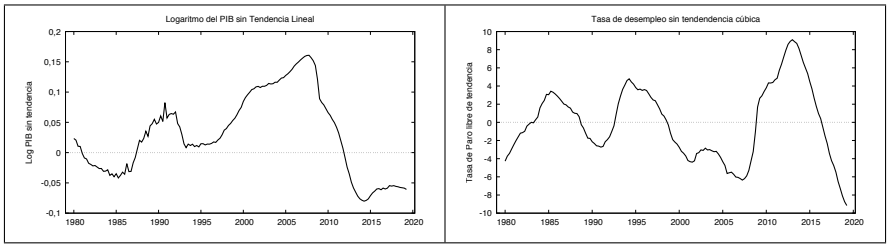
En efecto, hemos visto que el proceso paseo aleatorio es estacionario en media pero no en varianza ni en covarianza. El proceso más sencillo de tendencia estocástica que no es estacionario en media ni varianza es el denominado **paseo aleatorio con deriva**. Analíticamente el proceso es

$$Y_t = \beta_0 + Y_{t-1} + \varepsilon_t, \tag{11.10}$$

cuya única diferencia con el paseo aleatorio sin deriva, expresión (11.6), es la inclusión del término constante.

Analíticamente la inclusión o no de un término constante en series estacionarias no es importante; de hecho en los temas anteriores hemos restado a las observaciones su

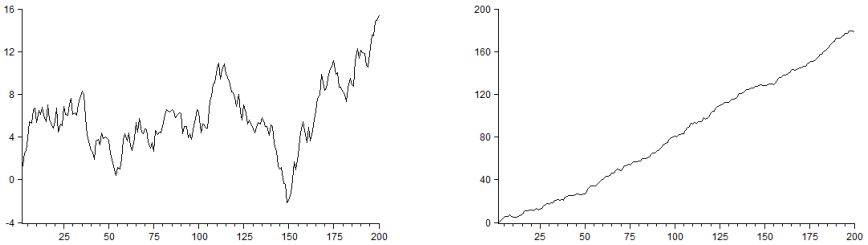
Figura 11.3: Series libres de tendencia determinista



media y hemos trabajado con procesos estacionarios con media cero sin pérdida de generalidad. Sin embargo, en los procesos no estacionarios si hay término constante, estos son importantes analíticamente y nos proporcionan una propiedad permanente del proceso no estacionario. Así si el término constante es igual a la unidad $\beta_0 = 1$, decimos que el proceso presenta una deriva unitaria y el proceso marca una tendencia lineal determinista con pendiente también unitaria. La Figura 11.2 reproduce un paseo aleatorio, gráfico a), y un paseo aleatorio con deriva unitaria, gráfico b).

Figura 11.2: Paseo aleatorio

a) $Y_t = Y_{t-1} + \varepsilon_t$ con $Y_0 = 0$ b) $Y_t = 1 + Y_{t-1} + \varepsilon_t$ con $Y_0 = 0$



En el gráfico a) paseo aleatorio, la serie es aproximadamente estacionaria en media pero la varianza va aumentando con el tiempo. En el gráfico b) paseo aleatorio con deriva unitaria, la tendencia es creciente y en consecuencia no es estacionario en media. Estas realizaciones artificiales nos recuerdan a las series de la tasa de desempleo y la del logaritmo del PIB. En efecto una tendencia estocástica del tipo «paseo aleatorio» para la tasa de desempleo y del tipo «paseo aleatorio con deriva» para el logaritmo del PIB parecen dos formas de modelizar las tendencias, respectivamente.

11.3 Eliminación de tendencias.

En el caso de series con tendencias deterministas, es decir cuando Y_t tiene una tendencia lineal, exponencial o cuadrática (o, en su caso, polinómica), entonces debe suceder que la serie libre de tendencia, $Y_t - \widehat{Y}_t = Z_t$, es estacionaria $[\mathbb{E}(Z_t) = 0,\ \text{var}(Z_t) = \sigma_Z^2, \rho_u = \rho_{-u}]$. En los paneles de la Figura 11.3 se aprecian las series del PIB y de la Tasa de Paro sin las correspondientes tendencias deterministas.

La parte indicada con la etiqueta «estacionario» podría ser cualquier proceso estacionario estudiado en los temas anteriores, por tanto podríamos indicarla, en términos generales, como $\Psi(B)\varepsilon_t$, siendo $\Psi(B)$ el correspondiente polinomio de retardos. De hecho cuando queramos evitar utilizar variables no estacionarias, podremos trabajar con las expresiones generales del tipo

$$\Phi(B)(Y_t - f(t, \boldsymbol{\beta})) = \Theta(B)\varepsilon_t$$

donde los polinomios respectivos son indicativos de la parte autorregresiva y de media móvil, respectivamente.

Un caso muy sencillo para la parte estacionaria y determinista es el de una tendencia lineal con ruido blanco

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t, \varepsilon_t \text{ es } iid, \tag{11.11}$$

y a partir del mismo podemos preguntarnos cuál es la contribución de la tendencia temporal sobre la variable Y: en este caso comprobamos que la contribución es exactamente la constante β_1:

$$\mathbb{E}(Y_t) - \mathbb{E}(Y_{t-1}) = \beta_0 + \beta_1 t - \beta_0 - \beta_1(t-1) = \beta_1,$$

es decir cada periodo (mes, año, trimestre,...) Y varía β_1 unidades.

Otro ejemplo algo más elaborado consiste en considerar un proceso AR(1) en la parte estacionaria:

$$Y_t = \beta_0 + \beta_1 t + \rho Y_{t-1} + \varepsilon_t, |\rho| < 1 \tag{11.12}$$

por tanto $\Theta(B) = 1$. En este caso, obviamente, el proceso no estacionario Y_t es «estacionario en tendencia» puesto que la parte $\rho Y_{t-1} + \varepsilon_t$, que denotaremos por Z_t, es estacionaria.

A partir de este modelo resulta ilustrativo observar el comportamiento de las predicciones s periodos hacia adelante:

$$\begin{aligned}
\mathbb{E}(Y_{t+s}|Y_t, Y_{t-1}, ...) &= \mathbb{E}(\beta_0 + \beta_1(t+s)|Y_t, Y_{t-1}, ...) + \mathbb{E}(Z_{t+s}|Y_t, Y_{t-1}, ...) \\
&= \beta_0 + \beta_1(t+s) + \mathbb{E}(Z_{t+s}|Y_t, Y_{t-1}, ...) \\
&= \beta_0 + \beta_1(t+s) + \mathbb{E}(Z_{t+s}|Z_t, Z_{t-1}, ...) \\
&= \beta_0 + \beta_1(t+s) + \rho^s Z_t
\end{aligned} \tag{11.13}$$

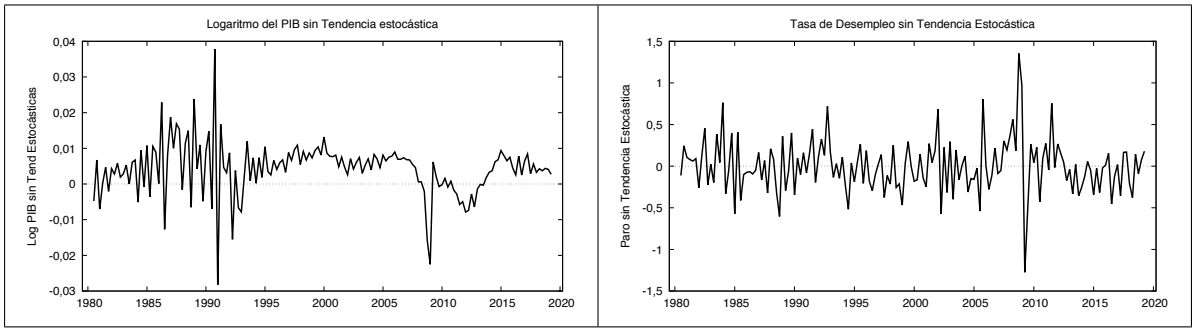
donde la penúltima igualdad es correcta puesto que los vectores $(Y_t, Y_{t-1}, ...)$ y $(Z_t, Z_{t-1}, ...)$ son informativamente idénticos, y la última igualdad se obtiene a partir de calcular la esperanza condicionada sobre $Z_{t+s} = \varepsilon_{t+s} + \rho\varepsilon_{t+s-1} + ... + \rho^{s-1}\varepsilon_{t-1} + \rho^s Z_t$.

De particular interés es comprobar que entonces la predicción s-periodos adelante, con un horizonte suficientemente largo, converge a la tendencia lineal $\beta_0 + \beta_1(t+s)$. En otros términos, podemos decir que los valores pasados y presentes de Y no afectan a la predicción. Como veremos posteriormente esto es un hecho diferencial respecto a los procesos con tendencias estocásticas.

Algo similar ocurre si $\beta_0 = \beta_1 = 0$. En esta situación

$$\mathbb{E}(Y_{t+s}|Y_t, Y_{t-1}, ...) = \rho^s Z_t = \rho^s(\rho Y_{t-1} + \varepsilon_t)$$

Figura 11.4: Series libres de tendencias estocásticas



que nos permite comprobar que: (i) el efecto de los shocks o innovaciones, ε_t, tienden a desparecer a medida que pasa el tiempo, es decir, tienen un efecto transitorio sobre la predicción (sobre la media condicionada); (ii) el efecto sobre la media condicionada (predicción) del valor inicial de Y_{t-1} también desaparece con el paso del tiempo.

Por último, y desde una perspectiva más aplicada, si las variables presentasen únicamente tendencia determinista lineal, la eliminación de dicho movimiento puede llevarse a cabo regresando dicha variable con respecto al tiempo

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 t.$$

Los residuos de esta estimación $\hat{u}_{yt}$ representarían entonces la serie libre de tendencia y, como tal, podría ser incluida en el modelo de regresión. Por ejemplo, en el caso de dos variables X e Y, la regresión se calcularía como:

$$\hat{u}_{yt} = \hat{\delta}_0 + \hat{\delta}_1 \hat{u}_{xt}.$$

Alternativamente podemos efectuar la regresión entre los valores originales, incluyendo el tiempo como un regresor adicional. Puede demostrarse que en este caso el estimador es el mismo, es decir

$$\hat{Y}_t = \hat{\gamma}_0 + \hat{\delta}_1 X_t + \hat{\delta}_2 t,$$

resultado que se generaliza para considerar el supuesto de que haya más variables explicativas. Por consiguiente, cuando hay variables con tendencias deterministas puede resultar conveniente incluir el tiempo como un regresor más en el modelo de regresión.

Una tendencia estocástica se elimina diferenciando la serie, como se deduce inmediatamente de la expresión (11.4). En la Figura 11.4 representamos las series sin las correspondientes tendencias estocásticas. Resulta ilustrativo comparar las series libres de tendencias en el caso de la modelización con tendencias estocásticas frente a las deterministas.

Ciertas propiedades relevantes del proceso paseo aleatorio con deriva las podemos considerar escribiendo el proceso como suma de variables aleatorias; realizando sustituciones sucesivas tenemos:

$$Y_t = t\beta_0 + Y_0 + \sum_{i=0}^{T-1} \varepsilon_{t-i}, \tag{11.14}$$

cuya esperanza es

$$\mathbb{E}(Y_t) = t\beta_0 + Y_0, \tag{11.15}$$

de manera que el valor esperado depende de t por lo que no es constante en el tiempo. Si el término constante es positivo $\beta_0 > 0$ la media es creciente, y si es negativo $\beta_0 < 0$ la media se reduce con el tiempo. Ahora podemos comprobar qué sucede si a este proceso le restamos una tendencia linal temporal, $t\beta_0$. En tal caso tendremos que

$$Y_t - t\beta_0 = Y_0 + u_t, \text{ donde } u_t = \textstyle\sum_{i=0}^{T-1} \varepsilon_{t-i},$$

de modo que la varianza del proceso de raíz unitaria (paseo aleatorio con deriva) cuando le restamos una tendencia determinista tiene una varianza $\sigma^2 t$, que crece con la fecha en la que tiene lugar la observación. Comprobamos entonces que en presencia de tendencia estocástica es poco afortunado quitar una tendencia determinista.

Siguiendo un proceso similar al utilizado para el proceso paseo aleatorio sin deriva llegamos a los mismos resultados para la varianza, autocovarianzas y autocorrelaciones del proceso paseo aleatorio con deriva. Por consiguiente, el proceso paseo aleatorio con deriva no es estacionario en varianza, $\text{var}(Y_t) = \sigma_\varepsilon^2 t$, la autocorrelación decrece linealmente, $\rho_u = [t/(t+u)]^{1/2}$, y la función de autocorrelación no depende solo del desfase, $\gamma_u \neq \gamma_{-u}$.

El valor del proceso para h periodos en el futuro teniendo en cuenta el valor de la variable actual Y_t es

$$Y_{t+h} = h\beta_0 + \varepsilon_{t+h} + \varepsilon_{t+h-1} + .. + \varepsilon_t + Y_t, \tag{11.16}$$

y su esperanza

$$\mathbb{E}(Y_{t+h} \mid Y_t) = h\beta_0 + Y_t, \tag{11.17}$$

por tanto, la mejor predicción es el último valor conocido Y_t más la deriva $h\beta_0$.

11.4 Contrastes de raíces unitarias

A tenor de lo elaborado hasta el momento en este tema, es evidente la importancia de saber si una serie es estacionaria o no, y en particular de si tiene una tendencia y en su caso de si la forma de la misma es la de una raíz unitaria. Hasta ahora hemos determinado si una serie temporal es estacionaria a partir de su gráfica y su función de

autocorrelación muestral, de tal manera que si la gráfica presenta un nivel estable en el tiempo, entonces decimos que parece estacionaria en media; y si la variabilidad es aproximadamente estable, entonces decimos que la serie temporal parece estacionaria en varianza; y si además la función de autocorrelación muestral decrece geométricamente, entonces decimos que la serie es estacionaria también en covarianza. En tales casos podemos concluir indicando que la serie temporal es compatible con un proceso estacionario, es decir, podría ser descrita y aproximada por un proceso estacionario de los contemplados y expuestos en apartados y temas anteriores.

También hemos visto la relevancia de distinguir entre procesos que se hacen estacionarios tras eleminar una tendencia determinista de aquellos en los que es necesario trabajar con la serie transformada en diferencias para hacerla estacionaria. Si partiendo de series no estacionarias erramos en la transformación necesaria a fin de hacerla estacionaria, tendremos potenciales problemas. Ante esta situación, una primera acción del económetra podría ser comparar los residuos de los modelos con ajustes de tendencia con los modelos que proceden de la diferenciación de la serie. Siguiendo la metodología clásica de Box y Jenkins preferiríamos la transformación que lleva a los residuos con una función de autocorrelación más simple.

Este procedimiento en ocasiones no es suficiente y si seguimos teniendo dudas sobre la existencia o no de una tendencia estocástica, dada su repercusión a todos los niveles (técnicos y de interpretación económica) podemos recurrir a procedimientos estadísticos formales que contrastan la existencia de tendencia estocástica frente a la hipótesis alternativa de que la serie es estacionaria.

Los contrastes más conocidos son los propuestos por Dickey-Fuller (DF) para el contraste de raíces unitarias. No es la única metodología estadística de contraste de raíces unitarias, pero es una de las más usuales, y tiene la ventaja de que los programas especializados incorporan todos sus contrastes y los calculan de forma rutinaria.

Inicialmente consideremos el caso más simple de modelo, el AR(1)

$$Y_t = \rho Y_{t-1} + \varepsilon_t, Y_0 = 0, \varepsilon_t \sim N(0, \sigma^2).$$

Cuando $\rho = 1$, el modelo es un paseo aleatorio sin deriva. La estimación MCO de ρ sería, como sabemos

$$\hat{\rho} = \left(\sum_{t=1}^{T} Y_{t-1} Y_t \right) \left(\sum_{t=1}^{T} Y_{t-1}^2 \right)^{-1}.$$

Dado que se trata -en este caso- de un modelo con errores homocedásticos y normales, sabemos que

$$\sqrt{T} \left(\hat{\rho} - \rho \right) \xrightarrow{d} N \left(0, \sigma^2 \left(\mathbb{E} \left(Y_{t-1}^2 \right) \right)^{-1} \right)$$

y dado que se trata de un AR(1), sabemos que estos procesos tienen una autocovarianza $\mathbb{E}\left(Y_{t-1}^2\right) = \sigma^2 / \left(1 - \rho^2\right)$, y por tanto se tiene que

$$\sqrt{T} \left(\hat{\rho} - \rho \right) \xrightarrow{d} N \left(0, \left(1 - \rho^2 \right) \right).$$

Bajo la hipótesis nula de raíz unitaria, obtendríamos entonces una distribución con varianza nula

$$\sqrt{T}\left(\hat{\rho}-1\right) \xrightarrow{p} 0,$$

esto es, a una distribución degenerada a un número que acumularía toda la densidad, y por tanto sería una distribución inútil para poder contrastar la hipótesis deseada. Es preciso multiplicar o escalar por T, y no por $\sqrt{T}$, para obtener una distribución **no degenerada**. De hecho la distribución a la que converge no es a una distribución estándar (conocida), sino que converge a un tipo de distribución no estándar que se denomina y tabulada por Dickey-Fuller (DF).

Para una hipótesis nula de paseo aleatorio sin deriva, el contraste sería del tipo t habitual, es decir

$$\frac{\hat{\rho}-1}{ee\left(\hat{\rho}\right)} \xrightarrow{d} DF_0,$$

digamos contraste DF en el Caso 1 o «sin término independiente ni tendencia».

Alternativamente, el modelo puede ser formulado como

$$\triangle Y_t = \left(\rho-1\right) Y_{t-1} + \varepsilon_t = \delta Y_{t-1} + \varepsilon_t$$

de modo que la hipótesis de raíz unitaria es

$$H_0 : \delta = 0 \text{ versus } H_1 : \delta < 0$$

es decir

$$H_0 : Y_t \text{ es } I(1)$$
$$H_1 : Y_t \text{ es } I(0)$$

Ahora el contraste estadístico sería

$$\tau = \frac{\hat{\delta}}{ee\left(\hat{\delta}\right)}.$$

En general, no es necesario usar el supuesto de que el error sea gausiano para llevar a cabo un contraste de DF_0. Únicamente lo hemos utilizado a efectos ilustrativos. De hecho las tablas relativas a DF_0 son aplicables con un error en forma de ruido blanco. Ya sea con ruido blanco o ruido blanco gausiano, la hipótesis nula es que el proceso estocástico es estacionario en diferencias, es decir, que tras realizar una diferencia el proceso se transforma en estacionario.

En términos un poco más generales en los que la serie a analizar presenta una media que no es cero, entonces deberíamos incorporar términos deterministas a la configuración de la regresión a estimar. Un problema importante, que también resolvieron Dickey y Fuller, es que al incluir la constante también cambia el valor del estadístico, por tanto el contraste se ve afectado.

Si incorporamos una constante a partir del modelo $AR\left(1\right)$ usual,

$$Y_t = \beta_0 + \rho Y_{t-1} + \varepsilon_t, \text{ con } \varepsilon_t \text{ ruido blanco}, \tag{11.18}$$

sabemos que el proceso es estacionario si $|\rho| < 1$, y cuando el parámetro tiene valor unitario, $\rho = 1$, el proceso se denomina paseo aleatorio con deriva (o sin ella) dependiendo del valor de β_0. Por tanto resulta natural plantear como hipótesis nula $\mathrm{H}_0 : \rho = 1$, frente a $\mathrm{H}_1 : \rho < 1$.

A partir de la expresión en diferencias que anteriormente hemos utilizado, podemos reescribir el modelo de la siguente manera

$$\begin{aligned} Y_t - Y_{t-1} &= \beta_0 + \rho Y_{t-1} - Y_{t-1} + \varepsilon_t; \\ \triangle Y_t &= \beta_0 + (\rho - 1) Y_{t-1} + \varepsilon_t; \\ \triangle Y_t &= \beta_0 + \delta Y_{t-1} + \varepsilon_t, \end{aligned} \tag{11.19}$$

cuya última ecuación es la expresión habitual del contraste DF.

Puesto que $\delta \equiv \rho - 1$, contrastar que $\delta = 0$, es lo mismo que $\rho = 1$, y si $\delta < 0$, entonces $\rho < 1$. Por consiguiente estamos considerando la hipótesis nula $\delta = 0$, frente $\mathrm{H}_1 : \delta < 0$. Obsérvese que la regresión planteada bajo la hipótesis nula implica que el regresando es $I(0)$ y que el regresor es $I(1)$ (a esto se le denomina regresión desequilibrada); sin embargo, bajo la hipótesis alternativa ambas variables son $I(0)$ y por tanto se vuelve a equilibrar.

El valor empírico del contraste DF se calcula de la forma habitual, $\frac{\hat{\delta}}{ee\left(\hat{\delta}\right)} = \frac{\hat{\rho}-1}{ee(\hat{\rho})}$, donde los errores estándar $ee\left(\widehat{\delta}\right)$ son los no robustos de MCO. Debemos tener en cuenta también que el contraste planteado es de una sola cola y rechazamos la hipótesis nula si el valor es más negativo o menor que el valor crítico de tablas de DF. El uso de tablas diferentes a las habituales es debido a que bajo la hipótesis nula estamos planteando una regresión desquilibrada. La presencia de un proceso con tendencia estocástica I(1), como hemos visto, hace que no sea aplicable el Teorema Central del Límite, por lo que asintóticamente no se converge a una distribución normal.

Si rechazamos la hipótesis nula, el constraste nos sugiere que la serie no tiene raíz unitaria, y entonces sería estacionaria. En particular, las tablas DF a utililzar son las indicadas como Caso 2 del test tipo t, y es necesario considerar que las mismas son válidas si $\beta_0 = 0$, es decir las tablas se elaboran considerando que el proceso verdadero es un paseo aleatorio sin deriva.

$$\frac{\hat{\delta}}{ee\left(\hat{\delta}\right)} = \frac{\hat{\rho} - 1}{ee\left(\hat{\rho}\right)} \xrightarrow{d} DF_1.$$

Sería posible y puede resultar interesante plantear la hipótesis nula compuesta,

$$\mathrm{H}_0 : \rho = 1, \beta_0 = 0.$$

En tal caso, el contraste estadístico sería del tipo F

$$F = \frac{(SCR_R - SCR_{NR})/r}{(SCR_{NR})/T - k}$$

donde los acrónomimos son los habituales. Nuevamente la distribución a utilizar es una distribución asintótica no estándar y también obtenida incialmente por Dickey y Fuller. Deberíamos utilizar la tabla indicada como Caso 1 del test tipo F dentro del Apéndice. Como resultado de dicho contraste, en el caso en que no se pueda rechazar, decimos que el proceso es estacionario en diferencias. En ocasiones a estos contrastes con distribuciones no estándar y en el caso de raíces unitarias, se les denomina pseudo-t estadístico y pseudo-F estadístico.

El contraste DF_1 se aplica en situaciones cuando el usuario tiene dudas sobre la estacionariedad, pero no hay una tendencia de largo plazo visible o alternativamente no hay razones teóricas para asumir dicha tendencia. Ejemplos típicos son los tipos de interés o la tasa de inflación.

Por el contrario, si el usuario viera razonable incorporar una tendencia, el modelo que acabamos de plantear podría albergar una hipótesis nula de paseo aleatorio con deriva, pero le permite incorporar una hipótesis alternativa de un proceso que se hiciera estacionario tras eliminar linealmente la tendencia. Para tal fin debemos considerar otro tercer caso, un caso en el que podamos incorporar una tendencia lineal

$$Y_t = \beta_0 + \alpha t + \rho Y_{t-1} + \varepsilon_t.$$

Si estimamos un modelo de este tipo, o de forma equivalente

$$\triangle Y_t = \beta_0 + \alpha t + \delta Y_{t-1} + \varepsilon_t, \delta = (\rho - 1),$$

la hipótesis nula sería

$$H_0 : \delta = 0 \text{ versus } H_1 : \delta < 0.$$

En tal caso el estadístico de DF también tendría una distribución asintótica distinta de las precedentes

$$\frac{\hat{\delta}}{ee\left(\hat{\delta}\right)} = \frac{\hat{\rho} - 1}{ee\left(\hat{\rho}\right)} \xrightarrow{d} DF_2$$

y por tanto las tablas a utilizar serían las del Caso 3.

Al igual que sucedía anteriormente, sería posible efectuar un contraste conjunto sobre $\alpha = 0, \rho = 1$, para ello usaríamos la correspondiente tabla de DF relativa a un contraste tipo F (caso 3).

Como el lector puede suponer que este tipo de contrastes deben usarse con criterio, y este puede resultar a veces confuso. Indicamos a continuación algunos usos estandarizados para los trabajos aplicados. Como norma general para contrastar la hipótesis nula de una raíz unitaria, lo adecuado es ajustar una especificación tal que represente una plausible descripción de los datos tanto bajo la hipótesis nula, como bajo la alternativa.

Si la serie original presenta tendencia, se deberían incluir como regresores el término independiente (constante) y el término de tendencia lineal temporal. Como hemos indicado, en ese caso la hipótesis nula contempla que la tendencia procede de un paseo aleatorio con deriva, mientras la alternativa es que el proceso tiene una tendencia temporal determinista junto con un proceso estacionario AR(1).

Si la serie no parece presentar tendencia y tiene un valor medio distinto de cero, deberíamos incluir un término constante en la regresión, si bien el modelo planteado bajo la hipótesis nula en este caso sería un paseo aleatorio sin deriva. Finalmente, si la serie parece fluctuar en torno al valor medio cero, no se considera necesario incluir ningún regresor adicional en la regresión, es decir, no incluimos ni constante ni término de tendencia.

El contraste DF es solo válido para un proceso $AR\,(1)$ y cuando la serie no responde a ese proceso, el contraste DF muestra autocorrelación de los errores. Para evitarlo y generalizar el contraste se utiliza el **contraste aumentado de Dickey-Fuller** (ADF) que consiste en añadir términos autorregresivos en el contraste DF hasta que desaparece la autocorrelación. El número de retardos utilizados se determina mediante el criterio de Akaike (eligiendo el número de retardos que minimiza su valor).

El caso básico con una posible estructura $AR(p)$ es

$$\triangle Y_t = \delta Y_{t-1} + \sum_{i=2}^{q} \gamma_i \triangle Y_{t-i} + \varepsilon_t$$

y el test es un contraste tipo t para

$$H_0 : \delta = 0 \text{ versus } H_1 : \delta < 0.$$

Los valores críticos para el parámetro δ serían los ya presentados con DF_0. Si quisiéramos llevar a cabo un contraste de significatividad sobre alguno de los parámetros γ_i utilizaríamos las distribuciones estándar asintóticas habituales. El motivo para esta diferencia es que cualquier hipótesis del tipo $\gamma_i = 0$ no introduce ninguna raíz unitaria.

Para llevar a cabo este tipo de contraste ADF es preciso incluir el número de retardos suficiente que asegure que los errores son IID. Es habitual empezar con un retardo amplio para el tipo de serie, e ir eliminando retardos irrelevantes.

En la práctica incluimos variables deterministas como la constante. Si ese fuera el caso, es decir, si estimamos

$$\triangle Y_t = \beta_0 + \delta Y_{t-1} + \sum_{i=2}^{q} \gamma_i \triangle Y_{t-i} + \varepsilon_t, \tag{11.20}$$

la hipótesis nula de tendencia estocástica o raíz unitaria sería $\mathrm{H}_0 : \delta = 0$, frente a la hipótesis alternativa de proceso estacionario, $\mathrm{H}_1 : \delta < 0$. Rechazamos la hipótesis nula si el valor empírico es menor o más negativo que el valor crítico y concluimos que la serie es estacionaria. Para ello utilizamos los mismos estadísticos y valores críticos,

en este caso DF_1. También aquí podemos contemplar hacer un contraste tipo F con la distribución ADF correspondiente para la hipótesis conjunta de $\beta_0 = \delta = 0$.

Cuando la serie presenta una tendencia clara (creciente o decreciente) la hipótesis alternativa de estacionaridad (ausencia de raíz unitaria) sin contemplar la posibilidad de estacionaridad en tendencia (determinista) no es adecuada y debe ser considerada. De manera que en este caso se añade al contraste ADF una tendencia determinista, analíticamente el contraste ADF con tendencia determinista es:

$$\triangle Y_t = \beta_0 + \alpha t + \delta Y_{t-1} + \sum_{i=2}^{q} \gamma_i \triangle Y_{t-i} + \varepsilon_t. \tag{11.21}$$

La hipótesis nula de tendencia estocástica o raíz unitaria es $\mathrm{H}_0 : \delta = 0$, frente a la hipótesis alternativa de proceso estacionario alrededor de una tendencia determinista, $\mathrm{H}_1 : \delta < 0$. Rechazamos la hipótesis nula si el valor empírico es menor o más negativo que el valor crítico y concluimos que la serie es estacionaria alrededor de una tendencia determinista. También aquí podemos contemplar hacer un contraste tipo F con la distribución ADF correspondiente para la hipótesis conjunta de $\alpha = \delta = 0$.

