Modelos
Lineales

Colección manuales uex - 56
(E.E.E.S.)

Jesús
Montanero Fernández

56

MANUALES UEX

56

(E.E.E.S.)
Espacio
Europeo
Educación
Superior

JESÚS MONTANERO FERNÁNDEZ

MODELOS LINEALES

UNIVERSIDAD DE EXTREMADURA

2008

La publicación del presente manual forma parte de las "Acciones para el Desarrollo del Espacio Europeo de Educación Superior en la Universidad de Extremadura Curso 2007/08" en el marco de la VI Convocatoria de Acciones para la Adaptación de la UEX al Espacio Europeo de Educación Superior (Proyectos Pilotos: modalidad A1) del Vicerrectorado de Calidad y Formación Continua y financiada por la Junta de Extremadura, el Ministerio de Educación y Ciencia y la Universidad de Extremadura.

UNIVERSIDAD DE EXTREMADURA

UNIÓN EUROPEA
Fondo Social Europeo

JUNTA DE EXTREMADURA

Edita

Universidad de Extremadura. Servicio de Publicaciones
C./ Caldereros, 2 - Planta 2ª - 10071 Cáceres (España)
Telf. 927 257 041 - Fax 927 257 046
publicac@unex.es
www.unex.es/publicaciones

ISSN 1135-870-X
ISBN 978-84-691-6344-3
Depósito Legal M-45.207-2008

Edición electrónica: Pedro Cid, S.A.
Teléf.: 914 786 125

Prólogo

El presente manual esta concebido como un apoyo a la docencia en una asignatura de segundo ciclo que puede cursarse tanto en la Licenciatura de Matemáticas como en la de Ciencias y Técnicas Estadísticas. El objetivo es que pueda ser entendido por alumnos con conocimientos básicos de Matemáticas en general y Estadística en particular.

Los aspectos formales de la materia han sido desarrollados con cierto detalle. En lo que respecta a las competencias cuya adquisición debe posibilitar esta asignatura, no es estrictamente necesaria la compresión exhaustiva de los mismos, aunque se antoje conveniente que el lector interesado tenga al menos un lugar donde acudir si quiere llevar a cabo un estudio más profundo de la materia, al margen de la bibliografía especializada. Por contra, el alumno debe tener en cuenta que el conocimiento teórico de estos contenidos debe complementarse con su aplicación mediante un programa estadístico. En la página web `http://kolmogorov.unex.es/jmf~/` se encuentra material al respecto.

También cabe resaltar que este manual se complementa con otro dedicado al Análisis Multivariante. De hecho podría considerarse como una primera parte o primer volumen de una serie de dos.

Introducción

El planteamiento y resolución de ecuaciones matemáticas tienen como objeto relacionar el comportamiento de una variable respuesta con el de una o varias variables explicativas. Podemos distinguir entre diversos tipos de ecuaciones: lineales, no lineales, diferenciales, etc. Nosotros estudiaremos fundamentalmente las primeras, es decir, consideraremos básicamente relaciones de tipo lineal entre la variable respuesta y las variables explicativas. ¿Por qué? Si bien es cierto que este tipo de relación se observa con relativa frecuencia en la naturaleza, hemos de reconocer, para ser honestos, que su principal virtud es su fácil manejo, su excelente y natural comportamiento desde el punto de vista formal, lo cual invita en no pocas ocasiones a considerar como lineales relaciones que sólo lo son aproximadamente, asumiendo en consecuencia cierto error como tributo a la sencillez del modelo. Cuando este error resulta excesivo es costumbre bastante habitual buscar cambios apropiados en las variables que permitan establecer relaciones aproximadamente lineales entre las variables transformadas. Podemos también añadir a las variables explicativas distintas potencias de grado superior de las mismas. De esta forma, las ecuaciones polinómicas quedan reducidas a un caso particular de ecuaciones lineales, lo cual permite cubrir aproximadamente un enorme campo de posibilidades. En definitiva, la solución a un problema de ecuaciones lineales y, en definitiva, la teoría del Álgebra Lineal, puede servirnos como referencia o punto de apoyo para la resolución de ecuaciones que, en principio, no los son.

Lo dicho hasta ahora puede encuadrarse en un marco determinista, donde las relaciones entre las variables sean siempre idénticas, independientemente del resultado concreto del experimento. Sin embargo, nosotros estamos dispuestos a admitir una variación o error de carácter aleatorio, lo cual conduce a considerar un modelo de tipo probabilístico. Dado que las distribuciones de probabilidad en juego no están especificadas por completo –de lo contrario, podríamos considerar el problema resuelto–, habría que hablar, para ser exactos, de un modelo estadístico, que denominaremos en lo sucesivo Modelo Lineal. Con frecuencia, se supone que el error del modelo, es decir, las diferencias entre el valor de la variable respuesta y el que predice la ecua-

ción lineal, sigue una distribución normal, lo cual convierte este modelo, denominado en ese caso Modelo Lineal Normal, en el mismo núcleo de la Estadística Paramétrica . El supuesto de normalidad es de gran utilidad a la hora de contrastar diversas hipótesis relativas a los parámetros o construir regiones de confianza para los mismos. Además, supone un argumento fundamental en la justificación de los tests de hipótesis y estimadores que se elaboran en la teoría.

Nuevamente nos encontramos ante la misma problemática. Aunque, efectivamente, se puedan observar en la práctica relaciones de tipo lineal salvo errores aleatorios normalmente distribuidos, la asunción del supuesto de normalidad no dejará de resultar al lector más suspicaz una artimaña para resolver problemas de carácter meramente técnico, y quizá no le falte buena parte de razón. Es mucho lo estudiado acerca de este delicado problema que, en buena lógica, podría disuadirnos del uso de los métodos Paramétricos en general y, ésa es, hoy en día, la opinión de buena parte de los estadísticos. No obstante, nos atrevemos aquí a romper una lanza en favor del supuesto de normalidad. Efectivamente, los métodos de Inferencia Estadística propios del modelo tienen un buen comportamiento asintótico aún obviando el supuesto de normalidad, es decir, que funcionan de manera *similar* al caso normal para muestras suficientemente grandes. No cabe duda de que detrás de esta afirmación debe estar –y así lo veremos– alguna versión del Teorema Central del Límite. El propio Teorema Central del Límite podría explicar la normalidad observada de hecho en muchos casos, en los cuales la variable respuesta podría ser la suma o conjunción de muchas variables independientes.

No obstante y yendo un poco más lejos, no parece del todo coherente extrañarse del uso del supuesto de normalidad cuando se ha asumido sin problemas el de linealidad, o cuando se afronta con absoluta naturalidad la inferencia acerca de la media y la varianzas (o matriz de varianzas-covarianzas). ¿Por qué? La pregunta debería ser más bien: ¿por qué estudiamos la media, la varianza o la covarianza? ¿No son éstos los parámetros que caracterizan la distribución normal (posiblemente multivariante)? Desde luego, si de una distribución desconocida suponemos su normalidad, conocer su media y varianza (o matriz e covarianzas en el caso multivariante) equivale a especificarla por completo, es decir, a convertir el problema estadístico en un problema meramente probabilístico, cosa que no ocurre en general. Si hablamos en términos muestrales, es desde luego continuo el uso que hacemos de la media y la varianza , lo cual podría justificarse mediante el hecho de que, conjuntamente, constituyen un estadístico suficiente y, además, completo. Pero esa afirmación es correcta precisamente bajo el supuesto de normalidad –por ejemplo para una muestra aleatoria simple de una distribución normal–. Más aún, es bien conocido que, dado un vector aleatorio

normal multivariante, las relaciones entre sus distintas componentes han de ser de tipo lineal. Con ello estamos llamando la atención sobre una vinculación natural entre los supuestos de normalidad y linealidad. Por todo ello, el objeto principal de nuestro estudio no será el Modelo Lineal sino, más concretamente, el Modelo Lineal Normal.

Lo primero que necesitamos aclarar en nuestra teoría es en qué sentido el Modelo Lineal formaliza los problemas cuya resolución nos atañe, que son, principalmente, el problema de regresión lineal, el de correlación lineal, el de análisis de la varianza y el de análisis de la covarianza. Podemos añadir a estos problemas otros similares que se encuadran en los denominados modelos lineales generalizados. En el primer capítulo se enuncian cuatro ejemplos que pretenden ilustrar los problemas mencionados anteriormente, a los que sigue una discusión acerca de su formalización mediante el modelo lineal, cuyo principal objetivo es la justificación de la bibliografía de referencia y el enfoque que hemos dado a esta materia.

Tras la reflexión inicial del capítulo 1, procederemos a desarrollar el programa en sí. Empezaremos con un capítulo dedicado a la distribución normal multivariante, haciendo especial hincapié en el caso esférico y en distribuciones derivadas de la misma, como son la χ^2, t de Student y F de Snedecor. También se analiza con cierto detenimiento la conexión existente entre los supuestos de normalidad y linealidad.

Es nuestra intención que este manual sea, en la mayor medida posible, *autocontenido*. Por ello hemos procurado demostrar los resultados que se exponen, si bien en algunos casos hemos considerado más conveniente remitir al lector a la oportuna referencia bibliográfica. Tal es el caso, por ejemplo, de todos los resultados *clásicos* en Probabilidad y Estadística que se precisan en esta teoría pero no son específicos de la misma. En general, las nociones y resultados previos que se requieren para afrontar nuestro estudio se exponen en el Apéndice. Se trata de una miscelánea de materias, la mayoría de las cuales pueden ser obviadas por el lector con conocimientos básicos en Probabilidad y Estadística. En la primera sección del mismo se recoge una selección de resultados del Álgebra matricial que serán de utilidad.

La piedra angular de nuestra teoría es el capítulo 3, donde se establecen una serie de resultados teóricos que serán de utilidad a la hora de estudiar los análisis de regresión y de la varianza en los capítulos 4 y 6, respectivamente. El capítulo 5, dedicado al modelo Correlación, es de es carácter netamente teórico y viene a complementar al tercero o al cuarto, según se entienda. Aunque hubiera sido más cómodo, desde el punto de vista técnico, incluirlo en la segunda parte, dedicada al Análisis Multivariante, hemos preferido presentarla en la primera para dar mayor coherencia al conjunto. El capítulo 7 se dedica al modelo lineal de rango no completo y el 8 a los modelos lineales generalizados.

Por otra parte, según se ha mencionado de pasada, este manual pretende ser un volumen previo a otro de dedicado al Análisis Multivariante. Obviamente, ambas materias comparten muchos contenidos pudiendo considerarse el estudio del Modelo Lineal un requisito previo al del Análisis Multivariante, aunque en ocasiones puede suceder lo contrario. Ambos volúmenes se conciben como complementarios y comparten la misma notación y filosofía, si bien el Análisis Multivariante presenta especial dificultad debido a la carencia de una verdadera cohesión lógica, al menos en la medida en que la posee el Modelo Lineal.

La referencia bibliográfica fundamental de ambos volúmenes es Arnold (1981). El título lo dice todo: *The Theory of Linear Models and Multivariate Annalysis*. En esta obra se basan sobre todo los capítulos 3 y 5 del presente volumen, así como el capítulo 2 del volumen dedicado al Análisis Multivariante. Recordamos que uno de los objetivos principales del capítulo 1 es justificar la elección de esta referencia bibliográfica como pilar para exponer la teoría que nos incumbe, en contraposición con otras formas de explicarla, más frecuentes, que podemos encontrar en multitud de libros de texto actuales y de referencias clásicas.

Índice general

Capítulo 1

Ejemplos y discusión

En esta primer capítulo vamos a exponer cuatro ejemplos, los cuales representan diferentes problemas que pueden formalizarse mediante el modelo lineal. Nos referimos a los problemas de Regresión Lineal, Correlación Lineal, Análisis de la Varianza y de la Covarianza y, por ultimo un problema de rango no completo. Nos hemos permitido la licencia de utilizar en las discusiones conceptos y notaciones propios de la Teoría de la Probabilidad y de la Estadística Matemática con los que el lector puede no estar familiarizado. No obstante, es nuestro propósito que cualquier duda al respecto quede aclarada en el capítulo 2 o en el Apéndice. Los datos correspondientes a los ejemplos podemos encontrarlos en formato SPSS en `http:/kolmogorov.unex.es/jmf∼/` .

1.1. Ejemplos

1. [`Linthurst Data`]: *Se pretende relacionar de manera precisa la producción de Biomasa de Espartina con la salinad, acidez y concentraciones de potasio, sodio y zinc del terreno donde ésta crece. Se tomaron un total de 45 mediciones de estas seis variables.*

2. [`Peso`]: *Se pretende establecer la relación existente entre la edad en semanas de un feto de entre 28 y 33 semanas y su peso. Para ello se midieron los pesos en gramos de 30 fetos, 5 de ellos de 28 semanas, 5 de 29, 5 de 30, 5 de 31, 5 de 32 y otros 5 de 33.*

3. [`Hipertensión`]: *Se desean comparar la efectividad de dos medicamentos A y B, junto con un placebo C, para combatir la hipertensión. Para ello se consideraron 30 pacientes hipertensos, 10 de los cuales fueron tratados con A, otros 10 con B y el resto con C. Pasado cierto tiempo se midió en cada caso el porcentaje*

de descenso de la presión arterial media –aquí el término media hace referencia a la semisuma entre la sistólica y diastólica–.

4. [`Complexión`]: *Se pretende establecer una relación clara entre la altura y el peso corporal en personas sanas dependiendo del tipo de complexión natural. Para ello se distingen tres complexiones, A, B y C y, para cada una de ellas, se toma una muestra de 10 individuos a los que se les miden ambas variables.*

1.2. Formalización

Procedamos a analizar los problemas de uno en uno para determinar qué modelo estadístico es el más apropiado para formalizarlos.

Problema de regresión lineal

En el problema uno, nuestros datos configuran seis vectores en $\mathbb{R}^{45}$, $\mathtt{Z}[j]$, con $1 \leq j \leq 5$, e Y, donde los cinco primeros hacen referencia a las variable explicativas (condiciones del terreno) y la última a la variable respuesta (biomasa). La componente i-ésima de cada vector corresponde al caso (individuo) i-ésimo del estudio. Supondremos que los valores correspondientes a las variables explicativas han sido determinados de antemano, siendo aleatorios los correspondientes a la variable respuesta, y que la relación entre la variable respuesta y las explicativas es lineal[1], es decir, que existen, $\beta_j \in \mathbb{R}$, $j = 0, 1, \ldots, 5$, tales que

$$Y_i = \beta_0 + \beta_1 \mathtt{Z}_i[1] + \beta_2 \mathtt{Z}_i[2] + \beta_3 \mathtt{Z}_i[3] + \beta_4 \mathtt{Z}_i[4] + \beta_5 \mathtt{Z}_i[5] + \varepsilon_i,$$

donde ε_i denota el error cometido, es decir, la diferencia entre el valor exacto de y y el que se obtiene a partir de las variables explicativas mediante la ecuación lineal. Consideraremos dichos errores como variables aleatorias incorreladas con media 0 y varianza finita común σ^2. Expresemos el modelo estadístico formulado en lenguaje matricial. Sean $\mathtt{X}$ la matriz 45×6 cuya primera columna está compuesta exclusivamente de unos (se denota por 1_{45}), siendo $\mathtt{Z}[j]$, donde $j = 1, \ldots, 5$, las restantes; β el vector (columna) compuesto por los β_j, desde $j = 0$ hasta 5, y $\mathcal{E}$ el vector aleatorio compuesto por las variables ε_i, desde $i = 1$ hasta 45. Entonces, se verifica que

$$Y = \mathtt{X}\beta + \mathcal{E},$$

[1]Deberíamos decir realmente afín, pues introducimos una constante en la ecuación.

siendo las componentes de $\mathcal{E}$ incorreladas con media 0 y varianza finita común σ^2. Si, además, consideramos que los errores están normalmente distribuidos, el modelo vendrá dado por un vector aleatorio Y que verifica que $Y = \mathtt{X}\beta + \mathcal{E}$, donde $\mathcal{E}$ sigue un modelo de distribución $N_{45}(0, \sigma^2\mathtt{Id})$ y β y σ^2, los parámetros del modelo, pueden ser cualquier elemento de $\mathbb{R}^6$ y $\mathbb{R}^+$, respectivamente. Se trata de un Modelo de Regresión Lineal Normal con término independiente[2], que puede expresarse, equivalentemente, de la siguiente forma:

$$Y \sim N_{45}(\mathtt{X}\beta, \sigma^2\mathtt{Id})$$

Las componentes de β se denominan coeficientes de regresión, y σ^2 puede interpretarse como una medida del error implícito a la ecuación lineal. Se supone también que la matriz $\mathtt{X}$ es de rango completo, es decir, que todas sus columna son linealmente independientes. En caso contrario, el valor del parámetro β no quedaría unívocamente determinado por la distribución de probabilidades dada.

¿Regresión o Correlación?

Nótese que, en el primer estudio, estamos considerando Y como un vector aleatorio mientras que $\mathtt{X}$ es una matriz constante, es decir, que suponemos que los datos de las variables explicativas son fijados de antemano, dependiendo del azar únicamente el resultado de la variable respuesta. No parece que este sea el diseño correspondiente al estudio 1, pero sí es exactamente lo que ocurre en el estudio número 2, donde se mide el peso del fémur en fetos con edades prestablecidas con el objetivo de establecer la relación entre ambas variables. Éste, y no aquél, sí que es un Modelo de Regresión, rigurosamente hablando. Discutiremos este asunto a continuación.

Efectivamente, parece claro que en el primer estudio, tanto las variables explicativas como la respuesta deben ser consideradas aleatorias. Cada unidad experimental de la muestra aporta realmente siete datos (uno más cinco), es decir, un vector aleatorio con valores en $\mathbb{R}^6$. Por lo tanto, las observaciones aleatorias no pertenecen a $\mathbb{R}^{45}$ sino que son matrices de orden 45×6. La primera columna de la matriz aleatoria es Y y la submatriz restante, Z. En lo que sigue, $\mathtt{Z}$ denotará una matriz fija de dimensiones 45×5, mientras que $\mathtt{X}$ y X serán las matrices fijas y aleatorias que se obtienen mediante

$$\mathtt{X} = (1_{45}|\mathtt{Z}), \qquad X = (1_{45}|Z)$$

Un Modelo de Correlación Lineal se corresponde con una muestra aleatoria simple de tamaño 45 en este caso $(Y|Z)$ de una distribución normal no degenerada en dimensión

[2]El término independiente puede eliminarse si se supone que la relación entre las variables es lineal en sentido estricto y no afín, como estamos considerando en principio.

6. En ese caso, veremos que las columnas de X son linealmente independientes con probabilidad 1, que las filas de Z constituyen una muestra aleatoria simple de una distribución normal en dimensión 5 y que Y y X se relacionan mediante

$$Y = X\beta + \mathcal{E},$$

siendo $\mathcal{E}$ un vector aleatorio de dimensión 45 de componentes normales, independientes , de media 0 y varianza común, y siendo $\mathcal{E}$ y Z independientes. En ese caso, los parámetros del modelo son la media y matriz de varianzas-covarianzas de las zetas, junto con β y la varianza común σ^2. Equivalentemente, se verifica que la distribución del vector aleatorio Y condicionada a que la submatriz aleatoria Z tome el valor $\mathtt{Z}$, sigue un modelo $N_{45}(\mathtt{X}\beta, \sigma^2\mathtt{Id})$. Es decir, el modelo de Correlación Lineal puede expresarse mediante

$$Y|Z = \mathtt{Z} \sim N_{45}(\mathtt{X}\beta, \sigma^2\mathtt{Id}), \qquad Z \sim N_{45}(\nu, \Xi)$$

Por lo tanto, el Modelo de Regresión Lineal Normal puede obtenerse *condicionando* en el Modelo de Correlación. Ocurre además que, si los problemas principales de inferencia relativos a los parámetro β y σ^2 del modelo de Regresión se abordan desde el Modelo de Correlación, se obtienen los mismos estadísticos que se derivan del de Regresión, y con las mismas distribuciones (pues éstas resultan no depender del valor concreto $\mathtt{Z}$ sobre el que se condiciona). Por lo tanto, los mencionado métodos de Inferencia conducen a las mismas conclusiones, bien se afronten desde el modelo de Regresión, es decir, con $\mathtt{X}$ fija, bien desde el modelo de Correlación, o sea, con X aleatoria[3]. Por ello, en la práctica, no supone problema alguno considerar, como en el estudio 1, un modelo de Regresión cuando no parece verosímil que las valores de las variables explicativas hayan sido fijado de antemano.

La principal ventaja del Modelo de Correlación estriba en que permite intercambiar las variables respuestas con las explicativas y realizar inferencias estadísticas acerca de los diversos coeficientes de correlación (simples, múltiples, canónicos y parciales). Además, un Modelo de Regresión Lineal en sentido estricto es poco factible con un numero elevado de variables explicativas, pues se trata de tomar, para cada valor concreto de las mismas, una muestra de la variable respuesta. Sin embargo, el hecho de controlar las variables explicativas, como en el Modelo de Regresión *puro*, evita la presencia de valores extremos potencialmente influyentes y permite contrastar por separado los supuestos del modelo.

[3]No obstante, veremos que existen ciertos matices que los diferencian, referentes únicamente a la justificación teórica de los mismos.

Análisis de la varianza

El tercer problema corresponde a lo que se denomina un Diseño Completamente Aleatorizado del Análisis de la Varianza. En esta ocasión, se toman 30 mediciones, que se supondrán independientes, de una variable respuesta y, 10 de ellas en cada uno de los tres grupos considerados (A, B y Placebo). El objeto del estudio es decidir si el uso de los medicamentos afectan a la distribución de la variable y (porcentaje de descenso de la presión arterial) y en qué sentido. En principio tendremos tres muestras, que supondremos aleatorias simples, todas ellas de tamaño 10, correspondientes a sendas distribuciones reales, de medias μ_i, $i = 1, 2, 3$, respectivamente. Se denotarán mediante Y_{ij}, donde el subíndice i, con valores entre 1, 2 y 3, hace referencia al medicamento (A, B y Placebo, respectivamente), mientras que j, entre 1 y 10, hace referencia al individuo en sí. Definimos entonces los errores $\varepsilon_{ij} = Y_{ij} - \mu_i$. Veamos entonces cómo expresamos el modelo.

En primer lugar, para cada $m \in \mathbb{N}$, 1_m y 0_m denotarán los vectores de $\mathbb{R}^m$ cuyas componentes son todas iguales a 1 y 0, respectivamente. En ese caso, se definen

$$\mathtt{v}_1 = \begin{pmatrix} 1_{10} \\ 0_{10} \\ 0_{10} \end{pmatrix}, \qquad \mathtt{v}_2 = \begin{pmatrix} 0_{10} \\ 1_{10} \\ 0_{10} \end{pmatrix}, \qquad \mathtt{v}_3 = \begin{pmatrix} 0_{10} \\ 0_{10} \\ 1_{10} \end{pmatrix}$$

De esta forma, si Y y $\mathcal{E}$ denotan los vectores de dimensión 30 que se obtiene componiendo ordenadamente las variables de la forma Y_{ij} y ε_{ij}, se tiene entonces que

$$Y = \sum_{i=1}^{3} \mu_i \cdot \mathtt{v}_i + \mathcal{E}.$$

Si asumimos que los errores se distribuyen según un modelo normal con varianza común σ^2, se verifica que $\mathcal{E} \sim N_{30}(0, \sigma^2 \mathtt{Id})$. Así pues, el modelo puede expresarse mediante

$$Y \sim N_{30}(\mu, \sigma^2 \mathtt{Id}),$$

donde μ puede ser cualquier vector del subespacio V de $\mathbb{R}^{30}$ generado por $\mathtt{v}_1$, $\mathtt{v}_2$ y $\mathtt{v}_3$, y σ^2 cualquier número positivo. El hecho de suponer normalidad e igualdad de las varianzas (homocedasticidad) simplifica sensiblemente el modelo. Pero además, bajo estos supuestos, la igualdad de las seis distribuciones consideradas se corresponde con la igualdad de las medias, es decir, que el contraste de hipótesis principal es un contraste de medias. Concretamente, queremos saber si el parámetro μ pertenece al subespacio de W generado por el vector 1_{30}. La igualdad entre, por ejemplo, las dos primeras distribuciones (es decir, entre los medicamentos A y B), se corresponde con

la hipótesis $\mu \in \langle \mathtt{v}_1 + \mathtt{v}_2 \rangle$. En general, estudiaremos contrastes del tipo $\mu \in W$, siendo W un subespacio de V.

A continuación, esclareceremos la relación existente entre los modelos de Regresión y de Análisis de la Varianza. Consideramos en el problema 1 el subespacio de $\mathbb{R}^{45}$ generado por las columnas de la matriz $\mathtt{X}$, de dimensión 6, y reparametricemos el experimento estadístico mediante $\mu = \mathtt{X}\beta$. Nótese que existe una correspondencia biunívoca entre μ y β dado que $\mathtt{X}$ es de rango completo. Podemos decir que el vector $\beta \in \mathbb{R}^6$ se compone de las coordenadas de la media μ de Y respecto de la base $\mathtt{X}$. De esta forma, el modelo de Regresión puede expresarse mediante

$$Y \sim N_{45}(\mu, \sigma^2 \mathtt{Id}),$$

donde μ puede ser cualquier valor del subespacio $V = \langle \mathtt{X} \rangle$, y σ^2 cualquier número positivo. Es decir, que no existe diferencia formal entre ambos estudios. Recíprocamente, la familia de distribuciones considerarada en el modelo de Análisis de la Varianza (problema 3) puede expresarse mediante coeficientes de regresión. Efectivamente, si en el tercer estudio definimos la matriz

$$\mathtt{X} = (1_{30} | \mathtt{v}_1 | \mathtt{v}_2) \tag{1.1}$$

ésta posee termino independiente y es tal que $V = \langle \mathtt{X} \rangle$. Definiendo β como las coordenadas de μ respecto de la base $\mathtt{X}$ tendremos

$$Y \sim N_{30}(\mathtt{X}\beta, \sigma^2 \mathtt{Id}). \tag{1.2}$$

Además, la hipótesis de igualdad de medias se traduce en la nulidad de los coeficientes de regresión correspondientes a los vectores $\mathtt{v}_1$ y $\mathtt{v}_2$ (todos salvo el del término independiente). Los vectores $\mathtt{v}_1$ y $\mathtt{v}_2$ que hemos construido desempeñan el mismo papel que las observaciones de las variables explicativas en Regresión, e indican únicamente a qué grupo pertenece cada individuo: un valor (1,0) indica que el paciente se ha tratado con el medicamento A, (0,1) corresponde a B y (0,0) al placebo. Estas columnas se denominarán observaciones de las variables ficticias. Así pues, un problema de análisis de la varianza (comparación de grupos) puede entenderse como un caso de regresión respecto a variables ficticias[4]

Análisis de la covarianza

El cuarto estudio es una mezcla entre los problemas de relación entre variables (peso y altura) y de diferenciación de grupos (contexturas). El objetivo en nuestro

[4]El hecho de que la variable respuesta no sea explicada por las variables ficticias (de asignación a grupo) equivale a que los grupos no se diferencias en la variable respuesta.

caso es establecer una relación diferente para cada contextura. En otras ocasiones se trata de un problema de comparación de grupos en el que se introduce una variable adicional que funciona como explicativa para controlar una posible fuente de variabilidad de la variable respuesta, de manera que queden más patentes las diferencias de los grupos respecto de la misma. En todo caso, la variable que actúa como explicativa se denomina *covariable*, mientras que la distingue entre grupos se denomina *factor*. Cuando el modelo cuenta exclusivamente con covariables se denomina modelo de regresión; cuando cuenta exclusivamente con factores se denomina de análisis de la varianza; cuando se mezclan factores y covariables, como es este caso, se denomina análisis de la covarianza.

Consideraremos el peso como variable respuesta y y la estatura como covariable z. Podemos descomponer el vector Y de manera análoga al estudio anterior. Lo mismo podemos hacer con la covariable $\mathtt{Z}$ (en este caso se trata de un vector, aunque pudiera ser perfectamente una matriz). También podemos construir de igual forma los vectores $\mathtt{v}_1$, $\mathtt{v}_2$ y $\mathtt{v}_3$. Supondremos que, para cada contextura, tenemos un modelo de Regresión lineal entre, todos independientes y con la misma varianza, es decir, que cada Y_{ij} se expresa de la forma

$$Y_{ij} = \beta_{0i} + \beta_{1i}\mathtt{Z}_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

siendo todos los errores ε_{ij} independientes. Nótese que, si el signo * denota el producto de dos vectores componente a componente y consideramos el subespacio lineal V de $\mathbb{R}^{30}$ generado de la forma

$$V = \langle \mathtt{v}_1, \mathtt{v}_2, \mathtt{v}_3, \mathtt{v}_1 * \mathtt{Z}, \mathtt{v}_2 * \mathtt{Z}, \mathtt{v}_3 * \mathtt{Z} \rangle$$

entonces, el modelo puede expresarse mediante

$$Y \sim N_{30}(\mu, \sigma^2 \mathtt{Id}),$$

donde μ es cualquier vector de V y σ^2 cualquier número positivo. Consideremos entonces la base de V

$$\mathtt{X} = (1_{30} \mid \mathtt{Z} \mid \mathtt{v}_1 \mid \mathtt{v}_2 \mid \mathtt{v}_1 * Z \mid \mathtt{v}_2 * Z)$$

y sea β el vector de coordenadas de μ respecto de $\mathtt{X}$. De esta manera, el modelo puede expresarse también mediante

$$Y \sim N_{30}(\mathtt{X}\beta, \sigma^2 \mathtt{Id}),$$

siendo β cualquier vector de $\mathbb{R}^6$ y σ^2 cualquier número positivo. Como vemos, podemos considerar nuevamente un modelo de Regresión Lineal con un término independiente, una variable explicativa denominada covariable, dos variables ficticias de

asignación a grupo y los productos de éstas con la covariable. Los coeficientes de estos ultimos se denominan *interacciones*. Veamos el porqué: si se denota

$$\beta = (\alpha, \gamma, \alpha_1, \alpha_2, \gamma_1, \gamma_2)'$$

tenemos las siguientes correspondencias

$$\begin{array}{rcl} \alpha & = & \beta_{03} \\ \gamma & = & \beta_{13} \\ \alpha_1 & = & \beta_{01} - \beta_{03} \\ \alpha_2 & = & \beta_{02} - \beta_{03} \\ \gamma_1 & = & \beta_{11} - \beta_{13} \\ \gamma_2 & = & \beta_{21} - \beta_{23} \end{array}$$

Por lo tanto, que las interacciones γ_1 y γ_2 sean nulas equivale a que las pendientes de las tres rectas sean idénticas, es decir, que la relación entre el incremento de la estatura y el del peso es la misma para las tres contexturas. En términos estadísticos diremos que peso y contextura no interaccionan mutuamente. La aceptación de dicha hipótesis (perfectamente contrastable en nuestro modelo) conduciría a un nuevo modelo más simple en el que se considerarían sólo las cuatro primeras columnas de $\mathbf{X}$. En dicho modelo sin interacción, cada observación Y_{ij} se expresa mediante

$$Y_{ij} = \beta_{0i} + \gamma \mathbf{Z}_{ij} + \varepsilon ij, \qquad \varepsilon_{ij} \sim N(0, \sigma^2),$$

siendo los errores ε_{ij} independientes, y en el mismo puede contrastarse la hipótesis inical

$$H_0 : \alpha_1 = \alpha_2 = 0 \ ^5$$

Su veracidad equivale a la igualdad de las tres rectas. Por contra, su falsedad quiere decir que, dado un valor concreto de la covariable estatura, tenemos, por término medio, distintos pesos en función de la contextura.

¿Rango completo o rango no completo?

Vamos a formalizar el tercer problema de una forma diferente. Supongamos que cada observación Y_{ij} descompone de la forma

$$Y_{ij} = \theta + \alpha_i + \varepsilon_{ij} \tag{1.3}$$

[5]Esta hipótesis puede contrastarse también en el modelo general, pero es aquí, en el modelo reducido, donde goza de mayor interés, según se ve a continuación.

donde $\varepsilon_{ij} \sim N(0, \sigma^2)$ y son independientes. Se supone que estos parámetros tienen una significado muy claro para nosotros: el parámetro θ representa aquello que tienen en común los tres medicamentos; el parámetro α_1 expresa la influencia particular que ejerce el medicamento A sobre la variable respuesta; lo mismo puede decirse de α_2 y α_3 en relación con los medicamentos B y C, respectivamente; lo dicho hasta ahora afecta exclusivamente a las medias, pues se supone que para cada medicamento existe una variabilidad de la respuesta explicada por el azar y cuantificada por σ^2, que es idéntica en los tres casos.

Desde el punto de vista formal, si se denota $\beta = (\theta, \alpha_1, \alpha_2, \alpha_3)'$, el modelo considerado es

$$Y \sim N_{\mathtt{n}}(\mathtt{X}\beta, \sigma^2 \mathtt{Id})$$

donde

$$\mathtt{X} = \left(\begin{array}{c|ccc} 1_{10} & 1_{10} & 0_{10} & 0_{10} \\ 1_{10} & 0_{10} & 1_{10} & 0_{10} \\ 1_{10} & 0_{10} & 0_{10} & 1_{10} \end{array}\right)$$

Respecto al modelo considerado en (4.13), la única diferencia estriba en un cambio en el parámetro. Efectivamente, mientras que en el caso anterior el vector β estaba compuesto por las coordenadas de la media respecto a la base (1.1) de V, en esta ocasión se trata de las coordenadas respecto a un nuevo sistema generador de V.

Por lo tanto, si entendemos modelo estadístico según la definición (9.31), se trata del mismo modelo que se consideró en (1.2). Sólo si nos acogemos a la definición de modelo estadístico que se expone en el capítulo 7 podemos reconocer una diferencia formal entre ambos modelos. La particularidad de éste radica en que la matriz $\mathtt{X}$ no es de rango completo, es decir, sus columnas no constituyen un sistema linealmente independiente. En consecuencia, dado un vector $\mathtt{v} \in \langle \mathtt{X} \rangle$, la ecuación $\mathtt{v} = \mathtt{X}\mathtt{b}$ presentará infinitas soluciones. En otras palabras, el parámetro β no está bien determinado y sólo podemos especificar una solución concreta si imponemos una restricción adicional, como puede ser $\sum_i \alpha_i = 0$ ó $\alpha_3 = 0$. Por cierto, que esta última conduciría al mismo modelo considerado en (1.1).

La distinción entre modelo de rango completo y modelo de rango no completo es muy sutil, por no decir inexistente. En todo caso, cualquier modelo de rango no completo se convierte en automáticamente en otro de rango completo cuando se imponen las oportunas restricciones. Podría decirse que dicha imposición conlleva una pérdida de generalidad. No obstante en el modelo de rango no completo se parte, como hemos dicho, de una matriz $\mathtt{X}$ cuyas columnas pueden ser en principio, linealmente dependientes, de ahí que se requiera del uso de inversas generalizadas para resolver un sistema de ecuaciones denominadas normales, lo cual introduce una sensible

complicación en la teoría. Ello no debería ser óbice para nosotros, dado los instrumentos de los que disponemos. No obstante, las soluciones a las ecuaciones normales constituyen una subvariedad afín, por lo cual, aunque se prescinde de restricciones previas sobre los parámetros del modelo, es necesaria la imposición de restricciones posteriores arbitrarias para encontrar una solución particular a dichas ecuaciones.

La diferencia no es pues de tipo formal sino de enfoque: cuando se plantea un modelo del tipo (1.3) sin ninguna restricción de los parámetros centramos nuestro interés en el significado intuitivo de los mismos y nos abandonamos, por así decirlo, a un algoritmo preestablecido para la obtención de soluciones concretas. El otro punto de vista se basa en tener claro a qué subespacio V pertenece la media μ pues β es sólo un parámetro contingente que expresa las coordenadas de μ respecto a cierta base $\mathbf{X}$ y que, en consecuencia, debe verificar de antemano una serie de restricciones de tipo lineal. Searle (1971) y Seber (1977), por citar referencias clásicas de sobras conocidas, entienden el Modelo Lineal desde el primer punto de vista, mientras que un claro exponente de la segunda visión es, sin duda, Arnold (1981). Estas dos tendencias no son contradictorias pero utilizan, como vemos, técnicas aparentemente distintas.

A nuestro entender, el uso de coordenadas tiene a su favor que proporciona algoritmos precisos a la hora de implementar los distintos métodos. Efectivamente, nosotros podemos entender perfectamente el concepto de subespacio lineal y sabemos que éste puede caracterizarse mediante una base o sistema generador, una matriz en definitiva. Pero sólo esto último es lo que, hablando coloquialmente, puede entender un ordenador. Se trata de un distinción que, lejos de ser ser de índole teórica, tiene un carácter eminentemente práctico. Otro punto a su favor podría ser una más que discutible ganancia en generalidad, dado que al no suponer que $\mathbf{X}$ sea de rango completo aspira a resolver cualquier ecuación lineal planteada en un contexto estadístico, lo cual permite afrontar como casos particulares los análisis de regresión, de la varianza y de la covarianza.

El planteamiento basado en V o en una base de V (con rango completo), asume cierta pérdida de generalidad para afrontar únicamente los análisis estadísticos anteriormente mencionados. Aquí, el uso de una herramienta fundamental del Algebra Lineal, como es la proyección ortogonal sobre V, permite establecer una teoría muy elegante y facilita una justificación profunda de los estimadores y tests de hipótesis obtenidos. Sin embargo, desde este punto de vista no pueden afrontarse modelos como (1.3) sin preocuparse de imponer previamente, ni problemas de regresión lineal donde el número de variables explicativas sea mayor que el número de individuos analizados, aunque conviene recalcar que esta situaciones no resulta en absoluto desable[6].

[6]Un problema de regresión con un demasiadas variables explicativas convendría afrontarlo me-

No obstante, un clara deficiencia del punto de vista de en Arnold(1981) podría quedar patente en algunos casos complejos del análisis de la varianza. Efectivamente, en estos estudios, es el parámetro en sí y no la media de la distribución lo que realmente interesa, pues el primero permite aislar las influencias que los distintos factores tienen en la segunda. El parámetro se define como la solución a un sistema de ecuaciones lineales no determinado, por lo que se precisa de la imposición de una serie de restricciones, como ya hemos dicho. No obstante, en el caso equilibrado vienen dadas de manera natural, lo cual conduce a un modelo de rango completo. Sin embargo, en los diseños no equilibrados con varios factores, no existen *a priori* argumentos para imponer una familia concreta de restricciones, de ahí que pueda resultar más coherente enfocar estos diseños desde un punto de vista más general: el Modelo Lineal Normal de Rango no Completo.

Así pues, hemos de decantarnos por la generalidad de planteamiento con coordenadas o por la elegancia del planteamiento sin coordenadas. Desde nuestro punto de vista, entendemos que la ganancia en generalidad del primer planteamiento es exigua en relación con la complicación que conlleva. El Modelo Lineal, según se entiende en Arnold (1981), es, en nuestra opinión, una de las teorías más redondas que pueden encontrarse en la Estadística clásica[7] y permite resolver la mayoría de los problemas *lineales* que se plantean en la práctica (regresión-correlación, análisis de la varianza y covarianza). Decimos esto teniendo en cuenta lo siguiente: primeramente, se puede considerar natural el imponer que un diseño de análisis de la varianza sea equilibrado, en cuyo caso disponemos de una solución directa del problema a partir de una serie de restricciones naturales; segundo, aunque en diseños no equilibrados se exige la imposición previa de restricciones artificiales sobre los parámetros, parametrizar el modelo mediante una matriz de rango no completo exigirá igualmente la imposición de restricciones artificiales, aunque en una fase posterior; tercero, resulta también razonable que el número de variables explicativas en un modelo de regresión-correlación sea menor que el número de unidades experimentales utilizadas en el estudio. Por ello, consideramos Arnold (1981) como referencia principal. No obstante, aunque haya quedado relegado a un segundo plano por las razones expuestas, dedicaremos un capítulo al denominado Modelo Lineal de Rango no Completo, para que el lector pueda valorar por sí mismo los argumentos expuestos anteriormente y optar por el procedimiento que considere oportuno.

diante técnicas de análisis de datos funcionales (Ferraty, Vieu (2006)).

[7]Entiéndase la distinción entre Probabilidad y Estadística.

1.3. Conclusión

A partir de todo lo dicho anteriormente, concluimos que los problemas de regresión y análisis de la varianza y covarianza, ya sea con rango completo o no completo, se formalizan mediante un mismo modelo que coincide, a su vez, con el modelo que se obtiene al condicionar sobre las variables explicativas en el modelo de Correlación. Ese modelo al que nos estamos refiriendo se denomina Modelo Lineal Normal, y viene dado por un vector aleatorio n-dimensional, Y que sigue una distribución $N_n(\mu, \sigma^2\mathtt{Id})$. Cuando no se suponga la normalidad, hablaremos de Modelo Lineal (a secas). No se establece ninguna restricción para la varianza σ^2, pero sí se impone una condición de tipo lineal a la media: que pertenezca a un subespacio lineal V de $\mathbb{R}^n$. Si $\mathtt{X}$ denota una matriz cuyas columnas constituyen un sistema generador de V, para cada $\mu \in V$ existirá algún vector β tal que $\mu = \mathtt{X}\beta$. En el caso de que $\mathtt{X}$ sea de rango completo, β será único. Por ello, dada $\mathtt{X}$, el modelo puede expresarse con la ayuda del parámetro β en lugar de μ.

El estudio de este modelo desde el punto de vista teórico es el objeto del capítulo 3. Posteriormente se aplicarán los resultados obtenidos a los diferentes problemas que formaliza.

Capítulo 2

Distribución Normal Multivariante

En este capítulo abordamos el estudio de una distribución que que viene a generalizar la conocida distribución normal unidimensional y que, por ende, desempeña un papel central en estadística multivariante . Se hará especial hincapié en la estrecha relación existente entre la normalidad y la linealidad, hipótesis fundamentales en nuestra teoría. Precisamente por ser el punto de partida del Modelo Lineal Normal, se estudiará con especial atención la distribución normal multivariante esférica, así como otras distribuciones obtenidas a partir de la misma, como son la χ^2, F-Senedecor, t-Student o Beta. Recordamos que la correcta compresión de este capítulo exige el conocimiento de diversas definiciones y resultados que se hayan en el Apéndice. Al final del capítulo se incluyen una serie de problemas referentes tanto a los contenidos del mismo como del mencionado Apéndice.

2.1. Definición y principales propiedades

Dados un vector $\mu \in \mathbb{R}^n$ y una matriz $\Sigma \in \mathcal{M}_{n\times n}$ simétrica y semidefinida positiva, se dice que un vector aleatorio $Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}^n$ sigue un modelo de distribución normal n-variante con media μ y matriz de covarianzas Σ (se denota $Y \sim N_n(\mu, \Sigma)$) cuando su correspondiente función característica es la siguiente

$$\varphi_Y(t) = \exp\left\{\mathbf{i}t'\mu - \frac{1}{2}t'\Sigma t\right\}, \quad t \in \mathbb{R}^n.$$

Un vector de este tipo puede construirse explícitamente como sigue: si Σ diagonaliza según el teorema 9.4 mediante

$$\Sigma = \Gamma\Delta\Gamma',$$

consideramos Z_i, $i = 1, \ldots, n$, independientes y con distribuciones normales de media 0 y varianza el elemento i-ésimo de la diagonal de Δ, δ_i^2, respectivamente. Si Z denota el vector aleatorio $(Z_1, \ldots, Z_n)'$, se tiene entonces que

$$Y = \mu + \Gamma Z \tag{2.1}$$

sigue la distribución deseada. Efectivamente, se verifica

$$\begin{aligned} \varphi_Z(t) &= \prod_{i=1}^{n} \varphi_{N(0,\delta_i^2)}(t_i) = \prod_{i=1}^{n} \exp\left\{-\frac{1}{2} t_i^2 \delta_i^2\right\} \\ &= \exp\left\{-\frac{1}{2} t' \Delta t\right\}. \end{aligned}$$

Luego,

$$\begin{aligned} \varphi_Y(t) &= \exp\{\mathbf{i} t'\mu\} \varphi_Z(\Gamma' t) = \exp\left\{\mathbf{i} t'\mu - \frac{1}{2} t' \Gamma \Delta \Gamma' t\right\} \\ &= \exp\left\{\mathbf{i} t'\mu - \frac{1}{2} t' \Sigma t\right\}. \end{aligned}$$

Dado que $\mathtt{E}[Z] = 0$ y $\mathtt{Cov}[Z] = \Delta$, se sigue de (9.11) que una distribución $N_n(\mu, \Sigma)$ tiene por media μ y por matriz de varianzas-covarianzas Σ. También es inmediato comprobar que presenta la siguiente función generatriz, bien definida en todo $\mathbb{R}^n$:

$$g_Y(t) = \exp\left\{t'\mu - \frac{1}{2} t' \Sigma t\right\}, \quad t \in \mathbb{R}^n.$$

En consecuencia, existen los momentos de cualquier orden de la distribución, que pueden calcularse mediante las sucesivas derivadas parciales de g en 0.

Es bien conocido que la normalidad en dimensión 1 se conserva ante transformaciones afines, es decir, que si a una distribución normal se le aplica una homotecia y una traslación, la distribución resultante sigue siendo normal. Operando con las funciones características podemos obtener de manera trivial el siguiente resultado que generaliza al anterior en el caso multivariante.

Proposición 2.1.

Dados $Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}^n$, tal que $Y \in N_n(\mu, \Sigma)$, $A \in \mathcal{M}_{n\times m}$ y $b \in \mathbb{R}^m$, se verifica

$$AY + b \sim N_m(A\mu + b, A\Sigma A').$$

De la proposición 2.1 se deduce que las n componentes de una normal n-variante son todas normales. Sin embargo, no podemos garantizar, en general, que n componentes normales configuren conjuntamente un vector n-normal, cosa que si sucede si

las componentes son independientes. Más adelante veremos un curioso contraejemplo. El siguiente resultado supone una interesante caracterización de la distribución normal multivariante.

Proposición 2.2.

Un vector aleatorio n-dimensional Y de media μ y matriz de varianzas-covarianzas Σ sigue una distribución n-normal si y sólo si la variable aleatoria real $u'X$ sigue una distribución $N(u'\mu, u'\Sigma u)$, para cada $u \in \mathbb{R}^n \backslash \{0\}$.

Demostración.
Supongamos que $u'Y$ sigue una distribución normal unidimensional, para cada $u \in \mathbb{R}^n \backslash \{0\}$, y sea $t \in \mathbb{R}^n$. Entonces

$$\varphi_Y(t) = \varphi_{tY}(1) = \varphi_{N(t'\mu, t'\Sigma t)}(1) = \exp\left\{ \mathbf{i} t'\mu - \frac{1}{2} t'\Sigma t \right\},$$

con lo cual acaba la prueba.

∎

Queremos decir, por lo tanto, que la distribución es n-normal cuando al proyectar sobre cualquier dirección de $\mathbb{R}^n$ obtenemos una normal en dimensión 1. Por otra parte, el siguiente resultado garantiza la equivalencia entre incorrelación e independencia bajo la hipótesis de normalidad multivariante.

Proposición 2.3.

Si $Y = (Y_1' Y_2')'$ sigue un modelo de distribución normal en dimensión $n_1 + n_2$ y $\Sigma_{12} = 0$, entonces Y_1 e Y_2 son independientes.

Demostración.
Efectivamente, supongamos que Y_1 e Y_2 son incorreladas. Entonces, la función característica de Y es la siguiente

$$\begin{aligned} \varphi_Y \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} &= \exp\left\{ \mathbf{i}(t_1' \mathbb{E}[Y_1] + t_2' \mathbb{E}[Y_2]) - \frac{1}{2}(t_1', t_2') \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \right\} \\ &= \varphi_{Y_1}(t_1) \cdot \varphi_{Y_2}(t_2). \end{aligned}$$

Teniendo en cuenta las propiedades fundamentales de la función característica, se acaba la prueba.

∎

Nótese que esta propiedad puede extenderse trivialmente a cualquier colección (no necesariamente dos) de subvectores de un vector aleatorio normal multivariante,

en particular, a cualquier subconjunto de componentes del mismo. Queremos decir lo siguiente: si $Y_{n(1)}, \ldots, Y_{n(k)}$ son componentes incorreladas de un vector n-normal, entonces son también independientes.

Con frecuencia suele suponerse que la matriz de covarianzas Σ de la normal es estrictamente definida positiva, es decir, no singular. En caso contrario se dice que la normal es degenerada, es decir, que está *sobredimensionada*[1] . En ese caso, estará contenida en una subvariedad afín de dimensión $n-1$, por lo que no estará dominada por la medida de Lebesgue en $\mathbb{R}^n$. En el caso no degenerado, tendrá sentido hablar de su densidad respecto a dicha medida.

Proposición 2.4.

Si $Y \sim N_n(\mu, \Sigma)$ con $\Sigma > 0$, entonces admite la siguiente densidad respecto a la medida de Lebesgue:

$$f(\mathtt{y}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left\{-\frac{1}{2}(\mathtt{y}-\mu)'\Sigma^{-1}(\mathtt{y}-\mu)\right\}, \quad \mathtt{y} \in \mathbb{R}^n. \tag{2.2}$$

Demostración.

Consideremos una construcción del tipo (2.1). En ese caso, el vector aleatorio Z admite una función de densidad f_Z respecto a la medida de Lebesgue al ser sus componentes independientes. Concretamente[2],

$$\begin{aligned} f_Z(\mathtt{z}) &= \prod_{i=1}^n f_{Z_i}(\mathtt{z}_i) = \prod_{i=1}^n f_{N(0,\delta_i^2)}(\mathtt{z}_i) \\ &= \frac{1}{(2\pi)^{n/2}\prod_{i=1}^n \delta_i} \exp\left\{\frac{1}{2}\frac{\mathtt{z}_i}{\delta_i^2}\right\} \\ &= \frac{1}{\sqrt{(2\phi)^n |\Delta|}} \exp\left\{-\frac{1}{2}\mathtt{z}'\Delta^{-1}\mathtt{z}\right\}. \end{aligned}$$

Por otra parte, si consideramos la transformación

$$\varphi : \mathtt{y} \in \mathbb{R}^n \mapsto \Gamma'(\mathtt{y}-\mu) \in \mathbb{R}^n,$$

cuyo jacobiano es Γ', se sigue del Teorema del Cambio de Variables[3] que la función de densidad de Y es

$$f_Y(\mathtt{y}) = |\Gamma'| f_Z(\varphi(\mathtt{y})), \quad \mathtt{y} \in \mathbb{R}^n.$$

[1] El objetivo del análisis de componentes principales es, precisamente, encontrar la manera de dar a la distribución su *verdadera* dimensión.

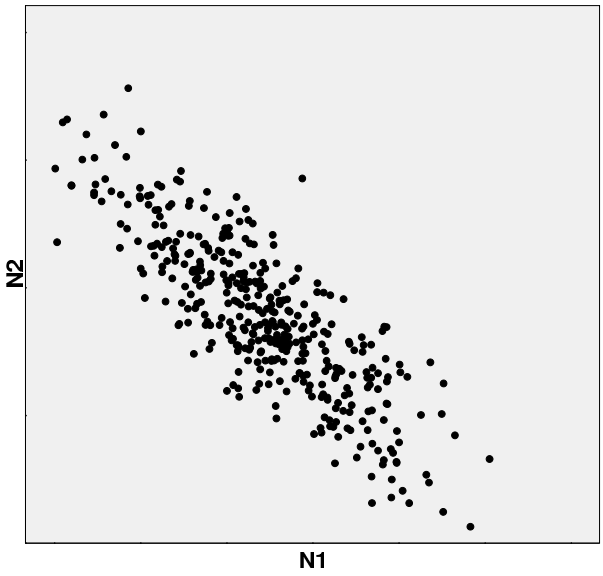
[2] Nótese que, al ser $\mathtt{rg}(\Sigma) = \mathtt{rg}(\Delta)$, $\delta_i > 0$ para todo $i = 1, \ldots, n$.

[3] Podemos encontrar una versión en Billingsley (1986), Th. 17.2.

Teniendo en cuenta que $|\Gamma'| = 1$ y $|\Delta| = |\Sigma|$, se tiene entonces

$$\begin{aligned} f_(\mathrm{y}) &= \frac{1}{\sqrt{(2\phi)^n|\Delta|}} \exp\left\{-\frac{1}{2}(\mathrm{y}-\mu)'\Gamma\Delta^{-1}\Gamma'(\mathrm{y}-\mu)\right\} \\ &= \frac{1}{\sqrt{(2\phi)^n|\Sigma|}} \exp\left\{-\frac{1}{2}(\mathrm{y}-\mu)'\Sigma^{-1}(\mathrm{y}-\mu)\right\}. \end{aligned}$$

El siguiente gráfico presenta una muestra aleatoria simple de tamaño 398 de un vector aleatorio $(N1, N2)'$ distribuido según un modelo 2-normal.



2.2. Normalidad y Linealidad

La siguiente propiedad establece una clara conexión entre los supuestos de normalidad y linealidad, arrojando luz sobre los modelos de Regresión y Correlación. Consideremos dos vectores aleatorios Y_1 e Y_2, de dimensiones n_1 y n_2, respectivamente. Construiremos una versión de la probabilidad condicional regular de Y_1 dado Y_2. bajo la hipótesis de $(n_1 + n_2)$-normalidad no degenerada de $Y = (Y_1', Y_2')'$. Supongamos que media y matriz de varianzas-covarianzas de Y descompone según (9.12) y consideremos loa parámetros α, β y $\Sigma_{11\cdot 2}$ definidos en (9.25), (9.26) y (9.14). Nótese que, en virtud del lema 9.7 y al ser $\Sigma > 0$, tiene sentido hablar de e $\Sigma_{11\cdot 2}$ y es definida positiva.

Proposición 2.5.
En las condiciones anteriores, se verifica

$$P^{Y_1|Y_2=\mathtt{y}_2} = N_{n_1}(\alpha + \beta\mathtt{y}_2, \Sigma_{11\cdot 2}), \qquad \forall \mathtt{y}_2 \in \mathbb{R}^{n_2}.$$

Demostración.
Consideremos la trasformación

$$\Phi : \begin{pmatrix} \mathtt{y}_1 \\ \mathtt{y}_2 \end{pmatrix} \in \mathbb{R}^{n_1+n_2} \mapsto \begin{pmatrix} \mathtt{Id} & -\beta \\ 0 & \mathtt{Id} \end{pmatrix} \begin{pmatrix} \mathtt{y}_1 \\ \mathtt{y}_2 \end{pmatrix} \in \mathbb{R}^{n_1+n_2},$$

cuyo jacobiano tiene por determinate 1. El Teorema del Cambio de Variables permite expresar la función de densidad de Y a partir de la de $\Phi \circ Y$ mediante

$$f_Y(\mathtt{y}) = f_{\Phi\circ Y}(\Phi(\mathtt{y})).$$

Si descomponemos en dos Φ en de acuerdo con las dimensiones de Y_1 e Y_2, se obtiene

$$\begin{pmatrix} \Phi_1 \\ \Phi_2 \end{pmatrix} \sim N_{n_1+n_2}\left(\begin{pmatrix} \alpha \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11\cdot 2} & 0 \\ 0 & \mathtt{Id} \end{pmatrix} \right).$$

Luego, se sigue de las proposiciones 2.1 y 2.3 que Φ_2 sigue el mismo modelo de distribución que Y_2 y es independiente de Φ_1. Por lo tanto, la densidad de Φ descompone en

$$f_\Phi(\phi_1, \phi_2) = f_{\Phi_1}(\phi_1) \cdot f_{\Phi_2}(\phi_2) \quad \forall (\phi_1, \phi_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}.$$

Dado que la densidad de la distribución condicional $P^{Y_1|Y_2=\mathtt{y}_2}$ se obtiene, según (9.28) mediante

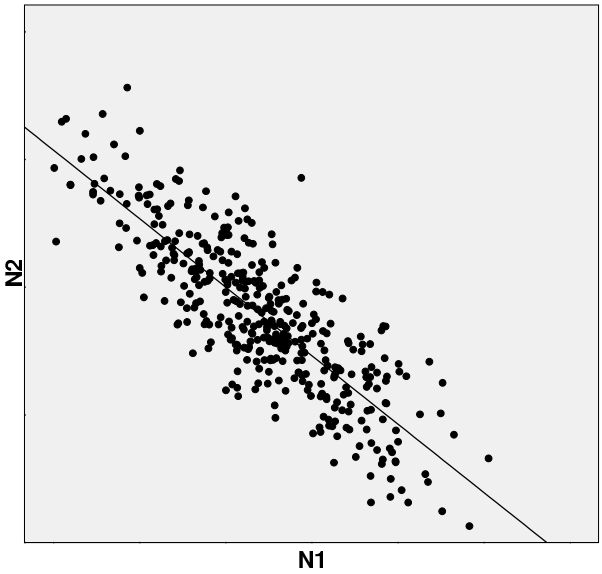
$$f_{Y_1|Y_2=\mathtt{y}_2}(\mathtt{y}_1) = \frac{f_Y(\mathtt{y}_1, \mathtt{y}_2)}{f_{Y_2}(\mathtt{y}_2)},$$

se sigue de lo anterior que

$$\begin{aligned} f_{Y_1|Y_2=\mathtt{y}_2}(\mathtt{y}_1) &= f_{\Phi_1}(\Phi_1(\mathtt{y}_1)) \\ &= \frac{1}{\sqrt{(2\pi)^n |\Sigma_{11\cdot 2}|}} \exp\left\{ -\frac{1}{2} (\mathtt{y}_1 - \alpha - \beta\mathtt{y}_2)' \Sigma_{11\cdot 2}^{-1} (\mathtt{y}_1 - \alpha - \beta\mathtt{y}_2) \right\}, \end{aligned}$$

con lo cual acaba la demostración. ∎

El siguiente gráfico ilustra el resultado anterior. La línea recta se aproximaría, dado que estamos trabajando con una muestra, a las esperanza condicional.



Podemos ir incluso algo más lejos. Para poder seguir la siguiente demostración se necesita tener presentes las propiedades fundamentales de la Esperanza Condicional.

Proposición 2.6.

En las condiciones anteriores, se verifica

$$Y_1 = \alpha + \beta Y_2 + \mathcal{E},$$

donde $\mathcal{E} \sim N_{n_1}(0, \Sigma_{11\cdot 2})$ y es independiente de Y_2.

Demostración.

Definamos $\mathcal{E} = Y_1 - (\alpha + \beta Y_2)$. En ese caso, se verifica, en virtud de (9.30), que

$$P^{\mathcal{E}|Y_2=\mathtt{y}_2} = \left(P^{Y_1|Y_2=\mathtt{y}_2}\right)^{g(\cdot,\mathtt{y}_2)},$$

donde

$$g(\cdot, \mathtt{y}_2) : \mathtt{y}_1 \in \mathbb{R}^{n_1} \mapsto \mathtt{y}_1 - (\alpha + \beta \mathtt{y}_2) \in \mathbb{R}^{n_1}.$$

Luego, de la proposición anterior se sigue que

$$P^{\mathcal{E}|Y_2=\mathtt{y}_2} = N_{n_1}(0, \Sigma_{11\cdot 2}), \qquad \forall \mathtt{y}_2 \in \mathbb{R}^{n_2}.$$

Al no depender del valor de $\mathtt{y}_2$ se concluye que $\mathcal{E}$ es independiente de Y_2 siendo su distribución marginal $N_{n_1}(0, \Sigma_{11\cdot 2})$.

∎

Así pues, entre dos vectores aleatorios que componen una distribución normal multivariante sólo es posible una relación lineal (o, mejor dicho, afín), salvo un error aleatorio independiente de media 0. Realmente, a esta conclusión podríamos haber llegado sólo con tener en cuenta que, si Y sigue una distribución norma multivariante, $Y_1-(\alpha+\beta Y_2)$ es incorrelada con Y_2 si, y sólo si, son independientes, como se demuetra en el apartado del Apéndice dedicado al concepto de Esperanza Condicional. Todo esto puede ilustrarse mediante un interesante ejemplo:

Ejemplo 2.1.
Consideremos tres variables aleatorias reales Y_1, Y_2 y X definidas sobre cierto espacio de probabilidad $(\Omega, \mathcal{A}, P)$ y verificando las siguientes condiciones:

(i) $Y_1 \sim N(0,1)$

(ii) $X \sim B(1, 0.5)$

(iii) Y_1 y X son independientes.

(iv) $Y_2 = (-1)^X Y_1$

Puede demostrarse sin dificultad que, en estas condiciones, $Y_2 \sim N(0,1)$ mientras que $P(Y_1 + Y_2 = 0) = 0.5$ luego, el vector aletorio $(Y_1, Y_2)'$ no puede ser 2-normal. Por lo tanto, se sigue de la proposición 2.1 que el vector aleatorio $(Y_1, Y_2)'$ no es 2-normal. Tenemos pues un ejemplo de vector aleatorio de componentes normales que, sin embargo, no es normal multivariante. Estas componentes no pueden ser por lo tanto independientes, cosa evidente en nuestro caso. De hecho, puede demostrarse sin dificultad que, si δ_z denota la distribución degenerada en un valor real z, entonces la distribución condicional de Y_1 dada Y_2 puede expresarse mediante

$$P^{Y_1|Y_2=y_2}(A) = \frac{1}{2}(\delta_{y_2} + \delta_{-y_2})$$

Basta pues aplicar (9.29) para demostrar que Y_1 e Y_2 son incorreladas. Tenemos pues un ejemplo de dos variables dependientes pero sin relación lineal. Por supuesto, ello sólo es posible si el vector que componen no es normal.

En general, $\Sigma_{11\cdot 2}$, que es la matriz de varianzas-covarianzas de $Y_1-(\alpha+\beta Y_2)$ o, lo que es lo mismo, de la distribución condicional de Y_1 dado Y_2 (no depende del valor concreto que tome Y_2), se denomina en el Apéndice matriz de varianzas-covarianzas parciales de las componentes de Y_1 dado Y_2, y se interpreta en este caso como la parte de la matriz de varianzas-covarianzas de Y_1 no explicada por Y_2. Si se denota por Y_{1i},

$i = 1, \ldots, n_1$, a las componentes de Y_1, se verifica, en virtud de la proposición 2.3, que un valor nulo de la covarianza parcial de Y_{1i} con Y_{1j} dado Y_2 equivale la independencia condicional ente Y_{1i} e Y_{1j} dado Y_2, y un valor nulo de la varianza parcial de Y_{1i} dado Y_2 supone una dependencia funcional de Y_{1i} respecto a Y_2. Ello significa, hablando en términos intuitivos, que conociendo el resultado de Y_2, el de Y_{1i} queda determinado sin margen alguno para el azar. En el caso $n_1 = 1$, obtenemos

$$Y_1 = \alpha + \beta Y_2 + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2_{11\cdot 2}),$$

donde

$$\sigma^2_{11\cdot 2} = \sigma^2_1 - \Sigma_{12}\Sigma^{-1}_{22}\Sigma_{21} = \sigma^2_1(1 - \rho^2_{12}).$$

Según hemos dicho anteriormente, una varianza parcial $\sigma^2_{11\cdot 2}$ nula, equivale a una dependencia funcional de Y_1 respecto a Y_2, y ρ^2_{12} puede interpretarse como la proporción de varianza de Y_1 explicada por Y_2.

2.3. Normal esférica y distribuciones relacionadas

Volviendo a la expresión (2.2), correspondiente a la densidad de una distribución normal multivariante no degenerada podemos apreciar que la densidad en el punto $\mathtt{y}$ depende exclusivamente de la distancia de Mahalanobis a la media de la distribución, es decir,

$$\Delta^2(\mathtt{y}, \mu) = (\mathtt{y} - \mu)'\Sigma^{-1}(\mathtt{y} - \mu).$$

En esas condiciones, el lugar geométrico de los puntos con una misma densidad es un elipsoide, cuya centro coincide con la media μ y cuya forma viene determinada por la matriz de varianzas-covarianzas Σ. Concretamente, los ejes del elipsoide quedan determinados por una base de autovectores de Σ y su excentricidad por la relación existente entre los autovalores. De hecho, puede demostrarse que los elipsoides son esferas si y sólo si los autovalores de Σ son idénticos, es decir, si Σ es de la forma $\sigma^2\mathtt{Id}$, para algún $\sigma^2 > 0$, en cuyo caso, la densidad en $\mathtt{y}$ dependerá únicamente del cuadrado de su distancia euclídea a la media $\|\mathtt{y} - \mu\|^2$. Por esa razón, la distribución $N_n(\mu, \sigma^2\mathtt{Id})$ se denomina normal multivariante esférica.

Ésta será la distribución de partida en el Modelo Lineal Normal. De hecho, salvo contadísimas excepciones[4], será el único tipo de distribución normal multivariante a estudiar en nuestra teoría. Su función de densidad es pues la siguiente

$$f_Y(\mathtt{y}) = \frac{1}{(2\pi\sigma)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathtt{y} - \mu\|^2\right\}. \tag{2.3}$$

[4]Concretamente, cuando se haga referencia a las distribuciones de los estimadores $\hat{\mu}$ y $\hat{\beta}$.

De las proposiciones 2.1 y 2.3 se sigue sin dificultad que, dados un vector aleatorio Y n-normal esférico y dos matrices $A \in \mathcal{M}_{m\times n}$ y $B \in \mathcal{M}_{k\times n}$, los vectores AY y BY son independientes si y sólo si $A'B = 0$. Como consecuencia inmediata se obtiene la siguiente proposición.

Proposición 2.7.
Si $Y \sim N_n(\mu, \sigma^2\mathtt{Id})$ y V_1, V_2 son subespacios lineales de $\mathbb{R}^n$ ortogonales entre sí, entonces $P_{V_1}Y$ y $P_{V_2}Y$ son independientes.

La familia de distribuciones normales esféricas (con restricciones de carácter lineal para la media) poseen excelentes propiedades estadísticas. En primer lugar, son familias exponenciales, por lo que la función de verosimilitud cumple con todas las condiciones de regularidad[5] que puedan exigirse en diversos teoremas que mencionaremos en nuestra teoría; podremos obtener de manera muy sencilla un estadístico suficiente y completo, lo cual hará posible una máxima reducción por suficiencia; son invariantes ante diversos grupos de transformaciones bimedibles, cosa que permitirá obtener profundas reducciones por invarianza[6], de una de las cuales resulta, por ejemplo, el test `F`; el Principio de Máxima Verosimilitud será aquí de fácil aplicación, conduciendo a la obtención del Estimador de Máxima Verosimilitud y el Test de la Razón de Verosimilitudes, etc.

Es especialmente llamativa la invarianza ante rotaciones que presenta cualquier distribución normal esférica de media 0, hasta el punto de que esta propiedad está cerca de caracterizar dicha distribución. Efectivamente, si $\Gamma \in \mathcal{O}_n$ y $Y \sim N_n(0, \sigma^2)$, con $\sigma^2 > 0$, entonces ΓY sigue exactamente la misma distribución. En Bilodeau (1999) podemos encontrar la demostración de una especie de recíproco, debida a Maxwell-Hershell.

Proposición 2.8.
Todo vector aleatorio n-dimensional con componentes independientes e invariante por rotaciones es n-normal esférico de media 0. Concretamente, si Y_1 denota la primera componente del mismo, el parámetro σ que caracteriza la distribución se obtiene mediante

$$\sigma = -\ln \varphi_{Y_1}(1).$$

Por último, una propiedad de demostración trivial que será de utilidad en el estudio de la distribución χ^2. Realmente, la tesis de la proposición es cierta para cualquier distribución de media μ y matriz de varianzas-covarianzas $\sigma^2\mathtt{Id}$.

[5] Continuidad, derivabilidad...
[6] Ver Apéndice.

Proposición 2.9.
Si $Y \sim N_n(\mu, \sigma^2 \mathtt{Id})$, entonces $\mathtt{E}\left[\|Y\|^2\right] = n\sigma^2 + \|\mu\|^2$.

A continuación abordaremos un breve estudio de cuatro distribuciones directamente derivadas de la normal esférica: χ^2, F-Snedecor, Beta y t-Student. Un estudio más detallado de las mismas con todas las demostraciones que quedarán pendientes puede encontrarse, por ejemplo, en Nogales (1998). En primer lugar, la distribución χ^2 central con n grados de libertad (se denota χ^2_n) está definida sobre $\mathbb{R}^+$ mediante la siguiente función de densidad[7]

$$g_n(\mathtt{y}) = [\Gamma(n/2)2^{n/2}]^{-1} e^{-\mathtt{y}/2} \mathtt{y}^{\frac{n}{2}-1} I_{(0,+\infty)}(\mathtt{y}). \tag{2.4}$$

Puede probarse que tiene por media n y por varianza $2n$. La distribución χ^2 no central con m grados de libertad y parámetro de no centralidad $\lambda > 0$ (se denota $\chi^2_m(\lambda)$) se define mediante la función de densidad

$$\sum_{n=0}^{\infty} P_n(\lambda) g_{2n+1}(\mathtt{y}),$$

donde

$$P_n(\lambda) = \lambda^n \frac{\mathtt{e}^{-\lambda}}{n!}, \quad \mathtt{n} \in \mathbb{N}.$$

Se obtiene, por lo tanto, a partir de una composición (producto generalizado) entre una distribución de Poisson en $\mathbb{N}$ y la familia de las distribuciones χ^2_n, cuando n recorre $\mathbb{N}$. La distribución χ^2 central se corresponde con el caso $\lambda = 0$. En general, dado $\gamma > 0$, la expresión $Y \sim \gamma\chi^2_m(\lambda)$ debe entenderse como $\gamma^{-1}Y \sim \chi^2_n(\lambda)$.

Puede demostrarse que, si $Y_1, \ldots, Y_n$ son variables aleatorias reales independientes tales que

$$Y_i \sim N(\mu_i, \sigma^2),\ i = 1, \ldots, n,\ \sigma^2 > 0,$$

entonces

$$\sigma^{-2} \sum_{i=1}^{n} Y_i^2 \sim \chi^2_n \left(\sigma^{-2} \sum_{i=1}^{n} \mu_i^2 \right).$$

En otras palabras, considerar una colección de variables en esas condiciones equivale a considerar un vector aleatorio $Y \sim N_n(\mu, \sigma^2 \mathtt{Id})$, para algún $\mu \in \mathbb{R}^n$ y $\sigma^2 > 0$, y estamos afirmando que

$$\|Y\|^2 \sim \sigma^2 \chi^2_n \left(\frac{\|\mu\|^2}{\sigma^2} \right).$$

[7] Recordemos previamente que la función $\Gamma(\cdot)$ se define mediante $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\mathtt{e}^{-x}dx$, donde $\alpha > 0$.

En consecuencia, debemos entender el modelo χ^2 no central como la distribución del cuadrado de la distancia euclídea al origen de un vector aleatorio normal esférico. La norma euclídea al cuadrado es una función positiva de gran importancia en nuestra teoría, debida fundamentalmente a su presencia en la función de densidad (2.3). De hecho, ya comentamos que la densidad depende de $\mathtt{y}$ a través del cuadrado de su distancia euclídea a la media. Ello se traducirá en el uso de esta función y, en consecuencia, del modelo χ^2, a la hora de estimar el parámetro σ^2, de reducir por suficiencia y, también, cuando se efectúe una reducción por invarianza respecto al grupo de las rotaciones, según se sigue del teorema 9.12.

Hemos afirmado que el modelo χ^2 no central surge de la necesidad de considerar la norma euclídea de un vector normal esférico. No obstante, podemos generalizar un poco más. Si E es un subespacio vectorial de $\mathbb{R}^n$ y Γ es una base ortonormal del mismo, se verifica trivialmente que $\|P_E Y\|^2 = \|\Gamma' Y\|^2$ y que $\|P_E \mu\|^2 = \|\Gamma' \mu\|^2$. Por lo tanto, se tiene

$$\|P_E Y\|^2 \sim \sigma^2 \chi^2_{\mathtt{dim}E}\left(\frac{\|P_E \mu\|^2}{\sigma^2}\right). \tag{2.5}$$

Así pues, el grado de libertad de la distribución coincide con la dimensión del subespacio. Obtendremos una χ^2 central cuando $\mathtt{E}[Y]$ sea ortogonal al subespacio sobre el cual se proyecta Y. Por lo tanto y en general, se sigue de lo anterior junto con la proposición 2.9, que la media de una distribución χ^2 no central se obtiene mediante

$$\mathtt{E}\left[\sigma^2 \chi^2_m\left(\lambda/\sigma^2\right)\right] = m\sigma^2 + \lambda. \tag{2.6}$$

Dadas dos variables aleatorias reales X_1 y X_2, positivas e independientes, con distribuciones $\chi^2_n(\lambda)$, siendo $\lambda \geq 0$, y χ^2_m, respectivamente, se define la distribución F-Snedecor no central con (n, m) grados de libertad y parámetro de no centralidad λ (de denota por $F_{n,m}(\lambda)$), como la que corresponde a la variable $(n^{-1}X_1)/(m^{-1}X_2)$. Puede demostrarse que su función de densidad es la siguiente:

$$f_{n,m,\lambda}(\mathtt{y}) = \frac{n}{m}\mathtt{e}^{-\lambda}\sum_{k=0}^{\infty} c_k \frac{\lambda^k}{k!}\frac{\left(\frac{n}{m}\mathtt{y}\right)^{\frac{n}{2}-1+k}}{\left(1+\frac{n}{m}\mathtt{y}\right)^{\frac{n+m}{2}+k}} I_{(0,+\infty)}(\mathtt{y}), \tag{2.7}$$

donde 0^0 se entiende como 1 y

$$c_k = \frac{\Gamma\left(\frac{1}{2}(n+m)+k\right)}{\Gamma\left(\frac{1}{2}n+k\right)\Gamma\left(\frac{1}{2}m\right)}, \quad k \in \mathbb{N}.$$

La distribución $F_{n,m}(0)$ se denomina F-Snedecor central con (n, m) grados de libertad, y se denota por $F_{n,m}$. Su función de densidad es pues la siguiente:

$$f_{n,m}(\mathtt{y}) = \frac{n^{\frac{n}{2}} m^{\frac{m}{2}} \Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{\mathtt{y}^{\frac{n}{2}-1}}{(n\mathtt{y}+m)^{\frac{n+m}{2}}} I_{(0,+\infty)}(\mathtt{y}).$$

En nuestro caso, si $Y \sim N_n(\mu, \sigma^2 \mathtt{Id})$ y dados dos subespacios ortogonales $V_1, V_2 \subset \mathbb{R}^n$ tales que $\mu \in V_2^{\perp}$, se verifica que

$$\frac{\mathtt{dim}V_2}{\mathtt{dim}V_1} \frac{\|P_{V_1} Y\|^2}{\|P_{V_2} Y\|^2} \sim F_{\mathtt{dim}V_1, \mathtt{dim}V_2} \left(\frac{\|P_{V_1}\mu\|^2}{\sigma^2} \right). \tag{2.8}$$

Así pues, la distribución F de Snedecor resulta de relacionar las distancias al origen de dos proyecciones sobre sendos subespacio ortogonales. Si $\mu \in V_1^{\perp} \cap V_2^{\perp}$ tendremos una distribución F central. Una operación de este tipo surgirá al reducir por invarianza en el proceso de obtención del test `F`. Otras distribuciones íntimamente relacionadas con la F-Snedecor central son la Beta y la t-Student.

La distribución Beta de parámetros $\alpha, \beta > 0$, que se denotará por $B(\alpha, \beta)$, se define mediante la función de densidad[8]

$$f_{\alpha,\beta}(\mathtt{y}) = \mathtt{B}(\alpha, \beta)^{-1} \mathtt{y}^{\alpha-1} (1-\mathtt{y})^{\beta-1} I_{(0,1)}(\mathtt{y}).$$

Se trata pues de una distribución sobre el intervalo $(0, 1)$. Presenta un estrecha relación con la distribución F-Snedecor central. Concretamente, se verifica

$$X \sim F(n, m) \;\Leftrightarrow\; \left(1 + \frac{n}{m} X\right)^{-1} \sim B\left(\frac{m}{2}, \frac{n}{2}\right). \tag{2.9}$$

La distribución t de student central con n grados de libertad (se denota por t_n) es la que corresponde al cociente $X_1/\sqrt{X_2/n}$, donde $X_1 \sim N(0,1)$ y $X_2 \sim \chi^2_n$, siendo ambas independientes. Su densidad es la siguiente:

$$f_n(\mathtt{y}) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)\left(1 + \frac{\mathtt{y}^2}{n}\right)^{-\frac{n+1}{2}}}.$$

La distribución t_n puede considerarse un caso particular de la distribución F-Snedecor central, concretamente $F_{1,n}$ dado que es la única distribución simétrica cuyo cuadrado es una $F_{1,n}$. En ese sentido decimos que $t_n^2 = F_{1,n}$.

Por último, comentaremos dos resultados de carácter estadístico acerca de la las familias de distribuciones χ^2 central y F-Snedecor no central, que serán de utilidad en las secciones 2.2 y 2.3. En él apartado del apéndice dedicado a los contrastes de hipótesis se define el concepto experimento estadístico con razón de verosimilitudes monótona. Puede demostrarse fácilmente, teniendo en cuenta (2.4) y (2.7), que los experimentos estadísticos

$$(\mathbb{R}^+, \mathcal{R}^+\{\sigma^2 \chi^2_n : \sigma^2 > 0\}), \qquad (\mathbb{R}^+, \mathcal{R}^+\{F_{n,m}(\lambda) : \lambda \geq 0\})$$

[8]Recordar que la función B se define mediante $\mathsf{B}(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx$, donde $\alpha, \beta > 0$.

presentan razón de verosimilitudes monótona. Ello se traduce en que, en todos los casos, las colas de las distribuciones no centrales *pesan* más que las de las centrales. Dado que las centrales se corresponderán con la hipótesis nula y las no centrales con la alternativa, los tests que plantearemos para resolver nuestros contrastes consistirán en rechazar la hipótesis nula cuando la observación se halle en una cola.

Cuestiones propuestas.

1. Demostrar que $\mathtt{rg}(AB) \leq \text{mín}\{\mathtt{rg}(A), \mathtt{rg}(B)\}$, y que si A es invertible, entonces $\mathtt{rg}(AB) = \mathtt{rg}(B)$.

2. Demostrar el corolario 9.5 del Apéndice

3. Demostrar la proposición 9.16 del Apéndice.

4. Probar que, si la mediana de una variable aleatoria integrable X está bien definida, se trata de la constante k que minimiza la distancia $\int |X - k|\, dP$.

5. Sea $X = (X_1, X_2, X_3)'$ tiene distribución normal con vector de medias $\mu = (-1, 0, 1)'$ y matriz de varianzas-covarianzas

$$\Sigma = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 4 & 0 \\ -1 & 0 & 3 \end{pmatrix}.$$

   Hallar:

   a) La distribución marginal de X_1 y la del vector $(X_1, X_2)^t$.

   b) La distribución condicional de X_1 dado $X_2 = \mathtt{x}_2$, $X_3 = \mathtt{x}_3$.

   c) Los coeficientes de correlación $\rho_{12}, \rho_{13}, \rho_{23}$

   d) La distribución de $Z = 4X_1 - 6X_2 + X_3$ y la del vector $(Z_1, Z_2)'$ siendo $Z_1 = 2X_2 + X_3$ y $Z_2 = X_1 - 3X_2 + X_3$.

6. Sea $X = (X_1, X_2, X_3, X_4)'$ tiene distribución normal con vector de medias

$$\mu = \begin{pmatrix} 2 \\ 1 \\ -1 \\ -3 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0 & 1 & -1 \\ 0 & 2 & 1 & 1 \\ 1 & 1 & 3 & 0 \\ -1 & 1 & 0 & 2 \end{pmatrix}.$$

   Hallar:

a) La distribución marginal de los vectores $(X_2, X_1, X_3)'$ y $(X_1, X_4)'$.

b) La distribución condicional de $(X_1, X_4)'$ dado $X_2 = \mathtt{x}_2$, $X_3 = \mathtt{x}_3$.

c) La distribución de $Z = 2X_1 - 6X_3 + 4X_4$ y la del vector $(Z_1, Z_2)'$ siendo $Z_1 = X_1 - 3X_4 + 4X_2$ y $Z_2 = X_3 + 2X_2 - X_1 + 2X_4$.

7. Hallar la media de la distribución $\chi^2_n(\lambda)$ y la varianza de la distribución χ^2_n.

8. Sean Q_1 y Q_2 independientes tales que $Q_1 \sim \chi^2_{n_1}$ y $Q_2 \sim \chi^2_{n_2}$. Probar que $Q_1 + Q_2 \sim \chi^2_{n_1+n_2}$.

9. Demostrar que la matriz de covarianzas parciales muestral puede expresarse según (9.62). Compárese dicha expresión con la que aparece en (9.14).

10. Se ha definido la matriz de correlaciones parciales como la matriz de correlaciones correspondiente a la matriz de varianzas-covarianzas (9.14). Probar que, en el caso tridimensional,

$$\rho_{12\cdot 3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1-\rho^2_{13}}\sqrt{1-\rho^2_{23}}}.$$

Obtener una expresión análoga para el coeficiente de correlación parcial muestral.

11. Demostrar (9.25).

12. Probar que, dados una variable aleatoria real Y_1 y un vector aleatorio Y_2, el coeficiente de correlación múltiple ρ^2_{12} es la máxima correlación lineal simple al cuadrado entre Y_1 y una variable aleatoria de la forma $a + bY_2$, que se alcanza en cualquier $a \in \mathbb{R}$, y $b = \beta$ definido en (9.25).

13. Probar que los coeficientes de correlación múltiple probabilístico y muestral son invariantes ante traslaciones y cambios de escala (homotecias)

14. Probar que la matriz de covarianzas de las variables tipificadas coincide con la matriz de correlaciones.

15. Es bien conocido que dos vectores aleatorios X e Y son independientes si, y sólo si, para cada suceso A en la imagen de Y existe una versión constante de $P(Y \in A|X)$, en cuyo caso coincidirá con $P(Y \in A)$. Ello implica que la función constante $\mathtt{E}[Y]$ es versión de $\mathtt{E}[Y|X]$, es decir, que la función de X que más se aproxima a Y en el sentido L^2 es la propia esperanza de Y. Probar mediante un

contraejemplo que el recíproco no es cierto, es decir, que podemos encontrar un par de variables aleatorias (reales, por ejemplo), tales que $\mathtt{E}[Y|X]$ sea constante pero no sean independientes.

16. Indicar un ejemplo de dos variables aleatorias reales que presenten dependencia funcional pero cuyo coeficiente de correlación sea tan pequeño como se desee.

17. ¿Cómo interpretar el hecho de que dos variables aleatorias sean incorreladas? ¿Y si se trata de dos vectores de $\mathbb{R}^n$?

18. Considérese un vector aleatorio $(X, Y, Z)'$ siguiendo un modelo de distribución

$$N_3\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 2 & -2 \\ 2 & 4 & 0 \\ -2 & 0 & 4 \end{pmatrix}\right)$$

Obtener la matriz de varianzas-covarianzas parciales de $(Y, Z)'$ dada X. Demostrar entonces que, en general, la independencia entre dos variables Y y Z no implica la independencia condicional entre las mismas dada una tercera variable X. ¿Cuál es en este ejemplo el coeficiente de correlación lineal parcial entre Y y Z dada X? ¿Cómo interpretamos ese hecho?

Capítulo 3

Modelo lineal de rango completo

En este capítulo abordaremos el estudio del Modelo Lineal desde un punto de vista completamente teórico, atendiendo principalmente a los problemas de Estimación y Test de Hipótesis. Dado que algunas de las propiedades de los estimadores y tests de hipótesis son ciertas sin necesidad de asumir la normalidad de la familia de distribuciones, distinguiremos entre Modelo Lineal y Modelo Lineal Normal, según se incluya o no, respectivamente, dicho supuesto. Como indicamos en el capítulo 1, abordaremos nuestra teoría considerando en principio el parámetro media. No obstante, el capítulo 7 se dedicará a un enfoque distinto del problema y del propio concepto de modelo estadístico, consistente en considerar como parte esencial del mismo un parámetro que es solución a un sistema de ecuaciones lineales que debe satisfacer la media, que viene pues dado por una matriz $\mathtt{X}$, posiblemente de rango no completo. Es lo que se denomina, por lo tanto, modelo lineal de rango no completo. El título de el capítulo que abordamos aquí se ha escogido por su contraposición a dicho planteamiento. El modelo a considerar fue comentado en el capítulo 1, aunque lo especificaremos con más detalle. Por último, recordamos que para entender lo que se expone en este capítulo se precisa el conocer previamente buena parte del contenido del capítulo a anterior y del apéndice.

Un Modelo Lineal consiste en considerar, dados $\mathtt{n} \in \mathbb{N}$, y V un subespacio vectorial de $\mathbb{R}^{\mathtt{n}}$ de dimensión menor que $\mathtt{n}$, un vector aleatorio $\mathtt{n}$-dimensional Y de manera que $\mathtt{E}[Y]$ pertenezca a V y que las componentes de $Y - \mathtt{E}[Y]$ sean independientes e idénticamente distribuidas según un modelo de probabilidad real $\mathsf{P}_{\mathcal{E}}$ con varianza finita. Podemos expresarlo así

$$Y = \mu + \mathcal{E}, \qquad \mu \in V, \quad \mathcal{E} \sim \mathsf{P}_{\mathcal{E}}^{\mathtt{n}},\ \mathtt{E}[P_{\mathcal{E}}] = 0,\ \mathtt{var}[\mathsf{P}_{\mathcal{E}}] = \sigma^2,\ \sigma^2 > 0. \tag{3.1}$$

En esas condiciones, μ es la esperanza del vector aleatorio Y y σ^2 es la varianza de cada una des sus componentes.

Si suponemos que la función generatriz de momentos de $\mathsf{P}_{\mathcal{E}}$ está bien definida en un entorno de 0, entonces existirán los momentos μ_k de cualquier orden $k \in \mathbb{N}$ y, en ese caso, dichos momentos caracterizarán, junto con μ, la distribución. Entonces el modelo estadístico puede expresarse con la ayuda del siguiente parámetro en $\mathbb{R} \times \mathbb{R}^{\mathbb{N}}$

$$\tilde{\theta} = \big[\mu, (\mu_2, \mu_3, \mu_4, ...)\big],$$

con las restricciones $\mu \in V$ y $\mu_k \geq 0$ para todo k par. Dado que $\sigma^2 = \mu_2$, el modelo puede expresarse también mediante el parámetro

$$\theta = \big[(\mu, \sigma^2), \mu_3, \mu_4, \dots\big] \tag{3.2}$$

El Modelo Lineal se dice Normal cuando se supone en (3.1) que $\mathsf{P}_{\mathcal{E}}$ es normal, es decir,

$$Y = \mu + \mathcal{E}, \quad \mathcal{E} \sim N_{\mathtt{n}}(0, \sigma^2 \mathtt{Id}), \quad \mu \in V, \ \ \sigma^2 > 0. \tag{3.3}$$

En ese caso, se da la particularidad de que los parámetros μ y σ^2 bastan para caracterizar las distribuciones consideradas. El Modelo (3.3) puede expresarse también mediante

$$Y \sim N_{\mathtt{n}}(\mu, \sigma^2 \mathtt{Id}), \qquad \mu \in V, \sigma^2 > 0. \tag{3.4}$$

La distribución normal multivariante esférica de media 0 goza de diversas propiedades que facilitarán enormemente nuestro estudio. Entre otras cosas, es, como ya comentamos en el capítulo anterior, invariante ante cualquier rotación. Es más: cualquier vector aleatorio $\mathtt{n}$-dimensional de componentes independientes y media 0 es invariante por rotaciones si y sólo si es normal esférico. Decimos esto teniendo en cuenta el papel que desempeña el Principio de Invarianza en nuestra teoría.

Efectivamente, un argumento de invarianza ante rotaciones conduce a calcular el módulo al cuadrado de un vector normal esférico, es decir, a la distribución χ^2. Igualmente, la invarianza ante homotecias nos impulsa considerar un cociente entre distribuciones χ^2, es decir, una distribución F-Snedecor. De esta forma, mediante sucesivas reducciones por invarianza, obtendremos el denominado test $\mathtt{F}$, que será UMP-invariante. Si bien es lo más común justificar el test $\mathtt{F}$ mediante el Principio de Máxima Verosimilitud, el hecho de ser uniformemente el más potente entre los tests invariantes puede suponer, en este caso, un argumento más poderoso, dado que, bajo ciertas condiciones de regularidad que aquí se cumplen[1], el Test de la Razón de Verosimilitudes es siempre invariante.

Respecto a los supuestos asumidos en el modelo, ya hemos discutido en la Introducción sobre lo delicado del problema. De todas formas, la incorrelación de lo

[1] Lehmann (1986), pag. 341.

errores puede ser analizada mediante el test de Rachas; el supuesto de homocedasticidad (igualdad de varianzas)puede ser contrastado mediante el test de Barlett (que estudiaremos en este capítulo), que requiere del supuesto de normalidad junto con un diseño determinado del experimento estadístico; el de normalidad puede ser contrastado por diversos tests (Kolmogorov-Smirnov, Shappiro-Wilks, D'Agostino) de que no siempre pueden aplicarse con la potencia deseada; existe también un test de linealidad que requiere del cumplimiento de los demás supuestos. Así pues, no será fácil en la práctica contar con una sólida justificación de todos los supuestos considerados, por lo que convendrá analizar los residuos, así como el comportamiento asintótico del modelo. También debemos estar capacitados para efectuar transformaciones de variables que nos aproximen a las condiciones teóricas del modelo o incluso, a aplicar métodos alternativos más robustos. Todo ello se verá en capítulos posteriores.

3.1. Estimación

El Modelo Lineal se define, fundamentalmente, imponiendo una serie de condiciones sobre la esperanza μ de Y y su matriz de varianzas-covarianzas, que resulta depender únicamente de un escalar positivo σ^2. Por lo tanto, dedicaremos esta sección al estudio de las dos primeras componentes del parámetro (3.2), es decir, de los estimandos μ y σ^2.

Si analizamos detenidamente los problemas planteados en la Introducción, muy especialmente el tercero, llegaremos seguramente a la conclusión de que es μ el parámetro principal, mientras que σ^2 es un parámetro secundario que cuantifica el error o desviación en sentido cuadrático respecto a una situación determinista. Si fuera conocido, cosa poco factible en la práctica, todo resultaría mucho más fácil. Parámetros de este tipo son calificados de *ruido* en la literatura anglosajona y de *fantasmas* en la francesa.

Teniendo en cuenta que la media μ pertenece por hipótesis al subespacio V y que resulta más natural pensar que nuestra observación es próxima a la media que lo contrario (estamos aplicando el principio de máxima verosimilitud), cabe considerar el siguiente estimador de μ.

$$\hat{\mu} = P_V Y \tag{3.5}$$

Se trata pues del estimador μ que minimiza la distancia euclídea

$$\|Y - \hat{\mu}\|^2 \tag{3.6}$$

Es decir, se trata de una solución mínimo-cuadrática. Este concepto se define en (7.8). Respecto a σ^2, si tenemos en cuenta (9.19), cabría considerar, al menos en

principio, el estimador $\hat{\sigma}^2 = \mathtt{n}^{-1}\|Y - P_V Y\|^2$. No obstante y por razones que quedarán patentes más adelante, se denotará mediante $\hat{\sigma}^2$ a cualquier estimador positivo que sea proporcional al cuadrado de a distancia euclídea entre Y y el estimador propuesto para μ, es decir,

$$\hat{\sigma}^2 \propto \|Y - P_V Y\|^2. \tag{3.7}$$

Es inmediato comprobar que $\hat{\mu}$ es un estimador insesgado de μ. Veamos que sucede lo mismo con $\hat{\sigma}^2$ si consideramos el factor de proporcionalidad $\mathtt{n} - \mathtt{dim}V$. Necesitamos un lema previo de demostración trivial.

Lema 3.1.

Si X es un vector aleatorio m-dimensional cuyas componentes son de cuadrado integrable, entonces

$$\mathtt{E}[\|X\|^2] = \|\mathtt{E}[X]\|^2 + \mathtt{tr}\big(\mathtt{Cov}[X]\big).$$

Proposición 3.2.

En las condiciones del Modelo Lineal, el siguiente estadístico es un estimador insesgado de σ^2

$$\hat{\sigma}^{2,\mathrm{I}} = \frac{1}{\mathtt{n} - \mathtt{dim}V}\|Y - P_V Y\|^2$$

Demostración.

En primer lugar, si X es un vector aleatorio $\mathtt{n}$-dimensional, se verifica

$$\begin{aligned} \mathtt{E}\left[\|X\|^2\right] &= \mathtt{E}\left[\sum_{i=1}^{\mathtt{n}} X_i^2\right] = \sum_{i=1}^{\mathtt{n}} \mathtt{E}\left[X_i^2\right] = \sum_{i=1}^{\mathtt{n}} \big(E[X_i]^2 + \mathtt{var}[X_i]\big) \\ &= \|\mathtt{E}[X]\|^2 + \mathtt{tr}(\mathtt{Cov}[X]) \end{aligned}$$

En nuestro caso, teniendo en cuenta el lema anterior, tenemos lo siguiente

$$\begin{aligned} \mathtt{E}\left[\hat{\sigma}^2\right] &= \frac{1}{\mathtt{n} - \mathtt{dim}V}\mathtt{E}\left[\|P_{V^\perp} Y\|^2\right] = \frac{1}{\mathtt{n} - \mathtt{dim}V}\left[\|E[P_{V^\perp} Y]\|^2 + \mathtt{tr}\big(\mathtt{Cov}[P_{V^\perp} Y]\big)\right] \\ &= \frac{1}{\mathtt{n} - \mathtt{dim}V}(\sigma^2 \mathtt{tr} P_{V^\perp}) = \sigma^2 \end{aligned}$$

∎

En general, no estamos en condiciones de garantizar que $\hat{\mu}$ sea el estimador insesgado de mínima varianza. No obstante, sí que los es, en cierto sentido, respecto a la familia de estimadores lineales insesgados.

Dado $a \in \mathbb{R}^{\mathtt{n}}$, decimos que un estadístico real T es un estimador lineal insesgado de $a'\mu$ cuando es una aplicación lineal, es decir, existe $b \in \mathbb{R}^{\mathtt{n}}$ tal que $T(Y) = b'Y$, verificándose además que $\mathtt{E}[T] = a'\mu$. Ello es equivalente a que a y b tengan idénticas proyecciones ortogonales sobre V. Efectivamente, $b'\mu = a'\mu$ para todo $\mu \in V$ si y sólo

si $\langle a-b, v\rangle = 0$ para todo $v \in V$, es decir, si y sólo si $(a-b) \in V^{\perp}$ o, lo que es lo mismo, $P_V a = P_V b$.

El Teorema de Gauss-Markov prueba que $\hat{\mu}$ es óptimo respecto a esta clase de estimadores.

Teorema 3.3.

Para todo $a \in \mathbb{R}^{\mathtt{n}}$, $a'\hat{\mu}$ es el estimador lineal insesgado de mínima varianza de $a'\mu$. [2]

Demostración.
Dado que $a'\hat{\mu} = (P_V a)'Y$ y $\mathtt{E}\,[a'\hat{\mu}] = a'\mu$, el estimador es lineal insesgado. Su varianza es la siguiente

$$\mathtt{var}[a'\hat{\mu}] = \mathtt{var}[a'P_V Y] = a'P_V \mathtt{Cov}(Y) P_V' a = \sigma^2 a' P_V a$$

Sea $T(Y) = b'Y$ cualquier estimador lineal insesgado de $a'\mu$, es decir, tal que $P_V b = P_V a$. Entonces,

$$\begin{aligned} \mathtt{var}[b'Y] &= b'\mathtt{cov}(Y)b = \sigma^2\|b\|^2 = \sigma^2\|b - P_V b\|^2 + \sigma^2\|P_V b\|^2 \\ &= \sigma^2\|b - P_V b\|^2 + \sigma^2 a' P_V a \geq \sigma^2 a' P_V a = \mathtt{var}[a'\hat{\mu}], \end{aligned}$$

verificándose la igualdad si y sólo si $b = P_V a$, es decir, si y sólo si $T(Y) = a'\hat{\mu}$. ∎

Hasta ahora no hemos supuesto la normalidad de la familia de distribuciones. Si hacemos uso de dicha hipótesis podemos deducir interesantes propiedades de los estimadores considerados, entre ellas una más completa justificación teórica de los mismos, como veremos a continuación.

Proposición 3.4.

Bajo las condiciones del Modelo Lineal Normal, $\hat{\mu}$ y $\hat{\sigma}^{2,\mathrm{I}}$ son independientes y tales que

$$\hat{\mu} \sim N_{\mathtt{n}}(\mu, \sigma^2 P_V), \qquad [\mathtt{n} - \mathtt{dim}V]\hat{\sigma}^{2,\mathrm{I}} \sim \sigma^2 \chi^2_{\mathtt{n}-\mathtt{dim}V}$$

La demostración es consecuencia inmediata de las propiedades fundamentales de la distribución normal multivariante esférica. Concretamente, de las proposiciones 2.1, 2.7 y 2.5. Nótese que, en particular, podemos afirmar que la media aritmética de una muestra aleatoria simple de una distribución normal es independiente de su varianza muestral. También hemos de advertir que, excepto en el caso $V = \mathbb{R}^{\mathtt{n}}$, la distribución de $\hat{\mu}$ es degenerada pues está contenida en una subvariedad afín cuya dimensión es el rango de P_V, es decir, $\mathtt{dim}V$. Por lo tanto, no está dominada por

[2] En ese caso, se denota por ELIMV.

la medida de Lebesgue en $\mathbb{R}^{\mathtt{n}}$. Esta situación no ocurre cuando consideramos las coordenadas de μ respecto de una base $\mathtt{X}$ de V, como se verá más adelante. Sigamos con otra interesante propiedad de estos estimadores.

Teorema 3.5.
El estadístico $(\hat{\mu}, \hat{\sigma}^2)$ es suficiente y completo para el Modelo Lineal Normal[3].

Demostración.
Supongamos que $\hat{\sigma}^2 = \lambda \|Y - P_V Y\|^2$, para cierto $\lambda > 0$. Nuestro modelo estadístico está dominado por la medida de Lebesgue, siendo su función de verosimilitud la siguiente:

$$\mathcal{L}(\mathtt{y}; \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{\mathtt{n}}{2}} \exp\left\{-\frac{1}{2}\frac{\|\mathtt{y} - \mu\|^2}{\sigma^2}\right\}, \qquad \mathtt{y} \in \mathbb{R}^{\mathtt{n}},\ (\mu, \sigma^2) \in V \times \mathbb{R}^+ \quad (3.8)$$

Dada $X \in \mathcal{M}_{\mathtt{n}\times\mathtt{dim}V}$ cuyas columnas constituyen una base ortonormal de V, consideremos la biyección $\phi : V \times \mathbb{R}^+ \longrightarrow \mathbb{R}^{\mathtt{dim}V} \times \mathbb{R}^+$, definida mediante $\phi(v, c) = (\frac{1}{c} X'v, \frac{1}{c})$, para todo $v \in V$ y $c > 0$. De esta forma, podemos expresar la familia de distribuciones del modelo con la ayuda del parámetro $\theta = (\theta_1, \theta_2)$, definido como $(\theta_1, \theta_2) := \phi(\mu, \sigma^2)$. Así mismo, consideremos el estadístico $S : \mathbb{R}^{\mathtt{n}} \longrightarrow \mathbb{R}^{\mathtt{dim}V} \times \mathbb{R}^-$, definido mediante $S(\mathtt{y}) = (X'\mathtt{y}, -\frac{1}{2}\|\mathtt{y}\|^2)$. De esta forma, si consideramos la función

$$h(\phi_1, \phi_2) = \left(\frac{\phi_2}{2\pi}\right)^{\frac{\mathtt{n}}{2}} \exp\left\{-\frac{1}{2}\frac{\|\phi_1\|^2}{\phi_2}\right\},$$

se verifica que la función de verosimilitud definida como función del nuevo parámetro ϕ es la siguiente

$$\tilde{\mathcal{L}}(\mathtt{y}; \phi) = h(\phi) \exp\left\{\langle S(\mathtt{y}), \phi\rangle\right\},$$

de lo cual se deduce, teniendo en cuenta el teorema 9.18, que el estadístico S es suficiente y completo. Consideremos, por último, la transformación bimedible $\varphi : \mathbb{R}^{\mathtt{dim}V} \times \mathbb{R}^- \longrightarrow \mathbb{R}^{\mathtt{dim}V} \times \mathbb{R}^+$ definida mediante $\varphi(a, b) = (Xa, -\lambda(2b + \|a\|^2))$, para cada $a \in \mathbb{R}^{\mathtt{dim}V}$ y $b < 0$. Se verifica entonces que

$$\left(\hat{\mu}, \hat{\sigma}^2\right) = \varphi \circ S.$$

Luego, $(\hat{\mu}, \hat{\sigma}^2)$ es igualmente un estadístico suficiente y completo. ∎

El hecho de que el estadístico sea suficiente tiene muy importantes implicaciones. Efectivamente, considerar el modelo imagen (reducir) no supondrá perdida alguna

[3]Se entiende que $\hat{\sigma}^2$ denota cualquier estimador de la forma (3.7).

de información (en el sentido de Fisher) relativa a los parámetros. Por ello es razonable considerar el modelo reducido para afrontar cualquier problema de Inferencia acerca de dichos parámetros[4]. Siendo además completo, se tiene que el estadístico es suficiente minimal, lo cual se traduce en que no es posible reducir más sin perder información.

Veamos implicaciones más concretas en lo que respecta al problema de Estimación (el de Tests de Hipótesis se verá más adelante). Que un estimador sea el de Máxima Verosimilitud supone una justificación bastante convincente, especialmente por las propiedades asintóticas que conlleva[5]. Precisamente, se prueba en Lehmann (1983) que, bajo ciertas condiciones de regularidad que en nuestro caso se cumplen, el Estimador de Máxima Verosimilitud ha de ser función de cualquier estadístico suficiente. Luego, dicho estimador, que determinaremos a continuación, ha de expresarse exclusivamente a través de nuestro estadístico.

Por otra parte, otra propiedad muy deseable para cualquier estimador es que sea insesgado y, mucho mejor, el de mínima varianza entre todos los insesgados. Ya hemos encontrado estimadores insesgados de μ y σ^2 son insesgados. Demostrar que son de mínima varianza es sencillo si se tiene en cuenta el resultado anterior junto con el Teorema de Lehmann-Scheffé.

Corolario 3.6.

En las condiciones del Modelo Lineal Normal, $\hat{\mu}$ y $\hat{\sigma}^{2,\mathrm{I}}$ son los EIMV de μ y σ^2, respectivamente[6].

Demostración.

Teniendo en cuenta que $\hat{\mu}$ y $\hat{\sigma}^{2,\mathrm{I}}$ son estimadores insesgados de μ y σ^2, respectivamente, y que son de cuadrado integrable[7] el teorema 9.19 garantiza que los estadísticos $\mathtt{E}\left[\hat{\mu}\middle|\left(\hat{\mu},\hat{\sigma}^{2,\mathrm{I}}\right)\right]$ y $\mathtt{E}\left[\hat{\sigma}^{2,\mathrm{I}}\middle|\left(\hat{\mu},\hat{\sigma}^{2,\mathrm{I}}\right)\right]$ son los únicos EIMV de μ y σ^2, respectivamente, de lo cual se deduce la tesis.

∎

A continuación probaremos que, en las condiciones del Modelo Lineal Normal, podemos hablar del Estimador de Máxima Verosimilitud o, abreviadamente, EMV. Dijimos antes que dicho estimador ha de expresarse como función del estadístico suficiente y completo obtenido en el teorema anterior. Efectivamente, consideremos

[4] Así podría formularse el Principio de Suficiencia.

[5] Cf. Fergusson (1996)

[6] Estamos afirmando de manera implícita, que son los únicos (esencialmente) EIMV.

[7] Esto es así porque sus distribuciones derivan de la normal multivariante esférica. Por otra parte, cuando decimos que $\hat{\mu}$ es de cuadrado integrable nos referimos a que todas sus componentes lo son.

el estimador de σ^2 que se obtiene de forma natural dividiendo por $\mathtt{n}$ en lugar de por $\mathtt{n}-\mathtt{dim}V$, es decir,

$$\hat{\sigma}^{2,\mathtt{MV}} = \frac{1}{\mathtt{n}}\|Y - P_V Y\|^2 = \tfrac{\mathtt{n}-\mathtt{dim}V}{\mathtt{n}}\ \hat{\sigma}^{2,\mathtt{I}} \tag{3.9}$$

En ese caso, se verifica lo siguiente:

Teorema 3.7.
Bajo las condiciones del Modelo Lineal Normal, $(\hat{\mu}, \hat{\sigma}^2)$ es el EMV de (μ, σ^2). Además, el valor que alcanza la función de verosimilitud en dicho estimador es $\left(2\pi \mathrm{e}\ \hat{\sigma}^{2,\mathtt{MV}}\right)^{-\mathtt{n}/2}$.

