MODELOS LINEALES

Francesc Carmona

Departament d'Estadística

UNIVERSITAT DE BARCELONA

Barcelona, 19 de diciembre de 2003

Prólogo

Las páginas que siguen constituyen una parte de las exposiciones teóricas y prácticas de asignaturas que se han impartido a lo largo de algunos años en varias licenciaturas y cursos de doctorado. En particular en la licenciatura de Matemáticas, la licenciatura de Biología y la diplomatura de Estadística de la Universidad de Barcelona. Se ha intentado un cierto equilibrio entre las explicaciones teóricas y los problemas prácticos. Sin embargo, nuestra intención siempre ha sido fundamentar sólidamente la utilización de los modelos lineales como base de las aplicaciones de la regresión, el análisis de la varianza y el diseño de experimentos. Por ello, en este libro la base matemática y estadística es considerable y creemos importante la correcta definición de los conceptos y la rigurosidad de las demostraciones. Una sólida base impedirá cometer ciertos errores, habituales cuando se aplican los procedimientos ciegamente.

Por otra parte, la aplicación práctica de los métodos de regresión y análisis de la varianza requiere la manipulación de muchos datos, a veces en gran cantidad, y el cálculo de algunas fórmulas matriciales o simples. Para ello es absolutamente imprescindible la utilización de algún programa de ordenador que nos facilite el trabajo. En una primera instancia es posible utilizar cualquier programa de hojas de cálculo que resulta sumamente didáctico. También se puede utilizar un paquete estadístico que seguramente estará preparado para ofrecer los resultados de cualquier modelo lineal estándar como ocurre con el paquete SPSS. En cambio, en este libro se ha optado por incluir algunos ejemplos con el programa R. Las razones son varias. En primer lugar, se trata de un programa que utiliza el lenguaje S, está orientado a objetos, tiene algunos módulos específicos para los modelos lineales y es programable. R utiliza un lenguaje de instrucciones y al principio puede resultar un poco duro en su aprendizaje, sin embargo superada la primera etapa de adaptación, su utilización abre todo un mundo de posibilidades, no sólo en los modelos lineales, sino en todo cálculo estadístico. Además, la razón más poderosa es que el proyecto R es GNU y, por tanto, de libre distribución. De modo que los estudiantes pueden instalar en su casa el programa R y practicar cuanto quieran sin coste económico alguno. Por otra parte, el paquete S-PLUS es una versión comercial con el mismo conjunto de instrucciones básicas.

El tratamiento de algunos temas tiene su origen en unos apuntes de C.M. Cuadras y Pedro Sánchez Algarra (1996) que amablemente han cedido para su actualización en este libro y a los que agradezco profundamente su colaboración. También es evidente que algunas demostraciones tienen su origen en el clásico libro de Seber.

Por último, este libro ha sido escrito mediante el procesador de textos científico LaTeX y presentado en formato electrónico. Gracias a ello este libro puede actualizarse con relativa facilidad. Se agradecerá cualquier la comunicación de cualquier errata, error o sugerencia.

Barcelona, 19 de diciembre de 2003.

Dr. Francesc Carmona

Índice general

Capítulo 1

Las condiciones

1.1. Introducción

Los métodos de la Matemática que estudian los fenómenos deterministas relacionan, por lo general, una variable dependiente con diversas variables independientes. El problema se reduce entonces a resolver un sistema lineal, una ecuación diferencial, un sistema no lineal, etc.. Sin embargo, la aplicación de los métodos cuantitativos a las Ciencias Experimentales ha revelado la poca fiabilidad de las relaciones deterministas. En tales Ciencias, el azar, la aleatoriedad, la variabilidad individual, las variables no controladas, etc. justifican el planteo, en términos muy generales, de la ecuación fundamental

"observación" = "modelo" + "error aleatorio"

El experimentador puede, fijando las condiciones de su experimento, especificar la estructura del modelo, pero siempre debe tener en cuenta el error aleatorio o desviación entre lo que observa y lo que espera observar según el modelo.

Los modelos de regresión utilizan la ecuación anterior fijando el modelo como una función lineal de unos parámetros. El objetivo consiste, casi siempre, en la predicción de valores mediante el modelo ajustado.

El *Análisis de la Varianza* es un método estadístico introducido por R.A. Fisher de gran utilidad en las Ciencias Experimentales, que permite controlar diferentes variables cualitativas y cuantitativas (llamadas factores), a través de un modelo lineal, suponiendo normalidad para el error aleatorio. Fisher(1938) definió este método como "la separación de la varianza atribuible a un grupo de la varianza atribuible a otros grupos". Como veremos, los tests en Análisis de la Varianza se construyen mediante estimaciones independientes de la varianza del error.

Ambos conjuntos de modelos se pueden abordar con una teoría común: los modelos lineales.

Iniciaremos este capítulo con un ejemplo de modelización de un problema y su aplicación práctica. A continuación explicaremos en qué consiste esencialmente el método de los mínimos cuadrados y estableceremos las condiciones para que este método sea válido para su utilización en Estadística.

1.2. Un ejemplo

En el libro de Sen and Srivastava en [66, pág. 2] se explica este ejemplo que nosotros hemos adaptado a las medidas europeas.

Sabemos que cuantos más coches circulan por una carretera, menor es la velocidad del tráfico. El estudio de este problema tiene como objetivo la mejora del transporte y la reducción del tiempo de viaje.

La tabla adjunta proporciona los datos de la densidad (en vehículos por km) y su correspondiente velocidad (en km por hora).

Dato	Densidad	Velocidad	Dato	Densidad	Velocidad
1	12,7	62,4	13	18,3	51,2
2	17,0	50,7	14	19,1	50,8
3	66,0	17,1	15	16,5	54,7
4	50,0	25,9	16	22,2	46,5
5	87,8	12,4	17	18,6	46,3
6	81,4	13,4	18	66,0	16,9
7	75,6	13,7	19	60,3	19,8
8	66,2	17,9	20	56,0	21,2
9	81,1	13,8	21	66,3	18,3
10	62,8	17,9	22	61,7	18,0
11	77,0	15,8	23	66,6	16,6
12	89,6	12,6	24	67,8	18,3

Cuadro 1.1: Datos del problema de tráfico

Como la congestión afecta a la velocidad, estamos interesados en determinar el efecto de la densidad en la velocidad. Por razones que explicaremos más adelante (ver ejercicio 9.2), tomaremos como variable dependiente la raíz cuadrada de la velocidad.