Los valores críticos para muestras grandes del contraste ADF los reproducimos a continuación:

Valores críticos del estadístico ADF para muestras grandes

	10 %	5 %	1 %
Con término constante	-2,57	-2,86	-3,43
Término constante y tendencia	-3,12	-3,41	-3,96

En el caso del logaritmo del PIB que hemos venido analizando, el contraste de ADF se implementaría considerando como hipótesis nula que tiene una raíz unitaria frente a la alternativa de que es estacionaria en torno a una tendencia temporal. La regresión ADF es

$$\begin{aligned}\widehat{\Delta logPIB_t} &= \underset{(0,093)}{0,167} + \underset{(0,000050)}{0,000075}t - \underset{(0,0079)}{0,014}\, logPIB_{t-1} + \underset{(0,078)}{0,0099}\Delta logPIB_{t-1} + \\ &\quad + \underset{(0,0706)}{0,4243}\Delta logPIB_{t-2} + \underset{(0,0788)}{0,2810}\Delta logPIB_{t-3}.\end{aligned}$$

El estadístico ADF tipo test, para contrastar la hipótesis de que el coeficiente del $logPIB_{t-1}$ es 0, es -1.772. Dado que el estadístico de contraste, -1.772 es menos negativo que -3.12, el contraste no permite rechazar la hipótesis nula. En función de estos resultados concluimos que el logaritmo del PIB para este periodo considerado tiene una raíz unitaria (es decir, una tendencia estocástica) frente a la posibilidad alternativa de que la variable fuera estacionaria alrededor de una tendencia determinista lineal.

En términos generales se hace necesario seleccionar un retardo, y buscando un proceso semiautomático para modelizar un proceso, un enfoque habitual es el conocido como **de lo general a lo específico**. La idea es comenzar con un retardo relativamente grande y reducir los retardos del modelo utilizando contrastes tipo t o tipo F. El proceso se repetiría hasta que el último retardo sea significativamente diferente de cero. Es

interesante observar que en el caso autorregresivo puro, este procedimiento producirá la verdadera longitud del retardo con una probabilidad asintótica unitaria, siempre que la elección inicial de la longitud del retarado incluya la longitud real.

Existen otros contrastes de raíces unitarias para cuando, al igual que en l caso ADF, el error está autocorrelacionado, algo bastante frecuente en las serie económicas. Un procedimiento propuesto por Phillips y Perron consiste en generalizar los tests DF

$$\triangle Y_t = f(t) + \delta Y_{t-1} + u_t$$

donde u_t está serialmente correlacionado y posiblemente se heterocedástico, y por otra parte $f(t)$ contiene la estructura determinista, similar a la contemplada en DF. No desarrollamos aquí estos tests, pero el lector podría ampliarlos fácilmente a partir de la excelente obra de Hamilton.

11.5 Cambio estructural

Al realizar pruebas de raíz unitaria, se debe tener especial cuidado si se sospecha que se ha producido un cambio estructural. Cuando hay rupturas estructurales los diferentes contrastes de Dickey-Fuller están sesgados hacia el "no rechazo" de una raíz unitaria. Consideremos que tenemos un proceso que antes y después de un T^* es estacionario. Por ejemplo,

$$Y_t = 0{,}5Y_{t-1} + \varepsilon_t + D, D = k > 0 \text{ si } t > T^* \text{ y } 0 \text{ en caso contrario}$$

Sin embargo estimamos

$$Y_t = \beta_0 + \rho Y_{t-1} + e_t$$

En tal caso, el coeficiente estimado $\hat{\rho}$ asociado a Y_{t-1} captura la propiedad de que los valores "bajos" de Y_t son seguidos por otros valores "bajos" y por otra parte valores "altos" son seguidos por otros valores "altos".

Por tanto si especificamos incorrectamente, mediante un paseo aleatorio con deriva, cuando en realidad estamos ante un proceso con un cambio estructural, el sesgo al estimar ρ implicará que el test Dickey-Fuller esté sesgado hacia la "aceptación" de la hipótesis nula de una raíz unitaria aun cuando la serie sea estacionaria pero con un cambio estructural en algún momento.

Unos contrastes para evitar este tipo de confusiones fueron propuestos por Perron en 1989 como una forma de contrastar la presencia de raíces unitarias en presencia de cambio estructural. Suponiendo que conocemos el momento temporal en el que se produce el cambio estructural, un primer test consiste en contrastar

$$H_0 : Y_t = \beta_0 + Y_{t-1} + \mu D_0 + \varepsilon_t, D_0 = 1 \text{ si } t = T^*, \ 0 \text{ en otros casos}$$

frente a

$$H_1 : Y_t = \beta_0 + \alpha t + \mu D_1 + \varepsilon_t, D_1 = 1 \text{ si } t > T^*, \ 0 \text{ en otros casos}$$

Para ello se estima

$$Y_t = \beta_0 + \rho Y_{t-1} + \alpha t + \mu D_1 + \sum_{i=1}^{k} \gamma_i \triangle Y_{t-i} + \varepsilon_t$$

y se contrasta la nula de $\rho = 1$, considerando que k es el número necesario de retardos para eliminar la correlación serial, y los valores críticos fueron obtenido por Perron y pueden ser consultados en Enders (2004).

Este esquema permite contrastar otras hipótesis nulas de interés. Por ejemplo, se puede contrastar la nula un cambio permanente en la deriva frente a un cambio en la pendiente de la tendencia:

$$H_0^* : Y_t = \beta_0 + Y_{t-1} + \mu D_1 + \varepsilon_t, D_1 = 1 \text{ si } t > T^*, \ 0 \text{ en otros casos}$$

frente a

$$H_1^* : Y_t = \beta_0 + \alpha t + \mu D_2 + \varepsilon_t, D_2 = t - T^* \text{ si } t > T^*, \ 0 \text{ en otros casos}$$

Incluso es factible contrastar un cambio en el nivel y en la deriva de una raíz unitaria combinando las hipótesis H_0 y H_0^*

$$H_0^{**} : Y_t = \beta_0 + Y_{t-1} + \mu_0 D_0 + \mu_1 D_1 + \varepsilon_t,$$

frente a

$$H_1^{**} : Y_t = \beta_0 + \alpha t + \mu_1 D_1 + \mu_2 D_2 + \varepsilon_t.$$

El problema que presenta la técnica de Perron es que requiere saber de antemano el momento temporal del cambio estructural. Trabajos posteriores a Perron han desarrollado una variedad de pruebas recursivas y secuenciales que endogeneizan el punto de ruptura, T^*. Los contrestes recursivos se calculan sobre submuestras $t = 1, ..., m$ para $m = m_0, ..., T$, siendo m_0 el valor inicial y T el tamaño de la muestra completa. Los tests secuenciales se calculan utilizando la muestra completa, aumentando secuencialmente la fecha de la ruptura hipotética (por lo tanto, utilizando diferentes variables ficticias).

11.6 Tendencias y descomposición por componentes inobservadas

En muchos contextos, resulta útil descomponer las series (económicas) en componentes o movimientos permanentes y no-permamentes (transitorio). Como hemos visto, esto puede ser complejo en el caso de que la serie tenga tendencia estocástica. En tal caso, sería difícil saber cuándo una serie está registrando valores por encima de la tendencia o por debajo.

Si el proceso, y por tanto la serie, es estacionaria en tendencia entonces es posible descomponer en una tendencia determinista y un componente estacionario (descomposición de Wold). Sin embargo, el método de Wold no es aplicable si el proceso temporal

es no estacionario. Beveridge y Nelson (BN) propusieron un método para descomponer una raíz unitaria, es decir, un proceso ARIMA(p,1,q) en componentes transitorio y permanente permitiendo que ambos sean estocásticos. Es decir, en una tendencia general más una parte irregular.

Consideremos un proceso $\{\triangle Y_t\}$ tal que

$$\triangle Y_t = \mu + a(B)\varepsilon_t,\ \varepsilon_t \sim IID(0, \sigma_\varepsilon^2), t = 1, ..., T \tag{11.22}$$

$$a(B) = a_0 + a_1 B + a_2 B^2 + ...,$$

donde μ es el coeficiente que indica la deriva y los coeficiente a_i son absolutamente sumables. En tal caso, el proceso no estacionario $\{Y_t\}$ se puede descomponer como

$$Y_t = Z_t + \xi_t \tag{11.23}$$

donde por un lado

$$Z_t = Z_{t-1} + \mu + u_t, u_t \sim IID(0, \sigma_u^2)$$

que es un paseo aleatorio con deriva, y por tanto los shocks a Z_t tienen efectos permanentes en Y_t, y por otro lado

$$\xi_t = c(B)v_t, v_t \sim IID(0, \sigma_v^2), c(B) = c_0 + c_1 B + c_2 B^2 + ...$$

es un componente irregular estacionario en el que los shocks a ξ_t generan un efecto en Y_t que termina desapareciendo a medida que t crece.

La cuestión es si podemos localizar la deriva, μ, junto con las sucesiones $\{u_t\}, \{v_t\}$ y $\{c_i\}$ de modo que sean compatibles con el proceso definido en (11.22).

A partir de la expresión (11.23) y de (11.22), se tiene

$$\Delta Y_t = \Delta Z_t + \Delta \xi_t = \mu + a(B)\varepsilon_t$$

y del lado izquierdo de la última ecuación, resulta que

$$\mu + u_t + (1 - B)c(B)v_t = \mu + a(B)\varepsilon_t.$$

Por tanto

$$u_t + (1 - B)c(B)v_t = a(B)\varepsilon_t. \tag{11.24}$$

Los dos lados de esta ecuación tiene funciones generadoras de autocovarianzas iguales puesto que tanto ΔY_t y ξ_t son estacionarias y $a(B)$ y $c(B)$ son sumables. Esto significa que ambos procesos son observacionalmente equivalentes. También significa que cualesquiera u_t, v_t que se acomoden a la correspondiente función generadora de autocovarianzas, serán procesos consistentes con (11.22). Lo que implica que habrá más de una solución posible, es decir, habrá más de una descomposición posible.

En el caso particular de la descomposición de BN, los autores asumen que u_t y v_t son perfectamente colineales ($u_t = \lambda v_t$). Sin embargo, no es necesario que estos shocks o innovaciones guarden esta relación de colinealidad. Es más sería factible que el shock u_t y el v_t no estuvieran correlacionados en absoluto. Este escenario precisamente nos

llevaría a un esquema de identificación distinto del de BN. De hecho sería el escenario de las componentes inobservadas en el que un proceso está compuesto de distintos (pero inobservables) componentes (piense el lector en los shocks u_t y v_t) y en son tales que $E(u_t v_t) = 0$. Pues bien, también es posible encontrar en este marco de componentes inobservadas una descomposición en términos permanentes e irregulares.

Naturalmente podríamos incluso pensar en casos en los que los shocks estén correlacionados, pero no perfectamente (como es el caso de la descomposición de BN). La cuestión es que si queremos identificar (en este caso descomponer) la serie, es imprescindible añadir alguna restricción en la relación entre los shocks de la parte estacionaria y los de la tendencia.

Desde un punto de vista más técnico, si volvemos a la expresión (11.24) y calculamos la función generatriz de autocovarianza[1] para los procesos representados a ambos lados del igual, se tiene

$$\sigma_u^2+(1-z)(1-z^{-1})c(z)c(z^{-1})\sigma_v^2+2\sigma_{u,v}(1-z)(1-z^{-1})c(z)c(z^{-1}) = \sigma^2 a(z)a(z^{-1}) \quad (11.25)$$

donde siendo

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} = IID\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix} \right]$$

de modo que bajo la restricción de BN: ($u_t = \lambda v_t$), se tiene

$$[\lambda + (1-z)c(z)][\lambda + (1-z^{-1})c(z^{-1})] = a(z)a(z^{-1}) \quad (11.26)$$

e igualando los términos constantes y los que tienen el mismo orden de $z$en ambos lados se obtiene

$$c_0 = a_0 - \lambda,$$

y

$$c_i = c_{i-1} + a_i, i = 1, 2, ...$$

o bien

$$c_i = c_0 + \sum_{j=1}^{i} a_j.$$

Por otra parte, la ecuación relativa a las funciones generatrices de autocovarianzas (11.26) se cumple para todo z, y en particular para $z = 1$, por tanto

$$\lambda^2 = a(1)^2$$

de manera que

$$\lambda = a(1) = \sum_{i=1}^{\infty} a_i.$$

[1] $G(z) = \sum_{j=-\infty}^{\infty} \gamma_j z^j$, es decir, se trata de una función que multiplica la autocovarianza j-ésima por un número (escalar complejo) z a la j-ésima potencia y hace la suma para todos los posible valores de j. En el caso particular de un MA infinito, se tiene que $G(z) = \sigma^2 a(z)a(z^{-1}), a(z) = \sum_{i=-\infty}^{\infty} a_i z^i$.

Estas ecuaciones nos permiten obtener

$$c_0 = -\sum_{j=1}^{\infty} a_j$$

así como

$$c_i = -\sum_{j=i+1}^{\infty} a_j.$$

Por tanto, bajo el supuesto de BN, podemos descomponer (11.23) de manera que

$$Z_t = Z_{t-1} + \mu + a(1)\varepsilon_t$$

y

$$\xi_t = c(B)v_t = (c_0 + c_1 B + c_2 B^2 + ...)\varepsilon_t$$

donde c_i está dado por $-\sum_{j=i+1}^{\infty} a_j$, como hemos visto.

Si alternativamente optamos por una descomposición de variable inobservables en el que el supuesto es que $\sigma_{uv} = 0$, entonces (11.25) se reduce a

$$\sigma_u^2 + (1-z)(1-z^{-1})c(z)c(z^{-1})\sigma_v^2 = \sigma^2 a(z)a(z^{-1})$$

que evaluada en $z = 1$ resulta en

$$\sigma_u^2 = \sigma^2 a(1)^2.$$

Sustituyendo esta última expresión en la ecuación anterior y a continuación dividiendo entre σ^2, se tiene

$$a(1)^2 + (1-z)(1-z^{-1})c(z)c(z^{-1})\frac{\sigma_v^2}{\sigma^2} = a(z)a(z^{-1})$$

de modo que si normalizamos a que $\frac{\sigma_v^2}{\sigma^2} = 1$, podemos proceder como en el caso de BN en (11.26), de modo que igualamos los términos de ambos lados y obtendríamos la descomposición deseada.

Por último, solo mencionar que merece la pena considerar otras descomposiciones existentes en la literatura que son incluso más atractivas ya que permiten estimar la correlación entre los componente u_t, v_t sin necesidad de hacer supuestos al respecto.

Tablas Dickey-Fuller

Caso 1. Ecuación de contraste sin término independiente ni tendencia

T	1 %	5 %	10 %
25	-2,66	-1,95	-1,60
50	-2,62	-1,95	-1,61
100	-2,60	-1,95	-1,61
250	-2,58	-1,95	-1,62
500	-2,58	-1,95	-1,62
∞	-2,58	-1,95	-1,62

Caso 2. Ecuación de contraste con término independiente y sin tendencia

T	1 %	5 %	10 %
25	-3,75	-3,00	-2,62
50	-3,58	-2,93	-2,60
100	-3,51	-2,89	-2,58
250	-3,46	-2,88	-2,57
500	-3,44	-2,87	-2,57
∞	-3,43	-2,86	-2,57

Caso 3. Ecuación de contraste con término independiente y con tendencia

T	1 %	5 %	10 %
25	-4,38	-3,60	-3,24
50	-4,15	-3,50	-3,24
100	-4,04	-3,45	-3,18
250	-3,99	-3,43	-3,15
500	-3,98	-3,42	-3,13
∞	-3,96	-3,41	-3,12

F test

Caso 1. $\triangle Y_t = \beta_0 + \delta Y_{t-1} + \varepsilon_t, H_0 : \beta_0 = 0 = \delta$

T	1 %	5 %	10 %
25	7,88	5,18	4,12
50	7,06	4,86	3,94
100	6,70	4,71	3,86
250	6,52	4,63	3,81
500	6,47	4,61	3,79
∞	6,43	4,59	3,78

Caso 2. $\triangle Y_t = \beta_0 + \alpha t + \delta Y_{t-1} + \varepsilon_t, \delta = (\rho - 1), H_0 : \beta_0 = 0 = \alpha = \delta = 0$

T	1 %	5 %	10 %
25	8,21	5,68	4,67
50	7,02	5,13	4,31
100	6,50	4,88	4,16
250	6,22	4,75	4,07
500	6,15	4,71	4,05
∞	6,09	4,68	4,03

Caso 3. $\triangle Y_t = \beta_0 + \alpha t + \delta Y_{t-1} + \varepsilon_t, \delta = (\rho - 1) . H_0 : \alpha = \delta = 0$

T	1 %	5 %	10 %
25	10,61	7,24	5,91
50	9,31	6,73	5,61
100	8,73	6,49	5,47
250	8,43	6,34	5,39
500	8,34	6,30	5,36
∞	8,27	6,25	5,34

Bibliografía complementaria

Matilla-García, M et al. 2017. Econometría y Predicción. McGraw Hill

Hamilton, XXX. Time Series Analysis.

Tema 12

Modelos de series temporales multiecuacionales

Este tema está elaborado como una adaptación de

Enders, W. Applied Econometric Time Series. 4 ed. Wiley. Capítulo 5 Stock y Watson, Introducción a la econometría. Capítulo 16. Así como de la bibliografía complementaria

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al Órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

- Análisis de Intervención y Funciones de Transferencia.
- Análisis VAR.
- Estimación e Identificación VAR.
- Función Impulso-Respuesta.
- VAR-Estructural

12.1 Modelos de series temporales multiecuacionales

En este tema pasamos del tratamiento de la dinámica temporal de una serie (tratamiento univariante) a considerar las relaciones dinámicas que se establecen entre varias series temporales (tratamiento multiecuacional). Esto nos permitirá ver cómo los efectos de las dececisones o acontecimiento económicos tienen consecuencias que se extienden a lo largo del tiempo, una cuestión práctica y relevante dentro del análisis econométrico aplicado. Nos interesa conocer cómo se distribuye el efecto de una causa a lo largo del tiempo.

Un cambio en nivel precios en una economía producirá sus correspondientes efecto después de que transcurra cierto tiempo; de modo que el efecto no se materializa de una sola vez, sino que se "distribuye" o se "transfiere" a lo largo del tiempo. De manera similar, cuando suben los impuestos sobre las rentas (sobre los ingresos), lo consumidores (los hogares) tienen menor renta disponible, lo que les lleva a reducir sus gastos en servicios y bienes, esto lleva a que se aminoren los beneficios de los oferentes, y estos reduzcan su demanda de insumos, y en consecuencia los beneficios de los productores de insumos, y así podríamos seguir hasta las últimas consecuencias. Lo relevante del que el efecto es que se prolonga a lo largo del tiempo, es decir, es como si el efecto de una causa fuera dinámico. Los efectos simultáneos, a difirencia de

los efectos distribuidos a lo largo del tiempo, son menos evidentes en las decisiones económicas, ya sean éstas las del consumidor o las del empresario; ya sean éstas de tipo microeconómico, o de tipo macroeconómico.

Lógicamente al modelizar econométricamente deberíamos tener en consideración los efectos dinámicos de la variables consideradas, por ejemplo: la toma de decisiones sobre consumo o inversión individuales que están sujetos en buena medida a hábitos del consumidor o a la verosimilitud en la percepción de cambios, o bien permanentes o bien transitorios, en variables económicas relevantes (pensemos en el nivel de ingresos o precios, que afecta a la toma de decisiones por parte de los agentes económicos). También hay aspectos de tipo contractual por los que compromisos contraídos no permiten cambiar instantáneamente ante acontecimientos exteriores: es difícil que una empresa cambie con rapidez las condiciones de producción si se encarece desorbitadamente una de las materias primas, o si la competencia explota una nueva tecnología. El mero coste de la información: hay decisiones económicas para muchos tipos de bienes o servicios en las que informarse consume un tiempo. Pensemos en algunos sectores productivos realmente dinámicos, como por ejemplo el mercado de tabletas electrónicas ("tablets"), que además se corresponde con un bien semi-duradero. En este tipo de sectores las decisiones de los agentes no son instantáneas, máxime si por ejemplo hay un escenario de presentación al mercado de un modelo más novedoso, lo que afecta a las expectativas sobre el precio de las existentes.

Dado que como hemos visto los efectos de los cambios en las variables no son siempre instantáneos, el objetivo prioritario es cómo modelizar la naturaleza dinámica de las relaciones económicas. Para ello veremos modelos que relacionen series estacionarias. Inicialmente presentamos que proceden de las ciencias físicas, conocidos como modelos de función de transferencia, y que en las ciencias económicas se conocen como modelos econométricos dinámicos. Describen cómo se transmiten los efectos desde una variable X_t a otra Y_t cuando no existen retroalimentación o causalidad bidireccional. Estos modelos resultan enormemente prácticos para evaluar la respuesta dinámica de políticas o medidas de intervención. También son modelos que funcionan probadamente mejorar la calidad de las predicciones respecto de la predicción univariante. La literatura respecto a las variables que actuán como "indicador adelantado de Y_t" son precisamente variables del tipo X_{t-k} siendo k un retardo temporamente cercano a t.

12.2 Análisis de Intervención y Funciones de Transferencia.

En términos generales, el análisis de intervención se refiere a la inclusión en un modelo de series temporales univariante de variables ficticias para representar sucesos que producen efectos deterministas. Se trata de una forma de ampliar la metodología univariante permitiendo que la evolución temporal de la variable dependiente pueda ser afectada por la evolución temporal de una variable exógena.

Suelen utilizarse variables ficticias para representar sucesos cualitativos que afectan a la serie dependiente, siendo entonces el **modelo de intervención** el siguiente:

$$Y_t = a_0 + A(L)Y_{t-1} + c_0 Z_t + B(L)\varepsilon_t, \tag{12.1}$$

Z_t : variable de intervención, L : operador retardos

La intervención puede darse de varias formas. Una de ellas es que la función Z_t recoga un **impulso**, es decir, una intervención puntual, de modo que solo en un momento $t = h$ la variable dependiente se vea afectada en c_0,

$$Z_t = \begin{cases} 1 & t = h \\ 0 & t \neq h \end{cases}$$

si bien los efectos pueden perdurar en el tiempo dada la naturaleza autorregresiva del proceso Y_t. En efecto, la variable dependiente tiene una estructura ARMA(p, q) con una intervención puntual en forma de impulso.

Otro modo de intervención es que la función Z_t recoga una intervención que tenga un efecto permanente sobre la serie (piense el lector en una subida de precios). En tal caso

$$Z_t = \begin{cases} 0 & t < h \\ 1 & t \geq h \end{cases}$$

donde los valores posteriores a h están afectados por una cantidad constante c_0, por este motivo se le denomina variable **escalón**.

También es posible que la intervención propiamente afecte a la variable dependiente con cierto retraso, digamos d. En tal caso habría que considerar retardar la variable de intervención adecuadamente, Z_{t-d}.

El modelo de intervención (12.1) puede extenderse de manera natural para contemplar no solo variables deterministas, sino cualquier otro tipo de variable exógena. La forma de implementar esta generalización es mediante un modelo del tipo

$$Y_t = a_0 + A(L)Y_{t-1} + C(L)Z_t + B(L)\varepsilon_t \tag{12.2}$$

donde al polinomio $C(L)$ se le denomina **función de transferencia** en la medida en que muestra cómo un movimiento en la variable exógena Z_t es transferido a la variable dependiente Y_t. Los sucesivos coeficientes del polinomio $C(L)$ se denominan «función impulso-respuesta». Este polinomio admite una representación admite una representación ARMA estacionaria e invertible, y por tanto podríamos expresarlo

$$D(L)Z_t = E(L)\varepsilon_{zt} \tag{12.3}$$

donde ε_{zt} es ruido blanco

Es fundamental que la variable Z_t sea exógena de manera que los shocks que afecten a la variable dependiente Y_t no estén relacionados con la trayectoria de $\{Z_t\}$, o puesto en otros términos: $E(\varepsilon_{zt}\varepsilon_t) = 0$. De este modo podrían trazarse los coeficientes impulso-respuesta de ε_t y ε_{zt} sobre Y_t a partir de las expresiones siguientes:

$$(1 - A(L)L)Y_t = a_0 + C(L)Z_t + B(L)\varepsilon_t$$

$$(1 - A(L)L)Y_t = a_0 + \frac{C(L)E(L)}{D(L)}\varepsilon_{zt} + B(L)\varepsilon_t$$

En efecto, a partir de ellas se obtienen los siguientes coeficientes impulso-respuesta:

$$\frac{\partial Y_t}{\partial \varepsilon_t} = \frac{B(L)}{(1 - A(L)L)}; \frac{\partial Y_t}{\partial \varepsilon_{zt}} = \frac{C(L)E(L)}{(1 - A(L)L)D(L)}.$$

Desde el punto de vista de la identificación del modelo de transferencia y su estimación, resulta útil estimar un modelo que se conoce como **modelo autorregresivo de retardos distribuidos** y que es de la siguiente forma

$$Y_t = \gamma_0 + \sum_{i=1}^{p} \gamma_i Y_{t-i} + \sum_{i=0}^{k} \beta_i Z_{t-i} + \varepsilon_t$$

desde el que se aprecia la relación con la expresión alternativa del modelo dada en (12.2). En su estimación e identificación pueden utilizarse los contrastes tipo t y tipo F para, desde una especificación de lo general a lo específico, eliminar coeficientes innecesarios. La condiciones de regularidad para aplicar este tipo de estimación es similar a los supuestos del modelo modelo de regresión con variables en forma de serie temporal[1]. Este tipo de modelos corren el riesgo de tener mucho parámetros a estimar (aunque en general no es el caso).

Alternativamente se podría estimar un modelo más parsimonioso si intentamos estimar directamente (12.2). Estimaríamos la expresión (12.3) y con los residuos del modelo, $\hat{\varepsilon}_{zt}$, y construiríamos la sucesión filtrada de Y_t

$$\left\{ \frac{D(L)}{E(L)} Y_t \right\} = \{Y_t^f\}$$

que serviría de base junto con los residuos filtrados $\hat{\varepsilon}_{zt}$ para por medio de un procedimiento basado en el correlograma cruzado entre $\hat{\varepsilon}_{zt}$ y Y_t^f e ir especificando un modelo parsimonioso. Existen varias propuestas en la literatura que indican pasos a seguir en al especificación del modelo de transferencia, sin haber ninguno claramente superior. El lector interesado puede encontrar algunas de ellas en la obra de Enders citada como referencia de este tema.

Más interesante desde el punto de vista econométrico es la flexibilidad que ofrecen estos modelos para tratar las relaciones dinámicas entre variables económicas. En efecto, la relación dinámica entre variables puede realizarse especificando que la variable dependiente es función tanto de valores contemporáneos como pasados de las variables explicativas

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + ... + \beta_{k+1} X_{t-k} + \varepsilon_t, \tag{12.4}$$

Si la variable dependiente fuera la inflación y la explicativa fuera los tipos de interés, la especificación del modelo está planteando que la actual tasa de inflación depende del tipo de interés actual y también de los tipos de interés mantenidos en periodos anteriores. Así pues un cambio en los tipos de interés ahora impactará sobre la inflación actual y sobre la futura. Lógicamente los efectos de un cambio de tipos se van diseminando

[1] El lector interesado puede revisar las condiciones de regularidad en Matilla-García et al. (2017).

progresivamente en la economía. A este tipo de modelos se les conoce como *modelo de retardos distribuidos*.

El coeficiente β_1 es el efecto contemporáneo de una variación unitaria en X_t sobre Y_t. El coeficiente de X_{t-1}, es el efecto sobre Y_t de una variación unitaria en X_{t-1} o, equivalentemente, el efecto sobre Y_{t+1} de una variación de X_t. En general, el coeficiente X_{t-k} es el efecto de una variación unitaria en X sobre Y en k periodos hacia adelante. Por tanto el efecto causal dinámico es el efecto de una variación unitaria en X_t sobre $Y_t, Y_{t+1}, .., Y_{t+k}$, y queda recogido en la sucesión de coeficientes $\beta_1, \beta_2, ..., \beta_{k+1}$.

La ecuación (12.4) nos permite calcular teóricamente los efectos causales en un caso de un cambio *transitorio* en la variable X y en el caso de un cambio *permanente*. Para verlo, consideremos que inicialmente $X_t = X$ es una constante y que en el momento t varía en una unidad, $X + 1$, pasando de nuevo al estado constante, X, en $t + 1$ y sucesivos periodos. Para entender básicamente la dinámica, consideremos momentáneamente que los errores poblacionales son nulos (obviamente esto es una simplificación que posteriormente eliminaremos), entonces tendríamos:

$$\begin{aligned}
Y_{t-1} &= \beta_0 + \beta_1 X + \beta_2 X + ... + \beta_{k+1} X \\
Y_t &= \beta_0 + \beta_1 (X + 1) + \beta_2 X + ... + \beta_{k+1} X \\
Y_{t+1} &= \beta_0 + \beta_1 X + \beta_2 (X + 1) + ... + \beta_{k+1} X \\
.. &.. \quad .. \\
Y_{t+k} &= \beta_0 + \beta_1 X + \beta_2 X + ... + \beta_{k+1} (X + 1) \\
Y_{t+k+1} &= \beta_0 + \beta_1 X + \beta_2 X + ... + \beta_{k+1} X,
\end{aligned}$$

de manera que el cambio en Y producido a causa de la variación unitaria en X en el momento t lo podríamos calcular fácilmente:

$$Y_t - Y_{t-1} = \beta_1.$$

Igualmente, el cambio en Y un periodo después del cambio sería

$$Y_{t+1} - Y_{t-1} = \beta_2,$$

y por tanto el cambio en Y tras k periodos después del cambio producido en t sería

$$Y_{t+k} - Y_{t-1} = \beta_{k+1}.$$

A partir del siguiente periodo, es decir, en el perido $k + 1$ los efectos habrían desaparecido, $Y_{t+k+1} - Y_{t-1} = 0$. A cada uno de estos efectos generados por una variación unitaria en X sobre Y tras k periodos se les denomina ***multiplicador dinámico*** del periodo k correspondiente. El gráfico de los retardos frente a los distintos coeficientes, multiplicadores dinámicos, nos daría una visión de cómo se distribuye el efecto causal esperado sobre Y ante una variación en el periodo t de X.