Demostración.
Consideremos nuevamente la función de verosimilitud $\mathcal{L}$ definida en (3.8). Supuesto fijo $\mathtt{y} \in \mathbb{R}^{\mathtt{n}}$ y teniendo en cuenta que $\|\mathtt{y}-\mu\|^2$ descompone en $\|\mathtt{y}-P_V\mathtt{y}\|^2+\|P_V\mathtt{y}-\mu\|^2$, se deduce fácilmente que $\mathcal{L}(\mathtt{y}; P_V\mathtt{y}, \sigma^2) \geq \mathcal{L}(\mathtt{y}; \mu, \sigma^2)$, para todo μ y σ^2. Maximicemos a continuación la función $f(\sigma) = \mathcal{L}(\mathtt{y}; P_V\mathtt{y}, \sigma^2)$. Para ello consideramos su primera derivada, que resulta ser

$$f'(\sigma) = f(\sigma)\sigma^{-1}\left(\frac{\|\mathtt{y} - P_V\mathtt{y}\|^2}{\sigma^2} - n\right),$$

que se anula si y sólo si $\sigma^2 = \mathtt{n}^{-1}\|\mathtt{y} - P_V\mathtt{y}\|^2$. Además, es fácil comprobar que la segunda derivada es negativa en ese punto, lo cual garantiza que f alcanza un máximo relativo en el mismo que, en estas condiciones, será absoluto. Por lo tanto, se verifica

$$\mathcal{L}(\mathtt{y}; \mu, \sigma^2) \leq \mathcal{L}(\mathtt{y}; , P_V\mathtt{y}, \sigma^2) \leq \mathcal{L}\left(\mathtt{y}; P_V\mathtt{y}, \frac{1}{\mathtt{n}}\|\mathtt{y} - P_V\mathtt{y}\|^2\right), \qquad \forall(\mu, \sigma^2) \in V \times \mathbb{R}^+.$$

Sustituyendo en $\mathcal{L}$ se obtiene el máximo indicado en la tesis.

∎

De este resultado se deduce la consistencia y eficiencia de ambos estimadores. Queda pues claro que los estimadores propuestos, especialmente $\hat{\mu}$, gozan de una excelente justificación teórica bajo las condiciones del Modelo Lineal Normal. Además, el teorema de Gauss-Markov garantiza su idoneidad, aunque respecto a un grupo de estimadores más restringido, prescindiendo del supuesto de normalidad. Pero sabemos que en Inferencia Estadística todo es relativo, y ésta no será la excepción, se cumplan o no las condiciones del Modelo Lineal Normal. Efectivamente, puede probarse fácilmente que

$$\mathtt{E}\left[\|\hat{\mu}\|^2\right] = \|\mu\|^2 + \mathtt{dim}V \cdot \sigma^2. \tag{3.10}$$

Es decir, que, por término medio, el EIMV proporciona una estimación *más larga* que el estimando μ. El EIMV de μ no es sino el estimador insesgado óptimo para todas

y cada una de las funciones de pérdida de la familia $\mathcal{W} = \{W_a : \ a \in \mathbb{R}^{\mathtt{n}}\}$, donde

$$W_a[v, (\mu, \sigma^2)] := \big(\langle a, v - \mu\rangle\big)^2, \qquad v \in V, \ (\mu, \sigma^2) \in V \times \mathbb{R}^+.$$

Si en vez de considerar la familia $\mathcal{W}$ consideramos una única función de pérdida W, definida de manera muy natural mediante

$$W[v, (\mu, \sigma^2)] := \frac{\|v - \mu\|^2}{\sigma^2}, \tag{3.11}$$

sucede que, cuando $\mathtt{dim}V > 2$, el EIMV de μ resulta ser inadmisible. De hecho, el siguiente estimador, proporcional al EIMV y denominado de James-Stein, resulta ser preferible al mismo para dicha función de pérdida:

$$\hat{\mu}^{\text{JS}} = \left(1 - \frac{(\mathtt{dim}V - 2)(\mathtt{n} - \mathtt{dim}V)}{\mathtt{n} - \mathtt{dim}V + 2} \frac{\hat{\sigma}^{2,\text{I}}}{\|\hat{\mu}\|^2}\right) \hat{\mu}.$$

Las propiedades de este nuevo estimador se estudian con mayor detenimiento en Arnold (1981), capítulo 11. En en el mismo capítulo se analiza también el denominado estimador de Ridge que, en el estudio de Regresión Lineal y en un marco teórico Bayesiano, puede mejorar en cierto sentido la estimación de μ cuando se observa multicolinealidad. No obstante y a pesar de todo, $\hat{\mu}$ será el único estimador de la media que consideraremos en la sucesivo.

A continuación, construiremos sendas regiones de confianza para los parámetros μ y σ^2 bajo los supuestos del Modelo Lineal Normal. Concretamente, se verifica lo siguiente:

Proposición 3.8.

En el Modelo Lineal Normal se verifica que, para cada $\alpha \in (0, 1)$, los conjuntos $\mathcal{E}_\alpha$ y $\mathcal{I}_\alpha$, definidos mediante

$$\mathcal{E}_\alpha = \left\{v \in V : \ \|v - \hat{\mu}\|^2 \leq \mathtt{dim}V \hat{\sigma}^{2,\text{I}} F^{\alpha}_{\mathtt{dim}V, n - \mathtt{dim}V}\right\} \tag{3.12}$$

$$\mathcal{I}_\alpha = \left\{z \in \mathbb{R}^+ : \ \frac{(\mathtt{n} - \mathtt{dim}V)}{\chi^{2,1-\alpha/2}_{\mathtt{n} - \mathtt{dim}V}} \ \hat{\sigma}^{2,\text{I}} \leq z \leq \frac{(\mathtt{n} - \mathtt{dim}V)}{\chi^{2,\alpha/2}_{\mathtt{n} - \mathtt{dim}V}} \ \hat{\sigma}^{2,\text{I}}\right\}, \tag{3.13}$$

constituyen sendas regiones de confianza al $(1-\alpha)\times 100\,\%$ para μ y σ^2, respectivamente.

Demostración.

Sabemos por la proposición 3.4 cuáles son las distribuciones de $\hat{\mu}$ y $\hat{\sigma}^{2,\text{I}}$, de lo cual se deduce inmediatamente el intervalo de confianza para σ^2. Respecto a μ, consideremos

una matriz Γ cuyas columnas constituyan una base ortonormal de V, y definamos el estadístico $T = \Gamma'\hat{\mu}$. En ese caso, se tiene que

$$T \sim N_{\mathtt{dim}V}(\Gamma'\mu, \sigma^2 \mathtt{Id}),$$

Siendo independiente de $\hat{\sigma}^{2,\mathrm{I}}$. Se tiene entonces que $\|T - \Gamma'\mu\|^2 \sim \sigma^2 \chi^2_{\mathtt{dim}V}$ y, en consecuencia,

$$\frac{1}{\mathtt{dim}V} \frac{\|T - \Gamma'\mu\|^2}{\hat{\sigma}^{2,\mathrm{I}}} \sim F_{\mathtt{dim}V, n-\mathtt{dim}V}.$$

Por lo tanto, cualesquiera que sean μ y σ^2, se verifica

$$P_{\mu,\sigma^2}\left[y \in \mathbb{R}^{\mathtt{dim}V} : \frac{1}{\hat{\sigma}^{2,\mathrm{I}}} \|y - T\|^2 \leq \mathtt{dim}V F^{\alpha}_{\mathtt{dim}V, n-\mathtt{dim}V}\right] = 1 - \alpha.^{8}$$

Teniendo en cuenta que todo $v \in V$ puede expresarse mediante $v = \Gamma y$, para un único $y \in \mathbb{R}^{\mathtt{dim}V}$, y que

$$\|\Gamma' v - \Gamma'\hat{\mu}\|^2 = (v - \hat{\mu})' P_V (v - \hat{\mu}) = \|v - \hat{\mu}\|^2,$$

se concluye.

∎

Podemos observar que, mientras la región de confianza para σ^2 es un intervalo positivo, la de μ es la intersección entre una esfera y el subespacio V. El centro de dicha esfera es el estimador puntual $\hat{\mu}$, mientras que el radio volumen es proporcional al estimador de σ^2.

El problema de Estimación está obviamente condicionado por la elección del parámetro, es decir, por la forma de caracterizar las distribuciones de la familia de probabilidades considerada, en nuestro caso mediante μ y σ^2. No obstante, dado que μ es un vector de V, puede resultar natural expresarla a través de sus coordenadas respecto de una base de dicho subespacio. Esta situación se dará, concretamente, cuando estudiemos el problema de Regresión Lineal, tal y como se comento en el capítulo de introducción. Efectivamente, en tal caso, partiremos de una matriz $\mathtt{X}$ de rango completo compuesta por los valores obtenidos en las variables explicativas junto con un término independiente. V será el subespacio generado por las columnas de $\mathtt{X}$ y el parámetro de interés no será la media μ en sí, sino su vector de coordenadas respecto de la base $\mathtt{X}$, que se corresponde con los coeficientes de las variables explicativas y el término independiente.

[8] El término P_{μ,σ^2} hace referencia, lógicamente, a la distribución $N_{\mathtt{n}}(\mu, \sigma^2 \mathtt{Id})$.

Así, hablando en términos generales, si $\mathtt{X}$ denota una matriz cuyas columnas constituyen una base de V, la ecuación lineal $\mu = \mathtt{X}b$ tendrá una única solución en $\mathbb{R}^{\mathtt{dim}V}$, concretamente el vector

$$\beta_{\mathtt{X}} = (\mathtt{X}'\mathtt{X})^{-1}\mathtt{X}'\mu. \tag{3.14}$$

De esta forma, las distribuciones de la familia pueden caracterizarse igualmente mediante los parámetros $\beta_{\mathtt{X}}$ y σ^2 y, teniendo en cuenta en todo caso la ecuación (3.14), lo dicho hasta el momento respecto a la estimación de (μ, σ^2) se traduce a la de $(\beta_{\mathtt{X}}, \sigma^2)$ de la siguiente forma.

Teorema 3.9.
En las condiciones del Modelo Lineal, sean $\mathtt{X}$ una base de V y $\hat{\beta}_{\mathtt{X}}$ el estadístico definido mediante

$$\hat{\beta}_{\mathtt{X}} = (\mathtt{X}'\mathtt{X})^{-1}\mathtt{X}'Y. \tag{3.15}$$

Entonces, se verifica:
(i) $\hat{\beta}_{\mathtt{X}}$ es un estimador insesgado de $\beta_{\mathtt{X}}$.
(ii) Para todo $b \in \mathbb{R}^{\mathtt{dim}V}$, $b'\hat{\beta}_{\mathtt{X}}$ es el estimador lineal insesgado de mínima varianza de $b'\beta_{\mathtt{X}}$.
Si, además, se verifican las condiciones del Modelo Lineal Normal, se tiene que:
(iii) $\hat{\beta}_{\mathtt{X}} \sim N_{\mathtt{dim}V}\left(\beta_{\mathtt{X}}, \sigma^2(\mathtt{X}'\mathtt{X})^{-1}\right)$, siendo independientes de $\hat{\sigma}^{2,\mathrm{I}}$.
(iv) El estadístico $\left(\hat{\beta}_{\mathtt{X}}, \hat{\sigma}^{2,\mathrm{I}}\right)$ es suficiente y completo.
(v) $\hat{\beta}_{\mathtt{X}}$ es el EIMV de $\beta_{\mathtt{X}}$.
(vi) $\left(\hat{\beta}_{\mathtt{X}}, \hat{\sigma}^{2,\mathrm{MV}}\right)$ es el EMV de $(\beta_{\mathtt{X}}, \sigma^2)$. Además, el valor que alcanza la función de verosimilitud en dicho estimador es $\left(2\pi \mathrm{e}\, \hat{\sigma}^{2,\mathrm{MV}}\right)^{-\mathtt{n}/2}$.
(vii) Dado $\alpha \in (0,1)$, el conjunto

$$\mathcal{E}_{\alpha,\mathtt{X}} = \left\{ b \in \mathbb{R}^{\mathtt{dim}V} : \left(b - \hat{\beta}_{\mathtt{X}}\right)' \mathtt{X}'\mathtt{X} \left(b - \hat{\beta}_{\mathtt{X}}\right) \leq \hat{\sigma}^{2,\mathrm{I}}\ \mathtt{dim}V F^{\alpha}_{\mathtt{dim}V, n-\mathtt{dim}V} \right\} \tag{3.16}$$

constituye una región de confianza al $(1-\alpha)\times 100\,\%$ para $\beta_{\mathtt{X}}$.

Nótese que, en un contexto determinista, es decir, si prescindiéramos del vector de errores $\mathcal{E}$, y siendo la matriz $\mathtt{X}$ de rango completo, existe solución a la ecuación lineal $Y = \mathtt{X}b$ si y sólo si $Y \in V$, en cuyo caso será única. El estimador (3.15) de $\beta_{\mathtt{X}}$ es, precisamente, el único que satisface

$$P_V Y = \mathtt{X}\hat{\beta}_{\mathtt{X}}. \tag{3.17}$$

3.2. Test F para la media.

El segundo problema de Inferencia Estadística que abordamos es el de contraste de hipótesis. En esta sección nos limitaremos a estudiar contrastes acerca del parámetro principal del modelo, μ. Concretamente, se considerarán hipótesis de tipo lineal. Nos referimos a lo siguiente: dado un subespacio lineal $W \subset V$, contrastaremos la hipótesis inicial

$$H_0 : \mu \in W \tag{3.18}$$

frente a su alternativa.

En capítulos siguientes veremos ejemplos de contrastes de este tipo en los diferentes problemas a estudiar. Supondremos, en todo caso, que se verifican los supuestos del Modelo Lineal Normal. En esas condiciones, tanto de la aplicación de los Principios de Suficiencia e Invarianza como del de Máxima Verosimilitud se deriva un mismo test, denominado frecuentemente por Anova[9] o, mejor, test F.

Dada la enorme trascendencia de este test, no basta probar que posee el nivel de significación α que se le supone, sino que conviene justificar su idoneidad a la luz de algún o algunos Principios Estadísticos. En ese sentido, no será difícil probar que el test F es el Test de la Razón de Verosimilitudes (TRV, para abreviar), lo cual, además de satisfacernos desde un punto de vista meramente filosófico, confiere al test importantes propiedades asintóticas[10]. No obstante, puede demostrarse que, en nuestras condiciones, el TRV es función de cualquier estadístico suficiente y es invariante ante cualquier grupo de transformaciones que deje a su vez invariantes tanto el experimento estadístico como el problema de contraste de hipótesis. De ahí que no sea una mera casualidad que el test F pueda justificarse también como test UMP-invariante a nivel α, es decir, es el test a nivel α más potente entre todos los invariantes a nivel α respecto de un grupo de transformaciones que especificaremos más adelante. Es más, el enunciado del lema fundamental de Neyman-Pearson (ver Apéndice) desvela una clara conexión entre la búsqueda de un test UMP y la del TRV, siempre y cuando se den ciertas condiciones que se cumplen en nuestro modelo. Por último, teniendo en cuenta que todo estadístico constante (en particular el que toma en todo caso el valor α) es invariante, se deduce que el test F será a su vez insesgado a nivel α, es decir, que su función potencia tomará valores no iferiores a α cuando $\mu \notin W$.

El lector interesado en seguir con rigor esta parte del capítulo debería estar familiarizado con con los fundamentos de la Teoría de la Decisión, así como con los

[9] Abreviatura de Analysis of Variance.
[10] Ver Fergusson (1996).

conceptos de Suficiencia, Completitud e Invarianza. Todo ello puede encontrarse en A.G. Nogales (1998). También aconsejamos ver previamente el apartado del Apéndice dedicado al Principio de Invarianza. Veamos pues cómo se obtiene el test `F`.

1. **Paso a forma canónica** En primer lugar, aplicaremos a nuestro modelo una transformación bimedible: concretamente un cambio de base en $\mathbb{R}^{\mathtt{n}}$. El objeto del mismo es estructurar el espacio de parámetros de manera natural en función de la hipótesis a contrastar. Para ello consideraremos tres matrices $\mathtt{X}_1$, $\mathtt{X}_2$ y $\mathtt{X}_3$, bases ortonormales de los subespacios ortogonales W, $V|W$ y $V^{\perp}$, respectivamente. Sea entonces la transformación bimedible φ de $(\mathbb{R}^{\mathtt{n}}, \mathcal{R}^{\mathtt{n}})$ en sí mismo, que hace corresponder a cada vector Y el vector $Z = \varphi(Y)$ definido mediante

$$Z = \begin{pmatrix} \mathtt{X}_1' \\ \mathtt{X}_2' \\ \mathtt{X}_3' \end{pmatrix} Y.$$

   El vector Z está compuesto por las coordenadas de Y respecto a una base ortonormal de $\mathbb{R}^{\mathtt{n}}$, la cual se descompone a su vez en bases de W, $V|W$ y $V^{\perp}$. Si se denota $Z_i = \mathtt{X}_i'Y$, $\nu_i = \mathtt{X}_i'\mu$, para $i = 1, 2, 3$, se tiene un nuevo modelo, que denominamos canónico, compuesto por tres vectores aleatorios independientes

$$\begin{aligned} Z_1 &\sim N_{\mathtt{dim}W}(\nu_1, \sigma^2\mathtt{Id}) \\ Z_2 &\sim N_{\mathtt{dim}V-\mathtt{dim}W}(\nu_2, \sigma^2\mathtt{Id}) \\ Z_3 &\sim N_{\mathtt{n}-\mathtt{dim}V}(0, \sigma^2\mathtt{Id}) \end{aligned}$$

   La familia de distribuciones puede expresarse pues con la ayuda del parámetro (ν_1, ν_2, σ^2), que recorre el espacio $\mathbb{R}^{\mathtt{dim}W} \times \mathbb{R}^{\mathtt{dim}V-\mathtt{dim}W} \times \mathbb{R}^+$. La hipótesis inicial (3.18) se traduce entonces en $H_0 : \nu_2 = 0$.

2. **Reducción por suficiencia.** En virtud del teorema 3.5, el estadístico $(\hat{\mu}, \hat{\sigma}^2)$ es suficiente y completo. Dado que

$$\hat{\mu} = \begin{pmatrix} \mathtt{X}_1 Z_1 \\ \mathtt{X}_2 Z_2 \end{pmatrix}, \qquad \hat{\sigma}^2 \propto \|Z_3\|^2,$$

   se verifica que $S = (Z_1, Z_2, \|Z_3\|^2)$ es, a su vez, un estadístico suficiente y completo respecto al modelo canónico. Sabemos que el considerar únicamente la imagen de dicho estadístico, lo cual se denomina reducción por suficiencia, no conlleva pérdida alguna de información en el sentido de Fisher y no afecta, como veremos más adelante, a la búsqueda de un test UMP a nivel α. Además,

al ser completo, la reducción por suficiencia es máxima, esto es, una reducción más profunda sí implicaría pérdida de información referente al parámetro. Las distribuciones del nuevo modelo reducido podrán expresarse, igual que en la fase anterior[11], con la ayuda del parámetro (ν_1, ν_2, σ^2). La hipótesis a contrastar sigue siendo $\nu_2 = 0$.

3. **Reducción por invarianza.** Dado que la reducción por suficiencia no simplifica de manera satisfactoria el modelo, llevaremos a cabo una reducción más profunda por invarianza. Para ello consideraremos el grupo de transformaciones bimedibles en el modelo canónico

$$G = \{\mathtt{g}_{k,O,\lambda} : k \in \mathbb{R}^{\mathtt{dim}W},\ O \in \mathcal{O}_{\mathtt{dim}V-\mathtt{dim}W},\ \lambda > 0\}\ ^{12},$$

siendo

$$g_{k,O,\lambda}\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} = \lambda \begin{pmatrix} Z_1 + k \\ OZ_2 \\ Z_3 \end{pmatrix}.$$

Puede comprobarse fácilmente que G deja invariante tanto el modelo como el problema de contraste de hipótesis considerado. Por ello, el Principio de Invarianza propone restringir la búsqueda de tests a aquellos que sean igualmente invariantes, y entre éstos seleccionar el mejor desde algún criterio establecido. En este caso y dado $\alpha \in (0, 1)$, encontraremos el test UMP-invariante a nivel α.

Dado que previamente hemos efectuado una reducción por suficiencia y que el estadístico suficiente S es trivialmente equivariante respecto a G, podemos considerar el grupo de transformaciones G^S que G induce de manera natural sobre el modelo imagen de S y buscar en dicho modelo un test ϕ_S UMP-invariante respecto a G^S a nivel α. De esta forma, el test $\phi_S \circ S$, definido sobre el modelo canónico, cumplirá la condición deseada. Vayamos por partes.

En primer lugar, el grupo G^S puede descomponerse en la suma de los subgrupos $G_1 = \{\mathtt{g}_k : k \in \mathbb{R}^{\mathtt{dim}W}\}$, $G_2 = \{\mathtt{g}_O : O \in \mathcal{O}_{\mathtt{dim}V-\mathtt{dim}W}\}$ y $G_3 = \{\mathtt{g}_\lambda : \lambda > 0\}$, donde

$$\mathtt{g}_k\begin{pmatrix} Z_1 \\ Z_2 \\ \|Z_3\|^2 \end{pmatrix} = \begin{pmatrix} Z_1 + k \\ Z_2 \\ \|Z_3\|^2 \end{pmatrix}, \quad \mathtt{g}_O\begin{pmatrix} Z_1 \\ Z_2 \\ \|Z_3\|^2 \end{pmatrix} = \begin{pmatrix} Z_1 \\ OZ_2 \\ \|Z_3\|^2 \end{pmatrix},$$

[11] Una reducción por suficiencia no puede implicar simplificación alguna en el espacio de parámetros.

[12] En general, el término $\mathcal{O}_m$ denotará el conjunto de las matrices cuadradas de orden m y ortogonales.

$$\mathrm{g}_\lambda \begin{pmatrix} Z_1 \\ Z_2 \\ \|Z_3\|^2 \end{pmatrix} = \begin{pmatrix} \lambda Z_1 \\ \lambda Z_2 \\ \lambda^2 \|Z_3\|^2 \end{pmatrix}.$$

Estos subgrupos verifican la propiedad (9.49). Nuestro primer objetivo es encontrar un estadístico invariante maximal respecto a G^S, así como el correspondiente invariante maximal para el espacio de parámetros. Aprovechando la descomposición de G^S, dicha búsqueda se realizará en tres etapas. En primer lugar, es obvio que el siguiente estadístico es un invariante maximal respecto a G_1.

$$M_1 \begin{pmatrix} Z_1 \\ Z_2 \\ \|Z_3\|^2 \end{pmatrix} = \begin{pmatrix} Z_2 \\ \|Z_3\|^2 \end{pmatrix}.$$

Además, el conjunto $\{(\nu_2, \sigma^2) : \nu_2 \in \mathbb{R}^{\mathtt{dim}V - \mathtt{dim}W}, \sigma^2 > 0\}$ es la imagen de un invariante maximal para el espacio de parámetros. Consideramos entonces el grupo $G_2^1 = \{\mathrm{g}_O^{M_1} : O \in \mathcal{O}_{\mathtt{dim}V - \mathtt{dim}W}\}$, definido mediante

$$\mathrm{g}_O^{M_1} \begin{pmatrix} Z_2 \\ \|Z_3\|^2 \end{pmatrix} = \begin{pmatrix} OZ_2 \\ \|Z_3\|^2 \end{pmatrix}.$$

En virtud del teorema 9.12, el estadístico M_2^1 definido mediante

$$M_2^1 \begin{pmatrix} Z_2 \\ \|Z_3\|^2 \end{pmatrix} = \begin{pmatrix} \|Z_2\|^2 \\ \|Z_3\|^2 \end{pmatrix}$$

es invariante maximal respecto a G_2^1 en el modelo imagen de M_1. El conjunto $\{(\|\nu_2\|^2, \sigma^2) : \nu_2 \in \mathbb{R}^{\mathtt{dim}V - \mathtt{dim}W}, \sigma^2 > 0\}$ es, a su vez, la imagen de un invariante maximal para el espacio de parámetros. Tomamos, por último, el grupo $G_3^{12} = \{\mathrm{g}_\lambda^{M_2^1 \circ M_1} : \lambda > 0\}$, definido mediante

$$\mathrm{g}_\lambda^{M_2^1 \circ M_1} \begin{pmatrix} \|Z_2\|^2 \\ \|Z_3\|^2 \end{pmatrix} = \lambda^2 \begin{pmatrix} \|Z_2\| \\ \|Z_3\|^2 \end{pmatrix}.$$

El estadístico M_3^{12} definido mediante

$$M_3^{12} \begin{pmatrix} \|Z_2\|^2 \\ \|Z_3\|^2 \end{pmatrix} = \delta_{\mathtt{n},V,W} \frac{\|Z_2\|^2}{\|Z_3\|^2}$$

es invariante maximal respecto a G_3^{12}. En la expresión anterior, $\delta_{\mathtt{n},V,W}$ puede ser cualquier número real no nulo. En nuestro caso, conviene tomar (ya veremos el porqué) $\delta_{\mathtt{n},V,W} = (\mathtt{n} - \mathtt{dim}V)/(\mathtt{dim}V - \mathtt{dim}W)$. Por su parte, un invariante

maximal respecto al espacio de parámetros nos lleva a considerar el parámetro $\theta = \|\nu_2\|^2/\sigma^2$, que recorre el espacio $[0, +\infty]$.

En definitiva, las distintas reducciones por suficiencia e invarianza conducen a considerar el modelo inducido por el estadístico invariante maximal M_3^{12}, concretamente

$$\big([0,+\infty], \mathcal{R}([0,+\infty], \{P_\theta : \theta \geq 0\}\big),$$

donde, para cada $\theta \geq 0$ y en virtud de (2.8), P_θ es la distribución F-Snedecor no central con grados de libertad $(\mathtt{dim}V - \mathtt{dim}W, n - \mathtt{dim}V)$ y parámetro de no centralidad θ Denótese por p_θ a la correspondiente densidad, cuya expresión explícita aparece en (2.7). La hipótesis a contrastar se traduce en $H_0 : \theta = 0$, frente a la alternativa $H_0 : \theta > 0$. Tal y como se afirma en la sección 1.2, para cada $\theta > 0$, la función $p_\theta(x)/p_0(x)$ es creciente en $x \geq 0$, es decir, que el modelo presenta razón de verosimilitudes monótona. En ese caso, se sigue de la proposición 9.20 que el test $\overline{\phi}$, definido sobre el modelo reducido final mediante

$$\overline{\phi}(x) = \begin{cases} 1 & \text{si } x > F^\alpha_{\mathtt{dim}V - \mathtt{dim}W, n - \mathtt{dim}V} \\ 0 & \text{si } x \leq F^\alpha_{\mathtt{dim}V - \mathtt{dim}W, n - \mathtt{dim}V} \end{cases}$$

es UMP a nivel α. Así pues, el test $\phi_S = \overline{\phi} \circ M_3^{12} \circ M_2^1 \circ M_1$, definido sobre el modelo reducido por suficiencia, es UMP-invariante a nivel α respecto al grupo G^S. Por lo tanto, el test $\phi_S \circ S$, definido sobre el modelo canónico, es UMP-invariante a nivel α respecto al grupo G. Para acabar, tomando $\mathtt{F} = \phi^S \circ S \circ \varphi$ deshacemos el cambio de variables φ inicial. El test $\mathtt{F}$ a nivel α, definido sobre el modelo original puede expresarse pues como sigue:

$$\mathtt{F}(Y) = \begin{cases} 1 & \text{si } F(Y) > F^\alpha_{\mathtt{dim}V - \mathtt{dim}W, n - \mathtt{dim}V} \\ 0 & \text{si } F(Y) \leq F^\alpha_{\mathtt{dim}V - \mathtt{dim}W, n - \mathtt{dim}V} \end{cases}, \tag{3.19}$$

siendo F el estadístico de contraste definido mediante

$$\begin{aligned} F(Y) &= M_3^{12} \circ M_2^1 \circ M_1 \circ S \circ \varphi(Y) && (3.20) \\ &= \frac{\mathtt{n} - \mathtt{dim}V}{\mathtt{dim}V - \mathtt{dim}W} \frac{\|P_{V|W} Y\|^2}{\|P_{V^\perp} Y\|^2} && (3.21) \\ &= \frac{\mathtt{n} - \mathtt{dim}V}{\mathtt{dim}V - \mathtt{dim}W} \frac{\|P_V Y - P_W Y\|^2}{\|Y - P_V Y\|^2} && (3.22) \\ &= \frac{1}{\mathtt{dim}V - \mathtt{dim}W} \frac{\|P_{V|W} Y\|^2}{\hat{\sigma}^{2,\mathtt{I}}}. && (3.23) \end{aligned}$$

En definitiva, hemos probado lo siguiente:

Teorema 3.10.
En las condiciones de Modelo Lineal Normal, dados $W \subset V$ y $\alpha \in (0,1)$, el test (3.19) es UMP-invariante [13] a nivel α para contrastar la hipótesis inicial $H_0 : \mu \in W$. En particular, es insesgado a nivel α.

La distribución del estadístico de contraste F respecto a $N_{\mathtt{n}}(\mu, \sigma^2 \mathtt{Id})$ depende de μ y σ^2 a través del parámetro del modelo reducido final

$$\theta = \frac{\|P_{V|W}\mu\|^2}{\sigma^2}. \tag{3.24}$$

En concreto, para cada distribución $N_{\mathtt{n}}(\mu, \sigma^2 \mathtt{Id})$, con $\mu \in V$ y $\sigma^2 > 0$, se tiene que

$$F \sim F_{\mathtt{dim}V-\mathtt{dim}W,n-\mathtt{dim}V}\left(\frac{\|P_{V|W}\mu\|^2}{\sigma^2}\right).$$

El caso nulo, $\mu \in W$, se corresponde con la situación $F \sim F_{\mathtt{dim}V-\mathtt{dim}W,n-\mathtt{dim}V}$. El término $\|P_V Y - P_W Y\|^2$, que aparece en el numerador de F, se denota con frecuencia en la literatura mediante `SCH` (siglas de suma cuadrática de la hipótesis), mientras que el término $\|Y - P_V Y\|^2$, que aparece en el denominador, se denota por `SCE` (suma cuadrática del error). El estadístico de contraste F resulta de dividir estos términos por las dimensiones (grados de libertad) de $V|W$ y $V^\perp$, respectivamente. De esta forma, en el denominador tenemos el EIMV de la varianza σ^2, suponiendo que μ pertenece a V, mientras que en el denominador aparece un estimador sesgado de la varianza, pues su esperanza es, en virtud de (2.6), $\sigma^2 + \|P_{V|W}\mu\|^2$. Por lo tanto, sólo si se supone cierta la hipótesis inicial, este estimador será insesgado, en cuyo caso, cabría esperar que el cociente F estuviera próximo a 1. Un valor muy alto del cociente entre estos dos estimadores de la varianza se interpreta como un desacuerdo entre los datos y la hipótesis nula. Por ello, es muy usual referirse al test (3.19) con el sobrenombre de Anova, abreviatura en inglés de Análisis de la Varianza. No obstante y para evitar confusiones con el Diseño de Experimentos, lo denominaremos en lo sucesivo test `F`.

Si analizamos detenidamente la expresión del invariante maximal F para el espacio de observaciones y del invariante maximal θ para el espacio de parámetros, detectaremos cierto paralelismo con la expresión de la función de densidad de la distribución normal multivariante esférica, lo cual no es de extrañar. Este ejemplo ilustra hasta qué punto cualquier propiedad relativa a la suficiencia e

[13] En este caso, queremos decir invariante respecto al grupo de transformaciones $G_\varphi = \{\mathtt{g} \circ \varphi : \mathtt{g} \in G\}$, que deja invariantes tanto el modelo como el problema de contraste de hipótesis originales.

invarianza es inherente a la formulación de la familia de distribuciones consideradas.

A continuación comprobaremos que el test F es también es el Test de la Razón de Verosimilitudes definido en el Apéndice. Ya comentamos allí que, bajo ciertas condiciones de regularidad, que se verifican bajo los supuestos de nuestro modelo, si G es un grupo de transformaciones que dejan invariante tanto el modelo como el problema de contraste de hipótesis, y existe el estadístico de la razón de verosimilitudes a nivel α, éste ha de ser equivalente a otro estadístico invariante respecto a G. De esta proposición se deduce que, en nuestro caso, existe un test equivalente TRV que puede expresarse a través de un estadístico de contraste invariante y, por lo tanto, función del estadístico de F, definido en (3.22). Más aún, podemos afirmar que el test F es el propio TRV:

Teorema 3.11.
En las condiciones del Modelo Lineal Normal, dados $W \subset V$ y $\alpha \in (0,1)$, el test (3.19) es el único test de la razón de verosimilitudes a nivel α para contrastar la hipótesis inicial $H_0 : \mu \in W$.

Demostración.
Recordemos que, en nuestro modelo la función de verosimilitud es la siguiente

$$\mathcal{L}(\mathtt{y};\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{\mathtt{n}}{2}} \exp\left\{-\frac{1}{2}\frac{\|\mathtt{y}-\mu\|^2}{\sigma^2}\right\}, \qquad \mathtt{y} \in \mathbb{R}^{\mathtt{n}},$$

donde $(\mu,\sigma^2) \in V \times \mathbb{R}^+$, y la hipótesis inicial a contrastar es $W \times \mathbb{R}^+$. Del teorema 3.7 se sigue que, para todo $\mathtt{y} \in \mathbb{R}^{\mathtt{n}}$,

$$\sup_{\mu\in V,\sigma^2>0} \mathcal{L}(\mathtt{y};\mu,\sigma^2) = \mathcal{L}\left(\mathtt{y}; P_V\mathtt{y}, \frac{1}{\mathtt{n}}\|P_{V^\perp}\mathtt{y}\|^2\right),$$

$$\sup_{\mu\in W,\sigma^2>0} \mathcal{L}(\mathtt{y};\mu,\sigma^2) = \mathcal{L}\left(\mathtt{y}; P_W\mathtt{y}, \frac{1}{\mathtt{n}}\|P_{W^\perp}\mathtt{y}\|^2\right)$$

En consecuencia, el estadístico de la razón de verosimilitudes es

$$RV(Y) = \left(\frac{\|P_{V^\perp}Y\|^2}{\|P_{W^\perp}Y\|^2}\right)^{\frac{\mathtt{n}}{2}}, \quad Y \in \mathbb{R}^{\mathtt{n}}. \tag{3.25}$$

Dado que que $W^\perp$ descompone en en la suma ortogonal $V^\perp \oplus V|W$, se tiene entonces que

$$RV^{\mathtt{n}/2} = \frac{1}{1 + \frac{\mathtt{dim}V - \mathtt{dim}W}{\mathtt{n}-\mathtt{dim}V}\ F}.$$

Luego, teniendo en cuenta (2.9), se deduce que $RV^{\mathtt{n}/2}$ sigue una distribución Beta de parámetros $(\mathtt{n}-\mathtt{dim}V)/2$ y $(\mathtt{dim}V-\mathtt{dim}W)/2)$ en el caso nulo. Por lo tanto, el único test de la razón de verosimilitudes es

$$TRV(Y)=\begin{cases}1 & \text{si } RV(Y)<\left[B\left(\frac{\mathtt{n}-\mathtt{dim}V}{2},\frac{\mathtt{dim}V-\mathtt{dim}W}{2}\right)^{1-\alpha}\right]^{2/n}\\ 0 & \text{si } RV(Y)\geq\left[B\left(\frac{\mathtt{n}-\mathtt{dim}V}{2},\frac{\mathtt{dim}V-\mathtt{dim}W}{2}\right)^{1-\alpha}\right]^{2/n}\end{cases}$$

Dado que la función $f(x)=(1+\lambda x)^{-2/n}$ es, para todo $\lambda>0$, una biyección decreciente de $[0,\infty]$ en $[0,1]$, se sigue que

$$\left[RV(Y)^{\mathtt{n}/2}<B\left(\frac{\mathtt{n}-\mathtt{dim}V}{2},\frac{\mathtt{dim}V-\mathtt{dim}W}{2}\right)^{1-\alpha}\right]\Leftrightarrow\left[F(Y)>F^{\alpha}_{\mathtt{dim}V-\mathtt{dim}W,n-\mathtt{dim}V}\right]$$

luego, el test anterior coincide con el test (3.19).

∎

Así pues, hemos demostrado que el test F a nivel α es insesgado, UMP-invariante y test de la razón de verosimilitudes. Al igual que en el problema de Estimación, veamos cómo se expresa el test F si parametrizamos el modelo mediante las coordenadas respecto a una base $\mathbf{X}$ de V, lo cual será de enorme utilidad cuando estudiemos el problema de Regresión Lineal. Consideremos pues una matriz $\mathbf{X}$ cuyas columnas constituyan una base de V. De esta forma, teniendo en cuenta la ecuación $\mu=\mathbf{X}\beta$, las distribuciones del modelo pueden caracterizarse mediante las coordenadas de la media respecto a $\mathbf{X}$, junto con σ^2. Dado un subespacio $W\subset V$, consideremos una matriz C de orden $n\times\mathtt{dim}V|W$ cuyas columnas constituyan una base del subespacio $V|W$. En ese caso, la hipótesis inicial $\mu\in W$ equivale a $C'\mathbf{X}\beta=0$, es decir, a $A_{\mathbf{X}}\beta=0$, siendo $A_{\mathbf{X}}=C'\mathbf{X}$, que es una matriz es de dimensiones $\mathtt{dim}V|W\times\mathtt{dim}V$ y rango $\mathtt{dim}V|W$.

Recíprocamente, dada una hipótesis inicial del tipo $A\beta=0$, siendo A una matriz de dimensiones $m\times\mathtt{dim}V$ y rango m (lo cual implica que $m\leq\mathtt{dim}V$), existe un subespacio $W_{\mathbf{X},A}$ de V de dimensión $\mathtt{dim}V-m$ tal que la hipótesis inicial anterior equivale a que $\mathbf{X}\beta$ pertenezca a $W_{\mathbf{X},A}$. Concretamente, se trata de la imagen del subespacio $\overline{W}$ de dimensión $\mathtt{dim}V-m$, constituido por los vectores b de $\mathbb{R}^{\mathtt{dim}V}$ tales que $Ab=0$, por la aplicación lineal inyectiva que a cada b en $\mathbb{R}^{\mathtt{dim}V}$ le asigna el vector $\mathbf{X}b$ de V.

Es decir, que contrastar hipótesis del tipo $\mu\in W$ equivale, en términos de β, a contrastar hipótesis del tipo $A\beta=0$, siendo A una matriz de orden $m\times\mathtt{dim}V$ y rango completo. De hecho, en Regresión Lineal expresaremos así las hipótesis iniciales. Conviene pues expresar también el estadístico de contraste del test F, que se ha

denotado por F, en función de $\mathtt{X}$ y de la matriz A correspondiente. Para ello es conveniente encontrar una base adecuada de $V|W_{\mathtt{X},A}$.

Lema 3.12.
Dada una matriz A de dimensiones $m \times \mathtt{dim}V$ y rango m, las columnas de la matriz $C = \mathtt{X}(\mathtt{X}'\mathtt{X})^{-1}A'$ constituyen una base del subespacio $V|W_{\mathtt{X},A}$.

Demostración.
Veamos que las columnas de C son linealmente independientes. En efecto, si existe un vector $g \in \mathbb{R}^m$, tal que $Cg = 0$, entonces, $A\mathtt{X}'Cg = 0$. Dado que AA' es una matriz cuadrada de orden m y rango m, podemos afirmar que

$$0 = (AA')^{-1}AX'Cg = (AA')^{-1}A\mathtt{X}'\mathtt{X}(\mathtt{X}'\mathtt{X})^{-1}A'g = g.$$

Por lo tanto, el rango de C es m. Falta probar que las columnas de C son ortogonales a $W_{\mathtt{X},A}$, es decir, que dado $b \in \mathbb{R}^m$ tal que $Ab = 0$, se verifica $(\mathtt{X}b)'C = (0, \ldots, 0)$. Efectivamente,

$$(\mathtt{X}b)'C = b'\mathtt{X}'\mathtt{X}(\mathtt{X}'\mathtt{X})^{-1}A' = b'A' = (0, \ldots, 0).$$

∎

Teorema 3.13.
Dada una matriz A de dimensiones $m \times \mathtt{dim}V$ y rango m, el estadístico de contraste del test F para contrastar la hipótesis inicial $H_0 : A\beta = 0$ es

$$F = \frac{\mathtt{n} - \mathtt{rg}(\mathtt{X})}{m} \; \frac{(A\hat{\beta})' \left[A(\mathtt{X}'\mathtt{X})^{-1}A'\right]^{-1} A\hat{\beta}}{\|Y\|^2 - Y'\mathtt{X}\hat{\beta}}, \tag{3.26}$$

con $\hat{\beta}$ definido según (3.15).

Demostración.
Se verifica por (3.21) y (3.22) que

$$F = \frac{\mathtt{n} - \mathtt{rg}(\mathtt{X})}{m} \; \frac{\|P_{V|W}Y\|^2}{\|Y\|^2 - \|P_VY\|^2}.$$

Sabemos que $P_{V|W}$ puede expresarse mediante $C(C'C)^{-1}C'$, para cualquier base C de $V|W$. Así pues, por el lema anterior y teniendo en cuenta que $\|P_{V|W}Y\|^2 = Y'P_{V|W}Y$, junto con la definición (3.15), obtenemos el numerador.

Respecto al denominador, basta notar que la proyección ortogonal sobre V puede expresarse mediante

$$P_V = \mathtt{X}(\mathtt{X}'\mathtt{X})^{-1}\mathtt{X}'$$

y que $\|P_VY\|^2 = Y'P_VY$. Entonces, por (3.15) se concluye.

∎

Nótese que la expresión (3.26) es más adecuada que (3.20) desde el punto de vista computacional. Éste puede ser un sólido argumento a la hora de justificar el uso de la versión coordenada del modelo lineal.

Para acabar con esta sección, nos preguntamos cómo se plantearía y resolvería en este marco teórico un contraste de tipo unilateral. Obviamente, no tiene sentido, en general, una hipótesis del tipo $\mu > \mu_0$ o $\mu < \mu_0$. Sin embargo, para cada $d \in V|W$, podemos considerar el parámetro $\langle d, \mu\rangle \in \mathbb{R}$ y, en consecuencia podemos contrastar la hipótesis unilateral $\langle d, \mu\rangle \leq 0$ frente a la alternativa $\langle d, \mu\rangle > 0$, o viceversa. En Arnold (1981)[14] se obtiene, mediante un paso a forma canónica, una reducción por suficiencia y dos por invarianza[15], el siguiente test UMP-invariante a nivel α:

$$\phi_d^+(Y) = \begin{cases} 1 & \text{si } t_d(Y) > t^\alpha_{\mathtt{n-v}} \\ 0 & \text{si } t_d(Y) \leq t^\alpha_{\mathtt{n-v}} \end{cases},$$

donde el estadístico de contraste t_d se define mediante

$$t_d(Y) = \frac{\langle d, \hat{\mu}\rangle}{\|d\|\hat{\sigma}}, \tag{3.27}$$

y $t_{\mathtt{n-v}}$ denota la distribución t-Student central con $n - \mathtt{v}$ grados de libertad.

3.3. Contrastes de Hipótesis para la varianza.

En esta breve sección se afronta, siguiendo el esquema lógico, el estudio del contraste de hipótesis relativas a la varianza σ^2. La brevedad de la misma se debe a razones de diversa índole. En primer lugar, desde un punto de vista práctico, interesan menos que los contrastes relativos a la media, pues ésta última constituye el parámetro principal del modelo. La varianza suele ser, por contra, un parámetro *fantasma* que no interesa en sí pero cuyo desconocimiento dificulta el estudio acerca de la media. La segunda razón es de carácter técnico pues, como se constatará en la próxima sección, los tests para la varianza presentan un comportamiento asintótico claramente peor que el test `F` para la media, lo cual se traducirá en una excesiva sensibilidad ante la frecuente violación del supuesto de normalidad.

En definitiva, dado $\sigma_0^2 > 0$, nos proponemos contrastar las siguiente hipótesis iniciales frente a sus correspondientes alternativas:

$$H_0^1 : \sigma = \sigma_0, \quad H_0^2 : \sigma \leq \sigma_0, \quad H_0^3 : \sigma \geq \sigma_0.$$

[14] capítulo 7, ejercicio B20

[15] El problema es invariante ante la acción de los grupos G_1 y G_3, considerados en el test `F`.

El procedimiento a seguir es, en principio, similar al llevado a cabo en la sección anterior: una reducción por suficiencia conduce a considerar el experimento estadístico inducido por el estadístico

$$(\hat{\mu}, \hat{\sigma}^{2,\mathrm{I}}) .$$

Además, se verifica, trivialmente, que tanto el experimento estadístico original como los tres problemas de contraste de hipótesis considerados permanecen invariantes ante cualquier traslación de coordenadas. Así pues, una reducción por invarianza conducen a considerar el estadístico invariante maximal $\hat{\sigma}^{2,\mathrm{I}}$, cuya distribución depende de (ν, σ^2) a través, únicamente, de σ^2. Concretamente, el estadístico T, definido mediante

$$T = (\mathtt{n} - \mathtt{dim}V)\frac{\hat{\sigma}^{2,\mathrm{I}}}{\sigma_0^2},$$

sigue una distribución $\chi^2_{\mathtt{n}-\mathtt{dim}V}$. Es fácil probar que el experimento estadístico inducido por T presenta razón de verosimilitudes monótona[16]. Por ello, los tests ϕ_2 y ϕ_3 definidos mediante

$$\phi_2(Y) = \begin{cases} 1 & \text{si } (\mathtt{n} - \mathtt{dim}V)\hat{\sigma}^{2,\mathrm{I}} > \sigma_0^2 \chi^{\alpha}_{\mathtt{n}-\mathtt{dim}V} \\ 0 & \text{si } (\mathtt{n} - \mathtt{dim}V)\hat{\sigma}^{2,\mathrm{I}} \leq \sigma_0^2 \chi^{\alpha}_{\mathtt{n}-\mathtt{dim}V} \end{cases} ,$$

$$\phi_3(Y) = \begin{cases} 1 & \text{si } (\mathtt{n} - \mathtt{dim}V)\hat{\sigma}^{2,\mathrm{I}} < \sigma_0^2 \chi^{1-\alpha}_{\mathtt{n}-\mathtt{dim}V} \\ 0 & \text{si } (\mathtt{n} - \mathtt{dim}V)\hat{\sigma}^{2,\mathrm{I}} \geq \sigma_0^2 \chi^{1-\alpha}_{\mathtt{n}-\mathtt{dim}V} \end{cases} ,$$

son UMP-invariantes a nivel α para contrastar las hipótesis iniciales H_0^2 y H_0^3, respectivamente. Sin embargo, cualquier test del tipo ϕ_2 o ϕ_3 no será siquiera insesgado para contrastar la hipótesis bilateral H_0^1, pues la función potencia del mismo será estrictamente creciente o decreciente, respectivamente[17].

Por otra parte y en virtud del teorema 3.7, el estadístico de la razón de verosimilitudes RV para contrastar las hipótesis iniciales H_0^i, $i = 1, 2, 3$, se expresa a través de T mediante

$$RV \propto T^{\mathtt{n}/2} \exp\left\{-\frac{1}{2}T\right\}.$$

Dado que la función $\varphi(x) := x^m \exp\{-x\}$ es creciente en $(0, m)$ y decreciente en $(m, +\infty)$, cualquier test del tipo (9.46) tendrá dos colas si lo expresamos en términos de T, de lo que se deduce que ϕ_2 y ϕ_3 no son tests de la razón de verosimilitudes para H_0^2 y H_0^3, respectivamente. No obstante, se prueba en Lehmann (1986) (sección

[16] Ver secciones 1.3 y 1.4

[17] Nogales (1998), pag. 185.

4.4) que una adecuada elección de las colas proporciona en un test insesgado a nivel α para contrastar la hipótesis inicial H_0^1, que será pues de la forma

$$\phi_1(Y) = \begin{cases} 1 & \text{si } \frac{\mathtt{n-dim}V}{\sigma_0^2}\hat{\sigma}^{2,\mathrm{I}} \in (0,A) \cup (B,+\infty) \\ 0 & \text{si } \frac{\mathtt{n-dim}V}{\sigma_0^2}\hat{\sigma}^{2,\mathrm{I}} \in [A,B] \end{cases}$$

para ciertos valores A y B[18], y que los test ϕ_1, ϕ_2 y ϕ_3 son UMP-insesgados a nivel α para contrastar las hipótesis H_0^1, H_0^2 y H_0^3, respectivamente.

3.4. Estudio asintótico del Modelo

En esta sección analizaremos el comportamiento de los estimadores y el test `F`, bajo las condiciones del Modelo Lineal (sin asumir en ningún momento normalidad), cuando el término `n` (que se corresponderá en la práctica con el número real de datos) tiende a infinito. Convendría repasar previamente las definiciones y resultados básicos de la Teoría Asintótica, en especial los distintos tipos de convergencias y sus relaciones, los conceptos de consistencia y eficiencia asintótica de un estimador, las Leyes de los Grandes Números y las diferentes versiones del Teorema Límite Central. Todo ello puede encontrarse, por ejemplo, en Ash (1972), Billingsley (1986), Fergusson (1996), Lehmann (1983) y Lehmann (1998). También recomendamos consultar el resumen que se encuentra en la última sección del Apéndice.

Hagamos previamente un inciso sobre una cuestión de carácter matricial. Dada una matriz (se admiten vectores) $A \in \mathcal{M}_{m\times k}$, de componentes a_{ij}, se define

$$m(A) = \max_{i,j} |a_{ij}|.$$

Si A es una matriz cuadrada de orden m, simétrica y semi definida positiva, existe, en virtud del teorema 9.5, una matriz B con las misma dimensiones tales que $A = B'B$. Si $b_1, \ldots, b_m$ denotan las columnas de B, se verifica

$$a_{ij} = \langle b_i, b_j \rangle, \quad a_{ii} = \|b_i\|^2, \quad a_{jj} = \|b_j\|^2.$$

Luego, por la desigualdad de Cauchy-Schwartz,

$$|a_{ij}| \leq \|b_i\| \cdot \|b_j\| = (|a_{ii}| \cdot |a_{jj}|)^{1/2} \leq \max_i |a_{ii}|.$$

Por lo tanto, en ese caso,

$$m(A) = \max_i |a_{ii}|.$$

[18]Existen resultados asintóticos que permiten aproximar A y B mediante $\chi^{2,1-\alpha/2}_{\mathtt{n-dim}V}$ y $\chi^{2,\alpha/2}_{\mathtt{n-dim}V}$ respectivamente.

También se verifica, trivialmente, que si $A \in \mathcal{M}_{m\times k}$ y $B \in \mathcal{M}_{k\times r}$,

$$m(AB) \leq km(A)m(B), \tag{3.28}$$

$$(m(A))^2 \leq m(AA'). \tag{3.29}$$

Teniendo en cuenta (3.28) junto con el teorema 9.4, se deduce que, si A es una matriz simétrica de orden k y D es la matriz diagonal constituida por sus autovalores, entonces

$$1/k^2 m(D) \leq m(A) \leq k^2 m(D). \tag{3.30}$$

Hasta ahora hemos trabajado con modelos en el cual el término $\mathtt{n}$ es fijo. Es lo que se denomina Modelo Exacto. Teniendo en cuenta que la Teoría Asintótica tiene como objeto estudiar la evolución de los distintos estimadores y tests de hipótesis en función de $\mathtt{n}$, es necesario construir un nuevo modelo, denominado Asíntótico, que, por así decirlo, englobe todos los experimentos exactos. En nuestro caso se definiría com sigue. Dada una sucesión $(V_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ de subespacios $\mathtt{v}$-dimensionales de $\mathbb{R}^n$, respectivamente, consideraremos el experimento estadístico constituido por una sucesión $(Z_i)_{i\in\mathbb{N}}$ de variables aleatorias que se descomponen de la siguiente forma

$$Z_i = \mu(i) + f_i, \quad i \in \mathbb{N},$$

donde $\mu(i) \in \mathbb{R}$ y $(f_i)_{i\in\mathbb{N}}$ es una secuencia de variables aleatorias independientes e idénticamente distribuidas con media 0 y varianza $\sigma^2 > 0$, y de tal forma que, para cada $\mathtt{n} \in \mathbb{N}$, el vector $\mu_{\mathtt{n}} = (\mu(1), \ldots, \mu(\mathtt{n}))'$ pertenece al subespacio $V_{\mathtt{n}}$. De esta forma, si se denota $Y_{\mathtt{n}} = (Z_1, \ldots, Z_{\mathtt{n}})$ y $e_{\mathtt{n}} = (f_1, \ldots, f_{\mathtt{n}})$, tendremos

$$Y_{\mathtt{n}} = \mu_{\mathtt{n}} + e_{\mathtt{n}}, \quad \mu_{\mathtt{n}} \in V_{\mathtt{n}}, \quad e_{\mathtt{n}} \sim \mathcal{P}_{\mathtt{n}},$$

siendo $\mathcal{P}_{\mathtt{n}}$ la familia compuesta por las potencias $\mathtt{n}$-ésimas de distribuciones de media 0 y varianza finita. Nótese que, para cada $n \in \mathbb{N}$, tenemos un Modelo Lineal Exacto en dimensión $\mathtt{n}$. Por lo tanto, tiene sentido hablar de los estimadores

$$\hat{\mu}_{\mathtt{n}} = P_{V_{\mathtt{n}}} Y_{\mathtt{n}}, \qquad \hat{\sigma}^{2,\mathtt{I}}{}_{\mathtt{n}} = \frac{1}{\mathtt{n}-\mathtt{v}} \|P_{V_{\mathtt{n}}^{\perp}} Y_{\mathtt{n}}\|^2 = \frac{1}{\mathtt{n}-\mathtt{v}} \|P_{V_{\mathtt{n}}^{\perp}} e_{\mathtt{n}}\|^2.$$

Así mismo y en lo que respecta a la problema de Contraste de Hipótesis, si consideramos una secuencia $(W_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ de subespacios $\mathtt{w}$-dimensionales de $(V_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$, respectivamente, tendrá sentido hablar del estadístico de contraste $F_{\mathtt{n}}$, definido en los términos de (3.23).

Nótese que, al contrario de lo que sucede en el Modelo Lineal Normal Exacto, el Modelo Lineal Asintótico no queda parametrizado por un vector media, μ, y una

varianza σ^2. Si acaso, podríamos hablar de una sucesión de medias $(\mu_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ y una varianza σ^2. Por ello, tiene aquí sentido hablar de una secuencia de estimadores consistente para σ^2, pero no para μ. Este problema, que afecta al estudio de Estimación, podría resolverse si consideráramos el Modelo Asintótico que resulta de imponer a $(\mu_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ la siguiente restricción: suponer que existe una sucesión $(\mathbf{X}_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ de bases de $(V_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$, de manera que $(\mu_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ verifica

$$\exists \beta \in \mathbb{R}^{\mathtt{v}} : \ \mu_{\mathtt{n}} = \mathbf{X}_{\mathtt{n}}\beta, \ \forall \mathtt{n} \in \mathbb{N}. \tag{3.31}$$

De esta forma, sí tendría sentido hablar de una secuencia de estimadores consistente para β. Consideremos, concretamente, la secuencia definida mediante

$$\hat{\beta}_{\mathtt{n}} = (\mathbf{X}'_{\mathtt{n}}\mathbf{X}_{\mathtt{n}})^{-1}\mathbf{X}'_{\mathtt{n}}Y_{\mathtt{n}}, \qquad \mathtt{n} \in \mathbb{N}.$$

Se verifica entonces lo siguiente.

Teorema 3.14.
En las condiciones anteriores, si se verifica la hipótesis

$$m\left(\mathbf{X}'_{\mathtt{n}}\mathbf{X}_{\mathtt{n}}\right) \longrightarrow \infty \tag{3.32}$$

la secuencia de estimadores $\left(\hat{\beta}_{\mathtt{n}}\right)_{\mathtt{n}\in\mathbb{N}}$ es consistente.

Demostración.
Tener en cuenta, primeramente, que

$$\mathtt{E}\left[\hat{\beta}_{\mathtt{n}}\right] = \beta, \qquad \mathtt{Cov}\left[\hat{\beta}_{\mathtt{n}}\right] = \sigma^2(\mathbf{X}'_{\mathtt{n}}\mathbf{X}_{\mathtt{n}})^{-1}, \quad \forall \mathtt{n} \in \mathbb{N}.$$

Por lo tanto, dado $\varepsilon > 0$, se sigue de la Desigualdad de Chebyshev[19] que

$$P\big(\big\|\hat{\beta}_{\mathtt{n}} - \beta\big\| > \varepsilon\big) \leq \frac{\sqrt{\mathtt{v}}\sigma^2 \cdot m((\mathbf{X}'_n\mathbf{X}_{\mathtt{n}})^{-1})}{\varepsilon}.$$

Sea $D_{\mathtt{n}}$ la matriz diagonal de los autovalores de $\mathbf{X}'_{\mathtt{n}}\mathbf{X}_{\mathtt{n}}$, para cada $\mathtt{n} \in \mathbb{N}$. Por el teorema 9.4, la matriz de los autovalores de $(\mathbf{X}'_{\mathtt{n}}\mathbf{X}_{\mathtt{n}})^{-1}$ será $D_{\mathtt{n}}^{-1}$. Luego, teniendo en cuenta (3.30), se verifica que $m((\mathbf{X}'_{\mathtt{n}}\mathbf{X}_{\mathtt{n}})^{-1}) \to 0$, lo cual concluye la prueba. ∎

Veamos qué podemos decir respecto a la estimación de σ^2.

[19] Si X en una variable aleatoria real con momento de segundo orden finito en un espacio de probabilidad y $\eta > 0$, entonces $P(|X - \mathtt{E}[X]| > \eta) \leq \mathtt{var}[X]/\eta^2$.

Teorema 3.15.
La secuencia $\left(\hat{\sigma}^{2,\mathrm{I}}{}_{\mathtt{n}}\right)_{\mathtt{n}\in\mathbb{N}}$ de estimadores de σ^2 es consistente.

Demostración.
Se verifica que

$$\frac{\mathtt{n}-\mathtt{v}}{\mathtt{n}}\,\hat{\sigma}^{2,\mathrm{I}} = \frac{\|P_{V_{\mathtt{n}}^{\perp}} e_{\mathtt{n}}\|^2}{\mathtt{n}} = \frac{\|e_{\mathtt{n}}\|^2}{\mathtt{n}} - \frac{\|P_{V_{\mathtt{n}}} e_{\mathtt{n}}\|^2}{\mathtt{n}}.$$

Teniendo en cuenta que, $\mathtt{E}[e_{\mathtt{n}}] = 0$ y $\mathtt{Cov}[e_{\mathtt{n}}] = \sigma^2\mathtt{Id}$, se deduce que $\mathtt{E}[P_{V_{\mathtt{n}}} e_{\mathtt{n}}] = 0$ y $\mathtt{Cov}[P_{V_{\mathtt{n}}} e_{\mathtt{n}}] = \sigma^2 P_{V_{\mathtt{n}}}$. Entonces, del lema 3.1 y de la proposición 9.16 se sigue que

$$\frac{\mathtt{E}\left[\|P_{V_{\mathtt{n}}} e_{\mathtt{n}}\|^2\right]}{\mathtt{n}} = \sigma^2 \frac{\mathtt{tr}(P_{V_{\mathtt{n}}})}{\mathtt{n}} = \sigma^2 \frac{\mathtt{v}}{\mathtt{n}}.$$

Dado $\varepsilon > 0$, se verifica triviamente, para todo $n \in \mathbb{N}$, que

$$P(\|P_{V_{\mathtt{n}}} e_{\mathtt{n}}\|^2/n > \varepsilon) < \frac{\mathtt{E}\left[\|P_{V_{\mathtt{n}}} e_{\mathtt{n}}\|^2\right]/n}{\varepsilon}.$$

En consecuencia,

$$\frac{\|P_{V_{\mathtt{n}}} e_{\mathtt{n}}\|^2}{\mathtt{n}} \xrightarrow{P} 0 \ ^{20}$$

Dado que $(f_i^2)_{i\in\mathbb{N}}$ constituye una sucesión de variables aleatorias iid de media σ^2, se verifica, en virtud de LDGN,

$$\frac{\|e_{\mathtt{n}}\|^2}{\mathtt{n}} = \frac{1}{\mathtt{n}} \sum_{i=1}^{\mathtt{n}} f_i^2 \xrightarrow{P} 0. \tag{3.33}$$

Entonces, se sigue del teorema 9.21 que

$$\hat{\sigma}^{2,\mathrm{I}}{}_{\mathtt{n}} = \frac{\mathtt{n}}{\mathtt{n}-\mathtt{v}} \left(\frac{\|e_{\mathtt{n}}\|^2}{\mathtt{n}} - \frac{\|P_{V_{\mathtt{n}}} e_{\mathtt{n}}\|^2}{\mathtt{n}} \right) \xrightarrow{P} 0.$$

∎

Obviamente, obtendremos también una secuencia consistente si utilizamos el estimador (3.9). El siguiente resultado se sigue del Teorema Central del Límite.

Teorema 3.16.
Sea $(a_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ una secuencia de vectores tales que $a_{\mathtt{n}} \in \mathbb{R}^{\mathtt{n}}$ y $\|a_{\mathtt{n}}\| = 1$, para todo $n \in \mathbb{N}$. Si $m(a_{\mathtt{n}}) \to 0$, entonces

$$a'_{\mathtt{n}} e_{\mathtt{n}} \xrightarrow{d} N(0, \sigma^2).$$

[20] Nótese que esta afirmación sigue siendo válida si sustituimos en el denominador $\mathtt{n}$ por $\sqrt{\mathtt{n}}$

Demostración.
Para cada $n \in \mathbb{N}$, consideremos la descomposición $a'_{\mathtt{n}} = (a_{\mathtt{n}1}, \ldots, a_{\mathtt{n}n})$, y sea $X_{\mathtt{n}i} = a_{\mathtt{n}i} f_i$, $1 \leq i \leq n$. En ese caso, $a'_{\mathtt{n}} e_{\mathtt{n}} = \sum_{i=1}^{\mathtt{n}} X_{\mathtt{n}i}$. Todos los $X_{\mathtt{n}i}$, $1 \leq i \leq n$, son independientes por serlo las f_i. Además,

$$\mathtt{E}[X_{\mathtt{n}i}] = 0, \quad \mathtt{var}[X_{\mathtt{n}i}] = a_{\mathtt{n}i}\sigma^2, \quad \sum_{i=1}^{\mathtt{n}} \mathtt{var}[X_{\mathtt{n}i}] = \sigma^2.$$

Por lo tanto, para demostrar la tesis basta probar que se verifica la hipótesis (9.66) del teorema 9.26, es decir,

$$C_{\mathtt{n}} = \frac{1}{\sigma^2} \sum_{i=1}^{\mathtt{n}} \mathtt{E}\big[X_{\mathtt{n}i}^2 I_\varepsilon(X_{\mathtt{n}i})\big] \to 0. \tag{3.34}$$

Efectivamente, si $m_{\mathtt{n}} = m(a_{\mathtt{n}})$, se verifica[21]

$$\begin{aligned} \sigma^2 C_{\mathtt{n}} &= \sum_{i=1}^{\mathtt{n}} \mathtt{E}\big[X_{\mathtt{n}i}^2 I_\varepsilon(X_{\mathtt{n}i})\big] = \sum_{i=1}^{\mathtt{n}} a_{\mathtt{n}i}^2 \mathtt{E}\big[f_i^2 I_{\varepsilon/a_{\mathtt{n}i}} \circ f_i\big] \leq \sum_{i=1}^{\mathtt{n}} a_{\mathtt{n}i}^2 \mathtt{E}\big[f_i^2 I_{\varepsilon/m_{\mathtt{n}}} \circ f_i\big] \\ &= \sum_{i=1}^{\mathtt{n}} a_{\mathtt{n}i}^2 \mathtt{E}\big[f_1^2 I_{\varepsilon/m_{\mathtt{n}}} \circ f_1\big] = \mathtt{E}\big[f_1^2 I_{\varepsilon/m_{\mathtt{n}}} \circ f_1\big]. \end{aligned}$$

Teniendo en cuenta que $|f_1^2 I_{\varepsilon/m_{\mathtt{n}}} \circ f_1| < f_1^2$ y que f_1^2 es integrable, se deduce del Teorema de la Convergencia Dominada que

$$\lim_{\mathtt{n}\to\infty} C_{\mathtt{n}} = \int \lim_{\mathtt{n}\to\infty} (f_1^2 I_{\varepsilon/m_{\mathtt{n}}} \circ f_1)\, dP.$$

dado que $m_{\mathtt{n}}$ converge a 0, por hipótesis, el integrando converge puntualmente a 0, con lo cual acabamos.

∎

Como consecuencia obtenemos el siguiente resultado, de gran utilidad tanto para el problema de Estimación como de Contraste de Hipótesis.

Lema 3.17.

Sea $(\Gamma_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ una sucesión de matrices de dimensión $\mathtt{n} \times u$, respectivamente, tales que $\Gamma'_{\mathtt{n}}\Gamma_{\mathtt{n}} = \mathtt{Id}$, para todo $\mathtt{n} \in \mathbb{N}$ y $m(\Gamma_{\mathtt{n}}\Gamma'_{\mathtt{n}})$ converge a 0. Entonces,

$$\Gamma'_{\mathtt{n}} e_{\mathtt{n}} \xrightarrow{d} N_u(0, \sigma^2 \mathtt{Id}).$$

[21] Cuando $a_{\mathtt{n}i}$ valga 0, considerar el sumando correspondiente como nulo.

Demostración.
Dado $c \in\in \mathbb{R}^u$ tal que $\|c\| = 1$, consideremos, para cada $n \in \mathbb{N}$, el vector $a_{\mathtt{n}} = \Gamma_{\mathtt{n}} c \in \mathbb{R}^{\mathtt{n}}$, verificando también $\|a_{\mathtt{n}}\| = 1$. Dado que $m(c) \leq 1$, se tiene que $m(a_{\mathtt{n}})$ converge a 0, pues

$$m(a_{\mathtt{n}}) \leq u \cdot m(c) \cdot m(\Gamma_{\mathtt{n}}) \leq u(m(\Gamma_{\mathtt{n}}\Gamma'_{\mathtt{n}}))^{1/2}.$$

Luego, por el teorema 3.16,

$$c'\Gamma'_{\mathtt{n}} e_{\mathtt{n}} \xrightarrow{d} N(0, \sigma^2).$$

Teniendo en cuenta la Astucia de Cramer-Wold (teorema 9.21-(x)), se concluye. ∎

El siguiente resultado, muy interesante desde el punto e vista de la Estimación, se obtiene como corolario del anterior.

Teorema 3.18.

Supongamos que se verifica (3.31) junto con la siguiente propiedad

$$\lim_{\mathtt{n}\to\infty} m\left(\mathtt{X}_{\mathtt{n}}(\mathtt{X}'_{\mathtt{n}}\mathtt{X}_{\mathtt{n}})^{-1}\mathtt{X}'_{\mathtt{n}}\right) = 0. \tag{3.35}$$

Entonces,

(i) $(\mathtt{X}'_{\mathtt{n}}\mathtt{X}_{\mathtt{n}})^{1/2}\left(\hat{\beta}_{\mathtt{n}} - \beta\right) \xrightarrow{d} N_{\mathtt{v}}(0, \sigma^2 \mathtt{Id})$.

(ii) Para todo $\alpha \in (0,1)$, $\lim_{\mathtt{n}\to\infty} P(\mathcal{E}^{\mathtt{n}}_{\alpha}) = 1 - \alpha$, donde

$$\mathcal{E}^{\mathtt{n}}_{\alpha} = \left\{ b \in \mathbb{R}^{\mathtt{v}} : \ \left(\hat{\beta}_{\mathtt{n}} - b\right)' \mathtt{X}'_{\mathtt{n}}\mathtt{X}_{\mathtt{n}} \left(\hat{\beta}_{\mathtt{n}} - b\right) \leq \hat{\sigma}^{2,\mathtt{I}}{}_{\mathtt{n}} \ \chi^{2,\alpha}_{\mathtt{v}} \right\} \tag{3.36}$$

Demostración.
(i) Si para cada $\mathtt{n} \in \mathbb{N}$ consideramos la matriz $\Gamma_{\mathtt{n}} = \mathtt{X}_{\mathtt{n}}(\mathtt{X}'_{\mathtt{n}}\mathtt{X}_{\mathtt{n}})^{-1/2}$, entonces $(\Gamma_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ satisface las hipótesis del lema anterior con $u = \mathtt{v}$. Por lo tanto,

$$\Gamma'_{\mathtt{n}} e_{\mathtt{n}} \xrightarrow{d} N_{\mathtt{v}}(0, \sigma^2).$$

Teniendo en cuenta que,

$$\hat{\beta}_n - \beta = (\mathtt{X}'_{\mathtt{n}}\mathtt{X}_{\mathtt{n}})^{-1}\mathtt{X}'_{\mathtt{n}}(Y_{\mathtt{n}} - \mu_n),$$

se deduce

$$(\mathtt{X}'_{\mathtt{n}}\mathtt{X}_{\mathtt{n}})^{1/2}\left(\hat{\beta}_{\mathtt{n}} - \beta\right) = \Gamma'_{\mathtt{n}} e_{\mathtt{n}},$$

con lo cual se acaba la primera parte.
(ii) Del apartado anterior se deduce que

$$\frac{1}{\hat{\sigma}^{2,\mathrm{I}}}\left(\hat{\beta}_{\mathtt{n}}-\beta\right)'\mathbf{X}_{\mathtt{n}}'\mathbf{X}_{\mathtt{n}}\left(\hat{\beta}_{\mathtt{n}}-\beta\right)\frac{\hat{\sigma}^{2,\mathrm{I}}}{\sigma^2}\xrightarrow{d}\chi^2_{\mathtt{v}}.$$

Teniendo en cuenta que $\hat{\sigma}^{2,\mathrm{I}}{}_{\mathtt{n}}$ converge a σ^2 en probabilidad, junto con el teorema 9.21-(ix), podemos despreciar el último factor del primer término y, aplicando el teorema 9.21-(ii) acabamos.

∎

Nótese que, de (i) se sigue que, para $\mathtt{n}$ *suficientemente* grande, el estadístico $\hat{\beta}_{\mathtt{n}}$ sigue aproximadamente un modelo de distribución $N_{\mathtt{v}}(\beta,\sigma^2(\mathbf{X}_{\mathtt{n}}'\mathbf{X}_{\mathtt{n}})^{-1})$. En ese sentido podemos decir que el la proposición (iii) del teorema 3.15 es asintóticamente válida para el Modelo Lineal, supuesto que se satisfaga la condición (3.35). Lo mismo puede decirse, por (ii), de la región de confianza (3.16).

Respecto al test `F`, que es el de la razón de verosimilitudes, sabemos, en virtud del teorema 9.28, que puede expresarse asintóticamente haciendo uso de la distribución χ^2 con $\mathtt{dim}V-\mathtt{dim}W$ grados de libertad. Veremos a continuación cómo podemos extender este resultado asintótico al Modelo Lineal (sin suponer normalidad).

Teorema 3.19.

Si $(U_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ es una sucesión de subespacios de $\mathbb{R}^{\mathtt{n}}$, respectivamente, de dimensión $u\in\mathbb{N}$, y tal que

$$m(P_{U_{\mathtt{n}}})\to 0, \tag{3.37}$$

Entonces

$$\frac{\|P_{U_{\mathtt{n}}}e_{\mathtt{n}}\|^2}{\sigma^2}\xrightarrow{d}\chi^2_u.$$

Demostración.
Es consecuencia directa del lema 3.17, considerando una base ortonormal de cada subespacio $(U_{\mathtt{n}})$, $\mathtt{n}\in\mathbb{N}$.

∎

La hipótesis (3.37), que desempeña un papel crucial en nuestra teoría, se conoce normalmente como Condición de Huber y puede considerarse una suerte de *traducción* de la condición de Lindemberg (9.66) al Modelo Lineal. En capítulos posteriores, cuando abordemos estudios más específicos como son la regresión lineal o el diseño de experimentos, veremos en qué se traduce dicha hipótesis para cada caso. Este resultado permitirá extender, en los términos de la Teoría Asintótica, el test `F` al Modelo Lineal (sin suponer normalidad). Si el modelo verifica la condición (3.31), la

condición de Huber equivale a (3.35), y confiere, como ya hemos visto, normalidad asintótica al estimador de β y validez asintótica a la región de confianza (3.16).

Teorema 3.20.
En las condiciones del Modelo Lineal Asintótico, si $(W_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ es una sucesión de subespacio lineales de $(V_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$, respectivamente, todos ellos de dimensión $\mathtt{w}$, y $(V_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ satisface la condición de Huber (3.37), entonces

$$F^*_{\mathtt{n}} = \frac{\|P_{V_{\mathtt{n}}|W_{\mathtt{n}}}(Y_{\mathtt{n}} - \mu_{\mathtt{n}})\|^2}{(\mathtt{v}-\mathtt{w})\hat{\sigma}^{2,\mathrm{I}}{}_{\mathtt{n}}} \xrightarrow{d} \frac{1}{\mathtt{v}-\mathtt{w}}\chi^2_{\mathtt{v}-\mathtt{w}}.$$

Demostración.
Dado que $P_{V_{\mathtt{n}}} = P_{W_{\mathtt{n}}} + P_{V_{\mathtt{n}}|W_{\mathtt{n}}}$, se tiene que $(V_{\mathtt{n}}|W_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ verifica igualmente la condición (3.37) luego, por el teorema 3.16,

$$\frac{\|P_{V_{\mathtt{n}}|W_{\mathtt{n}}}(Y_{\mathtt{n}} - \mu_{\mathtt{n}})\|^2}{\sigma^2} = \frac{\|P_{V_{\mathtt{n}}|W_{\mathtt{n}}}e_{\mathtt{n}}\|^2}{\sigma^2} \xrightarrow{d} \chi^2_{\mathtt{v}-\mathtt{w}}.$$

Por otro lado, se sigue del teorema 3.15 que

$$\frac{\hat{\sigma}^{2,\mathrm{I}}{}_{\mathtt{n}}}{\sigma^2} \xrightarrow{P} 1.$$

Aplicando el teorema 9.21, se obtiene

$$F^*_{\mathtt{n}} = \frac{\|P_{V_{\mathtt{n}}|W_{\mathtt{n}}}(Y_{\mathtt{n}} - \mu_{\mathtt{n}})\|^2}{(\mathtt{v}-\mathtt{w})\sigma}\frac{\hat{\sigma}^{2,\mathrm{I}}{}_{\mathtt{n}}}{\sigma^2} \xrightarrow{d} \frac{1}{\mathtt{v}-\mathtt{w}}\chi^2_{\mathtt{v}-\mathtt{w}}.$$

∎

Corolario 3.21.
En las condiciones del teorema anterior, y si $\mu_{\mathtt{n}} \in W_{\mathtt{n}}$ para todo $n \in \mathbb{N}$, se verifica

$$F_{\mathtt{n}} \xrightarrow{d} \frac{1}{\mathtt{v}-\mathtt{w}}\chi^2_{\mathtt{v}-\mathtt{w}}.$$

Demostración.
Basta aplicar el teorema anterior teniendo en cuenta que (3.23) y que $P_{V_{\mathtt{n}}|W_{\mathtt{n}}}\mu_{\mathtt{n}} = 0$, para todo $n \in \mathbb{N}$.

∎

En virtud de este resultado se verifica que, si se satisface la condición de Huber junto con la hipótesis nula ($\mu_{\mathtt{n}} \in W_{\mathtt{n}}$, para todo $n \in \mathbb{N}$), y se considera sucesión de tests $(\mathtt{F}_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$, definidos mediante

$$\mathtt{F}_{\mathtt{n}}(Y_{\mathtt{n}}) = \begin{cases} 1 & \text{si } F_{\mathtt{n}}(Y) > \frac{\chi^{2,\alpha}_{\mathtt{v}-\mathtt{w}}}{\mathtt{v}-\mathtt{w}} \\ 0 & \text{si } F_{\mathtt{n}}(Y) \leq \frac{\chi^{2,\alpha}_{\mathtt{v}-\mathtt{w}}}{\mathtt{v}-\mathtt{w}} \end{cases},$$

entonces

$$\lim_{\mathtt{n}\to\infty} P(\{F_{\mathtt{n}} = 1\}) = 1 - \alpha.$$

Por ello, si consideramos el contraste de hipótesis $H_0 : \mu \in W$ en un Modelo Lineal, siendo `n` *suficientemente* grande, se verifica que el nivel de significación del test `F` definido en (3.19) es *aproximadamente* igual a α. En ese sentido decimos que el test `F` es asintóticamente válido, aunque no se verifique el supuesto de normalidad, siempre y cuando se satisfaga la condición de Huber. Por otra parte, la distribución límite de F en el caso nulo corresponde, como cabría esperar, a la distribución asintótica del para el TRV, según se refleja en el teorema 9.28.

Podemos ir un poco más lejos. Se prueba en Arnold (1981)[22] que, si se verifica la condición (3.37) y, además, existe $\gamma > 0$ tal que

$$\lim_{\mathtt{n}\to\infty} \|P_{V_{\mathtt{n}}|W_{\mathtt{n}}}\mu_{\mathtt{n}}\| = \gamma, \tag{3.38}$$

entonces

$$F_{\mathtt{n}} \xrightarrow{d} \frac{1}{\mathtt{v}-\mathtt{w}}\chi^2_{\mathtt{v}-\mathtt{w}}\left(\frac{\gamma^2}{\sigma^2}\right).$$

Ello permite construir la función potencia asintótica para todos los valores del parámetro verificando la condición (3.38). Curiosamente, puede comprobarse que, si se plantea el contraste de la hipótesis $H_0 : \mu \in W$ suponiendo normalidad y varianza σ^2 conocida (estamos hablando pues de otro modelo), se obtiene[23] un test *óptimo* a nivel α (UMP-invariante) cuyo estadístico de contraste

$$F^* = \frac{\|P_{V|W}Y\|^2}{\sigma^2} \tag{3.39}$$

sigue una distribución

$$\chi^2_{\mathtt{dim}V-\mathtt{dim}W}\left(\frac{\|P_{V|W}\mu\|^2}{\sigma^2}\right), \quad \forall(\mu,\sigma^2) \in V \times \mathbb{R}^+.$$

En ese sentido podríamos decir que, si se cumple la condición de Huber, la potencia asintótica del test `F` en el Modelo Lineal para los valores del parámetro que verifican (3.38), en particular en el caso nulo, coincide con la del test óptimo que se obtiene suponiendo normalidad y varianza conocida[24].

Ya hemos visto cómo se comporta asintóticamente el Modelo Lineal en lo que respecta a los contraste de hipótesis sobre μ. Veamos ahora en qué medida el uso de

[22] Capítulo 10, ejercicio C1.

[23] Arnold (1981), sección 7.11

[24] Heurísticamente hablando, podríamos decir qu la violación de la normalidad y el desconocimiento de la varianza pueden ser, de alguna manera, obviados para muestras suficientemente grandes.

una cantidad suficientemente grande de datos puede permitirnos obviar el supuesto de normalidad a la hora de construir un test de hipótesis o un intervalo de confianza para σ^2. Supongamos que las variables f_i poseen momento de orden 4 y sea entonces δ el coeficiente definido mediante

$$\delta = \frac{\mathtt{E}[f_1^4]}{(\mathtt{E}[f_1^2])^2} = \frac{\mathtt{E}[f_1^4]}{\sigma^4}.^{25} \tag{3.40}$$

Teorema 3.22.
En las condiciones anteriores se verifica

$$\sqrt{\mathtt{n}}\left(\hat{\sigma}^{2,\mathrm{I}}{}_{\mathtt{n}} - \sigma^2\right) \xrightarrow{d} N\big(0, \sigma^4(\delta - 1)\big)$$

Demostración.
Se sigue la demostración del teorema 3.15, pero al llegar a (3.33) aplicamos TCL (caso iid) en lugar de LDGN, con lo cual se tiene que

$$\sqrt{\mathtt{n}}\left(\frac{\|e_{\mathtt{n}}\|^2}{\mathtt{n}} - \sigma^2\right) \xrightarrow{d} N\left(0, \sigma^4(\delta - 1)\right).$$

Entonces, teniendo en cuenta que $\mathtt{n}^{-1/2}\|P_{V_{\mathtt{n}}} e_{\mathtt{n}}\|^2$ converge a 0 en probabilidad y que $(\mathtt{n} - \mathtt{v})/\mathtt{n}$ converge a 1, basta aplicar el teorema 9.21-(ix) para concluir. ∎

Operando en la expresión obtenida pueden obtenerse, mediante la distribución $N(0,1)$, tests de hipótesis e intervalos de confianza con validez asintótica, siempre y cuando δ, denominado coeficiente de Kurtosis, sea conocido, cosa poco verosímil. Por ejemplo, puede comprobarse que, si f_1 sigue una distribución normal, entonces $\delta = 3$, con lo cual el problema estaría resuelto desde el punto de vista asintótico, lo cual no aporta mucho, puesto que el problema ya está resuelto también en el Modelo Exacto. No obstante, el resultado anterior tiene interesantes aplicaciones. A modo de ejemplo, haremos uso del mismo para construir el denominado test de Barlett de igualdad de varianzas, que será de utilidad en capítulos posteriores.

Consideremos k vectores aleatorios independientes

$$Y_i \sim N_{\mathtt{n}_i}(\mu_i, \sigma_i^2), \quad \mu_i \in V_i,\ \sigma_i^2 > 0, \quad i = 1, \ldots, k$$

siendo cada V_i un subespacio $\mathtt{v}_i$-dimensional de $\mathbb{R}^{\mathtt{n}_i}$. Supongamos que queremos contrastar la hipótesis inicial $H_0 : \sigma_1 = \ldots = \sigma_k$, lo cual permitiría componer un Modelo

[25] De la desigualdad de Holder se sigue trivialmente que $\delta \geq 1$.

Lineal Normal en dimensión $\mathtt{n} = \sum_i \mathtt{n}_i$. Una reducción por suficiencia y otra por invarianza[26] en el modelo producto nos llevan a considerar el estadístico

$$(\hat{\sigma}^{2,\mathrm{I}}{}_1, \ldots, \hat{\sigma}^{2,\mathrm{I}}{}_k)\,.$$

El teorema anterior (con $\delta = 3$) garantiza que

$$\sqrt{\mathtt{n}_i}\left(\hat{\sigma}^{2,\mathrm{I}}{}_{i,\mathtt{n}_i} - \sigma_i^2\right) \xrightarrow{d} N(0, 2\sigma_i^4), \quad i = 1, \ldots, k.$$

Si consideramos la transformación $g(x) = (\log x)/\sqrt{2}$ y aplicamos el Método Delta (teorema 9.27), se tiene que

$$\frac{\sqrt{\mathtt{n}_i}}{2}\left(\log \hat{\sigma}^{2,\mathrm{I}}{}_{i,\mathtt{n}_i} - \log \sigma_i^2\right) \xrightarrow{d} N(0,1), \quad i = 1, \ldots, k$$

siendo además secuencias independientes. Es decir,

$$\begin{pmatrix} \frac{\sqrt{\mathtt{n}_1}}{2}\left(\log \hat{\sigma}^{2,\mathrm{I}}{}_{1,\mathtt{n}_1} - \log \sigma_1^2\right) \\ \vdots \\ \frac{\sqrt{\mathtt{n}_k}}{2}\left(\log \hat{\sigma}^{2,\mathrm{I}}{}_{k,\mathtt{n}_k} - \log \sigma_k^2\right) \end{pmatrix} \xrightarrow{d} N_k(0, \mathtt{Id}). \tag{3.41}$$

Por lo tanto, para valores de $\mathtt{n}_1, \ldots, \mathtt{n}_k$ suficientemente grandes, se tiene que el vector aleatorio

$$\mathtt{T} = \begin{pmatrix} \frac{\sqrt{\mathtt{n}_1}}{2} \log \hat{\sigma}^{2,\mathrm{I}}{}_{1,\mathtt{n}_1} \\ \vdots \\ \frac{\sqrt{\mathtt{n}_k}}{2} \log \hat{\sigma}^{2,\mathrm{I}}{}_{k,\mathtt{n}_k} \end{pmatrix}$$

sigue, aproximadamente, una distribución $N_k(\theta, \mathtt{Id})$, donde θ puede ser cualquier vector de $\mathbb{R}^k$, pues su componente i-ésima es

$$\theta_i = \frac{\sqrt{\mathtt{n}_i}}{2} \log \sigma_i^2, \quad i = 1, \ldots, k.$$

Podemos pues considerar un nuevo modelo que se define mediante

$$Z \sim N_k(\theta, \mathtt{Id}), \quad \theta \in \mathbb{R}^k, \quad \sigma^2 > 0. \tag{3.42}$$

En este modelo podemos contrastar la hipótesis inicial

$$\theta \in W = \left\langle \begin{pmatrix} \sqrt{\mathtt{n}_1} \\ \vdots \\ \sqrt{\mathtt{n}_k} \end{pmatrix} \right\rangle.$$

[26] Respecto al grupo de las traslaciones.

Ya hemos comentado con anterioridad que en un modelo de este tipo (con varianza conocida), el contraste se resuelve de manera óptima mediante el estadístico (3.39). En nuestro caso,

$$F^* = \|P_{W^\perp} Z\|^2 \sim \chi^2_{k-1}\left(\|P_{W^\perp}\theta\|^2\right).$$

Dado $\alpha \in (0,1)$, el test siguiente es entonces UMP-invariante a nivel α en el modelo (3.42) para contrastar la hipótesis inicial $\theta \in W$:

$$\phi^* = \begin{cases} 1 & \text{si } F^* > \chi^{2,\alpha}_{k-1} \\ 0 & \text{si } F^* \leq \chi^{2,\alpha}_{k-1} \end{cases}.$$

Ahora bien, puede comprobarse fácilmente que, en virtud de (3.41), si H_0 es cierto, la distribución del estadístico $F^* \circ \mathtt{T}$ converge a χ^2_{k-1} cuando $\mathtt{n}_i$ tiende a infinito, para todo $i = 1, \ldots, k$. Por lo tanto, el nivel de significación del test $\phi = \phi^* \circ T$, construido a partir de un test óptimo a nivel α en el modelo límite (3.42), converge a α cuando $\mathtt{n}_i$ converge a infinito para todo $i = 1, \ldots, k$, es decir, que es asintóticamente válido. Falta sólo determinar una expresión más apropiada para el estadístico de contraste $F^* \circ \mathtt{T}$. Concretamente, consideremos $\mathtt{n}_i$, $i = 1, \ldots, k$ fijos. Si se denota

$$\overset{\bullet}{\sigma} = \left(\prod_{j=1}^{k} \hat{\sigma}_i^{\mathtt{n}_i}\right)^{\frac{1}{\mathtt{n}}},$$

el estadístico $F^* \circ \mathtt{T}$ se expresa mediante

$$\begin{aligned} F^* \circ \mathtt{T} &= \|P_{W^\perp}\mathtt{T}\|^2 = \sum_{i=1}^{k}\left(\mathtt{T}_i - \sqrt{\mathtt{n}_i}\frac{\sum_{j=1}^{k}\sqrt{\mathtt{n}_j}\mathtt{T}_j}{\mathtt{n}}\right)^2 \\ &= \sum_{i=1}^{k} \mathtt{n}_i \left(\log\frac{\hat{\sigma}_i}{\overset{\bullet}{\sigma}}\right)^2. \end{aligned}$$

Por lo tanto, el test de Barlett de igualdad de varianzas a nivel α es el siguiente

$$\phi = \begin{cases} 1 & \text{si } \sum_{i=1}^{k} \mathtt{n}_i \left(\log\frac{\hat{\sigma}_i}{\overset{\bullet}{\sigma}}\right)^2 > \chi^{2,\alpha}_{k-1} \\ 0 & \text{si } \sum_{i=1}^{k} \mathtt{n}_i \left(\log\frac{\hat{\sigma}_i}{\overset{\bullet}{\sigma}}\right)^2 \leq \chi^{2,\alpha}_{k-1} \end{cases}.$$

No obstante, hemos de recalcar que este test puede considerarse válido para muestras suficientemente grandes y suponiendo que se verifique la hipótesis de normalidad (recordemos que hemos supuesto $\delta = 3$). De hecho, el test resulta ser bastante sensible ante la violación de dicho supuesto, cosa bastante común en buena parte de los tests clásicos relativos a la varianza (o la matriz de varianzas-covarianzas en el caso multivariante).

3.5. Intervalos de confianza simultáneos

Para acabar el capítulo dedicado al análisis del Modelo Lineal desde un punto de vista puramente teórico, abordaremos el estudio general de las familias de intervalos de confianza simultáneos, lo cual nos conducirá a los métodos de Scheffé y Bonferroni, a los cuales se añadirá en el capítulo 6 el de Tuckey, de carácter más específico. Primeramente, hemos de aclarar el concepto en sí.

Dado un modelo estadístico $(\Omega, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$, un conjunto Λ de estimandos reales y $\alpha \in (0,1)$, una familia de intervalos de confianza simultáneos a nivel $1-\alpha$ para Λ es una colección de pares de estadísticos reales $\mathcal{I}_\Lambda^\alpha = \{(a_\lambda^\alpha, b_\lambda^\alpha) : \lambda \in \Lambda\}$, tal que

$$P_\theta\left(\omega \in \Omega :\ a_\lambda^\alpha(\omega) \leq \lambda(\theta) \leq b_\lambda^\alpha(\omega),\ \forall \lambda \in \Lambda\right) = 1-\alpha, \quad \forall \theta \in \Theta.$$

Consideremos un Modelo Lineal Normal

$$Y \sim N_{\mathtt{n}}(\mu, \sigma^2 \mathtt{Id}), \quad \mu \in V,\ \sigma^2 > 0,$$

y una hipótesis inicial $H_0 : \mu \in W$, para algún $W \in V$. Se denomina *contraste* a cualquier elemento del subespacio $V|W$. Nuestro objetivo es, dado $\alpha \in (0,1)$, construir una familia de intervalos de confianza simultáneos a nivel $1-\alpha$ para el conjunto $[V|W] = \{\lambda_d :\ d \in V|W\}$, donde

$$\lambda_d(\mu, \sigma^2) = \langle d, \mu \rangle \qquad , \forall d \in V|W,\ \forall (\mu, \sigma^2) \in V \times \mathbb{R}^+.$$

Necesitamos un lema previo.

Lema 3.23.
Si $x \in \mathbb{R}^n$ y $E \subset \mathbb{R}^n$, entonces

$$\sup_{e \in E \setminus \{0\}} \frac{\langle e, x \rangle^2}{\|e\|^2} = \|P_E x\|^2.$$

Demostración.
Dado $x \in \mathbb{R}^n$, se verifica trivialmente que $\langle x, e \rangle = \langle P_E x, e \rangle = \langle e, P_E x \rangle$, para todo $e \in E$. Luego, aplicando la Desigualdad de Cauchy-Schwartz a $\langle x, e \rangle^2$ se deduce que

$$\sup_{e \in E \setminus \{0\}} \frac{\langle e, x \rangle^2}{\|e\|^2} \leq \|P_E x\|^2.$$

La desigualdad contraria se obtiene valorando el cociente en el vector $e = P_E x$. ∎

Consideremos la familia $\mathcal{I}^{\alpha}_{[V|W]} = \{(a^{\alpha}_d, b^{\alpha}_d) : \ d \in V|W\}$ definida mediante

$$a^{\alpha}_d(Y) = \langle d, \hat{\mu}\rangle - \left[\mathtt{dim}V|W F^{\alpha}_{\mathtt{dim}V-\mathtt{dim}W,\mathtt{n}-\mathtt{dim}V}\right]^{1/2} \|d\|\hat{\sigma} \tag{3.43}$$

$$b^{\alpha}_d(Y) = \langle d, \hat{\mu}\rangle + \left[\mathtt{dim}V|W F^{\alpha}_{\mathtt{dim}V-\mathtt{dim}W,\mathtt{n}-\mathtt{dim}V}\right]^{1/2} \|d\|\hat{\sigma} \tag{3.44}$$

Teorema 3.24.
$\mathcal{I}^{\alpha}_{[V|W]}$ constituye una familia de intervalos de confianza simultáneos a nivel $1-\alpha$ para $[V|W]$.

Demostración.
Dado un valor fijo del parámetro (μ, σ^2), se verifica, en virtud del lema anterior,

$$P_{\mu,\sigma^2}\left(a^{\alpha}_d \leq \langle d, \mu\rangle \leq b^{\alpha}_d, \ \forall d \in V|W\right)$$

$$= P_{\mu,\sigma^2}\left(\frac{\langle d, \hat{\mu}-\mu\rangle^2}{(\mathtt{dim}V - \mathtt{dim}W)\hat{\sigma}^{2,\mathrm{I}}\|d\|^2} \leq F^{\alpha}_{\mathtt{dim}V-\mathtt{dim}W,\mathtt{n}-\mathtt{dim}V}, \ \forall d \in V|W\backslash\{0\}\right)$$

$$= P_{\mu,\sigma^2}\left(\sup_{d\in V|W\backslash\{0\}} \frac{\langle d, \hat{\mu}-\mu\rangle^2}{(\mathtt{dim}V - \mathtt{dim}W)\hat{\sigma}^{2,\mathrm{I}}\|d\|^2} \leq F^{\alpha}_{\mathtt{dim}V-\mathtt{dim}W,\mathtt{n}-\mathtt{dim}V}\right)$$

$$= P_{\mu,\sigma^2}\left(\frac{\left\|P_{V|W}(\hat{\mu}-\mu)\right\|^2}{(\mathtt{dim}V - \mathtt{dim}W)\hat{\sigma}^{2,\mathrm{I}}} \leq F^{\alpha}_{\mathtt{dim}V-\mathtt{dim}W,\mathtt{n}-\mathtt{dim}V}\right)$$

Teniendo en cuenta que

$$\frac{\left\|P_{V|W}(\hat{\mu}-\mu)\right\|^2}{(\mathtt{dim}V - \mathtt{dim}W)\hat{\sigma}^{2,\mathrm{I}}} \sim F_{\mathtt{dim}V-\mathtt{dim}W,\mathtt{n}-\mathtt{dim}V},$$

se concluye.

∎

Si deseamos contrastar la hipótesis inicial $H_0 : \mu \in W$, hemos de percatarnos de que H_0 es cierta si y sólo si, para cada $d \in V|W$, se satisface la hipótesis $H^d_0 : d'\mu = 0$. Como los estadístico definidos en (3.43) y (3.44) determinan un intervalo de confianza para $d'\mu$, podemos proponer el test consistente en aceptar la hipótesis inicial H_0 cuando el valor 0 quede dentro de los intervalos de confianza de la familia $\mathcal{I}^{\alpha}_{[V|W]}$. No obstante, ello equivaldría a afirmar que

$$\frac{1}{\mathtt{dim}V - \mathtt{dim}W}\frac{\left\|P_{V|W}\hat{\mu}\right\|^2}{\hat{\sigma}^{2,\mathrm{I}}} \leq F^{\alpha}_{\mathtt{dim}V-\mathtt{dim}W,\mathtt{n}-\mathtt{dim}V}.$$

Teniendo en cuenta (3.23), se deduce que el test propuesto es, precisamente, el test F. En ese sentido decimos que el test F a nivelα es consistente con la familia $\mathcal{I}^{\alpha}_{[V|W]}$ de intervalos de confianza simultáneos a nivel $1-\alpha$ para $[V|W]$, la cual se denominará en lo sucesivo, familia de Scheffé a nivel $1-\alpha$.

El problema de la familia de Scheffé es que, para que $d'\mu$ pertenezca al intervalo $(a^{\alpha}_d, b^{\alpha}_d)$, cualquiera que sea el contraste d elegido, es necesario que dichos intervalos sean más *conservadores* de lo deseado, es decir, demasiado amplios. Una solución a este problema puede ser seleccionar un subconjunto de contrastes particularmente *interesantes* y construir una familia de intervalos de confianza simultáneos para la misma. Tal es el caso de la familia de Tuckey, que estudiaremos en el capítulo 4. El método en cuestión se encuadra en el marco del Análisis de la Varianza, y consiste en seleccionar un tipo de contrastes denominados *comparaciones múltiples*. Presenta la desventaja de que deja de ser consistente con el test F y exige, teóricamente, que las diversas muestras tengan el mismo tamaño.

Existe otro método alternativo al de Tuckey, aunque válido en un contexto más general, para construir pseudo-familias de intervalos de confianza a u pseudo-nivel $1-\alpha$ para un subconjunto finito de $\mathcal{D} \subset V|W$. Aproximado a $1-\alpha$: el método de Bonferroni. Decimos pseudo-nivel $1-\alpha$ porque verifican

$$P_{\mu,\sigma^2}\left(A^{\alpha}_d \leq \langle d,\mu\rangle \leq B^{\alpha}_d(\omega),\ \forall d \in \mathcal{D}\right) \geq 1-\alpha, \quad \forall(\mu,\sigma^2) \in V \times \mathbb{R}^+. \tag{3.45}$$

Se basa en la conocida Desigualdad de Bonferroni

$$P(\cap_i A_i) \geq 1 - \sum_i P(A_i^c). \tag{3.46}$$

Teorema 3.25.

La familia siguiente verifica (3.45)
$$\begin{cases} A^{\alpha}_d = d'\hat{\mu} - \hat{\sigma}_{\mathtt{I}}\|d\| t^{\frac{\alpha}{2\mathtt{card}(\mathcal{D})}}_{\mathtt{n-dim}V} \\ B^{\alpha}_d = d'\hat{\mu} + \hat{\sigma}_{\mathtt{I}}\|d\| t^{\frac{\alpha}{2\mathtt{card}(\mathcal{D})}}_{\mathtt{n-dim}V} \end{cases}, d \in \mathcal{D}.$$

La demostración se deja como ejercicio. El principal problema del método de Bonferroni radica en si conservadurismo, a pesar de la precisión que se gana al seleccionar un subconjunto finito de $V|W$. No en vano la probabilidad de acierto es superior a $1-\alpha$.

Cuestiones propuestas

1. Demostrar el lema 3.1.

2. Demostrar la proposición 3.4.

3. Demostrar la igualdad (3.10).

4. Demostrar el teorema 3.9.

5. Teniendo en cuenta la Desigualdad de Chebyshev, probar que $kF_{k,m} \xrightarrow{d} \chi^2_k$.

6. Probar que, si se verifica la condición de Huber, la región de confianza (3.12) es asintóticamente válida para el Modelo Lineal.

7. La región de confianza (3.16) es un elipsoide. ¿Qué tiene que suceder para que sea un esfera? ¿Cómo se traduciría esa condición a un problema de Regresión Lineal?

8. Consideremos el modelo

$$Y \sim N_n(\mathtt{X}\beta, \sigma^2 \mathtt{Id}), \quad \beta \in \mathbb{R}^s,\ \sigma^2 > 0,$$

donde las columnas de $\mathtt{X}$, que se denotan mediante $\mathtt{X}_1, \ldots, \mathtt{X}_s$, constituyen un sistema ortonormal. Se desea contrastar la hipótesis inicial de que todas las componentes de β son idénticas. Probar que el estadístico de contraste del test `F` puede expresarse mediante

$$F = \frac{\mathtt{n} - s}{s-1} \cdot \frac{\sum_{j=1}^s (\mathtt{X}_j' Y)^2 - \mathtt{n}^{-1} \left(\sum_{j=1}^s \mathtt{X}_i' Y \right)^2}{Y'Y - \sum_{j=1}^s (\mathtt{X}_i' Y)^2}$$

Este estadístico se comparará con el cuantil $F^{\alpha}_{s-1,n-s}$.

9. Probar que el estadístico (3.27) sigue una distribución $t_{n-\mathtt{v}}$ cuando $d'\mu = 0$.

10. Demostrar que cualquier distribución normal presenta un coeficiente de Kurtosis $\delta = 3$.

11. Desarrollar la demostración del teorema 3.22

12. Siguiendo un procedimiento análogo al test de Barlett, obtener un test para contrastar la igualdad de las varianzas partiendo de distribuciones con Kurtosis conocido δ.

13. Demostrar el teorema 3.25.

14. **Método de mínimos cuadrados generalizado:** Dada una matriz $A \in \mathcal{M}_{\mathtt{n}\times\mathtt{n}}$ definida positiva, consideremos el modelo

$$Z \sim N_{\mathtt{n}}(\mu, \sigma^2 A), \quad \mu \in V \subset \mathbb{R}^n,\ \sigma^2 > 0.$$

Consideremos también un subespacio $W \subset V$. Encontrar entonces un estadístico suficiente y completo. Probar que el EIMV y EMV de μ es aquél que minimiza la distancia de mahalanobis

$$(Y - \hat{\mu})' A^{-1} (Y - \hat{\mu})$$

Encontrar, asimismo, el EIMV y EMV de σ^2 y un test UMP-invariante a nivel α para contrastar la hipótesis inicial $H_0 : \mu \in W$.
Indicación: Se aconseja considerar la transformación $Y = A^{-1/2}Z$, resolver los problemas anteriores en el nuevo modelo y deshacer el cambio.
Nota: Nótese que, en el caso ya estudiado, es decir, con $A = \mathtt{Id}$, el estimador de μ obtenido en la teoría es el que minimiza la distancia euclídea (3.6), por lo que se denomina solución por el método de mínimos cuadrados. En nuestro caso diremos que es una solución por el método de mínimos cuadrados generalizados. Si el modelo se parametriza a través de las coordenadas β de μ respecto de una base $\mathtt{X}$ de V, entonces nuestro problema se traduce a buscar el estimador $\hat{\beta}$ que minimice

$$\left(Y - \mathtt{X}\hat{\beta}\right)' A^{-1} \left(Y - \mathtt{X}\hat{\beta}\right)$$

Este problema será de utilidad a la hora de estudiar el método de Mínimos Cuadrados Ponderados en Regresión.

15. En las condiciones anteriores, probar que la solución $\hat{\beta}$ mínimo-cuadrática generalizada es la solución a la ecuación lineal

$$\mathtt{X}' A^{-1} \mathtt{X} \hat{\beta} = \mathtt{X}' A^{-1} Y \tag{3.47}$$

16. Dada una matriz $\mathtt{X} \in \mathcal{M}_{n\times k}$, consideremos el modelo $Y \sim N_n(\mathtt{X}\beta, \sigma^2)$, donde $\beta \in \mathbb{R}^k$ y $\sigma^2 > 0$. Determinar el test $\mathtt{F}$ a nivel α para contrastar la hipótesis inicial de que las dos primeras componentes de β son idénticas.

17. ¿Tiene validez asintótica la familia de intervalos de confianza simultáneos de Scheffé cuando se prescinde del supuesto de normalidad?

Capítulo 4

Regresión Lineal Múltiple

En el presente capítulo abordamos problemas como los que aparecen en los ejemplos 1 y 2 del capítulo 1. Es decir, consideramos una variable dependiente, y, que pretende ser explicada a partir de q variables explicativas, $z[1], \ldots, z[q]$, mediante una ecuación lineal. El hecho de que las variables explicativas sean, efectivamente, variables aleatorias o, por contra, predeterminadas de antemano, es lo que caracteriza a los Modelos de Correlación y Regresión, respectivamente. En este capítulo se considerarán fijos los valores correspondientes a $z[1], \ldots, z[q]$, respectivamente, que se denominarán vectores explicativos. El modelo de Correlación se estudiará en el siguiente capítulo. No obstante, adelantamos aquí, tal y como se comenta en la Introducción, que todos los problemas de Inferencia Estadísticos que se plantean en el Modelo de Regresión se resuelven de idéntica forma (salvo ciertos matices teóricos) desde el Modelo de Correlación.

La Regresión Lineal Múltiple se caracteriza porque admite varios vectores explicativas. Como caso particular, cuando es sólo uno, se denomina Regresión Lineal Simple. Si se consideran varias variables dependientes estaremos hablando de una Regresión Lineal Multivariante. Este último estudio no se trata aquí[1], aunque no añade dificultades considerables, al menos en lo que a Estimación se refiere.

En este capítulo pueden distinguirse claramente dos partes: la primera (secciones 1 y 2) dedicada al estudio del modelo sin considerar los posibles problemas prácticos que conlleva, bien sea por la violación de los supuestos del modelo, bien por las dificultades a la hora de extraer conclusiones. Por lo tanto, se trata en buena parte de la aplicación directa de los resultados obtenidos en el capítulo anterior. La segunda parte trata el diagnóstico y posibles soluciones a dichos problemas. En todo caso, el estudio puede complementarse con la bibliografía que se referencia a lo largo del capítulo. Dicho esto,

[1] Ver el volumen dedicado al Análisis Multivariante

empezaremos definiendo de forma precisa el Modelo y fijando la notación a seguir. Advertimos que muchos conceptos que aquí se definen se manejan e interpretan en el Apéndice. Asumimos la redundancia en aras de una mejor compresión de los mismos.

Consideraremos una vector aleatorio $Y = (Y_1, \ldots, Y_{\mathtt{n}})'$ que se expresa mediante

$$\begin{array}{ccccccccccc} Y_1 & = & \beta_0 & + & \beta_1 \mathtt{z}_1[1] & + & \ldots & + & \beta_q \mathtt{z}_1[q] & + & \varepsilon_1 \\ \vdots & & \vdots & & \vdots & & & & \vdots & & \vdots \\ Y_{\mathtt{n}} & = & \beta_0 & + & \beta_1 \mathtt{z}_{\mathtt{n}}[1] & + & \ldots & + & \beta_q \mathtt{z}_{\mathtt{n}}[q] & + & \varepsilon_{\mathtt{n}} \end{array}$$

donde $\beta = (\beta_0, \beta_1, \ldots, \beta_q)'$ puede ser, en principio, cualquier vector de $\mathbb{R}^{q+1}$ y ε_i, $i = 1 \ldots, \mathtt{n}$, son variables iid con distribución $N(0, \sigma^2)$, pudiendo ser σ^2 cualquier número positivo. Si se denota $\mathcal{E} = (\varepsilon_1, \ldots \varepsilon_{\mathtt{n}})'$ y

$$\mathtt{X} = \begin{pmatrix} 1 & \mathtt{z}_1[1] & \ldots & \mathtt{z}_1[q] \\ \vdots & \vdots & & \vdots \\ 1 & \mathtt{z}_{\mathtt{n}}[1] & \ldots & \mathtt{z}_{\mathtt{n}}[q] \end{pmatrix}$$

el modelo equivale a considerar un vector aleatorio Y tal que

$$Y = \mathtt{X}\beta + \mathcal{E}, \quad \mathcal{E} \sim N_{\mathtt{n}}(0, \sigma^2 \mathtt{Id}), \quad \beta \in \mathbb{R}^{q+1}, \ \sigma^2 > 0.$$

Se trata pues de un caso particular del Modelo Lineal Normal. Se supondrá por hipótesis que $\mathtt{rg}(\mathtt{X}) = q + 1$.