El gráfico 1.1 presenta la nube de puntos o diagrama de dispersión (*scatter plot*) con la variable independiente (densidad) en el eje horizontal y la variable dependiente (raíz cuadrada de la velocidad) en el eje vertical.

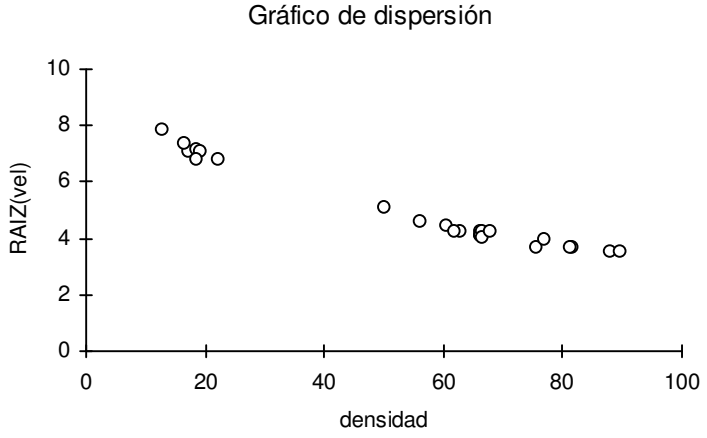


Figura 1.1: Nube de puntos del problema de tráfico

Como primera aproximación podríamos tomar, como modelo de ajuste, la recta que une dos puntos representativos, por ejemplo, los puntos $(12,7, \sqrt{62,4})$ y $(87,8, \sqrt{12,4})$. Dicha recta es $y = 8,6397 - 0,0583x$.

Inmediatamente nos proponemos hallar la mejor de las rectas, según algún criterio. Como veremos, el método de los mínimos cuadrados proporciona una recta, llamada recta de regresión, que goza de muy buenas propiedades. Este método consiste en hallar a y b tales que se minimice la suma de los errores al cuadrado.

$$\sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

En este caso la recta de regresión es $y = 8,0898 - 0,0566x$.

Para estudiar la bondad del ajuste se utilizan los residuos

$$e_i = y_i - \hat{y}_i$$

donde $\hat{y}_i = 8,0898 - 0,0566x_i$. Los gráficos de la figura 1.2 nos muestran estos residuos.

Para mejorar el modelo podemos añadir el término cuadrático y considerar el modelo parabólico

$$y_i = a + bx_i + cx_i^2$$

También aquí, el método de los mínimos cuadrados proporciona un ajuste que es óptimo en varios aspectos. Se trata de hallar los valores de a, b y c que minimizan la suma de los errores al cuadrado

$$\sum_{i=1}^{n}(y_i - (a + bx_i + cx_i^2))^2$$

El cálculo de estos valores con los datos del tráfico se deja como ejercicio (ver ejercicio 1.3).

La figura 1.3 muestra los gráficos de los residuos para el modelo parabólico.

Finalmente, podemos utilizar el modelo concreto que hemos obtenido para sustituir la velocidad en la ecuación

$$\text{flujo} = \text{velocidad} \times \text{densidad}$$

de modo que el flujo queda en función de la densidad. Por último, el máximo valor de esta función es la capacidad de la carretera.

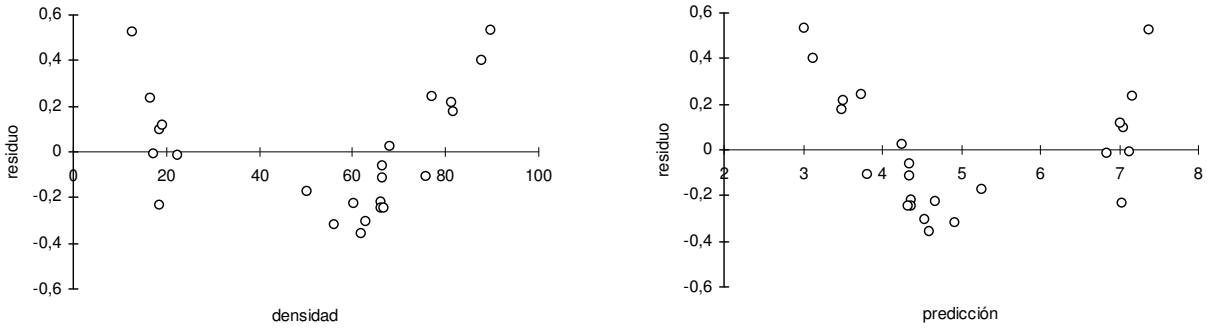


Figura 1.2: Gráficos de los residuos del modelo *recta de regresión*.

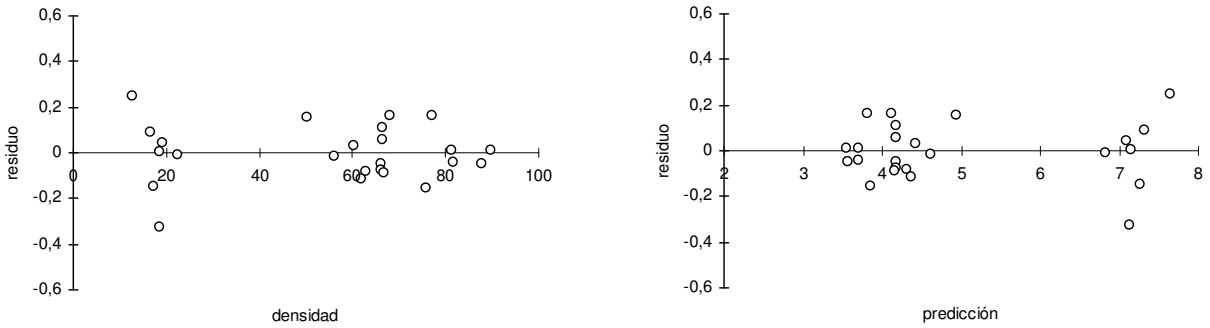


Figura 1.3: Gráficos de los residuos del modelo *parabólico*.