Alternativamente, si el cambio que se produce en el momento t es de carácter perma-

nente, entonces tendríamos

$$\begin{aligned} Y_{t-1} &= \beta_0 + \beta_1 X + \beta_2 X + ... + \beta_{k+1} X \\ Y_t &= \beta_0 + \beta_1 (X+1) + \beta_2 X + ... + \beta_{k+1} X \\ Y_{t+1} &= \beta_0 + \beta_1 (X+1) + \beta_2 (X+1) + ... + \beta_{k+1} X \\ .. &.. \quad .. \\ Y_{t+k} &= \beta_0 + \beta_1 (X+1) + \beta_2 (X+1) + ... + \beta_{k+1} (X+1). \end{aligned}$$

Ahora, después de un periodo desde el cambio, la variable Y ha variado

$$Y_{t+1} - Y_t = \beta_1 + \beta_2;$$

tras k periodos desde el cambio, la variable habrá cambiado

$$Y_{t+k} - Y_{t-1} = \beta_1 + \beta_2 + ... + \beta_{k+1} = \sum_{i=1}^{k+1} \beta_i,$$

que se denomina ***multiplicador dinámico acumulativo de largo plazo***. Este multiplicador puede utilizarse conjuntamente con el multiplicador dinámico, y conformar un multiplicador dinámico estandarizado:

$$\frac{\beta_1}{\sum \beta_i},$$

que nos indica la proporción de variación total imputable a primer periodo. Igualmente las sumas de sucesivos multiplicadores estandarizados nos informarían de la proporción del impacto de largo plazo imputable a un número consecutivo de periodos.

Estos modelos fácilmente son ampliables a lo que hemos denominado anteriormente *modelo autorregresivo de retardos distribuidos*. Las relaciones dinámicas entre las variables se establece mediante un modelo que considere a la variable dependiente retardada como una variable explicativa. Por ejemplo

$$Y_t = f(Y_{t-1}, X_t) + \varepsilon_t.$$

Siguiendo con el ejemplo de la tasa de inflación, este nuevo modelo plantea de la actual tasa de inflación depende (entre otras cosas) de cuál fue la tasa de inflación en el periodo anterior. De esta manera, y asumiendo que la relación entre las variables es positiva, los periodos de baja inflación vendrán seguidos de periodos de bajas tasas de inflación. Lógicamente podríamos mejorar fácilmente el modelo permitiendo covariar la variable dependiente también con valores pasados de la variable explicativa X, lo que nos conduciría al modelo

$$Y_t = f(Y_{t-1}, X_t, X_{t-1}, X_{t-2}, ...) + \varepsilon_t.$$

La existencia de autocorrelación en el error poblacional es algo intrínseco a los modelos con series temporales: los factores omitidos en el modelo, ya sea (12.4), están recogidos en el error, y estos factores es muy probable que estén autocorrelacionados consigo

mismos. Obviamente esto tendrá sus consecuencias a la hora de realizar inferencia sobre los parámetros del modelo, que por otra parte MCO puede estimar consistente e insesgadamente. Sin embargo, queremos llamar la atención sobre el hecho de que considerar autocorrelación en el modelo nos conduce a considerar de manera natural modelos del tipo ARD.

Los modelos ARD no son nuevos para el lector, de hecho hemos visto en el modelo de regresión con series temporales con varios predictores que, de cara a mejorar la predicción, un tipo de modelo competitivo es el que incorpora varios predictores aparte de los valores retardados de la variable de interés: ***modelo autorregresivo de retardos distribuidos*** (ARD).

En términos generales el modelo ARD(p,q) es

$$Y_t = \gamma_0 + \gamma_1 Y_{t-1} + \gamma_2 Y_{t-2} + ... + \gamma_p Y_{t-p} + \beta_1 X_t + \beta_2 X_{t-1} + ... + \beta_{k+1} X_{t-k} + \varepsilon_t \quad (12.5)$$

Consideremos ahora la autocorrelación dentro de un modelo RD. A partir de un modelo del tipo (12.4) con un error serialmente correlacionado usamos dicho modelo para obtener algunos estimadores:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + ... + \beta_{k+1} X_{t-k} + \varepsilon_t, \quad (12.6)$$

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + ... + \phi_p \varepsilon_{t-p} + \tilde{\varepsilon}_t,$$

con $\tilde{\varepsilon}_t$ no correlacionado con ε_t. Si por ejemplo $p = 1$ y desarrollamos la diferencia

$$\begin{aligned} Y_t - \phi_1 Y_{t-1} &= \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + ... + \beta_{k+1} X_{t-k} + \varepsilon_t \\ &\quad -\phi_1 \left(\beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_{k+1} X_{t-k-1} + \varepsilon_{t-1}\right) \\ &= \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + ... + \beta_{k+1} X_{t-k} \\ &\quad -\phi_1\beta_0 - \phi_1\beta_1 X_{t-1} - \phi_1\beta_2 X_{t-2} - ... - \phi_1\beta_{k+1} X_{t-k-1} + \tilde{\varepsilon}_t \quad (12.7) \\ &= \beta_0 - \phi_1\beta_0 + \beta_1 X_t + (\beta_2 X_{t-1} - \phi_1\beta_1 X_{t-1}) \\ &\quad +... + (\beta_{k+1} X_{t-k} - \phi_1\beta_k X_{t-k}) + \phi_1\beta_{k+1} X_{t-k-1} + \tilde{\varepsilon}_t. \end{aligned}$$

Por tanto

$$Y_t = \alpha_0 + \phi_1 Y_{t-1} + \varphi_0 X_t + \varphi_1 X_{t-1} + ... + \varphi_k X_{t-k} + \varphi_{k+1} X_{t-k-1} + \tilde{\varepsilon}_t, \quad (12.8)$$

donde $\alpha_0 = \beta_0 (1 - \phi_1)$, $\varphi_0 = \beta_1$, $\varphi_1 = \beta_2 - \phi_1\beta_1$, $\varphi_k = \beta_{k+1} - \phi_1\beta_k$ y $\varphi_{k+1} = \phi_1\beta_{k+1}$.

La ecuación (12.8) muestra que al considerar un simple proceso autorregresivo de orden 1 en el error, surge un modelo ARD(1, k+1). Por este motivo, a dicha ecuación se denomina *representación ARD del modelo de retardos distribuidos con errores autorregresivos*. La ecuación nos permite ver que al incluir como regresores el retardo de Y y un retardo adicional de X, entonces el término error está serialmente incorrelacionado, y por tanto se pueden utilizar los estimadores MCO habituales, sin necesidad de preocuparse de los efectos que genera la autocorrelación.

De forma equivalente, la ecuación (12.6) se puede reescribir de otro modo familiar y también útil. Consideremos la expresión (12.7) y reagrupemos los términos sacando a los

coeficientes β_j como factores comunes. Es decir, $\beta_j\left(X_{t-j}-\phi_j X_{t-j}\right)$ para los diferentes j considerados. Si simplemente definimos las variables en causi-diferencias $\tilde{Y}_t = Y_t - \phi_1 Y_{t-1}$ y $\tilde{X}_t = X_t - \phi_1 X_{t-1}$, $\tilde{X}_{t-1} = X_{t-1} - \phi_1 X_{t-2}$, etcétera, se obtiene

$$\tilde{Y}_t = \alpha_0 + \beta_1 \tilde{X}_t + \beta_2 \tilde{X}_{t-1} + ... + \beta_{k+1}\tilde{X}_{t-k} + \tilde{\varepsilon}_t, \tag{12.9}$$

donde lógicamente los errores son los mismos que los anteriores y, por tanto, no están tampoco correlacionados serialmente.

12.3 Análisis VAR

El tipo de modelos ARD de la sección anterior son muy útilies en la práctica econométrica, del mismo modo que lo han sido en el campo de la física y la ingeniería. Sin embargo, la condición de exogeneidad de las variables tipo Z_t o X_t del apartado anterior puede a llegar a ser muy restrictiva cuando se trata de variables de carácter económico en el que la interrelaciones son bastante frecuentes. En efecto, la secuencias de Y_t no debe afectar al devenir de las variables Z_t o X_t. Para que los coeficientes estimado del modelo ARD o de la función de transferencia $C(L)$ capturen el impacto de Z_t o X_t en Y_t de forma insesgada, debe suceder que Z_t o X_t esté incorrelacionada con los shocks que afectan a Y_t, es decir, ε_t para todo tipo de retardo. En el ejemplo de la inflación comentado anteriormente, es fácil que haya un proceso de retroalimentación de los tipos de interés a la inflación y de la inflación a los tipos de interés. Como respuesta a esta situación de retroalimentación o causalidad simultánea, Sims (1980) propuso una metodología alternativa a la denominada *macro-econometría tradicional*. A comienzos de la década de los setenta del siglo pasado, la metodología tradicional se basaba en la construcción de (grandes) modelos de ecuaciones simultáneas en los que las variables estaban divididas en dos grupos: endógenas o determinadas dentro del modelo, y exógenas. La estimación de estos modelos exigía que estuviesen identificados, lo que a su vez implicaba el cumplimiento de determinadas restricciones generalmente de exclusión (es decir, en cada una de las ecuaciones identificadas, debían excluirse una o varias variables). Estas restricciones no tenían mucho que ver con la teoría económica y eran contempladas con creciente escepticismo por una parte importante de la profesión. La división entre variables endógenas y exógenas también parecía arbitraria. Si a esto unimos el hecho de que los modelos multiecuacionales sufrieron un rotundo fracaso durante la crisis de los setenta, podemos entender el contexto en el que Sims planteó su alternativa metodológica.

Un VAR es un modelo multivariante que amplía el modelo univariante AR para estudiar conjuntamente las relaciones dinámica de dos o más series temporales. Para introducir el concepto, consideremos que solo tenemos dos variables X e Y. El VAR será entonces un modelo formado únicamente por dos ecuaciones. En la primera, X se hace depender de sus propios retardos y de los retardos de la otra variable, Y. Análogamente en la segunda ecuación la variable dependiente Y depende de los valores retardados de X e Y. Formalmente:

$$X_t = \alpha_{10} + \alpha_{11}X_{t-1} + ... + \alpha_{1p}X_{t-p} + \beta_{11}Y_{t-1} + ... + \beta_{1p}Y_{t-p} + u_{1t}$$

$$Y_t = \alpha_{20} + \alpha_{21}X_{t-1} + ... + \alpha_{2p}X_{t-p} + \beta_{21}Y_{t-1} + ... + \beta_{2p}Y_{t-p} + u_{2t}$$

Los supuestos del VAR son los mismos que formulamos para la regresión con series temporales, aplicados a cada una de las ecuaciones que lo conforman.

Si llamamos $\mathbf{w}_t$ al vector formado por X_t e Y_t, podemos escribir el VAR en notación matricial de la siguiente manera:

$$\mathbf{w}_t = \mathbf{A}_0 + \mathbf{A}_1\mathbf{w}_{t-1} + ... + \mathbf{A}_p\mathbf{w}_{t-p} + \mathbf{u}_t, \tag{12.10}$$

donde $\mathbf{u}_t$ es el vector de los errores tal que

$$\mathbb{E}(\mathbf{u}_t\mathbf{u}_s') = \sum \text{ si } s = t;\ 0 \text{ en el resto de los casos.}$$

$\mathbf{A}_0$ es el vector de los términos independientes y

$$\mathbf{A}_j = \begin{pmatrix} \alpha_{1j} & \beta_{1j} \\ \alpha_{2j} & \beta_{2j} \end{pmatrix}, j = 1, ..., p.$$

Siguiendo una regla análoga a la vista cuando estudiamos los modelos ARMA, el número de retardos incluidos en las ecuaciones del VAR (habitualmente los mismos) determina el orden del sistema. Así el orden de (12.10) será p, pues p es el retardo más largo. El VAR más sencillo que cabe imaginar es un VAR(1) con dos variables, cuya expresión sería:

$$\mathbf{w}_t = \mathbf{A}_0 + \mathbf{A}_1\mathbf{w}_{t-1} + \mathbf{u}_t =$$

$$= \begin{pmatrix} \alpha_{10} \\ \alpha_{20} \end{pmatrix} + \begin{pmatrix} \alpha_{11} & \beta_{11} \\ \alpha_{21} & \beta_{21} \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}. \tag{12.11}$$

Esta expresión también puede escribirse como

$$\left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} \alpha_{11} & \beta_{11} \\ \alpha_{21} & \beta_{21} \end{pmatrix} L \right] \begin{pmatrix} X_t - \alpha_{10} \\ Y_t - \alpha_{20} \end{pmatrix} = \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}$$

o bien

$$[\mathbf{I} - \mathbf{A}_1 L]\, \tilde{\mathbf{w}}_t = \mathbf{u}_t, \tilde{\mathbf{w}}_t = \begin{pmatrix} X_t - \alpha_{10} \\ Y_t - \alpha_{20} \end{pmatrix}$$

o también

$$\tilde{\mathbf{w}}_t = [\mathbf{I} - \mathbf{A}_1 L]^{-1}\, \mathbf{u}_t$$

Para que el proceso VAR tenga interés debe ser estacionario. Al igual que sucede con los procesos AR(1) deben cumplirse unas condiciones para la estacionaridad. En este caso, y sin entrar en los detalles técnicos, la condición es que las raíces de la ecuación

$$|\mathbf{I} - \mathbf{A}_1 L| = 0$$

deben estar fuera del círculo unidad. Obsérvese que en el caso de que $\mathbf{A}_1$ fuera diagonal, entonces tendríamos dos procesos univariantes AR(1) con shocks correlacionados, y los dos procesos serían estacionarios si $|\alpha_{11}| < 1, |\beta_{21}| < 1$.

En el caso VAR(p), los resultados se generalizan fácilmente, donde por simplicidad en la notación, y sin pérdida de generalidad, vamos a considerar que $\tilde{\mathbf{w}}_t = \mathbf{w}_t$

$$\mathbf{w}_t = \mathbf{A}_1\mathbf{w}_{t-1} + ... + \mathbf{A}_p\mathbf{w}_{t-p} + \mathbf{u}_t$$

que será estacionario si las raíces de la ecuación

$$\left|\mathbf{I} - \mathbf{A}_1 L - \mathbf{A}_2 L^2 - ... - \mathbf{A}_p L^p\right| = 0$$

están fuera del círculo unidad.

Implícitamente, considerar que las variables del vector $\mathbf{w}_t$ siguen un proceso VAR(p) implica que los p-retardos son suficientes para capaturar la relación dinámica entre las variables del vector $\mathbf{w}_t$. O puesto de una forma algo menos directa, pero importante: que los errores del modelo VAR no están relacionados con $\mathbf{w}_{t-p-1}, \mathbf{w}_{t-p-2}, ...$

Si junto con los términos autorregresivos, incorporamos en cada ecuación términos de medias móviles, estaremos ante un VARMA, es decir, un sistema en el que cada una de sus ecuaciones tiene términos autorregresivos y de medias móviles. Por ejemplo, un VARMA(p, q) expresado en forma matricial vendrá dado por:

$$\mathbf{w}_t = \mathbf{A}_1\mathbf{w}_{t-1} + ... + \mathbf{A}_p\mathbf{w}_{t-p} + \mathbf{u}_t - \mathbf{B}_1\mathbf{u}_{t-1} - ... - \mathbf{B}_q\mathbf{u}_{t-q},$$

es decir, tiene términos autorregresivos hasta el orden p y de medias móviles hasta el orden q. Si el proceso es estacionario lo podemos escribir

$$\mathbf{w}_t = \mathbf{A}_p(L)^{-1}\mathbf{B}_q(L)\mathbf{u}_t.$$

12.4 Estimación e Identificación VAR

En principio, una de las ventajas de los VAR con respecto a los modelos de ecuaciones simultáneas es que no requieren técnicas especiales de estimación. En efecto, como todas las variables explicativas son retardos del vector $\mathbf{w}$ y se asume que los errores son homocedásticos y no autocorrelados, la estimación de cada una de las ecuaciones del VAR puede llevarse a cabo por MCO. Los estimadores así obtenidos son consistentes y tienen una distribución asintótica normal, de manera que la inferencia estadística puede llevarse a cabo con los estadísticos t y F habituales.

Hemos visto que el número de retardos define el orden del VAR, pero ¿cómo se determina ese parámetro? En principio pueden incluirse gran cantidad de retardos con objeto de que los residuos tengan las propiedades deseables, pero conviene ser prudentes dado que la pérdida de grados de libertad derivada de la inclusión de retardos adicionales puede ser muy importante. Por ejemplo, un VAR(4) con tres variables tendrá 39 coeficientes (cada ecuación tendrá 3·4 =12 coeficientes más el término independiente),

pero si se incluye un retardo adicional, el número de coeficientes se eleva a 48, es decir, un retardo adicional implica la pérdida de 9 grados de libertad.

La selección de la longitud apropiada de los retardos suele hacerse basándose en criterios estadísticos. Puede emplearse un test de ratio de verosimilitud de la siguiente manera. Supongamos que deseamos contrastar la hipótesis nula de que el orden del VAR es p contra la alternativa de un VAR de orden q, con $q > p$. Estimamos tanto el VAR(p) como el VAR(q) y obtenemos para cada uno de ellos una estimación de la matriz de varianzas y covarianzas de los residuos, $\hat{\Sigma}_p$ y $\hat{\Sigma}_q$. Entonces el estadístico:

$$T\left[ln\left(det(\hat{\Sigma}_p)\right) - ln\left(det(\hat{\Sigma}_q)\right)\right] \tag{12.12}$$

se distribuye como una χ^2_r siendo r el número de restricciones impuestas bajo la hipótesis nula. Para tener en cuenta el sesgo en muestras pequeñas, Sims propuso utilizar en su lugar el estadístico:

$$(T-m)\left[ln\left(det(\hat{\Sigma}_p)\right) - ln\left(det(\hat{\Sigma}_q)\right)\right] \tag{12.13}$$

donde m es el número de parámetros a estimar bajo la hipótesis alternativa. Si el valor de (3) o (4) es superior al crítico en tablas se rechazará la hipótesis nula en favor de la alternativa, siendo q el orden del VAR. Podemos seguir con este procedimiento hasta que no podamos rechazar la hipótesis nula, en cuyo caso habríamos encontrado el orden adecuado.

El test de ratio de verosimilitud solo es aplicable cuando uno de los modelos es una versión restringida del otro. Además no tiene buenas propiedades en muestras pequeñas. Por ello suele recurrirse a diversos criterios de información. Por ejemplo, el criterio de información de Akaike (AIC) para un VAR(p) con k variables se calcula a partir de la expresión:

$$AIC(p) = ln\left[det\left(\Sigma_u\right)\right] + k(p+1)\frac{2}{T}, \tag{12.14}$$

donde Σ_u representa como antes la matriz de varianzas y covarianzas de los errores del VAR que se estima a partir de $\hat{u}_t$. Hay otros estadísticos. El criterio de información de Schwarz (SBC) tiene una expresión parecida a la anterior.

Calculado el AIC (o cualquier otro criterio) para distintos órdenes, elegiremos aquel que proporcione un valor menor.

12.5 Función Impulso-Respuesta

Las *funciones de respuesta al impulso* (FRI) son una de las herramientas que proporciona esta metodología. De la misma manera que un modelo autorregresivo univariante

admite una representación en forma de medias móviles, el VAR estable puede expresarse también como un sistema vectorial MA(∞). Por ejemplo, un VAR con dos variables podría expresarse como un VMA de la siguiente manera:

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \sum_{i=0}^{\infty} \begin{pmatrix} \varphi_{11}(i) & \varphi_{12}(i) \\ \varphi_{21}(i) & \varphi_{22}(i) \end{pmatrix} \begin{pmatrix} u_{1t-i} \\ u_{2t-i} \end{pmatrix} =$$

$$\mathbf{w}_t = \mathbf{u}_t + \boldsymbol{\psi}^1 \mathbf{u}_{t-1} + \boldsymbol{\psi}^2 \mathbf{u}_{t-2} + ... \tag{12.15}$$

En la expresión anterior $\varphi_{11}(1)$ y $\varphi_{12}(1)$ miden el impacto sobre x_t de un cambio unitario en cada uno de los elementos del vector **u** *un periodo después* de que dicho cambio haya tenido lugar. Análogamente $\varphi_{21}(1)$ y $\varphi_{22}(1)$ miden el impacto sobre y_t de dichos cambios. El resto de los coeficientes se interpretan de la misma forma.

Conocidos estos coeficientes, pueden calcularse los efectos a lo largo del tiempo ocasionados por un shock en alguna de las perturbaciones o innovaciones del sistema, es decir, las respuestas al impulso representado por ese shock

$$\frac{\partial \mathbf{w}_{t+s}}{\partial \mathbf{u}_t'}$$

que podemos calcular derivando a partir de

$$\mathbf{w}_{t+s} = \mathbf{u}_{t+s} + \boldsymbol{\psi}^1 \mathbf{u}_{t+s-1} + \boldsymbol{\psi}^2 \mathbf{u}_{t+s-2} + ... + \boldsymbol{\psi}^s \mathbf{u}_t + \boldsymbol{\psi}^{s+1} \mathbf{u}_{t-1} + ...$$

es decir,

$$\frac{\partial \mathbf{w}_{t+s}}{\partial \mathbf{u}_t'} = \boldsymbol{\psi}^s = [\psi_{ij}^s]_{n \times n}$$

que indica cómo reacciona la i-ésima variable del vector **w** cuando hay shock en la innovación j-ésima.

Volvamos al ejemplo del VAR(1) de dos variables. Supongamos que de la estimación del sistema obtenemos:

$$A_1 = \begin{pmatrix} 0{,}6 & 0{,}3 \\ -0{,}1 & 0{,}2 \end{pmatrix}.$$

En primer lugar habría que verificar que el VAR es estacionario[2], dado que no tiene sentido analizar las FRI en otro contexto. En este caso puede comprobarse que los valores propios de A_1 son menores que 1. Supongamos que la matriz de varianzas y covarianzas de las perturbaciones es:

$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 16 \end{pmatrix},$$

[2]El VAR será estacionario si los valores propios de la matriz A_1 son menores que la unidad.

y que los valores iniciales[3] son nulos $\mathbf{y}_0^T = \begin{pmatrix} 0 & 0 \end{pmatrix}$. Analicemos el efecto sobre la senda temporal de las variables del sistema de un shock de una desviación estándar[4] en el primer elemento del vector de las perturbaciones. Es decir, supondremos que en el periodo 1 $\mathbf{u}_1^T = (3, 0)$, volviendo a ser nulo dicho vector en los periodos siguientes. Entonces para el primer periodo se tiene:

$$\mathbf{y}_1 = \mathbf{A}_1\mathbf{y}_0 + \mathbf{u}_1 = \begin{pmatrix} 0{,}6 & 0{,}3 \\ -0{,}1 & 0{,}2 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \end{pmatrix}$$

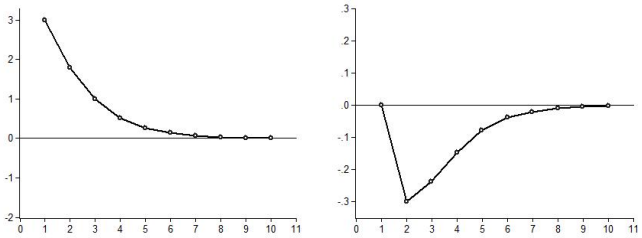
$$\mathbf{y}_2 = \mathbf{A}_1\mathbf{y}_1 + \mathbf{u}_2 = \begin{pmatrix} 0{,}6 & 0{,}3 \\ -0{,}1 & 0{,}2 \end{pmatrix}\begin{pmatrix} 3 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1{,}8 \\ -0{,}3 \end{pmatrix}$$

$$\mathbf{y}_3 = \mathbf{A}_1\mathbf{y}_2 + \mathbf{u}_3 = \begin{pmatrix} 0{,}6 & 0{,}3 \\ -0{,}1 & 0{,}2 \end{pmatrix}\begin{pmatrix} 1{,}8 \\ -0{,}3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0{,}99 \\ -0{,}24 \end{pmatrix}$$

$$\mathbf{y}_4 = \mathbf{A}_1\mathbf{y}_3 + \mathbf{u}_4 = \begin{pmatrix} 0{,}6 & 0{,}3 \\ -0{,}1 & 0{,}2 \end{pmatrix}\begin{pmatrix} 0{,}99 \\ -0{,}24 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0{,}522 \\ -0{,}147 \end{pmatrix}.$$

Operando sucesivamente, podemos obtener las respuestas de *x* e *y* al impulso inicial. En la Figura 12.1 se muestran las FRI. Puede comprobarse que al cabo de 10 periodos, el efecto del shock está prácticamente agotado.

Figura 12.1: Funciones de respuesta al impulso



12.6 VAR-Estructural(es)

En las secciones anteriores hemos presentado el VAR en lo que podemos denominar *forma estándar*. Esta forma está relaciona con un uso de los VAR con fines de prospectiva o predicción económica. Así definido el VAR no requiere mucho más que la elección del conjunto de variables a incluir en el sistema y la selección del orden más apropiado. El sistema estará casi con toda seguridad sobreparametrizado, pero en la medida en que imponer restricciones nulas inadecuadas puede implicar la pérdida de información

[3] Dado que lo que nos interesa es la evolución dinámica de las variables ante shocks, no es restrictivo en absoluto considerar unas condiciones iniciales dadas.

[4] Considerar shocks en términos de desviaciones típicas es muy habitual puesto que nos evita los problemas de las distintas unidades de medida.

importante, y teniendo en cuenta además que debido a la elevada multicolinealidad entre las variables explicativas, los tests tipo *t* no son una guía enteramente fiable, es preferible no reducir el modelo, al menos por el momento.

Además de la forma estándar, existe al menos otra presentación posible del VAR. En un VAR *estructural* (SVAR) se incluyen como variables explicativas, además de retardos de todas la variables, las propias variables contemporáneas. Esta configuración está orientada más hacia la explicación causal. Esta forma de expresar el VAR refleja que a priori se considera que las variables económicas modelizadas pueden estar simultáneamente relacionadas, sin necesidad de establecer una única relación causal, y en ese sentido todas las variables son tratadas de forma simétrica. La mayor diferencia conceptual entre usar VAR para pronosticar y usarlos para una modelización estructural del funcionamiento de algunas relaciones económicas es que la modelización estructural requiere supuestos muy específicos, derivados de la teoría económica y el conocimiento institucional, y de lo que es exógeno y lo que no lo es.

Esto lo podemos comprobar a partir del ejemplo con dos variables presentado en 12.11, escrito por comodidad con las variables expresadas en desviaciones con respecto a sus medias, se tendría

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} 0 & \gamma_{12} \\ \gamma_{21} & 0 \end{pmatrix} \begin{pmatrix} x_t \\ y_t \end{pmatrix} + \begin{pmatrix} \alpha_{11} & \beta_{11} \\ \alpha_{21} & \beta_{21} \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}. \tag{12.16}$$

La estructura de este sistema permite retroalimentaciones porque x_t e y_t pueden afectar una a la otra. Por ejemplo, γ_{21} es el efecto contemporáneo sobre y_t provocado por un cambio unitario de x_t, y α_{21} es el efecto sobre y_t tras un cambio unitario producido un periodo antes en la variable x_{t-1}. Asumiremos por el momento que tanto *x* como *y* son estacionarias, ε_1 *y* ε_2 son procesos de ruido blanco con varianzas σ_x^2 y σ_y^2, respectivamente, y que ambos están incorrelacionados entre sí. Tras una sencilla manipulación algebraica:

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} 1 & -\gamma_{12} \\ -\gamma_{21} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \alpha_{11} & \beta_{11} \\ \alpha_{21} & \beta_{21} \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} 1 & -\gamma_{12} \\ -\gamma_{21} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}, \tag{12.17}$$

es decir, podemos recuperar la forma denominada reducida (no estructural):

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \tag{12.18}$$

o, expresado en forma matricial compacta,

$$\mathbf{w}_t = \mathbf{\Gamma}_1 \mathbf{w}_{t-1} + \mathbf{u}_t.$$

Visto de esta manera, el sistema 12.18 es la *forma reducida del VAR estructural* 12.16, dado que tanto x como y están expresadas en función de las variables predeterminadas del sistema. Nótese además que bajo las condiciones de 12.16, $E(u_{1t}u_{2t})$ en 12.18 será en general distinta de cero.