La primera columna de la matriz $\mathtt{X}$ se denota por 1_n, y la submatriz restante por $\mathtt{Z}$. Siguiendo la notación introducida en el Apéndice, se denotan por $\mathtt{z}[j]$, $j = 1, \ldots, q$ los vectores columnas de $\mathtt{Z}$, que se denominarán vectores explicativos. Lo estadísticos $\overline{y}$, $\overline{Y}$, Y_0, $\overline{\mathtt{z}}[j]$,$\overline{\mathtt{z}}$, $\overline{\mathtt{Z}}$ y $\mathtt{Z}_0$ se definen también como en el Apéndice. Se denotan por $\mathtt{x}_i$ y $\mathtt{z}_i$, $i = 1, \ldots, \mathtt{n}$, los vectores filas traspuestos de $\mathtt{X}$ y $\mathtt{Z}$, respectivamente. Podemos hablar de la matriz de varianzas-covarianzas total muestral

$$S = \frac{1}{\mathtt{n}} \begin{pmatrix} s_y^2 & S_{y\mathtt{z}} \\ S_{\mathtt{z}y} & S_{\mathtt{zz}} \end{pmatrix} = \frac{1}{\mathtt{n}} \begin{pmatrix} Y_0'Y_0 & Y_0'\mathtt{Z}_0 \\ \mathtt{Z}_0'Y_0 & \mathtt{Z}_0'\mathtt{Z}_0 \end{pmatrix}. \tag{4.1}$$

La varianza parcial muestral, definida en (9.64), queda como sigue:

$$s^2_{y \cdot \mathtt{z}} = s_y^2 - S_{y\mathtt{z}} S_{\mathtt{zz}}^{-1} S_{\mathtt{z}y}. \tag{4.2}$$

Por último, se denota por $\underline{\beta}$ el vector de $(\beta_1, \ldots, \beta_q)'$, de manera que $\beta = \begin{pmatrix} \beta_0 \\ \underline{\beta} \end{pmatrix}$.

4.1. Estimaciones e intervalos de confianza.

Dado que el estudio de Regresión Lineal puede formalizarse mediante un Modelo Lineal Normal con $V = \langle \mathtt{X} \rangle$, los problema de Estimación y Contraste de Hipótesis han quedado resueltos, desde un punto de vista teórico, en el capítulo anterior. Únicamente hemos de aplicar los resultados allí obtenidos.

Estimación de β y σ^2.

Primeramente, en lo que se refiere al problema de Estimación, contamos con dos parámetros: $\beta \in \mathbb{R}^{q+1}$ [2] y $\sigma^2 > 0$. En virtud del teorema 3.9, el EIMV y EMV de β es

$$\hat{\beta} = (\mathtt{X}'\mathtt{X})^{-1}\mathtt{X}'Y. \tag{4.3}$$

Estamos pues hablando del único vector de $\mathbb{R}^{q+1}$ tal que

$$X\hat{\beta} = P_{\langle \mathtt{X} \rangle}Y.$$

Precisamente, $\mathtt{X}\hat{\beta}$ es el estimador de la media de Y, que en el capítulo anterior denotábamos por $\hat{\mu}$. No obstante, en este contexto y con el fin de coincidir en la notación con la mayor parte de la bibliografía recomendada, se denotará

$$\hat{Y} = \mathtt{X}\hat{\beta},$$

y sus componentes se denotarán por $\hat{Y}_1, \ldots, \hat{Y}_{\mathtt{n}}$, denominándose en lo sucesivo valores ajustados. La componentes de $\hat{\beta}$ se denotarán mediante $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_q$. El vector compuestos por todas ellas salvo $\hat{\beta}_0$ se denota por $\underline{\hat{\beta}}$. Siguiendo abreviadamente la notación introducida en el Apéndice, se tiene que

$$\mathtt{e} = Y - \hat{Y}.$$

Las componentes del vector anterior, que se denotarán por $\mathtt{e}_i$, $i = 1, \ldots, \mathtt{n}$, se denominan residuos de regresión. Se verifica entonces que

$$\|\mathtt{e}\|^2 = \|Y - \mathtt{X}\hat{\beta}\|^2 = \text{mín}\{\|Y - \mathtt{X}b\|^2 : \ b \in \mathbb{R}^{q+1}\}. \tag{4.4}$$

El EIMV de σ^2 es

$$\hat{\sigma}^{2,\mathrm{I}} = \frac{\|\mathtt{e}\|^2}{\mathtt{n} - (q+1)} \tag{4.5}$$

$$= \frac{1}{\mathtt{n} - (q+1)}\|Y - \mathtt{X}\hat{\beta}\|^2 \tag{4.6}$$

$$= \frac{1}{\mathtt{n} - (q+1)} \sum_{i=1}^{\mathtt{n}} \left[Y_i - \left(\hat{\beta}_0 + \mathtt{z}_i'\underline{\hat{\beta}} \right) \right]^2. \tag{4.7}$$

[2] Para mayor comodidad, suprimimos la notación $\beta_{\mathtt{X}}$ utilizada en el capítulo anterior.

Según (9.65), se tiene que $s^2_{y.\mathtt{Z}} = \mathtt{n}^{-1}\|\mathtt{e}\|^2$. De hecho, se trata del EMV de σ^2. Del teorema 3.9 se sigue también que

$$\hat{\beta} \sim N_{q+1}\left(\beta, \sigma^2(\mathtt{X}'\mathtt{X})^{-1}\right).$$

Luego, en particular,

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2\psi_{jj}), \quad j = 0, 1, \ldots, q, \tag{4.8}$$

donde ψ_{jj} denota el j-ésimo elemento de la diagonal de $(\mathtt{X}'\mathtt{X})^{-1}$. En (4.65) se da una expresión explícita de estos valores que dependerá, entre otros factores, del coeficiente de correlación múltiple de $\mathtt{z}[j]$ respecto al resto de vectores explicativos. Sabemos que el elipsoide (3.16) constituye una región de confianza a nivel $1-\alpha$ para β. Por su parte,

$$[\mathtt{n} - (q+1)]\hat{\sigma}^{2,\mathtt{I}} \sim \sigma^2\chi^2_{\mathtt{n}-(q+1)}, \tag{4.9}$$

siendo independiente del estimador de β, lo cual permite construir el siguiente intervalo de confianza a nivel $1-\alpha$ para β_i, $i = 0, 1, \ldots, q$.

$$\hat{\beta}_j \pm t^{\alpha/2}_{\mathtt{n}-(q+1)}\hat{\sigma}_{\mathtt{I}}\sqrt{\psi_{jj}} \tag{4.10}$$

En (3.13) podemos encontrar un intervalo de confianza para σ^2. Veamos cuál es el comportamiento asintótico de los estimadores a medida que vamos introduciendo más unidades experimentales en el estudio, es decir, a medida que se añaden nuevas filas a la matriz $(Y\,\mathtt{X})$ (en ese caso, tendremos un Modelo Asintótico del tipo (3.31)). Del teorema 3.14 se sigue que, si $m(\mathtt{X}'\mathtt{X}) \to \infty$ cuando el tamaño de muestra $\mathtt{n}$ tiende a infinito, el estimador de beta considerado es consistente. Pero la condición anterior se verifica trivialmente en nuestro caso, pues el primer elemento de la diagonal de $\mathtt{X}'\mathtt{X}$ coincide precisamente con el tamaño de muestra. Por otra parte, del teorema 3.15 se deduce la consistencia del estimador de σ^2. A continuación, intentaremos expresar los estimadores de β y σ^2 a partir de las medias muestrales y matrices de covarianzas, lo cual facilitará enormemente el estudio de los coeficientes de correlación. Realmente, hemos de advertir lo que viene a continuación no es sino un caso particular de lo estudiado en el Apéndice.

Primeramente, hay que tener en cuenta que $\hat{\beta}_0$ y $\underline{\hat{\beta}}$ son los únicos elemento de $\mathbb{R}$ y $\mathbb{R}^q$, respectivamente, tales que

$$P_{\langle \mathtt{X}\rangle}Y = \hat{\beta}_0 1_n + \mathtt{Z}\underline{\hat{\beta}}.$$

Dado que $\mathtt{Z}_0 = P_{\langle 1_n \rangle}\mathtt{Z}$, se tiene que $\langle \mathtt{X} \rangle = \langle 1_n \rangle \oplus \langle \mathtt{Z}_0 \rangle$, siendo dicha descomposición ortogonal. Por lo tanto, $P_{\langle \mathtt{X} \rangle}Y$ puede calcularse como sigue[3]

$$\begin{aligned} P_{\langle \mathtt{X} \rangle}Y &= P_{\langle 1_n \rangle}Y + P_{\langle \mathtt{Z}_0 \rangle}Y \\ &= P_{\langle 1_n \rangle}Y + P_{\langle \mathtt{Z}_0 \rangle}Y_0 \\ &= \overline{y}1_n + \mathtt{Z}_0(\mathtt{Z}_0'\mathtt{Z}_0)^{-1}\mathtt{Z}_0'Y_0 \\ &= \overline{y}1_n + (\mathtt{Z} - \overline{\mathtt{Z}})S_{\mathtt{ZZ}}^{-1}S_{\mathtt{Z}y} \\ &= \left(\overline{y} - \overline{\mathtt{z}}'S_{\mathtt{ZZ}}^{-1}S_{\mathtt{Z}y}\right)1_n + \mathtt{Z}S_{\mathtt{ZZ}}^{-1}S_{\mathtt{Z}y}. \end{aligned}$$

En consecuencia,

$$\underline{\hat{\beta}} = S_{\mathtt{ZZ}}^{-1}S_{\mathtt{Z}y}, \qquad \hat{\beta}_0 = \overline{y} - \overline{\mathtt{z}}'\underline{\hat{\beta}}. \tag{4.11}$$

Dado que $\underline{\hat{\beta}} = (\mathtt{Z}_0'\mathtt{Z}_0)^{-1}\mathtt{Z}_0'Y$, se sigue de lo anterior y de de (9.11) que $\underline{\hat{\beta}}$ e $\overline{y}$ son independientes y que

$$\underline{\hat{\beta}} \sim N_q\left(\underline{\beta}, \frac{\sigma^2}{n}S_{\mathtt{ZZ}}^{-1}\right), \tag{4.12}$$

lo cual será de gran utilidad cuando construyamos los intervalos de confianza para las predicciones. Un caso particular por su sencillez es el la Regresión Simple, donde tenemos

$$\underline{\hat{\beta}} = \frac{s_{\mathtt{Z}y}^2}{s_{\mathtt{Z}}^2}, \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{\mathtt{z}}.$$

Otro enfoque del problema

Lo que vemos a continuación es el planteamiento y solución del problema mediante la aplicación directa del criterio de mínimos cuadrados, sin hacer uso del concepto de proyección ortogonal. Obviamente, obtendremos una solución idéntica.

Dados las observaciones de la variable respuesta, $Y_1, \ldots, Y_{\mathtt{n}}$ y de los valores explicativos, $\mathtt{z}_1[1], \ldots, \mathtt{z}_{\mathtt{n}}[q]$, se trata de buscar los valores de $\beta_0, \beta_1, \ldots, \beta_q$ que minimizan la suma de cuadrados siguientes:

$$\sum_{i=1}^{\mathtt{n}} \left[Y_i - (\beta_0 + \beta_1\mathtt{z}_i[1] + \ldots \beta_q\mathtt{z}_i[q])\right]^2$$

El mínimo se busca haciendo uso de herramientas del Cálculo Diferencial. Concretamente, se buscan los valores donde las derivadas parciales respecto a los parámetros

[3]Realmente, la expresión de $P_{\langle \mathtt{X} \rangle}Y$ se obtuvo ya en (9.61). En consecuencia, el razonamiento que sigue puede omitirse.

se anulan. Es decir, se plantea el siguiente sistema de ecuaciones lineales:

$$0 = \sum_i \left[Y_i - (\beta_0 + \beta_1 \mathbf{z}_i[1] + \ldots + \beta_q \mathbf{z}_i[q])\right] \quad (4.13)$$

$$0 = \sum_i \left[Y_i - (\beta_0 + \beta_1 \mathbf{z}_i[1] + \ldots + \beta_q \mathbf{z}_i[q])\right] \mathbf{z}_i[j], \quad j = 1, \ldots, q \quad (4.14)$$

En consecuencia, para que el mínimo se alcance en $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_q)'$ es condición necesaria que

$$\mathbf{X}'(Y - \mathbf{X}\hat{\beta}) = 0^4$$

Es decir, buscamos una solución al sistema de ecuaciones lineales

$$\mathbf{X}'Y = \mathbf{X}'\mathbf{X}\hat{\beta} = 0$$

Si la matriz $\mathbf{X}$ es de rango completo, como suponemos en nuestro caso, la única solución es, precisamente, (4.3) [5]. El Hessiano es $2\mathbf{X}'\mathbf{X} > 0$, luego la única solución es, efectivamente, un mínimo.

Coeficiente de correlación múltiple

Recordemos que $\hat{\sigma}^{2,\mathtt{MV}} = s^2_{y\cdot\mathbf{Z}}$, y que el segundo término descompone de esta forma

$$s^2_{y\cdot\mathbf{Z}} = s^2_y - S_{y\mathbf{Z}} S^{-1}_{\mathbf{ZZ}} S_{\mathbf{Z}y}. \quad (4.15)$$

El segundo sumando del término de la derecha es la matriz de covarianzas total muestral de $P_{\langle \mathbf{Z}_0 \rangle} Y_0$, que equivale a la matriz de covarianzas total de $\hat{Y} = P_{\langle \mathbf{X} \rangle} Y$. En consecuencia, tenemos la siguiente descomposición de la varianza muestral de Y:

$$\begin{aligned} s^2_y &= s^2_{P_{\langle \mathbf{Z}_0 \rangle} Y} + s^2_{y\cdot\mathbf{Z}} \\ &= s^2_{\hat{Y}} + s^2_{y\cdot\mathbf{Z}}. \end{aligned}$$

Esta descomposición de s^2_y se corresponde con la siguiente descomposición ortogonal de $\langle 1_n \rangle^\perp$

$$\langle 1_n \rangle^\perp = \langle \mathbf{Z}_0 \rangle \oplus \langle \mathbf{X} \rangle^\perp = \langle \mathbf{X} \rangle | \langle 1_n \rangle \oplus \langle \mathbf{X} \rangle^\perp.$$

Así pues, tal y como se comenta en el Apéndice, $s^2_{y\cdot\mathbf{Z}}$ se interpreta como la parte de la variabilidad total de Y no explicada por la variabilidad total de $\mathbf{z}[1], \ldots, \mathbf{z}[q]$ mediante

[4] Nótese que se se impone la condición de ortogonalidad entre $Y - \mathbf{X}\hat{\beta}$ y $\mathbf{X}$, luego, estamos hablando de la protección ortogona de Y sobre $\langle \mathbf{X} \rangle$.

[5] Téngase en cuenta que la proyección ortogonal en $\mathbb{R}^{\mathbf{n}}$ de Y sobre $\langle \mathbf{x} \rangle$ minimiza la distancias euclídea del vector $\|Y - \mathbf{X}\hat{\beta}\|^2$.

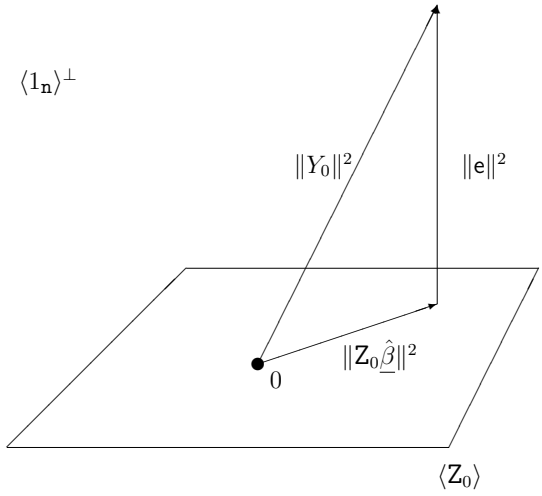
la regresión lineal, mientras que $s^2_{\hat{Y}}$ se interpretará como la parte de la variabilidad total de Y que sí es explicada por la regresión lineal respecto a $\mathtt{z}[1],\ldots,\mathtt{z}[q]$. Ello invita a definir el coeficiente de correlación múltiple muestral[6]

$$
\begin{aligned}
R^2_{y,\mathbf{z}} &= \frac{S_{y\mathbf{z}}S^{-1}_{\mathbf{zz}}S_{\mathbf{z}y}}{s_y^{-2}} && (4.16)\\
&= s^2_{\hat{Y}}/s^2_y && (4.17)\\
&= \frac{\|P_{\langle \mathbf{x}\rangle|\langle 1_n\rangle}Y\|^2}{\|P_{\langle 1_n\rangle^\perp}Y\|^2} && (4.18)\\
&= \frac{\|P_{\langle \mathbf{z}_0\rangle}Y_0\|^2}{\|Y_0\|^2} && (4.19)\\
&= \frac{\|\mathbf{Z}_0\hat{\underline{\beta}}\|^2}{\|Y_0\|^2} && (4.20)
\end{aligned}
$$

Puede interpretarse como la proporción de variabilidad total de Y explicada por la variabilidad total de $\mathtt{z}[1],\ldots,\mathtt{z}[q]$ mediante una regresión lineal. Esta interpretación en términos del lenguaje usual es, posiblemente, una extrapolación de lo que sucede, en términos probabilísticos, en el modelo normal multivariante, donde la varianza parcial es la varianza de la distribución condicional y por lo tanto, la parte de la varianza no explicada (linealmente en este caso) por el vector aleatorio que condiciona. En todo caso debe ser matizada para evitar confusiones.

Estamos descomponiendo la variablidad total de Y en un vector que es combinación lineal de las variabilidades totales de $\mathtt{z}[1],\ldots,\mathtt{z}[q]$ más otro, denominado residuo. Dado cualquier vector $\mathtt{u}\in\langle \mathbf{Z}_0\rangle$, podemos considerar la descompoción $Y=\mathtt{u}+(Y-\mathtt{u})$, pero no estamos dispuestos a admitir cualquier descomposición del vector Y, sino que buscamos el vector de $\langle Z_0\rangle$ más próximo según la distancia euclídea. En se sentido decimos que ese vector de $\langle Z_0\rangle$ es el que mejor explica la variabilidad de Y_0 y es el que conduce a una descomposición ortogonal con el residuo como diferencia, según se ve en la ilustración siguiente. Así pues, cuando hablamos de la parte de variabilidad de Y_0 explicada por la variabilidad total de $\mathtt{z}[1],\ldots,\mathtt{z}[q]$ nos estamos refiriendo implícitamente a dicho vector.

[6] Realmente, el parámetro que definimos a continuación se denomina coeficiente de determinación. El coeficiente de correlación múltiple es la raíz cuadrada del mismo.



Los comentarios anteriores pueden resultar banales pero, en lo relativo a la explicación, digamos coloquial, del coeficiente de correlación, una interpretación al pie de la letra en razonamientos de tipo heurístico puede conducir a errores conceptuales. Por ejemplo, ¿cómo es posible que dos variables incorreladas no lo sean condicionalmente dada una tercera? Si se pretende argumentar en términos de variabilidades explicadas difícilmente se logrará un razonamiento convincente: si la variabilidad de una no explica en absoluto la de la otra, ¿cómo es posible que una parte de la primera (residuo) explique otra parte de la segunda? Nuevamente, hemos de remitirnos a la consabida descomposición ortogonal para entender este hecho: es posible que los vectores originales sean ortogonales pero que sus residuos dada la tercera no lo sean.

En definitiva, de la ecuación (4.15) se deduce

$$s_{yy\cdot\mathbf{z}} = s_y^2(1 - R_{y,\mathbf{z}}^2). \tag{4.21}$$

El término de la izquierda es el estimador de máxima verosimilitud de σ^2. Veamos, no obstante, otra interesante caracterización de $R_{y,\mathbf{z}}^2$.

Proposición 4.1.

$R_{y,\mathbf{z}}^2 = r^2_{y,\mathbf{z}\underline{\hat{\beta}}} = \max\{r^2_{y,\mathbf{z}b} : b \in \mathbb{R}^q\}$.

Demostración.

Dado que los coeficientes de correlación simple y múltiple son invariantes ante traslaciones, podemos suponer, sin pérdida de generalidad, que $\overline{y} = 0$ y $\overline{\mathbf{z}} = 0$ o, lo que

es lo mismo, que $Y = Y_0$ y $\mathsf{Z} = \mathsf{Z}_0$. En ese caso y teniendo en cuenta (4.11), se sigue que

$$\begin{aligned} r^2_{y,\mathbf{z}\underline{\hat{\beta}}} &= \frac{s^2_{y,\mathbf{z}\underline{\hat{\beta}}}}{s^2_y \cdot \underline{\hat{\beta}}' S_{\mathbf{zz}} \underline{\hat{\beta}}} = \frac{\left(S_{y,\mathbf{z}} \cdot \underline{\hat{\beta}}\right)^2}{s^2_y \cdot \underline{\hat{\beta}}' S_{\mathbf{zz}} \underline{\hat{\beta}}} \\ &= \frac{\left(S_{y\mathbf{z}} S^{-1}_{\mathbf{zz}} S_{\mathbf{z}y}\right)^2}{s^2_y \left(S_{y\mathbf{z}} S^{-1}_{\mathbf{zz}} S_{\mathbf{z}y}\right)} = R^2_{y,\mathbf{z}}. \end{aligned}$$

Por otra parte, se sigue de (4.4), que

$$\|Y - \mathsf{Z}\underline{\hat{\beta}}\|^2 \leq \|Y - \lambda \mathsf{Z} b\|, \quad \forall b \in \mathbb{R}^q, \ \forall \lambda \in \mathbb{R}.$$

Operando en ambas expresiones y despejando el término $\|Y\|^2$, se tiene que

$$\|\mathsf{Z}\underline{\hat{\beta}}\|^2 - 2\langle Y, \mathsf{Z}\underline{\hat{\beta}}\rangle \leq \lambda^2 \|\mathsf{Z}b\|^2 - 2\lambda \langle Y, \mathsf{Z}b\rangle.$$

Por lo tanto,

$$\frac{2\langle Y, \mathsf{Z}\underline{\hat{\beta}}\rangle - \|\mathsf{Z}\underline{\hat{\beta}}\|^2}{\|Y\| \cdot \|\mathsf{Z}\underline{\hat{\beta}}\|} \geq \frac{2\lambda\langle Y, \mathsf{Z}b\rangle - \lambda^2\|\mathsf{Z}b\|^2}{\|Y\| \cdot \|\mathsf{Z}\underline{\hat{\beta}}\|}$$

Considerando entonces $\lambda = \|\mathsf{Z}\underline{\hat{\beta}}\|/\|\mathsf{Z}b\|$, se tiene que

$$r_{y,\mathbf{z}\underline{\hat{\beta}}} = \frac{\langle Y, \mathsf{Z}\underline{\hat{\beta}}\rangle}{\|Y\| \cdot \|\mathsf{Z}\underline{\hat{\beta}}\|} \geq \frac{\langle Y, \mathsf{Z}b\rangle}{\|Y\| \cdot \|\mathsf{Z}b\|} = r_{y,\mathbf{z}b},$$

con lo cual termina la demostración. ∎

Por tanto y como cabía esperar, la máxima correlación lineal entre Y y una combinación lineal de los vectores $\mathbf{z}[1], \ldots, \mathbf{z}[q]$, se alcanza precisamente con la ecuación de regresión, y su cuadrado es el coeficiente de correlación múltiple. Esta idea se puede generalizar al caso multivariante para construir los coeficientes de correlación canónica. Podemos garantizar un resultado completamente análogo para el coeficiente de correlación múltiple probabilístico (ejercicio 2.12).

Una propiedad del coeficiente de correlación múltiple que, desde cierto punto de vista, puede considerarse una patología, es el hecho de que al añadir al modelo un nuevo vector explicativo $\mathbf{z}[q+1]$, por inapropiado que éste sea, no se producirá disminución alguna del coeficiente de correlación múltiple. Es más, puede demostrarse (cuestión propuesta) que R^2 permanece invariante si y sólo si el coeficiente de correlación parcial entre Y y $\mathbf{z}[q+1]$ dados $\mathbf{z}[1], \ldots, \mathbf{z}[q]$ es nulo. Ello puede movernos a

definir otro coeficiente similar a R^2 pero que no presente esta propiedad. Nótese que (4.18) puede expresarse también así

$$R^2_{y,\mathbf{Z}} = 1 - \frac{\|P_{\langle \mathbf{X}\rangle^\perp} Y\|^2}{\|P_{\langle 1_n\rangle^\perp} Y\|^2}.$$

Teniendo en cuenta que

$$s_y^2 = \frac{1}{\mathtt{n}}\|P_{\langle 1_{\mathbf{n}}\rangle^\perp} Y\|^2, \qquad \hat{\sigma}^{2,\mathrm{I}} = \frac{1}{\mathtt{n}-(q+1)}\|P_{\langle \mathbf{X}\rangle^\perp} Y\|^2,$$

puede resultar natural definir el siguiente estadístico, denominado coeficiente de correlación múltiple corregido:

$$\overline{R}^2_{y,\mathbf{Z}} = 1 - \frac{\hat{\sigma}^{2,\mathrm{I}}}{s_y^2}.$$

La relación entre $R^2_{y,\mathbf{Z}}$ y $\overline{R}^2_{y,\mathbf{Z}}$ es la siguiente:

$$\overline{R}^2_{y,\mathbf{Z}} = 1 - \frac{\mathtt{n}}{\mathtt{n}-(q+1)}(1 - R^2_{y,\mathbf{Z}}). \quad [7] \tag{4.22}$$

Predicciones

Un estudio de Regresión Lineal Múltiple equivale a la búsqueda de una ecuación lineal que relacione la variable respuesta con las explicativas, lo cual se realiza normalmente con uno de los siguiente objetivos: conocer en qué medida influye en la respuesta cada uno de los vectores explicativos o predecir valores de la variable respuesta cuando se conocen los de los vectores explicativos. En este momento nos centramos en el segundo objetivo. Así pues, supongamos que tenemos una nueva unidad experimental, independiente de la muestra que se ha utilizado en la estimación de los parámetros β y σ^2, y que dicha unidad experimental aporta unos valores $\mathbf{z}_0 = (\mathbf{z}_0[1], \ldots, \mathbf{z}_0[q])'$ en los vectores explicativos. Se trata de predecir el valor Y_0 que presentará en la variable respuesta, suponiendo que se mantengan el patrón que rige nuestro modelo, es decir, que

$$Y_0 = \beta_0 + \beta_1 \mathbf{z}_0[1] + \ldots \beta_q \mathbf{z}_0[q] + \varepsilon_0, \quad \varepsilon_0 \sim N(0, \sigma^2).$$

En ese caso, tanto Y_0 como $\beta_0 + \mathbf{z}_0'\underline{\beta}$, que es el valor medio que cabe esperar para Y_0, pueden estimarse mediante

$$\hat{Y}_0 = \hat{\beta}_0 + \mathbf{z}_0'\hat{\underline{\beta}} = \overline{y} + (\mathbf{z}_0 - \overline{\mathbf{z}})'\hat{\underline{\beta}},$$

[7] El término $\mathtt{n}$ del numerador se sustituye por $\mathtt{n}-1$ si optamos por considerar el estimador insesgado $s_y^2 = (\mathtt{n}-1)^{-1}\|P_{\langle 1_n\rangle^\perp} Y\|^2$.

que, teniendo en cuenta (4.12), sigue un modelo de distribución

$$\hat{Y}_0 \sim N\left(\beta_0 + \mathbf{z}_0'\underline{\beta}\ ,\ \frac{\sigma^2}{\mathtt{n}}[1 + (\mathbf{z}_0 - \overline{\mathbf{z}})' S_{\mathbf{ZZ}}^{-1}(\mathbf{z}_0 - \overline{\mathbf{z}})]\right).$$

En lo sucesivo, se denotará

$$d^2(\mathbf{z}_0, \overline{\mathbf{z}}) = (\mathbf{z}_0 - \overline{\mathbf{z}})' S_{\mathbf{ZZ}}^{-1}(\mathbf{z}_0 - \overline{\mathbf{z}}) \tag{4.23}$$

a la distancia de Mahalanobis entre $\mathbf{z}_0$ y $\overline{\mathbf{z}}$. De la expresión anterior se obtiene el siguiente intervalo de confianza a nivel $1-\alpha$ para $\mathtt{E}[Y_0]$

$$\hat{Y}_0 \pm t^{\alpha}_{\mathtt{n}-(q+1)} \hat{\sigma}_{\mathtt{I}} \sqrt{\frac{1}{\mathtt{n}} + \frac{1}{\mathtt{n}} d^2(\mathbf{z}_0, \overline{\mathbf{z}})}. \tag{4.24}$$

Por otra parte, dado que Y_0 e $\hat{Y}_0$ son independientes, se verifica que $Y_0 - \hat{Y}_0$ sigue una distribución normal de media 0 y varianza $\sigma^2[1 + \mathtt{n}^{-1} + \mathtt{n}^{-1} d^2(\mathbf{z}_0, \overline{\mathbf{z}})]$. En consecuencia y teniendo en cuenta (4.21), podemos construir un intervalo de confianza a nivel $1-\alpha$ para el valor de Y_0 mediante

$$\hat{Y}_0 \pm t^{\alpha}_{\mathtt{n}-(q+1)} \mathtt{n}^{-1/2} [\mathtt{n} - (q+1)]^{1/2} \sqrt{s_y^2 (1 - R^2_{y,\mathbf{Z}}) \left[1 + \frac{1}{\mathtt{n}} + \frac{1}{\mathtt{n}} d^2(\mathbf{z}_0, \overline{\mathbf{z}})\right]}. \tag{4.25}$$

Si nos centramos en el término que queda dentro de la raíz cuadrada, podemos analizar los factores de los que depende la fiabilidad de la predicción $\hat{Y}_0$:

- Primeramente, de la varianza total de Y, s_y^2, de manera que cuanto mayor sea menos fiable resultará la predicción.
- De $R^2_{y,\mathbf{Z}}$, es decir, de la proporción de varianza explicada por la regresión, de manera que cuanto mayor sea más fiable resultará la predicción, lógicamente.
- De el tamaño de la muestra $\mathtt{n}$, de forma que cuanto mayor sea más fiable resultará la predicción.
- De la distancia de Mahalanois del punto $\mathbf{z}_0$ donde se realiza la predicción al *centroide* de la muestra. Curiosamente, cuanto más lejos esté $\mathbf{z}_0$ menos fiable resultará la predicción. Esto ha de servir para concienciarnos de que el problema de Regresión es de carácter local, es decir, que no deben exrtrapolarse los resultados lejos de la región de $\mathbb{R}^q$ donde se ha realizado el estudio.

4.2. Principales constrastes. Selección de variables.

Abordamos a continuación el problema de Contraste de Hipótesis. Podemos distinguir, en principio, contrastes relativos al parámetro μ y contrastes relativos a σ^2, aunque estos últimos, que se resuelven en la sección 2.3, gozan de menos interés que los primeros por razones que ya se detallaron el capítulo anterior. Así pues, nos centraremos en los contrastes de hipótesis referentes a β, que ya quedaron resueltos, desde un punto de vista teórico, en las secciones 3.2 y 3.4. Sabemos, concretamente, que para contrastar mediante el test `F` una hipótesis inicial del tipo

$$H_0 : A\beta = 0,$$

siendo A una matriz de dimensiones $m \times (q+1)$ y rango m, debemos comparar $F^{\alpha}_{m,\mathtt{n}-(q+1)}$ con el estadístico de contraste (3.26), que reproducimos a continuación:

$$F = \frac{1}{m}\, \frac{(A\hat{\beta})' \left[A(\mathtt{X}'\mathtt{X})^{-1}A'\right]^{-1} A\hat{\beta}}{\hat{\sigma}^{2,\mathtt{I}}}. \tag{4.26}$$

Vamos a destacar tres tipos de contrastes por su utilidad:

1. **Contraste de una ecuación:** en este apartado consideramos, en principio, el contraste de la hipótesis inicial $H_0 : \beta = 0$, que se corresponde con $A = \mathtt{Id}_{q+1}$. Por lo tanto, de (4.26) podemos obtener una expresión bastante explícita del estadístico de contraste
$$F = \frac{1}{q+1}\frac{\|\mathtt{X}\hat{\beta}\|^2}{\hat{\sigma}^{2,\mathtt{I}}}, \tag{4.27}$$
que ha de compararse con $F^{\alpha}_{q+1,\mathtt{n}-(q+1)}$. Este contraste no es muy útil en sí, pero sirve de instrumento a la hora de contrastar una hipótesis del tipo $H_0 : \beta = b$, para algún vector $b \in \mathbb{R}^{q+1}$ conocido. Es decir, cuando queremos contrastar si cierta ecuación de regresión predeterminada es aceptable teniendo en cuenta nuestros datos. En ese caso, debemos sustituir el vector Y por $Y^* = Y - \mathtt{X}b$ y contrastar la hipótesis $\beta = 0$ con los datos transformados (trasladados).

2. **Contraste total:** consideramos a continuación el contraste de la hipótesis inicial $H_0 : \underline{\beta} = 0$, que se corresponde con $A = (0_q|\mathtt{Id}_q)$. La veracidad de la misma equivale a la incapacidad de explicación de Y por parte de $\mathtt{z}[1], \ldots, \mathtt{z}[q]$. Dado que $\langle \mathtt{X} \rangle = \langle 1_{\mathtt{n}}|\mathtt{Z}_0 \rangle$, se sigue de (4.26) (ejercicio propuesto) que el estadístico de contraste puede expresarse mediante
$$F = \frac{\mathtt{n}-q}{q}\frac{\|(\mathtt{Z}-\overline{\mathtt{Z}})\underline{\hat{\beta}}\|^2}{\hat{\sigma}^{2,\mathtt{I}}}, \tag{4.28}$$

o bien en términos más generales mediante

$$F = \frac{1}{q}\frac{\|P_{\langle \mathbf{X}\rangle|\langle 1_{\mathtt{n}}\rangle}Y\|^2}{\|P_{\langle \mathbf{X}\rangle^{\perp}}Y\|^2}, \tag{4.29}$$

que ha de compararse con $F^{\alpha}_{q,\mathtt{n}-(q+1)}$. No obstante, si hacemos uso de (4.20), el estadístico de contraste puede expresarse también a través del coeficiente de correlación múltiple $R^2_{y\mathbf{Z}}$ mediante

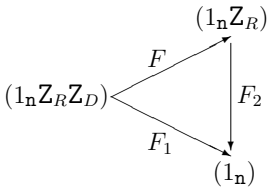
$$F = \frac{\mathtt{n}-(q+1)}{q}\frac{R^2_{y,\mathbf{Z}}}{1-R^2_{y,\mathbf{Z}}}. \tag{4.30}$$

La expresión anterior puede interpretarse fácilmente en términos intuitivos teniendo en cuenta el significado del coeficiente de correlación múltiple y que el test `F` aceptará la hipótesis inicial cuando éste sea próximo a cero.

3. **Contrastes parciales:** supongamos que la matriz $\mathbf{Z}$ se divide por columnas en dos submatrices, $\mathbf{Z}_R$ (con r columnas) y $\mathbf{Z}_D$ (con d columnas), y que el vector $\underline{\beta}$ se divide de manera análoga en dos subvectores β_R y β_D, compuestos respectivamente por los coeficientes de los vectores explicativos que conforman las submatrices $\mathbf{Z}_R$ y $\mathbf{Z}_D$. Nos interesamos ahora en el contraste de una hipótesis inicial del tipo $H_0 : \beta_D = 0$. La veracidad de la misma supone la nulidad de los vectores que componen $\mathbf{Z}_D$ para explicar la variabilidad de Y, lo cual induciría a eliminarlos y pasar de un modelo *completo* a otro *reducido*, en el cual sólo se tendría en cuenta la submatriz $\mathbf{Z}_R$. De (3.21) se sigue que el estadístico para contrastar dicha hipótesis es

$$F = \frac{\mathtt{n}-(q+1)}{d}\frac{\|P_{\langle \mathbf{X}\rangle|\langle 1_n \mathbf{Z}_R\rangle}Y\|^2}{\|P_{\langle \mathbf{X}\rangle^{\perp}}Y\|^2},$$

que se compara con $F^{\alpha}_{d,\mathtt{n}-(q+1)}$. Curiosamente, el estadístico del contraste parcial puede expresarse a través de los estadísticos de contraste total en los modelos completo y reducido. Efectivamente, denótense los mismos por F_1 y F_2, respectivamente, y considérese el siguiente diagrama:



Cada flecha del diagrama se interpreta como la reducción del modelo a la que conduciría la hipótesis inicial cuyo contraste se resuelve mediante el estadístico adjunto. Puede comprobarse (se deja como ejercicio), que

$$F = \frac{\mathtt{n}-(q+1)}{d}\left(\frac{1+\frac{q}{\mathtt{n}-(q+1)}F_1}{1+\frac{r}{\mathtt{n}-(r+1)}F_2}-1\right). \tag{4.31}$$

El contraste de hipótesis del tipo $H_0 : \beta_j = 0, j = 1, \ldots, q$ [8] es, desde el punto de vista práctico, el caso más interesante de contraste parcial. De (4.26) podemos obtener una expresión explícita del estadístico de contraste

$$F = \frac{\hat{\beta}_j^2}{\hat{\sigma}^{2,\mathrm{I}}\psi_{jj}}, \tag{4.32}$$

que se compara con $F^{\alpha}_{1,\mathtt{n}-(q+1)}$. Ello equivale a comparar con $t^{\alpha}_{\mathtt{n}-(q+1)}$ el estadístico

$$t = \frac{|\hat{\beta}_j|}{\hat{\sigma}_{\mathrm{I}}\sqrt{\psi_{jj}}}. \tag{4.33}$$

Para contrastar un hipótesis inicial del tipo $H_0 : \beta_j = b_j$, basta aplicar una traslación a los datos para obtener el estadístico de contraste

$$t = \frac{|\hat{\beta}_j - b_j|}{\hat{\sigma}_{\mathrm{I}}\sqrt{\psi_{jj}}}, \tag{4.34}$$

que se compararía con el mismo cuantil. Curiosamente, éste el test que se derivaría directamente de (4.8) y (4.9). Mediante un razonamiento análogo al realizado en (4.30) [9], podemos expresar (4.32) a través del coeficiente de correlación parcial entre Y y $\mathtt{z}[j]$ dados los demás vectores explicativos (que configuran una matriz $\mathtt{Z}_R$) mediante

$$F = [\mathtt{n}-(q+1)]\frac{r^2_{y,\mathtt{z}_{[j]}\bullet\mathtt{z}_R}}{1-r^2_{y,\mathtt{z}_{[j]}\bullet\mathtt{z}_R}}. \tag{4.35}$$

Esta expresión resulta muy intuitiva, pues significa que aceptamos la hipótesis inicial $H_0 : \beta_j = 0$ cuando $r_{y,\mathtt{z}_{[j]}\cdot\mathtt{z}_R}$ es próximo a cero, es decir, cuando, conocidos los valores correspondientes al resto de vectores explicativos, la variabilidad de $\mathtt{z}[j]$ aporta muy poco a la hora de explicar la variabilidad de Y.

[8] También puede incluirse β_0, pues el vector 1_n es a estos efectos un vector cualquiera, como pueden serlo $\mathtt{z}[1], \ldots, \mathtt{z}[q]$.

[9] Cuestión propuesta

Este tipo de contraste es de gran utilidad teniendo en cuenta que, antes de indagar acerca de la ecuación concreta que rige aproximadamente el comportamiento de la variable respuesta, conviene optimizar el modelo, desechando aquellas variables (vectores) explicativas que no tienen influencia significativa en la variable respuesta. La forma natural de realizar esta depuración sería, a simple vista, realizar los q contrastes parciales, uno para cada coeficiente, y eliminar las variables explicativas que no aporten resultados significativos. El problema de este método es que el hecho de eliminar o introducir una variable explicativa influye en los contrastes parciales de las otras. Así, por ejemplo, puede suceder que al eliminar una resulte significativa otra que no lo era en el modelo completo. Todo ello es debido a la *colinealidad* más o menos severa que suele afectar a las variables (vectores) explicativas. Este concepto se tratará más a fondo en una sección posterior. Por ello, se hace necesario el uso de algún algoritmo de selección de variables basado en los contrastes parciales, aunque más complejo. Comentaremos brevemente en qué consisten los métodos *forward*, *backward* y *stepwise*, junto con otros métodos no basados en los contrastes parciales. Un estudio ás detallado puede encontrarse en Rawlings et al. (1999).

El método forward o hacia delante consiste en considerar q modelos de regresión simple con Y como variable respuesta y cada uno de los vectores explicativos como único vector explicativo. Entrará en el modelo definitivo aquella cuyo contraste parcial, que equivale al total, sea más significativo. A continuación, se considerarán $q-1$ modelos de regresión añadiendo a la variable introducida cualquiera de las otras, y se realiza, en cada modelo, el contraste parcial para la variable candidata, entrando en el modelo definitivo aquélla que aporte un resultado más significativo[10]. El procedimiento continúa y se van añadiendo variables haste que ninguna de las candidatas aporte un resultado significativo en el contraste parcial. El límite de significación se conviene de antemano.

El método backward o hacia atrás parte del modelo completo, donde se realizan los q contrastes parciales para desechar la variable explicativa que presente un resultado menos significativo; a continuación se considera en el modelo reducido resultante los $q-1$ contrastes parciales y se desecha la variable menos significativa, y así sucesivamente hasta que todas las que quedan aportan un resultado significativo en el contraste parcial. El método stepwise o por pasos sucesivos es una combinación de los métodos forward y backward, pues cada vez que se introduce una nueva variable

[10]Coincide con aquella que aporte un resultado más significativo en el contrate total (cuestión propuesta).

por el método forward, depura mediante el método backward el modelo resultante.

Existen otros métodos no basados en os contrastes parciales consistente en buscar, para cada $q' \leq q$, el mejor modelo con q' vectores explicativos y escoger entonces un q' lo más pequeño posible siempre y cuando la *perdida* que conlleva la reducción sea también lo menor posible. La cuestión es dilucidar cómo se mide dicha pérdida y eso es en esencia lo que distingue unos métodos de otros. Así podemos analizar cuánto disminuye R^2 con el modelo reducido o considerar en su lugar el coeficiente de correlación múltiple ajustado. Podemos también considerar el aumento de $\|P_{\langle 1_{\mathtt{n}} \mathbf{Z}_R\rangle^\perp} Y\|^2$ respecto a $\|P_{\langle 1_{\mathtt{n}} \mathbf{Z}\rangle^\perp} Y\|^2$ para un modelo reducido $(1_{\mathtt{n}} \mathbf{Z}_R)$ con q' vectores explicativos. El método de Mallow, relacionado con el anterior, consisten en considerar el estadístico

$$C_{q'} = \frac{\|P_{\langle 1_{\mathtt{n}} \mathbf{Z}_R\rangle^\perp} Y\|^2}{\hat{\sigma}^{2,\mathtt{I}}} + 2q' - \mathtt{n}.$$

Si las variable excluida en el modelo reducido no son relevantes cabe esperar que $[\mathtt{n} - (q'+1)]^{-1} \|P_{\langle 1_{\mathtt{n}} \mathbf{Z}_R\rangle^\perp} Y\|^2$ tome un valor próximo a σ^2, con lo que $C_{q'}$ tomará un valor próximo a q'. De no ser así, $C_{q'}$ debería estar claramente por encima de q'. Así pues, para cada valor de q' se consideran todos los posible modelos reducidos y se escoge el que aporte un valor $C_{q'}$ menor. Entonces se escoge el menor q' tal que $C_{q'}$ sea lo suficiente próximo a q'.

4.3. Análisis de los supuestos del Modelo

Todas las inferencias realizadas hasta el momento se han efectuado suponiendo que se verifiquen los supuestos del modelo, que pueden desglosarse de la siguiente forma:

1. Independencia: Y_i, $i = 1, \ldots, \mathtt{n}$ son independientes.
2. Normalidad: Y_i sigue un modelo de distribución normal para $i = 1, \ldots, \mathtt{n}$.
3. Homocedasticidad: existe $\sigma^2 > 0$ tal que $\mathtt{var}[Y_i] = \sigma^2$, para todo $i = 1, \ldots, \mathtt{n}$.
4. Linealidad: existe $\beta \in \mathbb{R}^{q+1}$ tal que $\mathtt{E}[Y_i] = \mathbf{x}_i' \beta$, para todo $i = 1, \ldots, \mathtt{n}$.

Aunque, como veremos más adelante, existen técnicas para evaluar el cumplimiento del supuesto de independencia, diseñar un test de hipótesis para contrastarlo resulta especialmente dificultoso, dado que los tests suelen construirse partiendo precisamente de $\mathtt{n}$ unidades experimentales observaciones independientes. No obstante, el cumplimiento de este supuesto depende fundamentalmente de cómo se ha diseñado

la recogida de muestras, de manera que el investigador suele saber si sus unidades experimentales pueden considerarse (aproximadamente) independientes. En caso contrario, deberíamos optar por técnicas de análisis completamente diferentes a las que nos ocupa, como pueden ser el de series de tiempo o medidas repetidas.

La situación ideal se da cuando las denominadas variables o vectores explicativos son variables aleatorias, propiamente dicho, y la matriz (YZ) resultante puede considerarse una muestra aleatoria simple de tamaño $\mathtt{n}$ de una distribución $(q+1)$-normal multivariante. Ése es exactamente el modelo de Correlación Lineal[11] y, en ese caso, condicionando sobre el valor concreto de $\mathtt{Z}$ obtenido de la matriz aleatoria Z, se obtiene un modelo de Regresión Lineal con los cuatro supuestos anteriores. Por lo tanto, lo primero que deberíamos hacer es contrastar a normalidad multivariante de nuestro datos mediante un test de normalidad multivariante[12]. Si el resultado es significativo, puede entenderse como necesario, desde cierto punto de vista, un contraste de los supuestos de normalidad, homocedasticidad y linealidad, aunque esa visión es, como veremos, bastante discutible.

Respecto al supuesto de normalidad, hemos de advertir previamente que, si prescindimos del mismo, tendremos un Modelo Lineal cuyo comportamiento, tanto en el sentido exacto como asintótico, ha sido estudiado en el capítulo anterior. Así, desde el punto de vista exacto, podemos afirmar que el estimador propuesto para σ^2 es insesgado, mientras que el de β es lineal insesgado de mínima varianza. Desde el punto de vista asintótico, es decir, a medida que introducimos nuevas unidades experimentales (o sea, a medida que incorporamos a la matriz $(Y\mathtt{X})$ nuevas filas), sabemos que ambos estimadores son consistentes, puesto que la condición (3.32) se verifica trivialmente. Además, si se verifica la condición (3.35) de Huber, todas las inferencias realizas en las secciones anteriores son asintóticamente válidas para muestras suficientemente grandes. Lo que debemos hacer ahora, lógicamente, es estudiar en que se traduce exactamente la condición de Huber o, lo que es lo mismo, cuánto vale $m(\mathtt{X}(\mathtt{X}'\mathtt{X})^{-1}\mathtt{X}')$. Sabemos que

$$m(\mathtt{X}(\mathtt{X}'\mathtt{X})^{-1}\mathtt{X}') = \text{máx}\{\nu_{ii} : i = 1, \ldots, \mathtt{n}\}$$

donde ν_{ij}, $i, j = 1, \ldots, \mathtt{n}$, denotan las componentes de la matriz $P_{\langle \mathtt{X} \rangle} = \mathtt{X}(\mathtt{X}'\mathtt{X})^{-1}\mathtt{X}'$. Dado que $\langle 1_{\mathtt{n}} \rangle \subset \langle \mathtt{X} \rangle$ y aplicando la propiedad (3.29) con $A = P_{\langle \mathtt{X} \rangle}$, se sigue que

$$\frac{1}{\mathtt{n}} \leq \nu_{ii} \leq 1, \quad i = 1, \ldots, \mathtt{n}. \tag{4.36}$$

[11] Ver Arnold (1981).

[12] En Bilodeau & Brenner (1999) podemos encontrar una prueba de normalidad multivariante basado en el hecho de que las distancias de mahalanobis divididas por el tamaño muestral deben seguir una distribución Beta en el caso normal.

Nótese que, al ser $P_{\langle \mathtt{X} \rangle}^2 = P_{\langle \mathtt{X} \rangle}$, se tiene que

$$\nu_{ii}(1-\nu_{ii}) = \sum_{j \neq i} \nu_{ij}^2 \tag{4.37}$$

Para calcular explícitamente las componentes de $P_{\langle \mathtt{X} \rangle}$ consideraremos la descomposición ortogonal del subespacio vectorial $\langle \mathtt{X} \rangle = \langle 1_{\mathtt{n}} \rangle \oplus \langle \mathtt{Z}_0 \rangle$, de manera que

$$P_{\langle \mathtt{X} \rangle} = \frac{1}{\mathtt{n}} \left(\begin{array}{c|c} 1 & (\mathtt{z}_1 - \overline{\mathtt{z}})' \\ \vdots & \vdots \\ 1 & (\mathtt{z}_{\mathtt{n}} - \overline{\mathtt{z}})' \end{array} \right) \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & S_{\mathtt{zz}}^{-1} \end{array} \right) \left(\begin{array}{ccc} 1 & \ldots & 1 \\ \hline \mathtt{z}_1 - \overline{\mathtt{z}} & \ldots & \mathtt{z}_{\mathtt{n}} - \overline{\mathtt{z}} \end{array} \right)$$

En consecuencia, se verifica

$$\nu_{ii} = \frac{1}{\mathtt{n}} + \frac{d^2(\mathtt{z}_i, \overline{\mathtt{z}})}{\mathtt{n}}, \quad i = 1, \ldots, \mathtt{n}, \tag{4.38}$$

donde d^2 es la distancia de Mahalanobis definida en (4.23). Los elementos fuera de la diagonal pueden expresarse mediante

$$\nu_{ij} = \frac{1}{n} \left(1 + (\mathtt{z}_i - \overline{\mathtt{z}})' S_{\mathtt{zz}}^{-1} (\mathtt{z}_j - \overline{\mathtt{z}}) \right), \quad i \neq j. \tag{4.39}$$

De esta forma, la condición de Huber equivale a

$$\mathtt{n}^{-1} \max_{1 \leq i \leq \mathtt{n}} d^2(\mathtt{z}_i, \overline{\mathtt{z}}) \longrightarrow 0. \tag{4.40}$$

Esta condición, relacionada con la presencia de valores explicativos extremos, se interpreta de la siguiente forma: a medida que introducimos más datos, las distancias de mahalanobis de los vectores explicativos a su centroide puede ir aumentando, pero a ritmo menor que $\mathtt{n}$. Esto puede conseguirse de manera artificial si las variables explicativos están controladas en el diseño, es decir, si no son realmente variables aleatorias. Tal es nuestro caso. Cuando sean variables aleatorias, lo cual corresponde al modelo de correlación, que se estudiará en el próximo capítulo, la condición (4.40) se obtendrá de una forma bastante natural. Ello permite obviar el supuesto de normalidad para $\mathtt{n}$ suficientemente grande.

No obstante, aunque la violación del supuesto de normalidad no es en sí un problema grave, es preferible que no se produzca dada la vinculación existente entre los supuestos de normalidad y linealidad. Efectivamente, es muy frecuente que el incumplimiento del primero vaya acompañada de la violación del segundo, e incluso del supuesto de homocedasticidad. Si tenemos la intención de contrastar la normalidad, la homocedasticidad o la linealidad, hemos de tener en cuenta que el vector aleatorio

Y no es una muestra aleatoria simple de ninguna distribución, a menos que $\underline{\beta}$ sea nulo. De ahí que para poder efectuar el contraste sea necesario un modelo de regresión lineal muy particular, consistente en controlar el valor del vector explicativo y considerar para cada valor de éste una muestra aleatoria simple de valores de Y que presente ese valor concreto en los vectores explicativos. Obviamente, un diseño de este tipo sólo es viable en la práctica en un estudio de regresión simple, como sucede en el ejemplo 2 de la Introducción. El diseño al que nos referimos se denomina completamente aleatorizado y será estudiado en profundidad en el capítulo 6. Puede expresarse como sigue

$$\begin{array}{|ccccc|}
\hline
Y_{11} & = & \theta_1 & + & \varepsilon_{11} \\
\vdots & & \vdots & & \vdots \\
Y_{1\mathtt{n}_1} & = & \theta_1 & + & \varepsilon_{11\mathtt{n}_1} \\
\hline
\vdots & & \vdots & & \vdots \\
\vdots & & \vdots & & \vdots \\
\hline
Y_{k1} & = & \theta_1 & + & \varepsilon_{k1} \\
\vdots & & \vdots & & \vdots \\
Y_{k\mathtt{n}_k} & = & \theta_1 & + & \varepsilon_{k\mathtt{n}_k} \\
\hline
\end{array} \tag{4.41}$$

donde ε_{ij}, $i = 1, \ldots, k$ y $j = 1, \ldots \mathtt{n}_i$, son independientes con media 0 y varianza σ_i^2. En lo que sigue se denotará $\mathtt{n} = \sum_i \mathtt{n}_i$. En ese caso, se puede contrastar, para cada $i = 1, \ldots, k$, si Y_{ij}, $j = 1, \ldots, \mathtt{n}_i$, es una muestra aleatoria simple de una distribución normal. Para ello podemos hacer uso de diversos tests, como el de Kolmogorov-Smirnov-Lilliefords, el de Shappiro-Wilks, el test χ^2 o el de D'Agostino. No obstante, hemos de advertir que, para que estos tests tengan suficiente potencia en todos los casos es necesario que las muestras sean todas grandes, cosa poco factible en la práctica. De lo contrario, estaremos otorgando una enorme ventaja a la hipótesis inicial de normalidad. Si, a pesar de los inconvenientes comentados, estamos dispuestos a contrastar los supuestos, el procedimiento a seguir sería el siguiente: escoger un test de normalidad (el de D'Agostino es el más aconsejabe para muestras pequeñas) y aplicarlo a las k muestras. Si todos los resultados son no significativos, aceptaremos la hipótesis inicial de normalidad. Por lo tanto, podremos suponer que, en el modelo anterior, los términos ε_{ij} son todos normales.

A continuación procederíamos a contrastar la hipótesis inicial de igualdad de varianzas. Para ello contamos con el test de Barlett, estudiado en el capítulo anterior. Hay que advertir que éste test es bastante sensible ante la violación del supuesto de normalidad, de ahí que se precise un resultado no significativo en la fase anterior. No obstante, puede utilizarse un test más robusto como el de Levene. Si el resultado

es no significativo, podremos suponer que el modelo propuesto anteriormente es un modelo lineal normal $Y = \mu + \mathcal{E}$, donde $\mathcal{E} \sim N_{\mathtt{n}}(0, \sigma^2 \mathtt{Id})$, para algún $\sigma^2 > 0$, y $\mu \in V$, siendo V el subespacio de $\mathbb{R}^{\mathtt{n}}$ generado por los vectores

$$\mathtt{v}_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \hline \vdots \\ \vdots \\ \hline 0 \\ \vdots \\ 0 \end{pmatrix} \quad \ldots \quad \mathtt{v}_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \hline \vdots \\ \vdots \\ \hline 1 \\ \vdots \\ 1 \end{pmatrix}.$$

En este modelo, el supuesto de linealidad se corresponde con la hipótesis $\mu \in W$, donde

$$W = \left\langle 1_{\mathtt{n}} \middle| \sum_i \mathtt{z}_i \cdot \mathtt{v}_i \right\rangle \subset V.$$

Así pues, el modelo de regresión lineal puede considerarse un modelo reducido ($\mu \in W$) del modelo completo ($\mu \in V$). Por lo tanto, la linealidad se contrasta mediante el correspondiente test $\mathtt{F}$ a nivel α. Puede comprobarse que, en esta ocasión, el estadístico de contraste del mismo es el siguiente

$$F = \frac{(k-2)^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[\overline{y}_{i\cdot} - \left(\hat{\beta}_0 + \hat{\beta}_1 \mathtt{z}_i \right) \right]^2}{(\mathtt{n}-k)^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{y}_{i\cdot})^2}, \tag{4.42}$$

donde $\overline{y}_{i\cdot}$ denota, para cada $i = 1, \ldots, k$, la media aritmética o muestral del grupo i-ésimo. Este estadístico se comparará con $F^{\alpha}_{k-2,\mathtt{n}-k}$. El término del denominador, que el EIMV de σ^2 en el modelo completo, se denomina error puro de regresión.

Así pues, hemos visto un procedimiento para contrastar sucesivamente los supuestos de normalidad, homocedasticidad y linealidad del modelo de regresión. A este método se le pueden presentar diversas objeciones. En primer lugar, requiere de un diseño que sólo es factible en el caso de una regresión simple; en segundo lugar, para aplicar el test de linealidad es necesario suponer homocedasticidad y para el de homocedasticidad es necesario suponer la normalidad de cada uno de los k grupos, por lo cual, en el momento que aparezca un resultado significativo el modelo debería ser desechado. Que esto no suceda en muchas ocasiones suele deberse normalmente al hecho de que el número de datos por grupo no es lo suficientemente alto como para

que los tests utilizados tengan una potencia aceptable, privilegiándose enormemente las hipótesis iniciales de normalidad y homocedasticidad e, incluso, de linealidad. Por ello, el rigor que pretendíamos ganar contrastando los supuestos del modelo mediante sendos tests de hipótesis no es tal al no reunirse los requisitos mínimos para su aplicación.

4.4. Análisis de los residuos

El problema es, como cabía esperar, bastante delicado. Desde luego hemos de ser consciente que los supuestos de este modelo, como los de cualquier otro, son ideales, es decir, que hemos de asumir que, en la práctica, no se verificarán jamás. Por ello una alternativa al procedimiento anterior es renunciar al contraste de los supuestos en pro de una evaluación gráfica del desajuste existente entre el modelo teórico y los datos empíricos. De esta forma, si el desajuste se considera admisible se aplican los métodos estudiados. En caso contrario, se buscan transformaciones de las variables que permitan una mejor adecuación al modelo o bien se aplican procedimientos alternativos. Este análisis, que presenta por una importante componente de subjetividad, depende en buena medida del comportamiento asintótico del modelo y de la robustez de los métodos estudiados. En todo caso, hemos de tener en cuenta que los cuatro supuestos del modelo pueden expresarse en función de los errores $\varepsilon_i = Y_i - \mathtt{E}[Y_i]$, $i = 1, \ldots, \mathtt{n}$. Más concretamente, podemos definir, para cada vector β en $\mathbb{R}^{q+1}$, las variables aleatorias

$$\varepsilon_i^\beta = Y_i - \mathbf{x}_i'\beta, \quad i = 1, \ldots, \mathtt{n},$$

de manera que los supuestos, si β es el verdadero valor del parámetro, pueden expresarse así:

1. Independencia: ε_i^β, $i = 1, \ldots, \mathtt{n}$, son independientes.
2. Normalidad: ε_i^β sigue un modelo de distribución normal para $i = 1, \ldots, \mathtt{n}$.
3. Homocedasticidad: existe $\sigma^2 > 0$ tal que $\mathtt{var}[\varepsilon_i^\beta] = \sigma^2$, para todo $i = 1, \ldots, \mathtt{n}$.
4. Linealidad: $\mathtt{E}[\varepsilon_i^\beta] = 0$, para todo $i = 1, \ldots, \mathtt{n}$.

Es decir, que el cumplimiento de los cuatro supuestos equivales al hecho de que las observaciones ε_i^β, $i = 1, \ldots, \mathtt{n}$, constituyan una muestral aleatoria simple de una distribución normal de media 0. Dado que que estos valores son desconocidos por

serlo β, podemos estimarlos de manera natural mediante los denominados residuos[13]:

$$\mathtt{e}_i = Y_i - \hat{Y}_i = Y_i - \mathtt{x}_i'\hat{\beta}, \quad i = 1, \ldots, \mathtt{n}. \tag{4.43}$$

Como ya sabemos, estos residuos componen un vector $\mathtt{e} = (\mathtt{e}_1, \ldots, \mathtt{e}_\mathtt{n})'$ que verifica

$$\mathtt{e} = Y - P_{\langle \mathtt{X} \rangle} Y = P_{\langle \mathtt{X} \rangle^\perp} Y$$

cuya media aritmética es nula y cuya varianza es, por definición, la varianza parcial (4.2). Lo que hemos hecho es descomponer ortogonalmente el vector Y mediante

$$Y = \mathtt{X}\hat{\beta} + \mathtt{e}, \tag{4.44}$$

de manera que

$$\|Y\|^2 = \|\mathtt{X}\hat{\beta}\|^2 + \|\mathtt{e}\|^2.$$

La distribución del vector de residuos es, en virtud de la proposición 2.1, la siguiente:

$$\mathtt{e} \sim N_\mathtt{n}\big(0, \sigma^2[\mathtt{Id} - P_{\langle \mathtt{X} \rangle}]\big),$$

es decir, que

$$\mathtt{e}_i \sim N\big(0, \sigma^2[1 - \nu_{ii}]\big), \quad i = 1, \ldots, \mathtt{n},$$

verificándose además que $\mathtt{cov}[e_i, e_j] = -\nu_{ij}$ si i es distinto de j. Por lo tanto, los residuos no son incorrelados ni, por lo tanto, independientes. De hecho, puede probarse, teniendo en cuenta que $\mathtt{rg}(P_{\langle \mathtt{X} \rangle^\perp}) = \mathtt{n} - (q+1)$, que el vector aleatorio $\mathtt{e}$ está incluido con probabilidad 1 en un subespacio lineal de dimensión $\mathtt{n} - (q+1)$. De (4.38) se sigue que, para cada $i = 1, \ldots, \mathtt{n}$,

$$\mathtt{var}[\mathtt{e}_i] = \sigma^2(1 - \nu_{ii}) = \sigma^2 \frac{\mathtt{n} - 1}{\mathtt{n}} - \sigma^2 \frac{d^2(\mathtt{z}_i, \overline{\mathtt{z}})}{\mathtt{n}}. \tag{4.45}$$

Podemos observar que los residuos tampoco son homocedásticos, sino que su varianza depende de la distancia de mahalanobis del vector explicativo $\mathtt{z}_i$ correspondiente al centroide, de manera que cuanto mayor sea ésta menor será la varianza del residuo. El valor máximo se daría cuando $\mathtt{z}_i$ coincidiera con el centroide. Por contra, si ν_{ii} fuera igual a 1, la varianza del residuo sería nula, es decir, el valor de y pronosticado para $\mathtt{z}_i$ coincidirá con probabilidad 1 con el valor observado. Esta situación puede darse teóricamente. Teniendo en cuenta (4.37), equivale a que todos los ν_{ij}, para j distinto de i, sean nulos. Concretamente, en un análisis de regresión simple, puede probarse,

[13] Definidos ya en (9.55).

teniendo en cuenta (4.39), que ello equivale a que todos los vectores explicativos salvo $\mathtt{z}_i$ sean idénticos. En lo sucesivo supondremos que ese caso extremo no se verifica.

Por otra parte, la varianza de los residuos es menor que la varianza del modelo, lo cual era de esperar, teniendo en cuenta la descomposición ortogonal (4.44). No obstante, a medida que el número de unidades experimentales tiende a infinito, la primera converge a la segunda si, y sólo si, se verifica la condición de Huber.

Los residuos definidos anteriormente suelen denominarse residuos *brutos*, en contraposición con los residuos estandarizados que definimos a continuación. La nueva definición viene motivada por el hecho de que

$$\frac{\mathtt{e}_i}{\sigma\sqrt{1-\nu_{ii}}} \sim N(0,1), \quad i = 1, \ldots, \mathtt{n}.$$

Esto podría servirnos para plantear un test de bondad de ajuste al modelo de regresión, aun teniendo en cuenta que no se verifica la independencia. Dado que σ es desconocida, lo que se suele hacer en estos caso es sustituirla por un estimador insesgado de la misma. De esa forma, se definen los residuos estandarizados mediante

$$\mathtt{r}_i = \frac{\mathtt{e}_i}{\hat{\sigma}_{\mathrm{I}}\sqrt{1-\nu_{ii}}}, \quad i = 1, \ldots, \mathtt{n}.$$

En condiciones similares estos estadísticos seguirían una distribución $t_{\mathtt{n}-(q+1)}$. En esta ocasión eso no es correcto debido a que $\mathtt{e}_i$ no es independiente de $\hat{\sigma}_{\mathrm{I}}$. De hecho, recordemos que

$$\hat{\sigma}^{2,\mathrm{I}} = \frac{1}{\mathtt{n}-(q+1)} \sum_i \mathtt{e}_i^2.$$

Por lo tanto, si queremos obtener una distribución t-student nos vemos obligados a introducir unas sutiles variaciones.

En lo sucesivo y para cada $i = 1, \ldots, \mathtt{n}$, se denotarán mediante $\hat{\beta}(i)$ y $\hat{\sigma}^{2,\mathrm{I}}(i)$ los estimadores de β y σ^2, respectivamente, que se obtienen eliminado del modelo la i-ésima unidad experimental (es decir, la i-ésima fila de datos). Se define entonces

$$\hat{Y}(i) = \mathtt{X}\hat{\beta}(i).$$

Así mismo, $Y(i)$ y $\mathtt{X}(i)$ denotarán el vector aleatorio Y desprovisto de su componentes i-ésima y la matriz $\mathtt{X}$ desprovista de la fila i-ésima, respectivamente. Por último, en el modelo desprovisto de la unidad i-ésima se define el vector de residuos brutos mediante

$$\mathtt{e}(i) = Y(i) - \hat{Y}(i).$$

En esas condiciones, se define los residuos estudentizados mediante

$$\mathtt{t}_i = \frac{\mathtt{e}_i}{\hat{\sigma}_{\mathrm{I}}(i)\sqrt{1-\nu_{ii}}}, \quad i = 1, \ldots, \mathtt{n}.$$

Podríamos proponer también eliminar la influencia de la unidad i-ésima en el cálculo del residuo correspondiente, obteniéndose de esta forma los residuos estudentizados eliminados, que se definen mediante

$$\tilde{\mathtt{t}}_i = \frac{Y_i - \hat{Y}_i(i)}{\hat{\sigma}_{\mathtt{I}}(i)\sqrt{1-\nu_{ii}}}, \quad i = 1, \ldots, \mathtt{n}.$$

A continuación probaremos que, para cada $i = 1, \ldots, \mathtt{n}$, $\mathtt{t}_i$ sigue una distribución t-Student. De ahí su nombre.

Lema 4.2.
Con las notaciones anteriores se verifica que, para cada $i = 1, \ldots, \mathtt{n}$,

$$\hat{\beta} = \hat{\beta}(i) + \frac{\mathtt{e}_i}{1-\nu_{ii}}(\mathtt{X}'\mathtt{X})^{-1}\mathtt{x}_i. \tag{4.46}$$

Demostración.
Tener en cuenta, en primer lugar, que

$$\mathtt{X}'\mathtt{X} = \mathtt{X}(i)'\mathtt{X}(i) + \mathtt{x}_i\mathtt{x}_i', \quad \mathtt{X}'Y = \mathtt{X}(i)'Y(i) + \mathtt{x}_iY_i$$

y que $\mathtt{x}_i'(\mathtt{X}'\mathtt{X})^{-1}\mathtt{x}_i$ es igual a ν_{ii}, que es menor que 1. Por lo tanto, se sigue del lema 9.8 que

$$[\mathtt{X}(i)'\mathtt{X}(i)]^{-1} = [\mathtt{X}'\mathtt{X} - \mathtt{x}_i\mathtt{x}_i']^{-1} = (\mathtt{X}'\mathtt{X})^{-1} + \frac{(\mathtt{X}'\mathtt{X})^{-1}\mathtt{x}_i\mathtt{x}_i'(\mathtt{X}'\mathtt{X})^{-1}}{1-\nu_{ii}}.$$

En consecuencia,

$$\begin{aligned}
\hat{\beta}(i) &= [\mathtt{X}(i)'\mathtt{X}(i)]^{-1}\mathtt{X}(i)'Y(i) \\
&= \left[(\mathtt{X}'\mathtt{X})^{-1} + \frac{(\mathtt{X}'\mathtt{X})^{-1}\mathtt{x}_i\mathtt{x}_i'(\mathtt{X}'\mathtt{X})^{-1}}{1-\nu_{ii}}\right][\mathtt{X}'Y - \mathtt{x}_iY_i] \\
&= \hat{\beta} - (\mathtt{X}'\mathtt{X})^{-1}\mathtt{x}_iY_i + (1-\nu_{ii})^{-1}(\mathtt{X}'\mathtt{X})^{-1}\mathtt{x}_i\hat{Y}_i - \nu_{ii}(1-\nu_{ii})^{-1}(\mathtt{X}'\mathtt{X})^{-1}\mathtt{x}_iY_i \\
&= \hat{\beta} - (1-\nu_{ii})^{-1}(\mathtt{X}'\mathtt{X})^{-1}\mathtt{x}_iY_i + (1-\nu_{ii})^{-1}(\mathtt{X}'\mathtt{X})^{-1}\mathtt{x}_i\hat{Y}_i \\
&= \hat{\beta} - (\mathtt{X}'\mathtt{X})^{-1}\mathtt{x}_i\frac{Y_i - \hat{Y}_i}{1-\nu_{ii}},
\end{aligned}$$

de lo cual se obtiene la tesis. ∎

Teorema 4.3.
Para cada $i = 1, \ldots, \mathtt{n}$, se verifica lo siguiente

(i) $\tilde{\mathtt{t}}_i = (1-\nu_{ii})\mathtt{t}_i$.

(ii) $\mathtt{t}_i \sim t_{\mathtt{n}-(q+2)}$.

(iii) $[\mathtt{n}-(q+2)]\hat{\sigma}^{2,\mathrm{I}}(i) = [\mathtt{n}-(q+1)]\hat{\sigma}^{2,\mathrm{I}} - \frac{\mathtt{e}_i^2}{1-\nu_{ii}}$.

Demostración.
Si en la expresión (4.46) multiplicamos por $\mathtt{x}_i'$ por la izquierda obtenemos

$$\hat{Y}_i = \hat{Y}_i(i) + \frac{\nu_{ii}}{1-\nu_{ii}}\left(Y_i - \hat{Y}_i\right). \tag{4.47}$$

Por lo tanto,

$$\hat{Y}_i = \nu_{ii}Y_i + (1-\nu_{ii})\hat{Y}_i(i).$$

Luego,

$$\mathtt{e}_i = (1-\nu_{ii})\left(Y_i - \hat{Y}_i(i)\right), \tag{4.48}$$

De lo cual se sigue (i). Además, en virtud del teorema 3.9-(iii), se tiene que $\hat{\sigma}^{2,\mathrm{I}}(i)$ y $\mathtt{e}_i$ son independientes. Teniendo en cuenta que

$$\frac{\mathtt{e}_i}{\sigma\sqrt{1-\nu_{ii}}} \sim N(0,1), \qquad [\mathtt{n}-(q+2)]\sigma^{-2}\hat{\sigma}^{2,\mathrm{I}}(i) \sim \chi^2_{\mathtt{n}-(q+2)},$$

se obtiene la tesis (ii). Para probar (iii) multiplicamos en (4.46) por $\mathtt{x}_j'$, para $j \neq i$, obteniendo

$$\hat{Y}_j = \hat{Y}_j(i) + \frac{\nu_{ij}}{1-\nu_{ii}}\mathtt{e}_i.$$

En consecuencia,

$$\mathtt{e}_j(i) = \mathtt{e}_j + \frac{\nu_{ij}}{1-\nu_{ii}}\mathtt{e}_i.$$

Sumando los cuadrados cuando $j \neq i$ se obtiene

$$\sum_{j\neq i}\mathtt{e}_j(i)^2 = \sum_{j\neq i}\mathtt{e}_j^2 + \frac{\sum_{j\neq i}\nu_{ij}^2}{(1-\nu_{ii})^2}\mathtt{e}_i^2 + 2\frac{\mathtt{e}_i}{1-\nu_{ii}}\sum_{j\neq i}\nu_{ij}\mathtt{e}_j.$$

Teniendo en cuenta (4.37) y que, al pertenecer $\mathtt{e}$ al subespacio $\langle\mathbf{X}\rangle^{\perp}$, $\sum_{i=n}\nu_{ij}\mathtt{e}_j = 0$, se obtiene

$$\begin{aligned}\sum_{j\neq i}\mathtt{e}_j(i)^2 &= \sum_{j=1}^{n}\mathtt{e}_j^2 - \mathtt{e}_i^2 + \frac{\nu_{ii}}{1-\nu_{ii}}\mathtt{e}_i^2 - 2\frac{\nu_{ii}}{1-\nu_{ii}}\mathtt{e}_i^2 \\ &= \sum_{j=1}^{n}\mathtt{e}_j^2 - \frac{\mathtt{e}_i^2}{1-\nu_{ii}},\end{aligned}$$

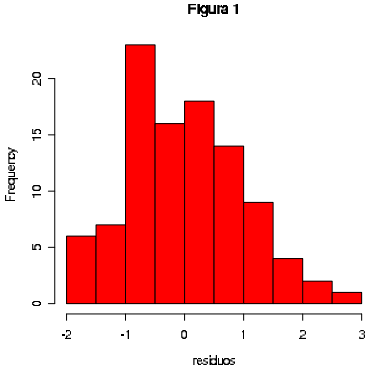
con lo cual se concluye. ∎

Este resultado permite proponer un test global de bondad de ajuste. Efectivamente, si los residuos estudentizados fueran independientes constituirían una muestra aleatoria simple de una distribución $t_{\mathtt{n}-(q+2)}$. Por lo tanto, un test de bondad de ajuste a nivel α a dicha distribución serviría para contrastar la hipótesis inicial de validez del modelo de regresión. En todo caso y en virtud de la desigualdad de Bonferroni (3.46), podemos proponer un test a nivel menor o igual que α, consistente en rechazar la hipótesis inicial cuando $|\mathtt{t}_i| > \mathtt{t}^{\alpha/2\mathtt{n}}_{\mathtt{n}-(q+2)}$, para algún valor de $i = 1, \ldots, \mathtt{n}$, es decir, cuando aparece algún residuo estudentizado muy extremo. Este método resulta claramente conservador, lo cual hace necesario un análisis gráfico de los residuos, ya sean brutos, estandarizados o estudentizados. La desventaja que presenta este tipo de estudio es la subjetividad que conlleva. A favor del mismo destacamos su mayor sensibilidad y que, en muchas ocasiones, arrojan pistas sobre las estrategias a seguir para conseguir un ajuste satisfactorio al modelo.

Desde luego, cabe esperar que la representación gráfica de los residuos estandarizados o estudentizados[14] sea semejante a la que correspondería a una campana de Gauss. Efectivamente, consideremos, por ejemplo[15], un modelo de regresión lineal con $\mathtt{n} = 100$ datos y tres variables explicativas independientes e idénticamente distribuidas según un modelo Uniforme[0,10].

$$Y_i = 5 + 2\mathtt{z}_i[1] + 4\mathtt{z}_i[2] + \mathtt{z}_i[3] + \varepsilon_i, \quad \varepsilon_i \text{ iid } N(0,4). \tag{4.49}$$

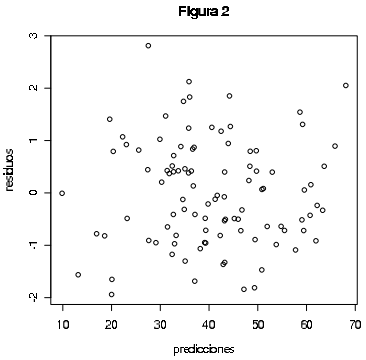
En las figuras 1 y 2 se presentan, respectivamente, el histograma de los residuos brutos tipificados[16] y el diagrama de dispersión simple de las predicciones $\hat{Y}_i$ (eje de abscisas) frente a dichos residuos (ejes de ordenadas).



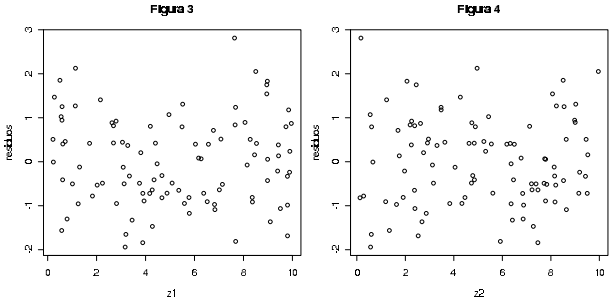
[14] Tener en cuenta que la distribución $t_{\mathtt{n}-(q+1)}$ es muy parecida a la distribución $N(0,1)$.

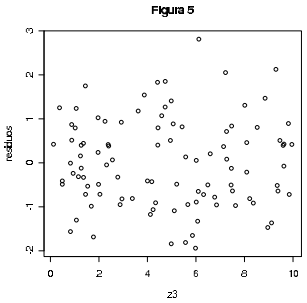
[15] Modelo simulado mediante el programa R.

[16] No coinciden exactamente con los residuos estandarizados ni estudentizados, pero las diferencias son prácticamente inapreciables en una análisis gráfico.



En el histograma se observa un buen ajuste a la campana de Gauss; en el diagrama de dispersión, no se aprecia ninguna tendencia cara en la nube de punto, sino que ésta se sitúa en torno al eje $y = 0$, con mayor densidad de puntos cuanto más cerca se esté de dicho eje con un nivel de dispersión similar. Dado que, en este caso, las predicciones se distribuyen uniformemente sobre el eje de las abscisas, se observa una banda de puntos con anchura uniforme. En general, la anchura de la misma irá en función de la concentración sobre el eje de las abscisas, pues cuanto más untos haya, más probable será obtener residuos extremos. Los gráficos de dispersión de los residuos frente a las distintas variables explicativas (figuras 3, 4 y 5) presentan características muy similares al de la figura 2.





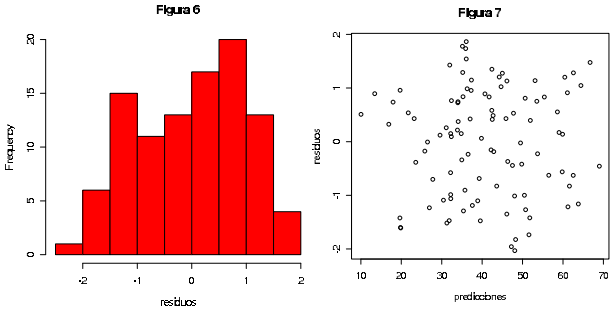
En definitiva, cuando se verifiquen los supuestos del modelo, se obtendrán gráficos como los que se han comentado. Por lo tanto, cuanto más nos desviemos de este tipo de gráficos, más patente será la violación de uno o varios de los supuestos. Para poder ilustrar la trascendencia de dichas violaciones en los métodos de inferencia considerados, indicaremos en cada la ecuación que se obtiene del modelo mediante el EIMV. En este primer caso es

$$y[1] \simeq 6{,}07 + 2{,}00z[1] + 3{,}86z[2] + 0{,}90z[3].$$

A continuación, vamos a ir introduciendo alteraciones en el modelo para ver como afectan a los gráficos de los residuos. En primer lugar, veamos qué sucede cuando se viola exclusivamente el supuesto de normalidad. Para ello, supondremos que las 100 unidades experimentales verifican la ecuación

$$Y_i = 5 + 2\mathbf{z}_i[1] + 4\mathbf{z}_i[2] + \mathbf{z}_i[3] + \varepsilon_i, \quad \varepsilon_i \text{ iid } \texttt{Unifome}(-4, 4). \tag{4.50}$$

Hemos de recordar que, según vimos en el capítulo anterior, este tipo de violación no debería tener gran trascendencia en las inferencias a realizar, siempre y cuando se verifique la condición de Huber y el tamaño de muestra sea los suficientemente grande. En las figura 6 y 7 se presentan de nuevo el histograma de los residuos y el gráfico de dispersión de los mismos frente a las predicciones.



En el histograma de los residuos brutos tipificados se aprecia un mayor aplastamiento que el que correspondería a una campana de Gauss (curtosis negativo). El gráfico de dispersión de los residuos frente a las predicciones no presenta diferencias claras respecto a la figura 2. Sólo mediante un análisis concienzudo se detecta una mayor concentración de puntos en torno a la recta $y = 0$ en el caso normal. En la figura 7, la distribución de los puntos es uniforme. Los diagramas de dispersión frente a las variables explicativas ofrecen una imagen complétamente análoga, por lo que se omiten en este caso. En este caso, se estima la siguiente ecuación:

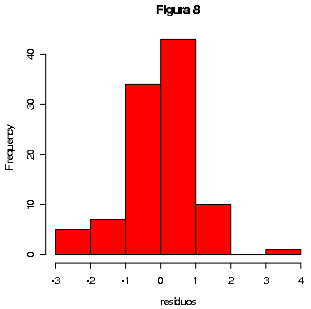
$$y[2] \simeq 6{,}13 + 1{,}84z[1] + 4{,}08z[2] + 0{,}92z[3].$$

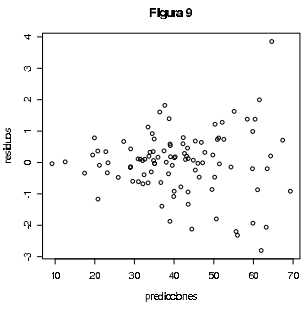
Como podemos observar, la violación de la normalidad que se ha considerado no es óbice para obtener una excelente aproximación a la verdadera ecuación que rige el modelo. Además, tiene escasa repercusión en el análisis gráfico de los residuos.

Introducimos una alteración que puede tener mayor trascendencia en el estudio: la violación del supuesto de homocedasticidad. Para ello simularemos el modelo (4.49), pero suponiendo que los errores ε_i son normales de media 0 y de desviación típica proporcional al valor de $\mathbf{z}[1]$. Es decir,

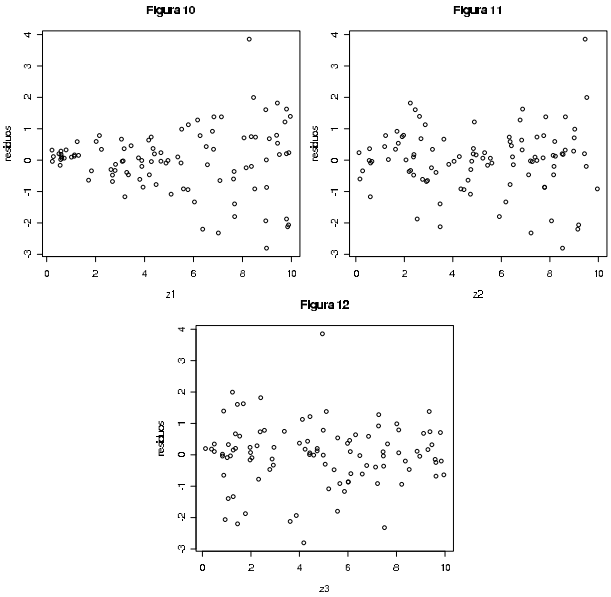
$$Y_i = 5 + 2\mathbf{z}_i[1] + 4\mathbf{z}_i[2] + \mathbf{z}_i[3] + \varepsilon_i, \quad \varepsilon_i \text{ iid } N(0, \mathbf{z}[1]^2). \tag{4.51}$$

Presentamos el histograma de residuos brutos tipificados y el diagrama de dispersión de los mismos frente a las predicciones.





En el histograma no se aprecia un desajuste evidente respecto a la campana de Gauss, aunque un análisis numérico delata un curtosis positivo. En el gráfico de dispersión se observa que la anchura de la nube de puntos crece pareja a la magnitud de las predicciones. Los gráficos de dispersión frente a las variables explicativas resultan en este caso concluyentes.



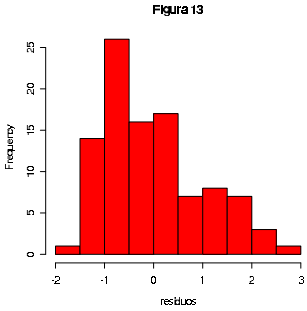
Efectivamente, en este queda perfectamente patente que la heterocedasticidad del modelo es únicamente achacable a la variable $z[1]$. En los gráficos restates no se aprecian anomalías, salvo un residuo extremo que se corresponde con un dato mal explicada por el modelo. La ecuación estimada es la siguiente:

$$y[3] \simeq 5{,}37 + 2{,}27z[1] + 3{,}74z[2] + 0{,}92z[3]$$

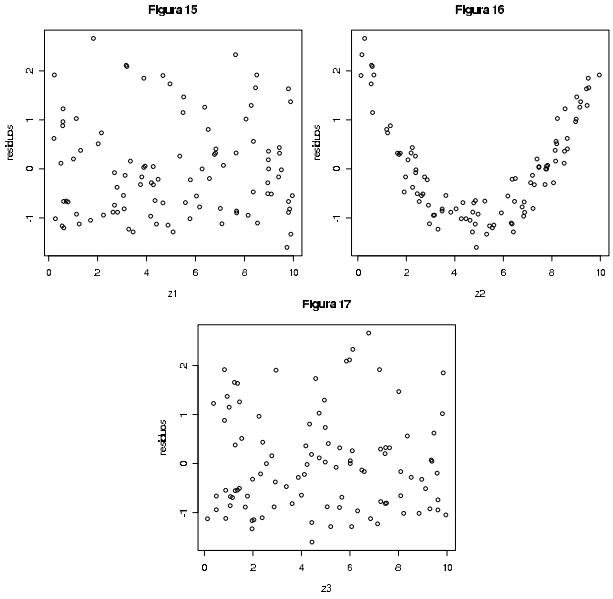
La diferencia respecto a la ecuación verdadera no es aún muy ostensible, al menos en este caso. A continuación, veamos qué sucede cuando se viola el supuesto de linealidad. Para ello simularemos el modelo

$$Y_i = 5 + 2\mathbf{z}_i[1] + 10\mathbf{z}_i[2]^2 + \mathbf{z}_i[4] + \varepsilon_i, \quad \varepsilon_i \text{ iid } N(0,4), \tag{4.52}$$

con un total de 100 unidades experimentales independientes. El histograma de los residuos y el gráfico de dispersión frente a la predicciones se muestran a continuación.



En el histograma se observa una clara asimetría con sesgo positivo. Lo más importante es que, al contrario que en los gráficos anteriores, el gráfico de dispersión presenta una clara tendencia, pues no se distribuye en torno al eje de abscisas de forma simétrica, sino que existe un patrón de comportamiento que puede hacernos recordar, en este caso, la forma de una parábola. Esta situación suele delatar el incumplimiento del supuesto de linealidad. Confrontamos a continuación los residuos con los distintos vectores explicativos con el objeto de detectar la variable o variables responsables de la falta de linealidad. En este caso, queda patente que se trata de `z`[2], tal y como se aprecia en las figuras siguientes.



Podemos apreciar, efectivamente, una clara forma de parábola cuando consideramos la variable $\mathbf{z}[2]$, lo cual revela una información valiosísima de cara a solucionar el desajuste (el desajuste se soluciona sustituyendo $\mathbf{z}[2]$ por su cuadrado). Hemos de empezar a tener muy claro que la correlación lineal entre los vectores explicativos (colinealidad) supone un pesado lastre en el análisis de regresión. De hecho, si las variables fueran no fueran incorreladas, no descubriríamos tan fácilmente que $\mathbf{z}[2]$ es la variable responsable de la no linealidad.

No obstante, hemos de advertir claramente que en el esquema que estamos siguiendo contamos con dos ventajas enormes a la hora de detectar violaciones del modelo: en primer lugar, éstas se introducen de manera aislada en cada caso; segundo, las variables explicativas son incorreladas. Este factor es fundamental pues, de no ser así, os resultaría muy difícil determinar qué variable es la reponsable de la heterocedasticidad o falta de linealidad. Cuando se da una relación lineal entre las variables explicativas, puede ser de utilidad el uso de gráficos parciales, que consisten en controlar todas las variables respuesta excepto una y enfrentar entre sí los residuos de la variable explicativa restante y la variable respuesta dadas las variable controladas. Así se elimina gráficamente el efecto de la relación lineal entre las variables explicativas. Como ya sabemos, el coeficiente de correlación entre ambos residuos es el coeficiente de correlación parcial. Precisamente, el test parcial para contrastar la hipótesis inicial $\beta_j = 0$ se basa únicamente en el valor de dicho coeficiente, es decir, que aporta un valor significativo cuando en el gráfico parcial se observa una correlación clara.

La ecuación estimada para este modelo es la siguiente:

$$y[4] \simeq -13{,}20 + 2{,}21z[1] + 10{,}04z[2] + 0{,}97z[3].$$

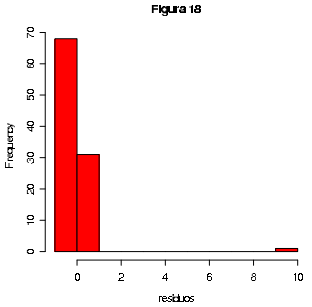
Como podemos observar, el efecto de la no linealidad se deja notar ostensiblemente en el coeficiente de $z[2]$. De existir multicolinealidad entre las variables explicativas, afectaría sin duda a las demás variables.

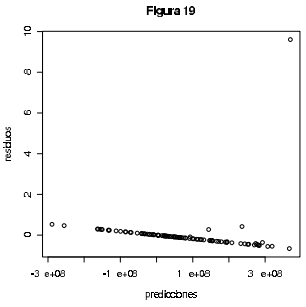
Es poco habitual, en la práctica, que se produzca una única violación aislada del modelo, ya sea por no normalidad, por heterocedasticidad o por no linealidad, como hemos visto hasta ahora mediante sendos ejemplos. Lo más frecuente es que se incumplan simultáneamente varios supuestos, por no decir todos. Por ejemplo, consideremos el modelo multiplicativo

$$Y_i = 5 \cdot \mathbf{z}_i[1]^2 \cdot \mathbf{z}_i[2]^4 \cdot \mathbf{z}_i[3] \cdot \varepsilon_i, \quad \varepsilon_i \text{ iid } LN(0,4) \ ^{17}. \tag{4.53}$$

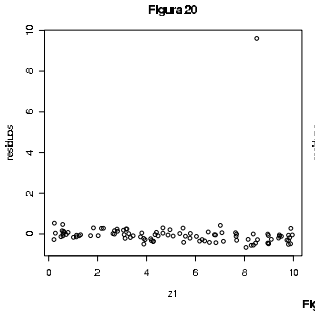
[17] Por $LN(\mu, \sigma^2)$ se denotarña la distribución positiva cuyo logaritmos es una normal de media μ y varianza σ^2.

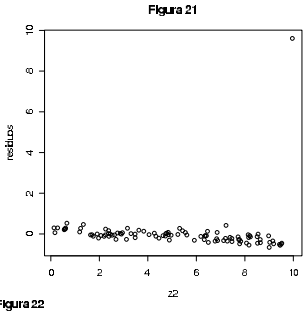
Veamos qué aspecto tienen el histograma de los residuos brutos tipificados y el diagrama de dispersión frente a las predicciones.

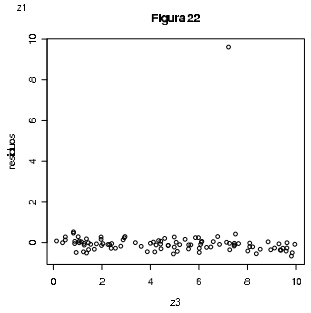




En el histograma podemos apreciar un fuerte sesgo positivo con al menos un valor muy extremo. Concretamente, podemos encontrar un residuo estudentizado con valor próximo a 10, lo cual se traduciría en un resultado significativo al aplicar el test de valores extremos. Este sesgo queda también patente en el diagrama de dispersión, junto con evidente tendencia de la nube de puntos. Los gráficos de dispersión frente a las variables explicativas son los siguientes:



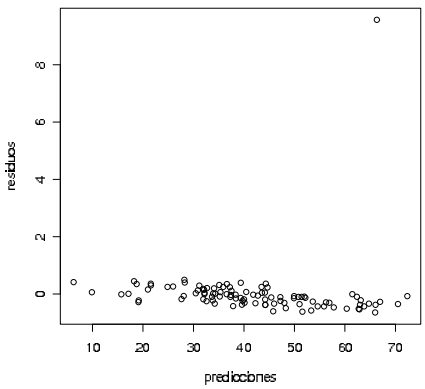




Desde luego, ante uno gráficos así no procede continuar, sin más, con el análisis de regresión estudiado en las dos primeras secciones. Nótese que la ecuación de regresión (lineal) no puede ser en modo alguno acertada. Sin embargo, en este caso, aunque el desajuste es evidente la solución es bien sencilla, pues si reemplazamos cada variable por su logaritmo obtenemos automáticamente el modelo (4.49).

La aparición en el modelo de unidades experimentales anómalas, como ha sido el caso (este problema se tratará más adelante) puede achacarse a una violación de los supuestos, un error en la toma de datos o, simplemente, al propio azar. Veamos hasta qué punto puede influir en el análisis gráfico de los residuos. Para ello, añadimos al modelo (4.49) una observación anómala, obteniéndose el siguiente diagrama de dispersión frente a las predicciones.

Figura 23



Podemos observar un residuo muy alto (el valor estudentizado es próximo a 10) y cierta tendencia lineal negativa en el resto de la nube. Si comparamos este gráfico con la figura 2 entenderemos hasta qué punto una única unidad experimental puede influir en el análisis de los residuos y, en general, en el de regresión.

4.5. Transformaciones de variables y MCP.

Cuando el análisis de los residuos delata una manifiesta violación de los supuestos del modelo, podemos optar por otro tipo de estudio, como puede ser una regresión no paramétrica o robusta, o bien por adecuar nuestros datos al modelo de regresión lineal mediante transformaciones de las variables en juego e, incluso, la adición de nuevos vectores explicativos. Por ejemplo, hemos visto que en la simulación (4.53),

los residuos evidencian una clara violación de los supuestos del modelo de regresión lineal. No obstante, si reemplazamos las variables originales, tanto las explicativas como la respuesta, por sus respectivos logaritmos, se verificará un ajuste perfecto al modelo. Esto sucede con cierta frecuencia, concretamente en los modelos en los cuales los vectores explicativos no tienen un efecto aditivo sino multiplicativo. En este caso, observamos una falta de normalidad y de homocedasticidad asociada a una falta de linealidad, de manera que al resolver la última se resuelven por añadidura las primeras.

Por desgracia, es bastante habitual que suceda lo contrario, es decir, que si aplicamos una transformación que permita verificar uno de los supuestos, deje de verificarse otro que, en principio, se satisfacía. Por ejemplo, si se satisface la linealidad y aplicamos una transformación a la variable respuesta (logaritmo, cuadrado,...) con objeto de conseguir normalidad, no es de extrañar que la relación lineal se rompa. El problema es pues bastante complicado, porque, aunque existen diversos métodos para verificar los supuestos por separado, necesitaríamos un algoritmo que permitiera verificarlos todos conjuntamente y que estuviera implementado en los programas estadísticos. Primeramente, debemos asumir que el ajuste no se conseguirá en multitud de ocasiones y, por tanto, debemos estar preparados para aplicar técnicas no paramétricas cuando sean necesarias. No obstante, proponemos, a modo orientativo, una serie de métodos que, aplicados aislada o conjuntamente, pueden lograr un ajuste satisfactorio al modelo de regresión. Muchos de ellos tiene un denominador común: de una forma u otra consisten en transformaciones de las variables, bien sea la respuesta, las explicativas o ambas.

1. **Método de Box-Cox:** este procedimiento se ideó, en principio, para obtener una trasformación de la variable respuesta que permita un ajuste satisfactorio a un distribución normal o, al menos, simétrica. Se basa en la idea de que una potencia con exponente mayor que 1 dispersa los datos elevados, por lo que puede eliminar un sesgo negativo. Por contra, una potencia con exponente menor que 1 o el propio logaritmo neperiano dispersan lo datos próximos a cero, por lo que pueden eliminar un sesgo positivo[18]. De esta forma, se considera la función ϕ, de $\mathbb{R} \times \mathbb{R}^+$ en $\mathbb{R}$ que asocia a cada λ en $\mathbb{R}$ y cada $x > 0$ el valor

$$\phi(\lambda, x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \ln x & \text{si } \lambda = 0 \end{cases}$$

[18]Para evitar problemas con potencias y logaritmos se supone que los datos son siempre positivos. Si sucede lo contrario, basta con trasladarlos inicialmente, restándoles el valor mínimo.

Como podemos apreciar, se ha efectuado una corrección sobre la función indicada anteriormente con el objeto de aportar regularidad a la transformación. Efectivamente, puede comprobarse, haciendo uso de la regla de L'Hopital, que la función ϕ, así definida, es continua. También es continua en $\mathbb{R}$ la derivada parcial $\partial\phi/\partial x$. Aunque, como hemos comentado, este método está originalmente orientado a conseguir normalidad, se utilizará para lograr un ajuste aproximado a todos los supuestos del modelo de regresión. En la práctica, la transformación se aplicará a una muestra de $\mathtt{n}$ datos, por lo que es necesario extender la definición a un vector $\mathtt{n}$-dimensional. Así, se define la función Φ de $\mathbb{R} \times (\mathbb{R}^+)^{\mathtt{n}}$ en $\mathbb{R}^n$ que asigna a cada λ real y cada vector $X = (x_1, \ldots, x_{\mathtt{n}})'$ el vector $\Phi(\lambda, X) = \big(\phi(\lambda, x_1), \ldots, \phi(\lambda, x_{\mathtt{n}})\big)'$.

El método, expresado en su forma más general, consiste en suponer que existe un valor λ de tal forma que el vector aleatorio $\Phi(\lambda, Y)$ sigue un modelo lineal normal. Por lo tanto, la media de $\Phi(\lambda, Y)$ debe estar restringida a cierto subespacio V de $\mathbb{R}^{\mathtt{n}}$ (queda excluido el propio $\mathbb{R}^{\mathtt{n}}$, pues en ese caso el modelo lineal es inviable). Por lo tanto, estamos considerando el siguiente modelo estadístico

$$Y \sim \big[N_{\mathtt{n}}(\mu, \sigma^2 \mathtt{Id})\big]^{(\Phi(\lambda, \cdot\ ))^{-1}}, \quad \lambda \in \mathbb{R},\ \mu \in V,\ \sigma^2 > 0.$$

El valor adecuado de λ se estima por el método de máxima verosimilitud, es decir, se escogerán los parámetros (es decir, la distribución) λ, μ y σ^2 que hagan más verosímil la observación Y. En virtud del teorema del cambio de variables[19], se tiene que la función de verosimilitud $\mathcal{L}$ del modelo se expresa mediante

$$\mathcal{L}(\mathtt{y}; \lambda, \mu, \sigma^2) = \mathcal{L}_0\left(\Phi(\lambda, \mathtt{y}); \mu, \sigma^2\right) \left(\prod_{i=1}^{\mathtt{n}} \mathtt{y}_i\right)^{\lambda - 1},$$

donde $\mathcal{L}_0$ denota la función de verosimilitud correspondiente al modelo lineal normal (modelo de regresión). De esta forma, dado $\lambda \in \mathbb{R}$, se sigue del teorema 3.9 que

$$\begin{aligned} \max_{\mu \in V,\ \sigma^2 > 0} \mathcal{L}(Y; \lambda, \mu, \sigma^2) &\propto \left(\hat{\sigma}^2\right)^{-\mathtt{n}/2} \left(\prod_{i=1}^{\mathtt{n}} \mathtt{y}_i\right)^{\lambda - 1} \\ &\propto \left(\frac{\hat{\sigma}}{\left(\overset{\bullet}{Y}\right)^{\lambda - 1}}\right)^{-\mathtt{n}}, \end{aligned}$$

[19] Se efectúa aquí ua razonamiento análogo al realizado en la demostración de la proposición 2.5.

donde $\overset{\bullet}{Y}$ denota la media geométrica de las componentes de Y. Fijo λ, el máximo se alcanza con los estimadores de máxima verosimilitud de μ y σ^2 calculados a partir de la observación $\Phi(\lambda, Y)$. La cuestión es, por tanto, encontrar el valor de λ que maximice esta función. El último término resulta de elevar a $-\mathtt{n}$ el estimador de σ que se obtendría si multiplicáramos escalarmente $\Phi(\lambda, Y)$ por la media geométrica de Y elevada a $(1-\lambda)$. En consecuencia, si consideraremos el vector $Y^{(\lambda)} = \left(Y_1^{(\lambda)}, \ldots, Y_{\mathtt{n}}^{(\lambda)}\right)'$, donde

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^{\lambda}-1}{\lambda \overset{\bullet}{Y}^{\lambda-1}} & \text{si } \lambda \neq 0 \\ \\ \overset{\bullet}{Y} \ln Y_i & \text{si } \lambda = 0 \end{cases},$$

el problema se reduce a buscar el valor de λ que minimice

$$\left\| Y^{(\lambda)} - P_V Y^{(\lambda)} \right\|^2 \tag{4.54}$$

La solución final al problema estará en función del subespacio V escogido o, lo que es lo mismo, de las restricciones impuestas a la media. Destacamos tres casos:

a) El más restrictivo es $V = \langle 1_{\mathtt{n}} \rangle$. En tal caso, estaremos afirmando que $\Phi(\lambda, Y)$ es una muestra aleatoria simple de una distribución normal. Salvo que se dé la total incorrelación entre la variable respuesta y los vectores explicativos, una situación de este tipo sólo puede plantearse en un problema de correlación (véase ejemplo 1 de la Introducción), es decir, aquél en el cual se eligen al azar y de forma independiente $\mathtt{n}$ unidades experimentales a las cuales se les miden q variables explicativas y una variable respuesta. En ese caso, este tipo de transformación puede aplicarse también a las distintas variables explicativas con el objeto de aproximarnos a las condiciones del modelo de correlación lineal (véase capítulo 4). También podemos buscar una transformación del vector aleatorio $(q+1)$-dimensional con la intención de conseguir una muestra aleatoria simple de una distribución $(q+1)$-normal, que es exactamente la condición de partida del modelo de correlación lineal. Para ello se utilizaría una versión multivariante del método de Box-Cox[20]. No obstante, este último procedimiento puede pecar de ambicioso.

[20] Ver volumen dedicado al Análisis Multivariante.

Sin embargo, en un modelo de regresión puro (véase ejemplo 2), en el que los vectores explicativos están controlados de antemano, los valores de la variable respuesta no pueden considerarse una muestra aleatoria simple de alguna distribución concreta, a menor quese dé la incorrelación total. Por ello debemos imponer otro tipo de restricciones.

En todo caso y teniendo en cuenta (4.54), el método propuesto consiste (cuestión propuesta) en encontrar el valor de λ que minimiza la varianza total muestral de $Y^{(\lambda)}$.

b) En un modelo de regresión propiamente dicho con una matriz $\mathbf{X}$, se puede considerar la restricción $\mu \in \langle \mathbf{X} \rangle$, es decir, $\mathtt{E}[\Phi(\lambda, Y)] = \mathbf{X}\beta$, para algún $\beta \in \mathbb{R}^{q+1}$. Ello equivale a buscar una transformación que verifique todos los supuestos del modelo de regresión: normalidad de cada observación, homocedasticidad y linealidad (además de inependencia). Teniendo en cuenta (4.54), el método consiste en encontrar el valor de λ que minimice

$$\left\| Y^{(\lambda)} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y^{(\lambda)} \right\|^2$$

c) Si el modelo de regresión es del tipo (4.41) o, para ser más preciso, como el del ejemplo 2 de la introducción, podemos obviar el supuesto de linealidad considerando $V = \langle \mathtt{v}_1, \ldots, \mathtt{v}_k \rangle$. En ese caso, estaremos buscando una transformación de los datos que permitan verificar los supuestos de normalidad (de cada observación) y homocedasticidad. Este modelo es menos restrictivo que los anteriores, por lo que se obtendrá un mayor máximo para la función de verosimilitud o, equivalentemente, un menor mínimo para (4.54) (queda como ejercicio determinar qué expresión se debe minimizar). Ello se traduce en una mejor aproximación al modelo buscado, lo cual es lógico dado que nuestras exigencias son menores.

Existe un problema de carácter técnico en el método que no hemos mencionado aún. Radica en la búsqueda del mínimo (4.54). En ese sentido, lo más habitual es escogerlo mediante un rastreo con diversos valores de λ. Si no disponemos de los medios adecuados, se aconseja tantear únicamente con los valores λ =-1, 0, 0.5, 1, 2. Es decir, considerando las funciones

$$\frac{1}{x}\,, \quad \ln x\,, \quad \sqrt{x}\,, \quad x\,, \quad x^2. \tag{4.55}$$

En la práctica, es difícil que se obtenga un buen ajuste co algún valor de λ si no se ha logrado con ninguno de estos cinco. Además, muchos autores rechazan el uso de transformaciones poco naturales pues desvirtúan la interpretación de os resultados en términos prácticos.