1.3. El modelo

Cuando en el ejemplo anterior ajustamos los datos a una recta, implícitamente estamos asumiendo la hipótesis de que los datos siguen un patrón lineal subyacente del tipo

$$y = \beta_0 + \beta_1 x$$

Pero el ajuste no es perfecto y contiene errores. La ecuación que define el modelo es

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \ldots, n$$

donde ϵ_i son los errores aleatorios. Éste es el modelo de *regresión simple* o con una sola variable independiente.

En el mismo ejemplo anterior, ajustamos mejor con el modelo

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \quad i = 1, \ldots, n$$

que continúa siendo un *modelo lineal*.

Un modelo es lineal si lo es para los parámetros. Por ejemplo, el modelo $\ln y_i = \beta_0 + \beta_1 \ln(x_i) + \epsilon_i$ es lineal, mientras que $y_i = \beta_0 \exp(-\beta_1 x_i)\epsilon_i$ no.

En general, suponemos que una cierta variable aleatoria Y es igual a un valor fijo η más una desviación aleatoria ϵ

$$Y = \eta + \epsilon$$

η representa la verdadera medida de la variable, es decir, la parte *determinista* de un experimento, que depende de ciertos factores cualitativos y variables cuantitativas que son controlables por el experimentador.

El término ϵ representa el *error*. Es la parte del modelo no controlable por el experimentador debido a múltiples causas aleatorias, inevitables en los datos que proceden de la Biología, Psicología, Economía, Medicina,... El error ϵ convierte la relación matemática $Y = \eta$ en la relación estadística $Y = \eta + \epsilon$, obligando a tratar el modelo desde la perspectiva del análisis estadístico.

En particular, los modelos de la forma

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \quad i = 1, \ldots, n$$

con $k > 1$ variables independientes, predictoras o regresoras, se llaman modelos de *regresión múltiple*. La variable cuyos datos observados son y_i es la llamada variable dependiente o respuesta.

Los parámetros β_j son desconocidos y nuestro objetivo principal es su estimación. En cuanto a los errores ϵ_i, su cálculo explícito nos permitirá, como veremos extensamente, la evaluación del modelo.

Observación:

En el modelo de regresión simple puede suceder que los datos x_i $i = 1, \ldots, n$ correspondan a los valores observados de una v.a. X o de una variable controlada no aleatoria. En cualquier caso, vamos a considerar los valores x_i como constantes y no como observaciones de una variable aleatoria.

En la regresión simple

$$Y = \phi(x) + \epsilon$$

donde Y es aleatoria y ϵ es aleatoria con $E(\epsilon) = 0$. De manera que, para cada valor $X = x$, Y es una v.a. con esperanza $\phi(x)$. Si asumimos

$$\phi(x) = E[Y|X = x] = \beta_0 + \beta_1 x$$

podemos proceder considerando las inferencias como condicionadas a los valores observados de X.

En cualquier caso, también en regresión múltiple, vamos a considerar los valores de las variables regresoras $X_1, \ldots, X_k$ como simplemente números.

1.4. El método de los mínimos cuadrados

La paternidad de este método se reparte entre Legendre que lo publicó en 1805 y Gauss que lo utilizó en 1795 y lo publicó en 1809.

Obviamente, cuanto menores son los residuos, mejor es el ajuste. De todos los posibles valores de los β_j, el método de los mínimos cuadrados selecciona aquellos que minimizan

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))^2$$

En el caso de la regresión lineal simple

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

de modo que derivando e igualando a cero, se obtienen los estimadores MC (mínimo-cuadráticos) ó LS (*least squares*)

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{aligned}$$

También se puede considerar el *modelo centrado*, que consiste en centrar los datos de la variable regresora

$$y_i = \gamma_0 + \beta_1 (x_i - \bar{x}) + \epsilon_i \quad i = 1, \ldots, n$$

La estimación *MC* de γ_0, β_1 es equivalente a la estimación de β_0, β_1, ya que $\gamma_0 = \beta_0 + \beta_1 \bar{x}$. De modo que $\hat{\gamma}_0 = \bar{y}$ y la estimación de β_1 es la misma que en el modelo anterior.

Con las estimaciones de los parámetros, podemos proceder al cálculo de predicciones $\hat{y}_i$ y residuos e_i

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x}) \\ e_i &= y_i - \hat{y}_i = y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}) \end{aligned}$$

Como consecuencia resulta que

$$\sum_{i=1}^{n} e_i = 0$$

lo que no ocurre en un modelo sin β_0.

Finalmente, si queremos una medida del ajuste de la regresión podemos pensar en la suma de cuadrados $\sum_{i=1}^{n} e_i^2$, pero es una medida que depende de las unidades de y_i al cuadrado. Si $\beta_0 \neq 0$, la medida que se utiliza es el coeficiente de determinación

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Sabemos que $0 \leq R^2 \leq 1$ y cuando $R^2 \approx 1$ el ajuste es bueno.

En el caso $\beta_0 = 0$, el coeficiente de determinación es

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2}$$

de modo que los modelos que carecen de término independiente no se pueden comparar con los que sí lo tienen.

1.5. Las condiciones de Gauss-Markov

Hasta aquí, el método de los mínimos cuadrados es analítico ¿dónde está la estadística?

A lo largo de los siguientes capítulos vamos a ver que un modelo estadístico y la imposición de algunas condiciones, hacen que podamos utilizar el modelo con toda la potencia de los métodos estadísticos y calibrar la bondad del ajuste desde esa óptica.

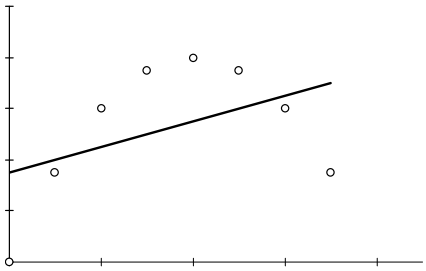
Una primera pregunta es ¿qué tan bueno es el método de los mínimos cuadrados para estimar los parámetros? La respuesta es que este método proporciona un buen ajuste y buenas predicciones si se verifican las condiciones de Gauss-Markov.