A diferencia de lo que sucedía en el VAR estándar, las ecuaciones de un VAR estructural no pueden ser estimadas por MCO, dado que no todos los regresores son exógenos: en la primera ecuación del sistema 12.16 y_t está correlacionado con ε_{1t} y lo mismo sucede con x_t y ε_{2t} en la segunda. Este problema podría solventarse utilizando la representación equivalente expresada en 12.18, dado que en la forma reducida todos los regresores son efectivamente exógenos (solo hay variables predeterminadas a la derecha de ambas ecuaciones). Sin embargo, para que esta forma de proceder resultase operativa, debería de ser posible obtener todos los parámetros de 12.16 a partir de las estimación de 12.18. ¿Es esto posible? La respuesta es claramente negativa, puesto que el número de parámetros en 12.18 es inferior al del sistema 12.16, como puede comprobar fácilmente el lector. En definitiva, nos enfrentamos a un típico problema de identificación: a menos que se impongan restricciones sobre el VAR estructural, no es posible identificar los parámetros del mismo.

¿Cuántas restricciones son necesarias para alcanzar la identificación? Puede comprobarse (aunque no lo haremos aquí) que en un VAR con k variables, es necesario imponer al menos $\frac{k^2-k}{2}$ restricciones. En el ejemplo que venimos manejando, esto significa que una sola restricción sería suficiente para identificar todos los parámetros de la forma estructural. Por ejemplo, supongamos que por nuestro conocimiento teórico admitimos que y tiene efectos contemporáneos sobre x, pero que no hay efectos contemporáneos de x sobre y. En términos prácticos ello significa que imponemos la restricción $\gamma_{21} = 0$. En nuestro caso, esto es todo lo que necesitamos para obtener la identificación. Además, esta restricción implica que el VAR estructural 12.16 se convierte en un sistema *recursivo*:

$$\begin{aligned} x_t &= \gamma_{12}y_t + \alpha_{11}x_{t-1} + \beta_{11}y_{t-1} + \alpha_{12}x_{t-2} + \beta_{12}y_{t-2} + \varepsilon_{1t}, \\ y_t &= \alpha_{21}x_{t-1} + \beta_{21}y_{t-1} + \alpha_{22}x_{t-2} + \beta_{22}y_{t-2} + \varepsilon_{2t}. \end{aligned} \tag{12.19}$$

Como es sabido, este tipo de sistemas sí pueden ser estimados por MCO. En efecto, las variables explicativas de la segunda ecuación son retardos de x e y, por lo que no plantean problemas de endogeneidad. Y en la primera ecuación tampoco hay ahora relación entre y y ε_{1t}.

En vista de la relación entre el VAR estructural y su forma reducida, cabe revisitar las funciones de respuesta la impulso. El problema de la presentación que realizamos en el apartado dedicado a las FRI es que no es verosímil que uno de los elementos de **u** cambie mientras el otro permanece inalterado. Si el VAR del ejemplo tratado en dicha sección es la forma reducida de un VAR estructural, los elementos de **u** estarán correlacionados entre sí. En efecto, de (12.16) y (12.17) se deduce que:

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \frac{\varepsilon_1+\gamma_{12}\varepsilon_2}{1-\gamma_{12}\gamma_{21}} \\ \frac{\varepsilon_2+\gamma_{21}\varepsilon_1}{1-\gamma_{12}\gamma_{21}} \end{pmatrix},$$

Tabla 12.1: Funciones de respuesta al impulso (errores ortogonales)

	x	y
1	2,20	−0,033
2	1,31	−0,227
3	0,72	−0,176
4	0,38	−0,107
5	0,19	−0,059
6	0,099	−0,031

de manera que, bajo los supuestos del VAR estructural:

$$E(u_1 u_2) = \frac{\gamma_{21}\sigma^2_{\varepsilon 1} + \gamma_{12}\sigma^2_{\varepsilon 2}}{(1-\gamma_{12}\gamma_{21})^2},$$

que será distinto de cero, y por tanto las perturbaciones del sistema no serán independientes.

Para solucionar este problema, lo que suele hacerse es generar un nuevo conjunto de perturbaciones ortogonales, que tendrán varianza constante (unitaria) y no estarán correlacionadas entre sí. El procedimiento podría ser como sigue:

En primer lugar hacemos $\varepsilon_1 = c_{11}u_1$, de manera que si ha de tener varianza unitaria, $c_{11} = 1/s_1$, siendo s_1 la desviación estándar muestral de u_1 . A continuación se efectúa la regresión de u_2 sobre u_1 obteniendo las discrepancias $\varepsilon^*_2 = u_2 - c_{21}u_1$, que por construcción estará incorrelacionada con u_1 y también con ε_1. Si llamamos $s_{2,1}$ al error estándar de la regresión anterior y hacemos $\varepsilon_2 = \varepsilon^*_2/s_{2,1}$, la transformación adecuada queda definida por:

$$\mathbf{P} = \begin{pmatrix} 1/s_1 & 0 \\ -c_{21}/s_{2,1} & 1/s_{2,1} \end{pmatrix},$$

de manera que,

$$\mathbf{u} = \mathbf{P}^{-1}\boldsymbol{\varepsilon}.$$

Apliquemos todo esto al ejemplo anterior y veamos qué sucede con el vector $\mathbf{u}$ cuando se produce un shock de una desviación estándar en el primer elemento de $\boldsymbol{\varepsilon}$ sin que se modifique el segundo. Dado que

$$\mathbf{P}^{-1} = \begin{pmatrix} s_1 & 0 \\ c_{21}s_1 & s_{2,1} \end{pmatrix},$$

tendremos que (se deja el cálculo al lector)

$$\mathbf{u} = \mathbf{P}^{-1}\boldsymbol{\varepsilon} = \begin{pmatrix} 3 & 0 \\ 0{,}444{\cdot}3 & 14{,}222 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ 1{,}3332 \end{pmatrix}.$$

Como puede comprobarse, ahora u_2 es distinto de 0. Los seis primeros valores de las nuevas FRI se recogen ahora en la Tabla 12.1.

Como hemos dicho esto es solo una forma de resolver el problema, existen otras maneras de imponer restricciones para identificar correctamente el model SVAR y por tanto de variar las FRI. Lo interesante es que las restricciones impuestas tengan un marcado carácter económico.

En términos más generales, podemos decir que hay un problema de identificación. Consideremos un modelo VAR(p)

$$\mathbf{w}_t = \mathbf{A}_1\mathbf{w}_{t-1} + ... + \mathbf{A}_p\mathbf{w}_{t-p} + \mathbf{u}_t$$

o bien

$$\mathbf{A}(L)\mathbf{w}_t = \mathbf{u}_t, \mathbf{A}(L) = I - \mathbf{A}_1 L - \mathbf{A}_2 L^2 - ... - \mathbf{A}_p L^p$$

siendo A_i los coeficientes poblacionales de la regresión de $\mathbf{w}_t$ sobre $\mathbf{w}_{t-1}, \mathbf{w}_{t-2}, ..., \mathbf{w}_{t-p}$. Los términos dentro del vector $\mathbf{u}_t$ en general no serán los shocks estructurales (es decir, no serán ε_{jt}), si lo fueran sería inmediato calcular las FRI simplemente utilizando la representación MA del VAR, $\mathbf{w}_t = \mathbf{A}(L)^{-1}\mathbf{u}_t$. Como hemos visto, en general $\mathbf{u}_t$ estará afectado por múltiples shocks: dado un trimestre determinado, el PIB variará de forma no esperada (prevista) por varias razones. Por ejemplo, si $\mathbf{w}$ contiene dos variables, como hemos considerado en los ejemplos, entonces

$$u_{1t} = R_{12}u_{2t} + \varepsilon_{1t}$$

$$u_{2t} = R_{21}u_{1t} + \varepsilon_{2t}$$

de modo que para identificar los elementos R_{ij} necesitamos hacer, por ejemplo, una restricción, digamos, $R_{12} = 0$, que nos llevaría a un tipo de descomposición que permitiría identificar a $\boldsymbol{R}$, donde $\boldsymbol{R}$ es una matriz cuadrada que permite generar los shocks estructurales a partir del vector $\mathbf{u}_t$

$$\boldsymbol{\varepsilon}_t = \boldsymbol{R}\mathbf{u}_t.$$

Así pues, si $\mathbf{A}(L), \boldsymbol{\Sigma}_u$ y $\boldsymbol{R}$ son temporalmente invariantes respecto a impulsos o cambios dentro del periodo considerado, entonces

$$\mathbf{B}(L)\mathbf{w}_t = \boldsymbol{R}\mathbf{u}_t = \boldsymbol{\varepsilon}_t, \mathbf{B}(L) \equiv \boldsymbol{R}\mathbf{A}(L)$$

que es el VAR estructural, cuya representación MA nos permitirá obtener las FRI estructurales:

$$\mathbf{w}_t = \mathbf{B}(L)^{-1}\boldsymbol{\varepsilon}_t = \mathbf{A}(L)^{-1}\mathbf{R}^{-1}\boldsymbol{\varepsilon}_t$$

de modo que

$$\frac{\partial \mathbf{w}_{t+s}}{\partial \boldsymbol{\varepsilon}_t'} = \boldsymbol{\Phi}^s = [\phi_{ij}^s]_{n\times n}, \boldsymbol{\Phi} \equiv \mathbf{A}(L)^{-1}\mathbf{R}^{-1}$$

Actualmente hay mucha literatura sobre cómo identificar los parámetros estructurales haciendo para ello restricciones de corto plazo (como la que hemos hecho anteriormente al asumir que $R_{12} = 0$) o largo plazo (por ejemplo respecto a la suma de los coeficiente AR) e incluso mediante variables instrumentales.

Por último, destacamo una herramienta también muy utilizada, quizás más que las FRI, es el análsis causal (en términos de predecibilidad establecido originalmente por Granger).

El ***análisis de causalidad de Granger*** estudia si los retardos de una determinada variable son de utilidad para elaborar pronósticos sobre otra. Si es así, decimos que la primera causa en el sentido de Granger a la segunda. Consideremos de nuevo el VAR(2) con dos variables presentado en (1), que volvemos a escribir expresado en desviaciones con respecto a la media:

$$x_t = \alpha_{11}x_{t-1} + \alpha_{12}x_{t-2} + \beta_{11}y_{t-1} + \beta_{12}y_{t-2} + u_{1t},$$
$$y_t = \alpha_{21}x_{t-1} + \alpha_{22}x_{t-2} + \beta_{21}y_{t-1} + \beta_{22}y_{t-2} + u_{2t}.$$

En base a la estimación de este sistema:

1. Si los coeficientes β_{1i} de la primera ecuación son estadísticamente significativos, mientras que los α_{2i} en la segunda no lo son, diremos que hay causalidad, en el sentido de Granger, de *y* a *x*.
2. Si los coeficientes α_{2i} de la segunda ecuación son estadísticamente significativos, mientras que los β_{1i} en la primera no lo son, diremos que hay causalidad, en el sentido de Granger, de *x* a *y*.
3. Si ambos conjuntos de coeficientes, α_{2i} y β_{1i}, son estadísticamente significativos, diremos que hay causalidad bidireccional en el sentido de Granger.
4. Si ninguno de los dos conjuntos de coficientes son estadísticamente significativos, no hay ninguna relación de causalidad.

Podemos generalizar este análisis sin problemas para contemplar más de dos variables. Para ello puede ser de utilidad emplear la siguiente notación para un VAR con *k* variables:

$$\begin{pmatrix} x_{1t} \\ x_{2t} \\ \vdots \\ x_{kt} \end{pmatrix} = \begin{pmatrix} A_{11}(L) & A_{12}(L) & \cdots & A_{1k}(L) \\ A_{21}(L) & A_{22}(L) & \cdots & A_{2k}(L) \\ \vdots & \vdots & \ddots & \vdots \\ A_{k1}(L) & A_{k2}(L) & \cdots & A_{kk}(L) \end{pmatrix} \begin{pmatrix} x_{1t-1} \\ x_{2t-1} \\ \vdots \\ x_{kt-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \\ \vdots \\ u_{kt} \end{pmatrix}, \qquad (12.20)$$

donde $A_{ij}(L)$ representa los coeficientes de los retardos de la variable *j* sobre la ecuación de la variable *i*. Entonces diremos que la variable *j* no causa, en el sentido de Granger, a la variable *i*, si no se puede rechazar la hipótesis nula de que todos los coeficientes de $A_{ij}(L)$, son estadísticamente iguales a cero.

Bibliografía complementaria

Matilla-García, M et al. 2017. Econometría y Predicción. McGraw Hill

Tema 13

Modelos de Cointegración y de Corrección del Error

Este tema está elaborado como una adaptación de

Enders, W. Applied Econometric Time Series. 4 ed. Wiley. Capítulo 6,

así como de la bibliografía complementaria.

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al Órgano convocante ni al Tribunal actuante.

- Modelos de Cointegración y de Corrección del Error.
- Cointegración y tendencias comunes.
- Cointegración y corrección del error.
- Tests de cointegración: Engle-Granger y Johansen.

13.1 Cointegración y tendencias comunes

En el tema anterior introdujimos la posibilidad de estudiar simultáneamente las interacciones dinámicas de dos o más series a través de un modelo VAR. Allí consideramos que las variables del sistema VAR eran I(0), pero ¿qué sucede si no lo son? Ya estudiamos para los procesos estocásticos univariantes que podemos eliminar la tendencia tomando la serie en diferencias, y también en un escenario VAR podríamos estudiar las relaciones dinámicas formando un VAR de variables en diferencias, sin embargo no siempre es lo adecuado (ni desde el punto de vista económico y ni econométrico) pues podríamos dejar de analizar aspectos relevantes de las relaciones dinámicas entre las variables consideradas. Este tema introduce precisamente el escenario en el que deberíamos trabajar con las series no estacionarias.

La regresión entre variables no estacionarias puede dar lugar al problema de las regresiones espurias. La mayoría de las series macroeconómicas son no estacionarias, lo que plantea un problema empírico importante en ciertas circunstancias. Por ejemplo, la teoría económica postula que hay una relación de equilibrio a largo plazo entre el consumo agregado y la renta disponible. Considérese la estimación de dicha función utilizando un sencillo modelo de regresión simple y datos trimestrales entre 1995 y 2012 (segundo trimestre), correspondientes a la economía norteamericana:

$$\hat{C}_t = \underset{(125,6)}{3253,6} + \underset{(0,012)}{0,49\, Y_t}, \quad R^2 = 0,96, \quad DW = 0,05.$$

A simple vista los resultados son bastante aceptables: tanto el signo como el valor de la propensión marginal al consumo son acordes con la teoría, el ajuste es muy elevado y la renta es altamentemente significativa, con un valor *p* prácticamente nulo. Pero un contraste ADF muestra que tanto consumo como renta disponible son I(1), es decir, series no estacionarias, por lo que podemos enfrentarnos al mencionado problema de regresión espuria. Añadamos que el valor del estadístico de Durbin y Watson es claramente incompatible con la hipótesis de no autocorrelación y significativamente menor que el valor del coeficiente de determinación, y tendremos todos los síntomas clásicos de la regresión espúrea o regresión sin sentido.

En estas condiciones se ha sugerido evitar el problema efectuando la regresión entre las series diferenciadas. La diferenciación de las series eliminaría la no estacionaridad, pero al coste de impedir estimar la supuesta relación de equilibrio a largo plazo, puesto que la regresión $\triangle C_t = \delta_0 + \delta_1 \triangle Y_t + \varepsilon_t$ sería una estimación de la relación entre variables a corto plazo.

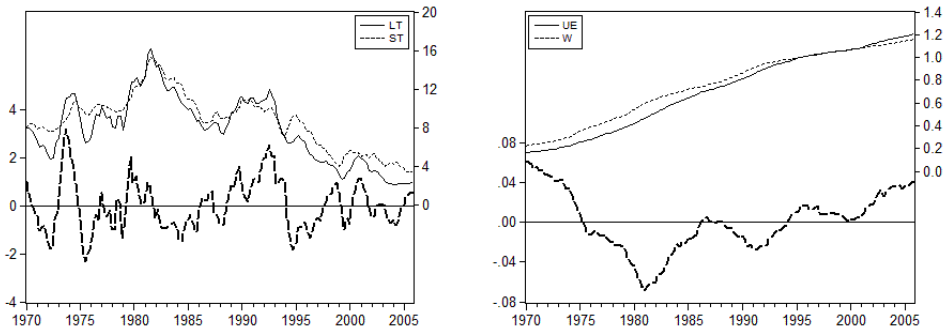
La pregunta es entonces si es posible estimar relaciones de equilibrio o de largo plazo entre variables no estacionarias; cuestión del máximo interés si tenemos en cuenta que, como hemos dicho, la mayor parte de las series económicas son en principio no estacionarias. La cointegración proporciona una respuesta a este interrogante. Supongamos que efectivamente existe una relación a largo plazo entre las variables Y (digamos gastos en consumo) y X (renta disponible) que podemos representar por:

$$Y_t = \alpha + \beta X_t + \varepsilon_t. \tag{13.1}$$

En la relación anterior el término de error $\varepsilon_t = Y_t - (\alpha + \beta X_t)$ puede interpretarse como la desviación del consumo de su relación de equilibrio a largo plazo dada por $\alpha + \beta X_t$, o *error de equilibrio*. Si dicha relación teórcia existe realmente, cualquier desviación del consumo respecto de $\alpha + \beta X_t$ ha de ser necesariamente transitoria. Es claro que en otro caso, es decir, si las desviaciones no se corrigiesen en un plazo relativamente breve, no podríamos sostener la existencia de dicha relación. En términos estadísticos, ello significa que ε_t no debería apartase mucho de la recta de ordenada nula, de hecho esperaríamos que la cruzara frecuentemente. Es decir, la existencia de la relación a largo plazo exige que el término error en 13.1 sea estacionario, a pesar de que las series de consumo y renta sean ambas integradas de orden uno[1]. Esto no sucede, por ejemplo, si dicho término tiene una raíz unitaria (una tendencia estocástica), dado que en ese caso los errores, lejos de eliminarse, se van acumulando en el tiempo. Es decir, que el cumplimiento de la teoría representada por la Ecuación 13.1, exige que aun siendo Y_t y X_t I(1), ha de existir una combinación lineal de las mismas ε_t, que sea estacionaria, o I(0). Esta es la idea fundamental del concepto de cointegración: dos series se dice que están *cointegradas*, si siendo ambas I(1), existe una combinación lineal entre las mismas que es

[1]Obsérvese que decimos condición necesaria, pero no suficiente. La existencia de una relación con sentido económico entre un conjunto de variables solo puede provenir de la teoría económica.

Figura 13.1: Series cointegradas (izquierda) y no cointegradas (derecha)



estacionaria. Nótese que la cointegración exige en este caso que ambas series sean I(1). Si una fuese I(1) y la otra I(0) no podría existir una combinación lineal estacionaria entre las mismas.

En estas condiciones, o sea si hay cointegración, se puede demostrar que la estimación MCO de 13.1 proporciona estimadores adecuados evitando por tanto el problema antes mencionado de las regresiones espurias. En concreto el estimador MCO de 13.1 no solo es consistente, sino *superconsistente*, es decir, converge con más rapidez de la habitual al verdadero parámetro poblacional.

La Figura 13.1 ilustra gráficamente lo anterior. En los dos paneles representamos dos series X e Y, integradas de primer orden (escala derecha), junto con los residuos de la regresión entre las mismas en trazo discontinuo (escala izquierda). En la Figura 13.1 de la parte izquierda se han representado los tipos de interés a corto y largo plazo en la Unión Europea, tal como aparecen en la base de datos AWM.

Los errores de desequilibrio representados por los residuos en la parte inferior del gráfico son claramente estacionarios, de manera que las desviaciones de la relación de equilibrio a largo plazo se corrigen con relativa rapidez. Podemos decir que X e Y están cointegradas. En el panel de la derecha, donde se han representado dos índices de precios (el deflactor de la UE y un índice mundial de precios, ambos obtenidos de la misma fuente) sucede lo contrario: los residuos son no estacionarios registrándose grandes desviaciones del equilibrio que además se mantienen de forma prolongada en el tiempo. En este caso las series no están cointegradas.

La figura precisamente nos invita a pensar que las series cointegradas comparten una tendencia estocástica común, mientras que eso no sucede en las no cointegradas. Pensemos en los determinantes del PIB per cápita, donde cada uno de los numerosos determinantes del mismo influyen sobre el valor observado. Si uno de esos determinantes, como es el caso del consumo per cápita, tiene una tendencia estocástica, también la tendrá el PIB per cápita. De hecho ambas variables (hemos visto ejemplos en otros temas) tienen cada una de ellas tendencia estocástica, por lo que incluso podemos decir que al estar teóricamente relacionado el consumo y el PIB per cápita, es factible que compartan una tendencia estocástica común. Al tratarse del PIB per cápita este podría contener más tendencias estocásticas y compartir o no con ellas una tendencia. Por

ejemplo la inversión per cápita podría compartir tendencia con el PIB per cápita.

La definición formal de cointegración desarrollada por Engle y Granger (1987) es la siguiente:

Se dice que dos series temporales Y_t y X_t están cointegradas de orden *d*, *b*, CI(*d*,*b*), donde $d \geq b$ y ambos son números enteros positivos, si:

1. Ambas son integradas de orden *d*.
2. Existe una combinación lineal de dichas variables $\beta_1 Y_t + \beta_2 X_t$ que es integrada de orden *d*-*b*.

El vector (β_1, β_2) recibe el nombre de vector de cointegración y además en este caso (solamente dos variables implicadas), dicho vector, una vez normalizado, es único.

Por defecto, nos referiremos en lo sucesivo, salvo que se diga lo contrario, a que *d*=*b*=1, es decir, CI(1,1) o series I(1) para las que existe una combinación lineal que las hace I(0). Por ejemplo, sean las series:

$$y_{1t} = w_{1t} + \varepsilon_{1t}$$
$$y_{2t} = w_{2t} + \varepsilon_{2t},$$

donde w_{1t} y w_{2t} son dos procesos de camino aleatorio representativos de la tendencia estocástica en cada una de las dos series y ε_{1t}, ε_{2t} los respectivos términos error. Si los procesos y_{1t} e y_{2t} están cointegrados, debe existir un vector de parámetros no nulos (β_1, β_2) tal que $\beta_1 y_{1t} + \beta_2 y_{2t}$ sea estacionario:

$$\beta_1 y_{1t} + \beta_2 y_{2t} = (\beta_1 w_{1t} + \beta_2 w_{2t}) + (\beta_1 \varepsilon_{1t} + \beta_2 \varepsilon_{2t}).$$

El último paréntesis es estacionario, al ser una combinación lineal de series estacionarias, y por tanto resulta que $(\beta_1 w_{1t} + \beta_2 w_{2t})$ debe ser también estacionario también. Sin embargo, este término es una combinación lineal de variables I(1) y la única forma de que sea I(0) es que se anule. Puesto que por hipótesis los parámetros del vector $\boldsymbol{\beta}$ son distintos de cero, se tiene que

$$\beta_1 w_{1t} + \beta_2 w_{2t} = 0 \implies w_{1t} = -\frac{\beta_2}{\beta_1} w_{2t}.$$

Es decir que (excepto por la constante $-\beta_2/\beta_1$) la tendencia estocástica de ambos procesos es la misma o decimos que es común. Observamos por tanto que cualquier otra combinación lineal de las dos variables contendría una tendencia, así pues el vector de cointegración es único salvo por la normalización escalar. Es decir, para otros $\beta_j, j = a, b$ distintos de $j = 1, 2$, se tiene que la combinación $\beta_a w_{1t} + \beta_b w_{2t}$ no será estacionaria.

Es inmediato ampliar la definición de cointegración, y por tanto también estas consideraciones de tendencias comunes, para albergar más variables en el modelo. En efecto, la definición ampliada sería la siguiente:

Decimos que los componentes de un vector $\mathbf{x}_t = (x_{1t}, ..., x_{kt})'$ están (o son) cointegrados de orden d, b, si:

1. Todos los elementos de $\mathbf{x}_t = (x_{1t}, ..., x_{kt})'$ son integrados de orden d
2. Si existe un vector de cointegración $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)$ tal que la combinación lineal

$$\boldsymbol{\beta}\mathbf{x}_t$$

es integrada de orden $(d - b), b > 0$.

Obsérvese que si $\boldsymbol{\beta}$ es un vector de cointegración, también lo será $\lambda\boldsymbol{\beta}, \lambda \neq 0$, por este motivo normalizamos haciendo que el primer coeficiente sea unitario, para lo cual hacemos que, por ejemplo, $\lambda = 1/\beta_1$.

Al considerar más de dos variables, es posible que se den situaciones donde dos variables sean cointegradas y la tercera no cointegre con las dos primeras pero sí lo haga con una combinación de las mismas. Por ejemplo, si x_1, x_2 son $CI(2, 1)$ y x_3 es $I(1)$, podría suceder que la combinación

$$\beta_1 x_{1t} + \beta_2 x_{2t}$$

esté cointegrada con una tercera variable, digamos, x_{3t} que debería de ser integrada de orden 1, I(1). Esta situación se denomina multicointegración.

En general, la cointegración ocurrirá siempre que la tendencia en una variable pueda ser expresada como un combinación lineal de tendencias en las demás variables. En efecto, sea

$$\mathbf{x}_t = \mathbf{w}_t + \boldsymbol{\varepsilon}_t$$

donde $\mathbf{x}_t = (x_{1t}, ..., x_{kt})'$, $\mathbf{w}_t = (w_{1t}, ..., w_{kt})'$ un vector de tendencias estocásticas y $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, ..., \varepsilon_{kt})'$ un vector de componentes estacionarias, entonces si una tendencia puede expresarse como combinación lineal de otras tendencias en el sistema, esto significaría que existe un vector $\boldsymbol{\beta}$ tal que

$$\beta_1 w_{1t} + \beta_2 w_{2t} + ... + \beta_k w_{kt} = 0.$$

Si premultiplicamos por dicho vector la expresión $\mathbf{x}_t = \mathbf{w}_t + \boldsymbol{\varepsilon}_t$, entonces

$$\boldsymbol{\beta}\mathbf{x}_t = \boldsymbol{\beta}\mathbf{w}_t + \boldsymbol{\beta}\boldsymbol{\varepsilon}_t$$

y puesto que $\boldsymbol{\beta}\mathbf{w}_t = 0$, sucede que $\boldsymbol{\beta}\mathbf{x}_t = \boldsymbol{\beta}\boldsymbol{\varepsilon}_t$ y por tanto la combinación lineal $\boldsymbol{\beta}\mathbf{x}_t$ sería, como dijimos, estacionaria.

13.2 Cointegración y corrección del error

Uno de los resultados más importantes en el análisis de cointegración es el denominado *teorema de representación de Granger* (Granger, 1986 y Engle y Granger, 1987). Según este teorema si dos series están cointegradas admitirán una representación en forma de *modelo de corrección de error* (que denotaremos por sus siglas en inglés ECM).

Un modelo de corrección de error es un modelo dinámico en el que se recogen conjuntamente tanto la relación a corto y a largo plazo entre las variables implicadas como el ajuste con el que se corrigen las hipóteticas desviaciones respecto del equilibrio a largo plazo, exigido por la hipotética relación de cointegración. Si seguimos considerando únicamente dos variables, un sencillo modelo de corrección de error podría tener la siguiente expresión:

$$\triangle y_t = \beta \triangle x_t - \gamma \left(y_{t-1} - \delta x_{t-1} \right) + v_t. \tag{13.2}$$

El ECM puede incluir en la parte derecha retardos de las diferencias de las variables, dummies u otras regresoras que de momento ignoraremos para hacer más fácil la exposición. El modelo puede verse como una reparametrización de la ecuación:

$$y_t = \phi_1 y_{t-1} + \varphi_1 x_t + \varphi_2 x_{t-1} + v_t. \tag{13.3}$$

Si en la expresión (13.2) hacemos $\gamma = 1 - \phi_1$ y $\delta = (\varphi_1 + \varphi_2)/(1 - \phi_1)$ entonces coincide con el anterior. Aunque esta última ecuación incluyese más retardos, siempre podría reparametrizarse como un ECM del estilo (13.2), si bien en este caso el modelo de corrección de error incluiría como explicativas retardos de las variables diferenciadas.

Examinemos ahora brevemente las características del modelo (13.2) en el supuesto de que *x* e *y* sean *CI*(1,1). En este caso las series son originalmente I(1), de manera que sus primeras diferencias han de ser estacionarias. El término error es estacionario por definición y, por tanto, para que la ecuación anterior tenga sentido, el término $y_{t-1} - \delta x_{t-1}$ ha de ser también estacionario (γ es una constante). Este último término es pues una combinación lineal entre variables I(1) que, como acabamos de señalar, es estacionaria. Dicha combinación lineal no es otra cosa que la ecuación de cointegración o relación a largo plazo entre las variables que, por hipótesis, habíamos supuesto cointegradas. Como el término entre paréntesis es el error de la ecuación de cointegración, es decir, el error de equilibrio, el ECM puede escribirse:

$$\triangle y_t = \beta \triangle x_t - \gamma \varepsilon_{t-1} + v_t. \tag{13.4}$$

La interpretación es que existe una relación a largo plazo entre las variables (están cointegradas) dada por $y_t = \delta x_t + \varepsilon_t$. Además el modelo recoge también la dinámica a corto, representada por las variables diferenciadas. Por supuesto a corto plazo pueden producirse desviaciones respecto a la relación a largo, pero si hay cointegración es necesario que estas se corrijan en un plazo razonable. En este sentido, el término γ mide la velocidad con la que se produce esa corrección y en consecuencia su valor debería estar comprendido entre 0 y 1. Por ejemplo, si en el periodo *t*-1 se ha producido una desviación positiva, es decir la cantidad observada de *y* es superior a la que correspondería de acuerdo con la relación de cointegración, en el periodo siguiente, es decir, en *t*, una parte importante de esa desviación debe ser compensada: la cantidad *y* en *t* será $\beta \triangle x_t$ menos la parte correspondiente a la mencionada corrección. En este caso ello se traducirá en restar a $\beta \triangle x_t$ la medida de esa compensación, dada por $\gamma \varepsilon_{t-1}$. De esta forma el mecanismo descrito estaría empujando *y* hacia su posición de equilibrio.