2. **Transformación de variables explicativas. Regresión polinómica:**

   Un desajuste debido a falta de linealidad puede eliminarse en ocasiones manipulando únicamente las variables explicativas o, mejor dicho, vectores explicativos. El hecho de operar únicamente sobre éstas permite conservar la normalidad y la homocedasticidad en el caso de que estos supuestos se verifiquen. Una estrategia en ese sentido puede ser tantear con las distintas transformaciones de (4.55) en cada uno de los vectores explicativos hasta conseguir un ajuste satisfactorio. No obstante, los gráficos de residuos frente a vectores explicativos pueden ofrecer pistas sobre qué variables transformar y el tipo de transformación a efectuar. Por ejemplo, en la simulación (4.52) se obtiene la linealidad considerando el cuadrado de $\mathbf{z}[2]$, cosa que puede intuirse a tenor de las figuras 8, 9 y 10. Esto resulta bastante claro dado que los vectores $\mathbf{z}[1]$, $\mathbf{z}[2]$ y $\mathbf{z}[3]$ son, en este caso, incorrelados.

   Por otra parte, del Teorema de Aproximación de Weierstrass se sigue que cualquier función continua puede aproximarse localmente por un polinomio. Ello nos lleva a considerar la posibilidad de añadir al modelo nuevos vectores explicativos que serán potencias enteras y productos de los ya existentes. De esta forma, una ecuación lineal en términos de estas nuevas variables equivale a una ecuación polinómica en términos de los vectores explicativos originales. Este tipo de estudio recibe el nombre de regresión polinómica. En el caso de la regresión simple resulta más fácil al no tener que introducir productos entre variables. Además, puede demostrarse fácilmente que, por muchas potencias de la variable explicativa que añadamos, el rango de la matriz resultante seguirá siendo completo. Una vez introducidos los distintos monomios y si se consigue un ajuste satisfactorio, puede depurarse el modelo mediante una selección de variables. De todas formas se aconseja no superar el grado 2 en una regresión polinómica.

   Los dos métodos considerados pueden combinarse si se realiza una regresión polinómica a partir de las variables transformadas, que pueden ser inversas, logaritmos o raíces cuadradas de las originales. Por otra parte, además de las transformaciones ya estudiadas existen otras más drásticas, como la transformación logística, que se estudia en el capítulo 8.

3. **Mínimos Cuadrados Ponderados (MCP):**

   Este procedimiento se plantea como una posible solución al problema de hete-

rocedasticidad. En un modelo del tipo siguiente

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_{\mathtt{n}} \end{pmatrix} \sim N_{\mathtt{n}} \left(\mathtt{X}\beta, \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{\mathtt{n}}^2 \end{pmatrix} \right), \quad \beta \in \mathbb{R}^{q+1}, \quad \sigma_1^2, \dots, \sigma_{\mathtt{n}}^2 > 0,$$

ni los estimadores propuestos en la primera sección, denominados mínimo-cuadráticos, ni los tests estudiados en la sección segunda poseen la idoneidad que les correspondería en un modelo homocedástico. No obstante, puede suceder que exista una función conocida g, de $\mathbb{R}^q$ en $\mathbb{R}^+$ tal que

$$\sigma_i \propto g(\mathtt{z}_i), \quad i = 1, \dots, \mathtt{n}.$$

En se caso, el modelo podría expresarse mediante

$$Y \sim N_{\mathtt{n}}(\mathtt{X}\beta, \sigma^2 D_g), \quad \beta \in \mathbb{R}^{q+1}, \quad \sigma^2 > 0, \tag{4.56}$$

donde

$$D_g = \begin{pmatrix} g^2(z_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & g^2(z_{\mathtt{n}}) \end{pmatrix}.$$

Estaríamos pues ante un modelo como el estudiado en el problema 3.14. En ese caso, tomando $Y^* = D_g^{-1/2} Y$ y $\mathtt{X}^* = D_g^{-1/2}\mathtt{X}$, se verifica que

$$Y^* \sim N_{\mathtt{n}}(\mathtt{X}^*\beta, \sigma^2), \quad \beta \in \mathbb{R}^{q+1}, \quad \sigma^2 > 0. \tag{4.57}$$

Se denomina estimador por mínimos cuadrados ponderados de β al estimador mínimo cuadrático de β para el modelo (4.57), es decir

$$\hat{\hat{\beta}} = (\mathtt{X}' D_g^{-1} \mathtt{X})^{-1} \mathtt{X}' D_g Y.$$

Puede probarse entonces que $\mathtt{X}\hat{\hat{\beta}}$ es el EIMV y EMV de $\mathtt{X}\beta$ en el modelo (4.56). Realmente, al considerar Y^* y $\mathtt{X}^*$ lo que estamos haciendo es dividir Y_i y x_i por el escalar $g(\mathtt{z}_i)$, ara todo $i = 1, \dots, \mathtt{n}$, es decir, se pondera cada unidad experimental de forma inversamente proporcional a la varianza que presenta, de ahí el nombre.

Hemos de advertir, no obstante, que el éxito de este método esta supeditado a una buena elección de la función g anterior, lo cual no es nada fácil.

Otros métodos para conseguir un satisfactorio ajuste al modelo de regresión pueden encontrarse en Rawlings et al. (1999). En todo caso, ante un problema tan complejo como este, convendría seguir algunas pautas orientativas, lo más concisas posibles. En vista de lo estudiado hasta ahora, nos aventuramos a proponer tres estrategias:

1. La primera es válida para modelos de correlación. Se trata de transformar todas las variables en juego mediante el método de Box-Cox o por simple tanteo para obtener, aproximadamente, muestras aleatorias simples de distribuciones normales. Esta situación nos aproximaría a las condiciones del modelo de correlación lineal. Tener en cuenta que, si éstas se dieran, la normalidad, homocedasticidad y linealidad de la distribución condicional se obtendrían automáticamente (ver capítulo 4).

2. La segunda es válida tanto para problemas de regresión pura como de correlación. Se trata de buscar primero la linealidad mediante transformaciones del tipo (4.55) para todas las variables o mediante regresión polinómica, para después buscar la homocedasticidad mediante MCP.

3. La tercera estrategia es válida únicamente para modelos de regresión pura. Consiste en intentar eliminar primero la heterocedasticidad mediante el método de Box-Cox (manipulando únicamente la variable respuesta) y, después, buscar la linealidad manipulando las variables explicativas. En todo caso, considerar únicamente transformaciones sencillas del tipo (4.55) o regresiones polinómicas.

Posiblemente, las dos primeras estrategias son las más factibles en la práctica No obstante y como dijimos al comienzo de la sección, conviene tener en cuenta también los distintos métodos alternativos de regresión, incluyendo los no paramétricos. Comentamos muy brevemente algunos de ellos.

En primer lugar, veamos el más natural desde el punto de vista teórico. Supongamos que $z = (z[1], \ldots, z[q])'$ es un vector aleatorio de manera que, conjuntamente con y, admiten una densidad respecto a ala medida de Lebesgue en $\mathbb{R}^{q+1}$. El objetivo de la regresión es encontrar el valor medio esperado para la variable y cuando se conocen el resultado de Z, es decir, $\mathtt{E}[y|z]$. En el caso de que y y z sigan conjuntamente un modelo de distribución $(q+1)$-normal, la esperanza condicional es, en virtud de la proposición 2.5, una función afín de z, lo cual nos llevaría a un modelo de Regresión. En general, se trata de una función cuyo valor en $z = \mathtt{z}$ es la media de la ditribución condicional de y respecto a $z = \mathtt{z}$, que puede calcularse a partir de la densidad $f_{y|z=\mathtt{z}}$ de dicha distribución condicional. Por su parte, esta densidad pude construirse

mediante

$$f_{y|z=\mathtt{z}}(\mathtt{y}) = \frac{f(\mathtt{y},\mathtt{z})}{f_z(\mathtt{z})}, \tag{4.58}$$

donde f y f_z denotan las densidades conjunta y marginal en z, respectivamente. Por lo tanto, el problema se reduce a estimar ambas densidades mediante el conocido Método del Núcleo, y los únicos inconvenientes son los inherentes a este método de estimación. Fundamentalmente, hemos de mencionar el problema de la elección del ancho de banda adecuado y, sobre todo, lo que en Estadística no Paramétrica se conoce como *maldición de la dimensión*: que el número de datos requerido para lograr una estimación satisfactoria de la densidad crece exponencialmente en relación con la dimensión considerada[21]

Otros métodos alternativos que aporta resultados muy satisfactorios son los de regresión local. Consisten en calcular una función de regresión (lineal o polinómica) en cada punto dependiendo de lo que se observe en un entorno del mismo. Se trata pues de un patrón de comportamiento cambiante que permite un ajuste muy satisfactorio en problemas complejos. Entre estos métodos podemos destacar el de Nadaraya-Watson, el de Gasser-Müller o los de regresión polinómica local. Para más información, ver Fan & Gijbels (1996).

4.6. Análisis de valores influyentes

En esta sección se abordará el diagnóstico de un problema que, si bien no ha de ser necesariamente consecuencia de la violación de los supuestos, puede acarrear mayores perjuicios que ésta. Se trata de la presencia de valores influyentes. Entendemos por valor influyente a una unidad experimental con una influencia determinante en el resultado del análisis, es decir, tal que su supresión del modelo provoca una cambio sustancial en la ecuación de regresión estimada, tal y como se observa en las figuras 24 y 25.

[21] Ver Silverman (1986), tabla 2.2.

Figura 24

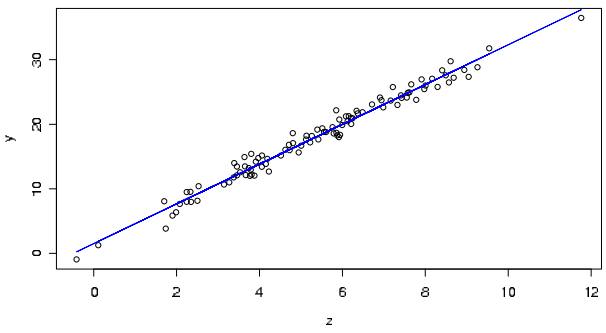
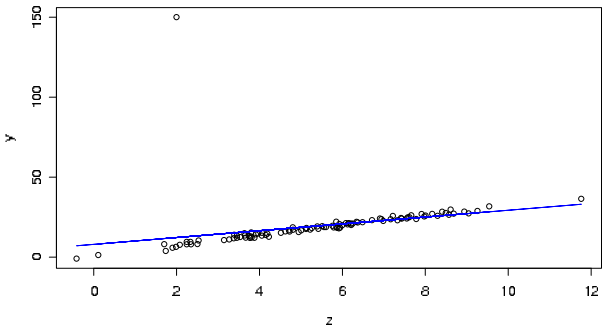


Figura 25



Una situación como la observada resulta inadmisible desde el punto de vista de la Inferencia Estadística, pues no parece razonable extraer conclusiones de carácter poblacional cuando vienen determinadas por un único individuo. Por ello, cuando se detecta algún o algunos valores influyentes debemos valorar dos circunstancias: primeramente, si esa influencia es debida a la falta de linealidad en el modelo, en cuyo caso debemos intentar conseguir un mejor ajuste, aplicando los métodos estudiados en la sección anterior; si no es el caso, debemos considerar la posibilidad de que el dato en sí constituya un error en la medición o en el proceso de tratamiento de la información, en cuyo caso debe ser eliminado. Si el dato es correcto, sería conveniente aplicar métodos de *Regresión Robusta*. En Carmona (2005) se presentan diversos procedimientos para construir una recta de regresión simple resistente ante la presencia de datos atípicos. En el caso de la regresión múltiple, el problema se resuelve ponderando negativamente los residuos de los datos atípicos[22].

[22] Ver Peña (1993).

De lo dicho anteriormente puede inferirse que el diagnóstico de datos influyentes ha de llevarse a cabo o con anterioridad o, a la sumo, paralelamente al análisis de los residuos. Veamos a continuación cuatro métodos para disgnosticar la presencia de valores influyentes:

1. **Distancias de Cook:** este método se basa en la idea de considerar influyente la unidad experimental i-ésima cuando existe una diferencia sustancial entre la estimación del vector β con dicha unidad y sin ella. De esta forma y teniendo en cuenta la región de confianza (3.16) para el parámetro β, definimos para la unidad i-ésima la siguiente distancia, denominada distancia de Cook:

$$D_i^2 = \frac{\big(\hat{\beta} - \hat{\beta}(i)\big)'\mathtt{X}'\mathtt{X}\big(\hat{\beta} - \hat{\beta}(i)\big)}{(q+1)\hat{\sigma}^{2,\mathtt{I}}} = \frac{\big\|\hat{Y} - \hat{Y}(i)\big\|^2}{(q+1)\hat{\sigma}^{2,\mathtt{I}}}.$$

Realmente, lo que estamos haciendo es determinar si la estimación de β sin la unidad i-ésima pertenece a la región de confianza para β construida con todas las unidades, incluida la i-ésimaq. A partir de esto, podríamos construir un test de hipótesis consistente en determinar si D_i^2 es mayor que $F^{\alpha}_{q+1,\mathtt{n}-q-1}$. No obstante, como la comparación se va a efectuar con todas las unidades experimentales, la Desigualdad de Bonferroni (tener en cuenta que las distancias de Cook no so independientes) induce a compara cada D_i^2 con $F^{\alpha/\mathtt{n}}_{q+1,\mathtt{n}-q-1}$, de forma que si alguna distancia de Cook supera dicho valor se diagnosticaría la presencia de valores influyentes. No obstante, dicho método resultaría enormemente conservador. En la práctica es muy común confrontar cada D_i^2 con el cuantil $F^{0,50}_{q+1,\mathtt{n}-q-1}$, de manera que los puntos que lo superen se consideran influyentes. Otros autores proponen consider como punto de corte $4/\mathtt{n}$.

Por otra parte, de (4.46) se sigue directamente que

$$D_i^2 = (q+1)^{-1} r_i^2 \frac{\nu_{ii}}{1-\nu_{ii}}, \quad i = 1, \ldots, \mathtt{n}. \tag{4.59}$$

Esta igualdad explica perfectamente el porqué de la influencia de una determinada unidad, pues vemos que la distancia de Cook es proporcional al producto de dos factores: el primero de ellos, $\nu_{ii}(1-\nu_{ii})^{-1}$ es mayor cuanto más extrema sea la observación $\mathtt{z}_i$ en el sentido de la distancia de Mahalanobis d^2 definida en (4.23). Efectivamente, según se sigue de (4.38), se tiene que

$$\nu_{ii}(1-\nu_{ii})^{-1} = \mathtt{f}(d^2(\mathtt{z}_i, \overline{\mathtt{z}})),$$

siendo $\mathtt{f}$ la función creciente de $[0, \mathtt{n}-1]$ en $\overline{\mathbb{R}}$ definida mediante

$$\mathtt{f}(x) = \frac{1+x}{\mathtt{n}-1-x}$$

Sabemos por (4.45) que, cuanto más extremos sea el dato, menor será la varianza del residuo $\mathtt{e}_i$. El caso extremo se da cuando $\nu_{ii} = 1$. En una regresión lineal simple, ello equivale a que todos los valores explicativos salvo $\mathtt{z}_i$ sean idénticos. En tal caso $d(\mathtt{z}_i, \overline{z}) = \mathtt{n} - 1$ y la varianza de $\mathtt{e}_i$ es nula. Por lo tanto, al ser la media del residuo nula en todo caso, la función pasa necesariamente por $(\mathtt{z}_i, Y_i)$, tal y como sucede en las figura 26 y 27.

Figura 26

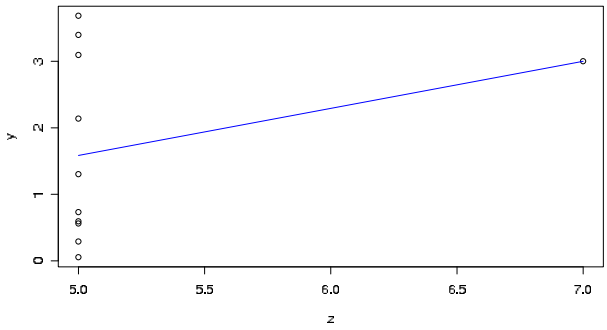
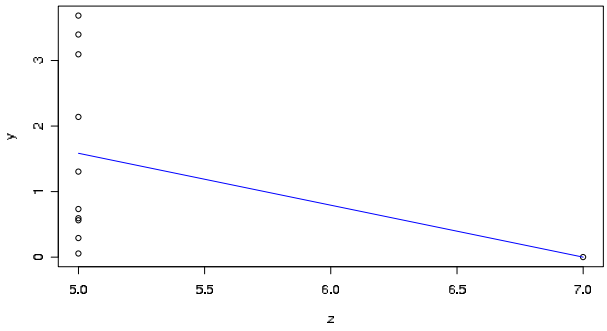
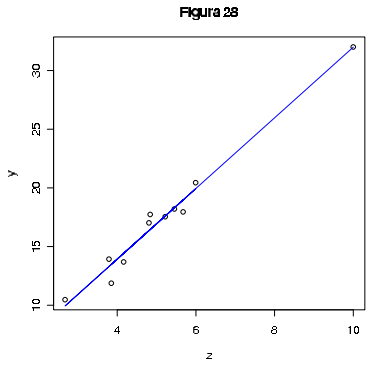


Figura 27



Razonando por continuidad deducimos que los valores distantes del centroide en términos relativos tienen residuos menores que los valores cercanos. En términos heurísticos, podríamos decir que se realiza un mayor esfuerzo por ajustar bien los datos extremos. En ese sentido podemos afirmar que estos valores poseen un gran peso en la regresión. Pero ello no quiere decir que sean de hecho decisivos pues, como podemos ver en (4.59), interviene un segundo factor que es el residuo estandarizado. Por lo tanto, una unidad experimental será más influyente

cuanto peor ajustada esté por la ecuación de regresión. Si ambos factores ($\mathtt{z}_i$ extremo y dato mal ajustado) concurren, como ocurre en la figura 25, el dato resulta ser enormemente influyente. Pero ello no tiene por qué suceder, como vemos en la figura 28.



2. **Dfbetas:** el planteamiento es similar al anterior pero considerando por separado los estimadores de las componentes de β, es decir, que un unidad experimental se considera influyente cuando su eliminación supone un cambio sustancial en alguna de las estimaciones de $\beta_0, \ldots, \beta_q$. Así, teniendo en cuenta en esta ocasión el intervalo de confianza para β obtenido en (4.10), definimos el estadístico

$$\mathtt{Df}\beta_j(i) = \frac{\hat{\beta}_j - \hat{\beta}_j(i)}{\hat{\sigma}_{\mathtt{I}}\sqrt{\psi_{jj}}}, \quad j = 0, \ldots, q, \ i = 1, \ldots, \mathtt{n}.$$

Siguiendo el midmo razonamiento que con las distancias de Cook, podríamos connfrontar los valores obtenidos con con $t^{\alpha}_{\mathtt{n}-q-1}$ para un análisis individual o mejor con $t^{\alpha/\mathtt{n}(q+1)}_{\mathtt{n}-q-1}$ para un análisis conjunto. Dicho método resulta muy conservador, por lo que en la práctica se utilizan puntos de corte más bajos[23].

3. **Dfajustados:** la idea también es similar al planteamiento de las distancias de Cook pero teniendo en cuenta las predicciones en lugar de las estimaciones de β. Al igual que en el caso de las Dfbetas, se realiza un estudio individual. De esta forma, teniendo en cuenta el intervalo de confianza para el valor medio

[23] Ver Rawlings et al. (1998), pag. 364.

esperado de una predicción obtenido en (4.24) [24], se define

$$\texttt{Dfadj}(i) = \frac{\hat{Y}_i - \hat{Y}_i(i)}{\hat{\sigma}_{\text{I}}(i)\sqrt{\nu_{ii}}}, \quad i = 1, \ldots, \texttt{n}.$$

Este valor podría confrontarse con $t^{\alpha}_{\texttt{n}-q-2}$ para una análisis individual o, mejor, con $t^{\alpha/\texttt{n}}_{\texttt{n}-q-2}$ para un análisis global. Para una análisis menos conservador utilizan otras cotas[25].

De (4.47) se sigue inmediatamente que

$$\texttt{Dfadj}(i) = t_i \sqrt{\frac{\nu_{ii}}{1-\nu_{ii}}}, \quad i = 1, \ldots, \texttt{n}. \tag{4.60}$$

Esta expresión permite interpretar $\texttt{Dfadj}(i)$ en los mismos términos que D_i^2, pero en términos del residuo estudentizado. De hecho, se puede establecer fácilmente la siguiente equivalencia:

$$D_i^2 = \big(\texttt{Dfadj}(i)\big)^2 \left(\frac{\hat{\sigma}^{2,\text{I}}(i)}{(q+1)\hat{\sigma}^{2,\text{I}}}\right). \tag{4.61}$$

4. **Covratios:** este método difiere sustancialmente de los tres anteriores. Se basa en e hecho de que, para cada $i = 1, \ldots, \texttt{n}$,

$$\texttt{Cov}\big[\hat{\beta}\big] = \sigma^2[\mathbf{X}'\mathbf{X}]^{-1}, \qquad \texttt{Cov}\big[\hat{\beta}(i)\big] = \sigma^2[\mathbf{X}(i)'\mathbf{X}(i)]^{-1}.$$

Se considera entonces el estadístico

$$\texttt{Covratio}(i) = \frac{|\hat{\sigma}^{2,\text{I}}(i)[\mathbf{X}(i)'\mathbf{X}(i)]^{-1}|}{\left|\hat{\sigma}^{2,\text{I}^2}[\mathbf{X}'\mathbf{X}]^{-1}\right|}.$$

Un valor distante de 1 se considera pues como signo de influencia de la unidad i-ésima.

Para todos los estadísticos introducidos podemos establecer otras cotas convencionales para determinar la influencia de una determinada unidad, al margen de las ya comentadas en los tres primeros métodos. El lector puede encontrarlas en Rawlings et al. (1998).

[24] Estamos hablando de la predicción en Y_i que se obtendría sin la participación de la unidad i-ésima en el modelo, es decir, partir de $\texttt{n}-1$ unidades experimentales.

[25] Ver Rawlings et al. (1998), pag. 363.

4.7. Multicolinealidad

Para acabar este capítulo abordamos el estudio de una situación que, aunque no puede considerarse una violación de los supuestos, puede acarrear muy serios problemas a la hora de extraer conclusiones. Nos referimos al problema de multicolinealidad, que se presenta cuando existe un alto grado de correlación lineal entre los vectores explicativos, lo cual puede implicar una elevada varianza en los estimadores de los respectivos coeficientes de regresión o una importante correlación entre los mismos. El hecho de que los estimadores presenten una elevada varianza puede considerarse negativo, al menos en principio, dado que resta fiabilidad a las estimaciones obtenidas. Lo mismo puede decirse de la correlación entre los estimadores, pues sería interesante que los distintos coeficientes se estimaran de forma totalmente independiente. No obstante, esto es bastante relativo, como ya veremos. La situación objetivamente indeseable se produce cuando estas circunstancias inducen a cometer importantes errores a la hora de determinar el grado de influencia de las variables explicativas en la variable respuesta.

El problema de multicolinealidad en regresión se trata tanto aquí como en el volumen dedicado al análisis multivariante. Quizás allí se puede abordar con mayor propiedad pues se suele hacer uso de las componentes principales para intentar solucionar el problema. En este caso haremos especial hincapié en las repercusiones de la multicolinealidad en los resultados de la regresión.

En todo momento hemos supuesto que la matriz $\mathbf{X}$ es de rango completo, es decir, que todas sus columnas son linealmente independientes. De no ser así, el parámetro β no quedaría unívocamente determinado, pues existirían infinitas soluciones a la ecuación $\mathtt{E}[Y] = \mathbf{X}\beta$. De hecho, el estimador propuesto para el mismo no podría calcularse pues la matriz $\mathbf{X}'\mathbf{X}$ no sería invertible. En tal caso se dice que estamos ante un Modelo Lineal de Rango no Completo. Este modelo se estudiará más adelante. Excluyendo esta situación, el problema se da cuando las columnas de $\mathbf{X}$ están próximas a la dependencia lineal, aunque ésta no se dé. Efectivamente, sabemos que la matriz de varianzas-covarianzas del estimador de β es la siguiente:

$$\mathtt{Cov}\big[\hat{\beta}\big] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Podemos observar cómo la misma se explica, por una parte, por la propia varianza del modelo, σ^2, y por otra, por la estructura de la matriz $\mathbf{X}$. Desde luego, si $\mathbf{X}$ fuera de rango no completo, el determinante de $\mathbf{X}'\mathbf{X}$ sería nulo. Razonando por continuidad, cuando más se aproximen las columnas de $\mathbf{X}$ a la situación de dependencia lineal, más se aproximará a 0 el determinante de la matriz $\mathbf{X}'\mathbf{X}$, lo cual implicará la existencia de valores muy altos en su inversa. No obstante, podemos ser mucho más explícitos

si consideramos la descomposición de β en β_0 y $\underline{\beta}$. Efectivamente, dado que

$$\underline{\hat{\beta}} = (\mathsf{Z}_0'\mathsf{Z}_0)^{-1}\mathsf{Z}_0'Y, \quad \hat{\beta}_0 = \overline{y} - \overline{\mathbf{z}}'\underline{\hat{\beta}},$$

se verifica

$$\mathtt{Cov}\left[\underline{\hat{\beta}}\right] = \frac{\sigma^2}{\mathtt{n}} S_{\mathtt{zz}}^{-1}, \quad \mathtt{var}\left[\hat{\beta}_0\right] = \sigma^2\left[\frac{1}{\mathtt{n}} + \frac{1}{\mathtt{n}} d^2(\overline{\mathtt{z}}, 0)\right]. \tag{4.62}$$

Luego, en primer lugar, hemos probado que $\psi_{00} = \mathtt{n}^{-1}[1 + d^2(\overline{\mathtt{z}}, 0)]$. Respecto a los demás coeficientes, se sigue del lema 9.7 que, para cada $j = 1, \ldots, q$, si Z_j denota la matriz Z despojada de la columna j-ésima, entonces

$$\mathtt{var}\big[\hat{\beta}_j\big] = \frac{\sigma^2}{\mathtt{n}}\left(s^2_{\mathtt{z}[j]} - S_{\mathtt{z}[j]\mathsf{Z}_j} S^{-1}_{\mathsf{Z}_j\mathsf{Z}_j} S_{\mathsf{Z}_j\mathtt{z}[j]}\right)^{-1}. \tag{4.63}$$

Teniendo en cuenta (4.16), resulta

$$\mathtt{var}\big[\hat{\beta}_j\big] = \sigma^2 \cdot \frac{1}{\mathtt{n}} \cdot \frac{1}{1 - R_j^2} \cdot \frac{1}{s^2_{\mathtt{z}[j]}}, \quad j = 1, \ldots, q, \tag{4.64}$$

donde R_j^2 denota abreviadamente el coeficiente de correlación múltiple de $\mathtt{z}[j]$ respecto al resto de vectores explicativos. Con esto queda demostrado que

$$\psi_{jj} = [\mathtt{n}(1 - R_j^2)s^2_{\mathtt{z}[j]}]^{-1}, \quad j = 1, \ldots, q. \tag{4.65}$$

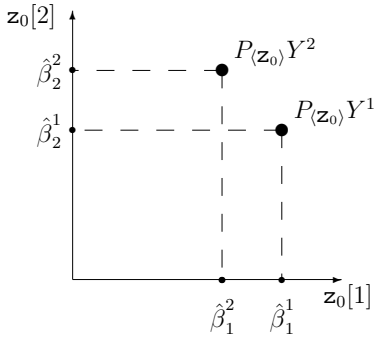
Una elevada varianza del estimador $\hat{\beta}_j$ puede conducir, según un análisis inicial, no demasiado reflexivo, a un resultado no significativo en el contraste parcial para β_j y, por lo tanto, a la eliminación de dicha variable. De hecho, uno de los más claros síntomas de multicolinealidad es la presencia de muchos resultados no significativos en los tests parciales. Esa apreciación no es errónea, aunque está sujeta a ciertas matizaciones, como veremos a continuación.

Para un análisis más exhaustivo, debemos estudiar detalladamente los distintos factores que intervienen en la expresión (4.64). En primer lugar, lógicamente, la propia varianza del modelo, σ^2; en segundo lugar, el tamaño de la muestra: cuanto mayor sea, menor será la varianza del estimador. No estamos afirmando que la varianza asintótica sea necesariamente nula, cosa que ocurre cuando ψ_{jj} converge a 0. Precisamente, que esto se verifique para todo $j = 0, 1, \ldots, q$, equivale a la proposición (3.32), que garantiza la consistencia del estimador de β.

El tercer factor en la ecuación (4.64) depende del grado de correlación lineal que $\mathtt{z}[j]$ presenta respecto al resto de vectores explicativos: cuanto más multicolinealidad exista, mayor será la varianza. De hecho, el término $(1 - R_j^2)^{-1}$ se denomina Factor de Inflación de la Varianza j-ésimo, abreviadamente FIV_j. Hemos de tener en cuenta

que el resultado de un test parcial depende únicamente del valor del coeficiente de correlación parcial entre la variable respuesta y la variable explicativa en cuestión, dadas el resto de variables explicativas. Que dicha variable pueda explicarse linealmente por las demás suele venir acompañado (aunque no necesariamente, según se ve en el ejercicio 3) con un bajo valor del coeficiente de correlación parcial.

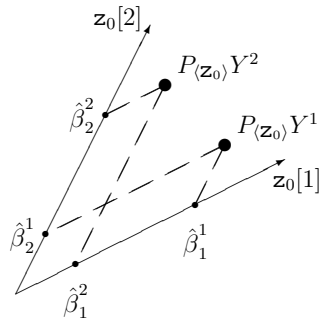
La visión geométrica puede ser fundamental en este caso. Los siguientes gráficos ilustran cómo una elevada correlación lineal entre dos vectores explicativos da lugar a una inflación en las varianzas de los estimadores. Supondremos que $q = 2$ y se denotarán por $\mathbf{z}_0[1]$ y $\mathbf{z}_0[2]$ las columnas primera y segunda de $\mathbf{Z}_0$, respectivamente. En la primera ilustración se presentan vectores explicativos incorrelados, lo cual equivale a que $\mathbf{z}_0[1]$ y $\mathbf{z}_0[2]$ sean perpendiculares. En este caso, los factores de inflación de la varianza son nulos.



Para una observación Y^1 de la variable respuesta (es decir, un vector $\mathtt{n}$-dimensional), obtenemos una proyección sobre el plano $\langle \mathbf{Z}_0 \rangle$, que podrá expresarse como una única combinación lineal de $\mathbf{z}_0[1]$ y $\mathbf{z}_0[2]$. Los coeficientes de dicha combinación serán las estimaciones de β_1 y β_2 para la observación Y^1. No obstante, las observaciones están sometidas a cierta variabilidad dado que son aleatorias. La magnitud de dicha variabilidad está determinada por el parámetro σ^2. Por lo tanto, una nueva ejecución del experimento proporcionará otra observación Y^2 cuya proporción sobre el plano $\langle \mathbf{Z}_0 \rangle$ será diferente, luego, diferentes serán también las estimaciones de los coeficientes. Podemos observar, no obstante, que una pequeña diferencia entre las observaciones se traduce en una pequeña diferencia entre las estimaciones. Ésta es la situación ideal, pues no se produce una inflación de la varianza debida a la correlación entre los vectores explicativos.

La situación contraria se ilustra en el siguiente diagrama. Hemos de tener en

cuenta que una alta correlación entre las variables explicativas se representa mediante dos vectores, $\mathtt{z}_0[1]$ y $\mathtt{z}_0[2]$ próximos a la dependencia lineal.



En este caso observamos cómo la misma variación en las observaciones produce una diferencia mucho mayor entre las estimaciones de los coeficientes. En esto consiste la inflación de la varianza. Las consecuencias de la misma pueden ser bastante graves en lo que se refiere a la optimización del modelo. Efectivamente, según la primera observación, sería $\mathtt{z}[1]$ la variable con mayor *peso* en la explicación de la variable respuesta, mientras que, según la segunda observación, la situación sería la contraria. Esto puede verse reflejado en los tests parciales, de forma que se considere no significativo un coeficiente (lo cual puede conllevar la eliminación de la correspondiente variable) que, con otra observación muy similar, sí lo sería. Esta especie de *discontinuidad* en la decisión no parece ser admisible desde el punto de vista de la Inferencia Estadística. Respecto a la covarianza entre los estimadores de β_1 y β_2, se sigue trivialmente de (4.12)

$$\mathtt{cov}[\hat{\beta}_1, \hat{\beta}_2] = -\frac{\sigma^2}{\mathtt{n}} \cdot \frac{1}{s_{\mathtt{z}[1]} \cdot s_{\mathtt{z}[2]}} \cdot \frac{r_{\mathtt{z}[1],\mathtt{z}[2]}}{1 - r^2_{\mathtt{z}[1],\mathtt{z}[2]}} \tag{4.66}$$

En consecuencia, si la correlación entre $\mathtt{z}[1]$ y $\mathtt{z}[2]$ es positiva y los signos de β_1 y β_2 coinciden, o bien si la correlación es negativa y los signos difieren, se verifica que, por término medio, un aumento en el valor absoluto de la estimación de β_1 va acompañado de una disminución en el valor absoluto de la de β_2, y, a efectos de los test parciales (ver (4.33)), esto es lo más importante a la hora de excluir una variable del modelo. Esa puede ser la situación que se da en la segunda ilustración. Por lo tanto, en esas condiciones, una sobrevaloración de una de las variables explicativas va acompañada de una minusvaloración de la otra. Que esta situación, con repercusiones muy negativas en el análisis, se dé o no, depende, insistimos, de la relación entre los signos de los coeficientes β_1 y β_2. Al ser éstos parámetros del modelo, hablar de la

probabilidad de que se dé esta circunstancia problemática sólo tiene sentido desde una perspectiva Bayesiana.

En el caso general, cuando tengamos q vectores explicativos, podemos obtener, a partir del lema 9.7, una expresión análoga en términos de las varianzas y coeficientes de correlación parciales. Concretamente, si consideramos i y j entre 1 y q y distintos entre sí, y se denota por Z_R la matriz Z desprovista de las columnas i-ésima y j-ésima, se verifica:

$$\mathtt{cov}\big[\hat{\beta}_i, \hat{\beta}_j\big] = -\frac{\sigma^2}{\mathtt{n}} \cdot \frac{1}{s_{\mathbf{z}[1]\bullet\mathbf{z}_R} \cdot s_{\mathbf{z}[2]\bullet\mathbf{z}_R}} \cdot \frac{r_{\mathbf{z}[1],\mathbf{z}[2]\bullet\mathbf{z}_R}}{1 - r^2_{\mathbf{z}[1],\mathbf{z}[2]\bullet\mathbf{z}_R}} \tag{4.67}$$

La interpretación es, por lo tanto, similar. En definitiva, el problema de multicolinealidad puede llevar a una situación en la cual el propio azar tenga demasiado peso a la hora escoger unas variables en detrimento de las otras.

Por último, se sigue de (4.64) que la varianza de $\hat{\beta}_j$ es inversamente proporcional a la varianza muestral de $\mathbf{z}[j]$. Ello se explica sencillamente por el hecho de que la varianza de $\mathbf{z}[j]$ coincide con el cuadrado de la longitud del vector $\mathbf{z}_0[j]$. Si éste es pequeño, lo coeficientes correspondientes serán grandes y su varianza también. De hecho, si, por ejemplo, $\mathbf{z}[j]$ expresa la medición en centímetros de cierta longitud, expresar los valores en metros equivale a dividir por cien la longitud de $\mathbf{z}_0[j]$ y, por lo tanto, a multiplicar por cien el estimador de su coeficiente. En particular, multiplicamos por cien su desviación típica. Este hecho no puede tener influencia en los contrastes parciales pues no suponen cambio alguno en los subespacios V del modelo ni W de la hipótesis inicial. Simplemente, estaremos manejando valores más elevados con varianzas más elevadas pero, en términos relativos, el grado de dispersión es el mismo. Hay que tener en cuenta que, el que la varianza muestral de $\mathbf{z}[j]$ sea próxima a 0, equivale a que el vector sea *casi* proporcional al término independiente $1_{\mathtt{n}}$, lo cual debe repercutir negativamente en la varianza del estimador y, por lo tanto, en la fiabilidad de la estimación. Pero que esta circunstancia tenga trascendencia real en el análisis de los resultados es discutible, al menos en lo que a los contrastes parciales se refiere. De hecho, basta tipificar los vectores explicativos para que este factor quede eliminado.

En conclusión, hemos analizado en qué sentido la multicolinealidad entre los vectores explicativos puede entorpecer la valoración de la importancia de las mismas a la hora de explicar la variable respuesta. Aunque no es éste el único problema que ocasiona, es posiblemente el más relevante pues afecta enormemente a la optimización del modelo. Existen diversas formas de detectar la multicolinealidad. Una de las más extendidas consiste en analizar los FIV's. Muchos autores consideran la presencia de algún FIV mayor que 10 como signo de un problema de multicolinealidad; también pueden analizarse los denominados Índices de Condicionamiento para detectar auto-

valores próximos a cero en $\mathtt{X}'\mathtt{X}$ (lo cual se corresponde con una situación próxima al rango no completo) y las matrices de Proporción de la Varianza[26]. La propia matriz de correlaciones de los vectores explicativos, $R_{\mathtt{Z}}$, o los gráficos de dispersión aportan una información muy valiosa. No obstante, en muchas ocasiones los resultados de los tests parciales pueden constituir signos claros de un problema de multicolinealidad.

Una vez diagnosticado el problema, la siguiente cuestión es cómo intentamos resolverlo. Recordemos que, realmente, lo que se exige de un estimador es que la matriz (9.42), conocida como error cuadrático medio, sea lo menor posible. Un estimador óptimo en ese sentido no puede encontrarse en la mayoría de los casos, por lo que es costumbre imponer la condición razonable de que el estimador sea insesgado y buscar entonces el que minimice el error cuadrático medio. En ese caso, se trata simplemente de minimizar la varianza, por lo que el estimador óptimo, si existe, se denomina insesgado de mínima varianza. Ese el el caso, como ya sabemos, del estimador de β Pero hemos de tener presente que se ha impuesto una condición muy restrictiva: que el estimador sea insesgado. Si el EIMV presenta una matriz de varianzas-covarianzas con valores elevados, como sucede cuando existe un problema de multicolinealidad, podemos buscar un estimador sesgado aunque con menor varianza, de manera que el error cuadrático medio disminuya sustancialmente. Eso es lo que se denomina una regresión sesgada.

Existen diversos métodos de estimación sesgada. Por ejemplo, en Arnold (1981) se estudia el denominado estimador *Ridge*, propuesto en Hoerl y Kennard (1970)

$$\hat{\beta}_{(k)} = (\mathtt{X}'\mathtt{X} + k\mathtt{Id})^{-1}\mathtt{X}'Y,$$

siendo k un número positivo seleccionada para minimizar el error cuadrático medio. Este procedimiento tiene una clara justificación teórica desde un punto de vista Bayesiano. No obstante, analizaremos con algo más de detenimiento otro método de estimación sesgada basado en el Análisis de Componentes Principales[27]. Antes de aplicar un técnica de este tipo es bastante común tipificar los vectores en juego, en este caso los explicativos, cosa que supondremos en lo que resta del capítulo. Por lo tanto, la matriz de covarianzas de $\mathtt{Z}$, $S_{\mathtt{Z}}$, coincidirá con la matriz de correlaciones, $R_{\mathtt{Z}}$.

El método en sí consiste en transformar los datos de manera que los factores de inflación de la varianza desaparezcan en favor de las varianzas de las vectores explicativos, que aumentan. Para ello, debemos encontrar una transformación en

[26] Hair et al. (1999).

[27] La descripción de esa técnica multivariante puede encontrarse, por ejemplo, en Rencher (1995), o también en el volumen dedicado al Análisis Multivariante.

las variables explicativas (rotación) que las haga incorreladas, lo cual se consigue mediante la diagonalización de la matriz de covarianzas según el teorema 9.4

$$S_{\mathtt{Z}} = \Gamma \Delta \Gamma',$$

donde Δ es la matriz diagonal de los autovalores ordenados de $S_{\mathtt{Z}}$, $\delta_1, \ldots, \delta_q$, y Γ es la matriz ortogonal cuyas columnas constituyen una base ortonormal de autovectores asociados, $g_1, \ldots, g_q$. A continuación, se proyectan los vectores $\mathtt{z}_i$ sobre los ejes determinados por los autovectores, de manera que se obtiene una nueva matriz explicativa

$$U = \mathtt{Z}\Gamma,$$

cuyas columnas, que se denotan por $u[1], \ldots, u[q]$, se denominan componentes principales. Esta transformación, consistente en aplicar una matriz ortogonal puede deshacerse mediante $\mathtt{Z} = U\Gamma'$. La ventaja que presentan las componentes principales es que son incorreladas, pues

$$S_U = \Delta.$$

Así pues, la regresión lineal respecto a $\mathtt{Z}$ puede convertirse en una reresión respecto a U si consideramos el parámetro $\gamma = \Gamma' \underline{\beta}$

$$\begin{aligned} Y &= \beta_0 1_{\mathtt{n}} + \mathtt{Z}\underline{\beta} + \mathcal{E} \\ &= \beta_0 1_{\mathtt{n}} + U\gamma + \mathcal{E}, \end{aligned}$$

donde $\mathcal{E}$ sigue un modelo de distribución $N_{\mathtt{n}}(0, \sigma^2 \mathtt{Id})$. El EIMV de γ es

$$\hat{\gamma} = (U'U)^{-1}U'Y = \Gamma' \hat{\underline{\beta}},$$

de manera que el estimador de $\underline{\beta}$ puede reconstruirse mediante

$$\hat{\underline{\beta}} = \hat{\gamma}. \tag{4.68}$$

Sin embargo,

$$\hat{\gamma} \sim N_q \left(\gamma, \frac{\sigma^2}{\mathtt{n}} \, \Delta^{-1} \right).$$

En consecuencia, los estimadores γ_j, $j = 1, \ldots, q$ son independientes, siendo su varianza

$$\mathtt{var}\left[\hat{\gamma}_j\right] = \frac{\sigma^2}{\mathtt{n}} \, \delta_j^{-1}. \tag{4.69}$$

Además, puede comprobarse que los estimadores $\hat{\gamma}_j$ coinciden con los que se obtendrían en cada caso con una regresión simple. Un diseño de este tipo, en el que los

vectores explicativos tienen media aritmética nula y son incorreladas, se denomina ortogonal. Podemos observar que la varianza del estimador es inversamente proporcional a la varianza de la correspondiente componente principal, sin que en este caso exista un factor de inflación de la varianza. Esto no debe inducirnos a pensar que hemos conseguido reducir la matriz de varianzas-covarianzas de los estimadores. De hecho, puede demostrarse fácilmente que, tanto la varianza generalizada[28] como la varianza total[29], permanecen invariantes cuando se consideran las componentes principales.

Consideremos una división de Δ en dos submatrices diagonales Δ_1 y Δ_2, lo cual induce una división análoga en la matriz Γ, en vector γ y en su estimador. De esta forma, se verifica

$$\mathtt{Cov}\left[\hat{\underline{\beta}}\right] = \frac{\sigma^2}{\mathtt{n}} (\Gamma_1 \Gamma_2) \begin{pmatrix} \Delta_1 & 0 \\ 0 & \Delta_2 \end{pmatrix}^{-1} \begin{pmatrix} \Gamma_1' \\ \Gamma_2' \end{pmatrix} \tag{4.70}$$

$$= \frac{\sigma^2}{\mathtt{n}} \Gamma_1 \Delta_1^{-1} \Gamma_1' + \frac{\sigma^2}{\mathtt{n}} \Gamma_2 \Delta_2^{-1} \Gamma_2'. \tag{4.71}$$

Además, $\hat{\underline{\beta}}$ descompone en

$$\hat{\underline{\beta}} = \Gamma_1 \hat{\gamma}_1 + \Gamma_2 \hat{\gamma}_2.$$

Si consideramos un nuevo estimador $\hat{\underline{\beta}}^*$ de $\underline{\beta}$ que se obtiene depreciando los coeficientes correspondientes a las componentes principales asociadas a Δ_2, es decir,

$$\hat{\underline{\beta}}^* = \Gamma_1 \hat{\gamma}_1,$$

se verificará lo siguiente:

$$\mathtt{Sesgo}\left[\hat{\underline{\beta}}^*\right] = \Gamma_2 \gamma_2, \qquad \mathtt{Cov}\left[\hat{\underline{\beta}}^*\right] = \mathtt{Cov}\left[\hat{\underline{\beta}}\right] - \frac{\sigma^2}{\mathtt{n}} \Gamma_2 \Delta_2^{-1} \Gamma_2'.$$

Así pues, si Δ_1 contiene los autovalores menores (que son las varianzas de las últimas componentes principales), al considerar este nuevo estimador de $\underline{\beta}$ conseguiremos una gran reducción en la matriz de varianzas-covarianzas. Por contra, el estimador obtenido será sesgado. Teniendo en cuenta (9.43), este procedimiento resulta rentable cuando el sesgo introducido es pequeño en relación con reducción en las varianzas, cosa que sucede cuando γ_2 es próximo a 0. Por lo tanto, la estrategia consiste en despreciar las componentes principales de menor varianza siempre y cuando su correspondiente coeficiente sea próximo a 0. Una decisión de este tipo puede basarse en

[28] Nos referimos al determinante de la matriz de varianza-covarianzas.

[29] Es decir, la suma de las varianzas de $\hat{\beta}_1, \ldots, \hat{\beta}_q$. o, lo que es lo mismo, la traza de la matriz de varianzas-covarianzas total.

los resultados de los test parciales. Mucho autores coinciden en considerar un nivel de significación mayor de lo habitual, por ejemplo 0.20, a la hora de aplicarlos. Por desgracia, no podemos garantizar que los tests parciales aporten resultados no significativos para las componentes principales de menor varianza, pero si esto sucede, cabrá confiar en una sustancial reducción de la matriz de varianzas-covarianzas y, por lo tanto, en una clara mejoría del análisis.

Queda pendiente una última cuestión. ¿En qué momento debe llevarse a acabo el diagnóstico de multicolinealidad? No estamos en condiciones de dar una respuesta clara pero parece razonable realizarlo una vez ajustados los datos a los supuestos del modelo, pues cualquier acción que emprendamos previa al análisis de los residuos puede quedar desbaratada después de aplicar transformaciones de las variables.

Cuestiones propuestas

1. Probar la igualdad (4.7).

2. Probar que $s^2_{\hat{Y}} = s^2_{P_{\langle \mathbf{Z}_0 \rangle} Y} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0 + \mathbf{z}_i\underline{\hat{\beta}} - \overline{y}\right)^2$.

3. Demostrar la siguiente igualdad

$$R^2_{y\bullet \mathbf{Z}[1],\ldots,\mathbf{Z}[q+1]} - R^2_{y\bullet \mathbf{Z}[1],\ldots,\mathbf{Z}[q]} = r^2_{y,\mathbf{Z}[q+1]\cdot\ \mathbf{Z}[1],\ldots,\mathbf{Z}[q]}\left(1 - R^2_{y\bullet \mathbf{Z}[1],\ldots,\mathbf{Z}[q]}\right).$$

   Demostrar la expresión equivalente en términos de los coeficientes probabilísticos.

4. Probar que el coeficiente de correlación múltiple puede obtenerse mediante

$$R^2_{y\bullet \mathbf{Z}[1],\ldots,\mathbf{Z}[q]} = \frac{\sum_{i=1}^{\mathtt{n}}\left(\hat{\beta}_0 + \hat{\beta}_1\mathbf{z}_i[1] + \ldots + \hat{\beta}_q\mathbf{z}_i[q] - \overline{y}\right)^2}{\sum_{i=1}^{\mathtt{n}}(Y_i - \overline{y})^2}$$

5. Probar que el coeficiente de correlación múltiple no puede disminuir al introducir un nuevo vector explicativo $\mathbf{z}[q+1]$, y que permanece constante si y sólo si el coeficiente de correlación parcial entre Y y $\mathbf{z}[q+1]$ dados $\mathbf{z}[j]$, $j = 1, \ldots, q$, es nulo.

6. Probar la igualdad (4.22).

7. Construir los intervalos de confianza (4.24) y (4.25). Estudiar el comportamiento asintótico de los mismos cuando se cumple la condición de Huber.

8. Considerar un Modelo de Regresión Lineal Múltiple con 4 vectores explicativos y `n` unidades experimentales. Construir el test `F` a nivel α para contrastar la hipótesis inicial

   a) $H_0^1 : \beta_1 = \beta_2$.

   b) $H_0^2 : \beta_1 + 2\beta_2 = 1$

   c) $H_0^3 : \begin{cases} \beta_1 = \beta_2 \\ \beta_3 = \beta_4 \end{cases}$

9. Obtener las expresiones (4.27), (4.28) y (4.30).

10. Obtener la expresión (4.31). Para ello es aconsejable expresar el estadístico de contraste del test `F` mediante (3.25).

11. Obtener la expresión (4.32).

12. Resolver el contraste $H_0 : \beta_i = b_i$, donde b_i es un valor real conocido.

13. Obtengamos una expresión análoga a (4.18) para el coeficiente de correlación parcial. Consideremos $Y_1, \mathtt{Z}_D \in \mathbb{R}^n$ y $\mathtt{Z} \in \mathcal{M}_{n\times q}$, y sea $\mathtt{X} = (1_n|\mathtt{Z}|\mathtt{Z}_D)$. Probar, teniendo en cuenta (9.63) y que $\langle \mathtt{X} \rangle$ descompone en la suma ortogonal $\langle 1_n \mathtt{Z} \rangle \oplus \langle \mathtt{Z}_D - P_{\langle 1_n \mathtt{Z} \rangle} \mathtt{Z}_D \rangle$, que
$$r^2_{Y,\mathtt{Z}_D \cdot \mathtt{Z}} = \frac{\|P_{\langle \mathtt{X} \rangle | \langle 1_n \mathtt{Z} \rangle} Y\|^2}{\|P_{\langle 1_n \mathtt{Z} \rangle^\perp} Y\|^2}.$$
   Nótese que, desde este punto de vista, el coeficiente de correlación simple puede entenderse como un caso particular del coeficiente de correlación parcial dado $\mathtt{Z} = 0$.

14. Obtener la expresión (4.35) para el estadístico de contraste de un test parcial.

15. Probar, teniendo en cuenta (4.31), que en el método de selección hacia adelante, la variable que aporta el resultado más significativo en el contraste parcial coincide con la que aporta el resultado más significativo en el contraste total.

16. Obtener el estadístico de contraste (4.42), correspondiente al test de linealidad.

17. Demostrar que $\overline{\mathtt{e}} = 0$ y $s^2_{\mathtt{e}} = \hat{\sigma}^{2,\mathtt{MV}}$.

18. Demostrar que, en el caso de la regresión lineal simple, $\mathtt{var}[\mathtt{e}_i] = 0$ equivale a que todos los vectores predictivos salvo $\mathtt{z}_i$ sean idénticos.

19. Demostrar que el método de Box-Cox para obtener una muestra aleatoria simple de una distribución normal consiste en encontrar el valor de λ que minimice $s^2_{Y^{(\lambda)}}$.

20. Probar que en una regresión polinómica simple se mantiene en todo caso el rango completo.

21. Obtener (4.60) y (4.61).

22. Obtener (4.62), (4.63) y (4.64).

23. ¿Qué semajanzas se dan entre la varianza de $\hat{\beta}_0$ y la de la predicción en $\mathbf{z}_i$? ¿Cómo puede interpretarse este hecho?

24. Obtener las covarianzas (4.66) y (4.67). Interpretar los resultados.

25. Probar que un diseño ortogonal, el estimador de β_j, $j = 1, \ldots, q$, coincide con el que se obtendría con una regresión simple respecto a la variable $\mathbf{z}[j]$.

26. ¿Por qué la tercera estrategia de transformación de variables propuesta es sólo válida en problemas de regresión pura? ¿Por qué la primera es válida únicamente en problemas de correlación?

27. ¿Por qué en las figura 27 y 28 la recta de regresión ha de pasar necesariamente por el dato extremo?

28. ¿Es cierto que la incorrelación entre dos variables implica la incorrelación parcial entre las mismas dada una tercera? En otras palabras: sean tres variables (vectores $\mathtt{n}$-dimensionales) x, y, z, tales que $r_{y,z} = 0$, ¿debe verificarse $r_{y,z\cdot x}$=0? Si es así demuétrese. En caso contrario presentar un contraejemplo mediante un programa estadístico.

29. Probar que, si $\mathbf{z}[1], \mathbf{z}[2], \mathbf{z}[3]$ son icorreladas, entonces

$$r^2_{y,\mathbf{z}[1]\cdot\mathbf{z}[2],\mathbf{z}[3]} \geq r^2_{y,\mathbf{z}[1]}$$

    ¿En qué condiciones se alcanzaría la igualdad?

30. ¿Puede cambiar el coeficiente de correlación múltiple cuando se lleva a cabo una regresión por componentes principales?

Capítulo 5

El Modelo de Correlación

Este capítulo, de carácter fundamentalmente teórico, viene a complementar los capítulos 3 y 4. La diferencia entre este capítulo y el anterior estriba únicamente en el hecho de que, mientras que en el modelo de Regresión los valores explicativos se consideran fijos, aquí se suponen observaciones correspondientes a variables aleatorias. No obstante, el objetivo principal que nos marcamos es dejar claro que, desde un punto de vista práctico, esta distinción no afecta sustancialmente a los fundamentales problemas de Estimación y Contraste de Hipótesis. Además, hemos de advertir que este estudio puede ser enfocado de forma más elegante desde el punto de vista del Análisis Multivariante. En todo caso, la clave del modelo podemos hallarla en el hecho conocido de que, entre las distintas componentes de un vector normal multivariante sólo cabe una relación de tipo lineal.

5.1. El Modelo

En este caso, consideraremos los $q+1$ vectores aleatorios siguientes:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{\mathtt{n}} \end{pmatrix}, \quad z[1] = \begin{pmatrix} z_1[1] \\ \vdots \\ z_{\mathtt{n}}[1] \end{pmatrix} \quad \dots \quad z[q] = \begin{pmatrix} z_1[q] \\ \vdots \\ z_{\mathtt{n}}[q] \end{pmatrix}$$

Se denotarán por Z y X, respectivamente, las matrices

$$\begin{pmatrix} z_1[1] & \dots & z_1[q] \\ \vdots & & \vdots \\ z_{\mathtt{n}}[1] & \dots & z_{\mathtt{n}}[q] \end{pmatrix}, \qquad X = (1_{\mathtt{n}}|Z).$$

Para cada $i = 1, \ldots, \mathtt{n}$ se denotará mediante Z_i el vector que se obtiene transponiendo la fila i-ésima de Z. De esta forma, Y_i denota el resultado de la variable respuesta para la i-ésima unidad experimental, mientras las componentes que Z_i serán los resultados de las variables explicativas para dicha unidad. Para cada $j = 1, \ldots, q$, los términos $\overline{z}[j]$ denotarán, respectivamente, las medias aritméticas de las $\mathtt{n}$ observaciones correspondientes a las variables $z[j]$. Igualmente, $\overline{y}$ denotará la media de y; $\overline{z}$ será el vector compuesto por las q medias $\overline{z}[j]$, $j = 1, \ldots, q$. El término M denotará la matriz de datos $(Y|Z)$. En ese caso, S_M denotará la matriz de varianzas-covarianzas total muestral $S_{(YZ)(YZ)}$, definida según (9.54). Ésta descompone de la siguiente forma

$$S_M = \begin{pmatrix} s_Y^2 & S_{YZ} \\ S_{ZY} & S_{ZZ} \end{pmatrix}.$$

El vector $\overline{\mathtt{m}}$ será igual a $(\overline{y}, \overline{z}')'$. Se supondrá en todo momento que el número de unidades experimentales, $\mathtt{n}$, es estrictamente mayor que el número de variables explicativas, q.

En definitiva, el Modelo de Correlación consiste en suponer la normalidad multivariante de la distribución conjunta, es decir, se considera

$$\begin{pmatrix} Y_1 \\ Z_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_{\mathtt{n}} \\ Z_{\mathtt{n}} \end{pmatrix} \text{ iid } N_{q+1}(\nu, \Xi), \quad \nu \in \mathbb{R}^{q+1}, \ \Xi > 0. \tag{5.1}$$

Hemos de tener en cuenta que, en este modelo, no existe diferencia formal entre la variable respuesta y las explicativas. De hecho, la variable y puede desempeñar el papel de explicativa, si lo deseamos. Además, si y eliminamos cualquiera de las variables en juego, tendremos un modelo de correlación con q variables, siempre que q sea mayor que 1. En caso contrario, nos quedaremos con un modelo lineal normal con un subespacio V unidimensional.

Si descomponemos ν y Ξ de acuerdo con la división entre variable respuesta y explicativas, podemos obtener, de manera manera análoga a (9.25) y (9.14), los siguientes parámetros:

$$\sigma^2 = \Xi_{11\cdot 2}, \quad \underline{\beta} = \Xi_{22}^{-1}\Xi_{21}, \quad \beta_0 = \nu_1 - \nu_2'\underline{\beta}, \quad \mu_Z = \nu_2, \quad \Sigma_{ZZ} = \Xi_{22}. \tag{5.2}$$

Estos términos, definidos a partir de ν y Ξ, pueden parametrizar el modelo (5.1), puesto que la transformación anterior es invertible. Efectivamente, podemos recons-

truir ν y Ξ mediante

$$\begin{aligned}
\Xi_{22} &= \Sigma_{ZZ}, && (5.3)\\
\Xi_{21} &= \Sigma_{ZZ}\underline{\beta}, && (5.4)\\
\Xi_{11} &= \sigma^2 + \underline{\beta}'\Sigma_{ZZ}\underline{\beta}, && (5.5)\\
\nu_2 &= \mu_Z && (5.6)\\
\nu_1 &= \beta_0 + \mu_Z'\underline{\beta}. && (5.7)
\end{aligned}$$

De esta forma, si se denota $\beta = (\beta_0, \underline{\beta}')'$, se verifica el siguiente resultado.

Teorema 5.1.
El Modelo de Correlación Lineal (5.1) puede expresarse de manera equivalente mediante

$$Y|Z = \mathtt{Z} \sim N_{\mathtt{n}}(\mathtt{X}\beta, \sigma^2\mathtt{Id}),\ Z_1, \ldots, Z_{\mathtt{n}} \text{ iid } N_q(\mu_Z, \Sigma_{ZZ}), \tag{5.8}$$

donde $\beta \in \mathbb{R}^{q+1}$, $\sigma^2 > 0$, $\mu_Z \in \mathbb{R}^q$ y $\Sigma_{ZZ} > 0$, y siendo $\mathtt{X} = (1_{\mathtt{n}}|\mathtt{Z})$. Además, sea cual sea la distribución de la familia considerada, se verifica que $\mathtt{rg}(\mathtt{X}) = q + 1$ con probabilidad 1.

Demostración.
La primera parte de la demostración se basa en el hecho de que la distribución conjunta de dos vectores aleatorios puede construirse como el producto generalizado entre la distribución marginal del segundo y la distribución condicional del primero dado el segundo. Concretamente y teniendo en cuenta que los vectores (Y_i, Z_i), $i = 1, \ldots, \mathtt{n}$, son independientes por hipótesis, junto con las proposiciones 2.1, 2.5, se sigue que

$$\begin{aligned}
P^{(Y,Z)} &= P^{Y|Z=\mathtt{Z}} \times P^Z = \left[\prod_{i=1}^{\mathtt{n}} P^{Y_i|Z_i=\mathtt{Z}_i}\right] \times \left[\prod_{i=1}^{\mathtt{n}} P^{Z_i}\right]\\
&= \left[\prod_{i=1}^{\mathtt{n}} \left[N(\beta_0 + \mathtt{Z}_i'\underline{\beta}, \sigma^2)\right] \times \left[\prod_{i=1}^{\mathtt{n}} N_q(\mu_Z, \Sigma_{ZZ})\right]\right.\\
&= N_{\mathtt{n}}(\mathtt{X}\beta, \sigma^2) \times \left[\prod_{i=1}^{\mathtt{n}} N_q(\mu_Z, \Sigma_{ZZ})\right],
\end{aligned}$$

donde el signo $\times$ denota el producto generalizado. Teniendo en cuenta que la transformación que permite obtener β, σ^2, μ_z y Σ_{ZZ} a partir de ν y Ξ es biunívoca, queda probada la equivalencia entre ambos modelos. La segunda parte del teorema se demostrará por inducción sobre q y teniendo en cuenta que la medida de Lebesgue en

$\mathbb{R}^{\mathtt{n}}$ de cualquier hiperplano del mismo es nula. En particular, será nula la probabilidad de un hiperplano si ésta está dominada por la medida de Lebesgue. De esta forma, si $q = 1$, se verifica que $\mathtt{rg}(\mathbf{X}) < 2$ si, y sólo si, $z[1]$ pertenece al subespacio $\langle 1_{\mathtt{n}} \rangle$, que es, a los sumo, un hiperplano de $\mathbb{R}^{\mathtt{n}}$ (recordar que estamos suponiendo, por hipótesis, que $\mathtt{n} > q$). Por lo tanto, la tesis queda probada cuando $q = 1$.

Supongámosla cierta para un cierto $q-1$ y veamos que lo es también para q. En ese caso, que $\mathtt{rg}(\mathbf{X})$ sea menor que $q+1$ equivale a que $z[q]$ pertenezca al subespacio generado por el vector $1_{\mathtt{n}}$ junto con los vectores aleatorios $z[j]$, $j = 1, \ldots, q-1$, que será, a lo sumo, un hiperplano. La distribución de $z[q]$ condicionada a la matriz aleatoria[1] constituida por los vectores aleatorios $z[j]$, $j = 1, \ldots, q-1$, es el producto de las respectivas distribuciones de $z_i[q]$ condicionadas a $(z_i[1], \ldots, z_i[q-1])$, $i = 1, \ldots, \mathtt{n}$. Aplicando en cada caso la proposición 2.5 y componiendo las distribuciones obtenidas, se obtiene que $z[q]$ condicionada a la matriz aleatoria $z[i]$, $i = 1, \ldots, q-1$, sigue un modelo de distribución $\mathtt{n}$-normal no degenerado y, por lo tanto, dominado por la medida de Lebesgue en $\mathbb{R}^{\mathtt{n}}$. Luego, fijos $\mathtt{z}[j]$, $j = 1, \ldots, q-1$, la probabilidad de que $z[q]$ pertenezca al subespacio $\langle 1_{\mathtt{n}}, \mathtt{z}[1], \ldots, \mathtt{z}[q-1] \rangle$ es nula. Aplicando (9.30) con $f = \mathtt{rg}(X)$ concluimos.

∎

En definitiva, dado un modelo de Correlación, al condicionar sobre las variables explicativas[2], es decir, cuando se consideran fijos los valores de éstas, se obtiene automáticamente un modelo de Regresión. Recíprocamente, si se añade el supuesto de q-normalidad de las variables explicativas, se recompone el modelo de Correlación. Un modelo más débil que el de Correlación se obtendría eliminando en (5.8) la hipótesis de normalidad, tanto de la distribución marginal de las variables explicativas como e la condicional para la respuesta, pero suponiendo que $Z_1, \ldots, Z_{\mathtt{n}}$ constituye una muestra aleatoria simple de una distribución dominada por la medida de Lebesgue en $\mathbb{R}^q$. Un modelo de ese tipo se considerará cuando se afronte el estudio asintótico. En ese caso, teniendo en cuenta (4.58) y aplicando un razonamiento análogo al de la demostración anterior, se deduciría también que $\mathtt{rg}(X) = q+1$, con probabilidad 1 (se deja como ejercicio). Por otra parte, si aplicamos nuevamente (9.30), se tiene que el modelo de Correlación puede expresarse también mediante $Y = (1_{\mathtt{n}}|Z)\beta + \mathcal{E}$, con $\mathcal{E} \sim N_{\mathtt{n}}(0, \sigma^2 \mathtt{Id})$ y $Z_1, \ldots, Z_{\mathtt{n}}$ una muestra aleatoria simple de $N_q(\mu_Z, \Sigma_{ZZ})$ independiente de $\mathcal{E}$.

[1] Realmente, una matriz puede entenderse como un vector dispuesto de una forma determinada, por lo que no es estrictamente necesario definir matriz aleatoria. No obstante, éste concepto y en particular el de normal matricial, se estudian en Arnold (1981), lo cual permite obtener de forma elegante diversos resultados propios del Análisis Multivariante.

[2] Obtenemos entonces lo que daremos en denominar modelo condicionado

Llegados a este punto, hemos de notar que cualquier estadístico $\overline{T}$ definido en el modelo de Regresión, que será de la forma $\overline{T}(Y)$, puesto que $\mathtt{Z}$ se considera constante, puede considerarse definido en el modelo de Correlación mediante $T(Y,Z)$, si consideramos Z variable. De esta forma, si $\overline{T}$ constituye un estimador de cierto estimando $\overline{\tau}$, T puede considerarse estimador del estimando τ, definido sobre el modelo de Correlación. Teniendo en cuenta el teorema anterior junto con (9.30), se sigue que la distribución de $\overline{T}$ en el modelo de Regresión coincide con la distribución condicional de T dada $Z=\mathtt{Z}$ en el de correlación. Efectivamente:

$$P^{T(Y,Z)|Z=\mathtt{Z}} = \left[P^{Y|Z=\mathtt{Z}}\right]^{T(\cdot,\mathtt{Z})} = \left[N_{\mathtt{n}}\big((1_{\mathtt{n}}|\mathtt{Z})\beta,\sigma^2\mathtt{Id}\big)\right]^{T(\cdot,\mathtt{Z})}.$$

No obstante, si dicha distribución no depende del valor $\mathtt{Z}$ considerado, entonces T y Z serán independientes y la distribución condicional coincidirá con la distribución marginal de T y, por lo tanto, con la de $\overline{T}$. Tal es el caso del estimador de la varianza (4.9), po lo que éste es insesgado en el modelo de Correlación y el intervalo de confianza para la misma construido en (3.13) sigue siendo válido. Respecto al estimador natural de β no puede decirse lo mismo, puesto que su distribución depende de $\mathtt{Z}$. Sin embargo, si β es el verdadero valor del parámetro, obtenemos las siguientes distribuciones marginales:

$$\hat{\beta} \sim N_{q+1}\big(\beta,\sigma^2(X'X)^{-1}\big), \qquad \frac{\left(\hat{\beta}-\beta\right)' X'X \left(\hat{\beta}-\beta\right)}{\hat{\sigma}^{2,\mathtt{I}}} \sim F_{q+1,\mathtt{n}-(q+1)}$$

Por lo tanto, el EIMV de β en el modelo de Regresión es insesgado en el de Correlación, y el elipsoide (3.12) sigue siendo una región de confianza a nivel $1-\alpha$. Mediante una razonamiento análogo podemos demostrar la validez de los intervalos de confianza (4.10), (4.24) y (4.25).

Además, la distribución nula del estadístico F correspondiente al test $\mathtt{F}$ a nivel α para contrastar cualquier hipótesis del tipo $H_0 : A\beta = 0$, es una F-Snedecor central que no depende en ningún caso del valor de $\mathtt{Z}$. Por lo tanto, el test $\mathtt{F}$ a nivel α es también válido desde el punto de vista del modelo de Correlación, en el sentido de que su nivel de significación es, verdaderamente, α.

5.2. Estimación y Contraste de Hipótesis

Hemos de advertir que, aun siendo importantes, los argumentos utilizados hasta el momento no son suficientes para justificar el uso en el modelo de Correlación de los métodos de Inferencia propios del modelo de Regresión pues, el hecho de que dichos métodos sean óptimos, según diversos criterios, bajo las condiciones del modelo de

Regresión, no garantiza, en principio, su optimalidad desde el punto de vista del de Correlación. Así, por ejemplo el estimador de β utilizado en el anterior capítulo se justifica como estimador insesgado de mínima varianza y de máxima verosimilitud. Visto desde el punto de vista del modelo de Correlación, sólo sabemos, por ahora, que es insesgado y que las regiones de confianza anteriores siguen siendo correctas. Igualmente, el test F se justifica en el modelo de Regresión como uniformemente más potente entre todos los test invariantes con nivel de significación menor o igual que α, además de ser el test de la razón de verosimilitudes con nivel de significación α. Hasta ahora, sólo hemos probado que, bajos las condiciones del modelo de Correlación, el nivel de significación del test es, efectivamente, α. Falta, po lo tanto, una justificación a nivel teórico análoga a la que se obtiene con los teoremas 3.9, 3.10 y 3.11. Siguiendo el mismo esquema de demostración que en el capítulo 2, empezaremos por obtener un estadístico suficiente y completo para el modelo.

Teorema 5.2.
El estadístico $(\overline{\mathtt{m}}, S_M)$ es suficiente y completo para el modelo de Correlación.

Demostración.
Al igual que en el teorema 3.5, nos situaremos en las condiciones del teorema 9.18. Si se denota $\mu = (\nu, .^{\mathtt{n}}., \nu)'$, la función de verosimilitud correspondiente al modelo (5.1) es, en virtud de la proposición 9.17, la siguiente

$$\mathcal{L}(\mathtt{Y}, \mathtt{Z}; \nu, \Xi) = \frac{1}{(2\pi)^{(q+1)\mathtt{n}/2}|\Xi|^{\mathtt{n}/2}} \exp\left\{-\frac{1}{2}\mathtt{tr}\left(\Xi^{-1}(M-\mu)'(M-\mu)\right)\right\}. \qquad (5.9)$$

Teniendo en cuenta que todas las columnas de la matriz μ pertenecen al subespacio $\langle 1_{\mathtt{n}} \rangle$ y que $P_{\langle 1_{\mathtt{n}} \rangle} = \mathtt{n}^{-1} 1_{\mathtt{n}} 1_{\mathtt{n}}'$, se tiene que

$$\mathcal{L}(\mathtt{Y}, \mathtt{Z}; \nu, \Xi) = h(\nu, \Xi) \exp\left\{-\frac{1}{2}\mathtt{tr}(\Xi^{-1}M'M) + \mathtt{tr}\left(\Xi^{-1}\nu \cdot \overline{\mathtt{m}}'\right)\right\},$$

donde

$$h(\nu, \Xi) = \frac{1}{(2\pi)^{(q+1)\mathtt{n}/2}|\Xi|^{\mathtt{n}/2}} \exp\{-\frac{1}{2}\mathtt{tr}\left(\Xi^{-1}\|\nu\|^2 1_{\mathtt{n}} 1_{\mathtt{n}}'\right)\}.$$

Definamos el parámetro $\theta = \Xi^{-1}\nu$, perteneciente a $\mathbb{R}^{q+1}$, y consideremos entonces los siguientes vectores:

$$\tilde{\Delta}_1 = \mathrm{diag}(\Sigma^{-1}) = \begin{pmatrix} \Delta_{11} \\ \vdots \\ \Delta_{q+1,q+1} \end{pmatrix} \in \mathbb{R}^{q+1}, \qquad \tilde{M}_1 = \mathrm{diag}(M) \in \mathbb{R}^{q+1},$$

$$\tilde{\Delta}_2 = \text{triangsup}(\Sigma^{-1}) = \begin{pmatrix} \Delta_{12} \\ \vdots \\ \Delta_{q,q+1} \end{pmatrix} \in \mathbb{R}^{q(q+1)/2}, \quad \tilde{M}_2 = \text{triangsup}(M) \in \mathbb{R}^{q(q+1)/2},$$

Entonces, se verifica

$$\mathtt{tr}(\Xi^{-1}M'M) = \tilde{\Delta}_1'\tilde{M}_1 + 2\tilde{\Delta}_2'\tilde{M}_2, \qquad \mathtt{tr}(\Xi^{-1}\nu \cdot \overline{\mathsf{m}}) = \theta'\overline{\mathsf{m}}.$$

Si consideramos las funciones Q y H definidas mediante

$$Q(\nu, \Xi) = \begin{pmatrix} \tilde{\Delta}_1 \\ \tilde{\Delta}_2 \\ \theta \end{pmatrix}, \qquad H(YZ) = \begin{pmatrix} -\frac{1}{2}\tilde{T}_1 \\ -\tilde{T}_2 \\ \overline{\mathsf{m}} \end{pmatrix},$$

se verifica que

$$\mathcal{L}(\mathtt{Y}, \mathtt{Z}; \nu, \Xi) = h(\nu, \Xi) \exp\left\{ [Q(\nu, \Xi)]' H(\mathtt{YZ}) \right\}.$$

Por lo tanto, estamos hablando de una estructura estadística de tipo exponencial y, aplicando el teorema de factorización de Neyman, se deduce que el estadístico H es suficiente. Además, puede comprobarse que el interior de $\{Q(\nu, \Xi) : \nu \in \mathbb{R}^{q+1},\ \Xi > 0\}$ es distinto del vacío[3]. Luego, en virtud del teorema 9.18, H es completo. Además, podemos encontrar fácilmente una biyección bimedible ϕ tal que $\phi(H) = (\overline{\mathsf{m}}, S_M)$, de manera que este último estadístico es, igualmente, suficiente y completo.

∎

Corolario 5.3.

El siguiente estadístico es suficiente y completo

$$\left(\hat{\beta}, \hat{\sigma}^{2,\text{MV}}, \overline{z}, S_{ZZ}\right) \tag{5.10}$$

completo.

Demostración.

Para probar la tesis basta encontrar una biyección ϕ que transforme $(\overline{\mathsf{m}}, S_M)$ en dicho estadístico. $\overline{z}$ y S_{ZZ} se obtiene de forma trivial, mientras que, teniendo en cuenta (4.11) y (4.15), se tiene que

$$\begin{aligned} \underline{\hat{\beta}} &= S_{ZZ}^{-1} S_{ZY}, \\ \hat{\beta}_0 &= \overline{y} - \overline{z}'\underline{\hat{\beta}}, \\ \hat{\sigma}^{2,\text{MV}} &= s_Y^2 - S_{YZ} S_{ZZ}^{-1} S_{ZY}. \end{aligned}$$

[3] Téngase en cuenta que, en general, el conjunto de las matrices $p \times p$ simétricas se corresponden, de manera natural, con $\mathbb{R}^{p(p+1)/2}$, y que el subconjunto de las matrices definidas positivas (es decir, aquellas cuyo p-ésimo autovalor es estrictamente positivo) se identifica entonces con un abierto, pues el p-ésimo autovalor es una función continua.

La transformación inversa se obtiene de manera análoga a la expresada en (5.3)-(5.7). ∎

Corolario 5.4.
$\hat{\beta}$ y $\hat{\sigma}^{2,\mathtt{I}}$ son los EIMV de β y σ^2, respectivamente.

Demostración.
Sabemos que ambos son insesgados, luego, teniendo en cuenta el corolario anterior junto con el teorema de Lehmann-Scheffé, se concluye. ∎

Teorema 5.5.
El estadístico $(\overline{\mathtt{m}}, S_M)$ es el EMV de (ν, Ξ) en el modelo de Correlación.

Demostración.
Consideremos la función de verosimilitud (5.9) y tengamos en cuenta que, si $\overline{M}$ denota la matriz $(\overline{\mathtt{m}}, \overset{\mathtt{n}}{\ldots}, \overline{\mathtt{m}})'$, entonces las columnas de la matriz $M - \overline{M}$ pertenecen a $\langle 1_{\mathtt{n}} \rangle$. Por lo tanto, $(M - \overline{M})'(\overline{M} - \mu) = 0$. Luego, se tiene que

$$\begin{aligned} \mathtt{tr}\left(\Xi^{-1}(M-\mu)'(M-\mu)\right) &= \mathtt{tr}\left(\Xi^{-1}(M-\overline{M})'(M-\overline{M})\right) \\ &+ \mathtt{tr}\left(\Xi^{-1}(\overline{M}-\mu)'(\overline{M}-\mu)\right) \end{aligned}$$

Puede demostrarse fácilmente que el último sumando no puede ser negativo. Luego, para valores de $\mathtt{Y}$, $\mathtt{Z}$ y Ξ fijos, la anterior expresión alcanza el mínimo (y la función de verosimilitud el máximo) cuando $\mu = \overline{M}$ o, equivalentemente, cuando $\nu = \overline{\mathtt{m}}$. Pues bien, dados $\mathtt{Y}$ y $\mathtt{Z}$, es decir, dado M, busquemos entonces el valor de Ξ que maximiza

$$\mathcal{L}(\mathtt{Y}, \mathtt{Z}; \overline{\mathtt{m}}, \Xi) = \frac{1}{(2\pi)^{(q+1)\mathtt{n}/2}|\Xi|^{\mathtt{n}/2}} \exp\left\{-\frac{1}{2}\mathtt{tr}\left(\Xi^{-1}(M-\overline{M})'(M-\overline{M})\right)\right\}.$$

Sea $A = (M - \overline{M})'(M - \overline{M})$, que es, con probabilidad 1, invertible[4] . Aplicando el teorema 9.15, se tiene que el máximo se alcanza cuando $\Xi = \frac{1}{\mathtt{n}}A$, que coincide con S_M. Recapitulando, tenemos que, dados $\mathtt{Y}$, $\mathtt{Z}$, ν y Ξ,

$$\mathcal{L}(\mathtt{Y}, \mathtt{Z}; \nu, \Xi) \leq \mathcal{L}\left(\mathtt{Y}, \mathtt{Z}; \overline{\mathtt{m}}, \Xi\right) \leq \mathcal{L}\left(\mathtt{Y}, \mathtt{Z}; \overline{\mathtt{m}}, S_Z\right),$$

lo cual acaba la prueba. ∎

[4]Para demostrarlo basta tener en cuenta que, el rango de dicha matriz coincide con el de $M - \overline{M}$, que es $q + 1$, pues, según un razonamiento análogo al del teorema 5.1, el rango de $(1_{\mathtt{n}}|M)$ es $q + 2$, con probabilidad 1.

Corolario 5.6.
$\hat{\beta}$ y $\hat{\sigma}^{2,\text{MV}}$ son los EMV de β y σ^2, respectivamente.

Demostración.
Basta tener en cuenta que, dada una observación, el EMV es el valor del parámetro o, mejor dicho, la distribución de la familia, que hace más verosímil la observación. Según el teorema anterior, dicha distribución se expresa mediante los parámetros $\nu = \overline{\mathsf{m}}$ y $\Xi = S_M$. Teniendo en cuenta la biyección (5.2) que permite expresar el modelo con la ayuda de los parámetros $\beta, \sigma^2, \nu_Z, \Sigma_{ZZ}$, junto (4.11) y (4.15), se concluye.

∎

A tenor de estos resultados, el uso de los estimadores $\hat{\beta}$ y $\hat{\sigma}^{2,\text{I}}$ en el modelo de Correlación queda plenamente justificado. Veamos a continuación qué sucede con el test F a nivel α para contrastar una hipótesis del tipo $H_0 : A\beta = 0$. Sabemos que el nivel de significación del test es correcto en el modelo de Correlación. Para buscar el test de la razón de verosimilitudes a nivel α, hemos de tener en cuenta que, en virtud del teorema 5.1, la función de verosimilitud del modelo descompone como producto de dos factores: uno correspondiente a un modelo de Regresión con Z fijo y otro, a un modelo de correlación con q variables. En ambos casos sabemos maximizar la función a partir de una observación dada. No obstante, a la hora de calcular el estadístico de la razón de verosimilitudes, RV, para un contraste del tipo H_0, los máximos de los segundos factores se despejan, con lo que el estadístico RV para este problema resulta ser igual al que aparece en el teorema 3.11. En definitiva, podemos afirmar lo siguiente:

Teorema 5.7.
El test F definido en (3.26) es el de la razón de verosimilitudes a nivel α para contrastar un hipótesis del tipo $H_0 : A\beta = 0$.

En el capítulo 2 también se justificó el test F como UMP-invariante a nivel α respecto a cierto grupo de trasformaciones bimedibles. Pues bien, se verifica también que, desde el punto de vista el modelo de Correlación, F es el test UMP-invariante a nivel α respecto a otro grupo de transformaciones G que, lógicamente, es diferente del utilizado para justificar el test F en el capítulo 2 [5]. Lo probaremos únicamente para el contraste de la hipótesis inicial $H_0 : \underline{\beta} = 0$. La demostración para el caso general podemos encontrarla en el capítulo 16 de Arnold (1981). En ambos casos, se sigue el mismo esquema de demostración que en el capítulo 3, es decir, una reducción por suficiencia, seguida de varias reducciones por invarianza, que conducen a un modelo

[5] Tener en cuenta que el espacio de observaciones es distinto, por lo que las transformaciones no pueden ser, en ningún caso, las mismas.

con razón de verosimilitudes monótona en el cual se aplica el lema fundamental de Neyman-Pearson. Efectivamente, se verifica lo siguiente:

Teorema 5.8.
El grupo

$$G = \left\{ g_{k,K,B,\lambda} : \ k \in \mathbb{R}, \ K \in \mathbb{R}^q, \ B \in \mathcal{M}_{q\times q} \text{ invertible }, \ \lambda > 0 \right\}$$

de trasformaciones bimedibles definidas mediante

$$g_{k,K,B,\lambda} \begin{pmatrix} Y_i \\ Z_i \end{pmatrix} = \begin{pmatrix} \lambda Y_i + k \\ B' Z_i + K \end{pmatrix}, \quad i = 1, \ldots, \mathtt{n},$$

deja invariante tanto el modelo de Correlación como el problema de contraste de hipótesis. Además, el test F es UMP-invariante respecto a G a nivel α para contrastar la hipótesis inicial $H_0 : \underline{\beta} = 0$ en el modelo de Correlación.

Demostración.
Comprobar que el grupo deja invariante tanto el modelo como el problema de constraste de hipótesis es trivial. También se puede comprobar fácilmente que el estadístico suficiente y completo (5.10), que se denotará abreviadamente por S, es G-equivariante, por lo que induce un nuevo grupo de transformaciones, G^S, traducidas en términos del mismo de la siguiente forma

$$g^S_{k,K,B,\lambda} \left(\hat{\beta}_0, \underline{\hat{\beta}}, \hat{\sigma}^{2,\mathrm{I}}, \overline{z}, S_{ZZ} \right) = \left(\lambda \hat{\beta}_0 + k - \lambda K' \underline{\hat{\beta}}, \lambda B^{-1} \underline{\hat{\beta}}, \lambda^2 \hat{\sigma}^{2,\mathrm{I}}, B' \overline{z} + K, B' S_{ZZ} B \right).$$

A su vez, el grupo G^S descompone en suma de tres subgrupos, G_1, G_2 y G_3, cuyos elementos se definen, respectivamente, de la siguiente forma:

$$\begin{aligned} g_{k,K} \left(\hat{\beta}_0, \underline{\hat{\beta}}, \hat{\sigma}^{2,\mathrm{I}}, \overline{z}, S_{ZZ} \right) &= \left(\hat{\beta}_0 + k - K' \underline{\hat{\beta}}, \hat{\beta}, \hat{\sigma}^{2,\mathrm{I}}, \overline{z} + K, S_{ZZ} \right), \\ g_B \left(\hat{\beta}_0, \underline{\hat{\beta}}, \hat{\sigma}^{2,\mathrm{I}}, \overline{z}, S_{ZZ} \right) &= \left(\hat{\beta}_0, B^{-1} \underline{\hat{\beta}}, \hat{\sigma}^{2,\mathrm{I}}, B' \overline{z}, B' S_{ZZ} B \right), \\ g_\lambda \left(\hat{\beta}_0, \underline{\hat{\beta}}, \hat{\sigma}^{2,\mathrm{I}}, \overline{z}, S_{ZZ} \right) &= \left(\lambda \hat{\beta}_0, \lambda \underline{\hat{\beta}}, \lambda^2 \hat{\sigma}^{2,\mathrm{I}}, \overline{z}, S_{ZZ} \right). \end{aligned}$$

Dado que estos grupos verifican la propiedad (9.49), podemos obtener un estadístico invariante maximal mediante los tres pasos siguientes: en primer lugar un estadístico G_1-invariante maximal es el siguiente

$$M_1 = \left(\underline{\hat{\beta}}, \hat{\sigma}^{2,\mathrm{I}}, S_{ZZ} \right).$$

Sobre el modelo imagen de M_1 consideramos las transformaciones inducidas por G_2, que se expresan mediante

$$g^{M_1}_B \left(\underline{\hat{\beta}}, \hat{\sigma}^{2,\mathrm{I}}, S_{ZZ} \right) = \left(B^{-1} \underline{\hat{\beta}}, \hat{\sigma}^{2,\mathrm{I}}, B' S_{zz} B \right).$$

En virtud del teorema 9.13, el estadístico

$$M_2^1 \circ M_1 = \left(\hat{\sigma}^{2,\mathrm{I}}, \underline{\hat{\beta}}' S_{ZZ}^{-1} \underline{\hat{\beta}}\right)$$

es $(G_1 \oplus G_2)$-invariante maximal. Consideremos entonces las transformaciones inducidas por G_3 en la imagen de $M_2^1 \circ M_1$, que se expresan mediante

$$g_\lambda^{M_2^1 \circ M_1}\left(\hat{\sigma}^{2,\mathrm{I}}, \underline{\hat{\beta}}' S_{ZZ}^{-1} \underline{\hat{\beta}}\right) = \left(\lambda^2 \hat{\sigma}^{2,\mathrm{I}}, \lambda^2 \underline{\hat{\beta}}' S_{ZZ}^{-1} \underline{\hat{\beta}}\right).$$

En este caso, el estadístico

$$M_{1,2,3} = \frac{\underline{\hat{\beta}}' S_{ZZ}^{-1} \underline{\hat{\beta}}}{\hat{\sigma}^{2,\mathrm{I}}}$$

es, trivialmente, G-invariante maximal. Por un razonamiento completamente análogo se deduce que la función

$$\theta = \frac{\underline{\beta}' \Sigma_{ZZ}^{-1} \underline{\beta}}{\sigma^2} \tag{5.11}$$

es un invariante maximal para el espacio de parámetros, es decir, que $M_{1,2,3}$ depende de β_0, $\underline{\beta}$, σ^2, ν_z y Σ_{ZZ} a través de θ. Puede comprobarse fácilmente (cuestión propuesta) que el estadístico de contraste del test `F` se expresa mediante

$$F = \frac{\mathtt{n}}{q}\, M_{1,2,3}, \tag{5.12}$$

y que F condicionado a Z sigue un modelo de distribución $F_{q,\mathtt{n}-(q+1)}(\mathtt{n}\theta)$. La hipótesis nula se traduce en $H_0 : \theta = 0$, en cuyo caso $F \sim F_{q,\mathtt{n}-(q+1)}$. Se sigue de la aplicación de las propiedades de la probabilidad condicional regular, junto con el teorema de Fubini, que la densidad de F admite a expresión

$$p_\theta(\mathsf{f}) = \int_{\mathbb{R}^{\mathtt{n}_q}} f_{q,\mathtt{n}-(q+1),\mathtt{n}\sigma^{-2}\beta' S_{ZZ}^{-1}\beta}(\mathsf{f},\mathsf{z})\, dP_{\Sigma_{ZZ}}(\mathsf{z}), \tag{5.13}$$

siendo $P_{\Sigma_{ZZ}} = [N_q(0, \Sigma_{ZZ})]^{\mathtt{n}}$ y $f_{q,\mathtt{n}-(q+1),\mathtt{n}\sigma^{-2}\beta' S_{ZZ}^{-1}\beta}$ la función de densidad de la distribución F-Snedecor con grados de libertad q y $\mathtt{n} - (q+1)$, y parámetro de no centralidad $\mathtt{n}\sigma^{-2}\beta' S_{ZZ}^{-1}\beta$. El hecho de que en (5.13) integremos respecto una distribución q-normal de media 0, se debe a que la distribución de F depende únicamente de θ y, por lo tanto, es la misma para cualquier valor del parámetro ν_Z.

En definitiva, si para cada $\theta > 0$ consideramos el cociente $T_\theta(\mathsf{f}) = p_\theta(\mathsf{f})/p_0(\mathsf{f})$ (el denominador entra en la integral), se deduce, al igual que en capítulo 3, que el mismo es creciente en $\mathsf{f} \geq 0$ y, por lo tanto, el modelo imagen presenta razón de verosimilitudes monótona. Luego, aplicando el Lema de Neyman-Pearson, se tiene que el test consistente en comparar F con $F_{q,\mathtt{n}-(q+1)}^{\alpha}$ es UMP-invariante a nivel α. ∎

En definitiva y hablando en términos prácticos, los resultados obtenidos justifican el hecho de que, al efectuarse las inferencias en un análisis de Regresión, poco importa si los vectores explicativos son valores fijos controlados en el experimento o, por contra, corresponden a valores concretos de q variables aleatorias explicativas. correspondientes a sus valores están controlados en el experimento. Por ello, en muchas ocasiones se habla simplemente de estudios de regresión-correlación. Esta afirmación admite, no obstante, importantes matices, fundamentalmente en lo que concierne a los supuestos del modelo. Los supuestos del modelo de Regresión se analizaron críticamente en el capítulo anterior. Veamos qué sucede con los del modelo de Correlación.

5.3. Supuestos del modelo. Estudio asintótico

Las hipótesis del modelo de Correlación pueden enumerarse, teniendo en cuenta (5.8), de la siguiente forma: independencia de las observaciones; normalidad, tanto de la distribución marginal de los vectores aleatorios explicativos, como de la distribución condicional de la variable respuesta; homocedasticidad de dicha distribución condicional y, por último, linealidad de la relación entre la variable respuesta y las explicativas. En primer lugar, hemos de tener en cuenta que si admitimos que nuestro datos constituyen una muestra aleatoria simple de cierta distribución $(q+1)$-dimensional, la independencia de los mismos se deduce automáticamente. Pero además, y a diferencia del modelo de Regresión, se obtiene también la hipótesis de homocedasticidad. Respecto al supuesto de normalidad, veamos qué sucede, desde un punto de vista asintótico, con los estimadores de β y σ^2, así como del test `F`, cuando éste no se verifica.

Consideraremos pues el modelo dado por un una secuencia infinita de variables aleatorias reales independientes, que se denota por $\mathcal{Y}$, y otra muestra aleatoria de tamaño infinito, $\mathcal{Z}$, de una distribución Q dominada por la medida de Lebesgue en $\mathbb{R}^q$, las componentes de la cual poseen momentos de orden 2 finitos. Si $Y^{\mathtt{n}}$ denota el vector aleatorio compuesto por los $\mathtt{n}$ primeros términos de $\mathcal{Y}$ y $Z^{\mathtt{n}}$ denota la matriz aleatoria de dimensión $\mathtt{n} \times q$ cuyas filas son las trasposiciones de los $\mathtt{n}$ primeros vectores de $\mathcal{Z}$ [6], se supone, por hipótesis, que existen $\beta \in \mathbb{R}^{q+1}$ y $\sigma^2 > 0$ tales que $Y^{\mathtt{n}} = (1_{\mathtt{n}}|Z^{\mathtt{n}})\beta + \mathcal{E}^{\mathtt{n}}$, siendo $\mathcal{E}^{\mathtt{n}}$ un $\mathtt{n}$-vector aleatorio cuyas componentes son independientes, de media 0 y varianza σ^2 [7]. La matriz $(1_{\mathtt{n}}|Z^{\mathtt{n}})$ se denotará por $X_{\mathtt{n}}$.

[6] En ese caso, se deduce que el rango de la matriz $(1_{\mathtt{n}}|Z^{\mathtt{n}})$ es $q+1$, con probabilidad 1.

[7] Por lo tanto, si impusiéramos la normalidad de $\mathcal{E}^{\mathtt{n}}$ y Q, tendríamos un modelo de Correlación para cada $\mathtt{n} \in \mathbb{N}$.

Nuestro objetivo es obtener resultados similares a los conseguidos en la sección 3.4. Para ello consideraremos, en todo caso, a la distribución de $Y^{\mathbf{n}}$ condicionada al valor $\mathcal{Z}$, que coincide con la distribución condicionada al valor de $Z^{\mathbf{n}}$. Ello nos sitúa, precisamente, en las condiciones de la sección 3.4.

En primer lugar, veamos que el estimador de β es insesgado y consistente. Para ello consideramos la distribución condicional del estimador de β dada $\mathcal{Z}$, lo cual nos conduce a las hipótesis del teorema 3.14. Dado que la esperanza del estimador de β, condicionada al valor de $\mathcal{Z}$, es constante e igual a β, también coincide con β la esperanza de la distribución marginal. Además, la condición (3.32) se satisface en todo caso, y se verifica que

$$P\big(\|\hat{\beta}-\beta\|>\varepsilon\big)=\int P\big(\|\hat{\beta}-\beta\|^2>\varepsilon|\mathcal{Z}\big)\,dP^{\mathcal{Z}}.$$

Dado que, en virtud del teorema 3.14, el integrando converge a 0, se sigue del Teorema de la Convergencia Dominada que la integral también lo hace. Por lo tanto, el estimador de β es consistente.

Respecto al estimador de σ^2, se sigue de la proposición 3.2 que es insesgado en el modelo condicional y, por lo tanto, insesgado también en modelo total. Teniendo en cuenta el teorema 3.15 y aplicando un razonamiento análogo al anterior, se deduce que el estimador es consistente.

Por otra parte, sabemos que la condición (3.35) equivale, al menos en este caso, a la condición (4.40), expresada en términos de las distancias de Mahalanobis para los valores explicativos. Puede demostrarse[8] que, en nuestras condiciones, la condición (4.40) se verifica con probabilidad 1. Por lo tanto, la tesis (i) del teorema 3.18 se verifica para la distribución condicional dada $\mathcal{Z}$. Por lo tanto, teniendo en cuenta la propia definición de convergencia en distribución y aplicando nuevamente el Teorema de la Convergencia Dominada, se obtiene la convergencia (i) en términos globales. En consecuencia, el elipsoide (3.36) constituye una región de confianza asintótica para el parámetro β. Por último, un razonamiento completamente análogo prueba la validez asintótica del test (3.26) para contrastar, con un nivel de significación α, la hipótesis inicial $H_0 : A\beta = 0$.

En definitiva, si obviamos el supuesto de normalidad (suponiendo que la distribución de las variables explicativas esté dominada por la medida de Lebesgue en $\mathbb{R}^q$ y es de cuadrado integrable) estamos en las mismas condiciones que en el modelo de Regresión: el comportamiento asintótico de los métodos de inferencia considerados es satisfactorio. Pero no debemos engañarnos, pues el problema más serio se encuentra

[8] Arnold, *Asymptotic Validity of F Test for the Ordinary Linear Model and Multiple Correlation Model*, Journal of the American Statistical Association, Dec. 1980, Vol. 75, 890-894.

en el supuesto de linealidad, estrechamente vinculado al de $(q+1)$-normalidad. Al igual que en el capítulo anterior, habría que considerar la posibilidad de trasformar de manera adecuada las variables para conseguir una relación lineal. No obstante, sería interesante disponer de un algoritmo que permitiera saber qué transformaciones considerar y cómo evaluar la efectividad de las mismas. En el caso del modelo de Correlación y a la vista de (5.1), parece razonable buscar transformaciones que confieran a nuestro vector aleatorio $(q+1)$-dimensional una distribución $(q+1)$-normal, en cuyo caso, el modelo se satisfaría plenamente. Lógicamente, el problema es difícil, pero podemos considerar una extensión multivariante del algoritmo de Box-Cox, estudiado en el capítulo anterior, con el objetivo de aproximarnos a esta situación No obstante, hemos de tener presente la posibilidad de resolver el problema mediante la estimación de las densidades marginales del vector de variables explicativas y de la conjunta, lo cual permite estimar la densidad de la distribución condicional, tal y como se indicó en el anterior capítulo.

5.4. Inferencias sobre los coeficientes de correlación

Aunque el estudio de los distintos coeficientes de correlación (múltiple, simple y parciales) es posible desde el punto de vista del modelo de Regresión, alcanza pleno sentido cuando las variables explicativas no están controladas sino que son aleatorias. En especial, cuando asumimos las hipótesis del modelo de Correlación, podemos expresar la distribución, tanto exacta como asintótica, de dichos coeficientes, lo cual es de gran utilidad de cara a la realización de inferencias sobre los mismos.

Consideraremos, en primer lugar, los coeficientes de correlación múltiple, tanto muestral como probabilístico. Realmente, no son éstos sino sus cuadrados, los denominados coeficientes de determinación, los coeficientes que más nos interesan, por razones que aclararemos. Del corolario 5.6 se sigue que $R^2_{\mathbf{Y}\cdot\mathbf{Z}}$ es el EMV de $\rho^2_{y\cdot z}$. Por lo tanto, se trata de un estimador consistente y asintóticamente eficiente (lo mismo sucede con los coeficientes de correlación simple y parcial). Por otra parte, se sigue de (4.30) y (5.12) que el estadístico invariante maximal para contrastar la hipótesis inicial $H_0 : \underline{\beta} = 0$ es proporcional a $R^2_{\mathbf{Y}\cdot\mathbf{Z}}/(1-R^2_{\mathbf{Y}\cdot\mathbf{Z}})$, tanto en el modelo de Regresión como en el de Correlación. Hemos de tener en cuenta que la función $\phi(x) = x(1-x)^{-1}$ constituye una biyección de $[0,1]$ en $[0,+\infty]$. Por lo tanto, el estadístico F depende de los datos únicamente a través de $R^2_{\mathbf{Y}\cdot\mathbf{Z}}$. Además, un valor de $R^2_{\mathbf{Y}\cdot\mathbf{Z}}$ *próximo* a 0 se traducirá en un resultado no significativo. En verdad, esto era de esperar, pues la hipótesis H_0, desde el punto de vista del modelo de Correlación, equivale a $\rho^2_{y\cdot z} = 0$.

Dado que se supone normalidad, dicha hipótesis equivale, a su vez, a la independencia entre la variable respuesta y el vector de variables explicativas. Además, la distribución de F depende únicamente del invariante maximal θ, definido en (5.11), que equivale, precisamente, a $\rho^2_{y\cdot z}/(1-\rho^2_{y\cdot z})$. Equivalentemente, podemos afirmar que la distribución de $R^2_{\mathtt{Y}\cdot\mathtt{Z}}$ depende únicamente de $\rho^2_{y\cdot z}$. En el caso $\rho^2_{y\cdot z}=0$, se verifica que F sigue una distribución $F_{q,\mathtt{n}-(q+1)}$. Luego, dado que $R^2_{\mathtt{Y}\cdot\mathtt{Z}} = q(\mathtt{n}-q)^{-1}F/(1+F)$, podemos obtener, aplicando el teorema del cambio de variables a una función del tipo (2.7), la densidad de la distribución de $R^2_{\mathtt{Y}\cdot\mathtt{Z}}$ bajo la hipótesis de independencia. Además, se sigue del teorema 3.20 que, bajo la hipótesis inicial de independencia,

$$\mathtt{n}\frac{R^2_{\mathtt{Y}\cdot\mathtt{Z}}}{1-R^2_{\mathtt{Y}\cdot\mathtt{Z}}} \xrightarrow{d} \chi^2_q. \tag{5.14}$$

Esta afirmación es válida para el modelo asintótico considerado en la sección anterior (sin suponer normalidad). En el caso general, basta aplicar el teorema del cambio de variables a la densidad (5.13) para obtener una función que dependerá del parámetro únicamente a partir de $\rho^2_{y\cdot z}$. Una expresión explícita de esta densidad puede encontrarse en Anderson (1958), capítulo 4. En Bilodeau (1999) se obtiene, además, la distribución asintótica de R^2 bajo el supuesto de normalidad y en el caso $\rho^2 \neq 0$. Concretamente, se verifica

$$\sqrt{\mathtt{n}}(R^2_{\mathtt{Y}\cdot\mathtt{Z}} - \rho^2_{y\cdot z}) \xrightarrow{d} N\big(0, 4\rho^2_{y\cdot z}(1-\rho^2_{y\cdot z})^2\big)$$

Esta expresión no resulta muy útil puesto que el parámetro desconocido $\rho^2_{y\cdot z}$ aparece en la distribución límite. No obstante, aplicando el teorema 9.27 con la función $\delta(x) = \sqrt{x}$, se verifica

$$\sqrt{n}(R_{\mathtt{Y}\cdot\mathtt{Z}} - \rho_{y\cdot z}) \xrightarrow{d} N\big(0, (1-\rho^2_{y\cdot z})^2\big) \tag{5.15}$$

Luego, aplicando nuevamente el teorema 9.27, pero con $\delta(x) = 2^{-1}\ln[(1+x)(1-x)^{-1}]$ en esta ocasión, se obtiene

$$\frac{\sqrt{\mathtt{n}}}{2}\left(\ln\frac{1+R_{\mathtt{Y}\cdot\mathtt{Z}}}{1-R_{\mathtt{Y}\cdot\mathtt{Z}}} - \ln\frac{1+\rho_{y\cdot z}}{1-\rho_{y\cdot z}}\right) \xrightarrow{d} N(0,1), \tag{5.16}$$

lo cual permite, por ejemplo, construir tests de hipótesis o intervalos de confianza aproximados para $\rho_{y\cdot z}$. Cuando $q=1$, es decir, cuando existe una única variable explicativa, estaremos hablando del coeficiente de correlación lineal simple. Ni que decir tiene que todo lo dicho anteriormente para el coeficiente de correlación múltiple es válido para el simple. En particular, se verifican (5.16) y, en el caso nulo, (5.14).

Para acabar, veamos qué podemos decir de los coeficientes de correlación parcial. Consideremos cualquiera de las variables explicativas, Z_j, y denótese por Z_R al resto

de las mismas. En ese caso, sabemos por (4.35) que, fijo $\mathtt{z}$, $r^2_{Y,\mathbf{Z}_j\bullet\mathbf{Z}_R}$ constituye un invariante maximal para contrastar la hipótesis inicial $H_0 : \beta_j = 0$ en el modelo de Regresión, lo cual no es de extrañar, teniendo en cuenta que H_0 equivale a $\rho^2_{y,\mathbf{Z}_j\bullet\mathbf{Z}_R} = 0$. Es más, en el modelo de Correlación, la hipótesis H_0 equivaldría a la independencia condicional entre Y y Z_j dadas Z_R, es decir, a la nulidad del coeficiente $\rho^2_{y,z_j\bullet z_R}$. Puede probarse, a partir de (3.24), que la distribución de

$$[\mathtt{n} - (q+1)]\,\frac{r^2_{Y,\mathbf{Z}_j\bullet\mathbf{Z}_R}}{1 - r^2_{Y,\mathbf{Z}_j\bullet\mathbf{Z}_R}}$$

condicionada a Z sigue un modelo $F_{1,\mathtt{n}-(q+1)}(\theta)$, donde

$$\theta = \mathtt{n}\,\frac{\beta_j' s^2_{\mathbf{Z}_j\bullet\mathbf{Z}_R}\beta_j}{\sigma^2}.$$

Por lo tanto, integrando la función $f_{1,\mathtt{n}-(q+1),\mathtt{n}\sigma^2\beta_j' s^2_{\mathbf{Z}_j\bullet\mathbf{Z}_R}\beta_j}$ respecto a la potencia $\mathtt{n}$-ésima de la distribución $N_q(0,\Sigma_{ZZ})$ y, aplicando el teorema del cambo de variables, obtenemos la densidad del coeficiente de correlación parcial al cuadrado. Puede demostrarse también[9] que la distribución del mismo depende del parámetro únicamente a través de $\rho^2_{y,z_j\bullet z_R}$. La forma explícita de esta densidad podemos encontrarla en Anderson (1958). Además, dado que, si condicionamos en Z_R, obtenemos un modelo de correlación simple y, en consecuencia, convergencias del tipo (5.14) y (5.16) a distribuciones que no dependen del propio Z_R, dichas convergencias se verifican también para la distribución conjunta, Es decir, que en el caso nulo, se tiene que

$$\mathtt{n}\,\frac{r^2_{Y,Z_j,\bullet Z_R}}{1 - r^2_{Y,Z_j,\bullet Z_R}} \to \chi^2_1,$$

y, en general,

$$\frac{\sqrt{\mathtt{n}}}{2}\left(\ln\frac{1 + r_{Y,Z_j,\bullet Z_R}}{1 - r_{Y,Z_j,\bullet Z_R}} - \ln\frac{1+\rho_{y,z_j\bullet z_R}}{1-\rho_{y,z_j\bullet z_R}}\right) \longrightarrow N(0,1).$$

Cuestiones propuestas

1. Probar que si consideramos el modelo que se obtiene eliminando en (5.8) la hipótesis de normalidad, pero suponiendo que $Z_1, \ldots, Z_{\mathtt{n}}$ constituye una muestra aleatoria simple de una distribución dominada por la medida de Lebesgue en $\mathbb{R}^q$, se verifica también que $\mathtt{rg}(X) = q + 1$, con probabilidad 1

[9] Para ello basta tener en cuenta que $s^2_{\mathbf{Z}_j\bullet\mathbf{Z}_R}$ sigue una distribución χ^2 y aplicar las propiedades de la misma.

2. Probar que $\hat{\beta}$ es insesgado en el modelo de Correlación, que el elipsoide (3.12) sigue siendo una región de confianza a nivel $1-\alpha$.

3. Probar que los intervalos de confianza (4.10), (4.24) y (4.25) siguen siendo validos.

4. Probar (5.12).

5. Describir la densidad del coeficiente de correlación parcial en el caso nulo.

Capítulo 6

Análisis de la Varianza

En este capítulo se proponen métodos para resolver problemas como el tercero y cuarto del capítulo 1. Si en el capítulo 4 estudiamos la posible influencia de q variables cuantitativas en la media de una variable respuesta y, en este consideraremos la influencia que puedan tener en la misma una o varias variables cualitativas, denominadas factores. Es decir, analizaremos en qué medida una división en subgrupos de la población afecta a la distribución de la variable y o, al menos, a su esperanza. A lo largo del capítulo estudiaremos diversos modelos o diseños con uno y dos factores. Es muy común, por cierto, denominar este tema mediante el epígrafe Diseño de Experimentos. No obstante, el título escogido se debe a que la resolución de los contrastes de hipótesis se realizará en todo caso mediante el test F, también denominado Anova, abreviatura de Análisis de la Varianza.