En el modelo lineal que hemos definido anteriormente, se supone que los errores ϵ_i son desviaciones que se comportan como variables aleatorias. Vamos a exigir que estos errores aleatorios verifiquen las siguientes condiciones:

1. $E(\epsilon_i) = 0 \qquad i = 1, \ldots, n$
2. $\text{var}(\epsilon_i) = \sigma^2 \qquad i = 1, \ldots, n$
3. $E(\epsilon_i \cdot \epsilon_j) = 0 \qquad \forall i \neq j$

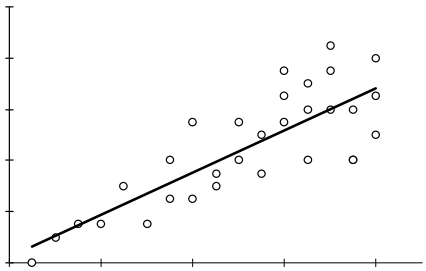
Veamos con detalle estas condiciones:

Primera condición $\boxed{E(\epsilon_i) = 0 \qquad i = 1, \ldots, n}$



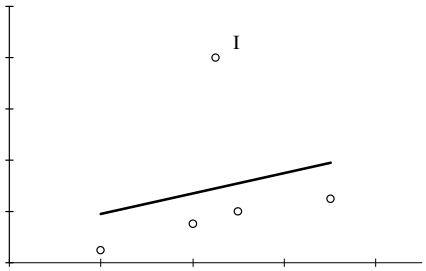
Se trata de una condición natural sobre un error.
De este modo nos aseguramos que $E(y_i) = \beta_0 + \beta_1 x_i$, el modelo lineal es correcto y la situación que representa el gráfico no se puede dar.

Segunda condición $\boxed{\text{var}(\epsilon_i) = E(\epsilon_i^2) = \sigma^2 \text{ constante} \qquad i = 1, \ldots, n}$

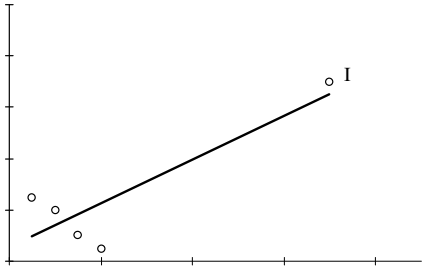


Es la propiedad de *homocedasticidad*.
En el gráfico se representa una situación anómala llamada de *heterocedasticidad*, en la que la $\text{var}(\epsilon_i)$ crece con x_i.
El parámetro desconocido σ^2 es la llamada varianza del modelo.

Otras situaciones extrañas, que también se pretende prevenir, son:



El punto I del gráfico representa un punto influyente y atípico (*outlier*). En general es un punto a estudiar, un error o incluso una violación de la primera condición.



El punto I del gráfico es claramente influyente, aunque no es atípico (outlier), ya que proporciona un residuo pequeño.

Tercera condición $\boxed{E(\epsilon_i \epsilon_j) = 0 \qquad \forall i \neq j}$

Las observaciones deben ser incorrelacionadas. Con dos puntos tenemos una recta de regresión. Con 20 copias de esos dos puntos, tenemos 40 puntos y la misma recta, poco fiable.

Tales condiciones pueden expresarse en forma matricial como

$$E(\boldsymbol{\epsilon}) = \mathbf{0} \qquad \text{var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$$

donde $E(\boldsymbol{\epsilon})$ es el vector de esperanzas matemáticas y $\text{var}(\boldsymbol{\epsilon})$ es la matriz de covarianzas de $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$.

Como demostraremos en los siguientes capítulos, la adopción de estas condiciones evitará teóricamente las situaciones anómalas que aquí hemos esquematizado.

1.6. Otros tipos de modelos lineales

Por suerte, con el mismo tratamiento podremos resolver otros modelos lineales, que aunque tienen diferentes objetivos, gozan de las mismas bases teóricas.

Por ejemplo, el Análisis de la Varianza con un factor (one-way Analysis of Variance), representado por el modelo lineal

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad \text{con } \epsilon_{ij} \sim N(0, \sigma^2) \text{ indep.},$$

se resuelve de forma similar al modelo de regresión.

El Análisis de la Covarianza, que utiliza como variables independientes tanto variables cuantitativas como factores, y el Análisis Multivariante de la Varianza, con varias variables dependientes, son dos de los análisis que generalizan el estudio y aplicaciones de los modelos lineales que vamos a investigar.

1.7. Algunas preguntas

Un típico problema de estadística consiste en estudiar la relación que existe, si existe, entre dos variables aleatorias X e Y. Por ejemplo, altura y peso, edad del hombre y la mujer en una pareja, longitud y anchura de unas hojas, temperatura y presión de un determinado volumen de gas.

Si tenemos n pares de observaciones (x_i, y_i) $i = 1, 2, \ldots, n$, podemos dibujar estos puntos en un gráfico o *scatter diagram* y tratar de ajustar una curva a los puntos de forma que los puntos se hallen lo más *cerca* posible de la curva. No podemos esperar un ajuste perfecto porque ambas variables están expuestas a fluctuaciones al azar debido a factores incontrolables. Incluso aunque en algunos casos pudiera existir una relación exacta entre variables físicas como temperatura y presión, también aparecerían fluctuaciones debidas a errores de medida.

Algunas cuestiones que podemos plantearnos en nuestras investigaciones son:

- Si existe un modelo físico teórico y lineal, podemos utilizar la regresión para estimar los parámetros.
- Si el modelo teórico no es lineal, se puede, en muchos casos, transformar en lineal. Por ejemplo:
$$PV^\gamma = c \quad \longrightarrow \quad \log P = \log c - \gamma \log V$$
- Si no es una recta, se puede estudiar un modelo de regresión polinómico. ¿De qué grado?

- En el modelo múltiple intervienen varias variables "predictoras" ¿son todas necesarias? ¿son linealmente independientes las llamadas "variables independientes"?
- ¿Se verifican realmente las condiciones de Gauss-Markov?
- ¿Qué ocurre si las variables predictoras son discretas?
- ¿Qué ocurre si la variable dependiente es discreta o una proporción?
- ¿Y si faltan algunos datos?
- ¿Qué hacemos con los puntos atípicos y los puntos influyentes?

Algunas de estas preguntas las iremos trabajando y resolviendo en los siguientes capítulos, otras pueden quedar para una posterior profundización.