Si el valor de y en $t-1$ fuese menor que el que corresponde al equilibrio a largo plazo, el sistema operaría en sentido contrario.

Podemos presentar estas ideas en el marco de un sencillo VAR bivariante, lo que nos servirá para introducir algún concepto adicional. Tomemos por ejemplo los gastos agregados en consumo de los hogares c_t y la renta disponible y_t para los que la teoría postula una relación de equilibrio a largo plazo. Si hay cointegración entre *c* e *y*, entonces si en un periodo concreto el consumo es elevado respecto a la correspondiente relación de equilibrio (es decir, hay desequilibrio), esa discrepancia se debe corregir en los periodos siguientes. El desequilibrio se puede compensar bien con una caída del consumo, o bien con un incremento de la renta o ambas a la vez. En cualquier caso, la dinámica a corto debería verse afectada por la situación de desequilibrio. Ello puede representarse con un modelo como el siguiente:

$$\begin{aligned}\triangle c_t &= -\alpha_c(c_{t-1} - \beta y_{t-1}) + \varepsilon_{1t}\\ \triangle y_t &= \alpha_y(c_{t-1} - \beta y_{t-1}) + \varepsilon_{2t}.\end{aligned}$$

En el modelo anterior, consumo y renta cambian como consecuencia de la existencia de errores de desequilibrio ($c_{t-1} - \beta y_{t-1} \neq 0$). Si la desviación es positiva $c_{t-1} - \beta y_{t-1} > 0$, el consumo caerá y/o la renta crecerá. Nada cambia en la interpretación del sencillo modelo anterior si se incluyen en el VAR términos adicionales, es decir:

$$\begin{aligned}\triangle c_t &= -\alpha_c(c_{t-1} - \beta y_{t-1}) + \lambda_{11}\triangle c_{t-1} + \delta_{11}\triangle y_{t-1} + \varepsilon_{1t}\\ \triangle y_t &= \alpha_y(c_{t-1} - \beta y_{t-1}) + \lambda_{21}\triangle c_{t-1} + \delta_{21}\triangle y_{t-1} + \varepsilon_{2t}.\end{aligned}$$

Los términos α_c y α_y miden la velocidad del ajuste y, si hay cointegración, al menos uno de ellos debe ser significativamente distinto de cero. Si ambos fuesen nulos, habría desaparecido la relación a largo plazo en el sistema anterior: no sería un modelo de corrección de error ni habría cointegración.

Ese modelo se puede ampliar para contemplar más variables. En este caso se tendría, expresado en forma matricial,

$$\triangle \mathbf{w}_t = \boldsymbol{\mu} + \boldsymbol{\pi}\mathbf{w}_{t-1} + \sum_{i=1}^{p} \boldsymbol{\pi}_i \triangle \mathbf{w}_{t-i} + \boldsymbol{\varepsilon}_t,$$

donde la matriz $\boldsymbol{\pi}$ no puede ser nula si existe algún vector de cointegración entre las variables incluidas en $\mathbf{w}$. Más adelante volveremos sobre esta representación del ECM. Retornemos por el momento a la representación del VAR bivariante entre consumo y renta disponible. En un sistema cointegrado de este tipo, en general las dos variables reaccionarán ante una situación de desequilibrio. Sin embargo, es posible que solo una de ellas lo haga. Por ejemplo α_y, el término que mide la velocidad del ajuste al equilibrio en la ecuación de renta, podría ser nulo, mientras que α_c no. En este caso

la renta no responde ante hipotéticos desequilibrios previos en el consumo y todo el ajuste correspondería a la primera ecuación. En estas circunstancias diremos que y_t es *débilmente exógena*.

Para estimar el ECM podemos seguir un procedimiento por etapas similar al que ya hemos visto para el contraste de cointegración. En primer lugar estimamos la Ecuación (13.1). Si las variables están cointegradas entonces los estimadores de los parámetros a largo plazo α y β serán consistentes. A continuación se salvan los residuos que son una estimación de los verdaderos errores de desequilibrio en (13.4). El segundo paso consiste en estimar (13.4). Para determinar si hay que incluir o no retardos de las variables diferenciadas y cuántos en caso afirmativo[2], podemos usar algún criterio del tipo AIC o SBC. En esta fase se obtienen por tanto las estimaciones de los parámetros a corto plazo así como un estimador de γ, que se interpreta como la velocidad del ajuste al equilibrio.

Engle y Granger (1987) han demostrado que, si existe cointegración, los estimadores MCO de esta ecuación son consistentes y asintóticamente eficientes. Asimismo se muestra la consistencia de los errores estándar de estos estimadores.

Como ilustración de todo lo anterior, estimaremos a continuación un ECM con los índices de precios industriales de Alemania y EE.UU. Los datos son mensuales, y corresponden al periodo 1981 - 1997 y están expresados en logaritmos. En primer lugar estudiamos el orden de integración de ambas series. El valor del estadístico ADF para los precios de Alemania y EE.UU. es -2,32 y -1,20 respectivamente. Por lo tanto no es posible rechazar la hipótesis de raíz unitaria. Sin embargo, la hipótesis de que las primeras diferencias de ambas variables es no estacionaria resulta claramente rechazada: ambas series son pues I(1).

A continuación estimamos la ecuación de cointegración, obteniendo (errores estándar entre paréntesis):

$$\hat{Y}_t = \underset{(0,03)}{1,85} + \underset{(,009)}{0,599} X_t,$$

siendo *y* los precios en Alemania y *x* en EE.UU. El contraste de raíces unitarias aplicado a los residuos de la regresión anterior propociona un valor ADF= -4,16, de forma que la hipótesis de raíz unitaria resulta claramente rechazada. Los residuos son estacionarios y por tanto las series de precios en ambos países están cointegradas. La relación de equilibrio a largo plazo vendría dada por la estimacion anterior, siendo la elasticidad 0,60, es decir que durante ese periodo, los precios crecieron menos en Alemania que en EE.UU. Con ello hemos cubierto la primera de las fases conducentes a la estimación del ECM.

La estimación de un ECM como el presentado en (13.4) es:

$$\triangle \hat{Y}_t = \underset{(,0002)}{,0008} + \underset{(0,02)}{0,25} \triangle X_t - \underset{(0,01)}{0,068} \hat{\varepsilon}_t, \quad R^2 = 0,32.$$

[2] A veces se incluyen también diferencias de otras variables I(1) que no aparecen en la relación a largo plazo.

Los estimadores tienen todos ellos los signos adecuados y son estadísticamente significativos. En cuanto a la magnitud de los mismos, la elasticidad a largo plazo sería como hemos dicho, aproximadamente 0,6. La elasticidad a corto plazo sería menor, 0,25, y el ajuste lento toda vez que el valor del EMC, -0,068, implica que se necesitan aproximadamente 5 trimestres para corregir un hipotético desequilibrio, todo ello suponiendo que los residuos de la ecuación de corrección de error tuvieran un comportamiento apropiado. En caso contrario habría que introducir más retardos de las variables hasta conseguirlo.

13.3 Tests de cointegración: Engle-Granger y Johansen

Si la dos variables están cointegradas, podemos estimar la relación estática a largo plazo mediante una simple ecuación de regresión mínimo cuadrática. Teniendo en cuenta la definición de cointegración y lo que estudiamos en temas anteriores, tampoco es difícil imaginar cómo podemos llevar a cabo dicho contraste. Hemos señalado que la condición para que dos series estén cointegradas es que los residuos de la ecuación 13.1 sean estacionarios.

En efecto, puesto que la condición para que las variables estén cointegradas es que el término de error sea estacionario, podemos utilizar su contrapartida empírica para contrastar la cointegración. Esta estrategia se conoce como aproximación de Engle y Granger e implica seguir los siguientes pasos:

1. Comprobar el orden de integración de las series implicadas. Podemos utilizar el test ADF para contrastar si ambas series son I(1). Si las dos resultan ser I(0) no tiene sentido hablar de cointegración. Por otro lado, si no son del mismo orden de integración, entonces tampoco pueden estar cointegradas. Solo si ambas son I(1) se continúa el proceso.

2. A no ser que la ecuación cointegración, $Y_t = \alpha + \beta X_t + \varepsilon_t$ representativa de la relación a largo plazo sea conocida, lo que no suele suceder en la práctica, el siguiente paso es estimarla. Para ello empleamos MCO. Ya hemos dicho que si X e Y están cointegradas, MCO proporciona estimadores superconsistentes de α y β. A partir de los estimadores MCO se obtiene la serie de los residuos estimados $e_t = \hat{\varepsilon}_t$. Esta serie es una estimación de las desviaciones respecto del equilibrio a largo plazo y, como hemos señalado, esta serie debe de ser estacionaria si realmente hay una relación de cointegración entre las variables implicadas.

3. Para contrastar la estacionariedad de la serie de e_t empleamos de nuevo un test ADF, es decir calculamos, $\triangle e_t = \delta e_{t-1} + \sum_{i=1}^{k} \lambda_i \triangle e_{t-i} + \varepsilon_t$ y procedemos, como hacíamos en los contrastes de este tipo, a contrastar la hipótesis nula $H_0 : \delta = 0$ contra la alternativa unilateral $H_1 : \delta < 0$. Si es posible rechazar esta hipótesis, entonces concluimos que las series están cointegradas. Por tanto, rechazaremos que las variables estén cointegradas si el $\hat{\delta}/ee(\hat{\delta})$ es mayor que el valor crítico en tablas para el nivel de significatividad elegido. Aquí surge un problema derivado del hecho de que e_t no representa realmente el error de desequilibrio, sino solo su estimación y dado el método empleado (MCO), el procedimiento estaría sesgado

Tabla 13.1: Test de raíces unitarias

Serie	ADF
LT	$-1{,}66$
ST	$-1{,}05$
$\triangle LT$	$-7{,}23$
$\triangle ST$	$-6{,}42$

hacia la estacionariedad de e_t. Para solventar este problema hemos de emplear tablas diferentes, en las que los valores críticos son más elevados en valor absoluto. Por ejemplo, para 100 observaciones y un nivel de significatividad del 5 %, el valor crítico es -3,39, mucho más negativo que -1,95, el valor crítico habitual en las tablas ADF para esta ecuación y nivel de significatividad (la ecuación de contraste no tiene término independiente).

Por ejemplo, tomemos las series representadas en el panel de la izquierda de la Figura 13.1. El contraste ADF aplicado a las mismas, arroja el resultado de la Tabla 13.1.

Ambas son pues I(1) de manera que tienen el mismo orden de integración.

El segundo paso es estimar la regresión, obteniéndose:

$$\widehat{LT_t} = -2{,}11 + 1{,}15ST_t.$$

Los residuos se representan gráficamente en la misma Figura 13.1 y visualmente tienen todo el aspecto de ser estacionarios. El contraste ADF aplicado a los mismos proporciona para el estadístico empírico de un valor de -4,69. El valor crítico en las tablas para el nivel del 5 % es (para 100 observaciones) -3,39 (-4,00 para el 1 %). Como el valor del estadístico de contraste es menor (más negativo) que el valor crítico, podemos rechazar la hipótesis nula ($H_0 : \delta = 0$) y por lo tanto los residuos son estacionarios, de manera que las series están cointegradas y el vector de cointegración es (-2,11, 1,15).

Podemos decir entonces que hay evidencia suficiente para sostener la existencia de una relación a largo plazo entre ambas variables. El procedimiento en dos pasos de Engle y Granger, aunque no exento de problemas, es una buena estrategia de contraste cuando solo hay dos variables implicadas. En la práctica hay más de dos variables implicadas. Especialmente con las relaciones económicas ya que a largo plazo suelen incluir más de dos variables. La teoría de la demanda incluye junto con el precio, al menos la renta; la demanda de dinero, la renta y el tipo de interés, etc. En estos casos u otros más complejos, el contraste de Engle y Granger no es apropiado. El problema fundamental es que en este tipo de ecuaciones puede haber más de una relación de equilibrio a largo y, aunque sigue siendo cierto que si los residuos son estacionarios existe una relación de cointegración, esta estrategia no nos permitiría distinguir la hipotética existencia de varias.

Por ejemplo, supongamos que tras estimar la ecuación $Y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \varepsilon_t$ donde todas las variables implicadas son I(1), se encuentra que los residuos son estacionarios. Entonces podemos afirmar que hay cointegración. Pero seríamos incapaces de

distinguir cuántas relaciones de cointegración hay. Una posibilidad es que haya una única combinación lineal entre las cuatro variables que sea estacionaria. Pero también podría ser que hubiera una relación de cointegración entre Y_t y X_{1t} y otra entre X_{2t} y X_{3t}. Sean v_1 y v_2 los residuos de tales relaciones

$$v_1 = Y_t - \delta_0 - \delta_1 X_{1t}$$
$$v_2 = X_{2t} - \lambda_0 - \lambda_1 X_{3t}.$$

Dado que hemos postulado la existencia de cointegración, dichos residuos han de ser I(0) y, por definición, cualquier combinación lineal de dos variables I(0), es también I(0). Por ejemplo, la suma de ambos residuos:

$$Y_t - \delta_0 - \delta_1 X_{1t} + X_{2t} - \lambda_0 - \lambda_1 X_{3t}.$$

también es estacionaria. De esta manera tenemos una combinación lineal estacionaria entre las cuatro variables, pero con el contraste descrito anteriormente no seríamos capaces de identificar todas las relaciones de cointegración.

Para solventar estos problemas se han desarrollado contrastes más apropiados, siendo probablemente el test de rango de cointegración de Johansen (1988) uno de los más utilizados[3]. El método es bastante más complejo, de manera que nos limitamos aquí a una somera explicación. Afortunadamente la práctica totalidad de los programas econométricos permiten llevar a cabo de forma rutinaria este test.

Johansen basa su metodología en los modelos VAR. Supongamos que deseamos estudiar las posibles relaciones de cointegración entre un grupo de *k* variables incluidas en el vector $\mathbf{W}_t$. Consideremos entonces el VAR

$$\mathbf{W}_t = \mathbf{C}\mathbf{X}_t + \sum_{i=1}^{p+1} \boldsymbol{h}_i \mathbf{W}_{t-i} + \mathbf{u}_t. \tag{13.5}$$

donde **W** es un vector de dimensión $k \times 1$ con las variables implicadas, que asumiremos I(1). Como hemos dicho, el caso más interesante en Economía es aquel en el que las variables son CI(1,1). La matriz **X** contiene variables como tendencias, dummy, etc., y puede o no ser incluida en la ecuación; **u** es el vector de los errores y $\boldsymbol{h}_i$ son matrices de dimensión $k \times k$. Si eliminamos **X** para simplificar la exposición, el sistema anterior puede ser reparametrizado como[4]:

[3] Pero no el único: podría emplearse también aquí un contraste basado en un ECM.

[4] Por ejemplo, en un VAR(2) se tendría:

$$\mathbf{w}_t = \mathbf{h}_1 \mathbf{w}_{t-1} + \mathbf{h}_2 \mathbf{w}_{\mathbf{t-2}} + \mathbf{u}_t;$$

si ahora sumamos y restamos $h_2 w_{t-1}$ se obtiene

$$\mathbf{w}_t = (\mathbf{h}_1 + \mathbf{h}_2)\mathbf{w}_{t-1} + \mathbf{h}_2 \triangle \mathbf{w}_{\mathbf{t-1}} + \mathbf{u}_t;$$

$$\triangle \mathbf{W}_t = \mathbf{\Pi W}_{t-1} + \sum_{i=1}^{p} \mathbf{\Pi}_i \triangle \mathbf{W}_{t-i} + \mathbf{u}_t. \tag{13.6}$$

Obsérvese la similitud entre 13.6 y la ecuación de contrastación de Dickey y Fuller. De la misma forma que allí nuestro interés estaba en el coeficiente de y_{t-1} con objeto de contrastar la existencia de una raíz unitaria, aquí nos centraremos en la matriz $\mathbf{\Pi}$, que referida a 13.6 viene dada por

$$\mathbf{\Pi} = \sum_{i=1}^{p+1} \boldsymbol{h}_i - \mathbf{I}_{kk}, \tag{13.7}$$

siendo $\mathbf{\Pi}_i = -\sum_{j=i+1}^{p} \mathbf{h}_j$.

En la ecuación 13.6 las variables diferenciadas y el vector de los errores son todos ellos estacionarios, de manera que para que la igualdad se cumpla, el término $\mathbf{\Pi W}_{t-1}$, donde las variables incluidas en $\mathbf{W}$ son I(1), debe ser asimismo estacionario. La estacionariedad de ese término implica que la matriz $\mathbf{\Pi}$ contiene los coeficientes necesarios para formar las combinaciones lineales estacionarias entre las variables de $\mathbf{W}$, es decir, cada fila de dicha matriz es un vector de cointegración. De hecho el rango de $\mathbf{\Pi}$ determina el número de relaciones de cointegración entre las k variables.

Por ejemplo, un caso elemental de 13.6 sería el VAR[5]:

$$\begin{pmatrix} \triangle x_t \\ \triangle y_t \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{21} \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}. \tag{13.8}$$

Si $\mathbf{\Pi} = \mathbf{0}$ entonces las ecuaciones del VAR están equilibradas en el sentido de que todos sus términos son estacionarios, pero en este caso no hay cointegración: si todos los elementos de $\mathbf{\Pi}$ son nulos, no puede decirse que haya una combinación lineal estacionaria entre x e y. Análogamente si $\mathbf{\Pi}$ no es de rango reducido, es decir si el rango es 2, entonces habría dos relaciones de cointegración, pero entre dos variables solo puede haber como máximo una relación de cointegración independiente, de manera que si el rango es 2 se deduce que ambas variables son I(0) y no tiene por tanto sentido hablar de cointegración.

El caso más interesante es aquel en el que el rango es 1, lo que quiere decir que las dos columnas de Π no son linealmente independientes. En este caso es posible factorizar la matriz $\mathbf{\Pi}$ como

$$\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^{\boldsymbol{T}}, \tag{13.9}$$

y restando $\mathbf{h}_1 + \mathbf{h}_2$ a ambos lados de la igualdad:

$$\triangle \mathbf{w}_t = (\mathbf{h}_1 + \mathbf{h}_2 - \mathbf{I})\mathbf{w}_{t-1} + \mathbf{h}_2 \triangle \mathbf{w_{t-1}} + \mathbf{u}_t = \boldsymbol{\pi} \mathbf{w}_{t-1} + \mathbf{h}_2 \triangle \mathbf{w}_{t-1} + \mathbf{u}_t.$$

[5]Con las variables en desviaciones con respecto a sus medias.

donde $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$ son ambas matrices de dimensión 2x1. Ahora el primer término a la derecha de la igualdad, $\boldsymbol{\Pi}\mathbf{W}_{t-1} = \boldsymbol{\alpha}\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{W}_{t-1}$ y este será estacionario si $\boldsymbol{\beta}\mathbf{W}_{t-1}$ es I(0), lo que significa que el vector $\boldsymbol{\beta}$ contiene los coeficientes de la combinación lineal estacionaria entre las dos variables I(1) de $\mathbf{W}$, es decir que $\boldsymbol{\beta}$ es el vector de cointegración. Por su parte los coeficientes de $\boldsymbol{\alpha}$ medirían la velocidad de ajuste al equilibrio en los ECM resultantes de la ecuación factorizada. Por ejemplo, supongamos que una vez efectuada la factorización el sistema (13.8) queda:

$$\begin{pmatrix} \triangle x_t \\ \triangle y_t \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \begin{pmatrix} \delta_1 & \delta_2 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}.$$

El vector de cointegración es $\begin{pmatrix} \delta_1 & \delta_2 \end{pmatrix}$ o, normalizado en x, $\begin{pmatrix} 1 & \frac{\delta_2}{\delta_1} \end{pmatrix}$. Entonces tenemos:

$$\triangle x_t = \lambda_1 \left(\delta_1 x_{t-1} + \delta_2 y_{t-1}\right) + \varepsilon_{1t} = \lambda_1\delta_1 \left(x_{t-1} + \frac{\delta_2}{\delta_1} y_{t-1}\right) + \varepsilon_{1t}$$

$$\triangle y_t = \lambda_2 \left(\delta_1 x_{t-1} + \delta_2 y_{t-1}\right) + \varepsilon_{2t} = \lambda_2\delta_1 \left(x_{t-1} + \frac{\delta_2}{\delta_1} y_{t-1}\right) + \varepsilon_{2t}.$$

Análogamente en el caso general con k variables en 13.6, el rango de la matriz $\boldsymbol{\Pi}$ indica el número de relaciones de cointegración independientes. Sea $r < k$ el rango de dicha matriz. Entonces dado que $\boldsymbol{\Pi}$ es de rango reducido, la factorizamos como $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^{\boldsymbol{T}}$, siendo $k \times r$ las dimensiones de las matrices $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$. Como antes, las r filas de la matriz $\boldsymbol{\beta}^{\boldsymbol{T}}$ son los vectores de cointegración del sistema.

Volvamos de nuevo a (13.6) y consideremos un ejemplo hipótetico. Supongamos que en un sistema con tres variables hemos obtenido:

$$\begin{pmatrix} \triangle y_{1t} \\ \triangle y_{2t} \\ \triangle y_{3t} \end{pmatrix} = \begin{pmatrix} -1/2 & 5/16 & -1/16 \\ 1/8 & -41/64 & 5/32 \\ 1/4 & 11/32 & -3/32 \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \\ y_{3t-1} \end{pmatrix} + \sum_{i=1}^{p} \boldsymbol{\Pi}_i \triangle \mathbf{W}_{t-i} + \mathbf{u}_t.$$

El último término de la ecuación puede ser ignorado sin pérdida de generalidad. A simple vista no es posible ver si las columnas (filas) de la matriz $\boldsymbol{\Pi}$ son linealmente independientes. Para calcular el rango podemos emplear la propiedad según la cual el rango de una matriz es igual al número de valores propios distintos de cero. Los autovalores de $\boldsymbol{\Pi}$ son en este caso (-0,79, -0,44, 0). Como solo dos de ellos son distintos de cero el rango es 2, es decir, la matriz es de rango reducido y podemos afirmar que hay dos relaciones de cointegración entre las variables del sistema. La matriz $\boldsymbol{\Pi}$ puede factorizarse como

$$\boldsymbol{\alpha} = \begin{pmatrix} -1/2 & 1/4 \\ 1/8 & -5/8 \\ 1/4 & 3/8 \end{pmatrix} \quad \boldsymbol{\beta}^{\boldsymbol{T}} = \begin{pmatrix} 1 & -1/8 & 0 \\ 0 & 1 & -1/4 \end{pmatrix},$$

de manera que las dos relaciones de cointegración serían $y_{1t} = 0{,}125y_{2t} + v_{1t}$ e $y_{2t} = 0{,}25y_{3t} + v_{2t}$ donde ambos vectores han sido normalizados en y_1 e y_2. En términos de ECM, ignorando como hemos dicho $\sum_{i=1}^{p} \mathbf{\Pi}_i \triangle \mathbf{W}_{t-i}$ tendríamos:

$$\triangle y_{1t} = -0{,}5v_{1t-1} + 0{,}25v_{2t-1} + u_{1t}$$

$$\triangle y_{2t} = 0{,}125v_{1t-1} - 0{,}625v_{2t-1} + u_{2t}$$

$$\triangle y_{3t} = 0{,}25v_{1t-1} + 0{,}375v_{2t-1} + u_{3t}.$$

El sistema (13.6) junto con la restricción expresada en (13.9) queda,

$$\triangle \mathbf{W}_t = \boldsymbol{\alpha}\boldsymbol{\beta}^{\boldsymbol{T}} \mathbf{W}_{t-1} + \sum_{i=1}^{p} \mathbf{\Pi}_i \triangle \mathbf{W}_{t-i} + \mathbf{u}_t, \tag{13.10}$$

que una vez estimado proporciona, como hemos visto, tanto las relaciones de cointegración como los parámetros de ajuste. El sistema anterior con las restricciones impuestas es no lineal y en consecuencia es necesario emplear algún procedimiento de estimación diferente del método de mínimos cuadrados ordinarios. Lo habitual es estimar esta ecuación por máxima verosimilitud. Además hay un problema de identificación derivado del hecho de que la factorización (13.9) no es única. El vector o vectores de cointegración no estarán identificados a menos que impongamos alguna normalización arbitraria, similar a la que hacemos implícitamente en cualquier modelo de regresión.

Como hemos señalado, el método de Johansen (1988 y 1992) está basado en la autorregresión vectorial. Por tanto el primer paso es estimar un VAR entre las variables que, según la teoría económica y/o el trabajo empírico previo, mantienen relaciones a largo plazo. El sistema 13.6 es, como hemos visto, una reparametrización del VAR entre las variables originales, donde adicionalmente pueden incluirse variables deterministas (término independiente, dummy, etc). Para determinar el orden del VAR original podemos emplear alguno de los criterios estadísticos señalados con anterioridad, asegurándonos de que los residuos del sistema cumplen las hipótesis necesarias. Si el orden del VAR entre las variables en niveles es *p*, el VAR en primeras será de orden *p*-1.

Inicialmente no habrá ninguna restricción sobre el rango de la matriz Π de manera que las matrices que la factorizan serán ambas de orden $k \times k$. Utilizaremos diversos test de hipótesis para contrastar restricciones de nulidad sobre los elementos de las mismas.

Puede demostrarse que la maximización de la función logarítmica de verosimilitud del modelo restringido conduce a

$$-\frac{kN}{2}(log2\pi + 1) - \frac{N}{2}\sum_{i=1}^{r} log(1 - \lambda_i),$$

siendo λ_i los autovalores de Π que podemos estimar como sigue.

En primer lugar estimamos por MCO las ecuaciones

$$\triangle \mathbf{W}_t = \sum_{i=1}^{p} \mathbf{\Pi}_i \triangle \mathbf{W}_{t-i} + \mathbf{u}_t$$

$$\mathbf{W}_{t-1} = \sum_{i=1}^{p} \mathbf{\Pi}_i \triangle \mathbf{W}_{t-i} + \mathbf{v}_t,$$

incluyendo si es oportuno una constante y una matriz **X** con variables adicionales. Dado que en **W** están las *k* variables del sistema, ello exigirá estimar 2*k* ecuaciones de regresión.

Se salvan los residuos de cada una de las regresiones y se calculan las matrices de varianzas y covarianzas,

$$\hat{\mathbf{\Sigma}}_{uu} = \frac{1}{N} \sum \hat{\mathbf{u}}\hat{\mathbf{u}}^T, \;\; \hat{\mathbf{\Sigma}}_{vv} = \frac{1}{N} \sum \hat{\mathbf{v}}\hat{\mathbf{v}}^T, \;\; \hat{\mathbf{\Sigma}}_{vu} = \frac{1}{N} \sum \hat{\mathbf{v}}\hat{\mathbf{u}}^T \;\; y \;\; \hat{\mathbf{\Sigma}}_{uv} = \hat{\mathbf{\Sigma}}_{vu}^T.$$

Entonces la matriz que nos interesa viene dada por:

$$\hat{\mathbf{\Pi}} = \hat{\mathbf{\Sigma}}_{vv}^{-1} \hat{\mathbf{\Sigma}}_{vu} \hat{\mathbf{\Sigma}}_{uu}^{-1} \hat{\mathbf{\Sigma}}_{uv}. \tag{13.11}$$

Conocida **Π** hallamos sus valores propios λ_i ordenándolos de mayor a menor. Un test de ratio de verosimilitud apropiado para contrastar el número de valores propios distintos de cero, es:

$$\lambda_{traza} = -N \sum_{i=r+1}^{k} log(1 - \lambda_i), \tag{13.12}$$

conocido en la literatura como *estadístico de la traza*. En este contraste la hipótesis nula es que el número de valores propios distintos de cero es menor o igual que r_0 contra la alternativa de que hay al menos $r_0 + 1$ de ellos distintos de cero, es decir:

$$H_0 : r \leq r_0, \;\; H_A : r \geq r_0 + 1.$$

Si no existe ninguna relación de cointegración entre las variables, entonces el rango de Π será nulo, o lo que es lo mismo, todos los valores propios serán nulos. Por lo tanto todos los términos $log(1 - \lambda_i)$ serán nulos y 13.12 también se anulará. Por el contrario, si un autovalor λ_1 es distinto de cero, entonces el término $log(1 - \lambda_1)$ será también distinto de cero y 13.12 ya no será nulo.

En la práctica lo que tenemos son estimaciones de **Π** y de sus autovalores. Una vez ordenadas las estimaciones de los valores propios, el test se lleva a cabo de forma secuencial:

1. Se comienza por contrastar la hipótesis $H_0 : r = 0$ contra la alternativa $H_A : r \geq 1$. Si esta hipótesis no puede ser rechazada, se detiene el proceso y se concluye que no hay relaciones de cointegración.
2. Si se rechaza la hipótesis nula anterior, continuamos con el contraste de $H_0 : r = 1$ contra $H_A : r \geq 2$. Si esta hipótesis no se puede rechazar se detiene el proceso, concluyendo que hay una relación de cointegración. Si no es así, continuamos con el mismo.
3. La última posibilidad, consiste en contrastar $H_0 : r = k - 1$ contra $H_A : r = k$. Si no es posible rechazar H_0 concluimos que hay *k*-1 relaciones de cointegración entre las *k* variables, deteniéndonos en este punto.
4. Si se rechazase la última hipótesis nula, habría que concluir que hay *k* relaciones de cointegración entre las *k* variables, lo que implicaría que todas ellas son estacionarias, por lo que carecería de sentido el análisis de cointegración.