El estudio del primer diseño, denominado Completamente Aleatorizado, tiene, indiscutiblemente, perfecto sentido desde el marco teórico establecido en el capítulo 3. Respecto a los demás diseños considerados en este capítulo, se hace necesaria la imposición de restricciones naturales sobre los tamaños de muestra considerados u otras, en principio arbitrarias, sobre los parámetros del modelo. Esta circunstancia puede llevarnos a enfocar el estudio desde un punto de vista teórico más general, es decir, partiendo de un Modelo Lineal de Rango no Completo, en el cual se hace uso del concepto de inversa generalizada de una matriz. Este problema se abordará en el capítulo 7.

Cualquiera de los diseños a estudiar puede formalizarse mediante un modelo de regresión lineal múltiple, por lo que todo lo estudiado en el capítulo 4 referente a la diagnosis y validación del modelo (tests de bondad de ajuste, análisis de los residuos, transformaciones de variables para mejorar el ajuste), así como los resultados asintóticos obtenidos en el capítulo 3, son aplicables aquí. No obstante, hemos de ad-

vertir que cualquier cambio en la variable respuesta debe afectar por igual a todos los niveles del factor o factores considerados. Además, la media de la variable transformada no será igual, en general, a la transformación de la media. Estos inconvenientes puede hacernos desistir en la búsqueda de transformaciones que permitan un adecuado ajuste del modelo, por lo que en ocasiones nos veremos obligados a apoyarnos en resultados de tipo asintótico, a buscar métodos alternativos o, sencillamente, a *confiar* en la robustez del método estadístico.

La última sección del capítulo está dedicada al estudio de ciertos diseños en los cuales uno o varios de los factores del modelo toman valores aleatorios en cierto espacio, en contraposición con los diseños estudiados en el resto del capítulo, donde los niveles de los factores se fijan de antemano. Hemos de anticipar aquí que, si bien las propuestas de solución a los principales problemas de Inferencia en un diseño con efectos aleatorios presenta bastantes similitudes con las correspondientes a diseños de efectos fijos, las primeras carecen de justificaciones teóricas de la solidez de las segundas.

Empezaremos pues estudiando el diseño más sencillo y natural, el Diseño Completamente Aleatorizado, que generaliza el diseño a partir del cual se obtiene el test de Student para el contraste de dos medias. Aprovecharemos este modelo para desarrollar las Comparaciones Múltiples y el Análisis de la Covarianza, aunque ambos temas tienen perfecto sentido en cualquiera de los demás modelos considerados en el capítulo.

6.1. Diseño Completamente Aleatorizado

Este diseño se utiliza para determinar la influencia de una factor cualitativo con $\mathtt{a}$ niveles en una variable respuesta y. Dado un entero positivo $\mathtt{a}$, se considera, para cada $i = 1, \ldots, \mathtt{a}$, una muestra aleatoria simple $Y_{i1}, \ldots, Y_{i\mathtt{n}_i}$ de una distribución normal de media μ_i, siendo independientes las muestras e idénticas las varianzas correspondientes a las mismas. Hablando en términos prácticos, se supone que la población estudiada es susceptible de dividirse en $\mathtt{a}$ partes en función del valor de la variable cualitativa o factor y que esta diferenciación puede traducirse únicamente en una diversificación de la media de la distribución. Cada muestra representa pues un valor o nivel del factor considerado. En lo que sigue, Y denotará el vector aleatorio compuesto por las $\mathtt{a}$ muestras ordenadas, $\mathtt{n}$ será la suma de los distintos tamaños de muestra; para cada entero positivo k, 0_k y 1_k denotarán, respectivamente, los vectores de $\mathbb{R}^k$ cuyas componentes son todas 0 y 1; para cada cada $i = 1, \ldots, \mathtt{a}$, $\mathtt{v}_i$ será el vector de $\mathbb{R}^n$ definido mediante $\mathtt{v}_i = (0'_{\mathtt{n}_1} \ldots 1'_{\mathtt{n}_i} \ldots 0'_{\mathtt{n}_\mathtt{a}})'$. En ese caso, estaremos hablando del

siguiente modelo lineal normal:

$$Y \sim N_{\mathtt{n}}(\mu, \sigma^2 \mathtt{Id}), \quad \mu \in V = \langle \mathtt{v}_1, \ldots, \mathtt{v}_{\mathtt{a}} \rangle, \ \sigma^2 > 0.$$

Que el factor no influya en la respuesta quiere decir que todos los niveles del mismo poseen una misma distribución, es decir, una misma media. Por lo tanto, el principal contraste a resolver es

$$H_0 : \mu_1 = \ldots = \mu_{\mathtt{a}},$$

contra su alternativa, es decir, que al menos un par de medias difieran entre sí. Es decir, la hipótesis inicial a considerar es $\mu \in \langle 1_{\mathtt{n}} \rangle$. Nótese que, al verificarse los supuesto de normalidad y homocedasticidad, la igualdad de medias equivale a la igualdad de distribuciones. Es lo más común que alguno de estos supuestos no se verifique o bien que se satisfaga sólo aproximadamente, en cuyo caso, la hipótesis inicial debe interpretarse como que el factor no influye por término medio en la respuesta.

Los problemas de estimación y contraste de hipótesis referente a este modelo ya ha sido en esencia resuelto en el capítulo 3. Para aplicar los resultados allí obtenidos es fundamental calcular, teniendo en cuenta (9.8), las matrices de las proyecciones ortogonales sobre V y $\langle 1_{\mathtt{n}} \rangle$. Así, si para cada par de enteros positivos k_1 y k_2, $1_{k_1 \times k_2}$ denota la matriz $k_1 \times k_2$ cuyas componentes son todas igual a 1, se verifica

$$P_V = \left(\begin{array}{c|c|c} \mathtt{n}_1^{-1} \cdot 1_{\mathtt{n}_1 \times \mathtt{n}_1} & \ldots & 0 \cdot 1_{\mathtt{n}_1 \times \mathtt{n}_{\mathtt{a}}} \\ \hline \vdots & \ddots & \vdots \\ \hline 0 \cdot 1_{\mathtt{n}_{\mathtt{a}} \times \mathtt{n}_1} & \ldots & \mathtt{n}_{\mathtt{a}}^{-1} \cdot 1_{\mathtt{n}_{\mathtt{a}} \times \mathtt{n}_{\mathtt{a}}} \end{array}\right), \qquad P_{\langle 1_{\mathtt{n}} \rangle} = \mathtt{n}^{-1} \cdot 1_{\mathtt{n} \times \mathtt{n}}. \tag{6.1}$$

Podemos descomponer Y en los tres siguientes sumando ortogonales:

$$Y = P_{\langle 1_{\mathtt{n}} \rangle} Y + P_{V | \langle 1_{\mathtt{n}} \rangle} Y + P_{V^\perp} Y \tag{6.2}$$

que, teniendo en cuenta, (6.1), queda como sigue

$$\left(\begin{array}{c} Y_{11} \\ \vdots \\ \hline Y_{1\mathtt{n}_1} \\ \hline \vdots \\ \vdots \\ \hline Y_{\mathtt{a}1} \\ \vdots \\ Y_{\mathtt{an}_{\mathtt{a}}} \end{array}\right) = \left(\begin{array}{c} \overline{y}_{\cdot\cdot} \\ \vdots \\ \hline \overline{y}_{\cdot\cdot} \\ \hline \vdots \\ \vdots \\ \hline \overline{y}_{\cdot\cdot} \\ \vdots \\ \overline{y}_{\cdot\cdot} \end{array}\right) + \left(\begin{array}{c} \overline{y}_{1\cdot} - \overline{y}_{\cdot\cdot} \\ \vdots \\ \hline \overline{y}_{1\cdot} - \overline{y}_{\cdot\cdot} \\ \hline \vdots \\ \vdots \\ \hline \overline{y}_{\mathtt{a}\cdot} - \overline{y}_{\cdot\cdot} \\ \vdots \\ \overline{y}_{\mathtt{a}\cdot} - \overline{y}_{\cdot\cdot} \end{array}\right) + \left(\begin{array}{c} Y_{11} - \overline{y}_{1\cdot} \\ \vdots \\ \hline Y_{1\mathtt{n}_1} - \overline{y}_{1\cdot} \\ \hline \vdots \\ \vdots \\ \hline Y_{\mathtt{a}1} - \overline{y}_{\mathtt{a}\cdot} \\ \vdots \\ Y_{\mathtt{an}_{\mathtt{a}}} - \overline{y}_{\mathtt{a}\cdot} \end{array}\right), \tag{6.3}$$

donde

$$\overline{y}_{\cdot\cdot} = \frac{1}{\mathtt{n}} \sum_{i=1}^{a} \sum_{j=1}^{\mathtt{n}_i} Y_{ij}, \qquad \overline{y}_{i\cdot} = \frac{1}{\mathtt{n}_i} \sum_{j=1}^{\mathtt{n}_i} Y_{ij}, \quad i = 1, \ldots, r.$$

Por lo tanto, se sigue del corolario 3.6 que los estimadores insesgados de mínima varianza de μ y σ^2 son, respectivamente,

$$\hat{\mu} = \begin{pmatrix} \overline{y}_{1\cdot} \\ \vdots \\ \overline{y}_{1\cdot} \\ \vdots \\ \vdots \\ \overline{y}_{\mathtt{a}\cdot} \\ \vdots \\ \overline{y}_{\mathtt{a}\cdot} \end{pmatrix}, \qquad \hat{\sigma}^{2,\mathtt{I}} = \frac{1}{\mathtt{n}-\mathtt{a}} \sum_{i=1}^{a} \sum_{j=1}^{\mathtt{n}_i} (Y_{ij} - \overline{y}_{i\cdot})^2.$$

Del teorema 3.7 se sigue que el EMV de μ es el mismo, mientras que el de σ^2 se obtiene dividiendo por $\mathtt{n}$ en lugar de $\mathtt{n}-\mathtt{a}$. Además, podemos hacer uso de la proposición 3.8 para construir regiones de confianza para μ y σ^2.

Por otra parte, en lo que respecta al contraste principal, se sigue de (6.3) que $\|P_{V|\langle 1_{\mathtt{n}}\rangle} Y\|^2 = \sum_{i=1}^{r} \mathtt{n}_i (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2$. Por lo tanto y según (3.23), el test F para contrastar la hipótesis inicial de igualdad de medias tendrá por estadístico de contraste

$$F = \frac{(\mathtt{a}-1)^{-1} \sum_{i=1}^{a} \mathtt{n}_i (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2}{(\mathtt{n}-\mathtt{a})^{-1} \sum_{i=1}^{a} \sum_{j=1}^{\mathtt{n}_i} (Y_{ij} - \overline{y}_{i\cdot})^2},$$

que seguirá un modelo de distribución

$$F_{\mathtt{a}-1,\mathtt{n}-\mathtt{a}} \left(\frac{\|P_{V|\langle 1_{\mathtt{n}}\rangle} \mu\|^2}{\sigma^2} \right).$$

Para calcular el parámetro de no centralidad basta tener en cuenta que $P_{\langle 1_{\mathtt{n}}\rangle}\mu$ es el vector cuyas componentes son todas iguales a $\overline{\mu} := a^{-1} \sum_{i=1}^{a} \mu_i$. Por lo tanto, se verifica

$$F \sim F_{\mathtt{a}-1,\mathtt{n}-\mathtt{a}} \left(\sigma^{-2} \sum_{i=1}^{\mathtt{a}} \mathtt{n}_i (\mu_i - \overline{\mu})^2 \right).$$

En definitiva, el test F a nivel α para contrastar H_0 consiste en comparar el estadístico F con $F^{\alpha}_{\mathtt{a}-1,\mathtt{n}-\mathtt{a}}$. Este test es, por lo tanto, insesgado, UMP-invariante y de razón de verosimilitudes. Un valor de F mayor que $F^{\alpha}_{\mathtt{a}-1,\mathtt{n}-\mathtt{a}}$ se interpretará como una influencia del factor sobre la media de la variable respuesta.

Análisis de la varianza y regresión

Este modelo puede parametrizarse también mediante coordenadas de la media respecto a una matriz $\mathtt{X} \in \mathcal{M}_{\mathtt{n}\times a}$, tal que sus columnas constituyan una base de V. dado que la principal hipótesis nula a contrastar es

$$H_0 : \mathtt{E}[Y] \in \langle 1_{\mathtt{n}} \rangle,$$

parece apropiado que el término independiente $1_{\mathtt{n}}$ esté incluido en la matriz $\mathtt{X}$, lo cual significa entender el análisis de la varianza como un problema de regresión lineal. Es decir, se trata de encontrar una matriz $\mathtt{Z} \in \mathcal{M}_{\mathtt{n}\times(\mathtt{a}-1)}$ tal que $\mathtt{X} = (1_{\mathtt{n}}|\mathtt{Z})$ sea una base de V. En tal caso, la hipótesis inicial H_0 anterior equivale, en los términos del capítulo 4, a

$$H_0 : \underline{\beta} = 0$$

Se trataría pues de un contraste total, según se ha denominado en la sección 4.2. El problema que se nos presenta es cómo elegir $\mathtt{Z}$ para completar una base de $V = \langle \mathtt{v}_1, \ldots, \mathtt{v}_{\mathtt{a}} \rangle$. Por ejemplo, la matriz

$$\mathtt{X} = (1_{\mathtt{n}}|\mathtt{v}_1 \ldots \mathtt{v}_{\mathtt{a}-1}) \tag{6.4}$$

verifica las condiciones requeridas. En ese caso, de la ecuación $\mu = \mathtt{X}\beta$ se sigue que

$$\begin{array}{rcl} \beta_0 & = & \mu_{\mathtt{a}} \\ \beta_1 & = & \mu_1 - \mu_{\mathtt{a}} \\ & \vdots & \\ \beta_{\mathtt{a}-1} & = & \mu_{\mathtt{a}-1} - \mu_{\mathtt{a}} \end{array} \tag{6.5}$$

No obstante, sería conveniente que la matriz $\mathtt{Z}$ escogida para parametrizar el modelo correspondiese a una descomposición natural del subespacio V. Podemos entender como natural una descomposición ortogonal del espacio. Esta calificación no se debe únicamente a criterios estéticos pues la descomposición en subespacios ortogonales facilita enormemente el trabajo de cara a la aplicación del test F, como veremos en los diseños multifactoriales. En nuestro caso estamos hablando, concretamente, de la descomposición

$$V = \langle 1_{\mathtt{n}} \rangle \oplus V|\langle 1_{\mathtt{n}} \rangle.$$

Es decir, que buscamos $\mathtt{Z} \in \mathcal{M}_{\mathtt{n}\times(\mathtt{a}-1)}$ cuyas columnas $\{e_1, \ldots, e_{\mathtt{a}-1}\}$ constituyan una base de $V|\langle 1_{\mathtt{n}} \rangle$. Para ello basta tener en cuenta que un vector e pertenece a $V|\langle 1_{\mathtt{n}} \rangle$ cuando puede expresarse mediante $e = \sum_{i=1}^{a} \alpha_i \mathtt{v}_i$, con la restricción $\sum_{i=1}^{a} \mathtt{n}_i \alpha_i = 0$. Por lo tanto, podemos expresarlo también mediante

$$e = \sum_{i=1}^{\mathtt{a}-1} \alpha_i (\mathtt{v}_i - \frac{\mathtt{n}_i}{\mathtt{n}_{\mathtt{a}}} \mathtt{v}_{\mathtt{a}}).$$

En consecuencia, la familia $e_i = \mathtt{v}_i - \mathtt{n}_\mathtt{a}^{-1}\mathtt{n}_i \cdot \mathtt{v}_\mathtt{a}$, $i = 1, \ldots, \mathtt{a}-1$, constituye una base de $V|\langle 1_\mathtt{n}\rangle$. Esto nos lleva a parametrizar el modelo de manera natural mediante la matriz

$$(1_\mathtt{n}|e_1 \ldots e_\mathtt{a}) \tag{6.6}$$

En ese caso, dado que $P_{\langle 1_\mathtt{n}\rangle}\mu = \beta_0 \cdot 1_\mathtt{n}$, se sigue que $\beta_0 = \overline{\mu}_{\cdot\cdot}$, siendo $\overline{\mu}_{\cdot\cdot} = \mathtt{n}^{-1}\sum_{i=1}^{a}\mathtt{n}_i\mu_i$. Si el diseño equilibrado, es decir, si $\mathtt{n}_1 = \ldots = \mathtt{n}_\mathtt{a}$, se tiene que $\overline{\mu}_{\cdot\cdot} = a^{-1}\sum_{i=1}^{a}\mu_i$. Este último parámetro se denota por $\overline{\mu}$ y es la media aritmética de las medias. En definitiva, si resolvemos la ecuación $\mu = \mathbf{X}\beta$ para la matriz (6.6) obtenemos todos los coeficientes de regresión:

$$\begin{array}{rcl} \beta_0 & = & \overline{\mu}_{\cdot\cdot} \\ \beta_1 & = & \mu_1 - \overline{\mu}_{\cdot\cdot} \\ & \vdots & \\ \beta_{\mathtt{a}-1} & = & \mu_{\mathtt{a}-1} - \overline{\mu}_{\cdot\cdot} \end{array} \tag{6.7}$$

Con mucha frecuencia, el diseño completamente aleatorizado (y en la misma línea todos los demás modelos del análisis de la varianza) se expresa de la forma

$$Y_{ij} = \theta + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \text{ independientes}, \tag{6.8}$$

en función de ciertos pararámetros $\theta, \alpha_1, \ldots, \alpha_\mathtt{a}, \sigma^2$. Expresar de esta forma un modelo del análisis de la varianza puede reportar algunos beneficios, especialmente en diseños con varios factores, como veremos más adelante. Es más, modelos más complicados como el de efectos aleatorios y los modelos mixtos se definen partiendo de una notación similar a ésta. En definitiva, es la notación que se utiliza comúnmente en el análisis de la varianza. Veamos qué relación existe entre ésta y la que hemos usado hasta el momento.

Las parametrizaciones estudiadas anteriormente se corresponden con $\theta = \beta_0$ y $\alpha_i = \beta_i$, para $i = 1, \ldots, \mathtt{a}-1$. En particular, (6.5) se corresponde con $\theta = \mu_\mathtt{a}$ y $\alpha_i = \mu_i - \mu_\mathtt{a}$, $i = 1, \ldots, \mathtt{a}$. Por lo tanto, considerar como base de V la matriz (6.4) equivale a expresar el modelo según (6.8) con la restricción $\alpha_\mathtt{a} = 0$. Sin embargo, la parametrización (6.7) se corresponde con $\theta = \overline{\mu}_{\cdot\cdot}$ y $\alpha_i = \mu_i - \overline{\mu}_{\cdot\cdot}$. Luego, considerar como base (6.6) equivale a imponer en (6.8) la restricción $\sum_{i=1}^{a}\mathtt{n}_i\alpha_i = 0$. Si el diseño es equilibrado, quedaría como $\sum_{i=1}^{a}\alpha_i = 0$.

En general, el sistema de ecuaciones lineales $\mu_i = \theta + \alpha_i$, $i = 1, \ldots, \mathtt{a}$ presenta una recta de soluciones en $\mathbb{R}^{a+1}$, por lo que el parámetro no queda determinado. Por lo tanto, para conseguir una solución única se hace necesario imponer una ecuación adicional al sistema, que puede ser una restricción lineal sobre los $\alpha_i's$. Eso es lo que, en definitiva, se ha hecho con las dos parametrizaciones consideradas. De todas

formas, el problema se resuelve de forma más general, al menos en principio, en el capítulo 7 dedicado al modelo lineal de rango no completo.

Para ilustrar lo expuesto anteriormente, podemos considerar un diseño completamente aleatorizado con tres niveles y cuatro datos por nivel. En ese caso, podemos parametrizar de diversas formas, por ejemplo según (6.5) o (6.7), que se corresponden con las restricciones $\alpha_3 = 0$ y $\sum_{i=1}^{3} \alpha_i = 0$, respectivamente. En todo caso, estaremos considerando las matrices siguientes:

$$\mathtt{X}_1 = \left(\begin{array}{c|cc} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \hline 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ \hline 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{array}\right) \qquad \mathtt{X}_2 = \left(\begin{array}{c|cc} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \hline 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ \hline 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{array}\right)$$

En ocasiones, podemos encontrarnos con un modelo no equilibrado parametrizado en función de una matriz del tipo $\mathtt{X}_2$ (manejando únicamente los valores 1,0 y -1). En ese caso, el término independiente es $\overline{\mu}$, que direfirá de $\overline{\mu}_{..}$ al no ser iguales los tamaños de muestra para los distintos niveles del factor. No se trata pues de una descomposición ortogonal de V.

Si expresamos el modelo según (6.8), el término α_i se interpreta como el efecto del nivel i-ésimo del factor sobre la media de la variable respuesta. De hecho, la distribución del estadístico F puede expresarse a partir de estos términos mediante

$$F \sim F_{\mathtt{a}-1,\mathtt{n}-\mathtt{a}}\left(\sigma^{-2}\sum_{i=1}^{a}\alpha_i^2\right).$$

La hipótesis H_0 equivale a $\alpha_1 = \ldots = \alpha_{\mathtt{a}} = 0$. De ser cierta, el estadístico F debe seguir una distribución $F_{\mathtt{a}-1,\mathtt{n}-\mathtt{a}}$, como ya sabíamos.

Las columnas de $\mathtt{X}$, excluyendo el término independiente, se denominan con frecuencia variables ficticias (en rigor, habría que hablar de vectores ficticios). Las puntuaciones obtenidas en las mismas determinan a qué nivel del factor corresponde una observación concreta. Dado que la igualdad de las medias equivale a $\underline{\beta} = 0$, sabemos

por (4.30) que el contraste de igualdad de medias puede dirimirse en función del coeficiente de correlación múltiple de la variable respuesta respecto a las variables ficticias. Además, se sigue de (4.18) que dicho coeficiente no depende de las variables ficticias escogidas y, por lo tanto, no depende de la parametrización concreta que se haya adoptado, cosa que era de esperar.

Normalidad y homocedasticidad

Por otra parte, cuando el supuesto de normalidad no se verifica, podemos justificar los métodos de inferencia anteriores mediante los resultados asintóticos estudiados en el capítulo 2, lo cual requiere el cumplimiento de la condición de Huber (3.37) por parte de la sucesión $(V_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$. En nuestro caso teniendo en cuenta (6.1), se sigue que la condición de Huber equivale a que $\mathtt{n}_i$ converja a infinito para todo $i = 1, \ldots, \mathtt{a}$. En términos prácticos, diríamos el test `F` sigue siendo válido (al menos su nivel de significación es aproximadamente correcto) aunque no se verifique el supuesto de normalidad, siempre y cuando las `a` muestras sean *suficientemente* grandes. Esta condición resulta, desde luego, bastante natural, lo cual es importante teniendo en cuenta que, en estas condiciones, el transformar la variable respuesta tiene una menor expectativa de éxito que en el análisis de regresión, puesto que la misma transformación debe servir para todos los niveles del factor.

Respecto a la violación del supuesto de homocedasticidad, podemos emplear, teniendo en cuenta que nuestro estudio puede entenderse como una análisis de regresión, la técnica de Mínimo Cuadrados Ponderados, estudiada en el capítulo anterior, siempre y cuando se conozca, aproximadamente, la relación entre las distintas varianzas del modelo. También podemos aplicar una transformación del tipo Box-Cox de las variables respuesta con el objeto de conseguir la normalidad y homocedasticidad de los datos transformados. No obstante, hemos de tener presente la existencia de procedimientos alternativos, como el test de Brown-Forsythe o el test no paramétrico de Kruskall-Walllis.

Comparaciones múltiples

Una vez realizado el contraste principal y si el resultado es significativo, conviene conocer entre qué niveles del factor existen diferencias en el valor medio de la variable respuesta. Se trata pues de contrastar hipótesis iniciales del tipo

$$H_0^{ij} : \mu_i = \mu_j, \qquad i \neq j.$$

Estos contrastes reciben el nombre de Comparaciones Múltiples. Dado que la hipótesis inicial anterior equivale a $\mu \in \langle \mathtt{v}_i, \mathtt{v}_j \rangle$, puede contrastarse mediante el test `F`, siendo su estadístico de contraste

$$F = \frac{\mathtt{n}_i(\overline{y}_{i\cdot} - \overline{y}_{i|j\cdot})^2 + \mathtt{n}_j(\overline{y}_{j\cdot} - \overline{y}_{i|j\cdot})^2}{\hat{\sigma}^{2,\mathtt{I}}},$$

donde $\overline{y}_{i|j\cdot}$ denota la media aritmética de las muestras i-ésima y j-ésima combinadas. F se confronta con el cuantil $F^{\alpha}_{1,\mathtt{n}-\mathtt{a}}$. No obstante, puede probarse que el estadístico de contraste se expresa también mediante el cuadrado del estadístico

$$t = \frac{\overline{y}_{i\cdot} - \overline{y}_{j\cdot}}{\sqrt{\frac{1}{\mathtt{n}_i} + \frac{1}{\mathtt{n}_j}}\,\hat{\sigma}_{\mathtt{I}}} \tag{6.9}$$

Por lo tanto, el test equivale a comparar $|t|$ con $t^{\alpha}_{\mathtt{n}-\mathtt{a}}$. Este es el denominado método LSD de Fisher. No obstante, sería deseable que este procedimiento fuera *consistente* con el contraste principal en el sentido de que éste aportara un resultado significativo si, y sólo si, alguna de las comparaciones múltiples lo fuera. Siendo menos ambiciosos, cabría entender, al menos, el conjunto de las $\mathtt{a}(\mathtt{a}-1)$ comparaciones múltiples desde un punto de vista global, de manera que, si las medias fueran todas iguales, la probabilidad de decidir H_1^{ij} para algún par $i \neq j$ sea α. Tal y como está planteado el test anterior, la probabilidad puede ser bastante mayor. Los tres métodos siguientes pretenden solucionar parcialmente el problema:

- **Método de Scheffé:** se basa en la familia de intervalos de confianza simultáneos de Scheffé, estudiada en el capítulo 2. Efectivamente, dado $\alpha \in (0,1)$, para cada vector $d \in V|\langle 1_{\mathtt{n}} \rangle$, se considera el siguiente intervalo para $d'\mu$

  $$d'\hat{\mu} \pm \sqrt{(\mathtt{a}-1)F^{\alpha}_{\mathtt{a}-1,\mathtt{n}-\mathtt{a}}\|d\|^2\hat{\sigma}^{2,\mathtt{I}}}$$

  De esta forma, el test `F` a nivel α para el contraste principal decide H_1 si, y sólo si, el valor 0 queda fuera del intervalo correspondiente a algún vector $d \in V|\langle 1_{\mathtt{n}} \rangle$. Teniendo en cuenta que la hipótesis H_0^{ij} se corresponde con $d'\mu = 0$, siendo $d = \mathtt{n}_i^{-1}\mathtt{v}_i - \mathtt{n}_j^{-1}\mathtt{v}_j$, que pertenece a $V|\langle 1_{\mathtt{n}} \rangle$, podemos considerar la siguiente familia de intervalos de confianza para las diferencias de medias $\mu_i - \mu_j$, $i \neq j$

  $$\overline{y}_{i\cdot} - \overline{y}_{j\cdot} \pm \hat{\sigma}\sqrt{(\mathtt{a}-1)\left(\frac{1}{\mathtt{n}_i} + \frac{1}{\mathtt{n}_j}\right)F^{\alpha}_{\mathtt{a}-1,\mathtt{n}-\mathtt{a}}},$$

  de manera que, si el valor 0 queda fuera de algún intervalo, el test `F` decide necesariamente H_1 en el contraste principal. Por lo tanto, si H_0 es correcta, la

probabilidad de que alguna comparación múltiple resulte significativa (es decir, que el 0 quede fuera de algún intervalo) es menor o igual que $1-\alpha$. Como vemos, el método de Scheffé nos aproxima a la solución buscada, aunque, por desgracia, peca de conservador.

- **Método de Bonferroni:** el conservadurismo del método de Scheffé se explica por el hecho de que la familia de intervalos se construye para que el test F sea consistente, no sólo con las comparaciones múltiples, sino con todos los contrastes de $V|\langle 1_{\mathtt{n}}\rangle$. El método de Bonferroni, basado en la desigualdad del mismo nombre, utiliza el estadístico (6.9) y corrige el valor de α a la hora de realizar las comparaciones, tomando $t^{\alpha/a(\mathtt{a}-1)}_{\mathtt{n}-\mathtt{a}}$ en lugar de $t^{\alpha}_{\mathtt{n}-\mathtt{a}}$. Por lo tanto estamos considerando la familia de intervalos de confianza

$$\overline{y}_{i\cdot}-\overline{y}_{j\cdot}\pm t^{\frac{\alpha}{a(\mathtt{a}-1)}}_{\mathtt{n}-\mathtt{a}}\sqrt{\frac{1}{\mathtt{n}_i}+\frac{1}{\mathtt{n}_j}}\,\hat{\sigma}.$$

  De esta forma, se sigue de (3.46) que, si todas las medias son iguales, la probabilidad de que el 0 quede fuera de algún intervalo es menor o igual que $1-\alpha$. El método sigue siendo pues conservador, aunque en la práctica se utiliza más que el de Scheffé cuando el número de niveles del factor, a, es bajo.

- **Método de Tuckey:** este método permite recuperar el nivel de significación α exacto, pero con la condición de que el diseño sea equilibrado, es decir, que los tamaños de muestras $\mathtt{n}_1,\ldots,\mathtt{n}_{\mathtt{a}}$ sean idénticos. El procedimiento se basa en el distribución del rango estudentizado, definida como sigue: dados k y s enteros positivos, se denota por $q_{k,s}$ la distribución de la variable aleatoria

$$q=\max_{i\neq j}\frac{|Z_i-Z_j|}{\sqrt{U/s}},$$

  calculada a partir de $Z_1,\ldots,Z_k$, variables aleatorias normales e independientes con media 0 y varianza σ^2, y U, variable aleatoria independiente de las anteriores con distribución $\sigma^2\chi^2_s$ central. Pude comprobarse fácilmente que, si en nuestro modelo todas las muestras tienen un mismo tamaño m, entonces

$$\max_{i\neq j}\sqrt{\mathtt{m}}\frac{|\overline{y}_{i\cdot}-\overline{y}_{j\cdot}-(\mu_i-\mu_j)|}{\hat{\sigma}}\sim q_{\mathtt{a},a(\mathtt{m}-1)}. \tag{6.10}$$

  Ello nos induce a considerar la siguiente familia de intervalos de confianza para las diferencias $\mu_i-\mu_j$, donde $i\neq j$,

$$\overline{y}_{i\cdot}-\overline{y}_{j\cdot}\pm q^{\alpha}_{\mathtt{a},\mathtt{m}(\mathtt{a}-1)}\frac{\hat{\sigma}}{\sqrt{\mathtt{m}}}.$$

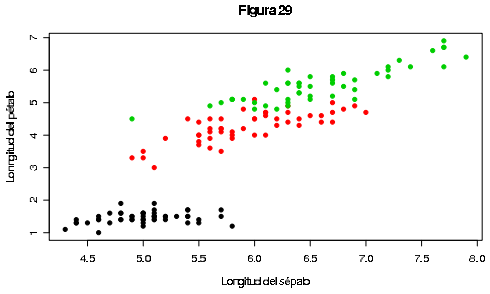
De esta forma, si todas las medias son iguales, la probabilidad de que el valor 0 quede fuera de algún intervalo es exactamente α.

Existen otros métodos para realizar las comparaciones múltiples. Podemos encontrarlos, por ejemplo, en Arnold (1981), capítulo 12. Además, estos procedimientos pueden extenderse, como veremos, al estudio de modelos con más de un factor.

6.2. Análisis de la Covarianza

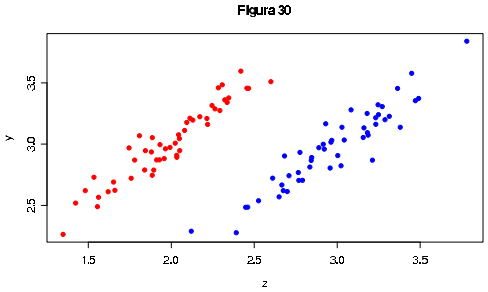
El análisis de la covarianza pude entenderse como una fusión entre los análisis de la varianza y de regresión múltiple. En principio, puede considerarse cualquier modelo del análisis de la varianza y un número indeterminado de variables explicativas. Sin embargo y con el objeto de facilitar la compresión, nos limitaremos a exponer aquí el análisis de la covarianza mezclando, por así decirlo, un diseño completamente aleatorizado con una regresión simple. La extrapolación al caso general puede realizarse sin dificultad.

En un estudio de regresión lineal (simple) puede existir un factor cualitativo, de manera que la relación entre las variables estudiadas puede variar, al menos en principio, en función del nivel del factor. Es decir, que existen diversos grupos y puede considerarse para cada grupo una regresión por separado. El objetivo del investigador puede ser el comparar las rectas de regresión de los distintos grupos. Tal puede ser el caso, por ejemplo, de los datos de Irisdata, donde se mide la anchura y la longitud de los pétalos y sépalos para muestras de tamaño 50 de tres especies de flores: Setosa (negro), Vesicolor (rojo) y Virgínica (verde). Parece razonable pensar que existe una correlación lineal entre la anchura (eje X) y la longitud (eje Y) de los sépalos, pero puede ser que esa relación dependa de la especie escogida. Eso es, efectivamente, lo que recoge el siguiente diagrama de dispersión:



Puede observarse que, al menos aparentemente, la relación entre el incremento de la anchura y de la longitud es similar en las especies virginica y vesicolor, aunque a la longitud de vesicolor habría que añadirle una cantidad adicional constante. Respecto al grupo setosa la cuestión parece más complicada: es posible que incluso la relación entre los incrementos sea diferente. Todas estas hipótesis pueden ser contrastadas, como veremos más adelante.

El estudio se puede contemplar también desde el punto de vista del análisis de la varianza. Por ejemplo, supongamos que nos somos capaces de detectar diferencias significativas entre las medias de una variable respuesta Y medida en dos grupos o niveles de un factor. Sin embargo, existe otra variable Z, denominada *covariable* y correlacionada linealmente con la anterior para los dos grupos, de tal manera que las pendientes de las respectivas rectas de regresión pueden considerarse iguales. Es el caso del ejemplo siguiente:



Si los términos independientes de las rectas son diferentes, como parece apreciarse en la figura, significará que, dado un valor fijo de Z, la variable respuesta toma, por término medio, distintos valores para los dos grupos. En concreto, el grupo de los puntos rojos tiende a tomar valores de Y más altos para un mismo valor Z de la covariable. Es decir, que aunque el factor no afecta a la esperanza de la variable Y, si afecta a la esperanza de Y condicionada al valor de Z.

Añadir a estos ejemplos más factores o más covariables no supone un cambio esencial en el análisis de los mismos. En todo caso y sea cual sea la intención del investigador, este tipo estudios se enmarca en un mismo modelo teórico: el Modelo Lineal. En un caso como el de la figura 30 con $\mathtt{a}$ grupos y una covariable, tendríamos un modelo del tipo

$$Y_{ij} = \eta_i + \gamma \mathbf{z}_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, \mathtt{a},\ j = 1, \ldots, \mathtt{n}_i.$$

El modelo, si se añaden los supuestos de independencia, normalidad y homocedasticidad, puede expresarse mediante

$$Y = \mathtt{X}\beta + \mathcal{E}, \quad \mathcal{E} \sim N_{\mathtt{n}}(0, \sigma^2), \quad \beta \in \mathbb{R}^{a+1},\ \sigma^2 > 0, \tag{6.11}$$

considerando distintas opciones para la matriz $\mathtt{X}$. Por ejemplo, por afinidad a la parametrización (6.4)-(6.5), podemos tomar $\mathtt{X} = (1_{\mathtt{n}}|\mathtt{v}_1, \ldots, \mathtt{v}_{\mathtt{a}-1}, \mathtt{z})$. En ese caso, se verifica

$$\begin{array}{rcl} \beta_0 & = & \eta_{\mathtt{a}} \\ \beta_1 & = & \eta_1 - \eta_{\mathtt{a}} \\ & \vdots & \\ \beta_{\mathtt{a}-1} & = & \eta_{\mathtt{a}-1} - \eta_{\mathtt{a}} \\ \beta_{\mathtt{a}} & = & \gamma \end{array} \tag{6.12}$$

Por lo tanto, la hipótesis $\beta_1 = \ldots = \beta_{\mathtt{a}-1} = 0$ equivale a que los términos independientes de las $\mathtt{a}$ rectas sean idénticos. La hipótesis $\beta_{\mathtt{a}} = 0$ equivale a que la covariable no explique en modo alguna la variablidad de Y, en cuyo caso el diseño utilizado no es el adecuado. Estas hipótesis pueden ser contrastadas fácilmente según hemos visto en el capítulo 4. Por afinidad a la parametrización (6.6)-(6.7), podemos tomar $\mathtt{X} = (1_{\mathtt{n}}|e_1, \ldots, e_{\mathtt{a}-1}, \mathtt{z})$. En ese caso, se tiene que

$$\begin{array}{rcl} \beta_0 & = & \overline{\eta}_{\cdot\cdot} \\ \beta_1 & = & \eta_1 - \overline{\eta}_{\cdot\cdot} \\ & \vdots & \\ \beta_{\mathtt{a}-1} & = & \eta_{\mathtt{a}-1} - \overline{\eta}_{\cdot\cdot} \\ \beta_{\mathtt{a}} & = & \gamma \end{array}, \tag{6.13}$$

siendo $\overline{\eta}_{\cdot\cdot} = \mathtt{n}^{-1} \sum_{i=1}^{a} \mathtt{n}_i \eta_i$. En ese caso, las hipótesis $\beta_1 = \ldots = \beta_{\mathtt{a}-1} = 0$ y $\beta_{\mathtt{a}} = 0$ coinciden con las de la parametrización anterior.

El diseño considerado en la figura 29 es algo más complejo, puesto que la pendiente de la recta puede variar en función del nivel del factor:

$$Y_{ij} = \eta_i + \gamma_i \mathtt{z}_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, \mathtt{a},\ j = 1, \ldots, \mathtt{n}_i.$$

Para formalizarlo consideramos la matriz $\tilde{\mathtt{X}}$ que se obtiene al añadir a la matriz $\mathtt{X}$ anterior las columnas $\{\mathtt{v}_1 * \mathtt{z}, \ldots, \mathtt{v}_{\mathtt{a}-1} * \mathtt{z}\}$, para la parametrización (6.12), o las columnas $\{e_1 * \mathtt{z}, \ldots, e_{\mathtt{a}-1} * z\}$ para (6.13). El signo * denota el producto de los vectores componente a componente. Las nuevas columnas añadidas se denomina interacciones. De esta forma, tendremos el modelo

$$Y = \tilde{\mathtt{X}}\beta + \mathcal{E}, \quad \mathcal{E} \sim N_{\mathtt{n}}(0, \sigma^2), \quad \beta \in \mathbb{R}^{2a},\ \sigma^2 > 0. \tag{6.14}$$

Los parámetros $\beta_0, \ldots, \beta_{\mathtt{a}-1}$ se interpretan como antes en ambas parametrizaciones. No obstante, en (6.12) se tiene que $\beta_{\mathtt{a}} = \gamma_{\mathtt{a}}$, y $\beta_{\mathtt{a}+i} = \gamma_i - \gamma_{\mathtt{a}}$, para $i = 1, \ldots, \mathtt{a}-1$. En (6.13) se verifica que $\beta_{\mathtt{a}} = \overline{\gamma}_{\cdot\cdot}$, y $\beta_{\mathtt{a}+i} = \gamma_i - \overline{\gamma}_{\cdot\cdot}$, para $i = 1, \ldots, \mathtt{a}-1$, siendo $\overline{\gamma}_{\cdot\cdot} = \mathtt{n}^{-1} \sum_{i=1}^{a} \mathtt{n}_i \gamma_i$. En todo caso, la hipótesis $\beta_{\mathtt{a}+1} = \ldots = \beta_{2\mathtt{a}-1} = 0$ significa la igualdad de las pendientes, lo cual se traduciría en un modelo del tipo (6.11). De hecho, esta hipótesis puede contrastarse antes de considerar dicho modelo. Si el resultado es no significativo, es costumbre habitual contrastar la hipótesis inicial de igualdad de términos independiente en el modelo reducido (6.11).

Como podemos ver, las principales hipótesis a contrastar no dependen del tipo de parametrización escogida, lo cual ocurre porque dichas hipótesis verifican la condición (9.44), es decir, que son contrastables.

En definitiva, el diseño completamente aleatorizado se resuelve introduciendo variables ficticias, que indican a qué nivel del factor pertenece la unidad experimental; en el problema de regresión lineal se introducen variables explicativas (covariables); en general, ambos tipos de variables, las ficticias y las covariables, pueden combinarse dando lugar a un análisis de la covarianza. Pueden considerarse, incluso, productos entre ambas, lo cual posibilita la existencia de interacción entre el factor y las covariables. Como ya hemos comentado, esto puede llevarse a cabo de igual modo en modelos multifactoriales, donde pueden considerarse, a su vez, productos o interacciones entre los factores e, incluso, interacciones entre las interacciones.

6.3. El test de Student como caso particular

En esta sección abordaremos el estudio de dos situaciones particulares, las más sencillas, del diseño completamente aleatorizado, concretamente, los casos $a = 1$ y $a = 2$, que se corresponden con el análisis de las medias de una y dos muestras de distribuciones normales. Ambos estudios se resuelven, como bien sabemos, mediante el denominado test de Student. Veremos cómo al aplicar las técnicas propias del modelo lineal obtenemos dicho test como caso particular del test `F`.

En primer lugar, analizaremos el caso $a = 1$, es decir, consideramos $Y_1, \ldots, Y_{\mathtt{n}}$ una muestra aleatoria simple de una distribución $N(\nu, \sigma^2)$, con media y varianza desconocidas. En ese caso, si se denota $Y = (Y_1, \ldots, Y_{\mathtt{n}})'$ y $\mu = (\nu, \ldots, \nu)'$, el modelo es el siguiente

$$Y \sim N_{\mathtt{n}}(\mu, \sigma^2 \mathtt{Id}), \qquad \mu \in \langle 1_{\mathtt{n}} \rangle, \ \sigma^2 > 0.$$

Se sigue del teorema 3.7 que los EMV de μ y σ^2 son, respectivamente, $(\overline{y}, \ldots, \overline{y})'$ y s_y^2. Del corolario 3.6 se sigue que $(\overline{y}, \ldots, \overline{y})'$ y $(\mathtt{n}-1)^{-1}\mathtt{n} s_y^2$ so los EIMV de μ y σ^2, respectivamente. Además, de la proposición 3.4 se sigue que los estadísticos $\overline{y}$ y s_y^2

son independientes, lo cual constituye, precisamente, la tesis del conocido teorema de Fisher.

Para resolver el contraste de la hipótesis inicial $H_0 : \nu = 0$, podemos hacer uso de test F, teniendo en cuenta que H_0 equivale a $\mu \in W = \{0\}$. En ese caso ,se obtiene sin dificultad

$$F = \frac{\mathtt{n}\overline{y}}{s_y^2} = \left(\frac{\overline{y}}{s_y/\sqrt{\mathtt{n}}}\right)^2,$$

que debe compararse con $F^{\alpha}_{1,\mathtt{n}-1}$, lo cual equivale a comparar con $t^{\alpha}_{\mathtt{n}-1}$ el estadístico de contraste

$$t = \frac{\overline{y}}{s_y/\sqrt{\mathtt{n}}}.$$

En la práctica, suelen considerarse contrastes de hipótesis iniciales del tipo $H_0 : \nu = \nu_0$, para algún valor ν_0 conocido. Este problema se resuelve considerando el modelo trasladado asociado a $Y_i^* = Y_i - \nu_0$, $i = 1, \ldots, \mathtt{n}$. En ese caso, el test F a nivel α consiste en comparar con $t^{\alpha}_{\mathtt{n}-1}$ el estadístico de contraste

$$t = \frac{\overline{y} - \nu_0}{s_y/\sqrt{\mathtt{n}}} \tag{6.15}$$

Éste es el denominado test de Student para una muestra. Realmente, no era estrictamente necesario recurrir al Modelo Lineal para llegar a este test, pero el hecho de obtenerlo mediante estas técnicas otorga mayor consistencia a nuestra teoría. Lo mismo puede decirse del intervalo de confianza para la media ν que se deriva de la región (3.12).

Respecto al supuesto de normalidad, la condición de Huber, que garantiza, en virtud del corolario 3.21, la validez asintótica del test de Student, es completamente vacua pues equivale a que n converja a infinito. Por lo tanto, para muestras suficientemente grandes podemos prescindir del supuesto de normalidad en el contraste de la media. A esta conclusión se podría haber llegado sin necesidad de aplicar el corolario 3.21. Hubiera bastado considerar resultados más básicos, como son la versión (9.67) del Terema Central del Límite, junto con el método de los momentos (teorema 9.24). Efectivamente, se verifica que, cuando n tiende a infinito, se verifica

$$\frac{\overline{y} - \nu}{\sigma/\sqrt{\mathtt{n}}} \xrightarrow{d} N(0,1), \qquad s_y \xrightarrow{P} \sigma.$$

Aplicando el teorema 9.21, se obtiene la convergencia del estadístico t de (6.15) a la distribución $N(0,1)$ en el caso nulo[1].

[1]Tener en cuenta también que la distribución t-Student con $\mathtt{n} - 1$ grados de libertad converge igualmente a la distribución $N(0,1)$ cuando n tiende a infinito.

Estudiamos, a continuación, el caso $a = 2$, es decir consideramos dos muestras aleatoria simples independientes, $Y_{11}, \ldots, Y_{1\mathtt{n}_1}$ y $Y_{21}, \ldots, Y_{2\mathtt{n}_2}$, correspondientes repectivamente a sendas distribuciones normales con idéntica varianza y medias μ_1 y μ_2 desconocidas. Si componemos las observaciones y las medias en sendos vectores Y y μ de dimensión $\mathtt{n} = \mathtt{n}_1 + \mathtt{n}_2$, obtenemos el modelo

$$Y \sim N_{\mathtt{n}}(\mu, \sigma^2 \mathtt{Id}), \qquad \mu \in \langle \mathtt{v}_1, \mathtt{v}_2 \rangle, \ \sigma^2 > 0.$$

Aplicando los resultados obtenidos en la primera sección obtenemos los EIMV de μ y σ^2 siguientes

$$\mu = \overline{y}_1 \cdot \mathtt{v}_1 + \overline{y}_2 \cdot \mathtt{v}_2, \qquad s_c^2 = \frac{(\mathtt{n}_1 - 1)s_{1,I}^2 + (\mathtt{n}_2 - 1)s_{2,I}^2}{\mathtt{n} - 2},$$

donde, para cada $j = 1, 2$, $\overline{y}_j$ y $s_{j,I}^2$ denotan los EIMV que se obtienen para cada muestra por separada según el modelo estudiado anteriormente. Podemos obtener de manera trivial una región de confianza para μ a partir de (3.12). La hipótesis inicial cuyo contraste puede resultar, en principio, más interesante, es $H_0 : \nu_1 = \nu_2$, que equivale a $\mu \in \langle 1_{\mathtt{n}} \rangle$. En ese caso, el test $\mathtt{F}$ a nivel α consiste en compara con $F_{1,\mathtt{n}-2}^{\alpha}$ el estadístico de contraste

$$F = \frac{\sum_{i=1}^{2} \mathtt{n}_i (\overline{y}_i - \overline{y})^2}{s_c^2},$$

donde $\overline{y}$ denota la media aritmética de los $\mathtt{n}$ datos. Teniendo en cuenta que $\overline{y} = \mathtt{n}^{-1}(\mathtt{n}_1 \overline{y}_1 + \mathtt{n}_2 \overline{y}_2)$, se deduce que el test $\mathtt{F}$ equivale a comparar co $t_{\mathtt{n}-2}^{\alpha}$ el estadístico de contraste

$$t = \frac{\overline{y}_1 - \overline{y}_2}{s_c \sqrt{\frac{1}{\mathtt{n}_1} + \frac{1}{\mathtt{n}_2}}}. \tag{6.16}$$

Nuevamente, estamos hablando pues del test de Student, en esta ocasión para dos muestras. Respecto a la validez asintótica del mismo presciendiendo del supuesto de normalidad, la condición de Huber se traduce, en este caso, a que tanto $\mathtt{n}_1$ como $\mathtt{n}_2$ converjan a infinito. Puede probarse también la validez asintótica haciendo uso del teorema 9.24 junto con (9.67). Veamos ahora cómo podemos obviar el supuesto de homocedasticidad.

Supongamos que $Y_{11}, \ldots Y_{1\mathtt{n}_1}$ es una muestra aleatoria simple de una distribución de media μ_1 y varianza σ_1^2, y que $Y_{21}, \ldots, Y_{2\mathtt{n}_2}$ es una muestra aleatoria simple, independiente de la anterior, de una distribución de media μ_2 y varianza σ_2^2. Supongamos que los tamaños muestrales convergen a infinito. En ese caso, podemos enunciar el siguiente resultado asintótico

Proposición 6.1.
En las condiciones anteriores, si $\mu_1 = \mu_2$ y $\frac{\mathtt{n}_1}{\mathtt{n}_2} \to 1$, se verifica que el estadístico (6.16) converge en distribución a $N(0,1)$.

Demostración.
Dado que $\mathtt{n}_1/\mathtt{n}_2 \to \infty$ y teniendo en cuenta el teorema 9.21 junto con (9.67), se verifica que

$$\sqrt{\mathtt{n}_1}(\overline{y}_1 - \mu_1) \xrightarrow{d} N(0, \sigma_1^2), \qquad \sqrt{\mathtt{n}_1}(\overline{y}_2 - \mu_2) \xrightarrow{d} N(0, \sigma_2^2). \tag{6.17}$$

Por lo tanto, si $\mu_1 = \mu_2$, se sigue nuevamente del teorema 9.21 que

$$\tau_{\mathtt{n}_1,\mathtt{n}_2} = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\frac{\sigma_1^2}{\mathtt{n}_1} + \frac{\sigma_2^2}{\mathtt{n}_2}}} \xrightarrow{d} N(0,1).$$

Por otra parte, el estadístico t de (6.16) puede expresarse mediante

$$t = \tau_{\mathtt{n}_1,\mathtt{n}_2} \cdot \frac{\sqrt{\frac{\mathtt{n}_1\mathtt{n}_2}{\mathtt{n}_1+\mathtt{n}_2} \cdot \left(\frac{\sigma_1^2}{\mathtt{n}_1} + \frac{\sigma_2^2}{\mathtt{n}_2}\right)}}{s_c}$$

Dado que, en virtud del teorema 9.24, s_i^2 converge en probabilidad a σ_i^2, para $i = 1, 2$, s_c^2 converge e probabilidad a $(\sigma_1^2 + \sigma_2^2)/2$ y, por lo tanto, el segundo factor converge en probabilidad a 1. Luego, por el teorema 9.21, se concluye.

∎

El resultado anterior garantiza la validez asintótica del test de Student aunque no se verifiquen los supuestos de normalidad y homocedasticidad, siempre y cuando $\mathtt{n}_1/\mathtt{n}_2$ converja a 1. En términos prácticos, diremos que el test puede considerarse válido cuando los tamaños de muestra sean lo suficiente grandes y los suficientemente parecidos. Esta forma de proceder se extrapola a cualquier diseño completamente aleatorizado. Es decir, que se procura que las muestras consideradas para cada nivel del factor sean lo mayores posibles y que no exista una gran desproporción entre sus tamaños. De todas formas, en el caso de dos muestras, contamos con procedimiento alternativos clásicos, de sobras conocidos, para el caso heterocedástico y el caso no normal, como son, respectivamente el test de Welch y el test no paramétrico de Mann-Whitney.

6.4. Diseño bifactorial equilibrado

En esta sección se estudiará la influencia de dos factores cualitativos, f_A con $\mathtt{a}$ niveles y f_B con $\mathtt{b}$ niveles, en la media de una variable respuesta y. Para ello,

consideraremos $\mathtt{a}\cdot\mathtt{b}$ muestras aleatorias simples, cada una de ellas correspondiendo a la combinación entre un determinado nivel del factor f_A, i, con otro del factor f_B, j. Se supondrá en todo caso que las $\mathtt{ab}$ muestras son del mismo tamaño, que se denota por $\mathtt{m}$. Por lo tanto el número total de datos es $\mathtt{n}=\mathtt{abm}$. El diseño puede representarse, esquemáticamente, como sigue:

		Factor B	
Factor A	$Y_{111},\ldots,Y_{11\mathtt{m}}$	.....	$Y_{1\mathtt{b}1},\ldots,Y_{1\mathtt{bm}}$
	$\vdots$		$\vdots$
	$Y_{\mathtt{a}11},\ldots,Y_{\mathtt{a}1\mathtt{m}}$	.....	$Y_{\mathtt{ab}1},\ldots,Y_{\mathtt{abm}}$

De esta manera, podemos asignar a la muestra correspondiente a los niveles i-ésimo y j-ésimo de los factores A y B, respectivamente, las coordenadas (i,j), que indica una celda de la cuadrícula. Una tercera coordenada, k, indicará la posición del dato en la celda corespondiente. Se supondrá, además, que todas las muestras son independientes y provienen de distribuciones normales con idéntica varianza. Por lo tanto, el modelo puede expresarse así:

$$Y_{ijk}=\mu_{ij}+\varepsilon_{ijk},\quad \varepsilon_{ijk}\sim N(0,\sigma^2)\text{ independientes,} \tag{6.18}$$

donde $i=1,\ldots,\mathtt{a}$, $j=1,\ldots,\mathtt{b}$ y $k=1,\ldots\mathtt{m}$. Si componemos todas las observaciones de las variable repuesta, ordenando las muestras por filas, obtenemos el vector aleatorio $\mathtt{abm}$-dimensional $Y=(Y_{111},\ldots,Y_{11\mathtt{m}},Y_{121},\ldots,Y_{\mathtt{abm}})'$, de media μ. Para cada celda (i,j) de la cuadrícula se considera el vector $\mathtt{v}_{ij}$ de $\mathbb{R}^{\mathtt{n}}$ cuyas componentes son todas nulas salvo las $\mathtt{m}$ correspondientes a la misma, que valen 1. Así, si V denota el subespacio $\mathtt{ab}$ dimensional del $\mathbb{R}^{\mathtt{n}}$ generado por los vectores $\mathtt{v}_{ij}$, para $i=1,\ldots,\mathtt{a}$ y $j=1,\ldots,\mathtt{b}$, el modelo puede expresarse mediante

$$Y=\mu+\mathcal{E},\quad \mathcal{E}\sim N_{\mathtt{n}}(0,\sigma^2\mathtt{Id}),\qquad \mu\in V,\ \sigma^2>0. \tag{6.19}$$

Así pues, se trata de un modelo lineal normal, que coincide con el que correspondería a un diseño completamente aleatorizado, es decir, co un único factor, pero con $\mathtt{ab}$ niveles. Por lo tanto, el problema de estimación de μ y σ^2 está ya resuelto: el valor correspondiente a las coordenadas ijk del estimador de μ, P_VY, es, para todo k de 1 a $\mathtt{m}$,

$$\overline{y}_{ij\cdot}=\mathtt{m}^{-1}\sum_{s=1}^{\mathtt{m}}Y_{ijs}$$

Por lo tanto, el EIMV de σ^2 es

$$\hat{\sigma}^{2,\mathrm{I}}=\frac{1}{\mathtt{ab}(\mathtt{m}-1)}\sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}\sum_{k=1}^{\mathtt{m}}(Y_{ijk}-\overline{y}_{ij\cdot})^2 \tag{6.20}$$

Nótese que, si obviamos el factor f_B, los cual equivale a agrupar las celdas por filas para obtener $\mathtt{a}$ muestras de tamaño $\mathtt{bm}$, el modelo resultando correspondería a un diseño completamente aleatorizado con $\mathtt{a}$ niveles. En ese caso, μ pertenece al subespacio $V_1 = \langle \mathtt{v}_{1\cdot}, \ldots, \mathtt{v}_{\mathtt{a}\cdot} \rangle$, siendo $\mathtt{v}_{i\cdot} = \sum_{j=1}^{\mathtt{b}} \mathtt{v}_{ij}$, $i = 1, \ldots, \mathtt{a}$. El estimador de μ en este modelo, $P_{V_1} Y$, posee en la posición ijk el valor

$$\overline{y}_{i\cdot\cdot} = (b\mathtt{m})^{-1} \sum_{j=1}^{\mathtt{b}} \sum_{s=1}^{\mathtt{m}} Y_{ijs}$$

Análogamente, si ignoramos el factor f_A, es decir, si agrupamos por columnas, obtenemos un modelo donde μ pertenece a $V_2 = \langle \mathtt{v}_{\cdot 1}, \ldots, \mathtt{v}_{\cdot b} \rangle$, siendo $\mathtt{v}_{\cdot j} = \sum_{i=1}^{\mathtt{a}} \mathtt{v}_{ij}$, $j = 1, \ldots, \mathtt{b}$. Igualmente, el estimador de μ para este este modelo, P_{V_2}, posee en la posición ijk el valor

$$\overline{y}_{\cdot j\cdot} = (\mathtt{am})^{-1} \sum_{i=1}^{\mathtt{a}} \sum_{s=1}^{\mathtt{m}} Y_{ijs}$$

Por último, si ignoramos ambos factores tendremos una única muestra aleatoria simple de tamaño $\mathtt{n}$, en cuyo caso el estimador de la media, $P_{\langle 1_{\mathtt{n}} \rangle} Y$, es el vector de $\mathbb{R}^{\mathtt{n}}$ cuyas componentes son todas iguales a

$$\overline{y}_{\cdots} = \sum_{i=1}^{\mathtt{a}} \sum_{j=1}^{\mathtt{b}} \sum_{k=1}^{\mathtt{m}} Y_{ijk}$$

Consideremos a continuación las medias aritméticas siguientes:

$$\begin{array}{rcll} \overline{\mu}_{\cdot\cdot} & = & (\mathtt{ab})^{-1} \sum_{i=1}^{\mathtt{a}} \sum_{j=1}^{\mathtt{b}} \mu_{ij} \ ; & \\ \overline{\mu}_{i\cdot} & = & \mathtt{b}^{-1} \sum_{j=1}^{\mathtt{b}} \mu_{ij} \ , & i = 1, \ldots, \mathtt{a}; \\ \overline{\mu}_{\cdot j} & = & \mathtt{a}^{-1} \sum_{i=1}^{\mathtt{a}} \mu_{ij} \ , & j = 1, \ldots, \mathtt{b}. \end{array}$$

Definimos entonces los siguientes parámetros:

$$\begin{array}{rcll} \theta & = & \overline{\mu}_{\cdot\cdot} \ ; & \\ \alpha_i & = & \overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot\cdot} \ , & i = 1, \ldots, \mathtt{a}; \\ \beta_j & = & \overline{\mu}_{\cdot j} - \overline{\mu}_{\cdot\cdot} \ , & j = 1, \ldots, \mathtt{b}; \\ (\alpha\beta)_{ij} & = & \mu_{ij} - \overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot j} + \overline{\mu}_{\cdot\cdot} \ , & i = 1, \ldots, \mathtt{a}, \ j = 1, \ldots, \mathtt{b}. \end{array}$$

Puede comprobarse, trivialmente, que se verifican las siguientes restricciones

$$\sum_{i=1}^{\mathtt{a}} \alpha_i = \sum_{j=1}^{\mathtt{b}} \beta_j = 0, \quad , \sum_{j=1}^{\mathtt{b}} (\alpha\beta)_{ij} = 0, \ i = 1, \ldots a, \quad \sum_{i=1}^{\mathtt{a}} (\alpha\beta)_{ij} = 0, \ j = 1, \ldots, \mathtt{b}. \tag{6.21}$$

De esta forma, (6.18) es equivalente al modelo

$$Y_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma^2), \tag{6.22}$$

con las restricciones expresadas en (6.21). Por lo tanto, estamos expresando el modelo bifactorial de manera análoga a como expresábamos el unifactorial en (6.8). Veremos a continuación que estos nuevos parámetros corresponden a una determinada descomposición de V en subespacios ortogonales:

Proposición 6.2.

La siguiente descomposición es ortogonal

$$V = \langle 1_{\mathtt{n}} \rangle \ \oplus \ V_1|\langle 1_{\mathtt{n}} \rangle \ \oplus \ V_2|\langle 1_{\mathtt{n}} \rangle \ \oplus \ V|(V_1 \oplus V_2) \tag{6.23}$$

Demostración.

Debemos probar únicamente que que $V_1|\langle 1_{\mathtt{n}} \rangle \perp V_2|\langle 1_{\mathtt{n}} \rangle$. Ciertamente, dos vectores cualesquiera, g_1 y g_2, pertenecientes a $V_1|\langle 1_{\mathtt{n}} \rangle$ y $V_2|\langle 1_{\mathtt{n}} \rangle$, respectivamente, pueden expresarse mediante $g_1 = \sum_{i=1}^{\mathtt{a}} x_i \mathtt{v}_{i\cdot}$ y $g_2 = \sum_{j=1}^{\mathtt{b}} y_j \mathtt{v}_{\cdot j}$. Al ser ortogonales a $1_{\mathtt{n}}$, se verifica que $\sum_{i=1}^{\mathtt{a}} x_i = \sum_{j=1}^{\mathtt{b}} y_j = 0$. En consecuencia,

$$g_1 = \sum_{i=1}^{\mathtt{a}-1} x_i(\mathtt{v}_{i\cdot} - \mathtt{v}_{\mathtt{a}\cdot}), \quad g_2 = \sum_{j=1}^{\mathtt{b}-1} y_j(\mathtt{v}_{\cdot j} - \mathtt{v}_{\cdot b})$$

Así pues,

$$g_1 * g_2 = \sum_{i=1}^{\mathtt{a}-1}\sum_{j=1}^{\mathtt{b}-1} x_i y_j(\mathtt{v}_{ij} - \mathtt{v}_{\mathtt{a}j} - \mathtt{v}_{ib} + \mathtt{v}_{\mathtt{ab}})$$

y, por lo tanto,

$$\langle g_1, g_2 \rangle = \sum_{i=1}^{\mathtt{a}-1}\sum_{j=1}^{\mathtt{b}-1} x_i y_j(\mathtt{m} - \mathtt{m} - \mathtt{m} + \mathtt{m}) = 0$$

∎

En todo caso, se verifica que

$$P_{\langle 1_{\mathtt{n}} \rangle}\mu = \overline{\mu}_{\cdot\cdot} \cdot 1_{\mathtt{n}}, \quad P_{V_1}\mu = \sum_{i=1}^{\mathtt{a}} \overline{\mu}_{i\cdot} \cdot \mathtt{v}_{i\cdot}, \quad P_{V_2}\mu = \sum_{j=1}^{\mathtt{b}} \overline{\mu}_{\cdot j} \cdot \mathtt{v}_{\cdot j}$$

Teniendo en cuenta que

$$P_{V_1|\langle 1_{\mathtt{n}} \rangle} = P_{V_1} - P_{\langle 1_{\mathtt{n}} \rangle}, \qquad P_{V_2|\langle 1_{\mathtt{n}} \rangle} = P_{V_2} - P_{\langle 1_{\mathtt{n}} \rangle} \tag{6.24}$$

y que

$$P_{V|(V_1\oplus V_2)} = P_V - (P_{\langle 1_{\mathtt{n}}\rangle} + P_{V_1|\langle 1_{\mathtt{n}}\rangle} + P_{V_2|\langle 1_{\mathtt{n}}\rangle}), \tag{6.25}$$

se sigue que ,

$$P_{V_1|\langle 1_{\mathtt{n}}\rangle}\mu = \sum_{i=1}^{\mathtt{a}} \alpha_i \mathtt{v}_{i\cdot}, \quad P_{V_2|\langle 1_{\mathtt{n}}\rangle}\mu = \sum_{j=1}^{\mathtt{b}} \beta_j \mathtt{v}_{\cdot j}, \quad P_{V|(V_1\oplus V_2)}\mu = \sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}} (\alpha\beta)_{ij}\mathtt{v}_{ij},$$

con los parámetros α_i, β_j y $(\alpha\beta)_{ij}$ expresados en el modelo (6.22) y, en consecuencia, con las restricciones expresadas en (6.21). En ese sentido decimos que la parametrización (6.22) obedecen a la descomposición ortogonal (6.23).

Obviamente, al igual que sucede en el diseño completamente aleatorizado cuando se considera la matriz (6.6), este diseño corresponde un modelo de regresión lineal múltiple a partir de cierta matriz $\mathtt{X}$. Se propone como ejercicio encontrar una forma concreta para la misma. Por otra parte, podemos considerar también la descomposición ortogonal del vector aleatorio $P_V Y$ en las proyecciones sobre los distintos subespacios.

$$P_V Y = P_{\langle 1_{\mathtt{n}}\rangle} Y + P_{V_1|\langle 1_{\mathtt{n}}\rangle} Y + P_{V_2|\langle 1_{\mathtt{n}}\rangle} + P_{V|(V_1\oplus V_2)} Y$$

Por un razonamiento análogo al anterior, la suma queda como sigue

$$P_V Y = \overline{y}_{\cdots} 1_{\mathtt{n}} + \sum_{i=1}^{\mathtt{a}} (\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots})\mathtt{v}_{i\cdot} + \sum_{j=1}^{\mathtt{b}} (\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdots})\mathtt{v}_{\cdot j} + \sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}} (\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdots})\mathtt{v}_{ij} \tag{6.26}$$

Además,

$$\|P_{V_1|\langle 1_{\mathtt{n}}\rangle} Y\|^2 = \mathtt{mb} \sum_{i=1}^{\mathtt{a}} (\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots})^2, \tag{6.27}$$

$$\|P_{V_2|\langle 1_{\mathtt{n}}\rangle} Y\|^2 = \mathtt{ma} \sum_{j=1}^{\mathtt{b}} (\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdots})^2, \tag{6.28}$$

$$\|P_{V|(V_1\oplus V_2)} Y\|^2 = \mathtt{m} \sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}} \left(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdots}\right)^2. \tag{6.29}$$

Estos resultados serán de gran utilidad a la hora e contrastar las siguientes hipótesis iniciales:

$$\begin{aligned} H_0^A &: \alpha_1 = \ldots = \alpha_{\mathtt{a}} = 0 \\ H_0^B &: \beta_1 = \ldots = \beta_{\mathtt{b}} = 0 \\ H_0^{AB} &: (\alpha\beta)_{11} = \ldots = (\alpha\beta)_{\mathtt{ab}} = 0 \end{aligned}$$

La hipótesis inicial H_0^A equivale a que μ pertenezca al subespacio

$$W = \langle 1_{\mathtt{n}} \rangle \oplus V_2 | \langle 1_{\mathtt{n}} \rangle \oplus V | (V_1 \oplus V_2)$$

Por lo tanto, teniendo en cuenta (3.23), (6.27) y (6.20), el test F a nivel α para contrastar la hipótesis inicial H_0^A consiste en comparar con $F^{\alpha}_{\mathtt{a}-1,\mathtt{ab}(\mathtt{m}-1)}$ el estadístico

$$F_A = \frac{\frac{1}{\mathtt{a}-1}\,\mathtt{mb}\sum_{i=1}^{\mathtt{a}}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot\cdot\cdot})^2}{\frac{1}{\mathtt{ab}(\mathtt{m}-1)}\sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}\sum_{k=1}^{\mathtt{m}}(Y_{ijk} - \overline{y}_{ij\cdot})^2}.$$

Igualmente, para contrastar H_0^B se compara con $F^{\alpha}_{\mathtt{b}-1,\mathtt{ab}(\mathtt{m}-1)}$ el estadístico

$$F_B = \frac{\frac{1}{\mathtt{b}-1}\,\mathtt{ma}\sum_{j=1}^{\mathtt{b}}(\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdot\cdot\cdot})^2}{\frac{1}{\mathtt{ab}(\mathtt{m}-1)}\sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}\sum_{k=1}^{\mathtt{m}}(Y_{ijk} - \overline{y}_{ij\cdot})^2}.$$

Por último, para contrastar H_0^{AB}, se compara con $F^{\alpha}_{(\mathtt{a}-1)(\mathtt{b}-1),\mathtt{ab}(\mathtt{m}-1)}$ el estadístico

$$F_{AB} = \frac{\frac{1}{(\mathtt{a}-1)(\mathtt{b}-1)}\,\mathtt{m}\sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}\left(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdot\cdot\cdot}\right)^2}{\frac{1}{\mathtt{ab}(\mathtt{m}-1)}\sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}\sum_{k=1}^{\mathtt{m}}(Y_{ijk} - \overline{y}_{ij\cdot})^2}.$$

Al igual que en el diseño completamente aleatorizado, podemos proceder a realizar distintas comparaciones múltiples. En primer lugar, podemos buscar una familia de intervalos de confianza, a a ser posible simultáneos, para los parámetros $\{\alpha_i - \alpha_{i'} : i \neq i'\}$. Hemos de tener en cuenta que, para cada par $i \neq i'$, se verifica

$$\overline{y}_{\mathtt{1}\cdot\cdot} - \overline{y}_{i'\cdot\cdot} \sim N(\alpha_i - \alpha_{i'}, (\mathtt{mb})^{-1}2\sigma^2). \tag{6.30}$$

En consecuencia, la familia intervalos de confianza por el método de Bonferroni es la siguiente

$$\alpha_i - \alpha_{i'} \in \overline{y}_{\mathtt{1}\cdot\cdot} - \overline{y}_{i'\cdot\cdot} \pm t^{\alpha/a(\mathtt{a}-1)}_{\mathtt{ab}(\mathtt{m}-1)}\hat{\sigma}_{\mathtt{I}}\sqrt{2(\mathtt{mb})^{-1}}, \quad i \neq i'. \tag{6.31}$$

Un razonamiento análogo conduce a la familia de intervalos de confianza simultáneos según el método de Tuckey:

$$\alpha_i - \alpha_{i'} \in \overline{y}_{\mathtt{1}\cdot\cdot} - \overline{y}_{i'\cdot\cdot} \pm q^{\alpha}_{\mathtt{a},\mathtt{ab}(\mathtt{m}-1)}\hat{\sigma}_{\mathtt{I}}\sqrt{(\mathtt{mb})^{-1}}, \quad i \neq i'. \tag{6.32}$$

Por último, se sigue del teorema 3.24 que la familia de intervalos Scheffé para estos contrates es

$$\alpha_i - \alpha_{i'} \in \overline{y}_{\mathtt{1}\cdot\cdot} - \overline{y}_{i'\cdot\cdot} \pm \hat{\sigma}_{\mathtt{I}}\sqrt{2(\mathtt{a}-1)(\mathtt{mb})^{-1}F^{\alpha}_{\mathtt{a}-1,\mathtt{ab}(\mathtt{m}-1)}}, \quad i \neq i'. \tag{6.33}$$

De manera completamente análoga (se deja como ejercicio), podemos construir las familias de intervalos de confianza de Bonferroni, Tuckey y Scheffé para el conjunto de $\{\beta_j - \beta_{j'} : \ j \neq j'\}$.

Los parámetros $(\alpha\beta)_{ij}$, $i = 1, \ldots, \mathtt{a}$, $j = 1, \ldots, \mathtt{b}$, se denominan interacciones. Si son todas nulas, es decir, si la hipótesis H_0^{AB} es verdadera, entonces estaremos hablando del siguiente modelo reducido

$$Y_{ijk} = \theta + \alpha_i + \beta_j + \varepsilon_{ijk}, \quad , \varepsilon_{ijk} \sim N(0, \sigma^2), \ \sum_{i=1}^{\mathtt{a}} \alpha_i = \sum_{j=1}^{\mathtt{b}} \beta_j = 0. \tag{6.34}$$

Este modelo se denomina modelo bifactorial sin interacción, y se caracteriza por cuplir la siguiente propiedad: para todo $i \neq i$ y $j \neq j'$, se verifica

$$\mu_{ij} - \mu_{ij'} = \mu_{i'j} - \mu_{i'j'}$$

Es decir, la variaciones de la media entre los distintos niveles del factor A no dependen del nivel del factor B considerado y viceversa. En este modelo, se verifica que μ pertenece al subespacio $V_1 \oplus V_2$. En consecuencia, se sigue de (6.25) que el EIMV de σ^2 es el siguiente:

$$\hat{\sigma}^{2,\mathrm{I}} = \frac{1}{\mathtt{n} - \mathtt{a} + \mathtt{b} - 1} \sum_{i=1}^{\mathtt{a}} \sum_{j=1}^{\mathtt{b}} \sum_{k=1}^{\mathtt{m}} \left(Y_{ijk} - \left(\overline{y}_{i\cdot\cdot} + \overline{y}_{\cdot j\cdot} - \overline{y}_{\cdots} \right) \right)^2 \tag{6.35}$$

Para contrastar las hipótesis iniciales

$$H_0^{A*} : \alpha_1 = \ldots = \alpha_{\mathtt{a}} = 0$$

$$H_0^{B*} : \beta_1 = \ldots = \beta_{\mathtt{b}} = 0$$

se manejan los mismos tests que se utilizan para contrastar en el modelo con interacción las hipótesis iniciales H_0^A y H_0^B, respectivamente, con la salvedad de que,en ambos casos, debe aparecer en el denominador de F la expresión (6.35) en lugar de (6.20), que se comparará con el cuantil $F^{\alpha}_{\mathtt{a}-1,\mathtt{n}-\mathtt{a}-\mathtt{ab}+1}$. Las familia de intervalos de confianza para $\{\alpha_i - \alpha_{i'} : \ i \neq i'\}$ y $\{\beta_j - \beta_{j'} : \ j \neq j'\}$ son idénticas a las del modelo con interacción salvo en los grados de libertad de los cuantiles utilizados. Concretamente, las familias de Bonferroni, Tuckey y Scheffé, se construirán, respectivamente, a partir de los cuantiles

$$t^{\alpha/a(\mathtt{a}-1)}_{\mathtt{n}-\mathtt{a}-\mathtt{b}+1}, \quad q^{\alpha}_{\mathtt{a},\mathtt{n}-\mathtt{a}-\mathtt{b}+1}, \quad F^{\alpha}_{\mathtt{a}-1,\mathtt{n}-\mathtt{a}-\mathtt{b}+1}$$

La veracidad de la hipótesis H_0^{A*} en el modelo sin interacción equivale al hecho de que la media de la variable respuesta no dependa del nivel del factor A, es decir, que dicho

factor no influye, por término medio, en la respuesta (no es una verdadera fuente de variabilidad). Lo mismo sucede, pero para el factor B, respecto a la hipótesis H_0^{B*}. Desde el punto de vista del modelo completo, es decir, con interacción, no está tan claro cómo contrastar si uno de los factores, por ejemplo A, influye en la media de la variable respuesta. En principio, podríamos considerar el contraste de la hipótesis inicial H_0^A, pero, en este caso, su veracidad equivaldría a que, para cada $i = 1, \ldots, \mathtt{a}$, $\overline{\mu}_{i\cdot}$ sea igual a $\overline{\mu}_{\cdot\cdot}$. Esto se parece a lo que queremos contrastar, pero no es exactamente lo que buscamos, de ahí que existan distintas aproximaciones a la hora de intentar resolver este problema, según se comenta en Arnold (1981), pp. 97 y 98. Por ejemplo, podemos contrastar la hipótesis inicial

$$H_0^{A,AB} : \alpha_1 = \alpha_{\mathtt{a}} = (\alpha\beta)_{11} = \ldots = (\alpha\beta)_{\mathtt{ab}} = 0,$$

pues, de ser cierta, tendríamos un modelo del tipo

$$Y_{ijk} = \theta + \beta_j + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma^2), \ \sum_{j=1}^{\mathtt{b}} \beta_j = 0,$$

donde la media no depende del nivel de f_A. El test $\mathtt{F}$ a nivel α para resolver este contraste se obtiene sumando los términos (6.27) y (6.29), correspondientes a proyecciones sobre subespacios ortogonales. Consiste pues en comparar con $F^{\alpha}_{(\mathtt{a}-1)\mathtt{b},\mathtt{ab}(\mathtt{m}-1)}$ el estadístico

$$F_{A,AB} = \frac{[(\mathtt{a}-1)\mathtt{b}]^{-1}\mathtt{m}\sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}(\overline{y}_{ij\cdot} - \overline{y}_{\cdot j\cdot})^2}{[\mathtt{ab}(\mathtt{m}-1)]^{-1}\sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}\sum_{k=1}^{\mathtt{m}}(Y_{ijk} - \overline{y}_{ij\cdot})^2}$$

De manera completamente análoga se procedería si se estudiara la influencia del factor B. Otra forma de proceder consiste en contrastar, primeramente, si las interacciones son nulas, es decir, si H_0^{AB} es verdadera. Si el resultado no es significativo, consideramos el modelo reducido sin interacción y contrastamos la hipótesis H_0^{A*} o H_0^{B*}, dependiendo del factor que estemos estudiando. La principal crítica a este método es que el hecho de obtener un resultado no significativo al contrastar la hipótesis H_0^{AB} no significa, ni mucho menos, que se haya probado su veracidad.

6.5. Diseños equilibrados con tres o más factores

En esta sección intentaremos generalizar el modelo bifactorial equilibrado al caso en el que exista un número arbitrario de factores. Para evitar un excesiva complejidad en la notación expondremos únicamente el modelo con tres factores, entendiendo que

con ello quedarán claras las claves para extrapolar el estudio al caso general. El uso de un tercer factor f_C con $\mathtt{c}$ niveles obliga a introducir un nuevo subíndice h, que toma valores desde 1 hasta $\mathtt{c}$. Así, nuestro modelo consiste en considerar

$$Y_{ijhk} = \mu_{ijh} + \varepsilon_{ijh} \sim N(0, \sigma^2) \text{ independientes.} \tag{6.36}$$

En este caso, el EIMV de σ^2 es

$$\hat{\sigma}^{2,\mathrm{I}} = [\mathtt{abc}(\mathtt{m}-1)]^{-1} \sum_{i=1}^{\mathtt{a}} \sum_{j=1}^{\mathtt{b}} \sum_{h=1}^{\mathtt{c}} \sum_{k=1}^{\mathtt{m}} (Y_{ijhk} - \overline{y}_{ijh\cdot})^2$$

El modelo puede expresarse también mediante $Y \sim N_{\mathtt{n}}(\mu, \sigma^2 \mathtt{Id})$, donde $\mathtt{n} = \mathtt{abcm}$ y μ pertenece al subespacio V generado por los vectores, $\{\mathtt{v}_{ijh} : \ 1 \leq i \leq \mathtt{a},\ 1 \leq j \leq \mathtt{b},\ 1 \leq h \leq \mathtt{c}\}$, siendo $\mathtt{v}_{ijh}$ el vector de $\mathbb{R}^{\mathtt{abcm}}$ cuyas componentes son todas nulas salvo las correspondientes a la celda ijh, que valen 1. De manera completamente análoga al diseño bifactorial, podemos definir los vectores siguientes

$$\mathtt{v}_{ij\cdot} = \sum_{h=1}^{\mathtt{c}} \mathtt{v}_{ijh}, \qquad 1 \leq i \leq \mathtt{a},\ 1 \leq j \leq \mathtt{b}, \tag{6.37}$$

$$\mathtt{v}_{i\cdot h} = \sum_{j=1}^{\mathtt{b}} \mathtt{v}_{ijh}, \qquad 1 \leq i \leq \mathtt{a},\ 1 \leq h \leq \mathtt{c}, \tag{6.38}$$

$$\mathtt{v}_{\cdot jh} = \sum_{i=1}^{\mathtt{a}} \mathtt{v}_{ijh}, \qquad 1 \leq j \leq \mathtt{b},\ 1 \leq h \leq \mathtt{c}, \tag{6.39}$$

$$\mathtt{v}_{i\cdot\cdot} = \sum_{j=1}^{\mathtt{b}} \sum_{h=1}^{\mathtt{c}} \mathtt{v}_{ijh}, \qquad 1 \leq i \leq \mathtt{a}, \tag{6.40}$$

$$\mathtt{v}_{\cdot j\cdot} = \sum_{i=1}^{\mathtt{a}} \sum_{h=1}^{\mathtt{c}} \mathtt{v}_{ijh}, \qquad 1 \leq j \leq \mathtt{b}, \tag{6.41}$$

$$\mathtt{v}_{\cdot\cdot h} = \sum_{i=1}^{\mathtt{a}} \sum_{j=1}^{\mathtt{b}} \mathtt{v}_{ijh}, \qquad 1 \leq h \leq \mathtt{c}, \tag{6.42}$$

$$1_{\mathtt{n}} = \sum_{i=1}^{\mathtt{a}} \sum_{j=1}^{\mathtt{b}} \sum_{h=1}^{\mathtt{c}} \mathtt{v}_{ijh}. \tag{6.43}$$

En lo que sigue, V_1, V_2, V_3, V_{12}, V_{13} y V_{23} denotarán los subespacios de V generados por las familias (6.37), (6.38), (6.39), (6.40), (6.41) y (6.42), respectivamente. De esta forma, V_1 será el subespacio que recorre μ cuando ignoramos los factores f_B y f_C,

es decir, cuando consideramos un diseño completamente aleatorizado con a niveles y bcm datos por nivel. De manera análoga se interpretan los subespacios V_2 y V_3. Así mismo, V_{12} es el subespacio que recorre μ si ignoramos el factor f_C, es decir, cuando a considermos un diseño bifactorial equilibrado con a niveles para un factor, b niveles para el otro y mc datos por celda, lo cual equivale a un diseño completamente aleatorizado con ab niveles y mc datos por nivel. De igual forma se interpretan V_{13} (se suprime el segundo factor) y V_{12} (se suprime el tercero).

Proposición 6.3.

El subespacio V descompone en la siguiente suma de subespacios ortogonales:

$$\begin{aligned} V &= \langle 1_{\mathtt{n}}\rangle \oplus V_1|\langle 1_{\mathtt{n}}\rangle \oplus V_2|\langle 1_{\mathtt{n}}\rangle \oplus V_3|\langle 1_{\mathtt{n}}\rangle \\ &\quad \oplus V_{12}|(V_1\oplus V_2) \oplus V_{13}|(V_1\oplus V_3) \oplus V_{23}|(V_2\oplus V_3) \\ &\quad \oplus V|(V_{12}\oplus V_{13}\oplus V_{23}) \end{aligned}$$

Demostración.

Utilizando los mismos argumentos que en el modelo bifactorial equilibrado se deduce que $V_1|\langle 1_{\mathtt{n}}\rangle$, $V_2|\langle 1_{\mathtt{n}}\rangle$ y $V_3|\langle 1_{\mathtt{n}}\rangle$ son ortogonales. El mismo razonamiento sirve para probar la ortogonalidad entre $V_3|\langle 1_{\mathtt{n}}\rangle$ y $V_{12}|\langle 1_{\mathtt{n}}\rangle$, pues basta considerar un modelo bifactorial equilibrado con un factor, f_{AB}, con ab niveles y otro, f_C, con c niveles. De esta forma queda probada también la ortogonalidad entre $V_3|\langle 1_{\mathtt{n}}\rangle$ y $V_{12}|(V_1\oplus V_2)$ y, análogamente, la ortogonalidad entre $V_2|\langle 1_{\mathtt{n}}\rangle$ y $V_{13}|(V_1\oplus V_3)$ y entre $V_1|\langle 1_{\mathtt{n}}\rangle$ y $V_{23}|(V_2\oplus V_3)$. Probemos a continuación la ortogonalidad entre $V_{12}|(V_1\oplus V_2)$ y $V_{13}|(V_1\oplus V_3)$: sendos vectores e_{12} y $e13$ de estos subespacios se expresan, respectivamente, mediante

$$e_{12} = \sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}} x_{ij}\mathtt{v}_{ij\cdot} \qquad e_{13} = \sum_{i=1}^{\mathtt{a}}\sum_{h=1}^{\mathtt{c}} z_{ij}\mathtt{v}_{i\cdot h}$$

La ortogonalidad respecto a $V_1\oplus V_2$ y $V_1\oplus V_2$ se caracteriza, respectivamente, mediante

$$\sum_{i=1}^{\mathtt{a}} x_{ij} = 0,\ \forall j = 1,\ldots,\mathtt{b}, \qquad \sum_{j=1}^{\mathtt{b}} x_{ij} = 0,\ \forall i = 1,\ldots,\mathtt{a},$$

$$\sum_{i=1}^{\mathtt{a}} z_{ih} = 0,\ \forall h = 1,\ldots,c, \qquad \sum_{h=1}^{\mathtt{c}} z_{ih} = 0,\ \forall i = 1,\ldots,\mathtt{a}.$$

En consecuencia, se verifica

$$\begin{aligned}\langle e_{12}, e_{13}\rangle &= \mathtt{m}\sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}\sum_{h=1}^{\mathtt{c}} x_{ij}z_{ih}\\ &= \mathtt{m}\sum_{i=1}^{\mathtt{a}}\left(\sum_{j=1}^{\mathtt{b}} x_{ij}\sum_{h=1}^{\mathtt{c}} z_{ih}\right) = 0\end{aligned}$$

Un razonamiento similar permite probar el resto de ortogonalidades entre los espacios de la segunda fila. Las ortogonalidades restantes son obvias por definición. También se verifica, por definición, que la suma resultante es V.

∎

Nótese que en la demostración de la proposición anterior se nos dice cómo se expresa explícitamente un vector correspondiente a un subespacio perteneciente a la segunda fila del enunciado. Por otra parte, sabemos que un elemento $\langle 1_{\mathtt{n}}\rangle$ es un vector constante, un elemento de $V_1|\langle 1_{\mathtt{n}}\rangle$ se expresará mediante $\sum_{i=1}^{\mathtt{a}} x_i \mathtt{v}_{i\cdot\cdot}$, donde $\sum_{i=1}^{\mathtt{a}} x_i = 0$. De forma análoga se expresan los elementos de $V_2|\langle 1_{\mathtt{n}}\rangle$ y $V_3|\langle 1_{\mathtt{n}}\rangle$. Por último, los elementos de $V|(V_{12}\oplus V_{13}\oplus V_{23})$ se expresan mediante $\sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}\sum_{h=1}^{\mathtt{c}} x_{ijh}\mathtt{v}_{ijh\cdot}$, con las restricciones

$$\sum_{h=1}^{\mathtt{c}} \mathtt{x}_{ijh} = 0,\ \forall(i,j), \quad \sum_{j=1}^{\mathtt{b}} x_{ijh} = 0,\ \forall(i,h), \quad \sum_{i=1}^{\mathtt{a}} x_{ijh} = 0,\ \forall(h,j).$$

En consecuencia, el modelo puede expresarse también mediante

$$Y_{ijhk} = \theta+\alpha_i+\beta_j+\gamma_h+(\alpha\beta)_{ij}+(\alpha\gamma)_{ih}+(\beta\gamma)_{jh}+(\alpha\beta\gamma)_{ijh}+\varepsilon_{ijhk}, \quad \varepsilon_{ijhk}\sim N(0,\sigma^2)$$

con las restricciones siguientes:

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_h \gamma_h = 0,$$

$$\sum_i (\alpha\beta)_{ij} = 0,\ \forall j, \quad \sum_j (\alpha\beta)_{ij} = 0,\ \forall i,$$

$$\sum_i (\alpha\gamma)_{ih} = 0,\ \forall h, \quad \sum_h (\alpha\gamma)_{ih} = 0,\ \forall i,$$

$$\sum_j (\beta\gamma)_{jh} = 0,\ \forall h, \quad \sum_h (\beta\gamma)_{jh} = 0,\ \forall j,$$

$$\sum_i (\alpha\beta\gamma)_{ijh} = 0,\ \forall(j,h), \quad \sum_j (\alpha\beta\gamma)_{ijh} = 0,\ \forall(i,h), \quad \sum_h (\alpha\beta\gamma)_{ijh} = 0,\ \forall(i,j).$$

Estos parámetros pueden relacionarse con las medias μ_{ijh} de la siguiente forma

$$\begin{aligned}
\theta &= \overline{\mu}_{...} \\
\alpha_i &= \overline{\mu}_{i..} - \overline{\mu}_{...} \\
\beta_j &= \overline{\mu}_{.j.} - \overline{\mu}_{...} \\
\gamma_h &= \overline{\mu}_{..h} - \overline{\mu}_{...} \\
(\alpha\beta)_{ij} &= \overline{\mu}_{ij.} - \overline{\mu}_{i..} - \overline{\mu}_{.j.} + \overline{\mu}_{...} \\
(\alpha\gamma)_{ih} &= \overline{\mu}_{i.h} - \overline{\mu}_{i..} - \overline{\mu}_{..h} + \overline{\mu}_{...} \\
(\beta\gamma)_{jh} &= \overline{\mu}_{.jh} - \overline{\mu}_{.j.} - \overline{\mu}_{..h} + \overline{\mu}_{...} \\
(\alpha\beta\gamma)_{ijh} &= \mu_{ijh} - \overline{\mu}_{ij.} - \overline{\mu}_{i.h} - \overline{\mu}_{.jh} + \overline{\mu}_{i..} + \overline{\mu}_{i.h} + \overline{\mu}_{.jh} - \overline{\mu}_{...},
\end{aligned}$$

donde las medias aritméticas anteriores se definen de manera análoga al caso bifactorial. De igual forma descompondrá el vector Y en suma ortogonal de proyecciones, de manera que se obtienen las siguientes sumas cuadráticas

$$\begin{aligned}
\|P_{V_1|\langle 1_{\mathtt{n}}\rangle}Y\|^2 &= \mathtt{mbc}\sum_{i=1}^{\mathtt{a}}(\overline{y}_{i..} - \overline{y}_{...})^2, \\
\|P_{V_2|\langle 1_{\mathtt{n}}\rangle}Y\|^2 &= \mathtt{mac}\sum_{j=1}^{\mathtt{b}}(\overline{y}_{.j.} - \overline{y}_{...})^2, \\
\|P_{V_3|\langle 1_{\mathtt{n}}\rangle}Y\|^2 &= \mathtt{mab}\sum_{h=1}^{\mathtt{c}}(\overline{y}_{..h} - \overline{y}_{...})^2, \\
\|P_{V_{12}|(V_1\oplus V_2)}Y\|^2 &= \mathtt{mc}\sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2, \\
\|P_{V_{13}|(V_1\oplus V_2)}Y\|^2 &= \mathtt{mb}\sum_{i=1}^{\mathtt{a}}\sum_{h=1}^{\mathtt{c}}(\overline{y}_{i.h} - \overline{y}_{i..} - \overline{y}_{..h} + \overline{y}_{...})^2, \\
\|P_{V_{23}|(V_1\oplus V_2)}Y\|^2 &= \mathtt{ma}\sum_{j=1}^{\mathtt{b}}\sum_{h=1}^{\mathtt{c}}(\overline{y}_{.jh} - \overline{y}_{.j.} - \overline{y}_{..h} + \overline{y}_{...})^2, \\
\|P_{V|(V_{12}\oplus V_{13}\oplus V_{23})}Y\|^2 &= \mathtt{m}\sum_{i,j,h}\left(\overline{y}_{ijh\cdot} - \overline{y}_{ij.} - \overline{y}_{i\cdot h} - \overline{y}_{.jh} + \overline{y}_{i..} + \overline{y}_{.j.} + \overline{y}_{..h} - \overline{y}_{...}\right)^2
\end{aligned}$$

Así, estamos en condiciones de contrastar un buen número de hipótesis iniciales. Por ejemplo, el test `F` a nivel α para contrastar la hipótesis inicial $\alpha_1 = \ldots = \alpha_{\mathtt{a}} = 0$ consiste en comparar con $F^{\alpha}_{\mathtt{a}-1,\mathtt{abc}(\mathtt{m}-1)}$ el estadístico

$$F_A = \frac{(\mathtt{a}-1)^{-1}\mathtt{mbc}\sum_{i=1}^{\mathtt{a}}(\overline{y}_{i..} - \overline{y}_{...})^2}{[\mathtt{abc}(\mathtt{m}-1)]^{-1}\sum_{i,j,h,k}\left(Y_{ijhk} - \overline{y}_{ijh\cdot}\right)^2}$$

El test F a nivel α para contrastar la hipótesis inicial $(\alpha\beta)_{11} = \ldots = (\alpha\beta)_{\mathtt{ab}} = 0$ consiste en comparar con $F^{\alpha}_{(\mathtt{a}-1)(\mathtt{b}-1),\mathtt{abc}(\mathtt{m}-1)}$ el estadístico

$$F_{AB} = \frac{(\mathtt{a}-1)(\mathtt{b}-1)^{-1}\mathtt{mc}\sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdots})^2}{[\mathtt{abc}(\mathtt{m}-1)]^{-1}\sum_{i,j,h,k}\left(Y_{ijhk} - \overline{y}_{ijh\cdot}\right)^2}$$

El test F a nivel α para contrastar la hipótesis inicial $(\alpha\beta\gamma)_{111} = \ldots = (\alpha\beta\gamma)_{\mathtt{abc}} = 0$ consiste en comparar con $F^{\alpha}_{(\mathtt{a}-1)(\mathtt{b}-1)(\mathtt{c}-1),\mathtt{abc}(\mathtt{m}-1)}$ el estadístico

$$F_{ABC} = \frac{\mathtt{m}\sum_{i,j,h}\left(\overline{y}_{ijh\cdot} - \overline{y}_{ij\cdot} - \overline{y}_{i\cdot h} - \overline{y}_{\cdot jh} + \overline{y}_{i\cdot\cdot} + \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdot\cdot h} - \overline{y}_{\cdots}\right)^2}{[\mathtt{abc}(\mathtt{m}-1)]^{-1}\sum_{i,j,h,k}\left(Y_{ijhk} - \overline{y}_{ijh\cdot}\right)^2}$$

Finalmente, pueden obtenerse de manera trivial (se deja como ejercicio)las familias de intervalos de confianza para $\{\alpha_i - \alpha_{i'} : i \neq i'\}$, $\{\beta_j - \beta_{j'} : j \neq j'\}$ y $\{\gamma_h - \gamma_{h'} : h \neq h'\}$, según los métodos de Bonferroni, Tuckey y Scheffé.

6.6. Diseños anidados o jerárquicos equilibrados

A continuación estudiaremos un diseño que tiene por objeto contrastar la influencia de dos factores, A y B, en la media de una variable respuesta, con la particularidad de que el factor B no es tal, en el sentido estricto de la palabra, sino que se define para cada nivel i del factor A, presentando en se caso un total de $\mathtt{b}_i$ niveles. Por ejemplo, supongamos que pretendemos evaluar si cierta variable biológica depende de la especie considerada. Para ello, se toman a especies sobre las que se mide la variable. No obstante, se desea también controlar el factor subespecie, bien por reducir el variabilidad achacable al azar o bien porque el contraste de su posible influencia sea interesante en sí mismo. Obviamente, el número de subespecies a considerar dependerá de la especie en cuestión. Por ello, el factor subespecie está subordinado al factor especie. En todo caso, para cada nivel i del factor A y cada nivel j_i del factor subordinado B [2], consideraremos m mediciones de la variable respuesta. Se trata pues de un diseño equilibrado. Si añadimos los supuestos típicos del modelo lineal normal (independencia, normalidad y homocedasticidad), tendremos el siguiente modelo

$$Y_{ij_ik} = \mu_{ij_i} + \varepsilon_{ij_i}, \quad \varepsilon_{ij_i} \sim N(0, \sigma^2) \text{ independientes.}$$

[2] Aunque no es estrictamente necesario, se expresa el nivel del factor B mediante el subíndice j_i, en lugar de j, con la intención de recalcar la subordinación al factor A y así diferenciar claramente este diseño del bifactorial.

Este modelo coincide con el que correspondería a un diseño completamente aleatorizado con $\sum_{i=1}^{\mathtt{a}} \mathtt{b}_i$ niveles y $\mathtt{m}$ observaciones por nivel. Por lo tanto, el EIMV de σ^2 es el siguiente

$$\hat{\sigma}^{2,\mathrm{I}} = \frac{1}{\sum_{i=1}^{\mathtt{a}} \mathtt{b}_i(\mathtt{m}-1)} \sum_{i=1}^{\mathtt{a}} \sum_{j_i=1}^{\mathtt{b}_i} \sum_{k=1}^{\mathtt{m}} (Y_{ij_ik} - \overline{y}_{ij_i\cdot})^2.$$

El número total de observaciones es $\mathtt{n} = \mathtt{m} \sum_{i=1}^{\mathtt{a}} \mathtt{b}_i$. Componiéndolas todas obtenemos la siguiente expresión del modelo

$$Y = \mu + \mathcal{E}, \qquad \mathcal{E} \sim N_{\mathtt{n}}(0, \sigma^2 \mathtt{Id}),\ \mu \in V,\ \sigma^2 > 0,$$

siendo V el subespacio generado por los vectores $\mathtt{v}_{ij_i}$, donde $i = 1, \ldots, \mathtt{a}$ y $j_i = 1, \ldots, \mathtt{b}_i$ (se definen de manera completamente análoga a la de las secciones anteriores). Si V_1 denota el subespacio generado por los vectores $\mathtt{v}_{1\cdot}, \ldots, \mathtt{v}_{\mathtt{a}\cdot}$, podemos considerar la descomposición ortogonal siguiente:

$$V = \langle 1_{\mathtt{n}} \rangle \ \oplus\ V_1 | \langle 1_{\mathtt{n}} \rangle \ \oplus V | V_1$$

Los vectores $V_1|\langle 1_{\mathtt{n}}\rangle$ se expresan de la forma $\sum_{i=1}^{\mathtt{a}} x_i \mathtt{v}_{i\cdot}$, con $\sum_{i=1}^{\mathtt{a}} x_i = 0$, mientras que los de $V|V_1$ se expresan de la forma $\sum_{i=1}^{\mathtt{a}} \sum_{j_i=1}^{\mathtt{b}_i} x_{ij_i} \mathtt{v}_{ij_i}$, con la restricción

$$\sum_{j_i=1}^{\mathtt{b}_i} x_{ij_i} = 0, \quad i = 1, \ldots, \mathtt{a}.$$

Por lo tanto, teniendo en cuenta la descomposición anterior, podemos, expresar el modelo mediante

$$Y_{ij_ik} = \theta + \alpha_i + \eta_{ij_i} + \varepsilon_{ij_ik}, \quad \sum_{i=1}^{\mathtt{a}} \alpha_i = 0,\ \sum_{j_i=1}^{\mathtt{b}_i} \eta_{ij_i} = 0,\ i = 1, \ldots, \mathtt{a}.$$

Los parámetros θ, α_i y η_{ij_i} pueden se relacionan con las medias mediante

$$\theta = \overline{\mu}_{\cdot\cdot}, \qquad \alpha_i = \overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot\cdot}, \qquad \eta_{ij_i} = \mu_{ij_i} - \overline{\mu}_{i\cdot}$$

El subespacio V_1 sería el que correspondería al diseño completamente aleatorizado que se obtiene ignorando el el factor subordinado. Por lo tanto, la proyección del vector Y sobre V_1 consiste en asignar a la posición ij_ik el valor $\overline{y}_{i\cdot\cdot}$. En consecuencia,

$$\begin{aligned} \|P_{V_1|\langle 1_{\mathtt{n}}\rangle} Y\|^2 &= \mathtt{m} \sum_{i=1}^{\mathtt{a}} \mathtt{b}_i (\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots})^2, \\ \|P_{V|V_1} Y\|^2 &= \mathtt{m} \sum_{i=1}^{\mathtt{a}} \sum_{j_i=1}^{\mathtt{b}_i} (\overline{y}_{ij_i\cdot} - \overline{y}_{i\cdot\cdot})^2. \end{aligned}$$

Así pues, estamos en condiciones de contratar las hipótesis iniciales

$$H_0^A : \alpha_1 = \ldots = \alpha_{\mathtt{a}} = 0$$

$$H_0^{AB} : \eta_{11} = \ldots = \eta_{\mathtt{ab_a}} = 0$$

El test F a nivel α para contrastar H_0^A consiste en comparar con $F^{\alpha}_{\mathtt{a}-1,(\mathtt{m}-1)\sum_{i=1}^{\mathtt{a}}\mathtt{b}_i}$ el estadístico

$$F_A = \frac{(\mathtt{a}-1)^{-1}\mathtt{m}\sum_{i=1}^{\mathtt{a}}\mathtt{b}_i(\overline{y}_{i\cdot\cdot}-\overline{y}_{\cdots})^2}{[(\mathtt{m}-1)\sum_{i=1}^{\mathtt{a}}\mathtt{b}_i]^{-1}\sum_{i=1}^{\mathtt{a}}\sum_{j_i=1}^{\mathtt{b}_i}\sum_{k=1}^{\mathtt{m}}(Y_{ij_ik}-\overline{y}_{ij_i\cdot})^2}$$

El test F a nivel α para contrastar H_0^B consiste en comparar con $F^{\alpha}_{\sum_{i=1}^{\mathtt{a}}\mathtt{b}_i-\mathtt{a},(\mathtt{m}-1)\sum_{i=1}^{\mathtt{a}}\mathtt{b}_i}$ el estadístico

$$F_{AB} = \frac{(\sum_{i=1}^{\mathtt{a}}\mathtt{b}_i-\mathtt{a})^{-1}\mathtt{m}\sum_{i=1}^{\mathtt{a}}\sum_{j_i=1}^{\mathtt{b}_i}(\overline{y}_{ij_i\cdot}-\overline{y}_{i\cdot\cdot})^2}{[(\mathtt{m}-1)\sum_{i=1}^{\mathtt{a}}\mathtt{b}_i]^{-1}\sum_{i=1}^{\mathtt{a}}\sum_{j_i=1}^{\mathtt{b}_i}\sum_{k=1}^{\mathtt{m}}(Y_{ij_ik}-\overline{y}_{ij_i\cdot})^2}$$

Se pueden obtener de manera trivial (se deja como ejercicio) comparaciones multiples para los efectos del factor A según los métodos de Bonferroni, Tuckey y Scheffé.

El aceptación de la hipótesis H_0^A no debe interpretarse como la no influencia del factor A en la media de la variable respuesta. Esta situación se correspondería más bien con a hipótesis $H_0^{A,AB} = H_0^A \bigwedge H_0^{AB}$. Esta hipótesis puede contrastarse directamente. También se puede optar por contrastar H_0^{AB} y, si el resultado no es significativo, realizar el contraste principal en el modelo reducido correspondiente aldiseño completamente aleatorizado para el factor A.

6.7. Bloques aleatorizados y cuadrados latinos

Los diseños por bloques aleatorizados tienen por objeto contrastar la influencia de un único factor, denominado factor principal o tratamiento, en la media de cierta variable respuesta. Sin embargo a diferencia del diseño completamente aleatorizado, se consideran simultáneamente uno o varios factores, denominados secundarios, sospechosos de ser constituir una fuente de variabilidad, con el objeto de reducir el grado de azar inherente al experimento (expresado por el parámetro σ^2), lo cual posibilitará, en principio, resultados más significativos. Se supondrá, por hipótesis, que los distintos factores considerados tienen efecto aditivo, es decir, que no se considerar ningún tipo de interacción en el modelo. Empezaremos considerando el diseño con un único factor secundario, que es el que se conoce propiamente como diseño en bloques

aleatorizados, para estudiar posteriormente el diseño con dos factores secundario, denominado de cuadrados greco-latinos.

En el diseño con un único factor secundario f_B, los $\mathtt{b}$ niveles del mismo se denominarán bloques. Para cada combinación ij entre los niveles de uno de los $\mathtt{a}$ niveles del factor principal y del secundario se tomará un único dato al azar, Y_{ij}. Al no contemplar ningún tipo de interacción entre bloques y tratamientos, el modelo asociado corresponde a un diseño bifactorial sin interacción con $\mathtt{m} = 1$ observación por celda, es decir,

$$Y_{ij} = \theta + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \text{ independientes}, \quad \sum_{i=1}^{\mathtt{a}} \alpha_i = \sum_{j=1}^{\mathtt{b}} \beta_j = 0.$$

Por lo tanto, podemos considerarlo resuelto desde un punto de vista teórico. Concretamente, el estimador de la varianza es

$$\hat{\sigma}^{2,\mathrm{I}} = \frac{1}{(\mathtt{a}-1)(\mathtt{b}-1)} \sum_{i=1}^{\mathtt{a}} \sum_{j=1}^{\mathtt{b}} \left(y_{ij} - \overline{y}_{i\cdot} - \overline{y}_{\cdot j} + \overline{y}_{\cdot\cdot}\right)^2$$

El test $\mathtt{F}$ a nivel α para contrastar la hipótesis inicial $H_0^A : \alpha_1 = \ldots = \alpha_{\mathtt{a}} = 0$ (es decir, que el tratamiento no tiene influencia, por término medio, en la variable respuesta), consiste en comparar con $F^{\alpha}_{\mathtt{a}-1,(\mathtt{a}-1)(\mathtt{b}-1)}$ el estadístico

$$F_A = \frac{(\mathtt{a}-1)^{-1}\mathtt{b}\sum_{i=1}^{\mathtt{a}}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2}{[(\mathtt{a}-1)(\mathtt{b}-1)]^{-1}\sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}\left(y_{ij} - \overline{y}_{i\cdot} - \overline{y}_{\cdot j} + \overline{y}_{\cdot\cdot}\right)^2}$$

Para contrastar la influencia del bloque en la media de l variable respuesta, se compara con $F^{\alpha}_{(\mathtt{b}-1),(\mathtt{a}-1)(\mathtt{b}-1)}$ el estadístico

$$F_B = \frac{(\mathtt{b}-1)^{-1}\mathtt{a}\sum_{j=1}^{\mathtt{b}}(\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot})^2}{[(\mathtt{a}-1)(\mathtt{b}-1)]^{-1}\sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}\left(y_{ij} - \overline{y}_{i\cdot} - \overline{y}_{\cdot j} + \overline{y}_{\cdot\cdot}\right)^2}$$

Un resultado significativo de éste test puede conducir a considerar un error la elección del factor secundario como fuente de variabilidad. Ello puede suponer que la significación al resultado del test para el tratamiento sea menor que la que se obtendría mediante un diseño completamente aleatorizado, dado que, al controlar los bloques, la varianza del modelo apenas disminuye, mientras que el grado de libertad por el que se divide sí.