1.8. Ejemplos con R

En esta sección vamos a ver como se calculan las regresiones que se han sugerido a partir del ejemplo inicial con los datos de la tabla 1.1.
En primer lugar procedemos a introducir los datos en los vectores correspondientes.

```
> dens<-c(12.7,17.0,66.0,50.0,87.8,81.4,75.6,66.2,81.1,62.8,77.0,89.6,
+ 18.3,19.1,16.5,22.2,18.6,66.0,60.3,56.0,66.3,61.7,66.6,67.8)
> vel<-c(62.4,50.7,17.1,25.9,12.4,13.4,13.7,17.9,13.8,17.9,15.8,12.6,
+ 51.2,50.8,54.7,46.5,46.3,16.9,19.8,21.2,18.3,18.0,16.6,18.3)
> rvel<-sqrt(vel)
```

Las siguientes instrucciones generan el gráfico de puntos para estos datos.

```
> par(pty="m")
> plot(dens,rvel,type="p",xlab="densidad",ylab="RAIZ(vel)")
```

El cálculo de la regresión simple se realiza con la función `lsfit(x,y)` que asignamos al objeto `recta.ls`

```
> recta.ls<-lsfit(dens,rvel)
```

Aunque esta última instrucción no muestra ninguna información en pantalla, ahora ya podemos utilizar su resultado. Por ejemplo, podemos añadir la recta de regresión al gráfico anterior.

```
> abline(recta.ls)
```

Los coeficientes de la recta son:

```
> recta.ls$coef
  Intercept           X
 8.08981299 -0.05662558
```

También se puede obtener una información más completa con la instrucción `ls.print`, aunque su resultado no se explicará hasta el capítulo correspondiente.

```
> ls.print(recta.ls, digits=4, print.it=T)
Residual Standard Error=0.2689
R-Square=0.9685
F-statistic (df=1, 22)=676.3944
p-value=0

          Estimate Std.Err  t-value Pr(>|t|)
Intercept   8.0898  0.1306  61.9295        0
X          -0.0566  0.0022 -26.0076        0
```

La estimación de la desviación estándar de los errores y otros elementos de diagnosis del modelo se obtienen con la función `ls.diag` como

```
> ls.diag(recta.ls)$std.dev
[1] 0.2689388
```

Con el vector de residuos y las predicciones se pueden dibujar unos gráficos similares a los de la figura 1.2. La instrucción `par(mfrow=c(1,2))` permite dos gráficos en la misma figura.

```
> e<-recta.ls$residuals
> par(mfrow=c(1,2))
> par(pty="s")
> plot(dens,e,type="p",xlab="densidad",ylab="residuos",ylim=c(-0.6,0.6))
> abline(h=0)
> pred<-rvel-e
> plot(pred,e,type="p",xlab="predicción",ylab="residuos",ylim=c(-0.6,0.6))
> abline(h=0)
```

Finalmente, podemos repetir los cálculos para el modelo parabólico. Simplemente debemos introducir los valores de la variable densidad y sus cuadrados en una matriz de datos. El resto es idéntico al modelo de regresión simple.

```
> matriz.frame<-data.frame(dens,dens^2)
> parabola.ls<-lsfit(matriz.frame,rvel)
> parabola.ls$coef
    Intercept          dens        dens.2
 8.8814208199 -0.1035152795  0.0004892585
> round(parabola.ls$coef,5)
Intercept      dens    dens.2
  8.88142  -0.10352   0.00049
> e<-parabola.ls$residuals
> par(mfrow=c(1,2))
> par(pty="s")
> plot(dens,e,type="p",xlab="densidad",ylab="residuos",ylim=c(-0.6,0.6))
> abline(h=0)
> pred<-rvel-e
> plot(pred,e,type="p",xlab="predicción",ylab="residuos",ylim=c(-0.6,0.6))
> abline(h=0)
```

Los gráficos serán muy similares a los de la figura 1.3.

En los siguientes capítulos veremos otras instrucciones de R, en especial la función `lm`, que permiten ajustar un modelo de regresión a unos datos.

1.9. Ejercicios

Ejercicio 1.1

Hallar las estimaciones de los parámetros en un modelo de regresión lineal simple, minimizando la suma de los cuadrados de los errores:

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Hallar una expresión para las predicciones $\hat{y}_i$ y los residuos $e_i = y_i - \hat{y}_i$.

Ejercicio 1.2

Hallar las estimaciones de los parámetros en un modelo de regresión parabólico, minimizando la suma de los cuadrados de los errores:

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$$

Hallar una expresión para las predicciones $\hat{y}_i$ y los residuos $e_i = y_i - \hat{y}_i$.

Ejercicio 1.3

Consideremos el problema de tráfico planteado en el apartado 1.2 de este capítulo, con la variable independiente *densidad* y la variable dependiente *raíz cuadrada de la velocidad*. Con los datos proporcionados en la tabla 1.1 realizar el siguiente proceso:

(a) Dibujar la nube de puntos y la recta que pasa por los puntos $(12{,}7, \sqrt{62{,}4})$ y $(87{,}8, \sqrt{12{,}4})$. Dibujar el gráfico de los residuos con la densidad y el gráfico con las predicciones. Calcular la suma de cuadrados de los residuos.

(b) Hallar la recta de regresión simple. Dibujar el gráfico de los residuos con la densidad y el gráfico con las predicciones. Calcular la suma de cuadrados de los residuos.

(c) Mejorar el modelo anterior considerando una regresión parabólica. Dibujar el gráfico de los residuos con la densidad y el gráfico con las predicciones. Calcular la suma de cuadrados de los residuos.

(d) Calcular la capacidad de la carretera o punto de máximo flujo. Recordar que *flujo* = *vel* × *densidad*.

Ejercicio 1.4

La siguiente tabla contiene los mejores tiempos conseguidos en algunas pruebas de velocidad en atletismo en los Juegos Olímpicos de Atlanta:

	hombres	mujeres
distancia	tiempo	
100	9,84	10,94
200	19,32	22,12
400	43,19	48,25
800	102,58	117,73
1500	215,78	240,83
5000	787,96	899,88
10000	1627,34	1861,63
42192	7956,00	8765,00

Si tomamos como variable regresora o independiente la distancia (metros) y como variable respuesta o dependiente el tiempo (segundos):

(a) Calcular la recta de regresión simple con los datos de los hombres y dibujarla. Dibujar el gráfico de los residuos con la distancia y el gráfico con las predicciones. Calcular la suma de cuadrados de los residuos y el R^2.