Cuando la hipótesis nula es que hay r_0 vectores de cointegración contra la alternativa de que hay $r_0 + 1$, entonces solo hay un término en el sumatorio de 13.12, siendo el estadístico:

$$\lambda_{máx} = -Nlog(1 - \lambda_{r_0+1}). \tag{13.13}$$

Esta versión del test recibe el nombre de *estadístico máximo*. La única diferencia entre ambos estadísticos es que ahora cambia la hipótesis alternativa que, en el caso del estadístico máximo, resulta restringida a que el rango sea una unidad mayor que la postulada por la hipótesis nula. Con ello se consigue mejorar la potencia del contraste. En todo caso, el procedimiento secuencial de contrastación es similar:

1. El primer paso será contrastar $H_0 : r = 0$ contra $H_A : r = 1$, empleando $\lambda_{máx} = -Nlog(1 - \lambda_1)$. Si esta hipótesis no se rechaza, se detiene el proceso no habiendo encontrado evidencia de cointegración.
2. A continuación contrastamos $H_0 : r \leq 1$ contra $H_A : r = 2$, siendo ahora el estadístico $\lambda_{máx} = -Nlog(1 - \lambda_2)$. Si no se rechaza, hemos hallado un vector de cointegración. En caso contrario continuamos con el procedimiento, de forma análoga.
3. La última posibilidad $H_0 : r \leq k - 1$ contra $H_A : r = k$, se contrastaría con $\lambda_{máx} = -Nlog(1 - \lambda_k)$.

Como en el caso del contraste ADF, la distribución de estos estadísticos no es estándar y sus valores han sido obtenidos por simulación.

Como ilustración, tomemos las series españolas de importaciones y producto interior bruto correspondientes al periodo 1983q1 - 1998q4, para las que se ha estimado un VAR(2) y a continuación obtengamos:

$$\hat{\mathbf{\Pi}} = \begin{pmatrix} -1{,}3573 & -0{,}4837 \\ 4{,}6731 & 1{,}6642 \end{pmatrix}.$$

Tabla 13.2: Contraste de cointegración: estadístico máximo

H. nula	H. alternativa	Est. máximo	Valor crítico (5 %)	valor p
$H_0 : r = 0$	$H_1 : r = 1$	23,05	14,26	0,002
$H_0 : r \leq 1$	$H_1 : r = 2$	0,29	3,84	0,59

Los valores propios de esta matriz son aproximadamente $\lambda_1 = 0{,}3024,\ \lambda_2 = 0{,}0045$. Por tanto el contraste de la hipótesis nula de que no existe ningún vector de cointegración (r = 0), contra la alternativa de que existen 2, empleando el estadístico de la traza, será:

$$-64\left[log\left(1-0{,}3024\right)+log(1-0{,}0045)\right]=24{,}07.$$

Como el valor crítico[6] al 5 % es aproximadamente 15.49, rechazamos la hipótesis nula y aceptamos que hay al menos un vector de cointegración. A continuación contrastaríamos $H_0 : r = 1$ contra la alternativa $H_A : r = 2$,

$$-64\left[log(1-0{,}0045)\right]=0{,}29.$$

Ahora el valor del estadístico de contraste es menor que el crítico al 5 % (3,84 en este caso), de manera que detendríamos aquí el procedimiento, concluyendo que existe una relación de cointegración (tampoco tendría sentido ya seguir dado que solo hay dos variables).

En cuanto al estadístico máximo, mostramos sus resultados en la Tabla 13.2. Como puede verse, en este caso ambos estadísticos llevan a la misma conclusión: hay un vector de cointegración.

Conviene señalar que, igual que sucedía con el test ADF, los valores críticos de estos contrastes son muy sensibles al tipo de ecuación empleada. En particular dependen de forma crucial de si se incluyen o no términos deterministas. Ello da lugar a un elevado número de tablas, lo que puede resultar confuso. Sin embargo, la situación suele ser más sencilla dado que dos posibilidades son con mucho las más frecuentes en la práctica.

Bibliografía complementaria

Matilla-García, M. et al. 2017. Econometría y Predicción. McGraw Hill

Engle, R. E. y Granger, C. W. (1987). «Cointegration and Error Correction: Representation, Estimation and Testing». Econometrica, 55, pp. 251-276

Johansen, S. (1988). «Statistical Analysis of Cointegration Vectors». Journal of Economic Dynamics and Control, vol. 12, pp.231-254.

[6] En este caso, los valores críticos corresponden a una ecuación con constante pero sin tendencia ni variables exógenas adicionales.

Johansen, S. (1992). «Determination of Cointegration Rank in the Presence of Linear Trends». Oxford Bulletin of Economics and Statistics, vol. 54, pp. 383-397.

Tema 14

Ajuste estacional, desagregación temporal y calibrado de series temporales.

Este tema está elaborado como una adaptación de:

Peña, Tiao and Tsay, eds. A Course in Time Series Analysis. Wiley 2001. Cap. 8.

Dagum and Cholette Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series. Springer 2006, Capítulo 1. Así como de la bibliografía complementaria.

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al Órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

- Introducción.
- Componentes deterministas, ajuste de calendario.
- Métodos no paramétricos y paramétricos de ajuste estacional, la descomposición canónica.
- Problemas prácticos.
- Desagregación temporal.
- Calibrado, benchmarking y reconciliación.

14.1 Introducción

Muchas series de tiempo económicas tienden a seguir un patrón regular a lo largo de cada año. Este tipo de comportamiento se conoce como variación estacional o estacionalidad. Puede ser el resultado de patrones climáticos estacionales regulares o de costumbres sociales tales como periodos festivos (días y "puentes"), vacaciones de verano y similares. Las series de tipo mensual, trimestral, semanal, a menudo se ven afectados por variaciones estacionales. La presencia de tales variaciones en las actividades socioeconómicas se reconoce desde hace mucho tiempo. De hecho, la estacionalidad suele representar la mayor parte de la variación total dentro del año. La estacionalidad se debe a que algunos meses, trimestres del año son más importantes en términos de actividad o nivel. Por ejemplo, el nivel de desempleo es generalmente más alto durante los meses de invierno y primavera y más bajo en los demás meses.

Así pues no es extraño que a muchas series de tiempo que se observan trimestralmen-

te, mensualmente, semanalmente o diariamente muestren alguna forma de estacionalidad, y esto puede tener implicaciones importantes para el trabajo econométrico aplicado.

La estacionalidad se origina en el clima y las estaciones convencionales, como eventos religiosos, sociales y cívicos, que se repiten de año en año. Las estaciones climáticas influyen en el comercio, la agricultura, los patrones de consumo de energía, la pesca, la minería y actividades relacionadas. Por ejemplo, el consumo de gasóleo para calefacción aumenta en invierno y el consumo de electricidad aumenta en los meses de verano debido al aire acondicionado.

Las fiestas institucionales como Navidad, Semana Santa, vacaciones, el año escolar y académico tienen un gran impacto en el comercio minorista y en el consumo de ciertos bienes y servicios, a saber, viajes en avión, ocupación hotelera, consumo de gasolina. Para determinar si una serie contiene estacionalidad, es suficiente identificar al menos un mes (o trimestre) que tiende a ser sistemáticamente más alto o más bajo que otros meses.

Las cuatro causas principales de la estacionalidad se atribuyen al clima, la composición del calendario, los principales plazos institucionales y las expectativas. La estacionalidad es en gran medida exógena al sistema económico, pero puede compensarse parcialmente con la intervención humana. Por ejemplo, la estacionalidad de la oferta monetaria puede controlarse mediante decisiones del banco central sobre las tasas de interés. En otros casos, los efectos pueden compensarse con el comercio internacional e interregional.

No tener en cuenta adecuadamente la estacionalidad puede hacer que hagamos inferencias incorrectas lo que es un serio incoveniente para el trabajo econométrico aplicado que utiliza datos de series de tiempo.

Hay dos puntos de vista bastante diferentes sobre la naturaleza de la estacionalidad en los datos económicos. Un punto de vista es que la variación estacional es una parte fundamental de muchas series de tiempo económicas y, cuando está presente, deberíamos intentar explicarla. Así, idealmente, un modelo econométrico para una variable dependiente y debería explicar cualquier variación estacional en él mediante la variación estacional en las variables independientes, quizás incluyendo variables meteorológicas o variables ficticias estacionales entre estas últimas.

Una segunda visión es que la estacionalidad es simplemente un tipo de ruido que contamina los datos económicos. No se puede esperar que la teoría económica explique este ruido, que en el caso de las variables independientes equivale a una especie de problema de errores en las variables. Por tanto, conviene utilizar lo que se denomina datos desestacionalizados o datos ajustados estacionalmente, es decir, datos que se han sido transformados de alguna forma para que supuestamente representen lo que habría sido la serie en ausencia de estacionalidad. Esta es una labor propiamente de las oficinas de estadística pública, y a estos efectos su labor consiste en publicar cifras ajustadas estacionalmente para muchas series. Este tema se centra en esta segunda visión.

La idea de ajustar estacionalmente una serie de tiempo para eliminar los efectos de la estacionalidad es intuitivamente atractiva. El ajuste estacional de una serie y_t tiene

sentido si para todo t podemos escribir $y_t = y^* + y_s$, donde y^* es una serie temporal que no contiene ninguna variación estacional, e y_s es una serie temporal que contiene únicamente la variación estacional. Es obvio que esta es una suposición fuerte, y que incluso en si fuera cierta, sería todo un reto dividir o separar ambos componentes. Analizamos a continuación las técnicas disponibles a estos efectos.

14.2 Componentes deterministas, ajuste de calendario

La estructura y composición del calendario afecta evidentemente a la actividad económica. Intentar desestacionalizar una serie requiere identificar este tipo de componentes de calendario y ajustar la serie convenientemente. Los efectos de calendario que generalmente se contemplan son:

- El diferente número de días hábiles en cada mes. El ajuste de días hábiles tiene como finalidad obtener una serie cuyos valores no dependen de la extensión del mes o trimestre, ni de la distribución del número de días hábiles.
- La diferente composición del número de días hábiles. Se recomienda comprobar si existen diferencias entre los distintos días laborables, sobre todo en series mensuales largas que miden la evolución de alguna variable o fenómeno en el que, a priori, el comportamiento puede ser diferente cada día de la semana.
- El efecto de año bisiesto. Cada cuatro años, el mes de febrero tiene 29 días, lo que puede afectar a la serie económica por dos motivos: la composición del mes varía (en días laborables o no laborables respecto a la media) y la duración de la se cambia el mes. El primer efecto se mide con el regresor para los días laborables (donde el día 29 debe considerarse laborable o no laborable, según corresponda); mientras que el segundo efecto debe cuantificarse con el regresor del año bisiesto.
- Las fiestas móviles, en el caso español la Semana Santa. El ajuste por fiestas móviles tiene como objetivo eliminar aquellos valores que se ven afectados por eventos que siguen un patrón complejo a lo largo de los años de la serie. La fiesta que más afecta a nuestra serie son las vacaciones de Semana Santa. Además, en este caso, este efecto es parcialmente estacional, ya que en promedio se celebra con más frecuencia en abril que en marzo. Dado que la parte estacional debe capturarse en el componente estacional, tampoco debe eliminarse con la corrección del efecto festivo de Semana Santa.

Estos efectos o de naturelza similar se pueden expresar mediante variables de regresión que comentaremos más adelante. El método o herramienta sugerida es utilizar un modelo regresión ARIMA (regARIMA), y aplicarlo con contrastes de significatividad y plausibilidad de los efectos

$$\mathbf{y}_t = \underbrace{\mathbf{W}_t\boldsymbol{\beta}}_{\text{COMPONENTE DETERMINISTA}} + \underbrace{\mathbf{x}_t}_{\text{COMPONENTE ESTOCÁSTICO}}$$

donde $\mathbf{y}_t$ es la serie observada, $\mathbf{W}_t$ es la matriz que en sus filas contendría los regresores que capturarían dichos efectos, y $\mathbf{x}_t$ es un modelo ARIMA, y será el objeto de estudio.

Consideremos que el tipo de serie mensual que estamos estudiando tiene dos elementos a ser considerados dentro del denominado efecto calendario: efecto Semana Santa (E) y el efecto ciclo semanal (CS) que el efecto del ciclo semanal relativo a patrones de series económicas que se encuentran dentro de la semana.

La modelización de estos elementos es de tipo determinista:

$$E_t = \gamma P_t$$

donde P_t expresa la proporción que representa la semana de Pascua en el mes t, y

$$CS_t = \delta D_t$$

siendo D_t = (número de lunes, martes, miércoles, jueves y viernes en el mes t) - (número de sábados y domingos en el mes t)*(5/2). El factor $5/2$ sirve para homogeneizar los dos elementos de la diferencia que da lugar a D_t.

Así el efecto calendario total sería la suma de efectos particulares asociados.

$$\mathbf{W}_t\boldsymbol{\beta} = E_t + CS_t = \gamma P_t + \delta D_t.$$

Como hemos dicho la cuantificación de este efecto se realiza mediante la identificación, estimación y diagnóstico de un modelo de regresión cuya perturbación sigue una representación autorregresiva, integrada y de medias móviles (ARIMA) que suele ser de tipo multiplicativo, es decir

$$\mathbf{y}_t = \mathbf{W}_t\boldsymbol{\beta} + \frac{\theta_q(B)\theta_Q(B^s)}{\phi_p(B)\phi_P(B)(1-B)^d(1-B^s)^D}a_t$$

donde $\phi_p(B)$ y $\theta_q(B)$ son, respectivamente, polinomios de orden p y q en el operador de retardo B, y $\varphi_P(B^s)$ y $\theta_Q(B^s)$ son polinomios de orden P y Q en B^s, con $s = 4$. Las expresiones $(1-B)d$ y $(1-B^s)D$ son operadores de diferenciación regular y estacional controlados por los parámetros enteros d y D, respectivamente. Por último, a_t es una secuencia de ruido blanco gaussiano con esperanza nula y varianza constante.

Una vez que se contrastan independientemente la significatividad estadística de las hipótesis nulas $\gamma = 0$ (ausencia de efectos Semana Santa) y $\delta = 0$ (ausencia de ciclo semanal), se obtendría la serie libre o corregida de efectos de calendario

$$x_t = y_t - \hat{\gamma}P_t - \hat{\delta}D_t, t = 1, ..., T$$

donde las estimaciones realizadas se suelen llevar a cabo mediante el programa TRAMO desarrollado por Maravall y Gómez en el Banco de España.

14.3 Métodos no paramétricos y paramétricos de ajuste estacional, la descomposición canónica.

El modelo identificado, estimado y diagnosticado en la sección anterior permite realizar una descomposición de la serie, posiblemente corregida de efectos de calendario, en

sus componentes subyacentes estocásticos de tendencia, estacionalidad e irregularidad, siguiendo los principios de descomposición canónica basada en modelos ARIMA.

Los dos enfoques más utilizados para el ajuste estacional de series de tiempo son el ajuste estacional basado en modelos ARIMA y el de filtros fijos. Anteriormente a este tipo de descomposición, se utilizaban los métodos de descomposición clásica, que en buena medida sirven de base para cierta parte de la lógica de algunas metodologías de descomposición, tanto la fundamentada en regresiones-ARIMA, como para la de filtros fijos. Explicaremos ahora brevemente en qué consiste y seguiremos con las técnicas más utilizadas actualmente.

Podemos representar cualquier serie de tiempo Y_t como la suma o producto de los tres componentes (el componente estacional S_t, tendencial T_t y residual I_t)Si adoptamos el esquema aditivo, la serie la escribimos como

$$Y_t = S_t + T_t + I_t, \tag{14.1}$$

y si el esquema es multiplicativo lo expresamos como

$$Y_t = S_t \cdot T_t \cdot I_t. \tag{14.2}$$

La forma aditiva es más adecuada cuando la variación alrededor de los componentes tendencial y estacional no varía con el nivel de la serie histórica. Cuando la variación de los componentes es proporcional al nivel de la serie, entonces es mejor utilizar el esquema multiplicativo, este esquema multiplicativo es muy común cuando nos referimos a series económicas.

Cuando utilizamos el modelo multiplicativo, en ocasiones se realiza la transformación logarítmica, es decir se aplica el esquema aditivo para la serie transformada en logaritmos. En efecto, aplicando logaritmos a (14.2) tenemos que

$$\ln(Y_t) = \ln(S_t \cdot T_t \cdot I_t) = \ln(S_t) + \ln(T_t) + \ln(I_t), \tag{14.3}$$

de manera que utilizar el esquema multiplicativo de la expresión (14.2) es equivalente a utilizar el esquema aditivo a la serie transformada en logaritmos, expresión (14.3).

Para estimar el componente tendencial utilizamos el método del promedio móvil centrado. Su expresión de cálculo, utilizando m valores, es

$$\hat{T}_t = \frac{1}{m} \sum_{i=-k}^{k} Y_{t+i}, \quad \text{donde} \quad m = 2k + 1. \tag{14.4}$$

Es decir, la estimación de la tendencia en el momento t la obtenemos promediando los valores de la serie de tiempo, utilizando k desfases hacia atrás y k valores adelante de su periodo central, t. Este promedio elimina en gran parte la aleatoriedad de la serie, dejando un componente de tendencia suavizado. Denominaremos al promedio móvil centrado de orden m como $m - MA$. El orden de la media móvil, m, determina la suavidad de la estimación. En general, un orden mayor implica una curva más suave.

Los promedios móviles simples son de orden impar. De esta forma conseguimos que sean simétricos en su punto medio t: en un promedio móvil de orden impar $m = 2k + 1$, hay k observaciones anteriores, y k posteriores a la observación que se promedia, t. Para realizar medias móviles centradas de orden par tenemos que aplicar una media móvil a la media móvil centrada.

Es muy usual utilizar medias móviles ponderadas y asignar distintas ponderaciones a los distintos desfases, su forma general es

$$\hat{T}_t = \sum_{i=-k}^{k} a_i Y_{t+i}, \tag{14.5}$$

con $k = (m-1)/2$ y ponderaciones o pesos dados por $(a_{-k}, ..., a_k)$, con suma unitaria $\left(\sum_{i=-k}^{k} a_k = 1\right)$ y simétricos, es decir, con $a_{-i} = a_i$. Algunas de las ponderaciones más ampliamente utilizadas las reproducimos en la Tabla 14.1.

Tabla 14.1: Ponderaciones más usuales. Medias móviles centradas

Nombre	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}
3-MA	0,333	0,333										
5 MA	0,200	0,200	0,200									
2x12-MA	0,083	0,083	0,083	0,083	0,083	0,083	0,042					
3x3-MA	0,333	0,222	0,111									
3x5-MA	0,200	0,200	0,133	0,067								
S15-MA	0,231	0,209	0,144	0,066	0,009	-0,016	-0,019	-0,009				
S21-MA	0,171	0,163	0,134	0,037	0,051	0,017	-0,006	-0,014	-0,014	-0,009	-0,003	
H5-MA	0,558	0,294	-0,073									
H9-MA	0,330	0,267	0,119	-0,010	-0,041							
H13-MA	0,240	0,214	0,147	0,066	0,000	-0,028	-0,019					
H23-MA	0,148	0,138	0,122	0,097	0,068	0,039	0,013	-0,005	-0,015	-0,016	-0,011	-0,004

S = Promedio móvil ponderado de Spencer
H = Promedio móvil ponderado de Henderson

En la descomposición clásica suponemos que el componente estacional es constante año a año. En consecuencia, elaboramos un índice estacional de m elementos (por ejemplo, $m = 4$ para datos trimestrales, $m = 12$ para datos mensuales y $m = 7$ para datos diarios).

Describimos a continuación cómo se realiza la descomposición clásica utilizando el esquema aditivo :

1. Si m es par, calculamos el componente de tendencia utilizando un $2 \times m - MA$ para obtener $\hat{T}_t$. Si m es impar, calculamos el componente de tendencia utilizando un $m - MA$.
2. Calculamos la serie sin tendencia: $Y_t - \hat{T}_t$.

3. Estimamos el componente estacional (trimestre, mes o día), promediando los valores sin tendencia de cada estación. Por ejemplo, el índice estacional de enero es el promedio de todos los valores de enero de la serie libre de tendencia. Estos índices estacionales los ajustamos posteriormente para garantizar que la suma de los m elementos sea nula. El componente estacional $\hat{S}_t$ lo obtenemos encadenando todos los índices estacionales para todos los años.
4. El componente residual lo calculamos restando de la serie los componentes estacional y de tendencia estimados anteriormente: $\hat{I}_t = Y_t - \hat{T}_t - \hat{S}_t$.

La descomposición clásica presenta graves problemas y por ello en la actualidad se utiliza de forma marginal. Algunos de ellos son los siguientes:

1. La estimación de tendencia elimina observaciones al principio y final de la serie. Por ejemplo, para $m = 12$, se pierden las seis primeras y últimas observaciones. En consecuencia, tampoco tenemos estimación del resto de componentes para esas observaciones.
2. El método de descomposición clásico asume que el componente estacional se repite año tras año. Para muchas series, esto es una suposición razonable, pero para algunas series largas no lo es. En ocasiones, los patrones estacionales van cambiando con el tiempo. Los métodos clásicos de descomposición no son capaces de capturar estos cambios estacionales en el tiempo.
3. A veces, algunos valores de la serie temporal pueden ser particularmente inusuales (por ejemplo, cuando ocurren conflictos laborales en las series de producción). El método clásico no es robusto a este tipo de valores inusuales.

Una forma de solventar estos problemas es utilizar los modelos actuales que hemos comentado al comienzo de esta sección. En particular nos centramos en el ajuste estacional basado en modelos ARIMA, que a nivel muy esquemático, consta de los siguientes pasos:

1. Se construye un modelo ARIMA para la serie a ajustar;
2. A partir del modelo se obtienen otros modelos para los componentes;
3. Se utiliza el filtro de Wiener-Kolmogorov para separar dichos componentes y, por último,
4. Se vuelven a agregar, pero excluyendo el componente estacional.

Los métodos basados en filtros fijos permiten descomponer la serie en componentes no observables mediante un procedimiento iterativo basado en alisados secuenciales. Este alisado se obtiene aplicando medias móviles. El INE utiliza el primero de los dos métodos, y por este motivo nos centramos en el mismo.

Este método considera que cada componente está gobernado por un modelo ARIMA que refleja sus principales propiedades teóricas; debiendo ser dichos modelos compatibles, en su conjunto, con el que caracteriza a la serie agregada x_t.

En general, si hay k componentes, el modelo consistirá en un conjunto de ecuaciones de

la forma

$$x_t = x_{1t} + ... + x_{kt}$$
$$\phi_i(B)x_{it} = \theta_i(B)a_{it}, i = 1, ..., k$$

donde $\phi_i(B)$ y $\theta_i(B)$ son polinomios finitos en B de órdenes p_i y q_i, sin raíces comunes y todas ellas fuera del círculo unidad, y la variable a_{it} is ruido blanco $(0, \sigma_i^2)$. Obsérvese que cada componente evolucionará según un modelo ARIMA de la forma

$$x_{it} = \frac{\theta_i(B)}{\phi_i(B)}a_{it} = \Psi_i(B)a_{it}, i = 1, ..., k$$

La agregación de modelos ARIMA genera un modelo ARIMA, y por tanto el agragado x_t seguirá también un modelo de la forma

$$\phi(B)x_t = \theta(B)a_t$$

donde la variable a_t is ruido blanco $(0, \sigma_a^2)$. Por lo tanto x_t sería compatible con la expresión utilizada anteriormente

$$x_t = \Psi(B)a_t = \frac{\theta(B)}{\phi(B)}a_t = \frac{\theta_q(B)\theta_Q(B^s)}{\phi_p(B)\phi_P(B)(1-B)^d(1-B^s)^D}a_t$$

Obsérvese que los componentes individuales deben ser compatibles con el agregado, por lo tanto

$$\frac{\theta(B)}{\phi(B)}a_t = \sum_{k=1}^{k}\frac{\theta_i(B)}{\phi_i(B)}a_{it}$$

Esta compatiblidad implica entonces que el polinomio $\phi(B)$ podría contener raíces unitarias, en particular este polinomio de la parte AR se puede factorizar como sigue

$$\phi(B) = \phi_1(B)\phi_2(B)...\phi_k(B) \tag{14.6}$$

Por otra parte, la parte MA se puede obtener de la relación siguiente

$$\theta(B)a_t = \sum_{k=1}^{k}\phi_{(i)}(B)\theta_i(B)a_{it} \tag{14.7}$$

donde $\phi_{(i)}(B) = \prod_{j=1,j\neq i}^{k}\phi_j(B)$, es decir, el producto de todos los $\phi_j(B), j = 1, ..., k$ sin incluir $\phi_i(B)$.

Las dos últimas expresiones (ecuaciones) son fundamentales para el desarrollo del procedimiento, ya que relacionan los operadores ARMA de la forma reducida de x_t con los correspondientes operadores de los componentes inobservables. Los primeros han sido estimados y los segundos pueden ser derivados a partir de éstos. Desafortunadamente, estas dos ecuaciones están sujetas al siguiente problema de identificación: existen potencialmente infinitas estructuras $\phi_i(B)$ compatibles con el modelo en forma reducida $\phi(B)$ que gobierna a x_t. La solución de este problema requiere la incorporación

de información adicional que solucione esta indeterminación. La metodología basada en modelos invoca al principio de descomposición canónica para alcanzar la identificación del sistema. Este principio establece que la descomposición adicional de cada componente como señal más ruido blanco es imposible, esto es, que el componente carece de información redundante: es señal pura o ruido blanco, sin mezcla posible. En términos formales este principio se materializa como sigue:

La descomposición

$$x_{it} = x_{it}^s + \xi_{it}, \xi_{it} \sim iid(0, \sigma_i)$$

implica $x_{it} = x_{it}^s$ (solo existe una señal) o bien $x_{it} = \xi_{it}$ (solo existe ruido).

Una de las consecuencias del principio de descomposición canónica es que los operadores MA de los modelos de los componentes no son invertibles, ya que poseen al menos una raíz sobre el círculo de radio unitario, lo que obliga a acomodar el análisis econométrico de los componentes estimados a este hecho.

Una vez aplicado el principio de descomposición canónica, las ecuaciones (14.6) y (14.7) permiten la determinación de los valores de $\phi_i(B)$ en función de los de $\phi(B)$ mediante, por ejemplo, el método de los momentos.

Hay dos técnicas o procedimientos para la especificación del los componentes y del modelo ARIMA general. Uno de ellos, conocido con enfoque estructural, especifica modelos para cada componente y luego lo unifica. Esto hace que los modelos ARIMA de los componentes tiendan a ser parsimoniosos, mientras que el modelo completo ARIMA no lo es. El otro procedimiento, que generalmente usa el INE, consiste en identificar el modelo ARIMA para la serie x_t, y derivar sus componentes de la estructura identificada. Este procedimiento hace que el modelo ARIMA de la serie x_t sea parsimonioso, mientras que no lo son los modelos ARIMA de los componentes.

Así pues siguiendo los procesos INE, una vez definidos los modelos teóricos para los componentes, estos han de ser estimados, esto es, hemos obtener series temporales para cada x_{it} a partir de los datos observados de x_t. Este proceso se realiza mediante el filtrado de x_t según:

$$\hat{x}_t = V_i(B, F)x_t$$

donde los filtros $V_i(B, F)$, $F = B^{-1}$, (F es el operador "adelante") pertenecen a la familia de Wiener-Kolmogorov, y pretenden minimizar el error cuadrático medio entre el estimador y el componente teórico. De esta forma, estos filtros se obtienen como solución del siguiente programa de optimización restringida:

$$\min_{\hat{x}_{it}} \mathbb{E}(x_{it} - \hat{x}_{it})^2$$

$$\text{sujeto a } x_{it} = \Psi_i(B)a_{it}$$

cuya solución es

$$\hat{x}_{it} = \frac{v_i}{v_a}\frac{\Psi_i(B)\Psi_i(F)}{\Psi(B)\Psi(F)}x_t = \kappa_i\pi(B)\pi(F)\phi_i(B)\phi_i(F)x_t$$

donde $\kappa_i = \sigma_i^2/\sigma_a^2$.

La expresión anterior representa la solución de filtrado adoptada por el enfoque basado en modelos expresados en forma reducida. Como características generales de este tipo de filtros se puede destacar que se trata de filtros lineales, simétricos, invariantes en el tiempo, con colas infinitas pero convergentes y que se derivan combinando la información suministrada por la forma reducida, $\pi(B)$, y la postulada para los componentes, $\phi_i(B)$.

