

Práctica 3: Análisis de la Covarianza

Modelos Lineales

Grados en Estadística y Matemáticas

1 Modelo de Análisis de la Covarianza

2 Ejemplo 1

3 Ejemplo 2

4 Ejemplo 3

Análisis de la Covarianza

El Análisis de la Covarianza se utiliza para determinar si una serie de variables cuantitativas y una serie de factores presentan influencia sobre una variable cuantitativa continua Y . Nos limitaremos al caso con **una sola variable cuantitativa X y un único factor F** .

El modelo que utilizaremos es un caso particular del Modelo de Regresión Lineal. Si $T_1, T_2, \dots, T_r$ son los niveles o tratamientos del factor, se crean las variables “dummy”:

$$FactorT_i = \begin{cases} 1 & \text{si } F = T_i \\ 0 & \text{si } F \neq T_i \end{cases} \quad i = 2, \dots, r$$

Modelo de Análisis de la Covarianza

Una vez definidas estas variables “dummy”, el modelo matemático de Análisis de la Covarianza se expresa del siguiente modo:

$$Y = \beta_0 + \beta_1 X + \tau_2 \text{Factor}T_2 + \cdots + \tau_r \text{Factor}T_r \\ + \delta_2 (X \cdot \text{Factor}T_2) + \cdots + \delta_r (X \cdot \text{Factor}T_r) + \mathcal{E}$$

En este modelo,

- β_0 y β_1 : coeficientes de la recta de regresión para el tratamiento T_1 (dicho tratamiento realiza una función de control).
- τ_i : efecto del tratamiento T_i sobre la ordenada en el origen de la recta de regresión.
- δ_i : efecto del tratamiento T_i sobre la pendiente de la recta de regresión.

Modelo de Análisis de la Covarianza

El contraste de interés en este modelo será:

$$H_0 : \delta_2 = \cdots = \delta_r = 0$$

$$H_1 : \delta_i \neq 0 \text{ para algún } i$$

Caso de aceptar H_1 , la ecuación de la recta de regresión será diferente para algún tratamiento. Calculamos la ecuación de la recta de regresión para el tratamiento T_i , $i = 2, \dots, r$ que será

$$Y = (\beta_0 + \tau_i) + (\beta_1 + \delta_i)X$$

Modelo de Análisis de la Covarianza

Si aceptamos H_0 , las rectas que ajustan Y frente a X serán paralelas o coincidentes para los r tratamientos. Pasaremos entonces al modelo:

$$Y = \beta_0 + \beta_1 X + \tau_2 \text{Factor}T_2 + \cdots + \tau_r \text{Factor}T_r + \mathcal{E}$$

En este modelo, el contraste de interés será:

$$H_0 : \tau_2 = \cdots = \tau_r = 0$$

$$H_1 : \tau_i \neq 0 \text{ para algún } i$$

Caso de aceptar H_1 , las ecuaciones de las rectas de regresión de los distintos tratamientos serán paralelas.

Si por el contrario aceptamos H_0 , el factor no tendrá influencia sobre Y y el problema de Análisis de la Covarianza se reducirá a un problema de Regresión Lineal Simple.

Ejemplo 1

Se seleccionan al azar cinco casas recientemente vendidas en tres distintas zonas residenciales (A, B y C) de cierta ciudad. Se compara el precio de venta de la casa (y) frente valor catastral de la propiedad (x).

Los datos, donde se dan el precio de venta y el valor catastral en miles de euros, se encuentran en el archivo `casas.dat`.

Se trata de ajustar una ecuación de regresión lineal y determinar si es válida para las tres zonas residenciales.

Ejemplo 1

En este ejemplo:

- $Y \equiv$ precio de venta de la casa
- $X \equiv$ valor catastral de la casa
- $F \equiv$ zona

Son necesarias dos variables “dummy”:

$$zonaB = \begin{cases} 1 & \text{si } zona = B \\ 0 & \text{si } zona \neq B \end{cases}, \quad zonaC = \begin{cases} 1 & \text{si } zona = C \\ 0 & \text{si } zona \neq C \end{cases}$$

El modelo de análisis de la covarianza queda:

$$Y = \beta_0 + \beta_1 X + \tau_2 zonaB + \tau_3 zonaC + \delta_2 (X \cdot zonaB) + \delta_3 (X \cdot zonaC) + \varepsilon$$

Ejemplo 1

En este modelo, se contrastará en primer lugar la hipótesis

$$H_0 : \delta_2 = \delta_3 = 0 \quad (\text{igual pendiente para las tres zonas})$$

$$H_1 : \delta_2 \neq 0 \text{ ó } \delta_3 \neq 0 \quad (\text{diferentes pendientes para las tres zonas})$$

Si los parámetros δ_2, δ_3 es significativo, la ecuación para la zona A quedaría:

$$Y = \beta_0 + \beta_1 X + \mathcal{E}$$

para la zona B

$$Y = (\beta_0 + \tau_2) + (\beta_1 + \delta_2)X + \mathcal{E}$$

y para la zona C

$$Y = (\beta_0 + \tau_3) + (\beta_1 + \delta_3)X + \mathcal{E}$$

Ejemplo 1

En caso contrario, pasaríamos a modelo

$$Y = \beta_0 + \beta_1 X + \tau_2 \text{zonaB} + \tau_3 \text{zonaC} + \mathcal{E}$$

donde se contrastaría la hipótesis

$H_0 : \tau_2 = \tau_3 = 0$ (ordenada en el origen igual para todas las rectas)