(b) Repetir el apartado anterior utilizando los logaritmos de las variables tiempo y distancia.

(c) Repetir los dos apartados anteriores utilizando los datos de las mujeres.

Capítulo 2

Estimación

2.1. Introducción

En primer lugar concretaremos la definición general de un modelo lineal y hallaremos la estimación por mínimos cuadrados de los parámetros del modelo.

Veremos que la estimación será única si la matriz de diseño es de rango máximo. En caso contrario, resulta importante definir el concepto de función paramétrica estimable y probar, para estas funciones, la unicidad del estimador mínimo-cuadrático, como estudiaremos en el siguiente capítulo.

Estudiaremos las propiedades de estos estimadores, entre las que destacaremos el *Teorema de Gauss-Markov* que demuestra que los estimadores mínimo-cuadráticos son los mejores, en el sentido de que son insesgados y de mínima varianza.

Además, con la introducción de la hipótesis de normalidad de los errores, podremos estudiar las distribuciones de los estimadores y de otros estadísticos, así como la relación con los estimadores de máxima verosimilitud.

Más adelante, trabajaremos la generalización del método de los mínimos cuadrados cuando la matriz de varianzas-covarianzas de los errores no es $\sigma^2\mathbf{I}$. Por otra parte, también profundizaremos el caso de matrices de diseño de rango no máximo.

2.2. El modelo lineal

Sea Y una variable aleatoria que fluctúa alrededor de un valor desconocido η, esto es

$$Y = \eta + \epsilon$$

donde ϵ es el error, de forma que η puede representar el valor verdadero e Y el valor observado.

Supongamos que η toma valores distintos de acuerdo con diferentes situaciones experimentales según el modelo lineal

$$\eta = \beta_1 x_1 + \cdots + \beta_m x_m$$

donde β_i son parámetros desconocidos y x_i son valores conocidos, cada uno de los cuales ilustra situaciones experimentales diferentes.

En general se tienen n observaciones de la variable Y. Diremos que $y_1, y_2, \ldots, y_n$ observaciones independientes de Y siguen un *modelo lineal* si

$$y_i = x_{i1}\beta_1 + \cdots + x_{im}\beta_m + \epsilon_i \qquad i = 1, \ldots, n$$

Estas observaciones de Y se pueden considerar variables aleatorias independientes y distribuidas como Y (son copias) o también realizaciones concretas (valores numéricos) para los cálculos.
La expresión del modelo lineal en forma matricial es

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

o en forma resumida

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1}$$

Los elementos que constituyen el modelo lineal son:

1. El vector de observaciones $\mathbf{Y} = (y_1, y_2, \ldots, y_n)'$.
2. El vector de parámetros $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_m)'$.
3. La matriz del modelo
$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{pmatrix}$$
cuyos elementos son conocidos.

   En problemas de regresión, $\mathbf{X}$ es la matriz de regresión. En los llamados diseños factoriales del Análisis de la Varianza, $\mathbf{X}$ recibe el nombre de matriz de diseño.
4. El vector de errores o desviaciones aleatorias $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$, donde ϵ_i es la desviación aleatoria de y_i.

Ejemplo 2.2.1

El modelo lineal más simple consiste en relacionar una variable aleatoria Y con una variable controlable x (no aleatoria), de modo que las observaciones de Y verifiquen

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n$$

Se dice que Y es la variable de predicción o dependiente y x es la variable predictora, por ejemplo Y es la respuesta de un fármaco a una dosis x. Hallar β_0 y β_1 es el clásico problema de regresión lineal simple.

Ejemplo 2.2.2

El modelo anterior se puede generalizar a situaciones en las cuales la relación sea polinómica.

Consideremos el modelo

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon \qquad i = 1, \ldots, n$$

Observemos que es lineal en los parámetros β_i. La matriz de diseño es

$$\begin{pmatrix} 1 & x_1 & \ldots & x_1^p \\ 1 & x_2 & \ldots & x_2^p \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & x_n^p \end{pmatrix}$$

Ejemplo 2.2.3

En general, cualquier variable Y puede relacionarse con dos o más variables control. Así, son modelos lineales:

a) $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

b) $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2 + \epsilon_i$

c) $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 \cos(x_{i2}) + \beta_3 \operatorname{sen}(x_{i2}) + \epsilon_i$

Sin embargo, no es modelo lineal

$$y_i = \beta_0 + \beta_1 \log(\beta_2 x_{i1}) + \beta_3 x_{i2}^{\beta_4} + \epsilon_i$$

Ejemplo 2.2.4

Supongamos que la producción Y de una planta depende de un factor F (fertilizante) y un factor B (bloque o conjunto de parcelas homogéneas). El llamado modelo del diseño del factor en bloques aleatorizados es

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

donde

μ *es una constante (media general)*
α_i *el efecto del fertilizante*
β_j *el efecto del bloque*

Si tenemos 2 fertilizantes y 3 bloques, tendremos en total $k = 2 \times 3 = 6$ situaciones experimentales y la siguiente matriz de diseño:

μ	α_1	α_2	β_1	β_2	β_3
1	1	0	1	0	0
1	0	1	1	0	0
1	1	0	0	1	0
1	0	1	0	1	0
1	1	0	0	0	1
1	0	1	0	0	1

La utilización del fertilizante 1 en el bloque 3 queda descrita a través de la fila 5 de $\mathbf{X}$.

Ejemplo 2.2.5

Para predecir la capacidad craneal C, en Antropología se utiliza la fórmula

$$C = \alpha L^{\beta_1} A^{\beta_2} H^{\beta_3}$$

donde L = longitud del cráneo, A = anchura parietal máxima y H = altura basio bregma. La fórmula anterior se convierte en un modelo lineal tomando logaritmos

$$\log C = \log \alpha + \beta_1 \log L + \beta_2 \log A + \beta_3 \log H$$

El parámetro α expresa el tamaño, mientras que los parámetros β expresan la forma del cráneo.