A modo ilustrativo, si consideramos por ejemplo que tenemos una serie trimestrales, $s = 4$, tenemos

$$x_t = \frac{(1-\theta_1 B)(1-\theta_4 B^4)}{(1-B)(1-B^4)} a_t; -1 < \theta_1 < 1, 0 < \theta_4 < 1$$

Serie que sobre la que vamos a realizar una descomposición canónica en tres componentes

- Tendencia-Ciclo (P)
- Estacional (S)
- Irregular (I)

De modo que

$$x_t = P_t + S_t + I_t,$$

entonces, según la metodología descrita, los componentes canónicos sería

$$P_t = \frac{(1+B)(1-\alpha B)}{(1-B)^2} a_{pt}, a_{pt} \sim \text{iid } N(0, v_p)$$

$$S_t = \frac{(1-B)(1-\delta_1 B - \delta_2 B^2)}{(1+B+B^2+B^3)} a_{st}, a_{st} \sim \text{iid } N(0, v_s)$$

$$I_t = a_{It}, a_{It} \sim \text{iid } N(0, v_I).$$

Los correspondientes filtros de Wiener-Kolmogorov que permiten estimar los componentes a partir de la muestra son:

$$\hat{P}_t = \frac{v_p}{v_a} \frac{U(B)(1+B)(1-\alpha B)}{(1-\theta_1 B)(1-\theta_4 B^4)} \frac{U(F)(1+F)(1-\alpha F)}{(1-\theta_1 F)(1-\theta_4 F^4)} x_t$$

$$\hat{S}_t = \frac{v_s}{v_a} \frac{(1-B)^3(1-\delta_1 B - \delta_2 B^2)}{(1-\theta_1 B)(1-\theta_4 B^4)} \frac{(1-F)^3(1-\delta_1 F - \delta_2 F^2)}{(1-\theta_1 F)(1-\theta_4 F^4)} x_t,$$

$$\hat{I}_t = \frac{v_I}{v_a} \frac{(1-B)^2 U(B)}{(1-\theta_1 B)(1-\theta_4 B^4)} \frac{(1-F)^2 U(F)}{(1-\theta_1 F)(1-\theta_4 F^4)} x_t$$

donde

$$U(B) = (1+B+B^2+B^3) = (1+B)(1+B^2).$$

Estos filtros conducen a los siguientes modelos

$$\widehat{P}_t = \frac{(1+B)(1-\alpha B)}{(1-B)^2} V_p(F) a_{pt} = \phi_p(B) V_p(F) a_{pt}$$

$$\widehat{S}_t = \frac{(1-B)(1-\delta_1 B - \delta_2 B^2)}{(1+B+B^2+B^3)} V_S(F) a_{st} = \phi_s(B) V_s(F) a_{st}$$

$$\widehat{I}_t = V_I(F) a_{It}$$

Una vez descompuesta la serie observada, es inmediato obtener la serie corregida de efectos calendario y de la estacionalidad

$$\hat{y}^{destacionalizada} = y_t - \hat{\gamma} P_t - \hat{\delta} D_t - \hat{S}_t, t = 1, ..., T$$

igualmente la serie ciclo-tendencia será el resultado de aplicar el filtro del componente (tendencia-ciclo) a la serie sin efecto calendario

$$\hat{y}^{ciclo-tendencia} = V_p(B, F)(y_t - \hat{\gamma} P_t - \hat{\delta} D_t), t = 1, ..., T.$$

Un método alternativo al presentado hasta ahora, esto es, el procedimiento TRAMO-SEATS, que es el utilizado en numerosas instituciones públicas relativas a las estadísticas europeas, es el denominado X13-ARIMA o sus versiones precedentes.

Se trata de uno de los métodos de descomposición más usuales para datos trimestrales y mensuales, acutalmente se conoce como $X-13-ARIMA$, tiene sus orígenes en los métodos desarrollados por la US Bureau of the Census. Este método es ampliamente utilizado por instituciones de todo el mundo. Las versiones anteriores fueron el $X-11$, $X-11-ARIMA$ y $X-13-ARIMA$.

El método se basa en la descomposición clásica. En particular, la estimación de tendencia incluye todas las observaciones de la serie, también permite que el componente estacional pueda variar lentamente con el tiempo, y cabe destacar que es relativamente robusto a las observaciones inusuales. Utiliza tanto efectos aditivos como multiplicativos, pero solo permite datos trimestrales y mensuales.

La referencia a la parte ARIMA indica la utilización de un modelo $ARIMA$ para proporcionar previsiones de la serie hacia adelante y hacia atrás en el tiempo. En consecuencia, cuando se aplica una media móvil para obtener la estimación de la tendencia, no hay pérdida de observaciones al comienzo y final de la serie, algo de lo que las descomposiciones clásicas adolecían.

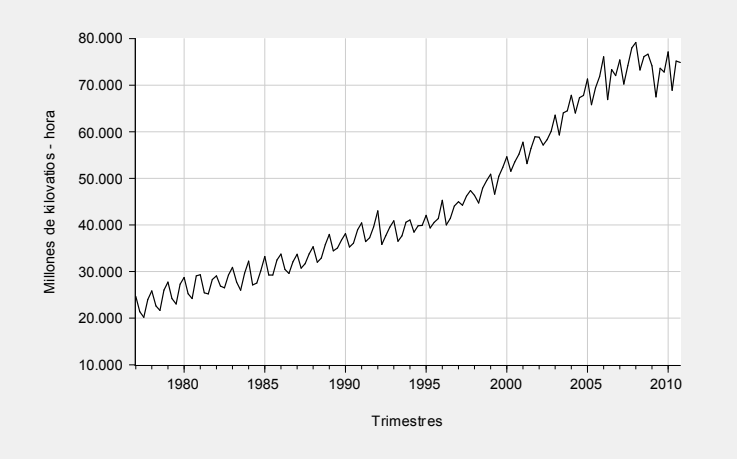
En el siguiente epígrafe ilustraremos el uso de ambos métodos, y abordaremos el problema práctico de descomponer el componente ciclo-tendencia en dos partes.

14.4 Problemas prácticos

Ilustramos inicialmente los problemas habituales a partir de un caso práctico, en concreto descompondremos la serie de producción eléctrica de España para el periodo 1977 a

2010. En particular, la producción eléctrica española entre el primer trimestre de 1977 y el último de 2010, en millones de kilovatios hora, se reproduce en el gráfico de la Figura 14.1. Observamos una tendencia creciente hasta la crisis financiera y una fuerte estacionalidad con máximos en el primer trimestre del año y mínimos en el segundo. Descompodremos la serie utilizando los métodos clásico, *X12-ARIMA* y *TRAMO-SEAT*, y finalmente realizaremos previsiones para los años 2011-13.

Figura 14.1: Producción de electricidad en España entre 1977 y 2010



En las Figuras 14.2, 14.3 y 14.4 mostramos los gráficos de los distintos componentes de la serie utilizando los tres métodos de descomposición vistos, descomposición clásica, *X12-ARIMA* y *TRAMO-SEATS*.

Figura 14.2: Producción de electricidad, descomposición clásica

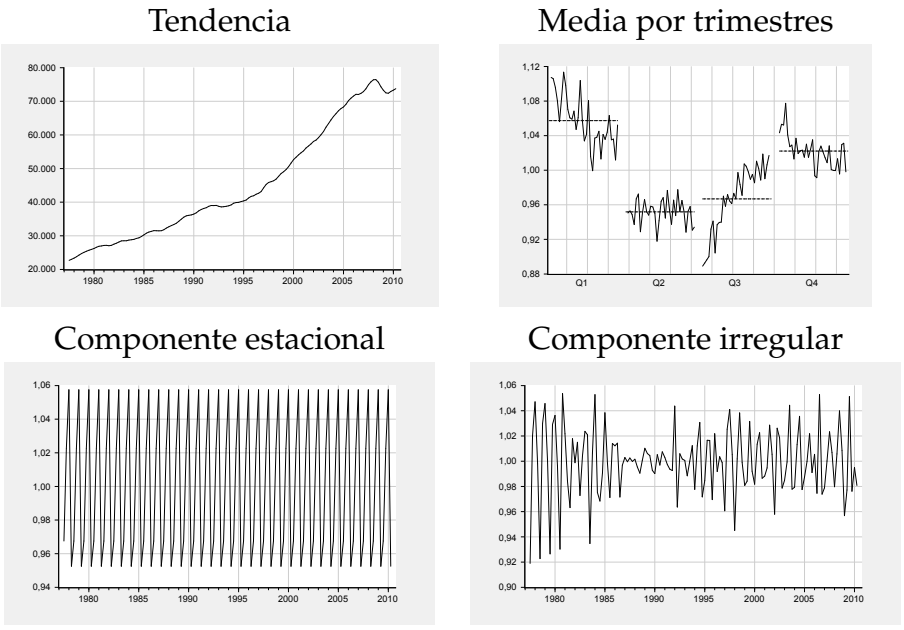


Figura 14.3: Producción de electricidad, descomposición *X12-ARIMA*

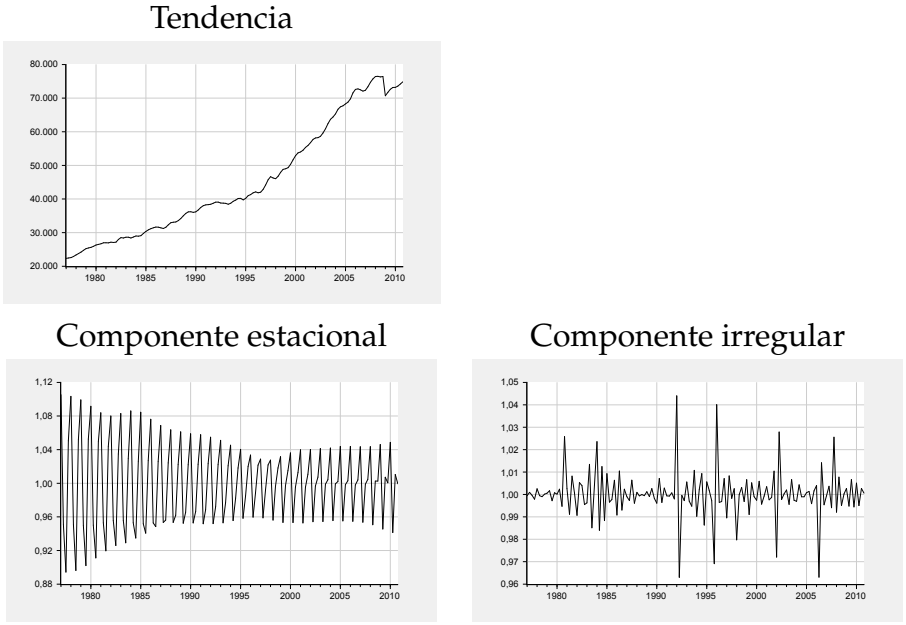
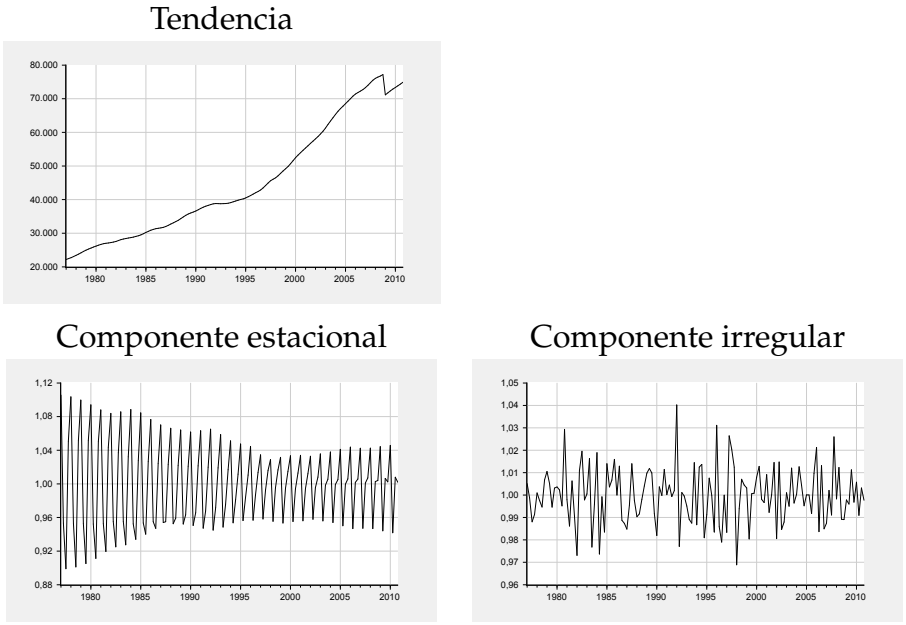


Figura 14.4: Producción de electricidad, descomposición *TRAMO-SEATS*



Con los tres métodos obtenemos tendencias similares, pero en la descomposición clásica perdemos la información de las dos primeras y últimas observaciones.

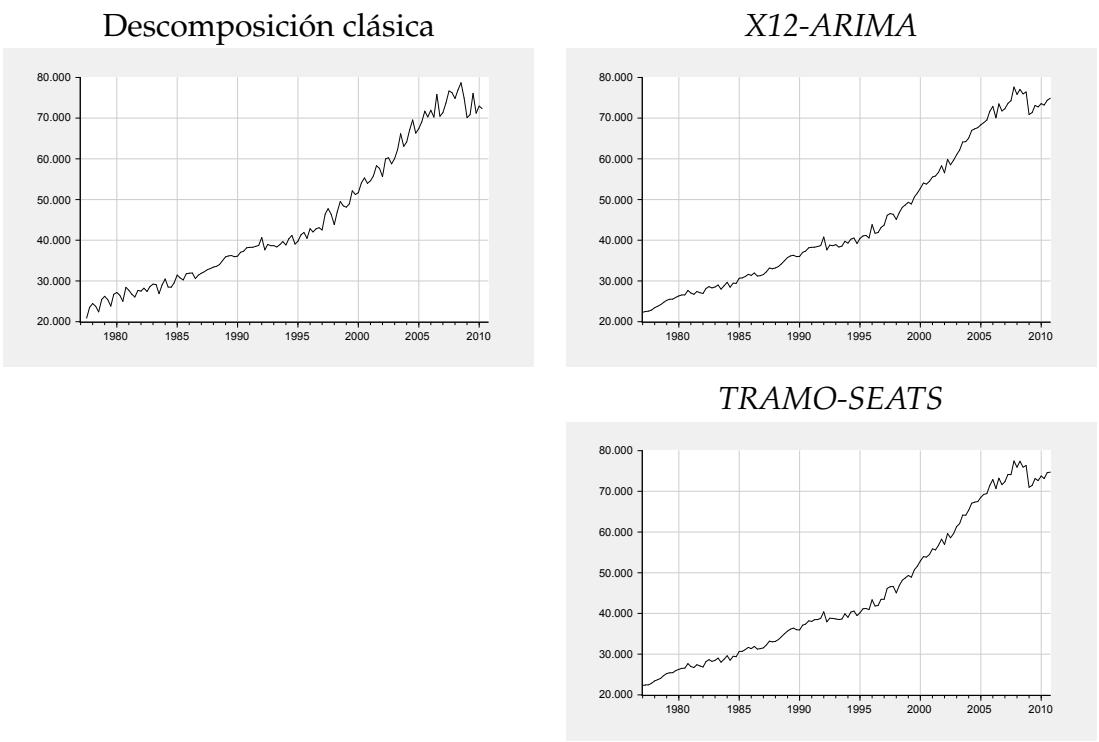
El componente estacional, sin embargo, muestra una clara diferencia entre la descomposición clásica y los otros métodos; el método clásico genera un patrón estacional constante para todo el periodo; el *X12-ARIMA* y *TRAMO-SEATS*, además de calcular

el componente estacional sin pérdida de información, muestra una reducción del componente hasta la segunda mitad del quinquenio 1995-2000 y un moderado crecimiento posterior.

En definitiva, la descomposición clásica presenta los problemas como la reducción de información en los extremos de los componentes y estacionalidad constante a lo largo de toda la serie. Los procedimientos *X12-ARIMA* y *TRAMO-SEATS* muestran, en general, descomposiciones similares y adecuadas a las características de las series observadas, y por ello son los métodos más utilizados para descomponer series de tiempo. Además el método *TRAMO-SEATS* presenta la ventaja de establecer de forma clara el proceso generador de los datos de la serie al estimar el modelo *ARIMA* subyacente de forma automática, lo que sin duda es una información relevante para el usuario no experto y un buen punto de partida para los más experimentados.

Multiplicando los componentes de tendencia y residual obtenemos las respectivas series desestacionalizadas, $\hat{A}_t = \hat{T}_t \cdot \hat{E}_t$. Series que utilizamos para predecir, excepto en el caso *TRAMO-SEATS* cuya predicción la realizaremos directamente sobre el modelo *ARIMA* subyacente de la serie original. En la Figura 14.5 se observan los gráficos de las tres series desestacionalizadas.

Figura 14.5: Electricidad, series desestacionalizadas



Observamos dos diferencias fundamentales. Por un lado a la serie clásica le faltan los dos últimos trimestres de 2010 (y los dos primeros de 1977), esto no tiene solución y tendremos que predecir a partir del tercer trimestre de 2010. Por otro, si nos fijamos en la serie clásica vemos que el componente estacional no ha desaparecido del todo; la razón se encuentra en que con este método el componente estacional se supone constante, y

como apreciamos claramente, en este caso este supuesto carece de fundamento. Estas son dos de las razones, sin duda importantes, por las que el método de descomposición clásico prácticamente se ha dejado de utilizar. Las series desestacionalizadas con los métodos *X12-ARIMA* y *TRAMO-SEATS* son muy similares.

Otro aspecto muy práctico y útil es que los métodos de descomposición son principalmente interesantes para el estudio de las series de tiempo, y el análisis de los cambios históricos, pero también se utilizan para pronosticar o predecir.

Para realizar los pronósticos de las series utilizamos modelos *ARIMA*. En el caso de la descomposición clásica, y por las razones aludidas en el párrafo anterior, permitiremos la incorporación de componentes estacionales, lo que solucionará, al menos en parte, el problema de considerar la estacionalidad constante en todo el periodo. La estimación para la serie desestacionalizada por el método clásico es

$$\triangle \ln\left(elec_t^{DC}\right) = \underset{\substack{(0,0021)\\ ***}}{0,0079} + \underset{\substack{(0,0709)\\ ***}}{0,5803}\triangle \ln\left(elec_{t-4}^{DC}\right) - \underset{\substack{(0,0681)\\ ***}}{0,6507}\widehat{\varepsilon}_{t-1} + \widehat{\varepsilon}_t, \tag{14.8}$$

donde $elec_t^{DC}$ es la serie desestacionalizada por el método clásico. El modelo presenta un fuerte componente estacional [$AR(4)$], lo que prueba que el componente estacional sigue presente en la serie presuntamente desestacionalizada. Todas las variables son significativas, incluso al 99 % de confianza. El correlograma de los residuos muestra la imagen empírica de un proceso puramente aleatorio (ruido blanco), de manera que damos el modelo por validado. El modelo para el método *X12-ARIMA* es

$$\triangle \ln\left(elec_t^{DX12}\right) = \underset{\substack{(0,0013)\\ ***}}{0,0090} - \underset{\substack{(0,0804)\\ ***}}{0,3827}\triangle \ln\left(elec_{t-1}^{DX12}\right) + \widehat{\varepsilon}_t, \tag{14.9}$$

donde $elec_t^{DX12}$ es la serie desestacionalizada por el método *X12-ARIMA*. El modelo presenta solo componentes regulares, de manera que el método *X12-ARIMA* parece eliminar correctamente el patrón estacional. Todos los parámetros son significativos, incluso al 99 % de confianza, y los residuos se comportan como ruido blanco, por consiguiente el modelo lo damos por correcto.

Utilizamos los modelos estimados para ambas series desestacionalizadas para predecir sus valores en el periodo 2011 - 2013 (en el caso de descomposición clásica también para los dos últimos trimestres de 2010).

En el caso del método *TRAMO-SEATS* el pronóstico lo realizaremos a partir del modelo subyacente $SARIMA(p,d,q)(P,D,Q)_s$ identificado por el algoritmo *TRAMO* (es decir, en este caso la previsión no se hace usualmente sobre la serie desestacionalizada sino sobre la serie original). El método presenta la ventaja para el usuario inexperto de que el propio algoritmo nos indica de forma automática el modelo subyacente, en este caso un $SARIMA(0,1,1)(0,1,1)_4$, cuya estimación es

$$\triangle\triangle^4 \ln\left(elec_t\right) = \underset{\substack{(0,0561)\\ ***}}{-0,4442}\widehat{\varepsilon}_{t-1} - \underset{\substack{(0,0569)\\ ***}}{0,5455}\widehat{\varepsilon}_{t-4} + \widehat{\varepsilon}_t, \tag{14.10}$$

donde $elec_t$ es la serie de producción de energía eléctrica original, de manera que el pronóstico con *TRAMO-SEATS*, se reduce a la estimación del modelo *ARIMA*. Todos

los parámetros son significativos, incluso al 99 % de confianza, y los errores estimados muestran un correlograma compatible con ruido blanco.

Para la previsión del componente estacional de los métodos clásico y *X12-ARIMA*, repetimos el patrón estacional del último año estimado. Es decir, utilizamos los valores del componente estacional del año 2010 para la predicción de los años 2011 a 2013.

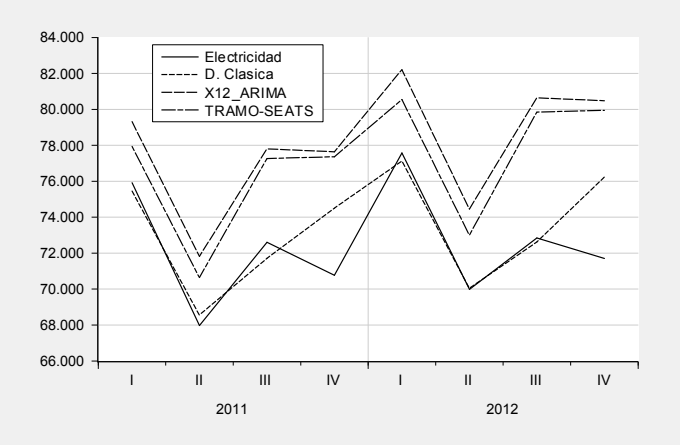
La Tabla 14.2 muestra las previsiones por los tres métodos así como los componentes estacionales y las series desestacionalizadas mediante la descomposición clásica y *X12-ARIMA*. También mostramos la producción de electricidad observada en esos años y en la última fila indicamos el porcentaje del error medio absoluto como medida de bondad del ajuste de la previsión realizada.

Tabla 14.2: Producción de electricidad. Previsión 2011-13

obs.	***Descomposición clásica***			***X12 - ARIMA***			***TRAMO-SEATS***	***Original***
	Deses.	C. Estacional	Previsión	Deses.	C. Estacional	Previsión		
2011Q1	71.355,24	1,0575	75.458,16	75.640,27	1,0486	79.319,94	77.941,22	75.914,80
2011Q2	71.975,44	0,9525	68.556,61	76.299,15	0,9414	71.824,45	70.646,68	67.968,40
2011Q3	74.122,88	0,9675	71.713,88	76.996,73	1,0104	77.796,88	77.258,40	72.600,70
2011Q4	72.861,23	1,0225	74.500,61	77.687,96	0,9993	77.637,01	77.357,49	70.767,60
2012Q1	72.933,95	1,0575	77.127,66	78.390,31	1,0486	82.203,75	80.547,85	77.575,40
2012Q2	73.545,12	0,9525	70.051,73	79.097,11	0,9414	74.458,31	73.009,35	69.987,10
2012Q3	75.059,56	0,9675	72.620,12	79.811,02	1,0104	80.640,40	79.842,19	72.843,60
2012Q4	74.562,89	1,0225	76.240,55	80.531,08	0,9993	80.478,26	79.944,60	71.704,80
2013Q1	74.854,39	1,0575	79.158,52	81.257,75	1,0486	85.210,69	83.241,66	73.143,00
2013Q2	75.468,10	0,9525	71.883,37	81.990,94	0,9414	77.182,43	75.451,05	65.604,40
2013Q3	76.620,20	0,9675	74.130,04	82.730,76	1,0104	83.590,49	82.512,40	71.305,90
2013Q4	76.579,64	1,0225	78.302,68	83.477,25	0,9993	83.422,50	82.618,23	70.781,70
PEMA			0,0397			0,0744	0,0588	

Las previsiones definitivas las calculamos multiplicando el componente estacional y la previsión de la serie desestacionalizada para los métodos clásico y *X12-ARIMA*. Para *TRAMO-SEATS* utilizamos directamente los resultados de la expresión (14.10). La Figura 14.6 muestra el gráfico de las predicciones junto con la serie original.

Figura 14.6: Previsión electricidad. 2011 - 2013



El pronóstico del método clásico es mejor (con error medio del 0,04 %), la razón se encuentra en que al eliminar las dos últimas observaciones de la serie desestacionalizada, gráfico derecho de la Figura 14.5, la serie no muestra claramente la incipiente recuperación posterior a la crisis de 2008 y lógicamente la proyección del modelo *ARIMA* predice una recuperación menor que en los casos de descomposición *X12-ARIMA* y *TRAMO-SEATS* donde sí se ve claramente la incipiente recuperación posterior a la crisis, ver la Figura 14.5. En consecuencia la descomposición *X12-ARIMA* y *TRAMO-SEATS* proyectan una mayor recuperación, recuperación que finalmente no se consolidó en los años siguientes. En todo caso, el gráfico también muestra que la previsión del componente estacional de la serie se reproduce mejor en las series obtenidas mediante el método *X12-ARIMA* y *TRAMO-SEATS*. La comparativa entre *X12-ARIMA* y *TRAMO-SEATS* muestra series muy similares con movimientos prácticamente paralelos; el pronóstico es mejor en el caso del método *TRAMO-SEATS* (con un error medio absoluto del 0,059 %), es decir, aquel en cuya previsión no hemos utilizado la descomposición por componentes. El último aspecto práctico que contemplamos es la descomposición del componente ciclo-tendencia en sus dos elementos naturales. Hasta ahora hemos tratado los componentes de tendencia T y ciclo C de forma conjunta, esto es así en todo el tema excepto en este apartado, donde analizaremos el filtro de Hodrick y Prescott cuya utilidad es precisamente, separar el ciclo de la tendencia.

El mayor problema para separar el ciclo de la tendencia se encuentra en la propia definición de ciclo económico, no exenta de subjetividad, y por ello lo normal es mantener ambos componentes conjuntamente, denominándolos como ciclo-tendencia o simplemente tendencia, como hacemos nosotros en el resto del tema.

En ocasiones el análisis económico, especialmente el macroeconómico, requiere la utilización de alguno de los dos componentes separadamente, el ciclo o la tendencia, y por ello dedicamos este epígrafe a analizar el filtro más extendido para la descomposición de series temporales en tendencia y ciclo.

Aunque no hay un consenso definitivo sobre la definición del ciclo económico, el NBER, a partir de Burns y Mitchel, define el ciclo económico de la siguiente manera:

> «*Los ciclos económicos son un tipo de fluctuaciones encontradas en la actividad económica agregada de las naciones que organizan su funcionamiento en empresas comerciales. Un ciclo consiste en expansiones que ocurren aproximadamente al mismo tiempo en muchas actividades económicas, seguidas generalmente de recesiones, contracciones y reactivaciones que se conectan con la fase de expansión del ciclo siguiente, esta secuencia es recurrente pero no periódica, la duración de los ciclos económicos varía entre más de un año y hasta diez o doce años, no son divisibles en ciclos más cortos.*

El filtro de Hodrick-Prescott (HP) es el más extendido en la literatura para separar los componentes de ciclo y tendencia. Su empleo se justifica por su linealidad, por estar bien definido independientemente de la serie a la que se aplica, exento de juicios subjetivos y fácil de replicar. La metodología de HP consiste en el filtrado del logaritmo de la serie extrayendo la tendencia y adquiriendo el componente cíclico, como la diferencia entre la serie y su componente permanente o tendencia. Para lograr tal separación, Hodrick y Prescott propusieron como medida de la variabilidad de la tendencia, la suma de los

cuadrados de sus segundas diferencias, con el fin de minimizar la variabilidad de la tasa de crecimiento del componente permanente. El filtro parte de la idea de que cualquier serie de tiempo, Y_t, en logaritmos y sin componente estacional, se puede escribir como la suma de la tendencia, T_t, y el ciclo, C_t. Es decir, que

$$Y_t = T_t + C_t, \qquad t = 1, \ldots, T. \tag{14.11}$$

Motivados por el criterio de variabilidad, Hodrick y Prescott propusieron el siguiente problema de minimización para encontrar la tendencia de una serie.

$$\text{mín}_{T_t} \sum_{t=1}^{T} (Y_t - T_t)^2 + \lambda \sum_{t=2}^{T-1} \left[(T_{t+1} - T_t) - (T_t - T_{t-1})\right]^2, \qquad t = 1, 2, ..., T \tag{14.12}$$

donde el primer componente de la ecuación (14.12) corresponde a las diferencias entre la serie original y la tendencia (es decir, el ciclo) elevada al cuadrado mientras el segundo componente, que se multiplica por λ, es la medida de suavizado de la serie, elevando al cuadrado la aceleración de la tendencia. A este respecto, λ es un número predeterminado, conocido como parámetro de suavización, cuya función principal en el problema de minimización es penalizar la suma de las segundas diferencias del componente permanente. Cuanto menor sea el parámetro, el componente permanente puede fluctuar más, y cuanto mayor sea éste, más se penalizan las fluctuaciones de la tendencia. Por lo tanto, cuanto mayor sea λ, la tendencia debe ser más suave. Cuando $\lambda \to \infty$, la tendencia se aproxima a su forma determinista $T_t = T_0 + \alpha t$, para una constante positiva α. Esta situación corresponde al caso en que la tendencia crece a una tasa constante (tendencia exponencial), en concordancia con la teoría neoclásica. Cuando $\lambda = 0$, no se penalizan las variaciones y por lo tanto la tendencia es la misma serie. Tal y como apuntan Hodrick y Prescott, si se cumple que:

$$\begin{aligned} &C_t \sim N(0, \sigma_1^2) \\ &(T_{t+1} - T_t) - (T_t - T_{t-1}) \sim N(0, \sigma_2^2). \end{aligned} \tag{14.13}$$

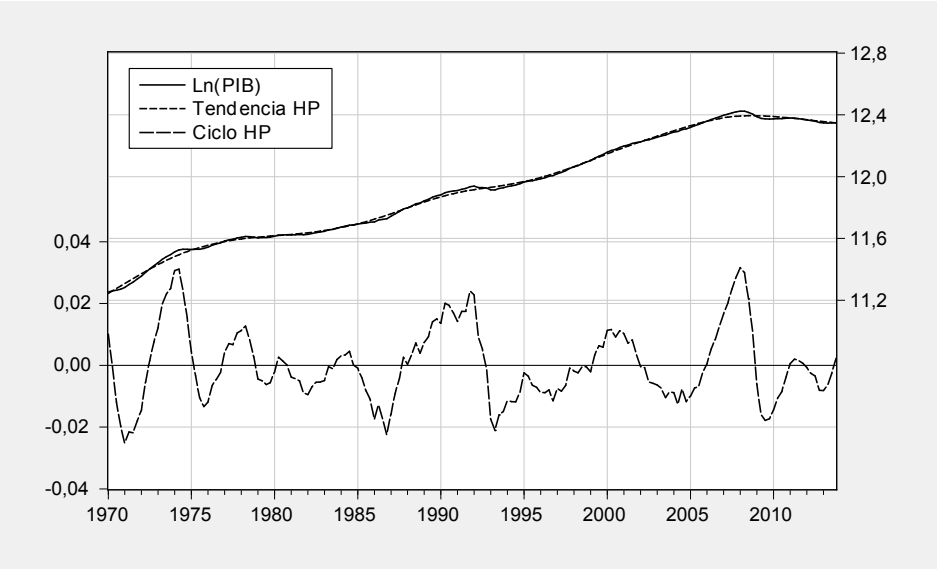
Es decir, si ciclo y segunda diferencia de la tendencia fueran variables normales e independientemente distribuidas, la solución al problema (14.12) correspondería al valor esperado de T_t dadas las observaciones, si y solo si $\lambda = (\sigma_1^2)/(\sigma_2^2)$. Sin embargo normalmente el ciclo y la segunda diferencia de la tendencia no se comportan de esta forma y el valor de λ es, sin duda, la cuestión fundamental a la hora de aplicar el filtro *HP*.