$H_1 : \tau_2 \neq 0 \text{ ó } \tau_3 \neq 0$ (ordenada en el origen distinta para todas rectas)

Ejemplo 1: datos de casas (continuación)

Estimar el precio de venta de una casa con valor catastral de 75 mil dolares, situada en la zona B.

```
predict (LinearModel.x, data.frame (x=75, zona="B"), interval="confidence")
```

```
predict (LinearModel.x, data.frame (x=75, zona="B"), interval="prediction")
```

Ejemplo 2

Se realizó un experimento para ver el rendimiento de la caña de azúcar (en toneladas por hectárea). Para ello se utilizaron 4 fertilizantes, destinándose 12 parcelas a cada uno de ellos. En cada parcela, además del rendimiento de la caña (Y) se hizo un recuento del número observado de tallos molederos (adultos).

Se trata de examinar el efecto de los fertilizantes sobre el rendimiento de caña, descontando el efecto del número de tallos molederos.

Ejemplo 2

En este ejemplo, el factor es el fertilizante:

- $F \equiv \text{Sex}$

Necesitaremos 3 variables “dummy”:

$$fertII = \begin{cases} 1 & \text{si } fert = II \\ 0 & \text{si } fert \neq II \end{cases}, \quad fertIII = \begin{cases} 1 & \text{si } fert = III \\ 0 & \text{si } fert \neq III \end{cases}$$

$$fertIV = \begin{cases} 1 & \text{si } fert = IV \\ 0 & \text{si } fert \neq IV \end{cases}$$

El modelo de análisis de la covarianza queda:

$$\begin{aligned} Y = & \beta_0 + \beta_1 X + \tau_2 fertII + \tau_3 fertIII + \tau_4 fertIV \\ & + \delta_2 (X \cdot fertII) + \delta_3 (X \cdot fertIII) + \delta_4 (X \cdot fertIV) + \mathcal{E} \end{aligned}$$

Ejemplo 2

En este modelo, se contrastará en primer lugar la hipótesis

$H_0 : \delta_2 = \delta_3 = \delta_4 = 0$ (igual pendiente para los 4 fertilizantes)

$H_1 : \delta_2, \delta_3 \text{ ó } \delta_4 \neq 0$ (diferente pendiente para algún fertilizante)

Si se acepta H_1 , la ecuación para cada fertilizante quedaría:

Fertilizante I $Y = \beta_0 + \beta_1 X + \mathcal{E}$

Fertilizante II $Y = (\beta_0 + \tau_2) + (\beta_1 + \delta_2)X + \mathcal{E}$

Fertilizante III $Y = (\beta_0 + \tau_3) + (\beta_1 + \delta_3)X + \mathcal{E}$

Fertilizante IV $Y = (\beta_0 + \tau_4) + (\beta_1 + \delta_4)X + \mathcal{E}$

Ejemplo 2

En caso contrario, pasaríamos a modelo

$$Y = \beta_0 + \beta_1 X + \tau_2 \text{fertIII} + \tau_3 \text{fertIII} + \tau_4 \text{fertIV}$$

donde se contrastaría la hipótesis

$H_0 : \tau_2 = \tau_3 = \tau_4 = 0$ (ordenada en el origen igual para las 4 rectas)

$H_1 : \tau_2, \tau_3 \text{ ó } \tau_4 \neq 0$ (ordenada en el origen distinta para alguna recta)

Ejemplo 3

En 1947 Fisher publicó un conjunto de datos que contenía el peso los pesos del corazón y del cuerpo de 47 gatas y 97 gatos. Dichos datos se encuentran en el archivo `cats.dat`.

A la vista de los datos, se trata de ajustar una ecuación lineal que relacione el peso del corazón con el peso corporal, decidiendo si la ecuación ajustada es igual o diferente para gatos y gatas.

Ejemplo 3

En este ejemplo:

- $Y \equiv \text{Hwt}$
- $X \equiv \text{Bwt}$
- $F \equiv \text{Sex}$

Sólo una variable “dummy” es necesaria:

$$\text{SexM} = \begin{cases} 1 & \text{si } \text{Sex} = M \\ 0 & \text{si } \text{Sex} = F \end{cases}$$

El modelo de análisis de la covarianza queda:

$$Y = \beta_0 + \beta_1 X + \tau \text{SexM} + \delta(X \cdot \text{SexM}) + \mathcal{E}$$

Ejemplo 3

En este modelo, se contrastará en primer lugar la hipótesis

$H_0 : \delta = 0$ (igual pendiente para hembras y machos)

$H_1 : \delta \neq 0$ (diferentes pendientes para hembras y machos)

Si el parámetro δ es significativo, la ecuación para las hembras quedaría

$$Y = \beta_0 + \beta_1 X + \mathcal{E}$$

y para los machos

$$Y = (\beta_0 + \tau) + (\beta_1 + \delta)X + \mathcal{E}$$

Ejemplo 3

En caso contrario, pasaríamos a modelo

$$Y = \beta_0 + \beta_1 X + \tau \text{Sex}M + \mathcal{E}$$

donde se contrastaría la hipótesis

$H_0 : \tau = 0$ (ordenada en el origen igual para ambas rectas)

$H_1 : \tau \neq 0$ (ordenada en el origen distinta para ambas rectas)