2.3. Suposiciones básicas del modelo lineal

En el modelo lineal definido en el apartado anterior, se supone que los errores ϵ_i son desviaciones que se comportan como variables aleatorias que verifican las condiciones de Gauss-Markov:

1. $E(\epsilon_i) = 0 \qquad i = 1, \ldots, n$
2. $\text{var}(\epsilon_i) = \sigma^2 \qquad i = 1, \ldots, n$
3. $E(\epsilon_i \cdot \epsilon_j) = 0 \qquad \forall i \neq j$

Como sabemos, la condición (2) es la llamada condición de *homocedasticidad* del modelo y el parámetro desconocido σ^2 es la llamada varianza del modelo. La condición (3) significa que las n desviaciones son mutuamente incorrelacionadas.

Estas condiciones pueden expresarse en forma matricial como

$$E(\boldsymbol{\epsilon}) = \mathbf{0} \qquad \text{var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$$

donde $E(\boldsymbol{\epsilon})$ es el vector de esperanzas matemáticas y $\text{var}(\boldsymbol{\epsilon})$ es la matriz de covarianzas de $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$.

Si además suponemos que cada ϵ_i es $N(0, \sigma)$ y que $\epsilon_1, \ldots, \epsilon_n$ son estocásticamente independientes, entonces diremos que el modelo definido es un *modelo lineal normal*. Así tendremos que

$$\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$$

es decir, $\mathbf{Y}$ sigue la distribución normal multivariante de vector de medias $\mathbf{X}\boldsymbol{\beta}$ y matriz de covarianzas $\sigma^2 \mathbf{I}_n$.

Se llama rango del diseño al rango de la matriz $\mathbf{X}$

$$r = \text{rango } \mathbf{X}$$

y es un elemento muy importante en la discusión de los modelos. Evidentemente $r \leq m$. El valor de r es el número efectivo de parámetros del diseño, en el sentido de que si $r < m$ es posible reparametrizar el modelo para que r sea igual al número de parámetros. En muchos casos el diseño verifica directamente que $r = m$ y entonces se dice que es de *rango máximo*.

El modelo lineal que verifique las condiciones aquí expuestas, salvo la normalidad, diremos que está bajo las condiciones de Gauss-Markov ordinarias.

2.4. Estimación de los parámetros

La estimación de los parámetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)'$ se hace con el criterio de los mínimos cuadrados. Se trata de hallar el conjunto de valores de los parámetros $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_m)'$ que minimicen la siguiente suma de cuadrados

$$\begin{aligned} \boldsymbol{\epsilon}'\boldsymbol{\epsilon} &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - \cdots - x_{im}\beta_m)^2 \end{aligned} \tag{2.2}$$

La estimación $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ la llamaremos estimación MC, abreviación de mínimo-cuadrática, o LS del inglés *least squares*.

Teorema 2.4.1

Toda estimación MC de $\boldsymbol{\beta}$ es solución de la ecuación

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y} \tag{2.3}$$

Demostración:
Si desarrollamos la suma de cuadrados $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$ tenemos

$$\begin{aligned} \boldsymbol{\epsilon}'\boldsymbol{\epsilon} &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

y si derivamos matricialmente respecto a $\boldsymbol{\beta}$ resulta

$$\frac{\partial \boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

De modo que, si igualamos a cero, obtenemos la ecuación enunciada en el teorema. ■

Las ecuaciones 2.3 reciben el nombre de *ecuaciones normales*.

Si el rango es máximo y $r = m$, entonces $\mathbf{X}'\mathbf{X}$ tiene inversa y la única solución de las ecuaciones normales es

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Si $r < m$ el sistema de ecuaciones 2.3 es indeterminado y su solución no es única. En estos casos, una posibilidad (ver Apéndice A) es considerar

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$$

donde $\mathbf{A}^{-} = (\mathbf{X}'\mathbf{X})^{-}$ es una g-inversa de $\mathbf{A} = \mathbf{X}'\mathbf{X}$, es decir, $\mathbf{A}^{-}$ verifica

$$\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$$

Entonces se puede demostrar que la solución general es

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} + (\mathbf{I} - \mathbf{A}^{-}\mathbf{A})\mathbf{z}$$

siendo $\mathbf{z}$ un vector paramétrico.

Ahora podemos definir la suma de cuadrados residual como

$$\text{SCR} = \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

Como veremos, SCR entendido como un estadístico función de la muestra $\mathbf{Y}$, desempeña un papel fundamental en el Análisis de la Varianza.
El modelo lineal $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, bajo las hipótesis de Gauss-Markov, verifica

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$

Teorema 2.4.2

Sea $\Omega = \langle \mathbf{X} \rangle \subset \mathbb{R}^n$ el subespacio vectorial generado por las columnas de $\mathbf{X}$ de dimensión $\dim\langle \mathbf{X} \rangle = r = \text{rango } \mathbf{X}$.
Entonces se verifica:

(i) $E(\mathbf{Y}) \in \langle \mathbf{X} \rangle$

(ii) Si $\widehat{\boldsymbol{\beta}}$ es una estimación MC, el vector de residuos $\mathbf{e} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ es ortogonal a $\langle \mathbf{X} \rangle$.

Demostración:
En efecto,

i) Si $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(m)}$ son las columnas de $\mathbf{X}$, entonces

$$E(\mathbf{Y}) = \mathbf{x}_{(1)}\beta_1 + \cdots + \mathbf{x}_{(m)}\beta_m \in \langle \mathbf{X} \rangle$$

ii) $\mathbf{X}'\mathbf{e} = \mathbf{X}'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{0}$ ■

Teorema 2.4.3

Para cualquier $\widehat{\boldsymbol{\beta}}$ solución MC de 2.3 se verifica que

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} \qquad \mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} \qquad \text{SCR} = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

son únicos.
Además

$$\text{SCR} = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} \tag{2.4}$$

Demostración:
Si desarrollamos la suma de cuadrados residual SCR resulta

$$\text{SCR} = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}$$

y como $\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$, obtenemos

$$\text{SCR} = \mathbf{Y}'\mathbf{Y} - 2\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$$