Nótese también que, en un modelo de este tipo, con un único dato por celda, no cabe siquiera contemplar la posibilidad de que exista interacción entre los factores,

pues ello nos conduciría a un modelo donde la dimensión de V coincidiría con el número de datos, con la cual no se podría siquiera estimar σ^2. Por otra parte, las comparaciones entre los niveles del tratamiento mediante los métodos de Bonferroni, Tuckey y Scheffé se realizan, respectivamente, mediante las siguientes familias:

- Bonferroni: $\alpha_i - \alpha_{i'} \in \overline{y}_{i\cdot} - \overline{y}_{i'\cdot} \pm t^{\alpha/\mathtt{a}(\mathtt{a}-1)}_{(\mathtt{a}-1)(\mathtt{b}-1)} \hat{\sigma}_{\text{I}} \sqrt{\frac{2}{\mathtt{b}}}$
- Tuckey: $\alpha_i - \alpha_{i'} \in \overline{y}_{i\cdot} - \overline{y}_{i'\cdot} \pm q^{\alpha}_{\mathtt{a},(\mathtt{a}-1)(\mathtt{b}-1)} \hat{\sigma}_{\text{I}} \sqrt{\frac{2}{\mathtt{b}}}$
- Scheffé: $\alpha_i - \alpha_{i'} \in \overline{y}_{i\cdot} - \overline{y}_{i'\cdot} \pm \hat{\sigma}_{\text{I}} \sqrt{\frac{2}{\mathtt{b}} F^{\alpha}_{\mathtt{a}-1,(\mathtt{a}-1)(\mathtt{b}-1)}}$

A continuación estudiaremos el caso en el que se introducen dos factores secundarios. En ese caso, aplicando la lógica anterior, deberíamos considerar los distintos niveles del tratamiento para cada combinación entre los niveles de los factores secundarios. No obstante y con el propósito de ahorrar datos, se considerará un diseño como el que sigue, denominado *diseño de cuadrados latinos*. En este caso, alteraremos ligeramente la notación, pues A y B denotarán los factores secundarios, mientras que T denotará el factor principal o tratamiento. El número de niveles $\mathtt{s}$ de T coincidirá con el número de niveles de A y B. Para cada nivel i del primer factor secundario, A, se considerará una única ejecución para cada unos de los niveles, t, de el tratamiento. Lo mismo sucederá para cada nivel j de B. De todas entre todas las formas de obtener un modelo así, se escogerá aleatoriamente una de ellas[3]. Veamos un ejemplo con $\mathtt{s} = 4$:

4×4	**B1**	**B2**	**B3**	**B4**
A1	T1	T2	T3	T4
A2	T2	T3	T4	T1
A3	T3	T4	T1	T2
A4	T4	T1	T2	T3

Tanto si se consideran las filas como las columnas, se pueden observar distintas permutaciones del conjunto $\{1, 2, 3, 4\}$. De esta forma, en vez de considerar 4^3 datos debemos recabar únicamente 4^2. Además, nos aseguramos de que cada nivel de A y cada nivel de B se someta a cada nivel del tratamiento, aunque sea una única vez.

Veamos cómo se formaliza este diseño. Primeramente, hemos de seleccionar dos subconjuntos, compuesto cada uno de ellos por $\mathtt{s}$ permutaciones distintas de los elementos de $\{1, \ldots, \mathtt{s}\}$, que se denotan por $\{\tau_{A,1}, \ldots, \tau_{A,\mathtt{s}}\}$ y $\{\tau_{B,1}, \ldots, \tau_{B,\mathtt{s}}\}$, y verificando que $\tau^{-1}_{A,i}(j) = \tau^{-1}_{B,j}(i)$, para todo par i, j. Precisamente, dicho número indica el

[3]En Peña (1986), pag. 130, se muestran las distintas posibilidades para los valores $\mathtt{s}$ de 3 a 8.

nivel del tratamiento que corresponderá a la combinación entre los niveles i-ésimo y j-ésimo de A y B, respectivamente., que se denota por $t(ij)$. Dicho de otra forma, dados los nivel i y t de A y T, respectivamente, $\tau_{A,i}(t)$ denota el único nivel j de B tal que t se aplica en la celda ij. Igualmente, $\tau_{B,j}(t)$ denota el único nivel i tal que t se aplica en la celda ij. En ese caso, el modelo correspondiente es, al menos en principio, el siguiente

$$Y_{ij,t(ij)} = \mu_{ij,t(ij)} + \varepsilon_{ij,t(ij)}, \quad \varepsilon_{ij,t(ij)} \sim N(0,\sigma^2) \text{ independientes.} \tag{6.44}$$

Por lo tanto, puede expresarse también mediante

$$Y = \mu + \mathcal{E},\ \mathcal{E} \sim N_{\mathtt{s}^2}(0, \sigma^2 \mathtt{Id}),\ \mu \in \mathbb{R}^{\mathtt{s}^2},\ \sigma^2 > 0.$$

Consideremos los subespacios de $\mathbb{R}^{\mathtt{s}}$ siguientes: V_1 y V_2, definidos de forma análoga al modelo bifactorial con $\mathtt{m} = 1$, y V_T, generado por la familia

$$\left\{ \sum_{i=1}^{\mathtt{s}} \mathtt{v}_{i,\tau_{A,i}(t)} :\ t = 1, \ldots, \mathtt{s} \right\} = \left\{ \sum_{j=1}^{\mathtt{s}} \mathtt{v}_{\tau_{B,j}(t),j} :\ t = 1, \ldots, \mathtt{s} \right\}$$

Proposición 6.4.
La siguiente descomposición es ortogonal

$$\mathbb{R}^{\mathtt{s}^2} = \langle 1_{\mathtt{s}^2} \rangle \ \oplus\ V_1 | \langle 1_{\mathtt{s}^2} \rangle \ \oplus\ V_2 | \langle 1_{\mathtt{s}^2} \rangle \ \oplus\ V_T | \langle 1_{\mathtt{s}^2} \rangle \ \oplus\ (V_1 \oplus V_2 \oplus V_3)^{\perp}$$

Demostración.
Basta demostrar que V_T es ortogonal a $V_1|\langle 1_{\mathtt{s}^2}\rangle$ y $V_2|\langle 1_{\mathtt{s}^2}\rangle$. Efectivamente, consideremos un vector de la forma $e_1 = \sum_{i=1}^{\mathtt{s}} x_i \mathtt{v}_{i\cdot}$, con $\sum_{i=1}^{\mathtt{s}} x_i = 0$, y otro de la forma $e_t = \sum_{i=1}^{\mathtt{s}} \mathtt{v}_{i,\tau_{A,i}(t)}$, para algún t entre 1 y $\mathtt{s}$. En ese caso, $e_1 * e_t = \sum_{i=1}^{\mathtt{s}} x_i \mathtt{v}_{i,\tau_{A,i}(t)}$. Luego, $\langle e_1, e_t\rangle = \sum_{i=1}^{\mathtt{s}} x_i = 0$. Así queda probado que $V_1|\langle 1_{\mathtt{s}^2}\rangle \perp V_T$. Para el caso de V_2, consideramos un vector de la forma $e_2 = \sum_{j=1}^{\mathtt{s}} z_j \mathtt{v}_{\cdot j}$, con $\sum_{j=1}^{\mathtt{s}} z_j = 0$, y otro vector $e_t = \sum_{j=1}^{\mathtt{s}} \mathtt{v}_{\tau_{B,j}(t),j}$, para algún t. En ese caso, $e_2 * e_t = \sum_{j=1}^{\mathtt{s}} z_j \mathtt{v}_{\tau_{B,j}(t),j}$. Por lo tanto, $\langle e_2, e_t\rangle = 0$.

∎

De acuerdo con esta descomposición ortogonal, el modelo (6.44) puede expresarse como sigue

$$Y_{t(ij),ij} = \theta + \alpha_i + \beta_j + \delta_t + \gamma_{ij} + \varepsilon_{t(ij),ij}, \quad \varepsilon_{t(ij),ij} \sim N(0,\sigma^2),$$

con las siguientes restricciones

$$\sum_{i=1}^{\mathtt{a}} \alpha_i = \sum_{j=1}^{\mathtt{b}} \beta_j = \sum_{t=1}^{\mathtt{s}} \delta_t = 0, \qquad \sum_{i=1}^{\mathtt{a}} \gamma_{i,\tau_{A,i}(t)} = 0,\ t = 1, \ldots, \mathtt{s}.$$

En lo que sigue, supondremos, por hipótesis, que los parámetros γ_{ij} son todos nulos[4], es decir, que suponemos, al igual que en el diseño por bloques aleatorizados, que se da una aditividad entre los efectos de los tres factores. Estamos pues considerando el modelo

$$Y_{t(ij),ij} = \theta + \alpha_i + \beta_j + \delta_t + \varepsilon_{t(ij),ij}, \qquad \sum_{i=1}^{\mathtt{a}} \alpha_i = \sum_{j=1}^{\mathtt{b}} \beta_j = \sum_{t=1}^{\mathtt{s}} \delta_t = 0.$$

Necesitamos calcular las proyecciones sobre los distintos subespacios considerados: el vector $P_{V_1}Y$ será el que toma en la cada posición ij el valor $\overline{y}_{i\cdot} = \mathtt{s}^{-1}\sum_{j=1}^{\mathtt{s}} Y_{ij}$; $P_{V_2}Y$ toma en la posición ij el valor $\overline{y}_{\cdot j} = \mathtt{s}^{-1}\sum_{i=1}^{\mathtt{s}} Y_{ij}$; P_{V_T} toma en la posición ij el valor $\overline{y}_t = \mathtt{s}^{-1}\sum_{i=1}^{\mathtt{s}} Y_{i,\tau_{A,i}(t)}$. En consecuencia, el EIMV de σ^2 se obtiene mediante

$$\hat{\sigma}^{2,\mathtt{I}} = \frac{1}{(\mathtt{s}-1)(\mathtt{s}-2)} \sum_{i=1}^{\mathtt{s}}\sum_{j=1}^{\mathtt{s}} (Y_{ij} - \overline{y}_{i\cdot} - \overline{y}_{\cdot j} - \overline{y}_t + 2\overline{y}_{\cdot\cdot})^2.$$

Además,

$$\begin{aligned} \|P_{V_1|\langle 1_{\mathtt{S}^2}\rangle} Y\|^2 &= \mathtt{s}\sum_{i=1}^{\mathtt{s}} (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2 \\ \|P_{V_2|\langle 1_{\mathtt{S}^2}\rangle} Y\|^2 &= \mathtt{s}\sum_{j=1}^{\mathtt{s}} (\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot})^2 \\ \|P_{V_T|\langle 1_{\mathtt{S}^2}\rangle} Y\|^2 &= \mathtt{s}\sum_{t=1}^{\mathtt{s}} (\overline{y}_t - \overline{y}_{\cdot\cdot})^2 \end{aligned}$$

Así, por ejemplo, el test $\mathtt{F}$ a nivel α para contrastar la hipótesis inicial $H_0^T : \delta_1 = \ldots = \delta_{\mathtt{s}} = 0$, consiste en comparar con $F^{\alpha}_{\mathtt{s}-1,(\mathtt{s}-1)(\mathtt{s}-2)}$ el estadístico

$$F_T = \frac{(\mathtt{s}-1)^{-1}\mathtt{s}\sum_{t=1}^{\mathtt{s}} (\overline{y}_t - \overline{y}_{\cdot\cdot})^2}{[(\mathtt{s}-1)(\mathtt{s}-2)]^{-1}\sum_{i=1}^{\mathtt{s}}\sum_{j=1}^{\mathtt{s}} (Y_{ij} - \overline{y}_{i\cdot} - \overline{y}_{\cdot j} - \overline{y}_t + 2\overline{y}_{\cdot\cdot})^2}$$

Un resultado significativo se interpretaría como una influencia de los distintos tipos de tratamientos en la variable respuesta. La influencia de los factores secundarios puede ser contrastada de forma análoga. Así mismo, puede construirse comparaciones múltiples para los tratamientos según los métodos de Bonferroni, Tuckey y Scheffé (se deja como ejercicio).

La idea del diseño de cuadrados latinos puede extenderse al caso de tres factores secundarios, obteniendo así el denominado diseño de cuadrados greco-latinos. Los detalles de este diseño se pueden consultar, por ejemplo, en Peña (1986).

[4]Si aplicamos ninguna restricción a la media, ésta podría ser cualquier vector de $\mathbb{R}^{\mathtt{s}^2}$ y el modelo considerado no sería siquiera lineal.

6.8. Diseños no equilibrados

A continuación, vamos a abordar un análisis crítico, desde una perspectiva global, de lo que hemos estudiado hasta ahora en el capítulo. Se trata de analizar la influencia de uno o varios factores cualitativos en la media de cierta variable respuesta. En el caso de un único factor, el estudio resulta trivial a partir de los resultados obtenidos en el capítulo 2, cosa que no ocurre cuando se consideran varios factores. En tal caso, el primer problema es cómo descomponer la media de cada observación, de manera que puedan contrastarse aisladamente la repercusión de cada factor en la media de la variable o las interacciones entre los distintos factores.

Pongamos por ejemplo el diseño bifactorial equilibrado, en el cual la media correspondiente a los nivel i-ésimo y j-ésimo de los factores A y B, espectivamemte, es μ_{ij}. Para poder aislar los efectos de los factores y la interacción entre los mismos, se considera una descomposición del tipo

$$\mu_{ij} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij}. \tag{6.45}$$

Descomposiciones de esta forma podemos encontrar muchas, puesto que los nuevos parámetros constituyen soluciones particulares a un sistema de $\mathtt{ab}$ ecuaciones lineales con $(\mathtt{a}+1)(\mathtt{b}+1)$ incógnitas[5]. Una solución particular, la que se adopta en el capítulo, se obtiene considerando

$$\theta = \overline{\mu}_{..} \qquad \alpha_i = \overline{\mu}_{i.} - \overline{\mu}_{..} \qquad \beta_j = \overline{\mu}_{.j} - \overline{\mu}_{..} \qquad (\alpha\beta)_{ij} = \mu_{ij} - \overline{\mu}_{i.} - \overline{\mu}_{.j} + \overline{\mu}_{..} \tag{6.46}$$

En ese caso, se verifican las siguientes restricciones

$$\sum_{i=1}^{\mathtt{a}} \alpha_i = 0, \quad \sum_{j=1}^{\mathtt{b}} \beta_j = 0, \quad \sum_{j=1}^{\mathtt{b}} (\alpha\beta)_{ij} = 0,\ i = 1, \ldots, \mathtt{a}, \ \sum_{i=1}^{\mathtt{a}} (\alpha\beta)_{ij} = 0,\ j = 1, \ldots, \mathtt{b}. \tag{6.47}$$

Realmente, estas restricciones, consideradas como vectores de $\mathbb{R}^{(\mathtt{a}+1)(\mathtt{b}+1)}$, no son linealmente independientes, es decir, son redundantes. De hecho, equivalen, por ejemplo, a las siguientes, que sí son linealmente independientes:

$$\sum_{i=1}^{\mathtt{a}} \alpha_i = 0, \quad \sum_{j=1}^{\mathtt{b}} \beta_j = 0, \quad \sum_{j=1}^{\mathtt{b}} (\alpha\beta)_{ij} = 0,\ i = 1, \ldots, \mathtt{a}, \ \sum_{i=1}^{\mathtt{a}} (\alpha\beta)_{ij} = 0,\ j = 1, \ldots, \mathtt{b}-1 \tag{6.48}$$

Imponer estas $\mathtt{a} + \mathtt{b} + 1$ restricciones equivale a añadir $\mathtt{a} + \mathtt{b} + 1$ ecuaciones lineales hasta completar un total de $(\mathtt{a} + 1)(\mathtt{b} + 1)$. Obtenemos así un sistema de ecuaciones

[5] En el próximo capítulo se describirá con precisión el espacio de soluciones

cuya única solución es (6.46). Así pues, hemos de tener claro que considerar una descomposición particular del tipo (6.45) equivale a imponer una familia de restricciones, como, por ejemplo, (6.48). Las descomposiciones particulares que hemos obtenido en los distintos diseños estudiados o, lo que es lo mismo, las restricciones consideradas en los mismos, obedecen, en todo caso, a una descomposición natural de V como suma directa de subespacios ortogonales y siguiendo un procedimiento inductivo. Así, recordamos que, en el diseño unifactorial, se considera la descomposición

$$V = \langle 1_{\mathbf{n}} \rangle \ \oplus \ V|\langle 1_{\mathbf{n}} \rangle.$$

En el bifactorial, tenemos

$$V = \langle 1_{\mathbf{n}} \rangle \ \oplus \ V_1|\langle 1_{\mathbf{n}} \rangle \ \oplus \ V_2|\langle 1_{\mathbf{n}} \rangle \ \oplus V|(V_1 \oplus V_2).$$

En el caso trifactorial, la descomposición es la siguiente

$$\begin{aligned} V \ = \ & \langle 1_{\mathbf{n}} \rangle \ \oplus V_1|\langle 1_{\mathbf{n}} \rangle \ \oplus V_2|\langle 1_{\mathbf{n}} \rangle \ \oplus V_3|\langle 1_{\mathbf{n}} \rangle \\ & \oplus V_{12}|(V_1 \oplus V_2) \ \oplus \ V_{13}|(V_1 \oplus V_3) \ \oplus \ V_{23}|(V_2 \oplus V_3) \\ & \oplus V|(V_{12} \oplus V_{13} \oplus V_{23}). \end{aligned}$$

En el diseño unifactorial, la perpendicularidad de la descomposición viene dada por la mima construcción. Sin embargo, en los diseños con dos o más factores, para garantizar la ortogonalidad ha sido preciso imponer la condición de que el diseño sea equilibrado. De esta forma, para todos los diseños estudiados en el capítulo, hemos obtenido una restricción de los parámetros que puede considerarse natural.

Por otra parte, cuando se planifica un diseño con el objeto de estudiar la influencia de uno o varios factores en una variable respuesta, el hecho de considerar un mismo número de observaciones por celda no sólo resulta razonable desde un punto de vista estético, sino que puede favorecer también la robustez del modelo. No obstante, dado que el proceso de recogida de datos no siempre se ajusta a nuestra expectativas, convendría estudiar el tratamiento adecuado de los datos cuando el diseño (con más de un factor) no sea equilibrado. En ese caso, a la hora de plantear una descomposición de la media del tipo (6.45), no contamos, al menos en principio, con ningún argumento para privilegiar una familia de restricciones en detrimento de las demás. Realmente, nada nos impide optar por las mismas soluciones seleccionadas en el diseño equilibrado pero, en este caso, los parámetros no se traducirían en términos de las medias de manera natural, como sucede en (6.46), por lo que la elección resultaría completamente arbitraria. Parece claro que un estudio coherente de los diseños no equilibrados debería partir de un análisis de todas las familias de restricciones a considerar o, lo que es lo mismo, de todos las soluciones al sistema de ecuaciones del tipo

(6.45). Para ello, debemos enfocar el problema desde un punto de vista más general, y eso es, precisamente, lo que nos lleva al estudio del Modelo Lineal de Rango no Completo, que se abordará en el capítulo 6.

6.9. Diseños con efectos aleatorios

Para acabar este capítulo abordamos el estudio de diseños del análisis de la varianza en los que los niveles o valores de uno o varios de los factores considerados no se restringen a una familia finita determinada de antemano, sino que se escoge un número determinado de niveles de manera aleatoria en un amplio espacio. El estudio formal de estos modelos es muy similar al de los modelos con efectos fijos, estudiados en el resto del capítulo. De hecho, podemos encontrar estimadores y tests muy similares a los propuestos en dichos modelos. No obstante, se trata en general de un teoría que no goza de la consistencia de la anterior, de ahí que las soluciones propuestas a los principales problemas de Inferencia carezcan en la mayoría de los casos de las sólidas justificaciones teóricas que poseían las soluciones correspondientes a modelos con efectos fijos. Además, las técnicas utilizadas en las demostraciones, aunque similares a las ya estudiadas, presentan diversas variaciones. Hemos optado por obviar dichas demostraciones con el objeto de no extendernos demasiado. Si el tema se expusiera con todo detalle debería configurar un capítulo aparte. En la presente sección nos limitaremos a presentación de los principales modelos y a la exposición de los resultados más relevantes de los mismos. El lector interesado puede encontrar la mayor parte de las demostraciones en el capítulo 15 de Arnold (1981). En Carmona (2005) podemos encontrar brevemente descrito algunos modelos más complejos. En todo caso consideraremos únicamente diseños equilibrados.

Un factor aleatorio

Situémonos en las condiciones de un diseño completamente aleatorizado equilibrado con $\mathtt{a}$ niveles o valores para el factor y $\mathtt{m}$ observaciones por celda ($\mathtt{n} = \mathtt{a} \cdot \mathtt{m}$ datos en total). En ese caso, para cada $i = 1, \ldots, \mathtt{a}$ y $j = 1, \ldots, \mathtt{m}$, la observación (ij)-ésima se expresa mediante

$$Y_{ij} = \theta + \alpha_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma^2) \text{ independientes}$$

con la restricción $\sum_{i=1}^{\mathtt{a}} \alpha_i = 0$. En primer lugar, supongamos que los distintos valores o niveles del factor no están determinados de antemano a la realización del experimento sino que son $\mathtt{a}$ valores independientes de una variable aleatoria sobre

un conjunto $\mathtt{A}$ de gran tamaño. De esta forma, la influencia particular del nivel del factor sobre la media de la observación (el término α_i en el diseño completamente aleatorizado) debe considerarse una variable aleatoria real que supondremos en todo caso normal de media 0 y varianza σ_a^2. Supondremos también que los valores de la misma son independientes de los errores ε_{ij}. En definitiva, la observación Y_{ij} se expresa mediante

$$Y_{ij} = \theta + a_i + \varepsilon_{ij},$$

donde todas las variables a_i yε_{ij} son independientes y tales que

$$a_i \sim N(0, \sigma_a^2), \qquad \varepsilon_{ij} \sim N(0, \sigma_e^2).$$

Puede demostrarse que este modelo se deriva de otro, quizás más intuitivo, que indicamos brevemente: si el nivel del factor se escoge aleatoriamente, la media de la distribución de la variable respuesta para el nivel del factor escogido puede también considerarse una variable aleatoria real. Supongamos por hipótesis que dicha distribución es normal con una cierta media θ y varianza σ_a^2. En ese caso, las medias de los niveles seleccionados, $m_1, \ldots, m_{\mathtt{a}}$, constituyen una muestra aleatoria simple de la distribución $N(\theta, \sigma_a^2)$. Supongamos también que se da la independencia condicional entre todos los Y_{ij} dados $(m_1, \ldots, m_{\mathtt{a}})$ y que la distribución condicional de cada Y_{ij} dados $(m_1, \ldots, m_{\mathtt{a}})$ sigue un modelo $N(0, \sigma_e^2)$. En tales condiciones se verifican los supuestos del modelo expresado anteriormente (cuestión propuesta).

En todo caso, nótese que el factor aleatorio influye en la media de las variable respuesta si, y sólo si, $\sigma_a^2 > 0$. El espacio de parámetros del modelo es

$$\theta \in \mathbb{R}, \quad \sigma_a^2 \geq 0, \quad \sigma_e^2 > 0.$$

Hemos de advertir que, si bien dos observaciones correspondientes a diferentes niveles del factor son incorreladas (y por lo tanto independientes), no lo son dos observaciones Y_{ij} e $Y_{ij'}$ correspondientes al mismo nivel. Concretamente,

$$\mathtt{cov}[Y_{ij}, Y_{ij'}] = \sigma_a^2.$$

En todo caso, se verifica que $\mathtt{var}[Y_{ij}] = \sigma_a^2 + \sigma_e^2$, de ahí que el coeficiente de correlación lineal entre Y_{ij} e $Y_{ij'}$, denominado también coeficiente de correlación intraclásica, sea

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

A la vista de estas expresiones se entiende por qué el estudio de los diseños con efectos aleatorios se denomina frecuentemente **análisis de las componentes de la varianza**.

Si seguimos el guión desarrollado en el estudio del modelo lineal normal, el primer objetivo es encontrar un estadístico suficiente y completo para este modelo. Puede demostrarse que el estadístico (U, S_1^2, S_2^2), donde

$$U = \overline{y}_{..}, \quad S_1^2 = \mathtt{m}\sum_{i=1}^{\mathtt{a}}(\overline{y}_{i.} - \overline{y}_{..})^2, \quad S_2^2 = \sum_{I1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{m}}(Y_{ij} - \overline{y}_{1.})^2,$$

verifica dichas condiciones. Como corolario inmediato, tenemos que U y $[\mathtt{a}(\mathtt{m}-1)]^{-1}S_2^2$ son EIMV de θ y σ_e^2, respectivamente. También puede demostrarse que, por desgracia, no existe un estimador insesgado no negativo de σ_a^2. Esto supone un serio inconveniente dado que parece razonable en todo caso exigir a un estimador que tome valores en la imagen del estimando correspondiente. El criterio de máxima verosimilitud ofrece estimadores más apropiados.

Efectivamente, puede demostrarse que los estimadores de máxima verosimilitud de θ, σ_a^2 y σ_e^2 son, respectivamente,

$$U, \quad \text{máx}\left[\frac{S_1^2}{\mathtt{ma}} - \frac{S_2^2}{\mathtt{ma}(\mathtt{m}-1)}, 0\right], \quad \text{mín}\left[\frac{S_2^2}{\mathtt{a}(\mathtt{m}-1)}, \frac{S_1^2 + S_2^2}{\mathtt{ma}}\right].$$

Nótese que el hecho de que $S_1^2 < (\mathtt{m}-1)^{-1}S_2^2$ constituye una evidencia intuitiva de $\sigma_a^2 = 0$, lo cual supone un argumento adicional para decantarnos por el EMV en detrimento de cualquier estimador insesgado.

En lo que respecta al contraste de la hipótesis inicial $H_0 : \sigma_a^2 = 0$, se verifica que el test F a nivel α para contrastar la hipótesis inicial $\alpha_1 = \ldots = \alpha_{\mathtt{a}} = 0$ en el diseño completamente aleatorizado es también UMP-invariante a nivel α para el cotraste de la hipótesis H_0, aunque al hablar de invarianza nos refiramos a un grupo de transformaciones diferente al considerado en el diseño con efectos fijos.

Por último, en Arnold (1981) podemos encontrar intervalos de confianza para algunos estimandos. Concretamente θ, σ_e^2, σ_a^2/σ_e^2 y $\mathtt{m}\sigma_a^2 + \sigma_e^2$.

Dos efectos aleatorios

Siguiendo el mismo esquema del caso anterior, vamos a reformular el diseño equilibrado para dos factores con interacción suponiendo que los niveles de ambos factores se escojan de manera aleatoria. El modelo que proponemos consiste en expresar cada observación Y_{ijk}, $i = 1, \ldots, \mathtt{a}$, $j = 1, \ldots, \mathtt{b}$ y $k = 1 \ldots, \mathtt{m}$, mediante

$$Y_{ijk} = \theta + a_i + b_j + d_{ij} + \varepsilon_{ijk},$$

donde todas las variables del tipo a_i, b_j, d_{ij} y ε_{ijk} son independientes y tales que

$$a_i \sim N(0, \sigma_a^2), \quad b_j \sim N(0, \sigma_b^2), \quad d_{ij} \sim N(0, \sigma_d^2), \quad \varepsilon_{ijk} \sim N(0, \sigma_e^2)$$

En este caso, el espacio de parámetros es

$$\theta \in \mathbb{R}, \qquad \sigma_a^2 \geq 0, \qquad \sigma_b^2 \geq 0, \qquad \sigma_d^2 \geq 0, \qquad \sigma_e^2 > 0$$

Al igual que sucede en el caso de un factor aleatorio, estas condiciones pueden deducirse a partir de otras más intuitivas expresadas en términos de distribuciones marginales y condicionales (ver Arnold (1981). Si se denota

$$U = \overline{y}_{\cdots}, \qquad S_1^2 = \mathtt{mb}\sum_{i=1}^{\mathtt{a}}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots})^2, \qquad S_2^2 = \mathtt{mb}\sum_{j=1}^{\mathtt{a}}(\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdots})^2,$$

$$S_3^2 = \mathtt{m}\sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdots})^2,$$

$$S_4^2 = \sum_{i=1}^{\mathtt{a}}\sum_{j=1}^{\mathtt{b}}\sum_{k=1}^{\mathtt{m}}(Y_{ijk} - \overline{y}_{ij\cdot})^2,$$

puede demostrarse que el estadístico $(U, S_1^2, S_2^2, S_3^2, S_4^2)$ es suficiente y completo para el modelo considerado. Como corolario obtenemos los EIMV para θ y σ_e^2, lo cual no es de excesiva utilidad puesto que cualquier estimador insesgado de alguno de los parámetros realmente interesantes, σ_a^2, σ_b^2 o σ_d^2 puede tomar valores negativos. Además, obtener los EMV para los mismos resulta demasiado complicado, con lo que hemos de conformarnos con proponer los siguientes estimadores sin aportar una clara justificación teórica para los mismos (salvo en el caso de θ y σ_e^2):

$$\hat{\theta} = \overline{y}_{\cdots}, \qquad \hat{\sigma}_a^2 = \text{máx}\left[\frac{S_1^2}{\mathtt{mb(a-1)}} - \frac{S_3^2}{\mathtt{mb(a-1)(b-1)}}, 0\right],$$

$$\hat{\sigma}_b^2 = \text{máx}\left[\frac{S_2^2}{\mathtt{ma(b-1)}} - \frac{S_3^2}{\mathtt{ma(b-1)(a-1)}}, 0\right],$$

$$\hat{\sigma}_d^2 = \text{máx}\left[\frac{S_3^2}{\mathtt{m(b-1)(a-1)}} - \frac{S_4^2}{\mathtt{mb(a-1)}}, 0\right], \qquad \hat{\sigma}_e^2 = \frac{S_4^2}{\mathtt{ab(m-1)}}$$

Es fácil construir intervalos de confianza exactos para diversos estimandos, pero no es posible para los parámetros de mayor interés: σ_a^2, σ_b^2 y σ_b^2. En Arnold (1981) se presentan intervalos aproximados.

Las principales hipótesis iniciales a contrastar son $H_0^A : \sigma_a^2 = 0$, $H_0^B : \sigma_b^2 = 0$ y $H_0^{AB} : \sigma_d^2 = 0$, así como todas sus posibles intersecciones. El modelo considerado y todos estos problemas de contrastes de hipótesis son invariantes ante el grupo de transformaciones $\{g_{u,w} : \ u \in \mathbb{R}, \ w > 0\}$ que actúan mediante $g_{u,w}(Y_{ijk}) =$

$wY_{ijk} + u$. Aunque no existe un estadístico UMP-invariante a nivel α para ninguno de los contrastes considerados, podemos optar por los siguientes tests invariantes a nivel α que describimos a continuación. Se consideran los siguientes estadísticos, F_1, F_2, y F_3 así como las funciones δ_1, δ_2 y δ_3 sobre el espacio de parámetros con valores en $[1, +\infty)$ siguientes:

$$F_1 = \frac{(\mathtt{b}-1)S_1^2}{S_2^2}, \qquad F_2 = \frac{(\mathtt{a}-1)S_2^2}{S_3^2}, \qquad F_3 = \frac{\mathtt{ab}(\mathtt{m}-1)S_3^3}{(\mathtt{a}-1)(\mathtt{b}-1)S_4^2}$$

$$\delta_1 = \frac{\mathtt{mb}\sigma_a^2 + \mathtt{m}\sigma_d^2 + \sigma_e^2}{\mathtt{m}\sigma_d^2 + \sigma_e^2}, \qquad \delta_2 = \frac{\mathtt{ma}\sigma_b^2 + \mathtt{m}\sigma_d^2 + \sigma_e^2}{\mathtt{m}\sigma_d^2 + \sigma_e^2}, \qquad \delta_3 = \frac{\mathtt{m}\sigma_d^2 + \sigma_e^2}{\sigma_e^2}$$

Puede demostrarse fácilmente que (F_1, F_2, F_3) y $(\delta_1, \delta_2, \delta_3)$ son sendos invariantes máximales para el espacio de observaciones y el de parámetros, respectivamente, correspondientes al grupo de trasformaciones anterior. Además, las distribuciones marginales de los tres estadísticos son las siguientes:

$$\delta_1^{-1}F_1 \sim F_{\mathtt{a}-1,(\mathtt{a}-1)(\mathtt{b}-1)}, \qquad \delta_2^{-1}F_2 \sim F_{\mathtt{b}-1,(\mathtt{a}-1)(\mathtt{b}-1)}, \qquad \delta_3^{-1}F_3 \sim F_{(\mathtt{a}-1)(\mathtt{b}-1),\mathtt{ab}(\mathtt{m}-1)}$$

Dado que los parámetros δ_2 y δ_3 no dependen de σ_a^2 y que $\delta_1 = 1$ si, y sólo si, $\sigma_a^2 = 0$, el siguiente test de hipótesis invariante a nivel α para contrastar la hipótesis inicial H_0^A puede resultar razonable:

$$\Phi_1(Y) = \begin{cases} 1 & \text{si} \quad F_1(Y) > F^{\alpha}_{\mathtt{a}-1,(\mathtt{a}-1)(\mathtt{b}-1)} \\ 0 & \text{si} \quad F_1(Y) \leq F^{\alpha}_{\mathtt{a}-1,(\mathtt{a}-1)(\mathtt{b}-1)} \end{cases}$$

Razonando de manera completamente análoga, obtenemos los siguientes tests a nivel α para constrastar las hipótesis iniciales H_0^B y H_0^{AB}, respectivamente:

$$\Phi_2(Y) = \begin{cases} 1 & \text{si} \quad F_2(Y) > F^{\alpha}_{\mathtt{b}-1,(\mathtt{a}-1)(\mathtt{b}-1)} \\ 0 & \text{si} \quad F_2(Y) \leq F^{\alpha}_{\mathtt{b}-1,(\mathtt{a}-1)(\mathtt{b}-1)} \end{cases}$$

$$\Phi_3(Y) = \begin{cases} 1 & \text{si} \quad F_3(Y) > F^{\alpha}_{(\mathtt{a}-1)(\mathtt{b}-1),\mathtt{ab}(\mathtt{m}-1)} \\ 0 & \text{si} \quad F_3(Y) \leq F^{\alpha}_{(\mathtt{a}-1)(\mathtt{b}-1),\mathtt{ab}(\mathtt{m}-1)} \end{cases}$$

La veracidad de la hipótesis inicial H_0^{AB} equivaldría a la nulidad de las variables d_{ij}, con lo que estaríamos hablando de un modelo sin interacciones entre los factores A y B. Este diseño, más sencillo, se desarrolla brevemente en en Carmona (2005). Para determinar si el factor aleatorio A influye en la media de la variable respuesta podemos contrastar la hipótesis inicial H_0^A en el modelo sin interacción, siempre y cuando se haya determinado previamente que no existe interacción entre los factores.

En todo caso podemos optar por contrastar en el modelo completo la hipótesis inicial $H_0^{A,AB}: \sigma_a^2 = \sigma_d^2 = 0$. Nótese que dicha hipótesis se verifica si, y sólo si, $\delta_1\delta_3 = 1$, y que $(\delta_1\delta_3)^{-1}F_1F_3 \sim F_{\mathtt{a}-1,\mathtt{ab}(\mathtt{m}-1)}$. Por lo tanto, el siguiente test es invariante a nivel α para contrastar la hipótesis inicial $H_0^{A,AB}$

$$\Phi_4(Y) = \begin{cases} 1 & \text{si} \quad F_1F_3(Y) > F^{\alpha}_{\mathtt{a}-1,\mathtt{ab}(\mathtt{m}-1)} \\ 0 & \text{si} \quad F_1F_3(Y) \leq F^{\alpha}_{\mathtt{a}-1,\mathtt{ab}(\mathtt{m}-1)} \end{cases}$$

Respecto al factor B procederíamos de manera completamente análoga. Nótese que los tests para contrastar las hipótesis iniciales H_0^{AB}, $H_0^{A,AB}$ y $H_0^{B,AB}$ en el modelo con efectos aleatorios coinciden con los que se propuestos para las hipótesis análogas en el modelo con efectos fijos. No ocurre lo mismo con las hipótesis iniciales H_0^A y H_0^B. Otra diferencia notable respecto a al modelo con efectos fijos es el hecho de que el modelo con interacción para efectos aleatorios sigue siendo viable con $\mathtt{m} = 1$ (una observación con celda), aunque el test ϕ_3 para contrastar la hipótesis inicial $\sigma_d^2 = 0$ no tendría sentido pues S_4^2 sería nulo.

Dos efectos mixtos

Para terminar esta sección, consideraremos el estudio de dos factores con interacción siendo uno de ellos aleatorio y el otro fijo. En la literatura se recogen dos versiones de este diseño, aunque veremos que son muy similares. El primer modelo consiste en expresar cada observación Y_{ijk}, $i = 1, \ldots, \mathtt{a}$, $j = 1, \ldots, \mathtt{b}$ y $k = 1 \ldots, \mathtt{m}$, mediante

$$Y_{ijk} = \theta + \alpha_i + b_j + d_{ij} + \varepsilon_{ijk},$$

donde $\sum_{i=1}^{\mathtt{a}} \alpha_i = 0$ y todas las variables del tipo b_j, d_{ij} y ε_{ijk} son independientes y tales que

$$b_j \sim N(0, \sigma_b^2), \quad d_{ij} \sim N(0, \sigma_d^2), \quad \varepsilon_{ijk} \sim N(0, \sigma_e^2)$$

En este caso, el espacio de parámetros es

$$\theta \in \mathbb{R}, \qquad (\alpha_1, \ldots, \alpha_{\mathtt{a}})' \in \langle 1_{\mathtt{a}} \rangle^{\perp}, \qquad \sigma_b^2 \geq 0, \qquad \sigma_d^2 \geq 0, \qquad \sigma_e^2 > 0$$

Se demuestra en Arnold (1981), que el estadístico $(\overline{y}_{1,\cdot\cdot}, \ldots, \overline{y}_{\mathtt{a},\cdot\cdot}, S_2^2, S_3^2, S_4^2)$, con S_2^2, S_3^2 y S_4^2 definidos como en el modelo anterior, es suficiente y completo. Por lo tanto, los EIMV de θ, α_i y σ_e^2 son, respectivamente, $\overline{y}_{\cdots}$, $\overline{y}_{i\cdot\cdot}$ y $[\mathtt{ab}(\mathtt{m}-1)]^{-1}S_4^2$. Nuevamente, no existen estimadores insesgados no negativos de los parámetros σ_b^2 y σ_d^2. No obstante, en la literatura se recogen los siguientes estimadores:

$$\hat{\sigma}_b^2 = \text{máx}\left\{\frac{1}{\mathtt{ma}}\left[\frac{S_2^2}{\mathtt{b}-1} - \frac{S_3^2}{(\mathtt{a}-1)(\mathtt{b}-1)}\right], 0\right\},$$

$$\hat{\sigma}_d^2 = \text{máx}\left\{\frac{S_3^2}{(\mathtt{a}-1)(\mathtt{b}-1)} - \frac{S_4^2}{\mathtt{ab}(\mathtt{m}-1)}, 0\right\}$$

Podemos construir fácilmente intervalos de confianza exactos para distintos estimandos (ver Arnold (1981)), no así para los parámetros σ_b^2 y σ_d^2, los más interesantes. En Arnold (1981) se construye a su vez una familia de intervalos de confianza simultáneos para los estimandos de la forma $\sum_{i=1}^{\mathtt{a}} w_i\alpha_i$, donde $\sum_{i=1}^{\mathtt{a}} w_i = 0$.

En cuanto al problema de contraste de hipótesis, hemos de distinguir los contrastes relativos a los parámetros σ_b^2 y σ_d^2 de los relativos a $\alpha_1, \ldots, \alpha_{\mathtt{a}}$. En ambos casos proponemos tests invariantes a nivel α pero respecto a distintos grupos de transformaciones (ver detalles en Arnold (1981)). En el primer caso, se proponen los test ϕ_2, ϕ_3 y ϕ_4 definidos en el modelo anterior para contrastar las hipótesis iniciales $\sigma_b^2 = 0$, $\sigma_d^2 = 0$ y $\sigma_b^2 = \sigma_d^2 = 0$, respectivamente. Para contrastar la hipótesis inicial $\alpha_1 = \ldots = \alpha_{\mathtt{a}} = 0$ se propone asimismo el test ϕ_1 del modelo anterior. En definitiva, se utilizan los mismos tests y se justifican también por invarianza, pero ante grupos de transformaciones distintas.

El segundo modelo consiste en expresar las observaciones Y_{ijk} de la forma

$$Y_{ijk} = \theta^* + \alpha_i^* + b_j^* + d_{ij}^* + \varepsilon_{ijk}^*$$

donde $\sum_{i=1}^{\mathtt{a}} \alpha_i = 0$ y b_j, d_{ij}^* y ε_{ijk} son variables aleatorias. Supondremos que, para todo $i = 1, \ldots, \mathtt{a}$, $\sum_{i=1}^{\mathtt{a}} d_{ij}^* = 0$; que $d_{ij} \sim N(0, \tau_d^2)$ para todo i y j; además, si de denota $\mathtt{d}_j^* = (d_{1j}^*, \ldots, d_{\mathtt{a}j}^*)'$, se supondrá que todos la b_j^*, ε_{ijk}^* y $\mathtt{d}_j^*$ son independientes; por último, se supone que

$$b_j \sim N(0, \tau_b^2), \qquad \varepsilon_{ijk}^* \sim N(0, \tau_e^2).$$

Los parámetros del modelo son pues

$$\theta^* \in \mathbb{R}, \qquad (\alpha_1^*, \ldots, \alpha_{\mathtt{a}}^*)' \in \langle 1_{\mathtt{a}}\rangle^{\perp}, \qquad \tau_b^2 \geq 0, \qquad \tau_d^2 \geq 0, \qquad \tau_e^2 > 0$$

Es fácil probar que, en estas condiciones, la distribución de $\mathtt{d}_j^*$ es la siguiente:

$$\mathtt{d}_j^* \sim N_{\mathtt{a}}\left(0_{\mathtt{a}}, \begin{pmatrix} 1 & -\frac{1}{\mathtt{a}-1} & \cdots & -\frac{1}{\mathtt{a}-1} \\ -\frac{1}{\mathtt{a}-1} & 1 & & -\frac{1}{\mathtt{a}-1} \\ -\frac{1}{\mathtt{a}-1} & -\frac{1}{\mathtt{a}-1} & \cdots & 1 \end{pmatrix}\right)$$

En Arnold (1981) se deducen los supuestos de este modelo a partir de una serie de hipótesis expresadas en términos más intuitivos, de manera análoga a los diseños con uno y dos factores aleatorios. También se prueba que el primer modelo mixto puede considerarse un caso particular del segundo salvo en el detalle de que debe imponerse

una restricción adicional en el espacio de parámetros. Concretamente, un modelo mixto tipo 1 con parámetros θ, $\alpha_1, \ldots, \alpha_{\mathtt{a}}$, σ_b^2, σ_d^2 y σ_e^2 equivale a un modelo mixto tipo 2 con parámetros

$$\theta^* = \theta, \quad \alpha_i^* = \alpha_i, \quad \tau_b^2 = \sigma_b^2 + \frac{1}{\mathtt{a}}\sigma_d^2, \quad \tau_d^2 = \frac{\mathtt{a}-1}{\mathtt{a}}\sigma_d^2, \quad \tau_e^2 = \sigma_e^2.$$

Por lo tanto, debe verificarse en todo caso que

$$(\mathtt{a}-1)\tau_b^2 \geq \tau_d^2.$$

De no ser por esta excepción podríamos afirmar que el modelo 2 es pues más general que el 1. Dada esta gran similitud, los resultados obtenidos para ambos modelos así como la propia forma de demostrarlos son muy similares. Los EIMV para θ^*, α_i^* y τ_e^2 son los mismos que para θ, α_i y σ_e^2 en el modelo 1. Para τ_b^* y τ_d^* proponemos los siguientes estimadores:

$$\hat{\tau}_b^2 = \text{máx}\left\{\frac{1}{\mathtt{ma}}\left[\frac{S_2^2}{\mathtt{a}-1} - \frac{S_4^2}{\mathtt{ab}(\mathtt{m}-1)}\right], 0\right\},$$

$$\hat{\tau}_d^2 = \text{máx}\left\{\frac{\mathtt{a}}{\mathtt{m}(\mathtt{a}-1)}\left[\frac{S_3^2}{(\mathtt{a}-1)(\mathtt{b}-1)} - \frac{S_4^2}{\mathtt{ab}(\mathtt{m}-1)}\right], 0\right\}$$

Respecto a la búsqueda de intervalos de confianza, estamos en la misma situación del modelo anterior: podemos construir una familia de intervalos de confianza simultáneos para los estimandos de la forma $\sum_{i=1}^{\mathtt{a}} w_i\alpha_i$, con $\sum_{i=1}^{\mathtt{a}} w_i = 0$, pero no intervalos de confianza exactos para τ_b^2 y τ_d^2. Por otra parte, todos los contrastes considerados en el modelo anterior se resuelven en éste mediante los mismos tests salvo el contraste de la hipótesis inicial $H_0^B : \tau_b^2 = 0$. El en este caso se rechazará dicha hipótesis cuando $F_1 > F^{\alpha}_{\mathtt{b}-1,\mathtt{ab}(\mathtt{m}-1)}$.

Recapitulando, hemos vistos en esta sección que el hecho de considerar efectos de tipo aleatorio supone, si acaso, sutiles modificaciones en lo que respecta al contrastes de las hipótesis iniciales más interesantes desde el punto de vista práctico. No obstante, se abre la puerta a la estimación de nuevos parámetros, que podemos denominar componentes de la varianza, que no tienen sentido en un modelo con efectos fijos. También hemos de recalcar que la estimación de los mismos presenta serias dificultades desde el punto de vista teórico.

Cuestiones propuestas

1. Probar que, en el diseño completamente aleatorizado, la condición de Huber equivale a que $\mathtt{n}_i \to \infty$, para todo $i = 1, \ldots, \mathtt{a}$.

2. Explicitar el algoritmo de Box-Cox para conseguir normalidad e igualdad de varianzas en un diseño completamente aleatorizado.

3. Obtener el estadístico de contraste (6.9).

4. Obtener (6.10) en el diseño completamente aleatorizado y equilibrado.

5. Expresar el EIMV de σ^2 para el modelo (6.14) a partir de los EIMV de las varianzas para los $\mathtt{a}$ modelos de regresión considerados (uno para cada nivel del factor).

6. Obtener, a partir de (3.12), un intervalo de confianza a nivel $1-\alpha$ para la media de una distribución normal, conocida una muestra aleatoria simple de tamaño $\mathtt{n}$ de la misma.

7. ¿En qué se traduce la condición de Huber en el modelo bifactorial equilibrado?

8. Considerar un modelo bifactorial equilibrado con $\mathtt{m}$ datos por celda, tres niveles para el primer factor y cuatro para el segundo. Probar que los parámetros θ, α_1, α_2, β_1, β_2, β_3, $(\alpha\beta)_{11}$, $(\alpha\beta)_{12}$, $(\alpha\beta)_{13}$, $(\alpha\beta)_{21}$, $(\alpha\beta)_{22}$ y $(\alpha\beta)_{23}$ constituyen, por ese orden, los coeficientes de regresión respecto a la matriz de diseño $\mathtt{X}$ siguiente[6]

$$\mathtt{X}=\left(\begin{array}{c|cc|ccc|ccc|ccc}
1_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} \\
1_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} \\
1_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} \\
1_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} \\
\hline
1_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} \\
1_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} \\
1_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} \\
1_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} \\
\hline
1_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & -1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & -1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} \\
1_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & -1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & -1_{\mathtt{m}} & 0_{\mathtt{m}} \\
1_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & 1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & -1_{\mathtt{m}} & 0_{\mathtt{m}} & 0_{\mathtt{m}} & -1_{\mathtt{m}} \\
1_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & -1_{\mathtt{m}} & 1_{\mathtt{m}} & 1_{\mathtt{m}} & 1_{\mathtt{m}} & 1_{\mathtt{m}} & 1_{\mathtt{m}} & 1_{\mathtt{m}}
\end{array}\right)$$

Es decir, que el diseño anterior puede formalizarse mediante un modelo de regresión lineal, $Y=\mathtt{X}\beta+\mathcal{E}$, respecto a unas variables ficticias que indican los

[6] Los términos $1_{\mathtt{m}}$ y $0_{\mathtt{m}}$ denotan los vectores de $\mathbb{R}^{\mathtt{m}}$ cuyas componentes son todas iguales a 1 y 0, respectivamente.

niveles de los factores a los que corresponde cada unidad experimental, junto con otras variables, construidas como producto de variables ficticias. Como indicación tener en cuenta que un vector de $V_1|\langle 1_{\mathtt{n}}\rangle$ se expresa mediante $\sum_{i=1}^{3} x_i \mathtt{v}_{i\cdot}$ con $\sum_{i=1}^{3} x_i = 0$, lo cual equivale a $\sum_{i=1}^{2} x_i(\mathtt{v}_i - \mathtt{v}_3)$.

9. Probar (6.30) y (6.31), (6.32) y (6.33).

10. Construir las familias de intervalos de confianza de Bonferroni, Tuckey y Scheffé para el conjunto de $\{\beta_j - \beta_{j'} : \ j \neq j'\}$.

11. Obtener (6.35) y las familias de intervalos de confianza de Bonferroni, Tuckey y Scheffé parael modelo bifactorial sin interacción y, en particular, para el diseño en bloques aleatorizados.

12. Obtener las comparaciones múltiples para el modelo trifactorial según los métodos de Bonferroni, Tuckey y Scheffé.

13. Proponer un algoritmo para contrastar en el diseño trifactorial equilibrado si factor f_A tiene influencia, por término medio, en la media de la variable respuesta.

14. Obtener las sumas cuadráticas que corresponden a un modelo con cuatro factores equilibrados.

15. Obtener las comparaciones múltiples para los efectos del factor principal en el diseño jerárquico.

16. Diseñar de forma clara un algoritmo para contrastar la influencia del factor principal en un diseño anidado equilibrado.

17. Obtener las comparaciones múltiples para los tratamientos en el diseño de cuadrados latinos.

18. Realizar una descomposición ortogonal de V para un diseño con cuatro factores equilibrado.

19. Probar que las condiciones del modelo de análisis de la varianza con un factor aleatoria se deriva de los supuestos intuitivos expuestos en la sección 9.

Capítulo 7

Modelo lineal de rango no completo

Hemos de advertir que el objeto de este capítulo no es un nuevo modelo si nos ceñimos a la definición de tal recogida en (9.31), sino una particular parametrización del modelo lineal estudiado en el capítulo 1. Este nuevo planteamiento se traducirá en la práctica en un método alternativo aunque equivalente para resolver los problemas ya estudiados, si bien puede resultar especialmente adecuado a la hora de afrontar análisis de la varianza complejos y no equilibrados o con vistas a su implementación en un programa informático.

7.1. El modelo

Efectivamente, según la definición de modelo estadístico dada en (9.31), el modelo lineal de rango no completo que estudiaremos a continuación no se distingue formalmente del modelo lineal definido en (3.1) y estudiado en el capítulo 3. En este capítulo estamos asumiendo pues una acepción diferente del término. Concretamente, estamos considerando como modelo estadístico un par compuesto por un espacio medible $(\Omega, \mathcal{A})$ y una aplicación sobreyectiva $P : \Theta \to \mathcal{P}$, siendo Θ un conjunto no vacío y $\mathcal{P}$ una familia de probabilidades sobre $(\Omega, \mathcal{A})$, que se denotará, en consecuencia, por $\{P_\theta : \theta \in \Theta\}$. Por lo tanto, desde esta nueva perspectiva, el parámetro deja de ser contingente para convertirse en una componente esencial del modelo.

Concretamente, sabemos que el modelo lineal puede parametrizarse por $\mu \in V$ y $\sigma^2 > 0$. No obstante, podemos considerar una base $\mathbf{X}$ de V y reemplazar el parámetros μ por sus coordenadas β respecto a dicha base, lo cual no supondrá ninguna alteración del modelo según la primera acepción. El parámetro β se expresa a partir de μ

mediante

$$\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mu. \tag{7.1}$$

Téngase en cuenta que $\mathbf{X}'\mathbf{X}$ es una matriz cuadrada de orden $\mathtt{dim}V$ y rango $\mathtt{dim}V$, por lo que es invertible.

Sin embargo, según la segunda acepción del término, el cambio de parámetro implica un cambio en el modelo estadístico. En este nuevo marco tiene sentido hablar de la siguiente generalización: consideraremos que la matriz $\mathbf{X}$ no es una base del subespacio que generan sus columnas, es decir, que sus columnas no tienen por qué ser linealmente independientes. En definitiva, estaremos hablando de una matriz $\mathbf{X} \in \mathcal{M}_{\mathbf{n}\times s}$ cuyo rango puede no ser completo. Esta generalización conlleva una clara complicación: no existe una identificación entre los parámetros β y μ, pues pueden existir, en principio, distintas soluciones al sistema $\mu = \mathbf{X}\beta$. Nótese que la expresión (7.1) no es válida en general, pues, si el rango de $\mathbf{X}$ no es completo, la matiz $\mathbf{X}'\mathbf{X}$ no es invertible.

Este planteamiento puede resultar adecuado para abordar la resolución de cualquier sistema de ecuaciones lineales desde un punto de vista estadístico. Sin embargo, debemos preguntarnos en qué situaciones de interés real puede considerarse un modelo parametrizado por una matriz de rango no completo. Podemos citar tres ejemplos. En primer lugar, un problema de regresión lineal cuyos vectores explicativos sean linealmente dependientes, o bien cuando el número de éstos sea mayor o igual que el de unidades experimentales. En ambas situaciones *patológicas*, el rango de la matriz $\mathbf{X}$ no puede ser completo. También puede aparecer una matriz $\mathbf{X}$ de rango no completo en cualquier diseño de experimentos, como ya dijimos en el capítulo anterior, aunque profundizaremos en este tema en la tercera parte del capítulo. En el capítulo 1 en la parte final de éste podemos encontrar una discusión más detallada sobre la conveniencia de utilizar este tipo de modelo.

Dado que la matriz $\mathbf{X}'\mathbf{X}$ no es necesariamente invertible, haremos uso de una generalización del concepto de inversa de ua matriz que abordaremos en profundidad en la primera parte de la sección. Este estudio, de carácter matricial, podría haberse abordado en la primera sección del segundo apéndice, pero lo hemos incluido aquí el por no alargar en exceso el Apéndice. La segunda parte está dedicada al planteamiento y resolución de los problemas de Estimación Puntual y Contraste de Hipótesis cuando el rango de $\mathbf{X}$ no es completo.

7.2. Inversa Generalizada de una Matriz

Se desarrolla aquí un concepto que, como su propio nombre indica, viene a generalizar el de inversa de una matriz cuadrada no singular, aunque es aplicable a cualquier matriz. Será de interés a la hora de determinar el conjunto de soluciones de cualquier sistema de ecuaciones lineales compatible, sea o no determinado, lo cual le confiere gran trascendencia en el estudio que llevamos a cabo en este capítulo.

En lo que sigue, A denotará una matriz de $\mathcal{M}_{m\times p}$ de rango r. Se dice que una matriz $G \in \mathcal{M}_{p\times m}$ es una inversa generalizada de A^- cuando verifica

$$AGA = A. \tag{7.2}$$

El subconjunto de $\mathcal{M}_{p\times m}$ constituido por todas las inversas generalizadas de A se denota por A^-. Desde luego, es inmediato comprobar que, si $p = m$ y A es no singular, el conjunto A^- está constituido únicamente por la matriz inversa de A, en cuyo caso nos permitiremos el abuso de denotar $A^- = A^{-1}$. En general, el conjunto A^- no es vacío. Para probarlo, basta considerar una descomposición de A según (9.5). En ese caso, la matriz G definida mediante

$$G = M\left(\begin{array}{c|c} D^{-1} & 0 \\ \hline 0 & 0 \end{array}\right)N'$$

verifica trivialmente la condición (7.2). Podemos ser aún más precisos y explicitar un algoritmo para la obtención de una inversa generalizada. Supondremos, en una primera instancia, que la matriz A puede expresarse mediante

$$A = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right),$$

donde A_{11} es una matriz cuadrada de orden r e invertible. En ese caso, la matriz $G \in \mathcal{M}_{p\times m}$ definida mediante

$$G = \left(\begin{array}{c|c} A_{11}^{-1} & 0 \\ \hline 0 & 0 \end{array}\right)$$

es una inversa generalizada. Para probarlo basta tener en cuenta que

$$AGA = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{12}A_{11}^{-1}A_{12} \end{pmatrix}.$$

Tener en cuenta que existe una matriz $K \in \mathcal{M}_{(m-r)\times r}$ tal que $A_{21} = KA_{11}$ y $A_{12} = KA_{22}$, de lo cual se sigue que $A_{22} = A_{21}A_{11}^{-1}A_{12}$. En el caso general, existe una

permutación Φ_1 de las filas y otra Φ_2 de las columnas, tal que la matriz $B = \Phi_1 A \Phi_2$ puede expresarse de la forma anterior. En ese caso, la matriz $\tilde{G}$ definida mediante

$$\tilde{G} = \left(\begin{array}{c|c} B_{11}^{-1} & 0 \\ \hline 0 & 0 \end{array}\right)$$

es una inversa generalizada de B. Dado que tanto Φ_1 como Φ_2 son matrices ortogonales, puede probarse fácilmente que $\Phi_2 \tilde{G} \Phi_1 \in A^-$. El algoritmo consiste pues en reordenar la matriz A para obtener un menor invertible de orden r en la primera posición, invertirlo, trasponerlo, rellenar con 0 el resto hasta completar una matriz $m \times p$, realizar las permutaciones de filas y columnas inversas y volver a trasponer.

De la definición de inversa generalizada se siguen una serie de propiedades inmediatas. Por ejemplo, es obvio que, si G es una inversa generalizada de A, G' lo será de A'. En particular, si A es una matriz cuadrada y simétrica y G es una inversa generalizada, entonces también lo es G'.

A continuación aprovecharemos este concepto para caracterizar el conjunto de soluciones de un sistema de ecuaciones lineales $Ax = \mathtt{y}$, donde $\mathtt{y} \in \mathbb{R}^m$ y $x \in \mathbb{R}^p$, siempre y cueando sea compatible, es decir, que exista alguna solución.

Lema 7.1.

Dada $G \in \mathcal{M}_{p\times m}$, las dos proposiciones siguientes son equivalentes:

(i) $\big[Ax = \mathtt{y}$ es compatible$\big] \Rightarrow \big[\mathtt{Gy}$ es una solución particular$\big]$

(ii) $G \in A^-$.

Demostración.

Supongamos que (i) es cierto y denótense por a_j, $j = 1, \ldots, p$, las columnas de A. En ese caso, la ecuación $Ax = a_j$ es compatible. Por lo tanto, alguna solución se expresará mediante $\mathtt{x} = Ga_j$. En consecuencia, $AGa_j = a_j$, para todo $j = 1, \ldots, p$ y (7.2) se verifica. Recíprocamente, si se verifica (7.2) y $A\mathtt{x} = \mathtt{y}$, entonces $AGA\mathtt{x} = AG\mathtt{y}$. Luego, $A(G\mathtt{y}) = \mathtt{y}$. Por lo tanto, el vector $\tilde{\mathtt{x}} = G\mathtt{y}$ es solución a la ecuación $Ax = \mathtt{y}$.

∎

Dadas $A \in \mathcal{M}_{m\times p}$, $\mathtt{y} \in \mathbb{R}^m$ tal que la ecuación $Ax = \mathtt{y}$ es compatible y $G \in A^-$, se define el siguiente subconjunto de $\mathbb{R}^p$

$$\mathcal{S}_{A,\mathtt{y}} = \{G\mathtt{y} + (GA - \mathtt{Id}_{p\times p})z : \ z \in \mathbb{R}^p\}.$$

Teorema 7.2.

En esas condiciones, $\mathcal{S}_{A,\mathtt{y}}$ es el conjunto de las soluciones a la ecuación $Ax = \mathtt{y}$. En particular, $S_{A,\mathtt{y}}$ no depende de la matriz $G \in A^-$ escogida.

Demostración.
Probar que cualquier elemento de $\mathcal{S}_{A,\mathtt{y}}$ es solución de la ecuación $Ax = \mathtt{y}$ es trivial. Recíprocamente, $A\mathtt{x} = \mathtt{y}$ implica que $\mathtt{x} = G\mathtt{y} + (GA - \mathtt{Id})(GA - \mathtt{Id})\mathtt{x}$.

∎

Describiremos a continuación el espacio $\mathcal{S}_{A,\mathtt{y}}$ de soluciones. Sea $H = GA \in \mathcal{M}_{p\times p}$. En ese caso, se verifica que

$$\mathcal{S}_{A,\mathtt{y}} = G\mathtt{y} + \mathcal{S}_{A,0} = G\mathtt{y} + \langle H - \mathtt{Id}\rangle, \tag{7.3}$$

que se trata de una subvariedad afín de $\mathbb{R}^p$. Respecto a la dimensión de la misma, se tiene lo siguiente:

Lema 7.3.

H verifica que $H^2 = H$ y $\mathtt{rg}(H) = r$, que los subespacios lineales $\langle H\rangle$ y $\langle \mathtt{Id} - H\rangle$ son perpendiculares y que $\mathtt{rg}(\mathtt{Id} - H) = p - r$.

Demostración.
Que $H^2 = H$ se sigue de (7.2). Además, dado que $\mathtt{rg}(GA) \leq \text{mín}\{\mathtt{rg}(G), \mathtt{rg}(A)\}$, se tiene que $\mathtt{rg}(H) \leq \mathtt{rg}(A)$. Aplicando el mismo razonamiento a $AH = AGA = A$, se deduce que $\mathtt{rg}(H) \geq \mathtt{rg}(A)$. Por otra parte, dado $z \in \mathbb{R}^p$, se deduce de lo anterior que $\langle(\mathtt{Id} - H)z, Hz\rangle = 0$, luego, $\langle H\rangle \perp \langle \mathtt{Id} - H\rangle$. Dado que $\langle \mathtt{Id} - H\rangle \oplus \langle H\rangle = \mathbb{R}^p$, se concluye.

∎

Teorema 7.4.

Dada $A \in \mathcal{M}_{m\times p}$ de rango r, se verifica

(i) *El espacio de soluciones $\mathcal{S}_{A,0}$ es un subespacio $(p - r)$-dimensional de $\mathbb{R}^p$. Por lo tanto, existen $p-r$ soluciones linealmente independientes para la ecuación $Ax = 0$.*

(ii) *Dado $\mathtt{y} \in \mathbb{R}^m\backslash\{0\}$ tal que la ecuación $Ax = \mathtt{y}$ es compatible, el espacio de soluciones $S_{A,\mathtt{y}}$ constituye una subvariedad afín $(p-r)$-dimensional de $\mathbb{R}^p$. Además, existen $p - r + 1$ soluciones linealmente independientes para la ecuación $Ax = \mathtt{y}$.*

Demostración.
El apartado (i) y la primera parte de (ii) se siguen directamente del lema anterior. Falta por demostrar que existen $p - r + 1$ soluciones lineales independientes para $Ax = \mathtt{y}$. Primeramente, $G\mathtt{y}$ es linealmente independiente de cualquier vector de $\mathcal{S}_{A,0} = \langle H - \mathtt{Id}\rangle$ pues, de lo contrario, se verificaría que $AG\mathtt{y} = 0$ y, dado que $\mathtt{y} = A\mathtt{x}$ para algún $\mathtt{x}$, ello implicaría, por (7.2), que $\mathtt{y} = 0$, en contra de la hipótesis. Por

lo tanto, si $\{\mathtt{x}_{0,1},\ldots,\mathtt{x}_{0,p-r}\}$ denota una base de $\mathcal{S}_{A,0}$, se trata de comprobar que $\{G\mathtt{y}, G\mathtt{y}+\mathtt{x}_{0,1},\ldots,G\mathtt{y}+\mathtt{x}_{0,p-r}\}$ es un conjunto de soluciones linealmente independientes. Efectivamente, dada una familia de números reales $\{\lambda_0,\lambda_1,\ldots,\lambda_{p-r}\}$ tal que $\lambda_0 G\mathtt{y}+\sum_{i=1}^{p-r}\lambda_i(G\mathtt{y}+\mathtt{x}_{0,i})=0$, se tiene que $\big(\sum_{i=0}^{p-r}\lambda_i\big)G\mathtt{y}+\sum_{i=1}^{p-r}\lambda_i\mathtt{x}_{0,i}=0$, lo cual implica $\lambda_i=0$, para todo $i=0,1,\ldots,p-r$.

∎

Este resultado podría considerarse, desde cierto punto de vista, como una versión más explícita del conocido Teorema de Rouche-Frobenius. Como caso particular, si las columnas de A son linealmente independientes y la ecuación $Ax=\mathtt{y}$ posee alguna solución, ésta es única y puede expresarse mediante $\mathtt{x}=G\mathtt{y}$, para cualquier $G\in A^-$. En particular, si A es una matriz cuadrada de orden m no singular e $\mathtt{y}\in\mathbb{R}^m$, la ecuación $Ax=\mathtt{y}$ tiene como única solución $\mathtt{x}=A^{-1}\mathtt{y}$.

El siguiente resultado será de utilizad a la hora de caracterizar funciones lineales estimables.

Corolario 7.5.

Un vector $k\in\mathbb{R}^p$ verifica que $k'\mathtt{x}$ es invariante para cualquier solución $\mathtt{x}$ de $Ax=\mathtt{y}$ si, y sólo si, $k\in\langle H\rangle$, siendo $H=GA$ para cualquier $G\in A^-$.

Demostración.
Basta tener en cuenta (7.3) junto co el hecho de que, por el lema 7.3, $\langle \mathtt{Id}-H\rangle^{\perp}=\langle H\rangle$.

∎

Ya sabemos que pueden existen varias matrices G verificando la propiedad (7.2). No obstante, si añadimos algunas hipótesis más, podemos garantizar la unicidad.

Teorema 7.6.

Dada $A\in\mathcal{M}_{m\times p}$, existe una única matriz $G\in\mathcal{M}_{p\times m}$ verificando

(i) $AGA=A$

(ii) $GAG=G$

(iii) $(GA)'=GA$

(iv) $(AG)'=AG$

Demostración.
Del teorema 9.5 se sigue que existen $B\in\mathbb{M}_{m\times r}$ y $C\in\mathbb{M}_{r\times p}$, ambas de rango r, tales que $A=BC$. En tal caso, tanto $B'B$ como CC' son invertibles y la matriz $G=C'(CC')^{-1}(B'B)^{-1}B'$ satisface trivialmente las condiciones requeridas. Veamos

que es la única. De (i) y (iii) se sigue que

$$AA'G = A. \tag{7.4}$$

Por un razonamiento completamente análogo se deduce que (ii)+(iv) implica

$$GG'A' = G'. \tag{7.5}$$

Puede probarse también, fácilmente, que (i)+(iv) y (ii)+(iii) implican, respectivamente

$$A'AG = A', \tag{7.6}$$

$$A'G'G = G. \tag{7.7}$$

Por lo tanto, si G_1, G_2 verifican las condiciones (i)-(iv), se sigue de (7.5) aplicado a G_1 y (7.6) aplicado a G_2 que $G_1 = G_1G_1'A' = G_1G_1'A'AG_2$. Aplicando nuevamente (7.5) a G_1, se deduce que $G_1 = G_1'AG_2$. Luego, por (7.7) aplicado a G_2, $G_1 = G_1AA'G_2'G_2$, que es igual, por (7.4) aplicado a G_1, a $A'G_2'G_2$. Aplicando nuevamente (7.7) a G_2, se deduce la unicidad.

∎

La matriz G verificando las condiciones del teorema se denomina inversa generalizada de Penrose, denotándose con frecuencia por $A^{(p)}$. Si verifica las condiciones (i) y (ii), se dice que es una inversa reflexiva generalizada. El conjunto formado por estas últimas se denota por $A^{(r)}$.

Conocemos, por (9.8), cómo se expresa la matriz de la proyección ortogonal sobre un subespacio a partir de una base de vectores $\mathtt{X}$ del mismo. Veamos cómo expresarla en el caso de que $\mathtt{X}$ sea un sistema generador de vectores, admitiendo la posibilidad de que sean linealmente dependientes.

Teorema 7.7.

Dada una matriz $\mathtt{X} \in \mathcal{M}_{m\times p}$, se verifica que $P_{\langle \mathtt{X}\rangle} = XGX'$, para cualquier $G \in (\mathtt{X}'\mathtt{X})^-$.

Demostración.

Primeramente, probaremos que $\mathtt{X}'\mathtt{X}D = 0$ implica $\mathtt{X}D = 0$. Efectivamente, basta considerar dos matrices $B\,C$ como en el teorema anterior tales que $\mathtt{X} = BC$. Entonces, $\mathtt{X}'\mathtt{X}D = 0$ implica $0 = C\mathtt{X}'\mathtt{X}D = (CC')(B'B)CD$. Al ser CC' y $B'B$ invertibles, se sigue que $CD = 0$ y, en particular, $\mathtt{X}D = BCD = 0$. Por otra parte, teniendo en cuenta que $G' \in (\mathtt{X}'\mathtt{X})^-$, se sigue de (7.2) que $\mathtt{X}'\mathtt{X}(G'\mathtt{X}'\mathtt{X} - \mathtt{Id}) = 0$. Luego, aplicando la primera parte, se deduce que $G'\mathtt{X}'\mathtt{X} = \mathtt{X}$ y , en particular, que $\mathtt{X}'\mathtt{X}G\mathtt{X}' = \mathtt{X}'$. Por lo tanto, dados $\mathtt{y} \in \mathbb{R}^m$ y $b \in \mathbb{R}^p$, se verifica que $\langle \mathtt{X}b, y - \mathtt{X}G\mathtt{X}'y\rangle = 0$, lo cual concluye la prueba.

∎

De este resultado, se sigue directamente que, para todo $G \in (\mathtt{X}'\mathtt{X})^-$, la matriz $\mathtt{X}G\mathtt{X}'$ es simétrica y su valor no depende del valor de G. Veamos más resultados relacionados con la inversa generaliza de $\mathtt{X}'\mathtt{X}$.

Lema 7.8.
$\mathtt{X}'\mathtt{X}B_1 = \mathtt{X}'\mathtt{X}B_2$ si, y sólo si, $\mathtt{X}B_1 = \mathtt{X}B_2$.

Demostración.
Denótese $Z = \mathtt{X}B_1 - \mathtt{X}B_2$. Si $\mathtt{X}'\mathtt{X}B_1 = \mathtt{X}'\mathtt{X}B_2$, se tiene, en particular, que es nula la matriz $(B_1' - B_2')(\mathtt{X}'\mathtt{X}B_1 - \mathtt{X}'\mathtt{X}B_2) = Z'Z$, en cuyo caso lo es también Z.

∎

Dada una matriz $\mathtt{X} \in \mathcal{M}_{m\times p}$ y un vector $\mathtt{y} \in \mathbb{R}^m$, la ecuación $\mathtt{X}b = \mathtt{y}$ es compatible, es decir, tiene solución exacta, si, y sólo si, $\mathtt{y} \in \langle \mathtt{X} \rangle$. En general, diremos que $\mathtt{b} \in \mathbb{R}^p$ es una solución mínimo-cuadrática [1] a la ecuación $\mathtt{X}b = \mathtt{y}$ cuando se verifica

$$\|\mathtt{X}\mathtt{b} - \mathtt{y}\| \leq \|\mathtt{X}b - \mathtt{y}\|, \quad \forall b \in \mathbb{R}^p. \tag{7.8}$$

Obviamente, $\mathtt{y} \in \langle \mathtt{X} \rangle$ si, y sólo si, las soluciones mínimo-cuadráticas coinciden con las exactas. El siguiente resultado es, posiblemente, el más importante de esta sección.

Teorema 7.9.
Dados $\mathtt{X} \in \mathbb{M}_{m\times p}$, $\mathtt{y} \in \mathbb{R}^m$, las soluciones mínimo-cuadráticas a la ecuación $\mathtt{X}b = \mathtt{y}$ coinciden con las soluciones exactas a la ecuación $\mathtt{X}b = P_{\langle \mathtt{X} \rangle}\mathtt{y}$, que coinciden a su vez con la soluciones exactas a la ecuación

$$\mathtt{X}'\mathtt{X}b = \mathtt{X}'\mathtt{y} \tag{7.9}$$

Además, dada cualquier $G \in (\mathtt{X}'\mathtt{X})^-$, el espacio de soluciones mínimo-cuadráticas es la subvariead afín $[p - \mathtt{rg}(\mathtt{X})]$-dimensional

$$G\mathtt{X}'\mathtt{y} + \langle G\mathtt{X}'\mathtt{X} - \mathtt{Id}_{p\times p} \rangle. \tag{7.10}$$

Demostración.
la primera parte de la tesis se sigue directamente del hecho de que

$$\|\mathtt{y} - P_{\langle \mathtt{X} \rangle}\mathtt{y}\| = \text{mín}\{\|\mathtt{y} - \mathtt{X}b\| : \ b \in \mathbb{R}^p\}.$$

En definitiva, se sigue del teorema 7.7 que las soluciones mínimo-cuadráticas a $\mathtt{X}b = \mathtt{y}$ coinciden con las soluciones exactas a la ecuación $\mathtt{X}b = \mathtt{X}G\mathtt{X}'\mathtt{y}$, para cualquier $G \in$

[1] El término *cuadrática* hace referencia a hecho de que la norma euclídea de un vector se define como la raíz cuadrada de la suma de los cuadrados de sus componentes.

$(\mathtt{X}'\mathtt{X})^-$. En virtud del lema anterior, dichas soluciones coinciden con las soluciones a la ecuación $\mathtt{X}'\mathtt{X}b = \mathtt{X}'\mathtt{X}G\mathtt{X}'\mathtt{y}$. El segundo término es igual a $\mathtt{X}'P_{\langle \mathtt{x} \rangle}\mathtt{y}$ que, teniendo en cuenta las propiedades fundamentales de la proyección ortogonal, coincide con $\mathtt{X}'\mathtt{y}$. El espacio de soluciones exactas a esta ecuación se obtiene haciendo uso del teorema 7.4. La dimensión de la subvariedad afín es $p - \mathtt{rg}(\mathtt{X}'\mathtt{X}) = p - \mathtt{rg}(\mathtt{X})$.

∎

Como consecuencia inmediata tenemos el siguiente resultado.

Corolario 7.10.
Si $\mathtt{X}$ es de rango completo, la única solución mínimo-cuadrática a la ecuación $\mathtt{X}b = \mathtt{y}$ es el vector $\mathtt{b} = (\mathtt{X}'\mathtt{X})^{-1}\mathtt{X}'\mathtt{y}$.

En definitiva, hemos probado que la búsqueda de soluciones mínimo cuadráticas al sistema de ecuaciones lineales $\mathtt{X}b = \mathtt{y}$ pasa por la resolución del sistema de ecuaciones (7.9), denominadas *normales* y, en consecuencia, según (7.10), por el cálculo de una inversa generalizada de la matriz $\mathtt{X}'\mathtt{X}$. Convendría pues disponer de una algoritmo para su obtención cuando el rango de $\mathtt{X}$ no sea completo. Proponemos aquí el siguiente.

Consideremos una matriz $R \in \mathcal{M}_{(p-\mathtt{rg}(\mathtt{x}))\times p}$ cuyas filas sean linealmente independientes entre sí y linealmente independientes de las filas de $\mathtt{X}'\mathtt{X}$. Lo mismo puede decirse entonces de las columnas de R' entre sí y en relación co las de $\mathtt{X}'\mathtt{X}$. Es importante tener en cuenta que si $R \in \mathcal{M}_{(p-\mathtt{rg}(\mathtt{x}))\times p}$ es una matriz cuyas filas sean linealmente independientes entre sí y linealmente independientes de las filas de $\mathtt{X}$, también son linealmente independientes de las de $\mathtt{X}'\mathtt{X}$. En ese caso, $Rb = 0$ puede entenderse como un conjunto de $p - \mathtt{rg}(\mathtt{X})$ restricciones a la ecuación $\mathtt{X}'\mathtt{X}b = \mathtt{X}'\mathtt{y}$.

En esas condiciones, la matriz

$$S = \begin{pmatrix} \mathtt{X}'\mathtt{X} & R' \\ R & 0 \end{pmatrix}$$

es cuadrada de orden $2p - \mathtt{rg}(\mathtt{X})$ e invertible. Denótese

$$S^{-1} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}.$$

Debe verificarse entonces las siguientes ecuaciones

$$\mathtt{X}'\mathtt{X}B_{11} + R'B_{21} = \mathtt{Id}, \tag{7.11}$$

$$\mathtt{X}'\mathtt{X}B_{12} + R'B_{22} = 0, \tag{7.12}$$

$$RB_{11} = 0. \tag{7.13}$$

Al ser las columnas de R' linealmente independientes de las de $\mathtt{X}'\mathtt{X}$, se verifica, por (7.12), que $B_{22} = 0$, luego, $B_{21}\mathtt{X}'\mathtt{X} = 0$. Si en (7.11) multiplicamos a la derecha por $\mathtt{X}'\mathtt{X}$ y aplicamos lo anterior, se tiene que $\mathtt{X}'\mathtt{X}B_{11}\mathtt{X}'\mathtt{X} = \mathtt{X}'\mathtt{X}$. Por lo tanto, $B_{11} \in (\mathtt{X}'\mathtt{X})^-$. Podemos probar también (cuestión propuesta), que B_{11} verifica también la condición (ii) del teorema 7.6, por lo que se trata realmente de una inversa generalizada reflexiva. En definitiva, la búsqueda de una solución particular pasa por encontrar una matriz de restricciones lineales R en las condiciones anteriores.

7.3. Estimación y Contraste de Hipótesis.

Consideremos el modelo $Y \sim \mathtt{X}\beta + \mathcal{E}$, donde $\mathtt{X} \in \mathcal{M}_{\mathtt{n}\times s}$, $\mathcal{E}$ es un $\mathtt{n}$-vector aleatorio de media 0 y matriz de varianzas-covarianzas $\sigma^2\mathtt{Id}$, siendo σ^2 un número positivo y β cualquier vector de $\mathbb{R}^s$. Afrontaremos en esta sección los problemas de Estimación Puntual y Contraste de Hipótesis desde un punto de vista general. En la sección siguiente estudiaremos un caso particular. También se consideraran otros ejemplos en las cuestiones propuestas.

Dado que el modelo considerado el el Modelo Lineal estudiado en el capítulo 3 con $V = \langle \mathtt{X} \rangle$, nada podemos añadir respecto la estimación de σ^2. Por lo tanto, se sigue de la proposición 3.2 junto con el teorema 7.7 que el siguiente estadístico es un estimador insesgado de σ^2

$$\hat{\sigma}^{2,\mathrm{I}} = [\mathtt{n} - \mathtt{rg}(\mathtt{X})]^{-1}\|Y - \mathtt{X}G\mathtt{X}'Y\|^2,$$

siendo G cualquier inversa generalizada de $\mathtt{X}'\mathtt{X}$. Es más, si $\mathcal{E}$ sigue una distribución normal, se sigue de los teoremas 3.6 y 3.7 que el estimador anterior es insesgado de mínima varianza y máxima verosilimitud. Además,

$$[\mathtt{n} - \mathtt{rg}(\mathtt{X})]\hat{\sigma}^{2,\mathrm{I}} \sim \sigma^2\chi^2_{\mathtt{n}-\mathtt{rg}(\mathtt{x})}.$$

Por último, todo lo expuesto en las secciones 3.3 y 3.4 en referencia al contraste de hipótesis y comportamiento asintótico para σ^2 sigue siendo válido.

Los problemas de Estimación y Contraste de Hipótesis para β plantean, sin embargo, una nueva dificultad: el hecho de que, en general, β no está perfectamente determinado sino que es cualquier solución a la ecuación $\mathtt{X}\beta = \mathtt{E}[Y]$. Consideraremos pues la estimación de funciones paramétricas de β estimables[2] y el contraste de hipótesis de beta contrastables[3]. Además, restringiremos el estudio de estimación a

[2] Ver (9.39).
[3] Ver (9.44).

funciones reales lineales, es decir, de la forma $a'\beta$, donde $a \in \mathbb{R}^s$. También se considerará únicamente el contraste de hipótesis del tipo $A\beta = 0$ [4], donde $A \in \mathcal{M}_{m\times s}$. Podemos asumir, sin pérdida de generalidad, que $\mathtt{rg}(A) = m$. A continuación, procederemos a caracterizar con precisión las funciones lineales estimables y las hipótesis lineales contrastables.

Lema 7.11.

Una función lineal $C\beta$, donde $C \in \mathcal{M}_{c\times s}$, es estimable si, y sólo si, existe $B \in \mathcal{M}_{c\times \mathtt{n}}$ tal que $C = B\mathtt{X}$.

Demostración.

Una implicación es trivial, pues si $C = B\mathtt{X}$, entonces $C\beta = B\mathtt{X}\beta = B\mathtt{E}[Y]$, en cuyo caso se verifica (9.39). Para probar el recíproco, consideremos $G \in (\mathtt{X}'\mathtt{X})^-$ y supongamos que $\mathtt{X}\beta_1 = \mathtt{X}\beta_2$, lo cual equivale, teniendo en cuenta (7.3), a que $\beta_1 - \beta_2$ pertenezca a $\langle C\mathtt{X}'\mathtt{X} - \mathtt{Id}\rangle$. Por lo tanto, si $A\beta_1 = A\beta_2$, las filas de A perteneceran al subespacio ortogonal al anterior, que, en virtud del lema 7.3, es $\langle G\mathtt{X}'\mathtt{X}\rangle$. Por lo tanto, existirá $D \in \mathcal{M}_{c\times s}$ tal que $A = DG\mathtt{X}'\mathtt{X}$, y la tesis se verifica tomando $B = DG\mathtt{X}'$. ∎

Teorema 7.12.

Dados $a \in \mathbb{R}^s$ y $A \in \mathcal{M}_{m\times s}$, se verifica lo siguiente:
(i) La función paramétrica $a'\beta$ es estimable si, y sólo si, existe $b \in \mathcal{M}_{1\times \mathtt{n}}$ tal que $a' = b\mathtt{X}$.
(ii) La hipótesis paramétrica $H_0 : A\beta = 0$ es contrastable si, y sólo si, existe $B \in \mathcal{M}_{m\times \mathtt{n}}$ tal que $A = B\mathtt{X}$.

Demostración.

El primer apartado es consecuencia directa del lema anterior. Respecto al segundo, supongamos que la hipótesis inicial $H_0 : A\beta = 0$ es contrastable y consideremos $\beta_1, \beta_2 \in \mathbb{R}^s$ tales que $\mathtt{X}\beta_1 = \mathtt{X}\beta_2$. En es ese caso, $\mathtt{X}(\beta_1 - \beta_2) = \mathtt{X}0$. Dado que $A0 = 0$, se verifica por hipótesis que $A(\beta_1 - \beta_2) = 0$. Por lo tanto, la función $A\beta$ es estimable y, aplicando el lema anterior, se concluye. ∎

Corolario 7.13.

Se verifica lo siguiente:
(i) Una función lineal real $a'\beta$ es estimable si, y sólo si, existe un estadístico lineal real T

[4] Los contrastes del tipo $A\beta = c$ puede resolverse a partir de éstos mediante una traslación del vector de observaciones.

tal que $\mathtt{E}_{(\beta,\sigma^2)}[T] = a'\beta$.
(ii) La función $a'\beta$ es estimable si, y sólo si, a' es una combinación lineal de las filas de $\mathtt{X}$. Cualquier combinación lineal de funciones lineales reales estimables de β es, a su vez, una función lineal real estimable de β.
(iii) El número máximo de funciones lineales reales estimables de β linealmente independientes es igual al rango de la matriz $\mathtt{X}$.

Demostración.
Para probar (i) supongamos que existe $c \in \mathbb{R}^{\mathtt{n}}$ tal que $\mathtt{E}_{\beta,\sigma^2)}[c'Y] = a'\beta$. En ese caso, la función $a'\beta$ verifica (9.38), es decir, es estimable. Recíprocamente, supongamos que $a = b\mathtt{X}$ para alguna matriz $b \in \mathcal{M}_{1\times\mathtt{n}}$. Dado $P_{\langle\mathtt{X}\rangle}Y$, que es un estimador insesgado de $\mathtt{E}[Y]$, se sigue que $bP_{\langle\mathtt{X}\rangle}Y$ es un estimador lineal insesgado de $a'\beta$.
La propiedad (ii) se sigue directamente del teorema anterior. Para probar (iii) basta tener en cuenta que las funciones lineales reales estimables de β se identifican, según el teorema anterior, con los vectores de $\mathbb{R}^s$ de la forma $\mathtt{X}'b$, donde $b \in \mathbb{R}^n$. El número de vectores de esta forma linealmente independiente es igual, obviamente, al rango de $\mathtt{X}$.

∎

Hemos de advertir que la proposición (i) del corolario anterior suele presentarse en la mayor parte de la literatura estadística como definición de función lineal real estimable de β.

El teorema 3.3, conocido como de Gauss-Markov, resuelve el problema de estimación de estimandos del tipo $c'\mathtt{E}[Y]$, donde $c \in \mathbb{R}^{\mathtt{n}}$, pues garantiza que el estadístico $c'P_{\langle\mathtt{X}\rangle}Y$ es el estimador lineal insesgado de mínima varianza. En consecuencia, si $a'\beta$ es estimable, existe $b \in \mathcal{M}_{1\times\mathtt{n}}$ tal que $a' = b\mathtt{X}$, luego, $a'\beta = b\mathtt{E}[Y]$ y el estadístico $T(Y) = bP_{\langle\mathtt{X}\rangle}Y$ será el estimador lineal insesgado de mínima varianza de $a'\beta$, cuya varianza es, precisamente, $\sigma^2 b'P_{\langle\mathtt{X}\rangle}b$. Si se verifica la $\mathtt{n}$-normalidad de $\mathcal{E}$, estaremos hablando del estimador insesgado de mínima varianza y el de máxima verosimilitud de $a'\beta$, cuya distribución será

$$T \sim N\big(a'\beta, \sigma^2 b'P_{\langle\mathtt{X}\rangle}b\big).$$

El parámetro β no puede considerarse estimable pues, en principio, no existe una única solución a la ecuación $\mathtt{X}\beta = \mathtt{E}[Y]$. No obstante, aunque no podemos hablar propiamente de estimadores insesgados de β, sí podemos buscar un estadístico $T : \mathbb{R}^{\mathtt{n}} \to \mathbb{R}^s$ tal que $\mathtt{X}T$ sea un estimador insesgado de $\mathtt{E}[Y]$, es decir, que satisfagan la ecuación $\mathtt{X}\mathtt{E}[T] = \mathtt{E}[Y]$. Teniendo en cuenta que $P_{\langle\mathtt{X}\rangle}Y$ es un *buen* estimador insesgado de $\mathtt{E}[Y]$, procederemos a buscar soluciones particulares a la ecuación

$$\mathtt{X}T = P_{\langle\mathtt{X}\rangle}Y. \tag{7.14}$$

Se denotará por $\hat{\beta}$ a cualquier estadístico que sea solución exacta al sistema de ecuaciones lineales anterior, lo cual equivale, según el teorema 7.9, a ser solución exacta al sistema de ecuaciones normales

$$\mathtt{X}'\mathtt{X}T = \mathtt{X}'Y \tag{7.15}$$

o, lo que es lo mismo, solución mínimo-cuadrática a la ecuación

$$\mathtt{X}T = Y.$$

En virtud del teorema 7.9, sabemos que, dada $G \in (\mathtt{X}'\mathtt{X})^-$, las soluciones a la ecuación (7.15) constituyen la siguiente subvariedad lineal de dimensión $s - \mathtt{rg}(\mathtt{X})$

$$G\mathtt{X}'Y + \langle G\mathtt{X}'\mathtt{X} - \mathtt{Id}_{s\times s}\rangle \tag{7.16}$$

Teniendo en cuenta el teorema 7.4-(ii) [5], existen $s - \mathtt{rg}(\mathtt{X}) + 1$ soluciones linealmente independientes. Dos soluciones cualesquiera difieren en un vector del subespacio $\langle G\mathtt{X}'\mathtt{X} - \mathtt{Id}_{s\times s}\rangle$. Por lo tanto, determinar una solución particular equivale a imponer $s - \mathtt{rg}(\mathtt{X})$ restricciones linealmenente independientes[6]. Si $\mathtt{X}$ es de rango completo, existe una única solución a (7.15), que coincide con el estimador (3.15) obtenido en el capítulo 3. El siguiente resultado es fundamental en lo que respecta al problema de estimación.

Teorema 7.14.
Si la función $a'\beta$ es estimable y $\hat{\beta}$ es una solución cualquiera a (7.15), $a'\hat{\beta}$ es el estimador lineal insesgado de mínima varianza de $a'\beta$. Si, además, $\mathcal{E}$ sigue un modelo de distribución $\mathtt{n}$-normal, será el estimador insesgado de mínima varianza y máxima verosimilitud.

Demostración.
Efectivamente, dado $b \in \mathbb{R}^{\mathtt{n}}$ tal que $a' = b\mathtt{X}$, se sigue de (7.14) que

$$a'\hat{\beta} = b\mathtt{X}\hat{\beta} = b'P_{\langle \mathtt{X}\rangle}Y,$$

luego, el teorema de Gauss-Markov prueba la primera afirmación. La segunda parte se sigue de (9.42), teniendo en cuenta la definición (9.4). ∎

[5] Si la distribución de Y está dominada por la medida de Lebesgue en $\mathbb{R}^{\mathtt{n}}$, la probabilidad de que Y pertenezca a $\langle \mathtt{X}\rangle^{\perp}$ es nula.

[6] Dado un vector $x \in \mathbb{R}^n$, entendemos por restricción a una hipótesis del tipo $y'x = 0$, donde $y \in \mathbb{R}^n$.

Respecto al contraste de hipótesis lineales contrastables, el problema también está resuelto en el capítulo 3. Efectivamente, una hipótesis inicial de la forma $H_0 : A\beta = 0$, siendo A una matriz $m \times s$ de rango m y tal que $A = B\mathbf{X}$ para cierta matriz $B \in \mathcal{M}_{m\times \mathtt{n}}$, puede expresarse mediante $H_0 : B\mathbf{E}[Y] = 0$. Así pues, el problema se reduce a contratar una hipótesis del tipo $H_0 : E[Y] \in W_{\mathbf{X},A}$, siendo $W_{\mathbf{X},A}$ cierto subespacio lineal de $\langle \mathbf{X}\rangle$. Como ya hemos afirmado, este problema se resuelve en el capítulo 3 mediante el test F. Lo único que podemos añadir es una expresión explícita del mismo a partir de las matrices $\mathbf{X}$ y A consideradas. En ese sentido, el siguiente resultado es una generalización del teorema 3.13.

Teorema 7.15.
Dados $G \in (\mathbf{X}'\mathbf{X})^-$ y A una matriz $m \times s$ de rango m tal que la hipótesis inicial $H_0 : A\beta = 0$ es contrastable, el test F a nivel α para constrastar H_0 consiste decidir H_1 cuando $F^{\alpha}_{m,\mathtt{n}-\mathtt{rg}(\mathbf{X})}$ es menor que el estadístico

$$F = m^{-1}\frac{(A\hat{\beta})'(AGA')^{-1}A\hat{\beta}}{\hat{\sigma}^{2,\mathrm{I}}}, \tag{7.17}$$

siendo $\hat{\beta}$ cualquier solución a la ecuación (7.15).

Demostración.
Si A es contrastable, existe $B \in \mathcal{M}_{m\times \mathtt{n}}$ tal que $A = B\mathbf{X} = (P_{\langle \mathbf{X}\rangle}B')'\mathbf{X}$. En ese caso, $P_{\langle \mathbf{X}\rangle}B'$ es una matriz $n \times m$ de rango m. Por lo tanto, sus columnas consituyen un conjunto de vectores linealmente independientes de $\langle \mathbf{X}\rangle$. Por otra parte, si $\mu = \mathbf{X}\beta$, se verifica $A\beta = 0$ si, y sólo si, $B\mu = 0$, lo cual equivale a $(P_{\langle \mathbf{X}\rangle}B')'\mu = 0$. En consecuencia, el conjunto de vectores anterior constituye una base de $\langle \mathbf{X}\rangle | W_{\mathbf{X},A}$, cuya dimensión es, por lo tanto, m. Teniendo en cuenta (9.8) junto con el teorema 7.7, se verifica

$$\begin{aligned} Y'P_{\langle \mathbf{X}\rangle | W_{\mathbf{X},A}}Y &= Y'P_{\langle \mathbf{X}\rangle}B'(BP_{\langle \mathbf{X}\rangle}B')^{-1}BP_{\langle \mathbf{X}\rangle}Y \\ &= Y'\mathbf{X}G\mathbf{X}'B'(B\mathbf{X}G\mathbf{X}'B')^{-1}B\mathbf{X}G\mathbf{X}'Y \\ &= (A\hat{\beta})'(AGA')^{-1}A\hat{\beta}, \end{aligned}$$

donde $\hat{\beta} = G\mathbf{X}'Y$. En ese caso, el estadístico de contraste $\mathtt{F} = \left(m\hat{\sigma}^{2,\mathrm{I}}\right)^{-1}\|P_{\langle \mathbf{X}\rangle | W_{\mathbf{X},A}}Y\|^2$ sigue trivialmente la expresión deseada. Para acabar, tener en cuenta que, para cualquier solución particular a (7.15), se verifica

$$A\hat{\beta} = B\mathbf{X}\hat{\beta} = BP_{\langle \mathbf{X}\rangle}Y.$$

Por lo tanto, la expresión anterior no depende de la solución $\hat{\beta}$ considerada. ∎

Obviamente, si $\mathbf{X}$ es de rango completo se obtiene la expresión (3.26). Lo más importante a nuestro entender es que todas las expresiones obtenidas en esta sección pueden implementarse dando lugar a algoritmos automáticos, cosa que no sucede si utilizamos el concepto abstracto de subespacio lineal.

7.4. Ejemplo: diseño bifactorial no equilibrado.

Como ya adelantamos en la última sección del capítulo anterior y en la introducción de éste, el Modelo Lineal parametrizado mediante una matriz de rango no completo puede se de utilidad cuando se aborda el modelo de regresión lineal donde los vectores explicativas son linealmente dependientes, o el número de estos no es superior al número de unidades experimentales; pero sobre todo puede resultar útil en el estudio de diseños no equilibrados en el análisis de la varianza. Dedicaremos esta sección a justificar dicha afirmación mediante la exposición de un diseño, similar al considerado en la sección 6.4.

Se estudia la influencia de dos factores cualitativos, f_A con $\mathtt{a}$ niveles y f_B con $\mathtt{b}$ niveles, en la media de una variable respuesta y. Para ello, consideraremos $\mathtt{ab}$ muestras aleatorias simples, cada una de ellas correspondiendo a la combinación entre un determinado nivel del factor f_A, i, con otro del factor f_B, j. Se denotará por $\mathtt{n}_{ij}$ el tamaño de la muestra (i,j)-ésima. No estamos suponiendo, por lo tanto, que el diseño sea equilibrado. El número total de datos es $\mathtt{n} = \sum_{i=1}^{\mathtt{a}} \sum_{j=1}^{\mathtt{b}} \mathtt{n}_{ij}$. El diseño puede representarse, esquemáticamente, como sigue:

Factor B

Factor A			
	$Y_{\mathtt{1}11}, \ldots, Y_{\mathtt{1}1\mathtt{n}_{11}}$	.....	$Y_{\mathtt{1b}1}, \ldots, Y_{\mathtt{1bn}_{\mathtt{1b}}}$
	$\vdots$		$\vdots$
	$Y_{\mathtt{a}11}, \ldots, Y_{\mathtt{a}1\mathtt{n}_{\mathtt{a}1}}$	.....	$Y_{\mathtt{ab}1}, \ldots, Y_{\mathtt{abn}_{\mathtt{ab}}}$

Hemos asignado la muestra correspondiente a los niveles i-ésimo y j-ésimo de los factores A y B, respectivamente, las coordenadas (i,j), que indica una celda de la cuadrícula. Una tercera coordenada, k, indicará la posición del dato en la celda correspondiente. Se supondrá, además, que todas las muestras son independientes y provienen de distribuciones normales con idéntica varianza. El modelo puede expresarse así:

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma^2) \text{ independientes}, \tag{7.18}$$

donde $i = 1, \ldots, \mathtt{a}$, $j = 1, \ldots, \mathtt{b}$ y $k = 1, \ldots, \mathtt{n}_{ij}$. Si componemos todas las observaciones de las variable repuesta, ordenando las muestras por filas, obtenemos el vector

aleatorio $\mathtt{n}$-dimensional $Y = (Y_{111}, \ldots, Y_{\mathtt{abn}_{\mathtt{ab}}})'$, de media μ. Para cada celda (i, j) de la cuadrícula se considera el vector $\mathtt{v}_{ij}$ de $\mathbb{R}^{\mathtt{n}}$ definido de manera análoga al capítulo anterior. Así, si V denota el subespacio $\mathtt{ab}$ dimensional del $\mathbb{R}^{\mathtt{n}}$ generado por los vectores $\mathtt{v}_{ij}$, $i = 1, \ldots, \mathtt{a}$, $j = 1, \ldots, \mathtt{b}$, el modelo puede expresarse mediante

$$Y = \mu + \mathcal{E}, \quad \mathcal{E} \sim N_{\mathtt{n}}(0, \sigma^2 \mathtt{Id}), \qquad \mu \in V,\ \sigma^2 > 0. \tag{7.19}$$

Se trata pues de un modelo lineal normal. Una descomposición del tipo (6.23), que lleva asociada la parametrización considerada en la sección 6.4 con la imposición de las restricciones (6.21), es posible en general, aunque no se puede garantizar la ortogonalidad entre los subespacios considerados (cuestión propuesta). Por ello, la familia de restricciones (6.21) no debe ser considerada necesariamente natural. No obstante, dado que el objetivo principal cuando se lleva a cabo un diseño de este tipo es determinar en qué medida influyen cada uno de los factores cualitativos y la interacción entre ambos en la media de la variable respuesta, sí resulta natural desde un punto de vista intuitivo proponer la siguiente descomposición para la media de la casillas (i, j)-ésima:

$$\mu_{ij} = \theta + \alpha_i + \alpha_j + (\alpha\beta)_{ij} \tag{7.20}$$

De esta forma, el parámetro θ se interpreta, siempre en términos intuitivos, como la aportación común a todos los niveles de los factores, el parámetro α_i como la aportación específica del nivel i-ésimo del primer factor, β_j como la aportación específica del nivel j-ésimo del segundo factor; por último, $(\alpha\beta)_{ij}$ se interpreta como la aportación a la media que resulta de combinar los niveles i-ésimo y j-ésimo del primer y segundo factor, respectivamente.

Para poder ser más explícitos, supongamos que $\mathtt{a} = 3$ y $\mathtt{b} = 2$. Denótese por $\mathtt{B}$ a al parámetro en $\mathbb{R}^{12}$ de componentes

$$\mathtt{B} = \big(\theta, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, (\alpha\beta)_{11}, \ldots, (\alpha\beta)_{32}\big)'$$

siendo solución a la ecuación (7.20). Equivalentemente, se verifica que $\mathtt{B}$ es solución a la ecuación

$$\mathtt{X}b = \mathtt{E}[Y],$$

siendo $\mathtt{X}$ la matriz en $\mathcal{M}_{\mathtt{n}\times 12}$ definida mediante

$$\mathtt{X} = \left(\begin{array}{c|ccc|cc|cccccc}
1_{\mathtt{n}_{11}} & 1_{\mathtt{n}_{11}} & 0 & 0 & 1_{\mathtt{n}_{11}} & 0 & 1_{\mathtt{n}_{11}} & 0 & 0 & 0 & 0 & 0 \\
1_{\mathtt{n}_{12}} & 1_{\mathtt{n}_{12}} & 0 & 0 & 0 & 1_{\mathtt{n}_{12}} & 0 & 1_{\mathtt{n}_{12}} & 0 & 0 & 0 & 0 \\
\hline
1_{\mathtt{n}_{21}} & 0 & 1_{\mathtt{n}_{21}} & 0 & 1_{\mathtt{n}_{21}} & 0 & 0 & 0 & 1_{\mathtt{n}_{21}} & 0 & 0 & 0 \\
1_{\mathtt{n}_{22}} & 0 & 1_{\mathtt{n}_{22}} & 0 & 0 & 1_{\mathtt{n}_{22}} & 0 & 0 & 0 & 1_{\mathtt{n}_{22}} & 0 & 0 \\
\hline
1_{\mathtt{n}_{31}} & 0 & 0 & 1_{\mathtt{n}_{31}} & 1_{\mathtt{n}_{31}} & 0 & 0 & 0 & 0 & 0 & 1_{\mathtt{n}_{31}} & 0 \\
1_{\mathtt{n}_{32}} & 0 & 0 & 1_{\mathtt{n}_{32}} & 0 & 1_{\mathtt{n}_{32}} & 0 & 0 & 0 & 0 & 0 & 1_{\mathtt{n}_{32}}
\end{array}\right)$$

Por lo tanto, el modelo puede expresarse de la forma

$$Y = \mathtt{XB} + \mathcal{E}, \qquad \mathcal{E} \sim N_{\mathtt{n}}(0, \sigma^2 \mathtt{Id}), \qquad \mathtt{B} \in \mathbb{R}^{12},\ \sigma^2 > 0.$$

De esta forma, descomponer la media de cada celda según (7.20) equivale a parametrizar el modelo a través de la matriz $\mathtt{X} \in \mathcal{M}_{\mathtt{n}\times 12}$ de rango 6. Del teorema 7.12, se sigue que las funciones lineales estimables de $\mathtt{B}$ son de la forma $a'\mathtt{B}$ para cualquier $a' \in \mathcal{M}_{1\times 12}$ que pueda expresarse como combinación lineal de las filas de $\mathtt{X}$. En ese caso, se verifica trivialmente (cuestión propuesta) que ninguno de los parámetros θ, α_i, β_j, $(\alpha\beta)_{ij}$, donde $i = 1, 2, 3$ y $j = 1, 2$, son estimables. Si son estimables, sin embargo, funciones del tipo

$$\beta_1 + (\alpha\beta)_{11} - \beta_2 - (\alpha\beta)_{12}, \tag{7.21}$$

que equivale a $\mu_{11} = \mu_{12}$, es decir, a que el factor B no afecta a la media del primer nivel del factor A. Igualmente, son estimables funciones del tipo

$$\alpha_1 + (\alpha\beta)_{11} - \alpha_2 - (\alpha\beta)_{21}, \tag{7.22}$$

que equivalen a $\mu_{11} = \mu_{21}$. En virtud del teorema (7.12), son contrastables las hipótesis del tipo $A\mathtt{B} = 0$, cuando las filas de A sean combinaciones lineales de las de $\mathtt{X}$. De esta forma, son contrastables la hipótesis

$$H_0^{A,AB} : \alpha_i + (\alpha\beta)_{ij} = \alpha_{i'} + (\alpha\beta)_{i'j}, \quad i \neq i',\ j = 1, 2,$$

$$H_0^{B,AB} : \beta_1 + (\alpha\beta)_{i1} = \beta_2 + (\alpha\beta)_{i2}, \quad i = 1, 2, 3.$$

Ambas pueden expresarse, en términos de la media, como sigue:

$$H_0^{A,AB} : \mu_{ij} = \mu_{i'j}, \quad i \neq i'\ j = 1, 2,$$

$$H_0^{B,AB} : \mu_{i1} = \mu_{i2}, \quad i = 1, 2, 3.$$

Por lo tanto, se traducen en la no influencia de los factores A y B, respectivamente, en la media de la variable respuesta, y esas son, precisamente, los contrastes que más interesan. Tanto los problemas de estimación como los de contraste de hipótesis podrían resolverse directamente con las técnicas estudiadas en el capítulo 3 [7], pues cabe formularlos en términos de la media μ. No obstante, la teoría de rango no completo permite generar de manera sencilla funciones lineales estimables e hipótesis

[7] Aunque no podamos ofrecer una expresión explícita para los tests como las que podemos encontrar en capítulo anterior, pues se basan en el cálculo de proyección ortogonal sobre el subespacio $V|W$ correspondiente.

contrastables expresadas a partir de unos parámetros que nos resultan intuitivos (θ, α_1, β_2, etc) y aporta automáticamente un algoritmo para la resolución del problema. Concretamente, para estimar funciones como las del tipo (7.21) y (7.22) haremos uso del teorema 7.14. Asimismo, para contrastar hipótesis como $H_0^{A,AB}$ y $H_0^{B,AB}$ utilizaremos el teorema 7.15. En ambos casos, precisamos de una solución mínimo-cuadrática a la ecuación $\mathtt{X}b = Y$ o, equivalentemente, una solución[8] exacta al sistema de ecuaciones normales

$$\mathtt{X}'\mathtt{X}b = \mathtt{X}'Y$$

que se denotará por $\hat{\mathtt{B}}$. En este caso, según se vio en la segunda sección, elegir una solución particular pasa por imponer un sistema de 6 restricciones linealmente independientes. Puede probarse que las restricciones (6.21), consideradas naturales en el caso equilibrado, son válidas también en un diseño no equilibrado, es decir, que su cumplimiento conduce a una única solución. Efectivamente, en nuestro caso se verifica trivialmente que las filas de la matriz R siguiente son linealmente independientes entre sí y linealmente independientes de las de $\mathtt{X}$

$$R = \left(\begin{array}{c|cc|cc|cccccc} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{array}\right)$$

Por lo tanto, podemos considerar la única solución

$$\hat{\mathtt{B}} = \left(\hat{\theta}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\beta}_1, \hat{\beta}_2, \widehat{\alpha\beta})_{11}, \ldots, (\widehat{\alpha\beta})_{32}\right)$$

al sistema de ecuaciones

$$\left(\begin{array}{c} \mathtt{X}'\mathtt{X} \\ R \end{array}\right) b = \left(\begin{array}{c} \mathtt{X}'Y \\ 0 \end{array}\right),$$

la cual verifica las restricciones

$$\sum_{i=1}^{3} \hat{\alpha}_i = 0, \quad \sum_{j=1}^{2} \hat{\beta}_j = 0, \quad \sum_{i=1}^{3} (\widehat{\alpha\beta})_{ij} = 0,\ j = 1,2, \quad \sum_{j=1}^{2} (\widehat{\alpha\beta})_{ij} = 0,\ i = 1,2,3.$$

Las restricciones anteriores suelen imponerse de manera habitual, lo cual no quiere decir que sean naturales.