Hay dos aspectos más a tomar en cuenta a la hora de aplicar el filtro. En primer lugar, el filtro *HP* debe ser aplicado a series desestacionalizadas para que el componente cíclico no sea contaminado con variaciones estacionales. En segundo lugar, la tendencia estimada adolece de ser muy sensible a shocks transitorios al final de la muestra. Para aminorar este inconveniente se recomienda hacer proyecciones de uno o dos años a partir de la serie original antes de calcular la tendencia aplicando el filtro *HP*.

Veamos ahora un ejemplo utilizando la serie del PIB español en millones de euros constantes de 2005 (serie desestacionalizada) entre el primer trimestre de 1970 y el

último de 2013. La descomposición de la serie en logaritmos, en tendencia y ciclo utilizando el filtro HP, con parámetro $\lambda = 1600$, la reproducimos en el gráfico de la Figura 14.7.

Figura 14.7: PIB trimestral desestacionalizado, millones de euros de 2005



Donde podemos observar el ciclo en la parte inferior del gráfico.

14.5 Desagregación temporal

Los datos de baja frecuencia (anuales) suelen ser detallados y de alta precisión, pero no son muy "oportunos" en el sentido de que tenemos que esperar bastante para obtenerlos y esto limita en ocaciones la toma de decisiones, los análisis basados en datos y el diseños e implementación de políticas. Por otro lado, los datos de alta frecuencia (subanuales) son menos detallados y precisos, pero más oportunos. De hecho, la producción de datos de alta frecuencia con el mismo nivel de detalle y precisión requeriría normalmente más recursos e impondría una carga de respuesta mayor para las empresas y las personas.

A nivel internacional, el problema se ve agravado por el hecho de que diferentes países producen datos con diferentes frecuencias, lo que complica las comparaciones. En algunos casos, diferentes países producen con la misma frecuencia pero con diferentes horarios, por ejemplo, cada cinco años pero no en los mismos años. Además, en algunos casos, los datos están espaciados irregularmente en el tiempo o muestran lagunas.

Los análisis socioeconómicos y ambientales actuales requieren un historial ininterrumpido de datos frecuentes y recientes sobre las variables de interés. Estudios relacionados con fenómenos a largo plazo, pensemos en los datos de efecto invernadero o los datos relativo a crecimiento de la población necesitan series muy largas. Por el contrario, los estudios relacionados con industrias de vanguardia de rápido crecimiento, como la

bioingeniería y las nanotecnologías, requieren datos cercanos al tiempo real, pero no precisa que se extiendan más allá de dos décadas.

Los investigadores y académicos se enfrentan a menudo a situaciones de datos con distintas frecuencias y con lagunas temporales. Esto complica o impide el desarrollo de modelos trimestrales (digamos) que proporcionan predicciones frecuentes y a tiempo. El benchmarking y la desagregación temporal abordan estos problemas mediante la estimación de valores de alta frecuencia a partir de datos de baja frecuencia y datos relacionados de alta frecuencia.

El problema de la **desagregación temporal**, también conocido como distribución temporal, suele estar asociado con las series de flujo, donde los datos anuales de baja frecuencia corresponden a las sumas anuales de los correspondientes datos de mayor frecuencia. La desagregación temporal intenta obtener estimaciones de alta frecuencia de una serie temporal que tan solo es observada en baja frecuencia y, a veces, contando también con información de alta frecuencia de algún indicador que tenga relación con la serie de baja frecuencia. La distribución temporal es una práctica estándar en el sistema de Cuentas Nacionales, debido al costo excesivo de la recolección frecuente de datos. Por tanto, los datos de alta frecuencia se obtienen de la distribución temporal de los datos anuales. Este proceso generalmente involucra indicadores que están disponibles en una base de alta frecuencia y se considera que se comportan como la variable objetivo.

Formalmente el problema de la desagregación temporal es el siguiente. Sea $Y = Y_T : T = 1..N$ la serie anual observada y $\mathbf{x} = x_{i,t,T} : i = 1...p, t = 1..,4, T = 1...N$ una matriz nxp cuyas filas recogen las n observaciones disponibles sobre p indicadores de frecuencia trimestral, siendo $p \geq 1$ y $n = 4N$. El problema de la desagregación temporal consiste en estimar una serie $\mathbf{y} = y_{tT} : t = 1.,4, T = 1..N$ que satisfaga la restricción temporal asociada a que la suma de los cuatro trimestres pertenecientes a un mismo año coincida con el total anual correspondiente:

$$\sum_{t=1}^{4} Y_{tT} = Y_T$$

donde la restricción longitudinal puede expresarse matricialmente como sigue

$$\mathbf{By} = \mathbf{Y}$$

donde

$$B = I_N \otimes f = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & ... & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & ... & 0 & 0 & 0 & 0 \\ ... & & ... & & & ... & & ... & ... & & & & ... \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & ... & 1 & 1 & 1 & 1 \end{bmatrix}$$

Donde hemo considerado $f = [1111]$, si bien puede tomar otras estructuras.

La relación funcional entre y e Y puede estar condicionada por la información contenida en los indicadores trimestrales x. En este caso se tienen los llamados "métodos de desagregación temporal basados en indicadores". Dentro de éstos existen dos enfoques principales: métodos de ajuste y métodos basados en modelos. Los primeros consideran

la estimación de y como la solución de un programa de optimización restringida mientras que los segundos plantean dicha estimación como un problema inferencial: dada la estructura del modelo, derivar estimadores lineales, insesgados y de varianza mínima (ELIO), que permitan obtener y en función de Y y de x, verificando al mismo tiempo la restricción longitu dinal.

Frecuentemente, el analista de series temporales tiene a su disposición uno o varios indicadores de alta frecuencia que están relacionados con la serie de baja frecuencia que se desea desagregar temporalmente. En consecuencia, la incorporación de la información contenida en dichos indicadores al proceso de desagregación mejorará su calidad ya que, por un lad proporciona una referencia explícita de evolución intraanual a la que debe ajustarse la serie trimestralizada; y por otro permite incluir elementos de alta frecuencia tales como estacionalidad, efectos de calendario, etc. que están ausentes de la serie anual debido a su frecuencia de muestreo.

Como hemos dicho hay dos métodos (ajuste y modelos) y la distinción entre ambos no es entendida como una barrera infranqueable. Ambos enfoques han de realizar hipótesis relativamente fuertes acerca de la serie trimestral inobservable. Los primeros lo hacen indirectamente al plantear qué medida de volatilidad se desea minimizar y, los segundos, al definir qué estructura gobierna la propiedades estocásticas de dicha serie.

El método de ajuste tiene la siguiente estructura: un programa de optimización cuadrática sujeto a restricciones lineales, donde la función objetivo representa una medida de volatilidad de la serie trimestral determinada a priori por el analista y las restricciones lineales recogen la consistencia cuantitativa entre las estimaciones trimestrales y el dato anual observado.

Sea $x = x_{i,t,T} : i = 1..p, t = 1.,4, T = 1..N$ una matriz nxp cuyos elementos recogen las observaciones disponibles sobre p indicadores de frecuencia trimestral, seleccionados para sustentar la desagregación temporal de Y. Pudiendo ser una columna de x un vector de unos, con lo que se puede conside rar la presencia de un término independiente en la relación entre agregado e indicador. La estimación trimestral y compatible con los totales anuales Y se obtiene como solución del siguiente programa de mínimo:

$$\min_{\mathbf{y},\boldsymbol{\beta}} \phi = (\mathbf{y} - \mathbf{x}\boldsymbol{\beta})' \mathbf{D}'\mathbf{D} (\mathbf{y} - \mathbf{x}\boldsymbol{\beta})$$

$$\text{sujeto a } \mathbf{B}\mathbf{y} = \mathbf{Y}$$

donde $\mathbf{D}$ es la versión matricial del operador $(1 - B)$

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & 0 & 0 & ... & ... & 0 \\ 0 & -1 & 1 & 0 & ... & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & -1 & 1 \end{bmatrix}$$

el lagrangiano asociado es

$$\varphi = \phi + 2\boldsymbol{\lambda}' (\mathbf{B}\mathbf{y} - \mathbf{Y}) = (\mathbf{y} - \mathbf{x}\boldsymbol{\beta})' \mathbf{D}'\mathbf{D} (\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) + 2\boldsymbol{\lambda}' (\mathbf{B}\mathbf{y} - \mathbf{Y})$$

con sistema p ecuaciones que reflejan las condiciones de primer orden

$$\frac{\partial \varphi}{\partial \mathbf{y}} = 2\mathbf{D}'\mathbf{D}\mathbf{y} - 2\mathbf{D}'\mathbf{D}\mathbf{x}\boldsymbol{\beta} + 2\mathbf{B}'\boldsymbol{\lambda} = \mathbf{0}$$

$$\frac{\partial \varphi}{\partial \boldsymbol{\beta}} = -2\mathbf{x}'\mathbf{D}'\mathbf{D}\mathbf{y} + 2\mathbf{D}'\mathbf{D}\mathbf{x}\boldsymbol{\beta} = \mathbf{0}$$

$$\frac{\partial \varphi}{\partial \boldsymbol{\lambda}} = 2\left(\mathbf{B}\mathbf{y} - \mathbf{Y}\right) = \mathbf{0}.$$

De este conjunto de ecuaciones se obtienen la expresión para estimar $\mathbf{y}$

$$\hat{\mathbf{y}}_{BA} = \mathbf{x}\hat{\boldsymbol{\beta}} + (\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}'[\mathbf{B}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}']^{-1}\left(\mathbf{Y} - \mathbf{B}\mathbf{x}\hat{\boldsymbol{\beta}}\right)$$

que requiere de la estimación, por separado, de $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{x}'\mathbf{B}'\left[\mathbf{B}(\mathbf{D}'\mathbf{D})\mathbf{D}'\right]\mathbf{B}\mathbf{x}\right]^{-1}\mathbf{x}'\mathbf{B}'\left[\mathbf{B}(\mathbf{D}'\mathbf{D})\mathbf{B}'\right]\mathbf{Y}$$

De este modo se obtiene que la serie trimestralizada es el resultado de agregar dos componentes: uno ligado linealmente al indicador y otro derivado de la distribución de la discrepancia anual existente entre el indicador (debidamente escalado) y el agregado.

Los procedimientos de desagregación temporal basados en modelos asumen que la serie trimestral inobservable y evoluciona según una estructura estadísticamente explícita que detalla sus propiedades estocásticas de manera completa.

Se asume que existe un modelo trimestral que relaciona un vector **x** de p indicadores con la serie trimestral inobservable **y**:

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{u}$$

donde **ß** es un vector de p parámetros constantes pero desconocidos y **u** es una perturbación estocástica de media nula y matriz de varianzas y covarianzas **v**.

Se asume que y satisface la restricción longitudinal habitual:

$$\mathbf{Y} = \mathbf{B}\mathbf{y}$$

Premultiplicando por **B** se obtiene el modelo anual que vincula la serie anual Y con el indicador X temporalmente agregado. De esta manera, se obtiene un modelo lineal que relaciona variables observables:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$$

En general, **U (=Bu)** no será ruido blanco, por lo que la estimación del modelo anterior ha de realizarse por mínimos cuadrados generalizados. Esta es la práctica usual del análisis de la coyuntura cuando se examina la congruencia entre agregados e indicadores.

La idea del procedimiento de trimestralización es definir un estimador lineal que satisfaga $\mathbf{Y} = \mathbf{B}\mathbf{y}$ (restricción longitudinal) y que sea compatible con $\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{u}$. Al ser lineal tiene la forma

$$\hat{\mathbf{y}} = \mathbf{A}\mathbf{Y}$$

La matriz **A**, que determina el estimador, se obtiene a partir de las restricciones impuestas de insesgadez y varianza mínima así como de minimizar la suma de las varianzas de los errores de estimación de cada uno de los trimestres. Esto nos conduce a

$$\mathbf{A} = \mathbf{vB'V}^{-1} + \mathbf{MX'V}^{-1}$$

donde $\mathbf{V} = \mathbf{BvB'}$ y M es la matriz nxp de multiplicadores de Lagrange del problema de optimización asociado. Operando esto nos conduce al estimador

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'V}^{-1}\mathbf{X})^{-1}(\mathbf{X'V}^{-1}\mathbf{Y})$$

que es el estimador de mínimos cuadrados generalizados del modelo anual que relaciona el agregado Y con el indicador X. Este estimdor nos conduce al estimador basado en modelos (BM) de **y**:

$$\hat{\mathbf{y}}_{BM} = \mathbf{x}\hat{\boldsymbol{\beta}} + \mathbf{vB'V}^{-1}\hat{\mathbf{U}} = \mathbf{x}\hat{\boldsymbol{\beta}} + \mathbf{L}\hat{\mathbf{U}}$$

es decir el estimador es la agregación de un término relacionado linealmente con el indicador y de un residuo anual trimestra lizado. La principal característica del estimador radica en la dependencia del filtro de desagregación temporal L de la forma del modelo trimestral y, en particular, de la estructura dinámica de sus perturbaciones.

Por otra parte, si no se dispone de indicadores de aproximación trimestral se pueden emplear métodos de desagregación temporal que sólo tienen en cuenta la información contenida en la serie anual Y. Son los llamados "métodos de desagregación temporal sin indicadores". En efecto, la desagregación temporal de series para las que no se dispone de indicadores de alta frecuencia ha dado lugar a una serie de métodos que combinan de una forma u otra la información muestral contenida en Y_T con determinadas consideraciones a priori acerca de las propiedades estocásticas de las serie desagregada, que han de ser suministradas por el analista.

Existen varios métodos, entre ellos está usar regresión con variables deterministas como una tendencia constante o lineal como indicador. A este respecto, destacamos un método que está formalmente relacionado con los expuestos hasta ahora.

Se plantea la estimación de y_t como la solución de un programa de minimización restringida en el que, por una parte, la función objetivo recoge determinadas consideraciones a priori sobre la evolución tendencial de la serie y_t objeto de estimación y, por otra parte, las restricciones reflejan la necesaria consistencia temporal característica de los problemas de desagregación temporal.

A esto efectos, se define la variable auxiliar

$$u_t = \nabla^d y_t, \nabla = (1 - B), d = 0, 1, 2$$

y se intenta minimizar la suma de cuadrados. Esta función objetivo es una manera de imponer a la serie estimada y_t una determinada estructura en su evolución tendencial. De esta manera, si el analista considera que debe poseer una tendencia estocástica integrada de orden uno aplicará la transformación $u_t = \nabla y_t$, es decir, $d = 1$. En consecuencia, un criterio frecuentemente utilizado en el ámbito apli cado consiste en seleccionar d en de

modo que refleje el mismo grado de diferenciación que la serie anual Y_T que se desea trimestralizar.

El problema de minimización será por lo tanto

$$\min_y \sum_2^N u_t^2 = \sum_2^N \left(\nabla^d y_t\right)^2$$

$$\text{sujeto a } \sum_{t=1}^4 y_{t,T} = Y_T$$

que de forma matricial se expresa del siguiente modo

$$\min_{\mathbf{y}} \mathbf{u}'\mathbf{u} = \mathbf{y}'\mathbf{D}'\mathbf{D}\mathbf{y}, \text{s.a. } \mathbf{B}\mathbf{y} = \mathbf{Y}$$

Tras forma el lagrangiano y aplicar las condiciones de primer orden, se obtienen las siguientes ecuaciones

$$\begin{bmatrix} \mathbf{D}'\mathbf{D} & \mathbf{B}' \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{Y} \end{bmatrix}$$

y despejando se llega a la expresión explícita para $\mathbf{y}$

$$\mathbf{y} = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}'[\mathbf{B}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}']^{-1}\mathbf{Y} = \mathbf{A}(\mathbf{D},\mathbf{B})\mathbf{Y}$$

En la ecuación anterior la expresión $\mathbf{A}(\mathbf{D},\mathbf{B})$ es una matriz nxN que, aplicada a la serie anual observada **Y** da lugar a la correspondiente serie trimestral **y**, cuantitativamente coherente con la primera (**By=Y**) y cuya volatilidad intertrimestral es lo menor posible. Esta matriz **A** depende de **B** y **D**, esto es, su estructura está condicionada por la naturaleza del problema de desagregación temporal que se plantea (**B**) y por el grado de suavidad que se desea imponer a la serie estimada (**D**). Naturalmente, como ya se ha comentado, esta última dependencia equivale a una determinada consideración a priori de las características tendenciales de **y**. Asimismo, la matriz **A** también puede interpretarse como un filtro lineal cuyos coeficientes varían con el tiempo, de forma que el filtro que se utiliza en los extremos de la serie no es el mismo que el que se aplica en el tramo central.

De forma más general, existe un método basado en este expuesto que considera que la serie trimestral inobservable sigue un proceso ARIMA(p,d,q), en lugar de un proceso ARIMA$(0,1,0)$.

14.6 Calibrado, benchmarking y reconciliación.

Las series temporales observadas se registran a una frecuencia constante: diaria, semanal, mensual, trimestral y anual. Como sabemos, las mediciones pueden tener su origen en registros administrativos, encuestas repetidas y censos con diferentes niveles de confiabilidad. Una opinión común entre los usuarios de series de tiempo es que los

datos son el resultado de compilaciones directas de las mediciones de una de estas fuentes. De hecho, antes de su publicación oficial por parte de los organismos de estadística, los datos de las series de tiempo están sujetos a varios ajustes para aumentar la eficiencia, reducir el sesgo, reemplazar los valores perdidos, facilitar el análisis y cambiar la frecuencia (por ejemplo, de anual a trimestral). Los ajustes más comunes son el benchmarking, la extracción de señales, la interpolación y la distribución temporal y la calendarización.

Por otra parte, los datos de las series temporales no siempre coinciden con los períodos del calendario. Para reducir la carga de respuesta a las encuestas y censos, las agencias de estadística a menudo aceptan datos con períodos fiscales. El momento y la duración de un valor informado que cubre períodos fiscales o de calendario dependen de las fechas de inicio y finalización del período de presentación de informes. La duración es el número de días del período del informe. El momento puede definirse en términos generales como la fecha intermedia del período de presentación de informes.

En muchas industrias o sectores, sin embargo, los años fiscales no son homogéneos y varían de una unidad a otra (encuestado). Como resultado, el momento y la duración de los valores informados difieren de una unidad a otra e incluso para una unidad determinada. Por tanto, los datos fiscales no son muy útiles. No es posible sumar de forma transversal los distintos años fiscales de un año calendario determinado porque no tienen el mismo tiempo. Un enfoque burdo de la calendarización es el procedimiento de asignación, que asigna datos fiscales a un año específico (u otro período calendario) de acuerdo con una regla arbitraria. Esta técnica distorsiona sistemáticamente el momento y la duración de las estimaciones de manera diferente en diferentes industrias y subestima la amplitud de los ciclos económicos. La técnica representa erróneamente los períodos de calendario objetivo, difumina la causalidad o asociaciones entre las variables socioeconómicas y conduce a un análisis contaminado.

Otro problema de las series oficiales es que una gran cantidad de las mismas pertenecen a un sistema de series clasificadas por atributos. Por ejemplo, las series de población activa se clasifican por provincia, edad, sexo, empleo a tiempo parcial y tiempo completo. En tales casos, la serie del sistema debe conciliarse para satisfacer ciertas restricciones de agregación transversales.

A este tipo de problemas en general se le denomina problema de la reconciliación y ha preocupado a economistas y estadísticos durante más de setenta años.

El problema del **benchmarking** surge cuando los datos de series temporales para la misma variable objetivo se miden a diferentes frecuencias con diferentes niveles de precisión. Normalmente, una fuente de datos es más frecuente que la otra: por ejemplo, series mensuales por un lado y series anuales por el otro, series mensuales y trimestrales, trimestrales y anuales, diarias y multisemanales... . La serie observada con mayor frecuencia suele ser menos "confiable" que la observada con menor frecuencia. Por esta razón, esta última se considera generalmente como referencia.

Por lo general, se observan discrepancias entre los puntos de referencia anuales (digamos) y las sumas anuales correspondientes de las series más frecuentes. En términos generales, el benchmarking consiste en combinar las fortalezas relativas de las dos

fuentes de datos. Más concretamente, consiste fundamentalmente en adoptar el movimiento de las series más frecuentes y el nivel de los benchmarks. La idea es que se puede utilizar información de origen administrativo o censal para mejorar estimaciones basadas en operaciones estadísticas que se repiten mensual o trimestralmente. Para simplificar la exposición, nos referiremos a la serie de alta frecuencia como la serie subanual y a la serie de baja frecuencia como la serie anual, a menos que sea necesario ser más específico. Las situaciones que requieren benchmarking son muy comunes en las agencias de estadística. En algunos casos, tanto la serie subanual como la anual se originan en encuestas o censos; en otros, ninguna de las dos series se origina en encuestas o solo una lo hace.

El benchmarking también se produce en el contexto del ajuste estacional que hemos visto en el apartado central de este tema. El ajuste estacional de una serie temporal mensual o trimestral provoca discrepancias entre las sumas anuales de la serie bruta (sin ajustar estacionalmente) y las sumas anuales correspondientes de la serie ajustada estacionalmente. Estas series ajustadas estacionalmente se comparan luego con las sumas anuales de la serie bruta.

La necesidad del benchmarking surge porque las sumas anuales de la serie subanual (alta frecuencia) no son iguales a los valores anuales correspondientes (baja frecuencia). En otras palabras, existen discrepancias anuales, d_m, entre los valores de referencia anuales y los valores subanuales

$$d_m = a_m - \sum_{t=t_{1m}}^{t_{Lm}} j_{mt} s_t, m = 1, ..., M$$

donde t_{1m} y t_{Lm} son respectivamente el primer y último período subanual cubierto por m-ésima referencia (benchmark), p. ej. trimestres 1 a 4 para el primer punto de referencia, 5 a 8 para el segundo, y así sucesivamente. Las cantidades j_{mt} son las fracciones de cobertura, aquí asumimos que son iguales a la unidad. Obsérvese que la cantidad $j_{m\bullet} = \sum_t j_{mt}$ es el número de períodos subanuales cubiertos por el m-ésimo punto de referencia.

Las discrepancias anuales se expresan más a menudo en términos de discrepancias anuales proporcionales:

$$d_m^{(p)} = a_m / \left(\sum_{t=t_{1m}}^{t_{Lm}} j_{mt} s_t \right), m = 1, ..., M$$

que podría ser menor, igual o mayor que la unidad.

El benchmarking generalmente consiste en imponer los valores anuales a los valores subanuales. Es decir, se modifica la serie subanual para que las sumas anuales de la serie subanual sean iguales al índice de referencia correspondiente. Es decir

$$a_m - \sum_{t=t_{1m}}^{t_{Lm}} j_{mt} \hat{\theta}_t = 0, m = 1, ..., M$$

siendo $\hat{\theta}_t$ la de la serie de referencia.

El benchmarking clásico supone que las referencias anuales son totalmente fiables y, por tanto, vinculantes, es decir, cumplen la restricción de la ecuación anterior. Una definición más amplia de benchmarking reconoce el hecho de que las referencias pueden observarse con error. Dichos puntos de referencia se denominan *no-vinculantes*, porque las series de puntos de referencia pueden no satisfacer necesariamente las restricciones indicadas. Así pues, el benchmarking consiste en combinar de manera óptima tanto los datos más frecuentes como los menos frecuentes para aumentar su fiabilidad. De este modo se puede obtener un conjunto mejorado de valores anuales tomando las sumas anuales de la serie de referencia

$$\hat{a}_m = \sum_{t=t_{1m}}^{t_{Lm}} j_{mt}\hat{\theta}_t = 0, m = 1, ..., M$$

El éxito de los métodos discutidos depende en gran medida de la disponibilidad de datos adecuados. Cada índice de referencia (benchmark) debe estar fechado exactamente, por medio de una fecha de inicio real y una fecha de finalización real, ambas definidas en términos de día, mes y año, p. Ej. 15 de abril de 2003 y 14 de abril de 2004. Esto es especialmente crítico en el caso de la calendarización. Debido a que la calendarización es un ajuste relativamente nuevo realizado por las agencias de estadística, las fechas de inicio y finalización de los datos reportados pueden ser inexactas o estar ausentes. A veces, la única fecha en el registro es la prevista para el registro por el diseño de la encuesta, por ejemplo, el año 2004, en lugar del 15 de abril de 2003 y el 14 de abril de 2004. Las brechas temporales en los puntos de referencia son aceptables, porque los puntos de referencia pueden estar disponibles cada dos años, o cada cinco años, o incluso de forma irregular. Por otro lado, estas brechas son inaceptables en la serie subanual que se va a comparar. Más específicamente, la serie mensual original (por ejemplo) debe proporcionar observaciones para todos los meses en el intervalo de meses cubierto por los puntos de referencia. Si la primera referencia tiene una fecha de inicio igual al 15 de enero de 1991 y la última referencia tiene una fecha de finalización igual al 14 de junio de 2005, la serie original sin referencia debe abarcar al menos desde el 1 de enero de 1991 hasta el 30 de junio de 2005.

Por último destacar que la gran mayoría de los datos de series temporales producidos por los organismos de estadística suelen formar parte de un sistema de series clasificadas por atributos. Por ejemplo, una serie de comercio minorista puede ser clasificada por grupo comercial (tipo de tienda) y provincia. En tales casos, la serie del sistema debe satisfacer restricciones de agregación transversal. En muchos casos, cada serie también debe sumar temporalmente sus índices de referencia anuales; estos requisitos se denominan restricciones de agregación temporal.

Algunos académicos inicialmente sugirieron el enfoque de ajuste proporcional iterativo como una aproximación a su optimización numérica de "mínimos cuadrados", probablemente justificada por la falta de potencia de cálculo en ese momento. Un problema importante resultó ser la cantidad de cálculos involucrados. Por lo tanto, con el fin de equilibrar las tablas de input-output, el método RAS se hizo muy popular. Sin embargo, tenía las siguientes deficiencias: no se tenían en cuenta los diversos grados de incertidumbre sobre las estimaciones iniciales y sus restricciones, una interpretación

económica dudosa de los ajustes prorrateados de las cuentas y un gran número de iteraciones antes de la convergencia.

La atención se centró nuevamente en el enfoque de mínimos cuadrados con procedimientos alternativos para minimizar una función de pérdida cuadrática restringida basada en el algoritmo de gradiente conjugado. Siendo esta línea metodológica una de las más utilizadas.

Bibliografía complementaria

Matilla-García, M et al. 2017. Econometría y Predicción. McGraw-Hill.

Quillis, E.M., 2018. Temporal disaggregation of economic time series: The view from the trenches. Statistica Neerlandica. doi: 10.1111/stan.12150