[8] Tener en cuenta que el espacio de soluciones es una subvariedad afín de dimensión 12-6=6 de $\mathbb{R}^{12}$.

A fin de cuentas y desde el punto de vista técnico, la diferencia entre abordar el problema con rango no completo en vez de completo (capítulo 3) estriba en calcular una solución a un sistema de ecuaciones no determinado en lugar de una matriz de proyección ortogonal sobre cierto subespacio a determinar. Se trata, en fin, de una distinción que bien podría obviarse, teniendo en cuenta que, hoy en día, los problemas estadísticos se resuelven en su totalidad mediante programas informáticos. No obstante, desde el punto de vista técnico, el computador debe entender a qué subespacio nos estamos refiriendo, para lo cual habremos de introducir cierta matriz $\mathtt{X}$, que en el modelo de rango no completo viene dada de partida. Esta ventaja la disfruta sólo el programador. Para el *usuario* del programa informático, el uso de rango no completo supone la posibilidad de manejar los parámetros del modelo sin necesidad de imponer previamente restricciones sobre los mismos que pueden resultar artificiales. No obstante, las restricciones, igualmente artificiales, deberán considerarse a la hora de seleccionar una solución particular a las ecuaciones normales.

Cuestiones propuestas

1. Demostrar que la matriz B_{11} obtenida como inversa generaliza da $\mathtt{X}'\mathtt{X}$, es reflexiva, es decir, verifica que $B_{11}\mathtt{X}'\mathtt{X}B_{11} = B_{11}$. (Indicación: considerar (7.11) y (7.13).

2. Dada una función lineal estimable $a'\beta$, construir un intervalo de confianza a nivel $1-\alpha$ para $a'\beta$, suponiendo la n-normalidad del vector aleatorio Y.

3. Desarrollar un diseño completamente aleatorizado (sección 6.1) mediante la parametrización $\mu_i = \theta + \alpha_i$, $i = 1, \ldots, r$, sin imponer ninguna restricción *a priori* sobre θ y $\alpha_1, \ldots, \alpha_r$.

4. Probar la validez de la descomposición (6.23) en el diseño bifactorial no equilibrado, aunque no se verifique la ortogonalidad entre todos los subespacios que la componen.

5. Si $\mu_{i\cdot}$ denota la media del nivel i-esimo del factor A, $i = 1, 2, 3$, construir la familia de intervalos de confianza simultáneos a nivel $1-\alpha$ de Bonferroni para las diferencias $\mu_{i\cdot} - \mu_{i'\cdot}$.

6. Establecer un algoritmo para estimar la función

$$f(\mathtt{B}) = 2\theta + \alpha_1 + \alpha_2 + 2\beta_1 + (\alpha\beta)_{11} + (\alpha\beta)_{12}$$

y para contrastar a nivel α la hipótesis inicial $H_0 : f(\mathtt{B}) = 0$.

7. Establecer algoritmo para resolver el contraste de la hipótesis inicial $H_0^{A,AB}$ que no precise del cálculo de inversa generalizada.

8. ¿Es contrastable la hipótesis $(\alpha\beta)_{11} = \ldots = (\alpha\beta)_{32}$ en diseño bifactorial 3×2?

9. ¿Cómo contrastar la existencia o no de interacción en un modelo bifactorial no equilibrado?

Capítulo 8

Modelos Lineales Generalizados

Este capítulo está dedicado a una familia de modelos que, salvo cierto detalle que comentaremos más adelante, vienen a generalizar el modelo lineal normal. El caso es que estos modelos aportan un procedimiento para resolver los problemas de regresión lineal y análisis de la varianza y covarianza, junto con otros nuevos, entre los que se encuentran la regresión de Poisson, la regresión logística o las tablas de contingencia.

Hemos de destacar que las pruebas de los resultados reposan fundamentalmente en el Cálculo Diferencial y la Teoría Asintótica, precisándose también cierto conocimiento de las familias exponenciales y la Teoría de la Información. Algunas de ellas se proponen como ejercicio para el lector. En la última sección se precisa un cierto dominio de los métodos numéricos para la aproximación a las raíces de una ecuación, como el de Newton-Raphson; también encontraremos en ella algunos razonamientos de tipo heurístico implícitamente presentes en la definición de *devianza*. En todo caso remitimos al lector interesado a la bibliografía recomendada para un estudio más completo del tema. Concretamente, en Dobson (1990) podemos encontrar una buena síntesis y aporta referencias más concretas, mientras que en Cox & Hinkley (1974) podemos consultar mejor ciertos detalles técnicos.

8.1. El modelo

La definición original de modelo lineal generalizado se debe a Nelder & Wedderburn (1972). Sea Y un vector aleatorio $\mathtt{n}$-dimensional de componentes $Y_1, \ldots, Y_{\mathtt{n}}$ independientes de medias $\mu_1, \ldots, \mu_{\mathtt{n}}$, respectivamente. Decimos que Y sigue un modelo lineal generalizado dada la matriz $\mathtt{X} \in \mathcal{M}_{\mathtt{n}\times s}$ de filas $\mathtt{X}_1', \ldots, \mathtt{X}_{\mathtt{n}}'$, cuando existe una función g monótona diferenciable tal que, para todo $i = 1, \ldots, \mathtt{n}$, se verifica:

(i) La distribución de Y_i es del tipo (9.36) con $\theta_i = \mu_i$, siendo $T = \mathtt{Id}$, $Q = b \circ g$

para alguna función real b, y c y d son las mismas para todo i.

(ii) Existe $\beta \in \mathbb{R}^s$ tal que $g(\mathtt{E}[Y_i]) = \mathtt{X}_i'\beta$.

En definitiva, se trata de un modelo dominado cuya función de verosimilitud puede expresarse, si se denota $\mathtt{Y} = (\mathtt{Y}_1, \ldots, \mathtt{Y}_\mathtt{n})'$, de la forma

$$\mathcal{L}(\beta, \mathtt{Y}) = \exp\left\{\langle B(\mathtt{X}'\beta), \mathtt{Y}\rangle + C(\beta) + D(\mathtt{Y})\right\} \tag{8.1}$$

siendo B la composición de $\mathtt{n}$ réplicas de b y C y D la suma de $\mathtt{n}$ réplicas de c y d, respectivamente. En todos los ejemplos que consideremos, salvo (8.9) y (8.21), tendremos que b y, por lo tanto B, serán la identidad, es decir, que la función de verosimilitud se expresará de la forma

$$\mathcal{L}(\beta, \mathtt{Y}) = \exp\left\{\langle \mathtt{X}'\beta, \mathtt{Y}\rangle + C(\beta) + D(\mathtt{Y})\right\} \tag{8.2}$$

La función g verificando las condiciones anteriores se denomina función de ligadura.

Al igual que el modelo lineal puede obtenerse condicionando en un modelo de correlación, muchos de los modelos lineales generalizados se obtendrán, como veremos, condicionando en otro modelo previo.

Por otra parte, si $\mathcal{L}_i$ denota la función de verosimilitud correspondiente a la componente Y_i y $l_i = \log \mathcal{L}_i$, se sigue que

$$l_i(\mu_i, \mathtt{Y}_i) = \mathtt{Y}_i \cdot [b \circ g](\mu_i) + c(\mu_i) + d(\mathtt{Y}_i) \tag{8.3}$$

El logaritmo l de la función de verosimilitud $\mathcal{L}$ se expresa a través de $\beta = (\beta_1, \ldots, \beta_s)'$ de la forma

$$l(\beta, \mathtt{Y}) = \sum_{i=1}^{\mathtt{n}} \left[\mathtt{Y}_i \cdot b(\mathtt{X}_i'\beta) + c\left(g^{-1}(\mathtt{X}_i'\beta)\right) + d(\mathtt{Y}_i)\right] \tag{8.4}$$

Sea $\mathtt{U}$ el vector aleatorio $\mathtt{n}$-dimensional de componentes

$$\mathtt{U}_i = \frac{dl_i}{d\mu_i} \qquad 1 \leq i \leq \mathtt{n} \tag{8.5}$$

En ese caso, se sigue de (9.35) que

$$\mathtt{E}[\mathtt{U}_i] = 0, \qquad \mathtt{var}[\mathtt{U}_i] = \mathtt{E}\left[-\frac{d\mathtt{U}}{d\mu_i}\right] \tag{8.6}$$

De (8.3) y (9.34) se sigue (cuestión propuesta) el siguiente resultado

Lema 8.1.
Para todo $i = 1, \ldots, \mathtt{n}$, se verifica

$$\begin{aligned} \mu_i &= -\frac{c'(\mu_i)}{b'(\mathtt{X}_i'\beta)\cdot g'(\mu_i)} \\ b'(\mathtt{X}_i'\beta) &= [\mathtt{var}[Y_i]\cdot g'(\mu_i)]^{-1} \end{aligned}$$

Si se denota $\mathtt{X}_i' = (\mathtt{X}_{i1}, \ldots, \mathtt{X}_{ij})$, se sigue de lo anterior (cuestión propuesta) lo siguiente

Proposición 8.2.

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{\mathtt{n}} \frac{(\mathtt{Y}_i - \mu_i)\mathtt{X}_{ij}}{\mathtt{var}[Y_i]\cdot g'(\mu_i)} \\ &= \sum_{i=1}^{\mathtt{n}} [\mathtt{Y}_i - g^{-1}(\mathtt{X}_i'\beta)]\cdot b'\left(g^{-1}(\mathtt{X}_i'\beta)\right)\cdot \mathtt{X}_{ij} \end{aligned}$$

Por su parte, se sigue (cuestión propuesta) del lema 8.1, la proposición 8.2 y (9.35) lo siguiente:

Proposición 8.3.
Las componentes de la matriz de información del modelo pueden expresarse mediante

$$\begin{aligned} \mathcal{I}_{jk} &= \sum_{i=1}^{\mathtt{n}} \frac{\mathtt{X}_{ij}\mathtt{X}_{ik}}{\mathtt{var}_{\mu_i}[Y_i]\cdot [g'(\mu_i)]^2} \\ &= \sum_{i=1}^{\mathtt{n}} \frac{\mathtt{X}_{ij}\mathtt{X}_{ik}}{\mathtt{var}_{\beta}[Y_i]\cdot \left[g'\left(g^{-1}(\mathtt{X}_i'\beta)\right)\right]^2} \end{aligned}$$

para $1 \leq j, k \leq s$

Corolario 8.4.
La matriz de información del modelo para β es la siguiente

$$\mathcal{I} = \mathtt{X}'W\mathtt{X}, \tag{8.7}$$

siendo W la matriz diagonal de componentes

$$w_{ii} = \left(\mathtt{var}_{\mu_i}[Y_i]\cdot [g'(\mu_i)]^2\right)^{-1}, \qquad 1 \leq i \leq \mathtt{n}$$

Este resultado será de gran utilidad tanto en la estimación de β como en el contraste de hipótesis.

8.2. Ejemplos

Veamos algunos modelos que pueden adaptarse a este formato, así como diversos problemas prácticos que pueden ser formalizados mediante estos modelos. El modelo lineal normal no es una estructura de este tipo dado que la distribución de los datos depende de la varianza, por lo que los algoritmos que estudiaremos a continuación no son, en principio, de aplicación en dicho modelo. Otra cosa es que la supongamos conocida. De hecho, si aplicamos el principio de sustitución y *arrastramos* ese parámetro hasta el final, dichos algoritmos conducen a las mismas estimaciones de β que se obtienen con el modelo lineal y al propio test F. Sólo en ese sentido podemos hablar de una generalización del modelo lineal normal.

Modelo lineal normal con varianza σ^2 conocida

Si conocemos el valor de la varianza σ^2 en un modelo lineal podemos dividir por σ cada dato Y_i de media μ_i, obteniendo $Y_i^* = \sigma^{-1}Y_i$, de media $\mu_i^* = \sigma^{-1}\mu_i$. Esta homotecia conduce a un nuevo modelo equivalente, concretamente $Y^* \sim N_{\mathtt{n}}(\mu^*, \mathtt{Id})$, siendo la densidad de cada componente la siguiente

$$\begin{aligned} f_{\mu_i^*}(\mathtt{y}_i^*) &= (2\pi)^{-1/2}\exp\left\{-\frac{1}{2}(\mathtt{y}_i^* - \mu_i^*)^2\right\} \\ &= (2\pi)^{-1/2}\exp\left\{-\frac{1}{2}(\mu_i^*)^2\right\}\exp\left\{\mu_i^* \cdot \mathtt{y}_i^*\right\}\exp\left\{-\frac{1}{2}(\mathtt{y}_i^*)^2\right\} \end{aligned}$$

Así pues, estamos hablando de un producto de $\mathtt{n}$ densidades del tipo (9.36) con

$$\theta_i = \mu_i^*, \quad , T(Y_i^*) = Y_i^*, \quad c(\mu_i^*) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}(\mu_i^*)^2,$$

$$Q(\mu_i^*) = \mu_i^*, \qquad d(Y_i^*) = -\frac{1}{2}(y_i^*)^2$$

Impongamos la restricción propia del modelo lineal de que, dados $\mathtt{X}_1, \ldots, \mathtt{X}_{\mathtt{n}} \in \mathbb{R}^s$, se verifique $\mu_i = \mathtt{X}_i'\beta$, para todo i, es decir, $\mu = \mathtt{X}\beta$. En los términos del modelo transformado se expresaría mediante $\mu^* = \mathtt{X}^*\beta$, siendo $\mathtt{X}^* = \sigma^{-1} \cdot \mathtt{X}$. la función de verosimilitud del modelo podrá expresarse de la forma (8.2). Concretamente

$$\mathcal{L}(\beta, \mathtt{Y}^*) = \exp\left\{\langle \mathtt{X}^*\beta, \mathtt{Y}^*\rangle - \frac{\mathtt{n}}{2}\log(2\pi) - \frac{1}{2}\|\mathtt{X}^*\beta\|^2 - \|\mathtt{Y}^*\|^2\right\} \tag{8.8}$$

Se trata de un modelo lineal generalizado dada $\mathtt{X}^*$ con función de ligadura $g = \mathtt{Id}$.

Regresión de Poisson

Consideremos $Y_1, \ldots, Y_{\mathtt{n}}$ variables aleatorias independientes distribuidas respectivamente según un modelo $P(\lambda_i)$, es decir, con densidad respecto a la medida cardinal

$$f_{\lambda_i}(\mathtt{y}_i) = \mathtt{e}^{-\lambda_i} \frac{\lambda_i^{\mathtt{y}_i}}{\mathtt{y}_i!} \qquad \mathtt{y}_i = 0, 1, \ldots$$

Esta densidad puede expresarse también mediante

$$f_{\lambda_i}(\mathtt{y}_i) = \exp\left\{\mathtt{y}_i \cdot \log \lambda_i - \lambda_i - \log(\mathtt{y}_i!)\right\}$$

Una situación como esta puede darse cuando cada Y_i es el número de sucesos contabilizados en un determinado periodo de tiempo, que se denota por i, en el que existe un fenómeno de *pérdida de memoria*[1].

Puede suceder que la media de esta distribución crezca de manera lineal en relación con cierta variable real Z, con valores $\mathtt{Z}_1, \ldots, \mathtt{Z}_{\mathtt{n}}$, es decir, que existan $\beta_0, \beta_1 \in \mathbb{R}$ tales que $\lambda_i = \beta_0 + \beta_1 \mathtt{Z}_i$, $i = 1, \ldots, \mathtt{n}$. En tal caso, si se denota $\mathtt{X}_i' = (1, \mathtt{Z}_i)$ y $\beta = (\beta_0, \beta_1)'$, la función de verosimilitud correspondientes a $(Y_1, \ldots, Y_{\mathtt{n}})$ se expresará de la forma (8.1), concretamente

$$\mathcal{L}_1(\beta, \mathtt{Y}) = \exp\left\{\sum_{i=1}^{\mathtt{n}} \left[\mathtt{Y}_i \cdot \log(\mathtt{X}_i'\beta) - \mathtt{X}_i'\beta - \log(\mathtt{y}_i!)\right]\right\} \tag{8.9}$$

Se trata de un modelo lineal generalizado dada $\mathtt{X} = (1_{\mathtt{n}}|\mathtt{Z})$ con función de ligadura $g(x) = \mathtt{Id}$ con $b(x) = \log(x)$.

En otras ocasiones podemos suponer un crecimiento exponencial de λ_i. Puede suceder, por ejemplo, cuando se contabilizan las muertes atribuibles a un enfermedad en una población grande durante un cierto intervalos de tiempos iguales y consecutivos, $i = 1, \ldots, \mathtt{n}$. Es decir, suponemos que existe un número β tal que

$$\lambda_i = i^\beta$$

En ese caso, $g(\lambda_i) = (\log i) \cdot \beta$. Por lo tanto, con estos supuestos, si se denota $\mathtt{X}_i = \log i$, la función de verosimilitud del modelo puede expresarse de la forma (8.2) mediante

$$\mathcal{L}_2(\beta, \mathtt{Y}) = \exp\left\{\langle \mathtt{X}\beta, \mathtt{Y}\rangle - \sum_{i=1}^{\mathtt{n}} \left[i^\beta - \log(\mathtt{y}_i!)\right]\right\} \tag{8.10}$$

Se trata de un modelo lineal generalizado dada $\mathtt{X} = (\log 1, \ldots, \log \mathtt{n})'$ con función de ligadura $g(x) = \log x$.

[1] Ver Nogales (1998).

Regresión logística

Una de las más importantes aplicaciones de los modelos lineales generalizados es la resolución de problemas de regresión con un vector explicativo q-dimensional $Z = (Z_1, \ldots, Z_q)'$ y una variable respuesta binaria Y. Supongamos que nuestra variable Y toma valores 1 ó 0. Si contamos con $\mathtt{n}$ réplicas independientes $(Y_1, Z_1), \ldots, (Y_{\mathtt{n}}, Z_{\mathtt{n}})$ y, para cada $1 \leq i \leq \mathtt{n}$, se denota $\pi_i = P(Y = 1 | Z_i = \mathtt{z}_i)$, se verifica que la función de verosimilitud del modelo condicional[2] de $(Y_1, \ldots, Y_{\mathtt{n}})$ dados $Z_i = \mathtt{z}_i$, $1 \leq i \leq \mathtt{n}$, es la siguiente

$$\mathcal{L}_{\mathtt{z}}\big(\pi_1, \ldots, \pi_{\mathtt{n}}, \mathtt{Y}_1, \ldots, \mathtt{Y}_{\mathtt{n}}\big) = \prod_{i=1}^{\mathtt{n}} (1 - \pi_i) \cdot \exp\left\{ \sum_{i=1}^{\mathtt{n}} \mathtt{Y}_i \cdot \log \frac{\pi_i}{1 - \pi_i} \right\} \tag{8.11}$$

Consideremos la función siguiente

$$g(x) = \log \frac{x}{1 - x}, \quad 0 \leq x \leq 1 \tag{8.12}$$

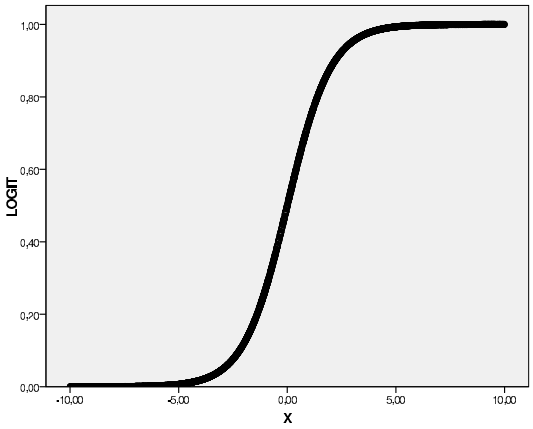
Denótese $\theta = (g(\pi_1), \ldots, g(\pi_{\mathtt{n}}))' \in \mathbb{R}^{\mathtt{n}}$ y $\mathtt{Y} = (\mathtt{Y}_1, \ldots, \mathtt{Y}_{\mathtt{n}})'$. En ese caso, podemos expresar (8.11) de forma canónica mediante

$$\mathcal{L}_{\mathtt{z}}(\theta, \mathtt{Y}) = \exp\left\{ \langle \theta, \mathtt{Y} \rangle + C^*(\theta) \right\} \tag{8.13}$$

siendo $C^*(\theta) = \sum_{i=1}^{\mathtt{n}} \log\left(1 - g^{-1}(\theta_i)\right)$. La función (8.12) es la inversa de la siguiente, que se denomina función logística:

$$\mathtt{L}(x) = \frac{\mathtt{e}^x}{1 + \mathtt{e}^x}, \qquad x \in \mathbb{R}$$

[2]Que está dominado por la medida uniforme en $\{0, 1\}$.



Las medias condicionales pueden por tanto obtenerse a partir de las componentes de θ mediante

$$\pi_i = \mathtt{L}(\theta_i), \qquad 1 \leq i \leq \mathtt{n} \tag{8.14}$$

El uso de esta función no responde únicamente a criterios estéticos, sino que puede venir dada por la aceptación de diversos supuestos, más o menos naturales, en diferentes problemas a resolver, y que conducen a un modelo lineal generalizado con la función $\mathtt{L}^{-1}$ desempeñando el papel de ligadura. Distinguimos dos situaciones.

- **Análisis discriminante:** supongamos que la distribución de Z condicionada a Y es

  $$P^{Z|Y=j} = N_q(\mu_j, \Sigma), \quad j = 0, 1$$

  Entiéndase Y como un factor aleatorio que distingue dos distribuciones normales con idéntica matriz de varianzas-covarianzas. Son los mismos supuestos que, en un análisis discriminante, permiten aplicar la estrategia de clasificación lineal de Fisher[3]. En ese caso, se sigue de la regla de Bayes (cuestión propuesta) que

  $$P(Y = 1|Z = \mathtt{z}) = \mathtt{L}\big(-(\beta_0 + \mathtt{z}\underline{\beta})\big) \tag{8.15}$$

  donde

  $$\begin{aligned} \beta_0 &= \log\frac{1-q}{q} + \mu_1'\Sigma^{-1}\mu_1 - \mu_0'\Sigma^{-1}\mu_0, \\ \underline{\beta} &= \Sigma^{-1}(\mu_0 - \mu_1). \end{aligned}$$

[3]Ver volumen dedicado al Análisis Multivariante.

Es decir, si se denota $\beta = (\beta_0, \underline{\beta}')'$ y $\mathtt{X}_i' = -(1, \mathtt{Z}_i')$, se sigue de (8.14) que

$$\theta_i = \mathtt{X}_i'\beta, \qquad 1 \leq i \leq \mathtt{n}$$

o, equivalentemente,

$$g(\pi_i) = \mathtt{X}_i'\beta, \qquad 1 \leq i \leq \mathtt{n}$$

En consecuencia, el modelo condicional dada la matriz explicativa $\mathtt{Z}$ verifica presenta una función de verosimilitud del tipo (8.2) con $\mathtt{X} = -(1_{\mathtt{n}}|\mathtt{Z})$ y $g = \mathtt{L}^{-1}$. Concretamente,

$$\mathcal{L}_{\mathtt{z}}(\beta, \mathtt{Y}) = \exp\left\{\langle \mathtt{X}\beta, \mathtt{Y}\rangle + \sum_{i=1}^{\mathtt{n}} \log[1 - \mathtt{L}(\mathtt{X}_i'\beta)]\right\} \tag{8.16}$$

Luego, estaremos hablando de un modelo lineal generalizado dada la matriz de regresión $\mathtt{X}$ con función de ligadura $\mathtt{L}^{-1}$. En definitiva, una buena estimación del parámetro β nos permitirá predecir con bastante exactitud la probabilidad de que Y tome el valor 0 ó 1 a partir de los valores obtenidos en Z.

- **Modelos de respuesta a una dosis:** estudiamos en este apartado la relación existente entre la dosis de una sustancia y la probabilidad de éxito de la misma. Puede tratarse de un medicamento o bien un veneno para animales o plantas; en el primer caso el éxito consistiría en la curación mientras que, en el segundo, sería la muerte del individuo. Desde el punto de vista histórico, se trata de una de las primeras aplicaciones de modelos derivados de la regresión lineal. Consiste pues, al igual que el análisis discriminate, en un modelo de regresión simple con una variable respuesta Y con valores en $\{0, 1\}$ y una variable explicativa Z con valores en $[0, +\infty)$.

  En la práctica es frecuente que el éxito sea imposible por debajo de un umbral mínimo de dosis c_1 y que sea seguro por encima de un umbral máximo c_2. También puede resultar natural que la probabilidad de éxito crezca de manera lineal entre ambos umbrales. Es decir,

$$P(Y = 1|Z = \mathtt{z}) = \begin{cases} 0 & \text{si } \mathtt{z} < c_1 \\ \frac{\mathtt{z}-c_1}{c_2-c_1} & \text{si } c_1 \leq \mathtt{z} \leq c_2 \\ 1 & \text{si } \mathtt{z} > c_2 \end{cases} \tag{8.17}$$

  Tener en cuenta que $(\mathtt{z} - c_1)(c_2 - c_1)^{-1} = \beta_0 + \beta_1\mathtt{z}$ para $\beta_0 = (c_1 - c_2)^{-1}$ y $\beta_1 = -c_1(c_1 - c_2)^{-1}$.

También podemos expresar (8.17) mediante

$$P(Y=1|Z=\mathtt{z}) = \int_{-\infty}^{\mathtt{z}} f(s)\, ds,$$

donde la función f, denominada **función de tolerancia**, se define mediante

$$f(s) = \begin{cases} \frac{1}{c_2-c_1} & \text{si } c_1 \le s \le c_2 \\ 0 & \text{en caso contrario} \end{cases} \tag{8.18}$$

Nótese que se trata, lógicamente, de una función de densidad. Si consideramos $\mathtt{n}$ réplicas independientes, $(Y_1, Z_1), \ldots, (Y_{\mathtt{n}}, \mathtt{Z}_{\mathtt{n}})$, el modelo condicional de $(Y_1, \ldots, Y_{\mathtt{n}})$ dado $Z_1 = \mathtt{z}_1, \ldots, Z_{\mathtt{n}} = \mathtt{z}_{\mathtt{n}}$ no puede considerarse lineal generalizado para $\mathtt{X} = (1_{\mathtt{n}}|\mathtt{Z})$ con la función de ligadura

$$g(x) = \int_{-\infty}^{x} f(s)\, ds \tag{8.19}$$

pues, en (8.17) sólo tenemos una relación lineal a trozos. Este problema puede resolverse reemplazando la función de tolerancia (8.18) por otra que sea continua. Para ese fin puede valernos una curva normal

$$f(s) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{s-\mu}{\sigma^2}\right)^2\right\} \tag{8.20}$$

Efectivamente, en ese caso, si Φ denota la función de distribución del modelo $N(0,1)$, se verifica en virtud del teorema de cambio ed variables,

$$P(Y_i = 1|Z_i = \mathtt{z}_i) = \Phi\left(\frac{\mathtt{z}_i - \mu}{\sigma^2}\right), \qquad 1 \le i \le \mathtt{n}$$

Por lo tanto, la función de verosimilitud del modelo condicional se expresará con la ayuda del parámetro $\beta = (\sigma^{-1}, \sigma^{-1}\mu)'$ de la forma (8.1). Concretamente

$$\mathcal{L}_{\mathtt{z}}(\beta, \mathtt{Y}_1, \ldots, \mathtt{Y}_{\mathtt{n}}) = \exp\left\{\sum_{i=1}^{\mathtt{n}}\left[\mathtt{Y}_i \cdot \log\frac{\Phi(\mathtt{X}_i'\beta)}{1-\Phi(\mathtt{X}_i'\beta)} + \log[1-\Phi(\mathtt{X}_i'\beta)]\right]\right\} \tag{8.21}$$

En este caso, estamos considerando las funciones b y g siguientes

$$b(x) = \log\frac{\Phi(x)}{1-\Phi(x)}, \qquad g(y) = \Phi^{-1}(y)$$

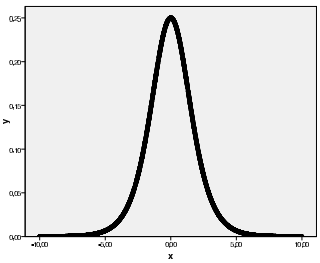
Los experimentos estadísticos de este tipo se denominan **modelos probit**. Estimar el parámetro β equivale a estimar los valores μ y σ de la función de

tolerancia. En los estudios relacionados con venenos el parámetro μ se denomina dosis letal media.

Sin embargo, podemos obtener un modelo lineal generalizado más sencillo si consideramos como función de tolerancia esta otra función de densidad, bastante similar a una curva normal, que depende de dos parámetros reales $\beta_0, \beta_1 \in \mathbb{R}$

$$f(s) = \frac{\beta_1 \exp\{\beta_0 + \beta_1 \cdot s\}}{[1 + \exp\{\beta_0 + \beta_1 \cdot s\}]^2}, \qquad s \in \mathbb{R} \tag{8.22}$$

Veamos qué aspecto tiene esta función para $\beta_0 = 0$ y $\beta_1 = 1$.



En ese caso, se verifica, para $1 \leq i \leq \mathtt{n}$

$$P(Y_i = 1 | Z_i = \mathtt{z}_i) = \int_{-\infty}^{\mathtt{z}_i} f(s)\, ds = \mathtt{L}(\beta_0 + \beta_1 \cdot \mathtt{z}_i)$$

Dado que

$$\mathtt{L}^{-1}(x) = \log \frac{x}{1-x}$$

Página 1

la función de verrosimilitud del modelo para $Y = (Y_1, \ldots, Y_{\mathtt{n}})$ condicionado a $Z_1 = \mathtt{z}_1, \ldots, Z_{\mathtt{n}} = \mathtt{z}_{\mathtt{n}}$ puede expresarse a través del parámetro $\beta = (\beta_0, \beta_1)'$ de la forma (8.16). Concretamente

$$\mathcal{L}_{\mathtt{z}}(\beta, \mathtt{Y}) = \exp\left\{ \langle \mathtt{X}\beta, \mathtt{Y} \rangle + \sum_{i=1}^{\mathtt{n}} \log[1 - \mathtt{L}(\mathtt{X}\beta)] \right\} \tag{8.23}$$

Este tipo de modelos basados en el uso de la función logística se denominan en la literatura estadística **modelos logit**.

8.3. Estudio asintótico

Según se desprende de lo estudiado hasta ahora, una de las tareas más importantes que debemos afrontar es la estimación del parámetro β del modelo (o del modelo condicional). La labor resultará mucho más difícil de lo que fue la estimación de μ o β en el modelo lineal normal o lineal a secas. En este caso buscaremos el estimador de máxima verosimilitud pues el procedimiento procura mejores estimaciones que las que obtendríamos según el método de mínimos cuadrados[4]. Es decir, buscaremos el valor de β que maximiza la función de verosimilitud (8.1) dados $\mathtt{Y}_1, \ldots, \mathtt{Y}_\mathtt{n}$. Concretamente, se denotará por $\hat{\beta}^{\text{MV}}$ cualquier vector de $\mathbb{R}^s$ cuyas componentes sean soluciones a las ecuaciones $\partial l/\partial\beta_j = 0, = 1 \leq j \leq s$. Esta sección se centra en las bien conocidas propiedades asintóticas de este estimador. Para un estudio exhaustivo recomendamos Lehmann (1983) y Ferguson (1996). La primera que debemos destacar es la consistencia del estimador. Efectivamente, según el teorema 2.2. del capítulo 6 de Lehmann (1983), queda garantizada la existencia de una secuencia de soluciones al sistema de ecuaciones anterior que converge en probabilidad a β. Partiendo de esta propiedad y utilizando herramientas básicas de la estadística asintótica, describiremos la distribución límite de $\hat{\beta}^{\text{MV}}$ en el modelo condicional. Las hipótesis que precisamos son las siguientes: supongamos que (Y_i, X_i), $i \in \mathbb{N}$, es una sucesión de vectores aleatorios $s+1$-dimensionales iid según un modelo de distribución $P_\beta^{Y|X} \times P^X$, para algún $\beta \in \mathbb{R}^s$, siendo $P_\beta^{Y|X=\mathtt{x}}$ una distribución dominada con función de verosimilitud del tipo

$$\mathcal{L}_{\mathtt{X}_i}(\beta, \mathtt{Y}_i) = \exp\{\mathtt{Y}_i \cdot q(\mathtt{X}_i'\beta) + c(\mathtt{X}_i'\beta) + d(\mathtt{Y}_i)\},$$

verificando q y c las condiciones de regularidad necesarias, y sea, para cada $\mathtt{n}$, $\hat{\beta}^{\text{MV}}{}_\mathtt{n} \in \mathbb{R}^s$ una solución al sistema de ecuaciones $U_j^\mathtt{n}(b) = 0$, para $1 \leq j \leq s$, siendo

$$U_j^\mathtt{n}(b) = \frac{\partial \log\left(\prod_{i=1}^\mathtt{n} \mathcal{L}(\beta, \mathtt{Y}_i)\right)}{\partial\beta_j}(b), \qquad b \in \mathbb{R}^s$$

$I_{(\beta),\mathtt{n}}$ y $I_{(\beta)}$ denotarán las matrices de información de los modelos condicionados con $\mathtt{n}$ y 1 datos, respectivamente.

Teorema 8.5.

En las condiciones anteriores, si $\mathcal{I}_\beta$ no es singular, se verifica la siguiente convergencia en distribución cuando $\mathtt{n}$ tiende a infinito:

$$\mathcal{I}_{(\beta),\mathtt{n}}^{1/2} \cdot (\hat{\beta}^{\text{MV}}{}_\mathtt{n} - \beta) \longrightarrow N_s(0, \mathtt{Id}) \tag{8.24}$$

[4] Ver Dobson (1990)

Demostración. Primeramente, se verifica trivialmente

$$I_{(\beta),\mathtt{n}} = \mathtt{n} \cdot I_{(\beta)}$$

Denótese $U^{\mathtt{n}} = (U_1^{\mathtt{n}}, \ldots, U_s^{\mathtt{n}})'$ y considérese un desarrollo de Taylor de grado 1 de $U^{\mathtt{n}}(\beta)$ en torno a $\hat{\beta}^{\text{MV}}$:

$$U^{\mathtt{n}}(\beta) = 0 + H_{\mathtt{n}}(\hat{\beta}^{\text{MV}})(\beta - \hat{\beta}^{\text{MV}}) + \frac{1}{2}(\beta - \hat{\beta}^{\text{MV}})' f(\hat{\beta}^{\text{MV}})(\beta - \hat{\beta}^{\text{MV}})$$

siendo

$$H_{\mathtt{n}}(b) = \begin{pmatrix} \frac{\partial^2 \log\left(\prod_{i=1}^{\mathtt{n}} \mathcal{L}(\beta, \mathtt{Y}_i)\right)}{\partial\beta_1 \partial\beta_1}(b) & \cdots & \frac{\partial^2 \log\left(\prod_{i=1}^{\mathtt{n}} \mathcal{L}(\beta, \mathtt{Y}_i)\right)}{\partial\beta_1 \partial\beta_s}(b) \\ \vdots & & \vdots \\ \frac{\partial^2 \log\left(\prod_{i=1}^{\mathtt{n}} \mathcal{L}(\beta, \mathtt{Y}_i)\right)}{\partial\beta_s \partial\beta_1}(b) & \cdots & \frac{\partial^2 \log\left(\prod_{i=1}^{\mathtt{n}} \mathcal{L}(\beta, \mathtt{Y}_i)\right)}{\partial\beta_s \partial\beta_s}(b) \end{pmatrix}$$

y siendo $f(b)$ una función con valores en $\mathcal{M}_{s\times s}$ que podemos suponer continua por las condiciones de regularidad de q y c. Por la consistencia de $\hat{\beta}^{\text{MV}}{}_{\mathtt{n}}$ podemos despreciar el último sumando del segundo término, pues converge a 0 en probabilidad. Lo expresamos así

$$\hat{\beta}^{\text{MV}}{}_{\mathtt{n}} - \beta = H_{\mathtt{n}}^{-1}(\hat{\beta}^{\text{MV}}) \cdot U^{\mathtt{n}}(\beta)$$

Teniendo en cuenta nuevamente la consistencia de $\hat{\beta}^{\text{MV}}{}_{\mathtt{n}}$ y aplicando la LDGN junto con el teorema 9.21 se deduce

$$\mathtt{n}^{-1} H_{\mathtt{n}}(\hat{\beta}^{\text{MV}}) \longrightarrow \mathcal{I}_\beta \tag{8.25}$$

Respecto al segundo factor, nótese que

$$U_j^{\mathtt{n}}(\beta) = \sum_{i=1}^{\mathtt{n}} a_i^j(\beta), \quad 1 \leq j \leq s,$$

siendo

$$a_i^j = \mathtt{Y}_i \cdot \mathtt{X}_{ij} \cdot \partial q/\partial\beta_j + \mathtt{X}_{ij} \cdot \partial c/\partial\beta_j$$

Los vectores aleatorios $(a_i^1, \ldots, a_i^s)'$, $i \in \mathbb{N}$, constituyen una sucesión iid con esperanza nula, por (9.34), y matriz de varianzas y covarianzas $\mathcal{I}_\beta$. Se sigue entonces del TCL iid multivariante, que

$$\mathtt{n}^{-1/2} U^{\mathtt{n}}(\beta) \longrightarrow N_s(0, \mathcal{I}_\beta) \tag{8.26}$$

Teniendo en cuenta (8.25), (8.26) junto con el teorema 9.21, se deduce (8.24).

□

Nótese que, para un tamaño de muestra $\mathtt{n}$ suficientemente grande, se verifica, aproximadamente,

$$\hat{\beta}^{\mathtt{MV}}{}_{\mathtt{n}} \sim N_s(\beta, \mathcal{I}_{\mathtt{n}}^{-1}) \tag{8.27}$$

Lo cual implica, en términos aproximados, insistimos, no sólo que $\hat{\beta}^{\mathtt{MV}}{}_{\mathtt{n}}$ sea insesgado, sino que su matriz de varianzas-covarianzas alcanza la cota mínima de Cramer-Rao[5], por lo que podríamos considerarlo como asintóticamente insesgado de mínima varianza. En ese sentido se dice que es un estimador asintóticamente eficiente (además de consistente). También podemos obtener como consecuencia inmediata el siguiente resultado:

Corolario 8.6.
En las condiciones anteriores se verifica

$$\mathtt{W} = \left(\hat{\beta}^{\mathtt{MV}}{}_{\mathtt{n}} - \beta\right)' \mathcal{I}_{\mathtt{n}} \left(\hat{\beta}^{\mathtt{MV}}{}_{\mathtt{n}} - \beta\right) \longrightarrow \chi_s^2 \tag{8.28}$$

Dado β_0 fijo, la función $\mathtt{W}$ se denomina **estadístico de Wald**. Conocida la matriz de información, (8.27) puede utilizarse, por ejemplo, para construir intervalos de confianza para las componentes de β. Concretamente, si ψ_{jk} denota la componente (j,k)-ésima de $\mathcal{I}_{\mathtt{n}}^{-1}$, serán de la forma

$$\hat{\beta}^{\mathtt{MV}j}{}_{\mathtt{n}} \pm z_\alpha \sqrt{\psi_{jj}}, \qquad 1 \leq j \leq s \tag{8.29}$$

Asimismo, a partir de (8.24), podemos construir regiones de confianza elípticas para β, concretamente

$$(\beta - \hat{\beta}^{\mathtt{MV}}{}_{\mathtt{n}})' \mathcal{I} (\beta - \hat{\beta}^{\mathtt{MV}}{}_{\mathtt{n}}) \leq \chi_s^{2,\alpha} \tag{8.30}$$

En las misma condiciones del teorema 8.5 y siguiendo razonamientos completamente análogos pero aplicados al logaritmo de la función de verosimilitud l, en lugar de a su derivada, y mediante un desarrollo de Taylor de orden 2, en lugar e 1, en torno a $\hat{\beta}^{\mathtt{MV}}{}_{\mathtt{n}}$, se obtiene el siguiente resultado cuya demostración queda como ejercicio

Teorema 8.7.
En las condiciones anteriores se verifica

$$2[l(\hat{\beta}^{\mathtt{MV}}{}_{\mathtt{n}}) - l(\beta)] \longrightarrow \chi_s^2$$

[5] Ver Lehmann(1983).

8.4. Estimación y contraste de de hipótesis

Todo lo dicho en la sección anterior tiene en principio un valor meramente teórico, veremos por qué. Se definió $\hat{\beta}^{\text{MV}}$ como cualquier vector de $\mathbb{R}^s$ cuyas componentes sean soluciones a las ecuaciones $\partial l/\partial\beta_j = 0$, $1 \leq j \leq s$. En las condiciones del primer ejemplo dedicado a un modelo lineal normal puede comprobarse sin dificultad que un vector b es solución al sistema de ecuaciones anteriores sil y sólo si, lo es del siguiente sistema de ecuaciones lineales:

$$\mathtt{X}'^{*}\mathtt{X}^{*}b = (\mathtt{X}^{*})'Y^{*} \tag{8.31}$$

Nótese que, si expresamos la ecuación en los términos originales (sin dividir por σ), ésta queda como sigue:

$$\mathtt{X}'\mathtt{X}b = \mathtt{X}'Y \tag{8.32}$$

Por lo tanto, para encontrar la solución final no es necesario conocer el valor de σ^2 pues no depende del mismo.

Pero esto no deja de ser una excepción pues, en general, puede tratase de un sistema de ecuaciones no lineales cuya solución deba obtenerse de manera aproximada mediante un método iterativo. Seguramente, lo más natural primera vista sea aplicar el procedimiento e Newton-Raphson. Concretamente, si U y H denotan respectivamente el vector y la matriz definidas en el teorema 8.5, y si $b^{(m-1)}$ denota una solución en la fase $(m-1)$-ésima, la solución *mejorada* en la fase m-ésima se obtiene mediante

$$b^{(m)} = b^{(m-1)} - \left(H[b^{(m-1)}]\right)^{-1} \cdot U[b^{(m-1)}] \tag{8.33}$$

Un procedimiento alternativo, más simple desde el punto de vista operativo, consiste en reemplazar H por su valor medio, es decir, $\mathcal{I}_{\mathtt{n}}$. De esta forma, (8.33) quedaría como sigue

$$b^{(m)} = b^{(m-1)} - \left(\mathcal{I}_{\mathtt{n},b^{(m-1)}}\right)^{-1} \cdot U[b^{(m-1)}]$$

Es decir,

$$\mathcal{I}_{\mathtt{n},b^{(m-1)}} b^{(m)} = \mathcal{I}_{\mathtt{n},b^{(m-1)}} b^{(m-1)} + U[b^{(m-1)}] \tag{8.34}$$

Se sigue entonces de las proposiciones 8.2 y 8.3 que, para $j = 1, \ldots, s$, la componente j-ésima del segundo término de la ecuación (8.34) puede expresarse así

$$\sum_{k=1}^{s}\sum_{i=1}^{\mathtt{n}} \frac{\mathtt{X}_{ij}\cdot\mathtt{X}_{ik}\cdot b_k^{(m-1)}}{\mathtt{var}_{b^{(m-1)}}[Y_i]\cdot\left[g'\left(g^{-1}(\mathtt{X}_i'b^{(m-1)})\right)\right]^2} + \sum_{i=1}^{\mathtt{n}} \frac{[Y_i - g^{-1}(\mathtt{X}_i'b^{(m-1)})]\cdot\mathtt{X}_{ij}}{\mathtt{var}_{b^{(m-1)}}[Y_i]\cdot g'\left[g^{-1}\left(\mathtt{X}_i'b^{(m-1)}\right)\right]}$$

En definitiva, si consideramos la matriz W definida en el corolario 8.4 y valorada en $b^{(m-1)}$, el vector $b^{(m)}$ será la solución al sistema de ecuaciones linales siguientete

$$\mathtt{X}'W\mathtt{X}b^{(m)} = \mathtt{X}'Wz, \tag{8.35}$$

siendo z el vector de $\mathbb{R}^{\mathtt{n}}$ de componentes

$$z_i = \sum_{k=1}^{s} \mathtt{X}_{ik} \cdot b_k^{(m-1)} + \left[Y_i - g^{-1}(\mathtt{X}_i' b^{(m-1)})\right] \cdot g'\left[g^{-1}\left(\mathtt{X}_i' b^{(m-1)}\right)\right], \quad 1 \leq i \leq \mathtt{n}$$

En ese sentido y teniendo en cuenta (3.47), puede entenderse $b^{(m)}$ como una especie de solución mínimo-cuadrática generalizada. La ecuación (8.35) es, en definitiva, la que debe resolverse en cada paso.

Puede demostrarse (cuestión propuesta) que, en las condiciones del primer ejemplo, dedicado al modelo lineal con varianza conocida, se tiene que $W = \mathtt{Id}$ y $\mathtt{z}_i = Y_i^*$, para todo i, por lo que estaremos buscando, para todo $m \in \mathbb{N}$, una solución al sistema de ecuaciones lineales (8.31), o bien a (8.32) si la expresamos en los términos originales. En consecuencia, estamos considerando la propia solución mínimo-cuadrática (7.9).

El modelo que estudiamos en este capítulo reposa en una serie de hipótesis, entre las que se encuentra que la existencia de $\mathtt{X} \in \mathbb{R}^s$ tal que $g(\mu_i) = \mathtt{X}_i'\beta$ para todo i. Vamos a proponer a continuación un procedimiento para contrastar dicha hipótesis, lo cual puede entenderse parcialmente como una prueba de bondad de ajuste. Es obvio que si $s = \mathtt{n}$, la hipótesis anterior es completamente vacua, pues cualquier base de $\mathbb{R}^{\mathtt{n}}$ proporcionará un ajuste perfecto, en cuyo caso los datos obtenidos tendrán una máxima verosimilitud. El término β y su EMV se denotarán en ese caso por $\beta_{\text{máx}}$ y $\hat{\beta}^{\text{MV}}{}_{max}$, respectivamente. Se obtendrá pues un máximo valor para $2l(\hat{\beta}^{\text{MV}}{}_{max})$. Lo que se espera, si el modelo que proponemos es correcto, es que la diferencia con el término $2l(\hat{\beta}^{\text{MV}})$ sea pequeña. Ello es un indicio de que el modelo reducido con s parámetros puede hacer suficientemente verosímiles nuestras observaciones. En definitiva, denominamos **devianza** a la diferencia

$$D = 2[l(\hat{\beta}^{\text{MV}}{}_{max}) - l(\hat{\beta}^{\text{MV}})],$$

es decir

$$\begin{aligned} D &= 2[l(\hat{\beta}^{\text{MV}}{}_{max}) - l(\beta_{max})] \\ &- 2[l(\hat{\beta}^{\text{MV}}) - l(\beta)] \\ &+ 2[l(\beta_{max}) - l(\beta)] \end{aligned}$$

Como vemos, D se obtiene sumando esa diferencia positiva constante (tercer sumando) a la que hacíamos alusión anteriormente dos términos aleatorios que se restan (primer y segundo sumandos). En virtud del teorema 8.7, el primer término sigue

aproximadamente una distribución $\chi^2_{\mathtt{n}}$, mientras que el término que se resta sigue aproximadamente una distribución χ^2_s. Llegamos al punto más conflictivo: si ambos son independientes, cosa que no sucede en general, dicha diferencia debe seguir, aproximadamente, un modelo de distribución $\chi^2_{\mathtt{n}-s}$. En ese caso, si el modelo es correcto, se espera que el último sumando sea próximo a 0 y que, por lo tanto, D siga aproximadamente una distribución $\chi^2_{\mathtt{n}-s}$, de manera que un valor de D por encima de $\chi^{\alpha}_{\mathtt{n}-s}$ puede conducirnos a desechar el modelo con s parámetros.

Este procedimiento puede utilizarse también para eliminar algunos de los parámetros del modelo, es decir, para contrastar hipótesis del tipo

$$H_0 : \beta_{r+1} = \ldots = \beta_s = 0$$

Efectivamente, si reducimos a r la dimensión de β obtendremos un nuevo parámetro y un nuevo EMV del mismo que se denotarán, respectivamente, por β_0 y $\hat{\beta}^{\text{MV}}{}_0$. En ese caso, si la hipótesis inicial es correcta, cabe esperar que la diferencia $2[l(\hat{\beta}^{\text{MV}}) - l(\hat{\beta}^{\text{MV}}{}_0)]$ sea próxima a 0. Dicha diferencia puede expresarse mediante

$$\triangle D = D_0 - D_1$$

donde D_0 expresa la devianza del modelo reducido y D_1 la del original. Si se dieran las condiciones de independencia adecuadas, cosa que sucede en el modelo lineal normal con varianza conocida, y el modelo reducido es correcto, cabría esperar que $\triangle D$ se distribuya aproximadamente según un modelo χ^2_{s-r}. Así pues, se puede optar por desechar la reducción cuando $\triangle D > \chi^{2,\alpha}_{s-r}$. Desde luego, no es necesario advertir al lector que considerar este tipo de procedimientos como un tests de hipótesis a todos los efectos se antoja bastante aventurado.

No es ése, sin embargo, el caso del caso del modelo lineal normal con varianza conocida, pues se verifica también la independencia entre D_1 y $\triangle D$, lo cual supone una aproximación a la distribución $F_{s-r,\mathtt{n}-s}$ de $\triangle D/D_1$ si el modelo reducido es correcto. De hecho, se puede comprobar (cuestión propuesta) que en dicho modelo, se obtiene una distribución F-Snedecor exacta. Además, ya hemos visto en su momento cómo se trabaja con este modelo: se dividen los datos originales por σ^2. En ese caso, el cociente anterior no depende del valor de σ^2, por lo que el procedimiento para contrastar la hipótesis anterior es igualmente viable en el caso general de que la varianza no se conozca. Curiosamente, puede comprobarse (cuestión propuesta) que este procedimiento es el propio test F. Para más detalles consultar Doob (1990). Queda pues claro que los procedimientos estudiados en este capítulo generalizan los ya vistos en los anteriores.

Cuestiones propuestas

1. Probar el lema 8.1.

2. Probar las proposiciones (8.2) y (8.3).

3. Probar (8.15).

4. Obtener las matrices de información para los modelos (8.8), (8.9), (8.10) y (8.16).

5. Obtener los intervalos de confianza (8.29) y la región de confianza (8.30).

6. Probar el teorema 8.7.

7. Probar que en el modelo de regresión lineal normal con $\beta_{r+1} = \ldots = \beta_s = 0$, se verifica
$$\frac{D_0 - D_1}{D_1} \sim F_{s-r,\mathtt{n}-s}$$
Probar que el procedimiento para contrastar la hipótesis anterior coincide con el test F.

8. Probar que, en las condiciones del modelo lineal general con varianza conocida, se tiene que $W = \mathtt{Id}$ y $\mathtt{z}_i = Y_i^*$, para todo i, por lo que el método (8.35) se reduce a buscar la solución mínimo-cuadrática (7.9).

9. ¿En qué aspectos relativos a la estimación y contraste de hipótesis podemos afirmar que los procedimientos estudiados en este capítulo generalizan los ya conocidos de los capítulos anteriores?

Capítulo 9

Apéndice

En este capítulo abordamos un sucinto estudio de una serie de temas que estimamos necesarios para el correcto seguimiento de nuestra teoría. En primer lugar, repasaremos una serie de definiciones y resultados fundamentales de la teoría matricial y, en definitiva, del Álgebra Lineal, cuya relación con el Modelo Lineal resulta obvia; a continuación, en las dos secciones siguientes, realizamos un brevísimo repaso de las nociones fundamentales de Probabilidad y Estadística, imprescindible para una exposición rigurosa de la materia; posteriormente, se expone someramente en qué consiste y cómo se aplica el principio de Invarianza, el cual tendrá una enorme trascendencia en la justificación del test F; por último, se presentan las nociones y resultados fundamentales de la teoría asintótica que se utilizarán para analizar el comportamiento límite de los estimadores y tests de hipótesis obtenidos en la teoría.

9.1. Resultados de Álgebra Matricial

En esta sección nos limitamos a exponer una serie de resultados relativos al Álgebra Lineal que serán de utilidad en nuestra teoría. Aparte de esto, podemos encontrar en la sección 7.1 un amplio estudio del concepto de inversa generalizada de una matriz. Recordamos, en primer lugar, algunas definiciones.

Dada una matriz $A \in \mathcal{M}_{n\times n}$ (entendemos que sus coeficientes son reales), $\delta \in \mathbb{C}$ se dice autovalor de A cuando es raíz del polinomio de grado n $p(x) = |A - x\mathtt{Id}|$, lo cual significa que existe un vector $e \in \mathbb{C}^n$ tal que $Ae = \delta e$. En ese caso, decimos que e es un autovector asociado al autovalor δ, lo cual vale para toda la recta $\langle e\rangle$.

Consideremos $y = (y_1, \ldots, y_n)'$ y $x = (x_1, \ldots, x_n)'$ dos vectores cualesquiera de

$\mathbb{R}^n$. Se dice que x e y son perpendiculares u ortogonales cuando

$$\sum_{i=1}^{n} x_i y_i = 0, \tag{9.1}$$

lo cual se denota mediante $x \perp y$. Se define la norma euclídea de cada vector mediante

$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{9.2}$$

y la distancia euclídea entre dos vectores mediante

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{9.3}$$

Por lo tanto, la región del espacio formada por los puntos cuya distancia respecto a x sea igual a un cierto número positivo k es un esfera. El cuadrado de la distancia puede expresarse de esta forma

$$\|y - x\|^2 = (y - x)'\mathtt{Id}(y - x).$$

Si sustituimos la matriz identidad por cualquier matriz simétrica definida positiva A, la región anterior será un elipsoide, cuyas características dependerán de los autovectores y autovalores de A (ver teorema de diagonalización). Una expresión de este tipo pueden encontrarse en la densidad de la distribución normal multivariante.

Una sistema de vectores de $\mathbb{R}^n$ se dice ortonormal cuando los vectores son de norma euclídea 1 y ortogonales entre sí. Una matriz $\Gamma \in \mathcal{M}_{n\times n}$ se dice ortogonal cuando Γ' es su inversa, lo cual equivale a afirmar que sus columnas constituyen una base ortonormal de $\mathbb{R}^n$. En ocasiones las denominaremos *rotaciones*, ya veremos por qué. El conjunto de todas las matrices ortogonales de orden n se denotará por $\mathcal{O}_n$. Dado un subespacio vectorial $V \subset \mathbb{R}^n$, $V^\perp$ denota el subespacio vectorial de dimensión $n - \mathtt{dim}V$ constituidos por todos los vectores ortogonales a V. Asimismo, si $W \subset V$, $V|W$ denotará el subespacio $V \cap W^\perp$, de dimensión $\mathtt{dim}V - \mathtt{dim}W$.

Una matriz $A \in \mathcal{M}_{n\times n}$ se dice semidefinida positiva cuando es simétrica[1] y verifica que $e'Ae \geq 0$, para todo $e \in \mathbb{R}^n$, en cuyo caso se denota $A \geq 0$. Esta definición permite establecer un preorden en $\mathcal{M}_{n\times n}$. Concretamente,

$$A \geq B \text{ cuando } x'Ax \geq x'Bx, \text{ para todo } x \in \mathbb{R}^n. \tag{9.4}$$

Decimos que A es definida positiva cuando verifica $e'Ae > 0$, para todo $e \in \mathbb{R}^n \backslash \{0\}$, en cuyo caso se denota $A > 0$.

[1] En rigor, no es necesario que la matriz sea simétrica para que sea definida positiva, pero en nuestra teoría lo supondremos siempre.

Diagonalización

Dada una matriz $A \in \mathcal{M}_{n\times n}$, se definen sus autovalores como las raíces del polinomio en x $|A - x\mathtt{Id}|$. Por lo tanto, $\lambda \in \mathbb{C}$ es un autovalor de A cuando existe un vector $e \in \mathbb{C}^n$ tal que $Ae = \lambda e$. En ese caso, se dice que e es un autovector de A asociado al autovalor λ. Si $\lambda \in \mathbb{R}$, podemos encontrar un autovector asociado de componentes reales. En nuestro caso, sólo consideraremos autovalores y autovectores de matrices simétricas, lo cual facilitará las cosas en virtud del siguiente resultado.

Lema 9.1.

Todos los autovalores de una matriz simétrica son reales.

Demostración.
Sea $A \in \mathcal{M}_{n\times n}$ simétrica y supongamos que existen $a, b \in \mathbb{R}$, con $b \neq 0$ tales que $a + b\mathbf{i}$ es raíz del polinomio en $p(x) = |A - x\mathtt{Id}|$. En ese caso, también lo será $a - b\mathbf{i}$. Luego, la matriz

$$\begin{aligned} B &= [A - (a + b\mathbf{i})\mathtt{Id}][A - (a - b\mathbf{i})\mathtt{Id}] \\ &= (A - a\mathtt{Id})^2 + b^2\mathtt{Id} \end{aligned}$$

es singular. Sea pues $x \neq 0$ tal que $Bx = 0$. Luego, con mayor razón, $x'Bx = 0$. Al ser A simétrica se tiene que, si $y = (A - a\mathtt{Id})x$,

$$\begin{aligned} 0 &= x'Bx = x'(A - a\mathtt{Id})'(A - a\mathtt{Id})x + b^2x'x \\ &= y'y + b^2x'x. \end{aligned}$$

Siendo el primer sumando del último término no negativo y el segundo estrictamente positivo, se llega a una contradicción.

∎

En consecuencia, dado que sólo consideraremos autovalores de matrices reales simétricas, tanto éstos como las componentes de sus autovectores serán reales. El resultado siguiente, cuya demostración es trivial, precede al más importante de esta sección.

Lema 9.2.

Si $A \in \mathcal{M}_{n\times n}$ simétrica y $\Gamma \in \mathcal{M}_{n\times n}$ ortogonal, los autovalores de A coinciden con los de $\Gamma' A \Gamma$.

El siguiente resultado, denominado Teorema de Diagonalización, permite expresar de forma natural cualquier matriz simétrica. Para la demostración de la segunda parte del mismo se precisa del Teorema de los Multiplicadores Finitos de Lagrange,

que presentamos previamente. Éste se divide en dos partes: la primera establece condiciones necesarias que debe verificar un extremos relativo condicionado; la segunda establece condiciones suficientes.

Teorema 9.3.
Sean n y m números naturales tales que $n < m$ y $\mathcal{U} \subset \mathbb{R}^m$ abierto. Consideremos las aplicaciones $\phi : \mathcal{U} \longrightarrow \mathbb{R}$ y $f : \mathcal{U} \longrightarrow \mathbb{R}^n$, ambas con derivadas parciales segunda continuas. Sean $M = \{x \in \mathcal{U} \colon f(x) = 0\}$ y $c \in M$. Supongamos que el rango de la matriz $\left(\frac{\partial f_i}{\partial x_k}(c)\right)$ es n, y que existe un vector $\lambda \in \mathbb{R}^n$ tal que $\bigtriangledown(\phi - \lambda' f)(c) = 0$. Entonces, para que $\phi_{|M}$ tenga un máximo (mínimo) relativo en c, es condición suficiente que $D^2 L_\lambda(c)(h,h) < 0$ (respectivamente > 0) cada vez que $h \in \mathbb{R}^m \setminus \{0\}$ verifique que $Df_i(c)(h) = 0,\ i = 1, \ldots, n$, donde $L_\lambda = \phi - \lambda' f$.

Obsérvese la analogía que guarda con las condiciones necesaria y suficiente para máximos y mínimos no condicionados. La primera parte (necesariedad) se obtiene como aplicación del teorema de la función implícita, mientras que la segundo (suficiencia) se deduce del teorema de Taylor. Para más detalles, consultar Fdez. Viñas II, pag. 126. Dicho esto, vamos a enunciar el teorema fundamental al que hacía alusión anteriormente.

Teorema 9.4 (Diagonalización)**.**
Si $A \in \mathcal{M}_{n\times n}$ simétrica, existe una matriz $n \times n$ ortogonal Γ y una matriz $n \times n$ diagonal Δ=diag$(\delta_1, \ldots, \delta_n)$, con $\delta_1 \geq \ldots \geq \delta_n$, tales que

$$A = \Gamma \Delta \Gamma'.$$

En ese caso, los δ_i's son los autovalores de A y las columnas γ_i's de Γ constituyen una base ortonormal de autovectores asociados, siendo igualmente válida cualquier otra base ortonormal de autovectores asociados. Se verifica, además, que

$$\delta_1 = \sup_{\alpha \in \mathbb{R}^n \setminus \{0\}} \frac{\alpha' A \alpha}{\|\alpha\|^2},$$

alcanzándose en $\alpha = \gamma_1$, y que, para cada $i = 2, \ldots, n$,

$$\delta_i = \sup_{\alpha \in \langle \gamma_1, \ldots, \gamma_{i-1} \rangle^\perp} \frac{\alpha' A \alpha}{\|\alpha\|^2},$$

alcanzándose el máximo en $\alpha = \gamma_i$.

Demostración.
Sean $\delta_1, \ldots, \delta_n$ los autovalores (reales) ordenados de A y γ_1 un autovector asociado

a δ_1 tal que $\|\gamma_1\| = 1$. Podemos considerar $e_2, \ldots, e_n \in \mathbb{R}^n$ tales que $\{\gamma_1, e_2, \ldots, e_n\}$ constituyan una base ortonormal de $\mathbb{R}^n$. Sea entonces $S_1 \in \mathcal{M}_{n\times n}$ cuyas columnas son los vectores de la base por el mismo orden. Si se denota $B_1 = (e_2 \ldots e_n) \in \mathcal{M}_{n\times(n-1)}$, se verifica, teniendo en cuenta que $S_1'AS_1$ es simétrica,

$$S_1'AS_1 = \left(\frac{\gamma_1'}{B_1'}\right) A(\gamma_1|B_1) = \left(\frac{\gamma_1'}{B_1'}\right) (\delta_1\gamma_1|AB_1) = \left(\begin{array}{c|c} \delta_1 & 0 \\ \hline 0 & B_1'AB_1 \end{array}\right).$$

Sea $A_1 = B_1'AB_1 \in \mathcal{M}_{(n-1)\times(n-1)}$ simétrica. Por el lema anterior, los autovalores de $S_1'AS_1$ coinciden con los de A. Luego, los autovalores de A_1 son $\delta_2, \ldots, \delta_n$. El proceso se repite análogamente con A_1, considerándose una descomposición de la forma

$$S_2'A_1S_2 = \left(\begin{array}{c|c} \delta_2 & 0 \\ \hline 0 & A_2 \end{array}\right),$$

siendo $S_2 \in \mathcal{M}_{(n-1)\times(n-1)}$ ortogonal, y así hasta agotar los n autovalores, tras lo cual, habremos obtenido una serie de matrices cuadradas ortogonales $S_1, \ldots, S_n$, donde cada S_i es de orden $n \times (n+1-i)$, tales que, si se define, $\Gamma_1 = S_1$ y, para cada $i = 2, \ldots, n$,

$$\Gamma_i = \left(\begin{array}{c|c} \mathtt{Id}_{i-1} & 0 \\ \hline 0 & S_i \end{array}\right) \in \mathcal{M}_{n\times n},$$

entonces

$$\Gamma_n' \cdot \ldots \cdot \Gamma_1'A\Gamma_1 \cdot \ldots \cdot \Gamma_n = \begin{pmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_n \end{pmatrix}.$$

Considerando $\Gamma = \Gamma_n' \cdot \ldots \cdot \Gamma_1'$, se tiene que $A = \Gamma D\Gamma'$, lo cual implica, además, que $A\Gamma = \Gamma D$, de lo que se deduce que las columnas de Γ constituyen una base ortonormal de autovectores asociados a los autovalores $\delta_1, \ldots, \delta_n$, respectivamente. Si Γ_* es otra base ortonormal de autovectores asociados, se verifica trivialmente que $\Gamma_*'A\Gamma_* = D$.

Veamos que

$$\delta_1 = \sup_{\alpha\in\mathbb{R}^p\setminus\{0\}} \frac{\alpha'A\alpha}{\|\alpha\|^2},$$

que coincide, trivialmente, com

$$\text{máx}\{\alpha'A\alpha\colon \alpha \in \mathbb{R}^n \ \wedge \ \|\alpha\| = 1\}.$$

Consideramos las funciones $\phi(\alpha) = \alpha'A\alpha$ y $f(\alpha) = \alpha'\alpha - 1$, y el conjunto $M = \{\alpha \in \mathbb{R}^n\colon f(\alpha) = 0\}$, que es compacto, por cual ϕ alcanza máximo relativo a M

en cierto elemento γ. Luego, por el teorema 9.3, existe un único $\delta \in \mathbb{R}$ tal que $\bigtriangledown(\phi - \delta' f)(\gamma) = 0$, es decir, $2(A\gamma - \delta\gamma) = 0$ y, por tanto, $A\gamma = \delta\gamma$. Por lo tanto, γ es un autovector asociado al autovalor δ. Realmente, si $x \in \mathbb{R}^n$ es un autovector de norma 1 asociado a un autovalor β, entonces $x'Ax = \beta$. Como la anterior función se maximiza en δ, se tiene que $\delta = \lambda_1$ y $\gamma = \gamma_1$. El siguiente paso es encontrar máx$\{\alpha' A\alpha \colon \|\alpha\| = 1 \ \wedge \ \alpha'\gamma_1 = 0\}$. Se trata pues de maximizar la función ϕ anterior pero restringida al compacto donde se anula la función

$$f(\alpha) = \begin{pmatrix} \alpha'\alpha - 1 \\ \alpha'\gamma \end{pmatrix}.$$

Aplicando el teorema 9.3 se deduce la existencia de $\delta, \theta \in \mathbb{R}$ tales que, si el máximo se alcanza en $\gamma \in \mathbb{R}^n$,

$$2A\gamma - 2\delta\gamma - \theta\gamma_1 = 0.$$

Por lo tanto, multiplicando por γ_1 se tiene que

$$2\gamma_1' A\gamma - \theta = 0.$$

Dado que $\gamma \in \langle\gamma_1\rangle^{\perp} = \langle\gamma_2, \ldots, \gamma_n\rangle$, y teniendo en cuenta que $\Gamma' A \Gamma = D$, se deduce que el primer sumando es nulo. Luego, $\theta = 0$ y estamos en definitiva en las mismas condiciones del primer paso. Por lo tanto, $\delta = \delta_2$ y $\gamma = \gamma_2$. El proceso se repite análogamente hasta completar los n autovalores.

∎

Obsérvese que, si los autovalores de la matriz son distintos, la descomposición es única salvo reflexiones de los autovectores. En caso contrario, será única salvo reflexiones y rotaciones de éstos. El siguiente corolario es inmediato:

Corolario 9.5. (i) Dos autovectores asociados a distintos autovalores de una matriz simétrica son ortogonales.

(ii) Si A es simétrica, su rango coincide con el número de autovalores no nulos.

(iii) Si $A \geq 0$, sus autovalores son todos no negativos. Si $A > 0$, son todos estrictamente positivos.

(iv) Si $A \geq 0$, existe[2] una matriz simétrica $A^{1/2}$ tal que $A = A^{1/2}A^{1/2}$. Si $A > 0$, existe también una matriz simétrica $A^{-1/2}$ tal que $A^{-1} = A^{-1/2}A^{-1/2}$.

(v) Si $A \geq 0$, existe una matriz X con las mismas dimensiones tal que $A = X'X$.

[2]En Arnold(1981) se prueba además la unicidad.

(vi) Dada $A \in \mathcal{M}_{n\times n}$ semidefinida positiva de rango r, existe $X \in \mathcal{M}_{n\times r}$ de rango r tal que $A = XX'$.

(vii) La traza de una matriz simétrica es la suma de sus autovalores y el determinante, el producto de los mismos.

El siguiente resultado, corolario del teorema 9.4, permite obtener un especie de diagonalización para cualquier matriz, sea o no simétrica.

Teorema 9.6.
Dadas $A \in \mathcal{M}_{m\times p}$ de rango r, existen una matriz $D = \mathsf{diag}(\lambda_1, \ldots, \lambda_r)$ con elementos positivos y ordenados de mayor a menor, y otras dos matrices $N \in \mathcal{O}_m$ y $M \in \mathcal{O}_p$ verificando

$$A = N\left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right) M'. \tag{9.5}$$

Demostración.
Sea $\Delta = \mathsf{diag}(d_1, \ldots, d_r, 0)$ la matriz diagonal de orden p de los autovalores ordenados de $A'A$ y H una matriz $p \times p$ cuyas columnas $h_1, \ldots, h_p$ constituyen una base ortonormal de autovectores respectivos. El teorema de diagonalización permite afirmar afirma que

$$A'A = H\Delta H'.$$

Consideremos Δ_r y H_r las submatrices de Δ y H constituidas respectivamente por los r primeros autovalores y sus correspondientes autovectores. Definamos

$$G_r = AH_r\Delta_r^{-1/2}.$$

Se verifica entonces que $G_r'G_r = \mathtt{Id}_r$. Por lo tanto, sus columnas pueden completarse hasta obtener una matriz ortogonal de orden m que se denota por G. En ese caso, si se denota $D = \Delta_r^{1/2}$, se tiene que

$$G'AH = \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right),$$

de lo cual se sigue que

$$A = G\left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right) H'. \qquad \blacksquare$$

Exponemos a continuación un resultado relacionado con la matriz de covarianzas parciales, de gran utilidad cuando se estudie el problema de multicolinealidad.

Lema 9.7.
Consideremos una matriz cuadrada

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}.$$

(i) Si S_{22} es invertible, entonces $|S| = |S_{22}| \cdot |S_{11} - S_{12}S_{22}^{-1}S_{21}|$.

(ii) Si $S > 0$, entonces $S_{22} > 0$. Además, si la inversa de S es

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix},$$

se verifica que $V_{11}^{-1} = S_{11} - S_{12}S_{22}^{-1}S_{21}$.

Demostración.
Denótese $S_{11\cdot 2} = S_{11} - S_{12}S_{22}^{-1}S_{21}$. Si S_{22} es invertible, se verifica que $S = TUT^*$, donde

$$T = \begin{pmatrix} \mathtt{Id} & S_{12}S_{22}^{-1} \\ 0 & \mathtt{Id} \end{pmatrix}, \quad U = \begin{pmatrix} S_{11\cdot 2} & 0 \\ 0 & S_{22} \end{pmatrix}, \quad T^* = \begin{pmatrix} \mathtt{Id} & 0 \\ S_{22}^{-1}S_{21} & \mathtt{Id} \end{pmatrix}.$$

Se tiene también que

$$U = \begin{pmatrix} S_{11\cdot 2} & 0 \\ 0 & \mathtt{Id} \end{pmatrix} \begin{pmatrix} \mathtt{Id} & 0 \\ 0 & S_{22} \end{pmatrix}.$$

En consecuencia,

$$|S| = |T| \cdot |U| \cdot |T^*| = |S_{22}| \cdot |S_{11\cdot 2}|,$$

con lo cual queda probado (i). Demostremos (ii): si $S > 0$, podemos expresarla de la forma $S = X'X$, teniendo X las mismas dimensiones que S. Descompongamos X por columnas en $(X_1|X_2)$. En ese caso, $S_{22} = X_2'X_2$. Además,

$$\begin{pmatrix} S_{12} \\ S_{22} \end{pmatrix} = X'X_2.$$

Se tiene entonces que

$$\mathtt{rg}\begin{pmatrix} S_{12} \\ S_{22} \end{pmatrix} \leq \mathtt{rg}(X_2) = \mathtt{rg}(S_{22}).$$

Luego, S_{22} es no singular. Se sigue entonces de la primera parte que también es invertible $S_{11\cdot 2}$. Puede comprobarse fácilmente que

$$T^{-1} = \begin{pmatrix} \mathtt{Id} & -S_{12}S_{22}^{-1} \\ 0 & \mathtt{Id} \end{pmatrix}, \quad U^{-1} = \begin{pmatrix} S_{11\cdot 2}^{-1} & 0 \\ 0 & S_{22}^{-1} \end{pmatrix}, \quad (T^*)^{-1} = \begin{pmatrix} \mathtt{Id} & 0 \\ -S_{22}^{-1}S_{21} & \mathtt{Id} \end{pmatrix}.$$

Por lo tanto, $V = S^{-1} = (T^*)^{-1}U^{-1}T^{-1}$. Operando se obtiene que $V_{11} = S_{11\cdot 2}^{-1}$. ∎

Miscelánea

A continuación expondremos una serie de resultados de diversa índole que servirán de herramienta en las teorías de Modelos Lineales y Análisis Multivariante. El siguiente lema, de carácter especialmente técnico, será de utilidad cuando abordemos el análisis de los residuos.

Lema 9.8.

Dados $A \in \mathcal{M}_{n\times n}$ definida positiva y $b \in \mathbb{R}^n$ tales que $b'A^{-1}b \neq 1$, se tiene que

$$(A - bb')^{-1} = A^{-1} + \left(1 - b'A^{-1}b\right)^{-1}\left(A^{-1}b\right)\left(b'A^{-1}\right).$$

Demostración.
Basta multiplicar la matriz $A - bb'$ por el término de la derecha y tener en cuenta que $b'A^{-1}b$ es número real y que, por lo tanto, $b(b'A^{-1}b)b'A^{-1}$ equivale a $(b'A^{-1}b)bb'A^{-1}$. ∎

A continuación dos resultados de interés en Análisis Multivariante:

Teorema 9.9.

Sean S y U matrices $p \times p$ simétricas, definida positiva y semidefinida positiva, respectivamente, y sea el polinomio en t $p(t) = |U - tS|$. Entonces, $p(t)$ tiene todas sus raíces reales y no negativas, $t_1 \geq \ldots \geq t_p$, verificándose que

$$t_1 = \max_{x\in\mathbb{R}^p\setminus\{0\}} \frac{x'Ux}{x'Sx}.$$

Además, existe una matriz $A \in \mathcal{M}_{p\times p}$ tal que

$$ASA' = \mathtt{Id}_p, \quad AUA' = \begin{pmatrix} t_1 & & 0 \\ & \ddots & \\ 0 & & t_p \end{pmatrix}.$$

Demostración.
Siendo $S > 0$, se verifica

$$|U - tS| = |S^{1/2}||S^{-1/2}US^{-1/2} - t\mathtt{Id}||S^{1/2}| = |S^{1/2}|^2||S^{-1/2}US^{-1/2} - t\mathtt{Id}|. \quad (9.6)$$

Dado que $S^{-1/2}US^{-1/2} \geq 0$, existen una matriz $p \times p$ ortogonal Γ y una matriz diagonal $D = \mathrm{diag}(t_1, \ldots, t_p)$ tales que

$$S^{-1/2}US^{-1/2} = \Gamma D\Gamma',$$

siendo $t_1, \ldots, t_p$ los autovalores ordenados de $S^{-1/2}US^{-1/2}$, que, por (9.6), coinciden con las raíces ordenadas $p(t)$. Además, serán todos no negativos, y si $U > 0$, serán estrictamente positivos. Por lo tanto

$$\Gamma' S^{-1/2} U S^{-1/2} \Gamma = D, \qquad \Gamma' S^{-1/2} S (\Gamma' S^{-1/2})' = \mathtt{Id}.$$

Luego, el teorema se satisface con $A = \Gamma' S^{-1/2}$. Además, en virtud del teorema 9.4 y considerando el cambio de variables $Z = S^{1/2}X$, se sigue que

$$t_1 = \max_{z \in \mathbb{R}^p \setminus \{0\}} \frac{z' S^{-1/2} U S^{-1/2} z}{\|z\|^2} = \max_{x \in \mathbb{R}^p \setminus \{0\}} \frac{x' U x}{x' S x},$$

lo cual acaba la demostración.

∎

Teorema 9.10.
Para toda $S \in \mathcal{M}_{p\times p}$ semidefinida positiva existe una matriz $C \in \mathcal{M}_{p\times p}$ triangular superior tal que $S = CC'$.

Demostración.
Sabemos que existe $B \in \mathcal{M}_{p\times p}$ tal que $S = BB'$. Entonces, para cada $\Gamma \in \mathcal{M}_{p\times p}$ ortogonal se tiene que $S = (B\Gamma)(B\Gamma)'$. Luego, basta probar que, para cada $B \in \mathcal{M}_{p\times p}$, existe Γ ortogonal tal que $B\Gamma$ es triangular superior. Si $b_1, \ldots, b_p \in \mathcal{M}_{1\times p}$ son las filas de B, construiremos Γ de tal manera que sus columnas, $\gamma_1, \ldots, \gamma_p \in \mathbb{R}^p$ sean de norma 1 y satisfagan

$$\gamma_1 \in \langle b_2', \ldots, b_p' \rangle^{\perp}, \qquad \gamma_i \in \langle \gamma_1', \ldots, \gamma_{i-1}', b_{i+1}', \ldots, b_p' \rangle^{\perp}, \quad \forall i = 2, \ldots, p.$$

Puede comprobarse fácilmente que Γ es ortogonal y $B\Gamma$ es triangular superior.

∎

El teorema siguiente se utiliza, por ejemplo, en la segunda reducción por invarianza para obtener el test F.

Lema 9.11.
Sean $X \in \mathcal{M}_{p\times k}$ de rango r y $U \in \mathcal{M}_{r\times k}$ de rango r tales que $X'X = U'U$. Entonces, existe una matriz $\Gamma \in \mathcal{M}_{p\times p}$ ortogonal tal que

$$\Gamma X = \begin{pmatrix} U \\ 0 \end{pmatrix}.$$

Demostración.
Consideremos el subespacio $V \subset \mathbb{R}^p$ generado por los vectores columnas de de X y

sea $R \in \mathcal{M}_{p\times(p-r)}$ cuyas columnas constituyen una base ortonormal de $V^{\perp}$. Dado que $\mathtt{rg}(U) = \mathtt{rg}(U'U) = r$, $U'U$ es una matriz invertible. Consideremos entonces

$$\Gamma = \begin{pmatrix} (U'U)^{-1}UX' \\ R' \end{pmatrix} \in \mathcal{M}_{p\times p}.$$

Esta matriz es ortogonal, pues

$$\Gamma\Gamma' = \begin{pmatrix} (U'U)^{-1}UU'UU'(U'U)^{-1} & (U'U)^{-1}UX'R \\ R'XU'(U'U)^{-1} & R'R \end{pmatrix} = \mathtt{Id}.$$

Además,

$$\Gamma X = \begin{pmatrix} (U'U)^{-1}UU'U \\ R'X \end{pmatrix} = \begin{pmatrix} U \\ 0 \end{pmatrix},$$

como queríamos demostrar.

∎

Teorema 9.12.
Sean $X, Y \in \mathcal{M}_{p\times k}$. Se verifica entonces que $X'X = Y'Y$ si, y sólo si, existe una matriz $\Gamma \in \mathcal{M}_{p\times p}$ ortogonal tal que $Y = \Gamma X$.

Demostración.
Obviamente, si $Y = \Gamma X$, entonces $Y'Y = X'X$. Veamos la otra implicación. Si $r = \mathtt{rg}(X)$, entonces $\mathtt{rg}(Y) = \mathtt{rg}(Y'Y) = \mathtt{rg}(X'X) = \mathtt{rg}(X) = r$. En virtud del corolario 9.5(vi), existe una matriz U $r \times k$ de rango r tal que $U'U = X'X = Y'Y$. Aplicando el lema anterior a $X'X$ y a $Y'Y$, se deduce la existencia de sendas matrices $p \times p$ ortogonales, Γ_1 y Γ_2, tales que

$$\Gamma_1 X = \begin{pmatrix} U \\ 0 \end{pmatrix} = \Gamma_2 Y.$$

Basta pues considerar $\Gamma = \Gamma_2'\Gamma_1$ para obtener el resultado deseado.

∎

Nótese que, si $k = 1$, estamos afirmando que $\|X\| = \|Y\|$ si, y sólo si, existe una matriz $\Gamma \in \mathcal{M}_{p\times p}$ ortogonal tal que $Y = \Gamma X$. Por ello se identifican las matrices ortogonales con las rotaciones y la norma euclídea constituye un invariante maximal para el grupo de las rotaciones. El siguiente resultado será de utilidad para justifiquemos el test F en el modelo de Correlación.

Teorema 9.13.
Sean $X, Y \in \mathcal{M}_{p\times k}$ y $S, T \in \mathcal{M}_{p\times p}$ definidas positivas. Si $X'S^{-1}X = Y'T^{-1}Y$, existe una matriz $A \in \mathcal{M}_{p\times p}$ invertible tal que $Y = AX$ y $T = ASA'$.

Demostración.
Aplicando el teorema anterior a $S^{-1/2}X$ y $T^{-1/2}Y$, se deduce la existencia de una matriz $\Gamma \in \mathcal{M}_{p\times p}$ ortogonal tal que $T^{-1/2}Y = \Gamma S^{-1/2}X$, es decir,

$$Y = (T^{1/2}\Gamma S^{-1/2})X.$$

Además,

$$(T^{1/2}\Gamma S^{-1/2})S(T^{1/2}\Gamma S^{-1/2})' = T.$$

Luego, considerando $A = T^{1/2}\Gamma S^{-1/2}$ obtenemos el resultado deseado.

∎

El siguiente teorema es de utilidad a la hora de encontrar el estimador de máxima verosimilitud en el modelo de correlación. Necesita un lema previo.

Lema 9.14.

Sea h una aplicación que asigna a cada matriz $U \in \mathcal{M}_{p\times p}$ definida positiva el número

$$h(U) = \frac{1}{|U|^{n/2}} \exp\left\{-\frac{1}{2}\mathtt{tr}(U^{-1})\right\}.$$

Entonces h alcanza el máximo en $U = \frac{1}{n}\mathtt{Id}$.

Demostración.
Si $t_1 \geq \ldots \geq t_p > 0$ denotan los autovalores ordenados de U^{-1}, $h(U)$ puede expresarse como

$$h(U) = \left(\prod_{i=1}^{p} t_i\right)^{n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{p} t_i\right\}.$$

Por lo tanto, h puede considerarse como una función $g(t_1, \ldots, t_p)$. Veamos que g alcanza un máximo en $(R^+)^p$. Dado que $g(t_1, \ldots, t_p) \to \infty$ cuando cada $t_i \to \infty$, $i = 1, \ldots, p$, podemos restringir la búsqueda del máximo a una región del tipo $(0, M]^p$. Teniendo en cuenta que g se anula cuando cualquier t_i vale 0, ello equivale a buscar el máximo en el compacto $[0, M]^p$. Siendo g continua, este máximo se alcanza con certeza en cierto punto, en el cual deben anularse las las derivadas parciales

$$\frac{\partial}{\partial t_i} g(t_1, \ldots, t_p) = \left(\frac{n}{2t_i} - \frac{1}{2}\right) g(t_1, \ldots, t_p), \quad i = 1, \ldots, p.$$

Dado que g no se anula en $(R^+)^p$, se tiene que

$$\frac{\partial}{\partial t_i} g(t_1, \ldots, t_p) = 0, \ \forall i = 1, \ldots, p \quad \Leftrightarrow \quad t_1 = \ldots = t_n = n.$$

Por lo tanto, el máximo se alcanza cuando todos los autovalores de U son iguales a $1/n$. Luego, por el teorema 9.4, se sigue que $U = (1/n)\mathtt{Id}$.

∎

Teorema 9.15.
Sean A una matriz $p \times p$ definida positiva y f la función que asigna a cada matriz U del mismo tipo el número $f(U) = \frac{1}{|U|^{n/2}} \exp\left\{-\frac{1}{2}\mathtt{tr}(U^{-1}A)\right\}$. Entonces, dicha función alcanza el máximo en $U = \frac{1}{n}A$.

Demostración.
Se verifica que

$$\begin{aligned} f(U) &= \frac{1}{|A^{1/2}|^n |A^{-1/2}UA^{-1/2}|^{n/2}} \exp\left\{-\frac{1}{2}\mathtt{tr}\left(A^{-1/2}UA^{-1/2}\right)^{-1}\right\} \\ &= \frac{1}{|A^{1/2}|^n} h\left(A^{-1/2}UA^{-1/2}\right), \end{aligned}$$

donde h se define como en el lema anterior. Por lo tanto, f alcanza el máximo cuando

$$A^{-1/2}UA^{-1/2} = \frac{1}{n}\mathtt{Id}$$

o, equivalentemente, cuando $U = \frac{1}{n}A$.

∎

Proyección Ortogonal

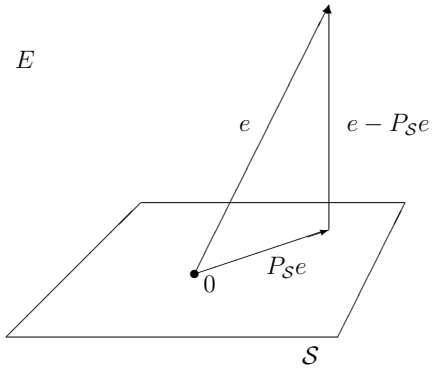
Se trata de un concepto definible no sólo en $\mathbb{R}^n$, sino en cualquier espacio de Hilbert. Dado E un espacio $\mathbb{R}$-vectorial[3], un producto interior sobre E es una aplicación $\langle\cdot,\cdot\rangle$ de $E \times E$ en $\mathbb{R}$ simétrica y tal que, para todo $e \in \mathbb{E}$, las aplicaciones $\langle e,\cdot\rangle$ y $\langle\cdot, e\rangle$ son lineales. En ese caso, se dice que e_1 y e_2 son perpendiculares u ortogonales cuando $\langle e_1, e_2\rangle = 0$, en cuyo caso se denota $e_1 \perp e_2$. Dado un subconjunto $\mathcal{S} \subset E$, se denota $e_1 \perp \mathcal{S}$ cuando $e_1 \perp e_2$ para todo $e_2 \in \mathcal{S}$. También se denota por $\mathcal{S}^\perp$ la familia de los vectores ortogonales a todos los de $\mathcal{S}$. Todo producto interior induce de forma natural una norma definida mediante $\|e\| = \langle e, e\rangle^{1/2}$, la cual induce a su vez una distancia $d(e_1, e_2) = \|e_1 - e_2\|$. Por último, dicha distancia induce una topología sobre E. Si el espacio topológico resultante es completo se dice de Hilbert. Como ejemplo tenemos el espacio $\mathbb{R}^n$ dotado del producto interior

$$\langle x, y\rangle = x'y = \sum_{i=1}^{n} x_i y_i, \tag{9.7}$$

[3]Podríamos considerar son problemas espacios $\mathbb{C}$-vectoriales.

denominado comúnmente producto escalar y del cual proviene la noción de ortogonalidad definida en (9.1) junto con la norma y distancia euclídeas definidas en (9.2) y (9.3), respectivamente. No será $\mathbb{R}^n$ el único caso espacio que manejemos. También se considerará mas adelante el espacio L^2 de las variables aleatorias de cuadrado integrable sobre un cierto espacio de probabilidad.

Dado un subespacio lineal cerrado $\mathcal{S} \subset E$, se define la proyección ortogonal sobre $\mathcal{S}$ como la aplicación $P_{\mathcal{S}}$ que asigna a cada vector $e \in E$ el único vector $s \in \mathcal{S}$ tal que $e - s \in \mathcal{S}^{\perp}$. Puede probarse[4] que se trata del vector de $\mathcal{S}$ más próximo a e según la distancia inducida por el producto interior. Dicha aplicación es lineal y sobreyectiva.



En el caso de un subespacio lineal $V \subset \mathbb{R}^n$ (dotado del producto escalar) de dimensión k, la aplicación P_V se identificará con una matriz $n \times n$ de rango k, que se denotará igualmente por P_V. Se verifica además, como probaremos a continuación, que dada $X \in \mathcal{M}_{n\times k}$ una base de V,

$$P_V = X(X'X)^{-1}X'. \tag{9.8}$$

La anterior expresión tiene sentido, pues $\mathtt{rg}(X) = \mathtt{rg}(X'X) = k$, es decir, $X'X$ es invertible. Así pues, dado $u \in \mathbb{R}^n$, se tiene que $X(X'X)^{-1}X'u \in V$. Además, dado cualquier $y \in \mathbb{R}^k$, se tiene que

$$\big\langle u - X(X'X)^{-1}X'u, Xy\big\rangle = u'Xy - u'X(X'X)^{-1}X'Xy = 0,$$

es decir, que $u - X(X'X)^{-1}X'u \in V^{\perp}$. Además, $X(X'X)^{-1}X'u$ es el único vector de V que lo verifica pues, si existiesen dos vectores $v_1, v_2 \in V$ tales que $u - v_1, u - v_2 \in V^{\perp}$, entonces se tendría que $v_1 - v_2 \in V \cap V^{\perp} = 0$. Además, dado que

$$\mathtt{rg}\left(X(X'X)^{-1}X'\right) = \mathtt{rg}(X) = k,$$

[4] Rudin (1979).

la aplicación es sobreyectiva. Por lo tanto, la proyección ortogonal está bien definida y es, efectivamente, una aplicación lineal sobreyectiva cuya matriz es (9.8). Nótese que, si X es una base ortonormal de V, entonces $P_V = XX'$.

La matriz P_V es simétrica e idempotente, es decir, verifica que $P_V^2 = P_V$. Puede demostrarse, recíprocamente (ver, por ejemplo, Arnold (1981)), que toda matriz $n \times n$ simétrica e idempotente de rango k es la matriz de la proyección ortogonal sobre el subespacio k-dimensional de $\mathbb{R}^n$ generado por sus vectores columna. Veamos algunas propiedades elementales de la proyección ortogonal en $\mathbb{R}^n$.

Proposición 9.16.
Sean $V, W \subset \mathbb{R}^n$, con $W \subset V$. Se verifica:

(i) $P_V = P_{V|W} + P_W$.

(ii) Para todo $y \in \mathbb{R}^n$, $\|P_V y\|^2 = \|P_W y\|^2 + \|P_{V|W} y\|^2$. En particular, $\|y\|^2 = \|P_V y\|^2 + \|P_{V^\perp} y\|^2$.

(iii) $P_V y = y$ sii $y \in V$.

(iv) $P_W \cdot P_V = P_W$.

(v) $\mathtt{tr} P_V = \mathtt{dim} V$.

(vi) $P_{V^\perp} = \mathtt{Id} - P_V$.

Obviamente, todas estas propiedades excepto (v) pueden extenderse a cualquier espacio de Hilbert. Asimismo, el concepto de proyección ortogonal posee pleno sentido cuando en lugar de subespacios lineales consideramos subvariedades afines. Así, puede demostrarse fácilmente que, dados un subsepacio lineal $V \subset \mathbb{R}^n$ y un vector $\mathtt{x} \in \mathbb{R}^n \backslash \{0\}$,

$$P_{x+V} u = x + P_V (u - x).$$

Hemos de tener en cuenta que, para cada $v \in V$, se verifica

$$P_{x+V} = P_{(x+v)+V}. \tag{9.9}$$

Por último, el producto interior definido en $\mathbb{R}^n$ puede extenderse a las matrices cuadradas de orden n como sigue. Dadas dos matrices $A, B \in \mathcal{M}_{n \times p}$, con componentes a_{ij} y b_{ij}, respectivamente, donde $i = 1, \ldots, n$ y $j = 1, \ldots, p,$, se verifica

$$\mathtt{tr}(A'B) = \sum_{i=1}^{n} \sum_{j=1}^{p} a_{ij} b_{ij},$$

es decir, $\mathtt{tr}(A'B)$ se entiende como el producto interior de los vectores np-dimensionale que se obtienen al leer las matrices de cualquier forma (pero de igual manera en ambas). En ese sentido, podemos afirma que la $\mathtt{tr}(A'B)$ generaliza el producto interior de dos vectores, de ahí que definamos

$$\langle A, B\rangle := \mathtt{tr}(A'B), \quad A, B \in \mathcal{M}_{n\times p}.$$

Por último, dadas A, B, C matrices cuadradas de orden n, se verifica que $\mathtt{tr}(ABC) = \mathtt{tr}(CAB) = \mathtt{tr}(BAC)$.

Proposición 9.17.
Dadas A, B y C, se verifica, siempre y cuando tengan sentido los productos, que

$$\mathtt{tr}(A'B) = \mathtt{tr}(B'A) = \mathtt{tr}(AB') = \mathtt{tr}(BA'),$$

$$\mathtt{tr}(ABC) = \mathtt{tr}(BCA) = \mathtt{tr}(CAB).$$

9.2. Generalidades sobre Probabilidad

En esta sección y en la siguiente presentamos una miscelánea de definiciones y resultados fundamentales que serán necesarios en nuestra teoría. Dado que la probabilidad se entiende formalmente como una medida de extensión 1, haremos uso de diversos conceptos y resultados de la Teoría de la Medida, que daremos por conocidos[5].

Definiciones básicas

Para empezar, un espacio medible es un par $(\Omega, \mathcal{A})$, donde Ω denota un conjunto no vacío y $\mathcal{A}$ una σ-álgebra de $\mathcal{P}(\Omega)$. Lo más común es que Ω sea un subconjunto de interior no vacío de $\mathbb{R}^n$, para algún $n \geq 1$, o una colección numerable de elementos, por ejemplo $\mathbb{N}$. En el primer caso, se considera normalmente la σ-álgebra de Borel, que es la generada por los conjuntos abiertos y se denota por $\mathcal{R}^n$ [6]; en el segundo, se considera $\mathcal{P}(\Omega)$.

Una probabilidad P sobre $(\Omega, \mathcal{A})$ es una medida positiva de extensión 1 sobre dicho espacio. La terna $(\Omega, \mathcal{A}, P)$ constituye un espacio de probabilidad. Una variable aleatoria será una función X medible de $(\Omega, \mathcal{A})$ en otro espacio $(\Omega_X, \mathcal{A}_X)$. Se dice

[5] Pueden consultarse, por ejemplo, en Ash (1972(, Billinsley (1986) o Nogales (1998).

[6] Coincide con el producto cartesiano n veces consigo misma de la σ-álgebra de Borel en $\mathbb{R}$, que se denota por $\mathcal{R}$.

real cuando el espacio de llegada es $\mathbb{R}$ (se entiende que $\mathbb{R}$ está provisto de σ-álgebra de Borel). En todo caso, X induce en el espacio de llegada una nueva probabilidad P^X, definida mediante $P^X(B) = P(X^{-1}(B))$, para todo $B \in \mathcal{A}_X$. Si X es real, la expresión $\mathtt{E}_P[X]$, denominada esperanza de X, hará referencia a la integral de X respecto de P, siempre y cuando exista. Esta definición puede hacerse extensiva a variables aleatorias con valores en $\mathbb{C}$, suponiendo $\mathbb{C}$ dotado de la σ-álgebra de Borel $\mathcal{R}^2$. Dado $k \in \mathbb{N}$, el momento de orden k de X se definirá como $\mathtt{E}_P[X^k]$, siempre y cuando exista. Se define la función característica de una variable aleatoria real X mediante

$$\varphi_X(t) = \mathtt{E}_P[\exp\{\mathbf{i}tX\}], \quad t \in \mathbb{R}.$$

Esta función, bien definida sobre toda la recta real y con valores complejos, viene a caracterizar, en virtud del Teorema de Inversión de Levy, a la probabilidad P^X. De manera análoga se define la función generatriz de momentos

$$g_X(t) = \mathtt{E}_P[\exp\{tX\}], \quad t \in \mathbb{R}.$$

Cuando esta función está bien definida en un entorno de 0, queda garantizada la existencia de todos los momentos de P^X, que se obtienen a partir de g_X mediante

$$\mathtt{E}_P[X^k] = g_X^{(k)}(0).$$

La función de distribución de X se define mediante

$$F_X(t) = P(X \leq t), \quad t \in \mathbb{R}.$$

Esta función es no decreciente, continua por la derecha y tal que $\lim_{t\to-\infty} F(t) = 0$ y $\lim_{t\to+\infty} F(t) = 1$. Al igual que la función característica, determina de manera unívoca la probabilidad P^X. Dado $\alpha \in (0,1)$, se denota por $[P^X]^\alpha$ al cualquier número real tal que $F_X\left([P^X]^\alpha\right) = 1-\alpha$, si es que existe. Si F_X es continua, $[P^X]^\alpha$ existirá y será único para cualquier valor de α. En general, las propiedades fundamentales de las tres funciones que hemos definido pueden encontrarse, por ejemplo, en Billingsley (1986).

Un n-vector aleatorio real es una función medible Y de $(\Omega, \mathcal{A}, P)$ en $\mathbb{R}^n$, que induce pues, de manera natural, una nueva probabilidad sobre $(\mathbb{R}^n, \mathcal{R}^n)$ denominada distribución de Y respecto a P y se denota por P^Y. Las funciones característica y generatriz pueden definirse entonces mediante

$$\varphi_Y(t) = \mathtt{E}_P[\exp\{\mathbf{i}\langle t, Y\rangle\}], \quad g_Y(t) = \mathtt{E}_P[\exp\{\langle t, Y\rangle\}], \quad t \in \mathbb{R}^n.$$

Las propiedades de las funciones característica e inversa se traducen de manera natural del caso unidimensional al multidimensional. Se dice que P^Y está dominada

por una medida σ-finita μ sobre $\mathbb{R}^n$ cuando todo suceso μ-nulo es P^Y-nulo. En tal caso, el teorema de Radom-Nykodin[7] garantiza la existencia de una función medible $f : \mathbb{R}^n \to \mathbb{R}^+$ tal que

$$P(A) = \int_A f(x)\, d\mu, \quad A \in \mathcal{R}^n$$

Una función en tales condiciones se denomina función de densidad y caracteriza plenamente la distribución P^Y. En la mayor parte de las ocasiones será la medida de Lebesgue[8] la que actúe como dominante y la integral anterior será la de Lebesgue. En otros casos, Y tendrá como imagen un conjunto finito o numerable, con lo cual la medida cardinal sobre dicho conjunto ejercerá como dominante y la función de densidad será la función indicador del mismo.

Por otra parte, se denota por $Y_1, \ldots, Y_n$ las componentes de Y, que son variables aleatorias reales. Así, para cada $i = 1, \ldots, n$, definimos como media de Y_i al parámetro $\mathtt{E}_P[Y_i]$, siempre y cuando exista. La media suele denotarse mediante la letra μ, seguida en este caso del correspondiente subíndice. Además, en la notación $\mathtt{E}_P$ suele eliminarse el subíndice P siempre y cuando no haya lugar a confusión. Igualmente, si Y_i posee momento de segundo orden finito, podemos definir el parámetro $\mathtt{var}[Y_i] = \mathtt{E}[(Y_i - \mu_i)^2]$, denominado varianza, que será positivo y finito. Suele denotarse mediante σ^2 seguida del correspondiente subíndice. Por otra parte, dado i y j entre 1 y n, si Y_i e Y_j poseen momentos de segundo orden finitos podemos definir la covarianza entre ambas mediante

$$\mathtt{cov}[Y_i, Y_j] = \mathtt{E}[(Y_i - \mu_i)(Y_j - \mu_j)].$$

Se denotará también mediante σ_{ij}. Obviamente, se tiene que $\sigma_{ii} = \sigma_i^2$. Además, se sigue de la desigualdad de Holder[9] que

$$-\sigma_i\sigma_j \leq \sigma_{ij} \leq \sigma_i\sigma_j,$$

lo cual invita a considerar el parámetro

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i\sigma_j} \in [-1, 1], \tag{9.10}$$

denominado coeficiente de correlación lineal simple. Las medias μ_i, $i = 1, \ldots, n$ componen un vector media que se denota por $\mathtt{E}[Y]$ [10] o, frecuentemente, por μ. Las

[7] Ver Billingsley (1986)
[8] Ver Billingsley (1986)
[9] Rudin (1979).
[10] Estamos entendiendo pues que la esperanza de un vector aleatorio es el vector formado por las esperanzas de sus componentes.

varianzas y covarianzas componen a su vez una matriz que se denota por $\mathtt{Cov}[Y]$ o, frecuentemente, mediante la letra Σ, y que puede definirse matricialmente mediante

$$\mathtt{Cov}[Y] = \mathtt{E}[(Y-\mu)(Y-\mu)'].$$

Esta matriz simétrica es semidefinida positiva. La suma de los elementos de su diagonal se denomina varianza total. De igual forma podemos hablar de una matriz de correlaciones que se define mediante $P = D_\Sigma^{-1}\Sigma D_\Sigma^{-1}$, siendo D_Σ la matriz diagonal constituida por las varianzas. Dados $A \in \mathcal{M}_{m\times n}$ y $b \in \mathbb{R}^n$, podemos considerar la transformación afín $AY + b$, de $(\Omega, \mathcal{A}, P)$ en $\mathbb{R}^m$. Puede comprobarse fácilmente que

$$\mathtt{E}[AY+b] = A\mathtt{E}[Y] + b, \qquad \mathtt{Cov}[AY+b] = A\mathtt{Cov}[Y]A'. \tag{9.11}$$

Dados dos vectores aleatorios Y_1 e Y_2 de $(\Omega, \mathcal{A}, P)$ en $\mathbb{R}^{n_1}$ y $\mathbb{R}^{n_2}$, respectivamente, decimos que son independientes cuando, para cada par de sucesos B_1 de $\mathcal{R}^{n_1}$ y B_2 de$\mathcal{R}^{n_2}$, se verifica que

$$P(Y_1 \in B_1,\ Y_2 \in B_2) = P(Y_1 \in B_1)P(Y_2 \in B_2).$$

La definición anterior sigue puede extenderse sin problemas al caso de k vectores aleatorios. Lo mismo ocurre con la que sigue: dados dos probabilidades P_1 y P_2 definidas sobre $(\Omega_1, \mathcal{A}_1)$ y $(\Omega_2, \mathcal{A}_2)$, respectivamente, se denota por $P_1 \times P_2$ la única probabilidad sobre el espacio producto $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \times \mathcal{A}_2)$ tal que

$$[P_1 \times P_2](A_1 \times A_2) = P_1(A_1)P_2(A_2), \quad \forall B_1 \in \mathcal{A}_1,\ \forall A_2 \in \mathcal{A}_2.$$

La existencia y unicidad de dicha probabilidad, denominada probabilidad producto, se deriva del Teorema de de la medida producto[11]. Este producto puede extenderse al caso en el que una de las probabilidades sea de transición: decimos que $\mathcal{L}$, definida sobre $\mathcal{A}_1 \times \Omega_2$ y con valores en $[0,1]$, es una probabilidad de transición cuando, para cada $A_1 \in \mathcal{A}_1$, la función $\mathcal{L}(A_1, \cdot)$ es medible y, además, para cada $x_2 \in \Omega_2$, la función $\mathcal{L}(\cdot, x_2)$ es una probabilidad. En ese caso, existe una única probabilidad $\mathcal{L} \times P_2$ sobre el espacio producto, denominada producto generalizado, tal que

$$[\mathcal{L} \times P_2](A_1 \times A_2) = \int_{A_2} \mathcal{L}(A_1, \cdot)\, dP_2, \quad \forall A_1 \in \mathcal{A}_1,\ \forall B_2 \in \mathcal{A}_2.$$

[11] Ver, por ejemplo, Billingsley (1986). Ver también el teorema de Fubini y el de la medida producto generalizado. La extensión al producto finito de probabilidades es trivial. En el caso infinito, el producto puede construirse teniendo en cuenta el Teorema de Extensión de Kolmogorov (Ash (1972)).

Obviamente, que dos vectores aleatorios Y_1 e Y_2 definidos en $(\Omega, \mathcal{A}, P)$ sean independientes equivale a que la distribución *conjunta* $P^{(Y_1,Y_2)}$ sea el producto de las distribuciones *marginales* P^{Y_1} y P^{Y_2}.

Consideremos el vector conjunto $Y = (Y_1', Y_2')'$, de $(\Omega, \mathcal{A}, P)$ en $\mathbb{R}^{n_1+n_2}$. Si Y_1 e Y_2 poseen momentos de orden 2 finitos podemos hablar de la media y matriz de varianzas-covarianzas de Y, que descomponen de la siguiente forma

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \tag{9.12}$$

Se dice que Y_1 e Y_2 son incorrelados cuando $\Sigma_{12} = 0$. Es inmediato comprobar que la independencia implica incorrelación, aunque el recíproco no es cierto en general. No obstante, sí que lo es bajo la hipótesis de normalidad multivariante, según se estudia en el capítulo 1, dedicado al estudio de dicha distribución.

Cuando la matriz Σ es definida positiva también lo es, en virtud del del lema 9.7, la matriz Σ_{22}, de ahí que tenga sentido definir la matriz

$$\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{9.13}$$

denominada matriz de varianzas-covarianzas parciales de Y_1 dado Y_2. En el caso $n_1 = 1$, estaremos hablando de un número no negativo

$$\sigma^2_{11\cdot 2} = \sigma^2_1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \tag{9.14}$$

que denominaremos varianza parcial de Y_1 dado Y_2. En ese caso, se define también el siguiente parámetro

$$\rho^2_{12} = \frac{1}{\sigma^2_1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{9.15}$$

denominado coeficiente de correlación lineal múltiple (al cuadrado) de Y_1 respecto a Y_2. En el caso $n_2 = 1$ estaremos hablando del coeficiente de correlación lineal simple definido en (9.10). A continuación intentaremos ofrecer una interpretación geométrica de todos los parámetros definidos.

Interpretación geométrica de los parámetros

Es bastante habitual en Matemáticas en general, y en Probabilidad y Estadística en particular, cuantificar los errores evaluando los cuadrados de las diferencias. Esta forma de proceder, a todas luces razonable, fue propuesta por el propio Gauss a finales del siglo XVIII. Se conoce como técnica de mínimos cuadrado. El propio

Gauss demostró en 1829 un resultado conocido como Teorema de Gauss-Markov[12] que explica el éxito de esta técnica.

No obstante, nuestra intención aquí es aclarar que esta forma de proceder posee una sencilla justificación en un marco formal más general: el los espacios de Hilbert. El ejemplo más inmediato de espacio de Hilbert es el propio $\mathbb{R}^n$ dotado del producto escalar. Esta consideración será de utilidad a la hora de interpretar los parámetros muestrales (estadísticos). El otro espacio de Hilbert a tener en cuenta y el que nos atañe en esta sección es L^2. Dado un espacio de probabilidad $(\Omega, \mathcal{A}, P)$, se denota por $L^2(\Omega, \mathcal{A}, P)$ el conjunto de las variables aleatorias[13] reales de cuadrado integrable (es decir, de varianza finita). En dicho espacio podemos considerar el producto interior definido mediante

$$\langle f, g\rangle = \int_\Omega fg\, dP, \quad f, g \in L_2. \tag{9.16}$$

La desigualdad de Holder garantiza que dicha integral existe y es finita. El producto interior induce una noción de ortogonalidad y una norma sobre L^2 definida mediante

$$\|f\|_2 = \left(\int_\Omega f^2\, dP\right)^{1/2} \tag{9.17}$$

que induce, a su vez, una métrica en L_2 que se denotará por $\mathtt{d}_2$ y que hace completo el espacio. Si consideramos el espacio de los p-vectores aleatorios cuyas componentes poseen cuadrados integrables, podemos definir, para cada par $\mathtt{f} = (\mathtt{f}_i)_{i\leq p}$ y $\mathtt{g} = (\mathtt{g}_i)_{i\leq p}$, el producto interior

$$\langle \mathtt{f}, \mathtt{g}\rangle_p = \int \mathtt{f}'\mathtt{g}\, dP = \sum_{i=1}^{p} \langle \mathtt{f}_i, \mathtt{g}_i\rangle. \tag{9.18}$$

Este producto induce igualmente una norma y una métrica $\mathtt{d}_{2,p}$ en dicho espacio. Interpretaremos los parámetros probabilísticos considerados anteriormente a la luz de estas definiciones.

En primer lugar, es obvio que, si cualquiera de las variables aleatorias f o g posee media 0, la ortogonalidad equivale a la incorrelación. Además, la esperanza o media de cualquier función f en L_2 puede entenderse como la proyección ortogonal de f sobre el subespacio de las funciones constantes, que se denotará por $\langle 1\rangle$, siendo su varianza la distancia al cuadrado entre f y su proyección, que es mínima. Efectivamente, basta demostrar que $f - \mathtt{E}_P[f]$ es ortogonal a cualquier función constante, lo cual se sigue directamente de la propia definición de $\mathtt{E}_P[f]$. Por lo tanto, se verifica que

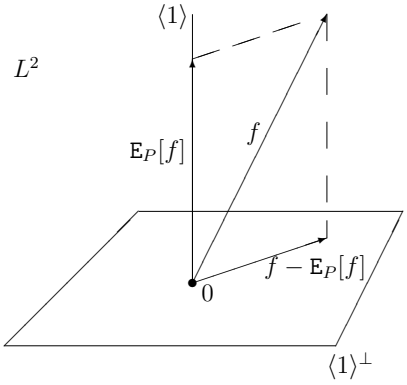
[12] Ver teorema teorema 3.3.

[13] Se identifican los vectores que difieren en un suceso de probabilidad nula.

$\|f - \mathtt{E}[f]\| < \|f - k\|$ para cualquier otra función constante k [14] Es decir, que la media puede entenderse como la variable constante más próxima (según la métrica anterior) a la nuestra. La diferencia existente entre la situación real (aleatoria) y la que correspondería a un fenómeno determinista (constante) queda recogida mediante la variable aleatoria

$$f - P_{\langle 1\rangle} f = P_{\langle 1\rangle^\perp} f = f - \mathtt{E}_P[f],$$

que podemos denominar variabilidad total. La varianza es el tamaño al cuadrado (norma al cuadrado) de la variabilidad total y pretende pues cuantificar dicha diferencia.



En el caso multivariante, se denota por $\mathtt{K}_p$ el subespacio de los p-vectores aleatorios constantes. El vector constante cuyas componentes sean más próximas en sentido $\mathtt{d}_2$ a las del vector aleatorio $\mathtt{f}$ es $P_{\langle \mathtt{K}_p\rangle}\mathtt{f} = \mathtt{E}_P[\mathtt{f}]$. La diferencia entre ambos es $\mathtt{f} - \mathtt{E}_P[\mathtt{f}]$, cuya componente i-ésima es $P_{\langle 1\rangle^\perp}\mathtt{f}_i$, para $i = 1, \ldots, p$. Esta discrepancia entre $\mathtt{f}$ y la situación determinista puede cuantificarse mediante la distancia $\mathtt{d}_{2,p}$ entre ambos que se denomina varianza multivariante total de $\mathtt{f}$. Concretamente,

$$\mathtt{var}_T[\mathtt{f}] = \mathtt{E}_P[\|\mathtt{f} - \mathtt{E}_P[\mathtt{f}]\|^2] = \sum_{i=1}^{p} \mathtt{var}[\mathtt{f}_i] \tag{9.19}$$

Nótese que este parámetro supone una generalización multivariante de la varianza. Los productos interiores entre las componentes del vector variabilidad total son las

[14] Algo análogo podemos decir respecto a la mediana (si es que está bien definida) en el contexto del espacio L_1 de funciones integrables. Concretamente, se trata de la constante k que minimiza la distancia $\int |f - k|\, dP$, siendo el mínimo $\mathtt{E}_P[f] - 1$.

covarianzas. Así pues, dos variables aleatorias son incorreladas cuando sus proyecciones sobre $\langle 1\rangle^{\perp}$ son perpendiculares según el producto interior definido en (9.25). Posteriormente interpretaremos este hecho en términos del problema de regresión lineal. Por otra parte, aplicando la desigualdad de Holder[15], se tiene que la covarianza al cuadrado es menor o igual que el producto de las varianzas, lo cual invita a definir el coeficiente de correlación lineal simple que, a la postre, tendrá una interpretación más clara que la de la covarianza. En definitiva,

$$P_{\langle 1\rangle} f \equiv \mathtt{E}[f], \quad \mathtt{d}_2^2(f, \mathtt{E}[f]) = \mathtt{var}[f]. \tag{9.20}$$

$$P_{\mathtt{K}_p}\mathtt{f} \equiv \mathtt{E}[\mathtt{f}], \quad \mathtt{d}_{2,p}^2(\mathtt{f}, \mathtt{E}[\mathtt{f}]) = \mathtt{var}_T[\mathtt{f}]. \tag{9.21}$$

$$\left\langle P_{\langle 1\rangle^{\perp}}\mathtt{f}_i, P_{\langle 1\rangle^{\perp}}\mathtt{f}_j \right\rangle = \mathtt{cov}[\mathtt{f}_i, \mathtt{f}_j], \quad i,j = 1,\ldots,p. \tag{9.22}$$

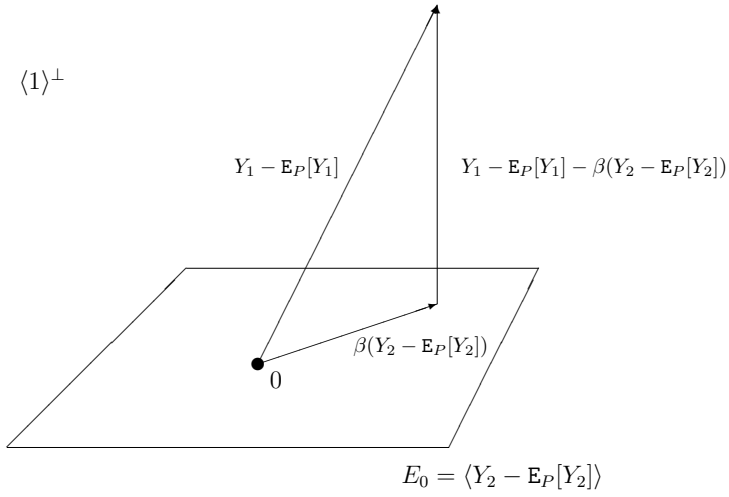
$$\begin{pmatrix} \mathtt{var}[\mathtt{f}_1] & \ldots & \mathtt{cov}[\mathtt{f}_1, \mathtt{f}_p] \\ \vdots & \ddots & \vdots \\ \mathtt{cov}[\mathtt{f}_p, \mathtt{f}_1] & \ldots & \mathtt{var}[\mathtt{f}_p] \end{pmatrix} = \mathtt{Cov}[\mathtt{f}]. \tag{9.23}$$

Dados una variable aleatoria Y_1 y un q-vector aleatorio Y_2 con matriz de varianzas-covarianzas conjunta $\Sigma > 0$. Sabemos que tanto Y_1 como Y_2 se descomponen ortogonalmente en sendas funciones constantes, las respectivas medias, más sus variabilidades totales, $Y_1 - \mathtt{E}_P[Y_1]$ y $Y_2 - \mathtt{E}_P[Y_2]$, repectivamente. Queremos saber en qué medida la variabilidad total de Y_1 puede ser explicada como combinación lineal de la de Y_2.

Se trata de la proyección ortogonal de $Y_1 - \mathtt{E}_P[Y_1]$ sobre el subespacio $E_0 \subset L^2$ compuesto por las funciones de la forma $\beta(Y_2 - \mathtt{E}_P[Y_2])$, para algún $\beta \in \mathcal{M}_{1\times q}$. Se denotará también mediante $\langle Y_2 - \mathtt{E}_P[Y_2]\rangle$. En definitiva, buscamos pues el valor de β tal que

$$Y_1 - \mathtt{E}_P[Y_1] - \beta(Y_2 - \mathtt{E}_P[Y_2]) \ \perp \ Y_2 - \mathtt{E}_P[Y_2] \tag{9.24}$$

[15] Caso particular de la de Cauchy-Schwarz



De (9.24) se sigue que β es la solución a la ecuación

$$\Sigma_{12} = \beta\Sigma_{22},$$

es decir,

$$\beta = \Sigma_{12}\Sigma_{22}^{-1}. \tag{9.25}$$

Aplicando las propiedades de la proyección ortogonal se tiene entonces que la combinación afín de las componentes de Y_2 que más se aproxima en el sentido $\mathtt{d}_2$ a Y_1[16] es $\alpha + \beta Y_2$, siendo

$$\alpha = \mathtt{E}[Y_1] - \beta\mathtt{E}[Y_2] \tag{9.26}$$

Si Y_1 es un p-vector aleatorio, podemos razonar de igual forma y por separado para cada una de sus componentes, de manera que β será una matriz $p \times q$ y α un vector p-dimensional. El vector aleatorio $Y_1 - (\alpha + \beta Y_2) = Y_1 - \mathtt{E}_P(Y_1) - \beta(Y_2 - \mathtt{E}_P[Y_2])$, recoge la parte de la variabilidad total de Y_1 no explicada linealmente por la variabilidad total de Y_2. Ésta es constante (es decir, estaríamos hablando de una situación determinista) si, y sólo si, es nula, en cuyo caso Y_1 quedaría determinado por el valor de Y_2 mediante la relación afín anterior. Ello invita a considerar la matriz de varianzas-covarianzas de dicha diferencia. Teniendo en cuenta la ilustración anterior,

[16] Es decir, la proyección de Y_1 sobre el subespacio $\langle 1|Y_2\rangle$

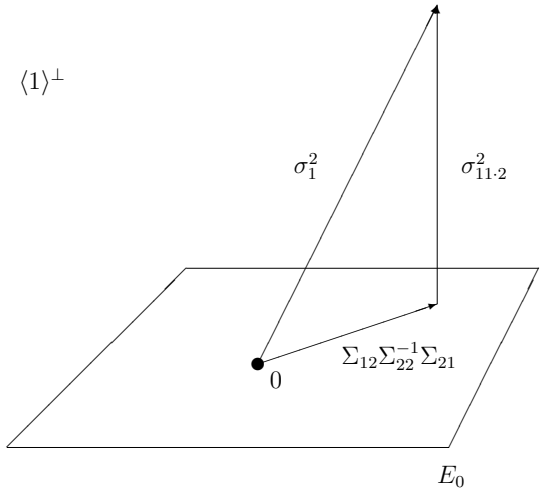
puede obtenerse mediante:

$$\begin{aligned}\mathtt{Cov}_P[Y_1-(\alpha+\beta Y_2)] &= \big\langle Y_1-\mathtt{E}_P(Y_1), Y_1-\mathtt{E}_P(Y_1)-\beta(Y_2-\mathtt{E}_P[Y_2])\big\rangle\\ &= \big\langle Y_1-\mathtt{E}_P(Y_1), Y_1-\mathtt{E}_P(Y_1)\big\rangle\\ &+ \beta\big\langle Y_2-\mathtt{E}_P(Y_2), Y_2-\mathtt{E}_P(Y_2)\big\rangle\beta'\\ &= \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\end{aligned}$$

Estamos hablando pues de la a matriz de varianzas-covarianzas parciales, definida en (9.13). Ésta es menor o igual que la matriz de varianza-covarianzas de Y_1 en el sentido del preorden definido en (9.4).

En el caso $p = 1$ tendremos la varianza parcial, que será menor o igual que la varianza total de Y_1. Analizando los dos casos extremos tenemos, primeramente, que un valor nulo de la varianza parcial se corresponderá con una dependencia afín perfecta (determinista) de Y_1 respecto a Y_2; por contra, un valor de la varianza parcial igual al de la varianza total se corresponde con $\beta = 0$ y $\alpha = \mathtt{E}[Y_1]$. En tal caso, la variabilidad total de las componentes de Y_2 no sirve en absoluto para *explicar* linealmente la variabilidad total de Y_1. Este hecho se corresponde con el caso $\Sigma_{12} = 0$. De esta forma podemos interpretar la incorrelación entre variables aleatorias. En general, el término $\rho^2_{1\cdot 2}$ se interpreta como la proporción de variabilidad total de Y_1 *explicada*[17] linealmente por la variabilidad total de Y_2. Este coeficiente generaliza el de correlación lineal simple definido en (9.10), en el sentido de que el primero es el cuadrado del segundo cuando $q = 1$. Para ilustrarlo, se expresan en el siguiente gráfico las normas al cuadrado de los vectores (varianzas).

[17] Esta interpretación heurística del coeficiente de correlación, muy frecuente en nuestra teoría, será comentada y matizada en el capítulo 3.



Nótese que, por la ortogonalidad de la descomposición, se verifica que

$$\sigma_1^2 = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} + \sigma_{11\cdot 2}^2$$

De esta manera, el coeficiente de correlación múltiple al cuadrado que se define como el cociente

$$\rho_{1\cdot 2}^2 = \frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_1^2},$$

se interpreta, como hemos dicho anteriormente, como la proporción de la varianza de Y_1 explicada linealmente por Y_2, mientras que la parte no explicada es

$$\sigma_{11\cdot 2}^2 = \sigma_1^2(1 - \rho_{1\cdot 2}^2)$$

Además, puede demostrarse (cuestión propuesta) que ρ_{12}^2 es la máxima correlación lineal simple al cuadrado entre Y_1 y una variable aleatoria de la forma bY_2, con $b \in \mathcal{M}_{1\times q}$, que se alcanza en $b = \beta$.

Esperanza condicional

El estudio de los parámetros anteriores tendrá mayor alcance a la luz de los conceptos de esperanza condicional, probabilidad condicional regular e independencia condicional, que introducimos a continuación. Dadas una variable aleatoria Z, de $(\Omega, \mathcal{A}, P)$ en $(\Omega_Z, \mathcal{A}_Z)$, y una variable aleatoria real Y no negativa o integrable, se

define $\mathtt{E}_P[Y|Z]$ como la clase de variables aleatorias reales definidas sobre $(\Omega_Z, \mathcal{A}_Z)$ verificando la propiedad[18]

$$\int_B g\, dP^Z = \int_{Z^{-1}(B)} Y\, dP, \quad \forall B \in \mathcal{A}_Z.$$

Puede probarse[19] que, si E_1 denota el subespacio lineal cerrado de $L^2(\Omega, \mathcal{A}, P)$ constituido por las funciones de la forma $f \circ Z$, para alguna variable aleatoria $f : (\Omega_Z, \mathcal{A}_Z) \to \mathbb{R}$, se verifica que

$$\mathtt{E}[Y|Z] \circ Z = P_{E_1} Y, \tag{9.27}$$

es decir, la esperanza condicional es la función de Z que más se aproxima a Y en los términos de la distancia $\mathtt{d}_2$ definida en (9.17). Podríamos hablar pues de la mejor aproximación mínimo-cuadrática.

Si Y es un n-vector aleatorio real, queda garantizada la existencia de una probabilidad de transición $P^{Y|Z}$, de $\Omega_Z \times \mathcal{R}^n$ en $[0,1]$, tal que, para cada $A \in \mathbb{R}^n$, $P^{Y|Z}(\cdot, A)$ es una versión de $P[Y \in A|Z]$, es decir, de $\mathtt{E}[I_{Y^{-1}(A)}|Z]$. Una función en esas condiciones se denomina versión de la probabilidad condicional regular de Y dada Z. Las propiedades de la misma pueden estudiarse con detalle en Billingsley (1986). Mencionaremos aquí tres de ellas: en primer lugar, la esperanza condicional de Y dada Z es la media de la variable $P^{Y|Z=\mathtt{z}}$, para cualquier versión probabilidad condicional regular; la distribución conjunta de Y y Z se reconstruye como producto generalizado entre $P^{Y|Z}$ y P^Z; por último, Y y Z son independientes si, y sólo si, podemos encontrar una versión de $P^{Y|Z}$ constante en Z.

Puede probarse fácilmente que, si P^Y y P^Z están dominadas por sendas medidas σ-finitas μ_1 y μ_2, siendo f_Y y f_Z sus respectivas densidades, entonces $P^{(Y,Z)}$ está dominada por la medida producto $\mu_1 \times \mu_2$. Además, si se denota por f la correspondiente función de densidad, la siguiente función, bien definida P^Z-c.s., constituye una densidad de $P^{Y|Z=\mathtt{z}}$ respecto a μ_1:

$$f_{Y|Z=\mathtt{z}}(\mathtt{y}) = \frac{f(\mathtt{y}, \mathtt{z})}{f_Z(\mathtt{z})} \tag{9.28}$$

Por otra parte, si Y descompone en dos subvectores, Y_1 e Y_2, de dimensiones p y q, respectivamente, se dice que Y_1 e Y_2 son condicionalmente independientes dado Z,

[18] El Teorema de Radom-Nicodym garantiza la existencia de esta familia de funciones. Además, las funciones en tales condiciones constituyen una clase de equivalencia en el conjunto de las funciones $\mathcal{A}_Z$-medibles, pues dos cualesquiera serán iguales P^Z-casi seguro, es decir, salvo en un conjunto de $\mathcal{A}_Z$ de probabilidad nula. Por otra parte, si Y es un n-vector aleatorio de componentes $Y_1, \ldots, Y_n$, se define $\mathtt{E}[Y|Z] = (\mathtt{E}[Y_1|Z], \ldots, \mathtt{E}[Y_n|Z])'$, cuando tenga sentido. En general, las propiedades fundamentales de la Esperanza Condicional pueden estudiarse en Ash (1972) o Nogales (1998).

[19] Ver Nogales (1998).

lo cual se denota mediante $Y_1 \perp\!\!\!\perp Y_2|Z$, cuando se puede construir una versión de la probabilidad condicional regular de Y dada Z mediante

$$P^{Y|Z=\mathtt{z}} = P^{Y_1|Z=\mathtt{z}} \times P^{Y_2|Z=\mathtt{z}}, \quad \mathtt{z} \in \Omega_Z,$$

lo cual equivale afirmar que se pude construir una versión de la probabilidad condicional regular de Y_1 dadas Y_2 y Z mediante

$$P^{Y_1|Y_2=\mathtt{y}_2,Z=\mathtt{z}} = P^{Y_1|Z=\mathtt{z}}, \quad (\mathtt{y}_2, \mathtt{z}) \in \mathbb{R}^{n_1} \times \Omega_Z.$$

Ello viene a significar, en términos heurísticos que, conocido el valor que toma Z, el hecho de conocer también el valor de Y_2 no condiciona el resultado de Y_1. En general no es cierto que la independencia entre dos variables aleatorias implique la independencia condicional entre las mismas dada otra tercera variable[20]. Una interesante propiedad de la probabilidad condicional de la que se hace uso muy a menudo es la siguiente: en las condiciones anteriores, si f es variable aleatoria real definida sobre $\mathbb{R}^{n_1+n_2}$, se verifica que

$$\mathtt{E}[f \circ (Y_1, Y_2)|Y_2 = \mathtt{y}_2] = \int_{\mathbb{R}^{n_2}} f(\cdot, \mathtt{y}_2)\, dP^{Y_1|Y_2=\mathtt{y}_2}, \tag{9.29}$$

donde $f(\cdot, \mathtt{y}_2)$ es la variable aleatoria real que asigna a cada $\mathtt{y}_1 \in \mathbb{R}^{n_1}$ el número $f(\mathtt{y}_1, \mathtt{y}_2)$.

Si $\left(P^{Y_1|Y_2=\mathtt{y}_2}\right)^{f(\cdot,\mathtt{y}_2)}$ denota la distribución de dicha variable respecto de $P^{Y_1|Y_2=\mathtt{y}_2}$, se tiene como corolario inmediato que

$$P^{f\circ(Y_1,Y_2)|Y_2=\mathtt{y}_2} = \left(P^{Y_1|Y_2=\mathtt{y}_2}\right)^{f(\cdot,\mathtt{y}_2)}, \quad (\mathtt{y}_1, \mathtt{y}_2) \in \mathbb{R}^{n_1+n_2}. \tag{9.30}$$

Si la probabilidad de $f \circ Y$ condicionada a Y_2 resulta no depender de el valor que tome esta última, se deduce que ambas son independientes, coincidiendo la distribución condicional anterior con la propia distribución marginal de $f \circ Y$. [21]

Por último, vamos a añadir algunos comentarios a las conclusiones obtenidas en el apartado anterior. Sean de nuevo Y_1 e Y_2 una variable aleatoria real y un q-vector aleatorio, respectivamente. Recordemos que E_0 denota el subespacio cerrado

[20] Véase el ejercicio 18 al final del capítulo.

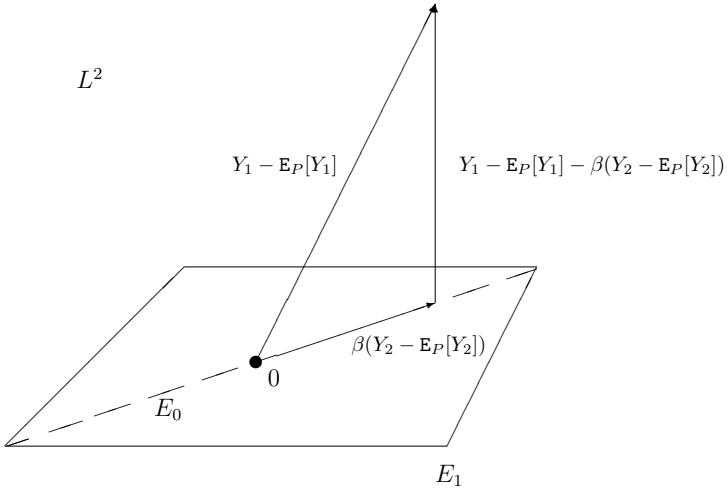
[21] Esta situación ocurre, por ejemplo, en el Modelo de Correlación Lineal. Este Modelo tiene la propiedad de que, al condicionar sobre un valor concreto de las variables explicativas, se obtiene un Modelo de Regresión Lineal. Según hemos dicho, cualquier variable definida en el modelo condicional, es decir, el de Regresión, cuya distribución no dependa del valor concreto de las variables explicativas (F-Snedecor o χ^2centrales, por ejemplo), será independiente de éstas y tendrá la misma distribución si se considera desde el modelo inicial, es decir, el de Correlación.

de $L^2(\Omega, \mathcal{A}, P)$ constituido por las combinaciones lineales de las componentes de $Y_2 - \mathtt{E}_P[Y_2]$, y sea E_1 el subespacio cerrado compuesto por las funciones medibles de Y_2 o, equivalentemente, de $Y_2 - \mathtt{E}_P[Y_2]$. En ese caso se verifica que $E_0 \subset E_1$. Obviamente, que la función $f : (\mathbb{R}^q, \mathcal{R}^q) \to \mathbb{R}$ que minimiza la distancia $\mathtt{d}_2$ entre $Y_1 - \mathtt{E}_P[Y_1]$ y $f \circ (Y_2 - \mathtt{E}_P[Y_2])$ sea lineal equivale a que las proyecciones de $Y_1 - \mathtt{E}_P[Y_1]$ sobre E_1 y E_0 coincidan. Según (9.24), la diferencia $Y_1 - \mathtt{E}_P[Y_1] - \beta(Y_2 - \mathtt{E}_P[Y_2])$ es ortogonal a $Y_2 - \mathtt{E}_P[Y_2]$, es decir, son incorreladas, pues las medias son nulas. Supongamos por un momento que la probabilidad P es tal que la incorrelación (ortogonalidad de las variabilidades totales) implica la independencia, cosa que sucede si el vector $(Y_1' Y_2')'$ es normal multivariante. Entonces, con mayor razón, se tendría que

$$Y_1 - \mathtt{E}_P[Y_1] - \beta(Y_2 - \mathtt{E}_P[Y_2]) \ \perp \ f \circ (Y_2 - \mathtt{E}_P[Y_2]),$$

para toda variable aleatoria real f sobre $\mathbb{R}^q$. En consecuencia,

$$P_{E_0}(Y_1 - \mathtt{E}_P[Y_1]) = P_{E_1}(Y_1 - \mathtt{E}_P[Y_1]).$$



Teniendo en cuenta la descomposición ortogonal $E_1 = \langle 1 \rangle \oplus E_1 | \langle 1 \rangle$, se concluiría que

$$\mathtt{E}[Y_1 | Y_2] \circ Y_2 = \alpha + \beta Y_2$$

Por lo tanto, la función de Y_2 más próxima en términos de $\mathtt{d}_2$ a Y_1 sería es una transformación lineal de la variabilidad total de Y_2 o, lo que es lo mismo una transformación a afín de Y_2, concretamente, $\alpha + \beta Y_2$.

En ese caso, las varianzas y covarianzas parciales podrían entenderse como la parte la matriz de varianzas-covarianzas de Y_1 no explicada por Y_2. Decimos por Y_2 y no por la relación lineal (afín, si queremos ser más precisos) con Y_2, dado que, en estas condiciones (recordamos, cuando incorrelación equivale a independencia), la relación con Y_2 es afín. Este pequeño matiz otorgará pleno sentido a la matriz de varianzas-covarianzas parciales y, en consecuencia, al coeficiente de correlación lineal múltiple (o canónicos) en el caso normal multivariante, donde esta condición se verifica, según la proposición (2.3) . Además, la matriz de varianza-covarianzas parciales se relacionará en la sección dedicada al estudio de la distribución normal multivariante con el concepto de independencia condicional introducido anteriormente.

9.3. Generalidades sobre Estadística

Lo dicho en la sección anterior se enmarca en un contexto meramente probabilístico, pues la distribución se supone conocida. La Estadística se sitúa en una fase anterior, en la cual la distribución de probabilidades no se conoce. En ese caso, tras imponer una serie de restricciones razonables más o menos fuertes a la misma, tendremos una familia de distribuciones candidatas. Todo el trabajo estadístico va encaminado, de una u otra forma, a determinar la *verdadera* distribución. Así pues, el punto de partida formal será un par compuesto por un espacio medible y una familia de probabilidades sobre el mismo. Definimos[22] experimento estadístico (también estructura estadística o modelo estadístico) como un terna de la forma

$$(\Omega, \mathcal{A}, \mathcal{P}), \tag{9.31}$$

siendo $\mathcal{P}$ una familia de probabilidades sobre $(\Omega, \mathcal{A})$. Con frecuencia, la familia $\mathcal{P}$ se expresa con la ayuda de cierto conjunto Θ y una función sobreyectiva $P_- : \Theta \to \mathcal{P}$, que asigna a cada θ de Θ la distribución P_θ, de forma que el modelo estadístico se escribe de la forma

$$(\Omega, \mathcal{A}, \{P_\theta : \theta \in \Theta\}) \tag{9.32}$$

Los conjuntos Ω y Θ se denominan, en ese caso, espacio de observaciones y espacio de parámetros, respectivamente. Realmente, el objeto del estudio estadístico no suele ser un espacio de probabilidad abstracto sino un n-vector aleatorio real Y, donde $n \geq 1$, definido sobre un cierto espacio de probabilidad $(\underline{\Omega}, \underline{\mathcal{A}}, \underline{P})$, cuya distribución $\underline{P}^Y$ es desconocida aunque se supone perteneciente a una familia $\mathcal{P}$ de distribuciones sobre $\mathbb{R}^n$, lo cual conduce a considerar el modelo $(\mathbb{R}^n, \mathcal{R}^n, \mathcal{P})$. Por ello, nos permitiremos

[22] Esta definición es discutible. De hecho, en el capítulo 6 trabajamos con una definición alternativa.

la licencia de expresar también dicho modelo mediante $Y \sim \mathtt{P}$, $\mathtt{P} \in \mathcal{P}$, o bien, cuando $\mathcal{P}$ esté parametrizada, mediante $Y \sim P_\theta$, $\theta \in \Theta$. En concreto, en nuestro estudio el espacio de observaciones será siempre un subconjunto de interior no vacío de $\mathbb{R}^n$, para algún $n \in \mathbb{N}$ y las distribuciones de la familia estarán dominadas por la medida de Lebesgue en $\mathbb{R}^n$. En general, cuando la familia está dominada por una medida σ-finita, las probabilidades quedan caracterizadas, en virtud del Teorema de Radom-Nikodym, por sus correspondientes densidades $\{p_\theta : \theta \in \Theta\}$. En ese caso, suele considerarse una única función, denominada función de verosimilitud, definida sobre $\Omega \times \Theta$ mediante

$$\mathcal{L} : (\theta; \omega) \in \Omega \times \Theta \mapsto p_\theta(\omega).$$

En estas condiciones, una variable aleatoria S definida en nuestro modelo (que en el contexto de la Estadística se denomina estadístico) se dice suficiente[23] cuando existe una función $\tilde{\mathcal{L}}$ tal que

$$\mathcal{L}(\theta; \omega) = \tilde{\mathcal{L}}(\theta; S(\omega)).$$

Se entiende pues que la *información* referente al parámetro que contiene la observación ω queda perfectamente resumida en $S(\omega)$. Sería interesante comentar aquí diferentes aproximaciones a la idea de Información, aunque nos conformaremos con presentar la definición de Fisher, que es la que mejor casa con esta definición de suficiencia.

Sea $(\Omega, \mathcal{A}, \mathcal{P})$ un modelo estadístico dominado tal que $\mathcal{P}$ se expresa con la ayuda de un parámetro $\theta \in \Theta$, siendo Θ un abierto de $\mathbb{R}^s$. En el caso de que la función de verosilimitud $\mathcal{L}$ verifique las condiciones de regularidad necesarias, se define la información asociada al modelo para el parámetro θ como la función $\mathcal{I} : \Theta \to \mathcal{M}_{s\times s}$ siguiente

$$\mathcal{I}(\theta) = \mathtt{Cov}_\theta[V_\theta], \tag{9.33}$$

siendo

$$V_\theta(\omega) = \left(\frac{\partial \log \mathcal{L}(\omega, \theta)}{\partial \theta_1}, \ldots, \frac{\partial \log \mathcal{L}(\omega, \theta)}{\partial \theta_s} \right)'$$

Puede demostrarse sin dificultad que

$$\mathtt{E}_\theta \left[\frac{\partial \log \mathcal{L}}{\partial \theta_j} \right] = 0, \qquad 1 \leq j \leq s \tag{9.34}$$

y que las componentes de la matriz de información pueden obtenerse mediante

$$\mathcal{I}_{jk} = -\mathtt{E}_\theta \left[\frac{\partial^2 log \mathcal{L}}{\partial \theta_j \partial \theta_k} \right], \qquad 1 \leq j, k, \leq s \tag{9.35}$$

[23] La definición que se presenta aquí tiene sentido únicamente en el caso dominado. En general, se dice que un estadístico S es Suficiente cuando para cada $A \in \mathcal{A}$, $\cap_{\mathtt{P}\in\mathcal{P}} \mathtt{E}_\mathtt{P}[I_A|S] \neq \emptyset$. El Teorema de factorización Neyman-Halmos-Savage permite la traducción al caso dominado.

También puede demostrarse fácilmente que, efectivamente, que en el caso dominado y con las condiciones de regularidad necesarias un estadístico suficiente S conduce a un nuevo modelo reducido en el que la información de Fisher permanece invariante.

Esta y otras definiciones de información, como la de Kullback[24], al igual que otros muchos conceptos con los que trabajaremos, como el caso de la suficiencia, el principio de máxima verosimilitud, etcétera, son de fácil manejo cuando el modelo estudiado es de tipo exponencial. Decimos que un modelo estadístico dominado es exponencial cuando puede expresarse con la ayuda de cierto parámetro $\theta \in \Theta$ mediante dos funciones T y Q con valores en $\mathbb{R}^s$ definidas sobre $(\Omega, \mathcal{A})$ y Θ, respectivamente, y otras dos h y C definidas respectivamente sobre los mismos espacios con valores en $\mathbb{R}^+$, tales que

$$\mathcal{L}(\theta; \omega) = \exp\left\{\langle Q(\theta), T(\omega)\rangle + c(\theta) + d(\omega)\right\} \tag{9.36}$$

En ese caso, se sigue directamente del teorema de factorización que el estadístico T es suficiente. Como ejemplos de modelos exponenciales podemos citar las familias normales, binomiales y de Poisson. El modelo lineal normal es un ejemplo de modelo exponencial. Puede probarse fácilmente que, mediante una modificación adecuada del parámetro y de la medida dominante, la función de verosimilitud puede expresarse de manera canónica mediante

$$\mathcal{L}^*(\theta^*; \omega) = \exp\left\{\langle \theta^*, T(\omega)\rangle + c^*(\theta^*)\right\} \tag{9.37}$$

Expresar el modelo de esa forma es de enorme utilidad a la hora de buscar un estadístico completo. El concepto de completitud es, en cierta forma, complementario al de suficiencia. Se dice que un estadístico X con valores en $\mathbb{R}^k$ es completo cuando, para cada variable aleatoria real g definida sobre $\mathbb{R}^k$, se verifica

$$\left[\mathrm{E}_\theta[g] = 0, \quad \forall \theta \in \Theta\right] \;\Rightarrow\; \left[g = 0 \; P_\theta^X - \text{casi seguro}, \quad \forall \theta \in \Theta\right]$$

Decimos que suficiencia y completitud son propiedades complementarias porque de la coincidencia de ambas pueden extraerse interesantes beneficios, como veremos más adelante.

Teorema 9.18.

En un modelo estadístico del tipo (9.37) con Θ es de interior no vacío en $\mathbb{R}^s$, el estadístico T es, además de suficiente, completo.

Remitimos al lector interesado en los conceptos de Suficiencia, Información y Completitud, así como en el estudio de las familias exponenciales, a las referencias Lehmann (1986) y Nogales (1998).

[24] Ver Nogales (1998).

Problema de Estimación

Ya hemos comentado que el propósito final de la Estadística es determinar cuál es, de entre una familia de candidatas, la verdadera probabilidad que rige un fenómeno aleatorio. A este objetivo podemos aproximarnos mediante dos tipos de estudios: el de Estimación y el de Contraste de Hipótesis. El primer problema consiste en, dada una función $\tilde{g}$, denominada estimando, definida sobre $\mathcal{P}$ y con valores en cierto conjunto Δ, encontrar un estadístico T, denominado estimador, con valores en Δ, de manera que, si $\mathtt{P}$ es la verdadera distribución y ω es la observación del experimento, $T(\omega)$ sea *próximo* a $\tilde{g}(\mathtt{P})$.

Como ya sabemos, la familia de distribuciones $\mathcal{P}$ suele expresarse con la ayuda de un espacio de parámetros Θ. Si la identificación se realiza mediante una biyección, existe una única función paramétrica (es decir, definida sobre el espacio de parámetros Θ) $g : \Theta \to \Delta$ tal que

$$g = \tilde{g} \circ P_{\_} \tag{9.38}$$

En general, es decir, si no se supone que la aplicación $P_{\_}$ es inyectiva[25], una función paramétrica g se dice estimable cuando existe un estimando $\tilde{g} : \mathcal{P} \to \Delta$ verificando (9.38). Luego, una función paramétrica g se dice estimable cuando se verifica

$$[P_{\theta_1} = P_{\theta_2}] \;\Rightarrow\; [g(\theta_1) = g(\theta_2)] \tag{9.39}$$

Por otra parte, debemos especificar qué entendemos por *proximidad*. Por ejemplo, si $\Delta = \mathbb{R}$, es muy frecuente considerar la función de *pérdida* cuadrática W, denominada función de pérdida y definida mediante $W(\delta_1, \delta_2) = (\delta_1 - \delta_2)^2$. De esta forma, el problema estadístico consiste en encontrar, si es posible, el estimador T tal que, para cada $\theta \in \Theta$, haga mínimo el denominado error cuadrático medio

$$\mathtt{E}_\theta[W(T, g(\theta)] = E_\theta[(T - g(\theta))^2]. \tag{9.40}$$

Esta forma de proceder es acorde con la técnica de mínimos cuadrados, de ahí su popularidad, aunque no sea la única función de pérdida a considerar[26]. Obviamente, se verifica la siguiente descomposición:

$$E_\theta[(T - g(\theta))^2] = \big(\mathtt{E}_\theta[T] - g(\theta)\big)^2 + \mathtt{var}_\theta[T]. \tag{9.41}$$

El término $\mathtt{E}_\theta[T] - g(\theta)$ se denomina sesgo de T. Cuando es nulo para cada θ se dice que T es un estimador insesgado de g, es decir, que, por término medio, la estimación

[25] Como sucede en e capítulo 6. De hecho, el estudio del modelo lineal de rango no completo es la causa de esta duscusión.

[26] Considerar, por ejemplo, la función de pérdida (3.11).

es correcta en todo caso. Si restringimos la búsqueda de estimadores apropiados a la familia de estimadores insesgados, entonces, (9.41) coincide con $\mathtt{var}_\theta[T]$. Por lo tanto, con esta restricción, nuestro propósito será encontrar el estimador insesgado de mínima varianza (EIMV, para abreviar), si existe, y será óptimo entre una clase de estimadores verificando una propiedad (el ser insesgado) muy razonable, aunque fuertemente restrictiva.

Si $\Delta = \mathbb{R}^k$, podemos generalizar lo anterior considerando la familia $\mathcal{W} = \{W_y : y \in \mathbb{R}^k\}$, siendo W_y la función de pérdida definida mediante $W_y(\delta_1, \delta_2) = \langle y, \delta_1 - \delta_1\rangle^2$. Así, el problema en dimensión k consiste en encontrar el estimador T que, para cada $\theta \in \Theta$, minimice

$$E_\theta\big[\big(T - g(\theta)\big)\big(T - g(\theta)\big)'\big]. \tag{9.42}$$

Al hablar de minimizar estamos refiriéndonos al preorden definido en $\mathcal{M}_{k\times k}$ mediante (9.4). La expresión anterior descompone de forma análoga a (9.41)

$$E_\theta\big[\big(T - g(\theta)\big)\big(T - g(\theta)\big)'\big] = \big(\mathsf{Sesgo}_\theta[T]\big)\big(\mathsf{Sesgo}_\theta[T]\big)' + \mathtt{Cov}_\theta[T]. \tag{9.43}$$

Si imponemos la condición de que el estimador sea insesgado, se trata de buscar aquél que, para cada $\theta \in \Theta$, minimize la matriz de varianzas-covarianzas, por lo que dicho estimador, si existe, se denominará igualmente EIMV. No obstante, pueden considerarse otras funciones de pérdida, por ejemplo (3.11), según las cuales el EIMV pierda su condición de estimador óptimo. El Teorema de Lehmann-Scheffé, cuya demostración puede encontrarse en Nogales (1998), permite obtener el EIMV a partir de un estimador insesgado y un estadístico suficiente y completo.

Teorema 9.19.
Dado un T estimador insesgado y de cuadrado integrable de un estimando g, y un estadístico S suficiente y completo, el estadístico[27] $\mathtt{E}[T|S] \circ S$ es el único[28] EIMV de g.

Otro método para buscar un estimador adecuado del estimando $g = \mathtt{Id}$ es el de Máxima Verosimilitud. Se define el estimador de máxima verosimilitud (EMV, para abreviar), como aquél que hace corresponder a cada observación $\omega \in \Omega$ el valor de θ que maximice $\mathcal{L}(\theta; \omega)$. Por lo tanto, para poder hablar del EMV, dicho máximo debe existir de manera que podamos construir una función medible. El EMV, cuando existe, presenta excelentes propiedades asintóticas, tal y como se comenta en la sección 4 o en el capítulo 8.

[27] Nótese que se habla de una única una versión de $E_\theta[T|S]$ común a cualquier valor del parámetro. Ello es posible por ser S suficiente.

[28] Cualquier otro difiere de éste en un suceso nulo para todas las probabilidades de la familia.

Problema de Contraste de Hipótesis

Un problema de Contraste de Hipótesis consiste en considerar un subconjunto propio $\mathcal{P}_0 \subset \mathcal{P}$ y decidir si la verdadera distribución pertenece a $\mathcal{P}_0$. Es decir, se trata de aceptar o rechazar una hipótesis inicial $H_0 : \mathtt{P} \in \mathcal{P}_0$. Si la familia $\mathcal{P}$ se expresa con la ayuda de un espacio espacio de parámetros Θ, se denotará por Θ_0 la antiimagen por P_- de $\mathcal{P}_0$, de forma que la hipótesis inicial se denota mediante $H_0 : \theta \in \Theta_0$. En general, diremos que una hipótesis inicial Θ_0 es contrastable cuando es la antiimagen por P_- de algún subconjunto propio $\mathcal{P}_0 \subset \mathcal{P}$, es decir, cuando se verifica

$$\left[P_{\theta_1} = P_{\theta_2}\right] \implies \left[\theta_1 \in \Theta_0 \Leftrightarrow \theta_2 \in \Theta_0\right] \tag{9.44}$$

La decisión se tomará en función del resultado de un test no aleatorio[29] $\phi : (\Omega, \mathcal{A}) \to \{0, 1\}$, donde los valores 0 y 1 se interpretan como la aceptación y el rechazo, respectivamente, de la hipótesis inicial o nula. La función potencia del test se define sobre el espacio de parámetros mediante $\beta_\phi(\theta) = \mathtt{E}_\theta[\phi]$. En consecuencia, de un *buen* test cabe exigir que su función potencia sea baja en Θ_0 y alta en su complementario. El test óptimo sería aquél cuya función potencia fuera mínima en todo Θ_0 y máxima en Θ_0^c. Lógicamente, un test en esas condiciones sólo existirá en caso triviales, por lo que debemos rebajar nuestra pretensiones. Un procedimiento muy usual a la hora de buscar un test adecuado establecido por Neyman y Pearson consiste en lo siguiente: se denomina nivel de significación del test al supremo de la función potencia en Θ_0. Se fija entonces un número $\alpha \in (0, 1)$, a ser posible pequeño (el valor más utilizado es 0.05), y se trata de encontrar el test que maximice la función potencia en Θ_0^c entre aquéllos cuyo nivel de significación sea, a lo sumo, α. Si existe, se denomina test uniformemente más potente a nivel α (UMP a nivel α, para abreviar). En muchas ocasiones, no existe (o no sabemos encontrar) un test en esas condiciones, por lo que se hace necesario imponer restricciones adicionales sobre los tests a considerar, por ejemplo que sean invariantes (ver el siguiente apartado) o insesgados (es decir, que β_ϕ sea mayor o igual que α en Θ_0^c). Un test uniformemente más potente entre los invariantes se dice UMP-invariante y es necesariamente insesgado.

El Lema fundamental de Neyman-Pearson cuyo enunciado completo y demostración podemos encontrar en Nogales (1998), pp. 180-182, puede considerarse el pilar básico en la construcción de tests UMP. Se enuncia para un experimento estadístico con familia de probabilidades binaria $\{P_0, P_1\}$ y dominada, siendo p_0 y p_1 las respectivas densidades. En esas condiciones, el test UMP a nivel α para contrastar la hipótesis inicial $\{P_0\}$ consiste en rechazar al hipótesis inicial si, y sólo si, la observa-

[29] En nuestra teoría no consideraremos test aleatorios, con valores en $[0, 1]$.

ción ω satisface

$$p_1(\omega) > k \cdot p_0(\omega) \tag{9.45}$$

donde la constante k se escoge de manera que el nivel de significación del test sea α.

Este resultado puede extenderse a experimentos estadísticos con razón de verosimilitud monótona, como puede ser el caso del Modelo Lineal Normal tras dos reducciones consecutivas por Suficiencia e Invarianza. Concretamente, dado un experimento estadístico de la forma $(\mathbb{R}, \mathcal{R}, \{P_\theta : \theta \in \Theta \subset \mathbb{R}\})$ y dominada por la medida de Lebesgue, decimos que posee razón de verosimilitud monótona cuando, para cada par $\theta_1 < \theta_2$, la función $p_{\theta_2}/p_{\theta_1}$ es no decreciente. En ese caso, el siguiente resultado, cuya demostración podemos encontrar en Nogales (1998), pp. 180-186, permite obtener un test UMP a nivel α en un problema de contraste de una hipótesis unilateral $\{\theta \leq \theta_0\}$ frente a su alternativa $\{\theta > \theta_0\}$.

Proposición 9.20.

En las condiciones anteriores, el test ϕ definido mediante

$$\phi_(\omega) = \begin{cases} 1 & \text{si } \omega > C \\ 0 & \text{si } \omega \leq C \end{cases},$$

es UMP a nivel $\alpha = P_{\theta_0}\big((C, +\infty)\big)$.

Al igual que ocurre en el problema de Estimación, podemos acogernos al Principio de Máxima Verosimilitud para construir un test de hipótesis muy natural y con interesantes propiedades asintóticas (ver sección 4). Consideremos un modelo estadístico dominado $\big(\Omega, \mathcal{A}, \{P_\theta : \theta \in \Theta\}\big)$ con función de verosimilitud $\mathcal{L}$, y supongamos que queremos contrastar una hipótesis inicial $\Theta_0 \subset \Theta$. Se denomina Razón de Verosimilitudes (RV, para abreviar) a la función siguiente

$$RV(\omega) := \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\omega; \theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\omega; \theta)}, \quad \omega \in \Omega.$$

Se trata pues de una aplicación definida sobre el espacio de observaciones Ω con valores en $[0, 1]$. Supongamos que existe (es decir, que ambos supremos se alcanzan) y que es $\mathcal{A}$-medible. En ese caso, un test de la razón de verosimilitudes a nivel $\alpha \in (0, 1)$ es un test de la forma

$$\phi(\omega) = \begin{cases} 1 & \text{si } RV(\omega) < C \\ 0 & \text{si } RV(\omega) \geq C \end{cases}, \tag{9.46}$$

donde C es una constante tal que

$$\sup_{\theta \in \Theta_0} P_\theta(R_V < C) = \alpha. \tag{9.47}$$

En particular, si existe una probabilidad P sobre $[0,1]$ tal que $P_\theta^{RV} = \mathsf{P}$, para todo $\theta \in \Theta_0$, y existe $\mathsf{P}^{1-\alpha}$, entonces $\mathsf{P}^{1-\alpha}$ es la única constante que verifica (9.47). Por lo tanto, el único test de la razón de verosimilitudes a nivel α será TRV, definido mediante

$$TRV(\omega) = \begin{cases} 1 & \text{si } RV(\omega) < \mathsf{P}^{1-\alpha} \\ 0 & \text{si } RV(\omega) \geq \mathsf{P}^{1-\alpha} \end{cases}, \tag{9.48}$$

Hay que tener en cuenta que, según el Lema Fundamental de Neyman-Pearson, más concretamente en virtud de (9.45), el test UMP a nivel α para contrastar una hipótesis unitaria en una familia binaria dominada es el que cabría esperar de la aplicación inmediata del Principio de Máxima verosimilitud. Por ello, no es de extrañar que el test de la razón de verosimilitudes resulte a su vez UMP, al menos dentro de una subclase de tests, como pueden ser los invariantes. Y es que también podemos establecer condiciones naturales que propician la concordancia entre el Principio de Invarianza y el de Máxima Verosimilitud. Efectivamente, podemos enunciar la siguiente propiedad, que se prueba en Lehmann (1983), página 341, aunque requiere de cierto dominio de los concepto de Invarianza (ver el siguiente apartado) y casi-invarianza (ver Lehmann (1983)): si $\{P_\theta : \theta \in \Theta\}$ es una familia de probabilidades sobre $\mathbb{R}^n$ dominada por la medida de Lebesgue, y G es un grupo de transformaciones dotado de una topología que lo hace localmente compacto, que actúa mediblemente sobre $(\mathbb{R}^n, \mathcal{R}^n)$ dejando invariantes tanto el experimento estadístico como la hipótesis inicial Θ_0, el estadístico RV es, si existe, igual, salvo un suceso nulo para toda la familia $\{P_\theta : \theta \in \Theta\}$, a otro invariante.

En consecuencia, si buscamos un test óptimo entre los invariantes o equivalentes a invariantes, el TRV es un firme candidato. Dado que la búsqueda del test UMP parte del Lema fundamental de Neyman-Pearson, no es de extrañar que sea el propio TRV el elegido. De hecho, así sucede en el Modelo Lineal Normal, según se demuestra en el capítulo 3. Los resultados allí obtenido se antojan bastante previsibles a la luz de las propiedades que acabamos de comentar.

Nótese, por último, que el test TRV y en general todos los tests que aparecerán en nuestra teoría, están compuestos por dos elementos: un estadístico denominado de contraste, RV en este caso, y un cuantil de cierta distribución, denominado valor teórico.

Invarianza y Contraste de Hipótesis

En esta sección vamos a estudiar los aspectos relativos al Principio de Invarianza que son fundamentales para la justificación del test `F`. Por lo tanto, consideraremos únicamente el problema de Contraste de Hipótesis. El Principio de Invarianza en

relación con el problema de Estimación se estudia, por ejemplo, en Arnold (1981) o Lehmann (1983).

Consideremos un experimento estadístico $(\Omega, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$ y un grupo G de transformaciones bimedibles de $(\Omega, \mathcal{A})$ en sí mismo. De esta forma, dado $\theta \in \Theta$, cada transformación $\mathtt{g} \in G$ induce de manera natural una probabilidad $P_\theta^{\mathtt{g}}$ sobre el espacio medible $(\Omega, \mathcal{A})$. En el conjunto Ω podemos establecer pues la siguiente relación: dados $\omega, \omega' \in \Omega$, decimos que $\omega \sim \omega'$ cuando existe $\mathtt{g} \in G$ tal que $\omega' = \mathtt{g}(\omega)$. Al ser G un grupo, esta relación es de equivalencia. Se denota por $[\omega]$ a la clase de equivalencia u órbita del elemento $\omega \in \Omega$.

Decimos que G deja invariante el experimento cuando, para toda transformación $\mathtt{g} \in G$, $\{P_\theta : \theta \in \Theta\} = \{P_\theta^{\mathtt{g}} : \theta \in \Theta\}$. En se caso, cada transformación $\mathtt{g} \in G$ induce una biyección $\overline{\mathtt{g}}$ de Θ en sí mismo, definida tal que $P_\theta^{\mathtt{g}} = P_{\overline{\mathtt{g}}(\theta)}$, para todo $\theta \in \Theta$. El conjunto de biyecciones $\overline{G} = \{\overline{\mathtt{g}} : \mathtt{g} \in G\}$ tiene, a su vez, estructura de grupo respecto de la operación composición, lo cual induce una partición del espacio de parámetros Θ en clases de equivalencia u órbitas.

Se dice que un estadístico T definido sobre $(\Omega, \mathcal{A}, \{P_\theta : \theta \in \theta\})$ y con valores en cualquier espacio medible $(\Omega', \mathcal{A}')$ es G-invariante cuando es constante sobre cada orbita de Ω, es decir, cuando $T \circ \mathtt{g} = T$, para todo $\mathtt{g} \in G$. Se dice G-invariante maximal cuando, además, toma valores distintos sobre órbitas distintas. En ese caso, será igual, salvo una biyección, a la proyección de Ω sobre el conjunto cociente $\Omega/\sim$. Se verifica entonces que, si M es un estadístico G-invariante maximal con valores en $(\Omega'', \mathcal{A}'')$ y T es un estadístico con valores en $(\Omega'', \mathcal{A}'')$, T es G-invariante si y sólo si existe una aplicación[30] h de Ω'' en Ω' tal que $T = h \circ M$.

De igual forma podemos hablar de aplicaciones $\overline{G}$-invariantes y $\overline{G}$-invariantes maximales en el espacio de parámetros Θ. Puede demostrarse fácilmente que si M y v son G-invariante maximal y $\overline{G}$-invariante maximal, respectivamente, se verifica, para cada par $\theta_1, \theta_2 \in \Theta$, la proposición $[v(\theta_1) = v(\theta_2)] \Rightarrow [P_{\theta_1}^M = P_{\theta_2}^M]$. Es decir, las distribuciones inducidas por un estadístico G-invariante maximal dependen del parámetro a través de cualquier aplicación $\overline{G}$-invariante maximal.

Si consideramos el problema de contrastar una hipótesis inicial, es decir, un subconjunto $\Theta_0 \subset \Theta$ frente a su alternativa, decimos que el grupo G deja invariante el problema de contraste de hipótesis cuando, para todo $g \in G$. $\overline{\mathtt{g}}(\Theta_0) = \Theta_0$. El Principio de Invarianza viene a proponer soluciones invariantes a problemas invariantes. Es decir, si ninguna transformación de G altera el experimento ni la hipótesis a contrastar, parece razonable solucionar el problema mediante un test que sea igualmente

[30] Si $(\Omega, \mathcal{A})$ y $(\Omega'', \mathcal{A}'')$ son espacios de Borel, podemos garantizar la medibilidad de h (ver Florens et al. (1990), secc. 8.2.2).

invariante. Dicho test será pues función de un estadístico invariante maximal. Por lo tanto, el primer objetivo será encontrar un invariante maximal respecto al grupo G de trasformaciones. El experimento imagen de dicho dicho estadístico, que puede entenderse como un paso al cociente, constituye lo que se denomina reducción por invarianza, y supondrá no sólo una simplificación en el espacio de observaciones sino también del de parámetros. De hecho, este último quedará reducido a la imagen de una aplicación $\overline{G}$-invariante maximal. De ello se deduce que, salvo en el caso trivial $G = \{\mathtt{Id}\}$, una reducción por invarianza no puede serlo a la vez por suficiencia, porque en el segundo caso no es posible una simplificación del espacio de parámetros. Así pues, la reducción por invarianza conlleva cierta pérdida de información, en el sentido de Fisher, pero se entiende que la información que se desecha no es relevante en el problema de contraste de hipótesis que se plantea.

No obstante, es lo más común, y así sucede en nuestra teoría, combinar ambos tipos de reducciones. El procedimiento estándar es empezar con una reducción por suficiencia, pues no implicará pérdida alguna de información. Si la simplificación no es satisfactoria, procederemos a reducir por invarianza. Decimos que éste es el procedimiento habitual aunque puede demostrarse que, en ciertas ocasiones, en particular en nuestra teoría, ambas reducciones pueden permutar. En todo caso, si se aplica una reducción por suficiencia seguida de otra por invarianza, es conveniente, en aras de una mayor coherencia estadística en la solución final, que exista cierta compatibilidad entre el estadístico suficiente y el grupo de transformaciones. Concretamente, decimos que un estadístico S definido sobre el experimento original y con valores en $(\Omega^S, \mathcal{A}^S)$ es G-equivariante cuando es sobreyectivo y verifica la proposición

$$[S(\omega) = S(\omega')] \;\Rightarrow\; [(S(\mathtt{g}(\omega)) = S(\mathtt{g}(\omega')),\ \forall \mathtt{g} \in G].$$

En ese caso, S induce un nuevo grupo de transformaciones $G^S = \{\mathtt{g}^S \colon \mathtt{g} \in G\}$ en el espacio de llegada, tal que, para cada $\mathtt{g} \in G$, $\mathtt{g}^S \circ S = S \circ \mathtt{g}$. Si suponemos que G deja invariante el problema de contraste de hipótesis $\Theta_0 \in \Theta$ y S es suficiente y G-equivariante, entonces G^S deja invariante el experimento $(\Omega^S, \mathcal{A}^S, \{P_\theta^S : \theta \in \Theta\})$ y el mismo problema de contraste de hipótesis planteado en el nuevo experimento.

A la hora de justificar el test `F` a nivel α en nuestra teoría, hubiera sido ideal que fuera UMP (uniformemente más potente) a nivel α. Ello no ha sido posible, de ahí que hayamos buscado un grupo que deja invariante tanto el experimento como el problema de contraste de hipótesis. El principio de invarianza propone pues considerar como únicas posibles soluciones a los test invariantes respecto a dicho grupo. Así pues, nuestro objetivo se reduce a encontrar un test UMP-invariante a nivel α, es decir, uniformemente más potente entre todos los invariantes a nivel α. Una reducción previa

mediante un estadístico suficiente y equivariante sirve únicamente para facilitar la búsqueda, ya que puede demostrarse[31] que, bajo cierta condiciones de regularidad que se satisfacen en nuestra teoría, se verifica que, si $\phi^S : (\Omega^S, \mathcal{A}^S) \longrightarrow [0,1]$ es un test UMP-G^S-invariante a nivel α, entonces $\phi^S \circ S$ es UMP-G-invariante a nivel α. Es decir, a partir de un tes UMP-invariante en el experimento imagen del estadístico suficiente obtenemos un test UMP-invariante en el experimento de partida.

En nuestra teoría podemos encontrar no una, sino tres reducciones sucesivas por invarianza respecto a sendos grupos de transformaciones distintas. Se trata en realidad de una única reducción respecto a un grupo mayor, que es suma de los anteriores. El hecho de distinguir varias, es decir, varios subgrupos, facilita, como veremos a continuación, la búsqueda de un invariante maximal para ese grupo más complejo. En lo que sigue, eludiremos, para abreviar, cualquier cuestión referida a la medibilidad de las aplicaciones[32].

Sea $\mathtt{G}$ el grupo de todas las biyecciones de Ω en Ω, y $G_1, \ldots, G_n$ subgrupos de $\mathtt{G}$ que dejan invariante el experimento. Supongamos además que se verifica la siguiente propiedad

$$\forall i, j \in \{1, \ldots, n\},\ \forall \mathtt{g}_i \in G_i,\ \forall \mathtt{g}_j \in G_j,\ \ \exists \mathtt{g}_j^* \in G_j\ :\ \mathtt{g}_i \circ \mathtt{g}_j = \mathtt{g}_j^* \circ \mathtt{g}_i. \tag{9.49}$$

Denótese por $G_1 \oplus \ldots \oplus G_n$ el menor subgrupo de $\mathtt{G}$ conteniendo a $G_1, \ldots, G_n$. Puede demostrarse fácilmente que

$$G_1 \oplus \ldots \oplus G_n = \left\{\mathtt{g}_1 \circ \ldots \circ \mathtt{g}_n : \mathtt{g}_i \in G_i,\ i = 1, \ldots, n\right\}.$$

Por la propiedad (9.49), puede considerarse cualquier permutación. Se verifica entonces que $(G_1 \oplus \ldots \oplus G_n)$ deja a su vez invariante el experimento y que cualquier aplicación $T : \Omega \longrightarrow \Omega'$ es $(G_1 \oplus \ldots \oplus G_n)$-invariante si y sólo si es G_i-invariante para todo $i = 1, \ldots, n$.

Veamos que una aplicación $(G_1 \oplus \ldots \oplus G_n)$-invariante maximal puede obtenerse por etapas: consideremos M_1 de Ω en Ω_1, G_1-invariante maximal y sobreyectivo. Definimos entonces, para cada $\mathtt{g}_2 \in G_2$, la aplicación $\mathtt{g}_2^1$ de Ω_1 en sí mismo tal que $\mathtt{g}_2^1 \circ M_1 = M_1 \circ \mathtt{g}_2$. Al verificarse (9.49), la transformación está bien definida. Además, se verifica que $(\mathtt{g}_1 \circ \mathtt{g}_2)^1 = \mathtt{g}_1^1 \circ \mathtt{g}_2^1$. Por lo tanto, el conjunto $G_2^1 = \{\mathtt{g}_2^1 : \mathtt{g}_2 \in G_2\}$ constituye un grupo de biyecciones de Ω_1 en sí mismo. Veamos entonces que, si $M_2^1 : \Omega_1 \longrightarrow \Omega_2$, es G_2^1-invariante maximal, entonces $M_2^1 \circ M_1$ es $(G_1 \oplus G_2)$-invariante maximal.

[31] Lehmann (1986), pp. 297-301.

[32] Estas cuestiones pendientes quedan totalmente resueltas introduciendo el concepto de estabilidad de un estadístico. En lo que respecta a nuestra teoría, las propiedades requeridas para garantizar la medibilidad se verifican trivialmente.

En efecto, se tiene que $M_2^1 M_1(\omega) = M_2^1 M_1(\omega')$ si y sólo si existe $g_2 \in G_2$ tal que $M_1(\omega') = g_2^1\,(M_1(\omega)) = M_1(g_2(\omega))$, lo cual equivale a afirmar que existe $g_1 \in G_1$ tal que $\omega' = g_1(g_2(\omega))$, es decir, que $\omega' = g(\omega)$ para algún $g \in G_1 \oplus G_n$, como queríamos probar.

En general, podemos considerar un proceso de este tipo:

$$\Omega \xrightarrow{M_1} \Omega_1 \xrightarrow{M_2^1} \Omega_2 \xrightarrow{M_3^{12}} \ldots \Omega_{n-1} \xrightarrow{M_n^{1\ldots n-1}} \Omega_n.$$

Por un razonamiento de tipo inductivo, podemos concluir que $M_n^{1\ldots n-1} \circ \ldots \circ M_1$ es $(G_1 \oplus \ldots \oplus G_n)$-invariante maximal. Si se considera un orden alternativo en los grupos, la afirmación sigue siendo válida, y todo lo dicho se hace extensivo al espacio de parámetros.

Parámetros muestrales y su interpretación.

Anteriormente se han definido los parámetros media y matriz de varianzas-covarianzas, correspondientes a un n-vector aleatorio sobre un espacio probabilístico. A continuación, procederemos a definir conceptos análogos desde el punto de vista estadístico, es decir, aplicaciones definidas sobre un experimento estadístico (que se denominan, como ya sabemos, estadísticos) que funcionarán como estimadores de los primeros. Estos parámetros se denominarán muestrales pues se construirán a partir de los valores de la muestra seleccionada. El término *parámetro* suele reservarse en Estadística para referirnos a números propios de las distribuciones de probabilidad, como la media, varianza, etc. Desde ese punto de vista, se incurriría en un error al denominar parámetro a un número que se calcula a partir de una muestra de datos, como pueden ser a media muestral o la varianza muestral. Hemos de tener en cuenta, no obstante, lo siguiente.

El espacio de observaciones del experimento estadístico dado por la muestra es $\mathbb{R}^n$, donde n denota el número de unidades experimentales que se manejan. En un marco multivariante, es decir, cuando se estudian no una sino p variables sobre la población considerada, el espacio se observaciones será $\mathbb{R}^{np}$, es decir, $\mathcal{M}_{n\times p}$. Veamos porqué: en la práctica, el proceso estadístico consiste en seleccionar n unidades experimentales $\omega_1, \ldots, \omega_n$ en la población, en cada una de las cuales se mide la o las variables consideradas. De la observación de una variable y, definida sobre la población estudiada, en la n-upla de unidades experimentales resulta como muestra una n-upla de número reales $y_1, \ldots, y_n$, es decir, un vector de $\mathbb{R}^n$. En general, si se observan p variables $y[1], \ldots, y[p]$, obtenemos como muestra una matriz de dimensiones $n \times p$ como la que

aparece a continuación

$$\mathsf{Y} = \begin{pmatrix} y_1[1] & \dots & y_1[p] \\ \vdots & & \vdots \\ y_n[1] & \dots & y_n[p] \end{pmatrix}.$$

El término entre corchetes hace referencia a la variable y el subíndice a la unidad experimental sobre la que se mide. Para todo $j = 1, \dots, p$, $\mathsf{Y}[j]$ denotará el vector columna j-ésimo, es decir, el resultante de medir la variable j-ésima en las n unidades experimentales.

Los parámetros probabilísticos han sido definidos en el contexto del espacio de Hilbert L^2 de las variables aleatorias de cuadrado integrable definidas sobre el espacio de probabilidad en cuestión. Los muestrales se definirán en el marco del espacio, también de Hilbert, $\mathbb{R}^n$. La conexión entre ambos espacios, tan diferentes a simple vista, es la siguiente.

Una vez llevado a cabo el experimento, es decir, cuando se posee la n-upla de unidades experimentales y el vector de datos, matriz en el caso multivariante, que resultan de la medición de la o las variables sobre los mismos, podemos tomar el conjunto $\Omega_n = \{\omega, \dots, \omega_n\}$ dotado de la sigma-álgebra de sus partes, $\mathcal{P}(\Omega_n)$. Sobre este espacio de medida consideramos la probabilidad uniforme P_n, la cual induce una distribución en $\mathbb{R}$ si estudiamos una única variable, y sobre $\mathbb{R}^p$ en el caso multivariante, que asigna probabilidad n^{-1} a cada número (respectivamente vector) de la muestra. Dicha distribución se denomina distribución empírica. Tenemos pues un espacio tipo L^2, concretamente las funciones de cuadrado integrable sobre $(\Omega_n, \mathcal{P}(\Omega_n), P_n)$, dotado de un producto interior que se define como sigue: dadas dos variables $y[j]$ e $y[k]$,

$$\begin{aligned} \langle y[j], y[k] \rangle_{L^2} &= \int y[j] \cdot y[k] \, dP_n = \sum_{i=1}^{n} y[j](\omega_i) \cdot y[k](\omega_i) \cdot P_n(\omega_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} y_i[j] y_i[k] \propto \langle Y[j], Y[k] \rangle_{\mathbb{R}^n}. \end{aligned}$$

Por lo tanto, se trata, salvo la constante n^{-1}, del producto escalar en $\mathbb{R}^n$ entre los vectores asociados. Los parámetros muestrales se definen como los parámetros probabilísticos correspondientes a dicha distribución. Por ello, los parámetros que se estudian es este apartado pueden considerarse casos particulares de (9.20), (9.21), (9.22), (9.23) (9.14), (9.23) y (9.15). Los parámetros probabilísticos definidos anteriormente pueden interpretarse en términos del producto interior. Así pues, como caso particular, los parámetros muestrales se interpretarán en términos del producto escalar (y, por lo tanto, de la norma euclídea) en $\mathbb{R}^n$.

En definitiva, estamos trabajando paralelamente con dos espacios tipo L^2 aunque asociados a distintos espacios probabilísticos: $(\Omega, \mathcal{A}, \mathcal{P})$, que corresponde al fenómeno aleatorio estudiado, y $(\Omega_n, \mathcal{P}_n, P_n)$, que corresponde a una muestra del mismo. La Teoría de la Probabilidad recoge bajo la denominación común de Leyes de los Grandes Números diversos resultados (ver, por ejemplo los teoremas 9.22, 9.23 y 9.24) que nos hablan de la convergencia de parámetros muestrales a sus análogos probabilísticos, a medida que el tamaño de muestra n tiende a infinito y siempre y cuando ésta sea aleatoria simple, es decir, una secuencia independiente e idénticamente distribuida según el modelo probabilístico P. También el teorema 9.25 nos ayuda a entender la relación existente entre los productos interiores (9.7) y (9.16) cuando de muestras aleatorias simples se trata. En definitiva, podríamos habar en términos heurísticos de una aproximación del espacio de Hilbert $\mathbb{R}^n$, asociado a las muestra aleatorias simples de tamaño n, al espacio de Hilbert L^2 asociado a las variables aleatorias.

Visto esto, parece claro que lo que sigue debería omitirse si pretendiéramos ser concisos en este punto, pero no es el caso. Esta visión de los parámetros muestrales como casos particulares de los probabilísticos puede resultar interesante desde el punto de vista formal, pero la intuición parece ir en sentido contrario. Seguramente resulte más comprensible definir estos parámetros en $\mathbb{R}^n$ e interpretarlos mediante las nociones usuales de perpendicularidad y distancia entre vectores, para después extender el concepto a espacios de probabilidad abstractos. En todo caso, el objetivo de este apartado es dejar patente el paralelismo entre los parámetros muestrales y sus análogos probabilísticos, tanto en la definición como en la interpretación, de ahí que las propiedades de unos sean automáticamente heredadas por los otros. Así pues, se definen la media, varianza y covarianza muestrales mediante

$$\begin{aligned}
\overline{\mathsf{Y}[j]} &= \frac{1}{n}\sum_{i=1}^{n} y_i[j], \quad j = 1, \ldots, p. \\
s^2_{\mathsf{Y}[j]} &= \frac{1}{n}\sum_{i=1}^{n} (y_i[j] - \overline{\mathsf{Y}[j]})^2, \quad j = 1, \ldots, p. \\
s_{\mathsf{Y}[j],\mathsf{Y}[k]} &= \frac{1}{n}\sum_{i=1}^{n} (y_i[j] - \overline{\mathsf{Y}[j]})(y_i[k] - \overline{\mathsf{Y}[k]}), \quad j, k = 1, \ldots, p,
\end{aligned}$$

respectivamente. Nótese que $s^2_{\mathsf{Y}[j]} = s_{\mathsf{Y}[j],\mathsf{Y}[j]}$. Además, de la Desigualdad de Cauchy-Swartz se sigue que

$$|s_{\mathsf{Y}[j],\mathsf{Y}[k]}| \leq s_{\mathsf{Y}[j]} s_{\mathsf{Y}[k]},$$

lo cual induce a definir, para todo par $j, k = 1, \ldots, p$, el parámetro

$$r_{jk} = \frac{s_{\mathsf{Y}[j],\mathsf{Y}[k]}}{s_{\mathsf{Y}[j]} s_{\mathsf{Y}[k]}}, \tag{9.50}$$

denominado coeficiente de correlación muestral simple, que tomará un valor entre -1 y 1. Es muy frecuente que, por razones que atañen al problema de Estimación, las varianza y covarianzas aparezcan divididas por $n-1$ en lugar de n. Por el momento, pasaremos por alto ese detalle. Con las medias podemos componer un vector media muestral $\overline{\mathsf{y}} = (\overline{\mathsf{Y}[1]}, \ldots, \overline{\mathsf{Y}[p]})'$. Si este vector (columna) se repite p veces tenemos una matriz $n \times p$ que se denota por $\overline{\mathsf{Y}}$. Igualmente, tanto las varianzas y covarianzas como las correlaciones componen sendas matrices denominadas matriz de varianzas-covarianzas total muestral y matriz de correlaciones, respectivamente, que se denotará mediante S_{Y} y R_{Y}. La matriz de varianzas-covarianzas totales muestral puede expresarse mediante

$$S_{\mathsf{Y}} = \frac{1}{n}(\mathsf{Y} - \overline{\mathsf{Y}})'(\mathsf{Y} - \overline{\mathsf{Y}})$$

En ocasiones se utilizará la notación S_{YY} con el objeto de evitar confusiones. Es fácil demostrar que, para cada, $j = 1, \ldots, p$,

$$\overline{\mathsf{Y}[j]} \cdot 1_n = P_{\langle 1_n \rangle} \mathsf{Y}[j],$$

donde 1_n denota el vector de $\mathbb{R}^n$ cuyas componentes son todas igual a 1 [33]. Es decir, el vector de $\mathbb{R}^n$ de componentes iguales cuya distancia euclídea a $\mathsf{Y}[j]$ sea mínima es su media muestral repetida n veces. Nótese que, en una situación determinista, el vector $\mathsf{Y}[j]$ sería constante, es decir, pertenecería al subespacio $\langle 1_n \rangle$. Eso se correspondería con una variabilidad total de $\mathsf{Y}[j]$ nula, entendiendo por variabilidad total de $\mathsf{Y}[j]$ el vector $\mathsf{Y}[j] - \overline{\mathsf{Y}[j]}1_n$. De esta forma, la discrepancia entre la situación real y la que correspondería a un modelo determinista, puede cuantificarse, de alguna forma, mediante la distancia euclídea entre $\mathsf{Y}[j]$ y $\overline{\mathsf{Y}}[j]1_n$, es decir, mediante la norma euclídea de la variabilidad total, y eso es precisamente lo que se mide con la varianza muestral[34].

$$\begin{aligned} s^2_{\mathsf{Y}[j]} &= \frac{1}{n}\|\mathsf{Y}[j] - \overline{\mathsf{Y}[j]}1_n\|^2 = \frac{1}{n}\|\mathsf{Y}[j] - P_{\langle 1_n \rangle}\mathsf{Y}[j]\|^2 \\ &= \frac{1}{n}\|P_{\langle 1_n \rangle^\perp}\mathsf{Y}[j]\|^2 = \frac{1}{n}\mathsf{Y}[j]' P_{\langle 1_n \rangle^\perp}\mathsf{Y}[j], \quad j = 1, \ldots, p. \end{aligned}$$

La última expresión puede utilizarse en el caso multivariante y, de hecho, la matriz de varianzas-covarianzas totales puede expresarse mediante

$$S_{\mathsf{Y}} = \frac{1}{n}\mathsf{Y}' P_{\langle 1_n \rangle^\perp}\mathsf{Y}. \tag{9.51}$$

[33] Nótese la similitud con la media probabilistica, que determina la proyección ortogonal de una variable aleatoria sobre el subespacio de las funciones constantes.

[34] Nótese, de nuevo, la analogía con la varianza poblacional, que expresa la distancia al cuadrado entre una variable y su proyección sobre el subespacio de funciones constantes.

Si definimos la variabilidad total de la matriz Y mediante $\mathsf{Y}_0 = \mathsf{Y} - \overline{\mathsf{Y}}$, es decir, $P_{\langle 1_n \rangle^\perp}\mathsf{Y}$, la matriz anterior puede expresarse también de la forma

$$S_{\mathsf{Y}} = \frac{1}{n}\mathsf{Y}_0'\mathsf{Y}_0 \tag{9.52}$$

Si B, C, son matrices de orden $p \times q$ y $n \times q$, puede comprobarse, trivialmente el análogo muestral de la proposición 9.11

$$\overline{\mathsf{Y}B + C} = \overline{\mathsf{Y}}B + C, \quad S_{\mathsf{Y}B+C} = B'S_{\mathsf{Y}}B. \tag{9.53}$$

Cuando se distinguen dos grupos de variables y y z, de forma que la matriz de datos se expresa mediante YZ, la matriz de covarianzas totales queda dividida, naturalmente, en cuatro partes, que se denotarán de la siguiente forma

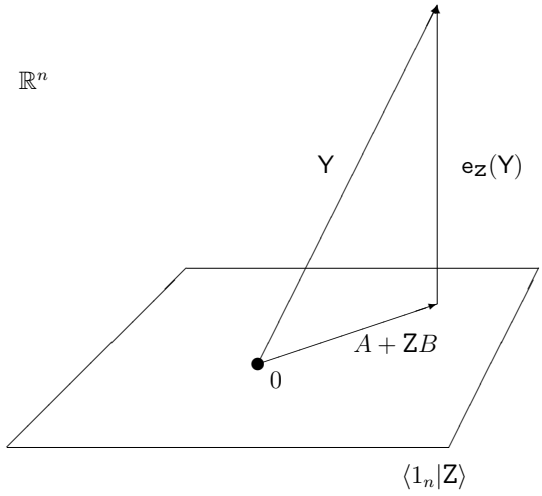
$$S_{(\mathsf{YZ})(\mathsf{YZ})} = \begin{pmatrix} S_{\mathsf{YY}} & S_{\mathsf{YZ}} \\ S_{\mathsf{ZY}} & S_{\mathsf{ZZ}} \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \mathsf{Y}_0'\mathsf{Y}_0 & \mathsf{Y}_0'\mathsf{Z}_0 \\ \mathsf{Z}_0'\mathsf{Y}_0 & \mathsf{Z}_0'\mathsf{Z}_0 \end{pmatrix}. \tag{9.54}$$

Tanto Y como Z presentarán cierta variabilidad total, pero la variabilidad de Z, que se denota por Z_0, podría explicar buena parte de la variabilidad total de Y, que se denota por Y_0, si esta última fuera, aproximadamente, una combinación lineal de la primera, lo cual equivaldría a que Y fuese una combinación afín (con término independiente) de Z . El caso ideal se daría cuando Y_0 perteneciera a $\langle \mathsf{Z}_0 \rangle$ [35] o, equivalentemente, cuando $\mathsf{Y}_0 - P_{\langle \mathsf{Z}_0 \rangle}\mathsf{Y}_0$ fuese nulo. Puede comprobarse fácilmente que este último término es igual a $\mathsf{Y} - P_{\langle 1_n \mathsf{Z} \rangle}\mathsf{Y}$, es decir, a $P_{\langle 1_n \mathsf{Z} \rangle^\perp}\mathsf{Y}$, que pertenece a $\langle 1_n \rangle^\perp$. En lo sucesivo, dicho término se denominará residuo de Y dado Z, denotándose

$$\mathsf{e}_{\mathsf{Z}}(\mathsf{Y}) = \mathsf{Y} - P_{\langle 1_n \mathsf{Z} \rangle}\mathsf{Y} \tag{9.55}$$

En el capítulo 4 se denotará abreviadamente por e. Veamos una ilustración correspondiente al caso $p = 1$.

[35] Entendemos que una matriz pertenece a cierto subespacio vectorial cuando cada una uno de sus vectores columna pertenecen a dicho subespacio.



Los valores de A y B se calcularán a continuación buscando la ortogonalidad de la descomposición. En definitiva, hemos afirmado que la variabilidad total de Y es explicada por la de Z mediante una relación lineal entre ambas cuando el residuo es constante (en cuyo caso sería nulo). Luego, razonando como antes, expresaremos la discrepancia entre esta situación y la que se da en la realidad mediante la matriz de varianzas-covarianzas totales de $\mathsf{e_Z(Y)}$, que se denominará a partir de ahora matriz de varianzas-covarianzas parciales de Y dada Z, denotándose por $S_{\mathsf{Y}\cdot\mathsf{Z}}$. Así pues,

$$\begin{aligned} S_{\mathsf{Y}\cdot\mathsf{Z}} &:= S_{\mathsf{e_Z(Y)}} && (9.56)\\ &= \frac{1}{n}\mathsf{e_Z(Y)}' P_{\langle 1_n\rangle^\perp}\mathsf{e_Z(Y)} && (9.57)\\ &= \frac{1}{n}\mathsf{e_Z(Y)}'\mathsf{e_Z(Y)}. && (9.58)\end{aligned}$$

Es obvio que esta matriz es invariante ante traslaciones[36]. Dado que Y descompone en suma ortogonal de $\mathsf{e_Z(Y)}$ y $P_{\langle 1_n\mathsf{Z}\rangle}\mathsf{Y}$, se deduce de (9.51) que

$$S_{\mathsf{Y}} = S_{\mathsf{Y}\cdot\mathsf{Z}} + S_{P_{\langle 1_n\mathsf{Z}\rangle}\mathsf{Y}}.$$

Por lo tanto, $S_{\mathsf{Y}\cdot\mathsf{Z}}$ es menor o igual que S_{Y} según el preorden (9.4). En particular, los elementos de la diagonal, denominados varianzas parciales, son menores o iguales

[36] Es decir, que el sumar una constante a cada componente de una columna de Y o Z no afecta al cálculo de $S_{\mathsf{Y}\cdot\mathsf{Z}}$.

que las correspondientes varianzas totales, cosa lógica dado que las primeras vienen a expresar la parte de la variabilidad total de Y no explicada por la la variabilidad total de Z mediante una relación lineal. Por otra parte, el término $P_{\langle 1_n \mathsf{Z}\rangle}\mathsf{Y}$ se denotará en el capítulo 4 mediante $\hat{\mathsf{Y}}$. Puede descomponerse trivialmente en $\overline{\mathsf{Y}} + P_{\langle \mathsf{Z}_0\rangle}\mathsf{Y}_0$. El último término es, por definición, la única matriz $n \times p$ de la forma $\mathsf{Z}_0 B$, con $B \in \mathcal{M}_{q\times p}$, tal que las columnas de $\mathsf{Y}_0 - \mathsf{Z}_0 B$ son perpendiculares a las de Z_0, es decir, tal que

$$(\mathsf{Y}_0 - \mathsf{Z}_0 B)'\mathsf{Z}_0 = 0$$

En ese caso, se obtiene trivialmente

$$B = S_{\mathsf{ZZ}}^{-1} S_{\mathsf{ZY}} \tag{9.59}$$

Por lo tanto, si se denota

$$A = \overline{\mathsf{Y}} - \overline{\mathsf{Z}} B, \tag{9.60}$$

se tiene que

$$P_{\langle 1_n \mathsf{Z}\rangle}\mathsf{Y} = A + \mathsf{Z}B \tag{9.61}$$

Nótese que los razonamientos utilizados son completamente análogos a los considerados para obtener (9.25) y (9.26) en el espacio L^2 de las variables aleatorias de cuadrado integrables. La matriz $S_{\hat{\mathsf{Y}}}$ viene a expresar la parte de la variabilidad total de Y que sí es explicada por la variabilidad total de Z mediante una relación lineal, y puede calcularse como sigue

$$\begin{aligned} S_{\hat{\mathsf{Y}}} &= S_{\overline{\mathsf{Y}}+\mathsf{Z}_0 B} = S_{\mathsf{Z}_0 B} = \frac{1}{n} B'\mathsf{Z}_0 P'_{\langle 1_n\rangle^\perp}\mathsf{Z}_0 B \\ &= \frac{1}{n} B'\mathsf{Z}_0'\mathsf{Z}_0 B = S_{\mathsf{YZ}} S_{\mathsf{ZZ}}^{-1} S_{\mathsf{ZZ}} S_{\mathsf{ZZ}}^{-1} S_{\mathsf{ZY}} \\ &= S_{\mathsf{YZ}} S_{\mathsf{ZZ}}^{-1} S_{\mathsf{ZY}}. \end{aligned}$$

En definitiva, la matriz de varianzas-covarianzas parciales puede expresarse a través (9.54) de mediante

$$S_{\mathsf{Y}\cdot\mathsf{Z}} = S_{\mathsf{YY}} - S_{\mathsf{YZ}} S_{\mathsf{ZZ}}^{-1} S_{\mathsf{ZY}}. \tag{9.62}$$

A partir de esta matriz de varianzas-covarianzas podemos construir una matriz de correlaciones denominadas parciales, que serán invariantes ante traslaciones y homotecias. Así, si Y posee dos columnas Y_1 y Y_2 se define el coeficiente de correlación parcial entre Y_1 e Y_2 dado Z mediante

$$r_{\mathsf{Y}_1,\mathsf{Y}_2\bullet\mathsf{Z}} = \frac{s_{\mathsf{e}_\mathsf{Z}(\mathsf{Y}_1),\mathsf{e}_\mathsf{Z}(\mathsf{Y}_2)}}{s_{\mathsf{e}_\mathsf{Z}(\mathsf{Y}_1)} s_{\mathsf{e}_\mathsf{Z}(\mathsf{Y}_2)}} = \frac{\langle \mathsf{e}_\mathsf{Z}(\mathsf{Y}_1), \mathsf{e}_\mathsf{Z}(\mathsf{Y}_2)\rangle}{\|\mathsf{e}_\mathsf{Z}(\mathsf{Y}_1)\| \cdot \|\mathsf{e}_\mathsf{Z}(\mathsf{Y}_2)\|}. \tag{9.63}$$

Para obtener a última expresión basta considerar (9.55). Cuando y es unitario, es decir, cuando Y es un vector de $\mathbb{R}^n$, cabe habar únicamente de la varianza parcial de Y dado Z, que es número positivo

$$s^2_{\mathsf{Y}\cdot\mathsf{Z}} = s^2_{\mathsf{Y}} - S_{\mathsf{YZ}}S^{-1}_{\mathsf{ZZ}}S_{\mathsf{ZY}} \tag{9.64}$$

que, insistimos, se corresponde con la parte de la variabilidad total de y que no es explicada linealmente por la variabilidad total de Z. Teniendo en cuenta (9.58), puede expresarse mediante

$$s^2_{\mathsf{Y}\cdot\mathsf{Z}} = \frac{1}{n}\|\mathsf{e}_{\mathsf{z}}(\mathsf{Y})\|^2 \tag{9.65}$$

Todo ello invita a definir el coeficiente de correlación múltiple muestral de y respecto a z mediante

$$R^2_{\mathsf{Y},\mathsf{Z}} = \frac{S_{\mathsf{YZ}}S^{-1}_{\mathsf{ZZ}}S_{\mathsf{ZY}}}{s^2_{\mathsf{Y}}},$$

de tal forma que

$$\frac{s^2_{\mathsf{Y}\cdot\mathsf{Z}}}{s^2_{\mathsf{Y}}} = 1 - R^2_{\mathsf{Y},\mathsf{Z}}.$$

Por lo tanto, el coeficiente de correlación múltiple expresa la proporción de la variabilidad total de Y explicada linealmente por la variabilidad total de Z. El caso $R^2_{\mathsf{Y},\mathsf{Z}} = 1$ equivale a una relación afín perfecta entre el vector Y y la matriz Z. Si z también es unitario, el coeficiente de correlación múltiple muestral de y respecto a z es igual al cuadrado del coeficiente de correlación simple definido en (9.50). Obviamente, el coeficiente de correlación múltiple es invariante ante traslaciones y homotecias.

9.4. Algunos elementos de Teoría Asintótica.

En esta sección repasamos algunos conceptos y resultados límite fundamentales (en el sentido de importantes, no de triviales) de la Teoría de la Probabilidad y de la Estadística en general, de los que se hace uso en los capítulos 3 y 8. En primer lugar definiremos las convergencias en probabilidad, en distribución y casi seguro, y estableceremos relaciones entre las mismas. Recomendamos consultar, en todo caso, bibliografía complementaria, por ejemplo Billingsley (1986), Fergusson (1996) o Lehmann (1998).

Dados X y $(X_n)_{n\in\mathbb{N}}$, vector aleatorio y sucesión de vectores aleatorios, respectivamente, definidos sobre un espacio de probabilidad $(\Omega, \mathcal{A}, P)$ con valores en $\mathbb{R}^m$, se dice que $(X_n)_{n\in\mathbb{N}}$ converge en casi seguro a X cuando existe un suceso $N \in \mathcal{A}$ P-nulo, tal que la sucesión $\big(X_n(\omega)\big)_{n\in\mathbb{N}}$ converge a $X(\omega)$, para todo ω en el complemtario de

N, en cuyo caso, se denota $X_n \xrightarrow{c.s.} X$. Se dice que $(X_n)_{n\in\mathbb{N}}$ converge en probabilidad a X cuando la sucesión $\big(P(\|X_n - X\| > \varepsilon)\big)_{n\in\mathbb{N}}$ converge a 0 para todo $\varepsilon > 0$, en cuyo caso se denota $X_n \xrightarrow{P} X$.

Dadas P_0 y $(P_n)_{n\in\mathbb{N}}$, distribución y sucesión de distribuciones, respectivamente, sobre $\mathbb{R}^m$, decimos que $(P_n)_{n\in\mathbb{N}}$ converge en distribución a P cuando $\mathtt{E}_{P_n}[f]$ converge a $\mathtt{E}_{P_0}[f]$, para toda función f de $\mathbb{R}^m$ en $\mathbb{R}$ medible, continua y acotada[37]. En ese caso, se denota $P_n \xrightarrow{d} P_0$. Si X y $(X_n)_{n\in\mathbb{N}}$ son como antes, se dice que $(X_n)_{n\in\mathbb{N}}$ converge en distribución a X cuando $P^{X_n} \xrightarrow{d} P^X$, denotándose $X_n \xrightarrow{d} X$.

Teorema 9.21. (i) La convergencia en distribución equivale a la convergencia de las respectivas funciones características en todo punto de $\mathbb{R}^m$.

(ii) Si $m = 1$, la convergencia en distribución de $(X_n)_{n\in\mathbb{N}}$ a X equivale a la convergencia de las respectivas funciones de distribución F_n a la función de distribución de F de X en cada punto de continuidad de esta última. En ese caso, si, además, son continuas F y F_n, para cada $n \in \mathbb{N}$ [38], se da también una convergencia entre las funciones inversas.

(iii) La convergencia casi-seguro implica convergencia en probabilidad.

(iv) La convergencia en probabilidad implica convergencia en distribución.

(v) Si dos sucesiones de variables aleatorias convergen en probabilidad a sendas constantes, las sucesiones de las sumas y productos convergen, respectivamente, a la suma y producto de dichas constantes.

(vi) La convergencia en distribución a una constante implica convergencia en probabilidad.

(vii) Si $f \in \mathcal{C}(\mathbb{R}^p)$ y $(X_n)_{n\in\mathbb{N}}$ converge en distribución a X, $(f(X_n))_{n\in\mathbb{N}}$ converge en distribución a $f(X)$.

(viii) Si f es continua en a y $(X_n)_{n\in\mathbb{N}}$ converge en distribución a una constante a, $(f(X_n))_{n\in\mathbb{N}}$ converge en distribución a $f(a)$.

[37] Si consideramos medidas finitas con signo sobre un compacto K de $\mathbb{R}^m$, en virtud del Teorema de Representación de Riesz (Ash (1972), Th. 4.3.13), podemos identificar dichas medidas con los las funciones reales, lineales y continuas sobre $\mathcal{C}(K)$, y la convergencia en distribución con la convergencia respecto a la topología débil*.

[38] En ese caso podemos hablar de la inversas de cada una de ellas

(ix) Si $(X_n)_{n\in\mathbb{N}}$, $(U_n)_{n\in\mathbb{N}}$ y $(V_n)_{n\in\mathbb{N}}$ convergen en distribución a X, a (cte.) y 1, respectivamente,

(a) $X_n + U_n \xrightarrow{d} X + a$.

(b) $X_n \cdot U_n \xrightarrow{d} aX$

(c) $\frac{X_n}{V_n} \xrightarrow{d} X$

(x) [Astucia de Cramer-Wold] $(X_n)_{n\in\mathbb{N}}$ converge en distribución a $N_p(\theta, \Sigma)$ si y sólo si $(c'X_n)_{n\in\mathbb{N}}$ converge en distribución a $N(c'\mu, c'\Sigma c)$, para todo $c \in \mathbb{R}^p$ tal que $\|c\| = 1$.

(xi) Si $(P_n)_n$ y $(Q_n)_n$, secuencias de probabilidades sobre $\mathbb{R}^{m_1}$ y $\mathbb{R}^{m_2}$, convergen en distribución a P y Q, respectivamente, $(P_n \times Q_n)_n$ converge en distribución a $P \times Q$.

El siguiente resultado es consecuencia directa de la desigualdad de Chebyshev.

Teorema 9.22.
Si $(X_n)_{n\in\mathbb{N}}$ es una sucesión de variables aleatorias reales con varianzas uniformemente acotadas por cierta constante $M > 0$, se verifica que

$$\frac{\sum_{i=1}^{n}(X_n - \mathtt{E}[X_n])}{n} \xrightarrow{P} 0.$$

En el caso de que las variables X_n, $n \in \mathbb{N}$, sean independientes e idénticamente distribuidas (iid), con varianza finita y media μ, podemos deducir que la media muestral $\overline{X}_n$ converge en probabilidad a μ. Puede probarse que la convergencia en probabilidad a la media se sigue verificando aunque la distribución no sea de cuadrado integrable. Este resultado suele denominarse Ley Débil de los Grandes Números (LDGN). El siguiente, denominado Ley Fuerte de los Grandes Números (LFGN), caso iid, va aún más lejos.

Teorema 9.23.
Si $(X_n)_{n\in\mathbb{N}}$ es una sucesión de variables aleatorias reales iid con media $\mu \in [-\infty, +\infty]$, se verifica que

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{c.s.} \mu.$$

El resultado anterior desempeña un papel crucial en la Teoría de Probabilidad. Podemos encontrar su demostración en Billigsley (1986). A partir del mismo podemos probar el Teorema de Glivenko-Cantelli[39] que garantiza, en el caso unidimensional, la

[39] Ver Nogales (1998).

convergencia uniforme de la función de distribución empírica de una muestra aleatoria simple de tamaño n a la función de distribución probabilística F. Algunos autores lo denominan Teorema Fundamental de la Estadística Matemática, con lo que queda patente su trascendencia.

Por otra parte y teniendo en cuenta que la media es el momento de orden 1 de la distribución, no preguntamos si el resultado sigue siendo válido para un momento de cualquier orden. La respuesta, afirmativa, la encontramos en el siguiente resultado, conocido como método de los momentos[40]

Teorema 9.24.
Si $(X_n)_{n\in\mathbb{N}}$ es una sucesión de variables aleatorias reales iid correspondientes a una distribución con momento de orden k finito, se verifica que

$$M_n^j = \frac{1}{n}\sum_{i=1}^{n} X_i^j \xrightarrow{c.s.} \mathtt{E}[X_1^j], \quad \forall j = 1, \ldots, k.$$

Además, si h es una función continua de $\mathbb{R}^k$ en $\mathbb{R}$, la secuencia $h\circ(M_n^1, \ldots, M_n^k)$ converge casi seguro a $h \circ (\mathtt{E}[X_1], \ldots, \mathtt{E}[X^k])$. En particular, si la distribución es de cuadrado integrable, la varianza muestral definida en la sección 1.4 converge casi seguro a la varianza de la distribución.

La convergencia casi segura se verifica también, en el caso multidimensional, para la media del producto de las componentes, siempre y cuando ambas sean de cuadrado integrable. La demostración es inmediata partiendo del teorema 9.23.

Teorema 9.25.
Si $(X_n, Y_n)_{n\in\mathbb{N}}$ es una sucesión de vectores aleatorios bidimensionales iid correspondientes a una distribución cuyas componentes son de cuadrado integrable, se verifica que

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i \xrightarrow{c.s.} \mathtt{E}[X_1 Y_1].$$

Como corolario inmediato se obtiene la convergencia c.s. de la covarianza muestral a la covarianza probabilística y, en consecuencia, del coeficiente de correlación.

También desempeña un papel crucial el siguiente resultado, denominado Teorema Central del Límite (TCL). Como ya sabemos, esta denominación no hace referencia a un único teorema sino a una colección de resultados, teniendo todo ellos en común la convergencia la la distribución normal de ciertas sumas de variables. El siguiente

[40]Realmente, el método de los momentos más general. Su enunciado y demostración pueden encontrarse en Nogales (1998).

teorema puede considerarse consecuencia directa de la versión de Lindemberg-Feller, que podemos encontrar demostrada, por ejemplo, en Billigsley (1986) o en Ash (1972).

Teorema 9.26.
Consideremos una colección de variables aleatorias reales $\{X_{ni} : (n,i) \in \mathbb{N} \times \{1,\ldots,n\}$, tal que, para cada $n \in \mathbb{N}$, se verifica que $\mathtt{E}[X_{ni}] = 0$ y $\mathtt{var}[X_{ni}] = \sigma_{ni}^2$, siendo las X_{ni} independientes para $i = 1,\ldots,n$. Denótese, para cada $n \in \mathbb{N}$, $\tau_n^2 = \sum_{i=1}^n \sigma_{ni}^2$. Supongamos además que, para todo $\varepsilon > 0$, se verifica

$$\lim_{n\to\infty} \frac{1}{\tau_n^2} \sum_{i=1}^n \mathtt{E}\big[X_{ni}^2 I_\varepsilon(X_{ni})\big] = 0. \ ^{41} \tag{9.66}$$

Entonces,

$$\frac{1}{\tau_n} \sum_{i=1}^n X_{ni} \xrightarrow{d} N(0,1).$$

Como caso particular, si $(X_i)_{i\in\mathbb{N}}$ es una sucesión de variables aleatorias iid con media μ y varianza σ^2, se verifica que

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0,\sigma^2). \tag{9.67}$$

La denominada astucia de Cramer-Wold permite extender este resultado al caso multivariante (ver Nogales (1998)). Otro resultado de enorme interés, cuya demostración podemos encontrar en Arnold (1981), pag. 152, es el siguiente:

Teorema 9.27.
Si $\sqrt{n}(T_n - a) \xrightarrow{d} N(0,b^2)$ y δ es una función real diferenciable, entonces

$$\sqrt{n}(\delta(T_n) - \delta(a)) \xrightarrow{d} N(0,(\delta'(a))^2 b^2).$$

Este teorema permite establecer un procedimiento, que suele denominarse método Delta, que de alguna manera viene a complementar el TCL en aquellos casos en los cuales la distribución asintótica obtenida depende del parámetro estudiado.

Un concepto propio de la teoría asintótica, importante a la hora de justificar un estimador, es el de consistencia. Para poder formular su definición así como otras propiedades relativas al problema de contraste de hipótesis, es necesario establecer primeramente un marco formal apropiado. Consideremos un experimento estadístico $(\Omega^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}}, \{P_\theta : \theta \in \Theta\})$. Nótese que P_θ denota una probabilidad sobre el espacio

[41] I_ε denota la función que toma valor 0 en $[0,\varepsilon]$ y 1 en $(\varepsilon,+\infty)$.

producto $(\Omega^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$. Como ya hemos comentado, el Teorema de Existencia de Kolmogorov[42] permite construir un experimento de este tipo a partir de experimentos en dimensión n, siempre que se dé la adecuada *consistencia*. De hecho, será así como procedamos en la sección dedicada al estudio asintótico del Modelo Lineal. No queremos decir que las probabilidades de la familia sean de la forma $\mathsf{P}_\theta^{\mathbb{N}}$, para cierta distribución real P_θ. Si ello sucediera, estaríamos considerando una muestra aleatoria simple de tamaño infinito de dicha distribución. Ese caso particular (caso iid) resulta de más fácil manejo y permite, entre otras cosas, deducir la consistencia y normalidad asintótica de EMV [43]. Dado que la matriz de varianzas-covarianzas de dicha distribución normal límite es la inversa de la matriz de Información, alcanza la cota óptima de Cramer-Rao, lo que se traduce en la *eficiencia asintótica* del EMV[44].

Dado un estimando g sobre Θ, que supondremos con valores en $\mathbb{R}^m$, una secuencia de estimadores $(T_n)_{n\in\mathbb{N}}$ de g será una sucesión de estadísticos sobre el experimento anterior y con valores en $\mathbb{R}^m$ tales que, para todo $n \in \mathbb{N}$, existe una función medible $\tilde{T}_n$, definida sobre $(\Omega^n, \mathcal{A}^n)$, verificando $T_n = \tilde{T}_n \circ \pi_n$, donde π_n denota la proyección natural sobre Ω^n. En los mismos términos podemos hablar de una secuencia de test de hipótesis con valores en $[0, 1]$. Decimos que $(T_n)_{n\in\mathbb{N}}$ es una secuencia de estimadores consistente cuando, para todo $\theta \in \Theta$, la sucesión $(T_n)_{n\in\mathbb{N}}$ converge en probabilidad a $g(\theta)$.

De la aplicación del Principio de Máxima Verosimilitud se derivan importantes propiedades asintóticas, no sólo desde el punto de vista de la Estimación, como ya hemos mencionado, sino también desde el punto de vista del Contraste de Hipótesis, como veremos a continuación. Efectivamente, supongamos que Θ es un conjunto abierto de $\mathbb{R}^s$ y deseamos contrastar la hipótesis inicial $\Theta_0 \subset \Theta$, que su vez es un subespacio vectorial c-dimensional de Θ, con $c < s$. Podemos suponer, sin pérdida de generalidad, que el parámetro θ descompone en $(\theta_1, \ldots, \theta_s)$ y la hipótesis inicial a contrastar es $H_0 : \theta_1 = \ldots = \theta_{s-c} = 0$. Supongamos que, para cada $n \in \mathbb{N}$, podemos construir el test de la razón de verosimilitudes, con estadístico de contraste RV_n, en cada caso. En Fergusson (1996) se prueba lo siguiente:

Teorema 9.28.
Con las condiciones de regularidad[45] necesarias se verifica, para todo $\theta \in \Theta_0$,

$$-2 \log RV_n \xrightarrow{d} \chi^2_{s-c}$$

[42] Ash(1972).
[43] Lehmann (1983), Cap. 6, Corolario 2.1 y Teorema 2.3.
[44] Lehmann (1983), Cap. 6. Corolario 2.3
[45] Nos referimos a hipótesis relacionadas con la continuidad y derivabilidad de las funciones de densidad. Para más detalles, consultar Lehmann (1983) o Fergusson (1996).

Así pues, para un tamaño muestral suficientemente grande y teniendo en cuenta el teorema 9.21-(ii), puede construirse el test de la razón de verosimilitudes a nivel α de manera aproximada, sin necesidad de conocer la distribución nula exacta del estadístico RV, mediante

$$TRV(\omega) = \begin{cases} 1 & \text{si } -2\log RV(\omega) > \chi^{2,\alpha}_{s-c} \\ 0 & \text{si } -2\log RV(\omega) \leq \chi^{2,\alpha}_{s-c} \end{cases}$$

Biliografía

T.W. Anderson (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley.

S.F. Arnold (1981). *The Theory of Linear Models and Multivariate Analysis.* Wiley.

S.F. Arnold (1980). *Asymptotic Validity of F test fr the Ordinary Linear Model and Multiple Correlation Model.* J.A.S.A., **75**, 890-894.

R.B. Ash (1972). *Real Analysis and Probability.* Academic Press.

P. Billingsley (1986.) *Probability and Measure.* Wiley.

M. Bilodeau & D. Brenner (1999). *Theory of Multivariate Statistics.* Springer Verlag.

F. Carmona (2005). *Modelos Lineales.* Universidad de Barcelona.

D.R. Cox & D.V. Hinkley (1974). *Theoretical Statistics.* Chapman & Hall .

A.J. Dobson (1990). *An Introduction to Generalized Linear Models.* Chapman & Hall.

J. Fan & I. Gijbels (1996). *Loca Polynomial Modelling and Its Applications.* Chapman & Hall.

T.S. Ferguson (1996). *A Course in Large Sample Theory.* Chapman & Hall.

F. Ferraty & P. Vieu (2006). *Nonparametric Modelling for Functional Data.*

Springer.

J.P. Florens, M. Mouchart & J.M. Rolin. *Elements of Bayesian Statistics.* Marcel Decker.

J.F. Hair, R.E. Anderson, R.L. Tatham & C.B. Black (1999). *Análisis Multivariante.* Prentice Hall.

A. Hoerl & R. Kennard (1970) *Ridge regresion: biased estimation for nonorthogonal problems.* Technometrics, **19**, 275-284.

A.C. Lehmann (1983). *Theory of Point Estimation.* Wiley.

A.C. Lehmann (1986). *Testing Statistical Hypotesis.* Wiley.

A.C. Lehmann (1998). *Elements of Large Sample Theory.* Wiley.

J.A. Nelder & R.W.M. Wedderburn (1972) *Generalized Linear Moldels.* J, R. Statist. Soc. A, **135**, 370-84.

A.G. Nogales (1988) *Estadística Matemática.* Servicio de Publicaciones Uex.

D. Peña & S. Rivera (1986). *Estadística. Modelos y Métodos.* Alianza Editorial.

J.O. Rawlings, S.G. Pantula & D.A. Dickey (1998). *Applied Regression Analysis.* Springer Verlag.

A.C. Rencher (1995). *Methods of Multivariate Analysis.* John Wiley & Sons.

W. Rudin (1979). *Análisis Real y Complejo.* Alhambra.

S.R. Searle (1971). *Linear Models.* Wiley.

G.A.F. Seber (1977). *Linear Regression.* Wiley.

B.W. Silverman (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall.

Índice alfabético