Consideremos ahora los vectores $\widehat{\mathbf{Y}}_1 = \mathbf{X}\widehat{\boldsymbol{\beta}}_1$ y $\widehat{\mathbf{Y}}_2 = \mathbf{X}\widehat{\boldsymbol{\beta}}_2$, donde $\widehat{\boldsymbol{\beta}}_1$ y $\widehat{\boldsymbol{\beta}}_2$ son dos soluciones MC. Entonces $\widehat{\mathbf{Y}}_1$ y $\widehat{\mathbf{Y}}_2$ pertenecen al subespacio $\langle \mathbf{X} \rangle$ generado por las columnas de $\mathbf{X}$ y su diferencia $\widehat{\mathbf{Y}}_1 - \widehat{\mathbf{Y}}_2$ también. Por otra parte, observamos que

$$\mathbf{X}'(\widehat{\mathbf{Y}}_1 - \widehat{\mathbf{Y}}_2) = \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}_1 - \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}_2 = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{Y} = \mathbf{0}$$

de modo que $\widehat{\mathbf{Y}}_1 - \widehat{\mathbf{Y}}_2$ pertenece al ortogonal de $\langle \mathbf{X} \rangle$. Así pues, necesariamente $\widehat{\mathbf{Y}}_1 - \widehat{\mathbf{Y}}_2 = \mathbf{0}$ y el vector de errores $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}_1 = \mathbf{Y} - \widehat{\mathbf{Y}}_2$ es único.
En consecuencia, la suma de cuadrados de los errores SCR también es única. ■

Interpretación geométrica

El modelo teórico es

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{\theta} + \boldsymbol{\epsilon} \qquad \text{si } \boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$$

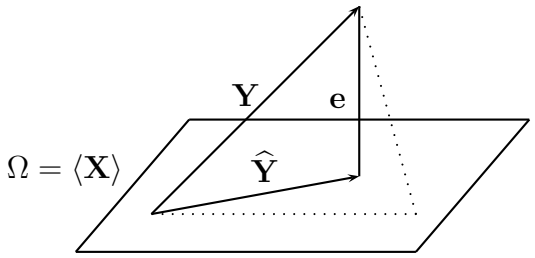
Entonces $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\theta}$ significa que el valor esperado de $\mathbf{Y}$ pertenece al subespacio $\Omega = \langle \mathbf{X} \rangle$ y para estimar los parámetros $\boldsymbol{\beta}$ debemos minimizar

$$\boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \|\mathbf{Y} - \boldsymbol{\theta}\|^2 \qquad \text{con } \boldsymbol{\theta} \in \Omega = \langle \mathbf{X} \rangle$$

Como el vector concreto de observaciones $\mathbf{Y}$ se puede considerar un vector de $\mathbb{R}^n$, el problema anterior se puede resolver en términos geométricos. Así se sabe que cuando $\boldsymbol{\theta} \in \Omega$, $\|\mathbf{Y} - \boldsymbol{\theta}\|^2$ es mínimo para $\boldsymbol{\theta} = \widehat{\mathbf{Y}} = \mathbf{P}\mathbf{Y}$, donde $\mathbf{P}$ es la matriz de la proyección ortogonal en $\Omega = \langle \mathbf{X} \rangle$ (ver Apéndice B). La estimación MC es equivalente a hallar la proyección ortogonal $\widehat{\mathbf{Y}}$ de $\mathbf{Y}$ sobre $\langle \mathbf{X} \rangle$, es decir, la norma euclídea de $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$ es mínima:

$$\text{SCR} = \mathbf{e}'\mathbf{e} = \|\mathbf{e}\|^2 = \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2$$

Se comprende que cualquier otra proyección no ortogonal daría una solución menos adecuada.



Como $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$ es ortogonal a Ω, se verifica que

$$\mathbf{X}'(\mathbf{Y} - \widehat{\mathbf{Y}}) = \mathbf{0} \qquad \text{ó} \qquad \mathbf{X}'\widehat{\mathbf{Y}} = \mathbf{X}'\mathbf{Y}$$

donde $\widehat{\mathbf{Y}}$ está determinada por ser la única proyección ortogonal de $\mathbf{Y}$ en Ω. Cuando las columnas de $\mathbf{X}$ son linealmente independientes, forman una base y existe un único vector $\widehat{\boldsymbol{\beta}}$ tal que $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ de manera que

$$\mathbf{X}'\widehat{\mathbf{Y}} = \mathbf{X}'\mathbf{Y} \quad \Rightarrow \quad \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$

son las ecuaciones normales. En caso contrario, es decir, cuando las columnas de $\mathbf{X}$ son dependientes no podemos concretar una solución única para los parámetros $\boldsymbol{\beta}$. Sin embargo todas las soluciones deben verificar la siguiente propiedad.

Teorema 2.4.4

$\widehat{\boldsymbol{\beta}}$ es una estimación MC de $\boldsymbol{\beta}$ si y sólo si $\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{Y}$, donde $\mathbf{P}$ es la proyección ortogonal en $\Omega = \langle \mathbf{X} \rangle$

Demostración:
Una estimación $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ es MC si y sólo si

$$(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) = \min_{\boldsymbol{\beta}} (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$$

Sea $\widetilde{\boldsymbol{\beta}}$ una estimación cualquiera de $\boldsymbol{\beta}$, entonces

$$\begin{aligned}(\mathbf{Y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}) &= (\mathbf{Y}-\mathbf{PY}+\mathbf{PY}-\mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{PY}+\mathbf{PY}-\mathbf{X}\widetilde{\boldsymbol{\beta}}) \\ &= (\mathbf{Y}-\mathbf{PY})'(\mathbf{Y}-\mathbf{PY}) + (\mathbf{Y}-\mathbf{PY})'(\mathbf{PY}-\mathbf{X}\widetilde{\boldsymbol{\beta}}) \\ &\quad + (\mathbf{PY}-\mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{PY}) + (\mathbf{PY}-\mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{PY}-\mathbf{X}\widetilde{\boldsymbol{\beta}})\end{aligned}$$

Sin embargo

$$(\mathbf{Y}-\mathbf{PY})'(\mathbf{PY}-\mathbf{X}\widetilde{\boldsymbol{\beta}}) = \mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{PY} - \mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{X}\widetilde{\boldsymbol{\beta}} = \mathbf{0}$$

ya que $\mathbf{P}$ es idempotente y además $\mathbf{PX} = \mathbf{X}$. De forma que

$$(\mathbf{Y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}) = (\mathbf{Y}-\mathbf{PY})'(\mathbf{Y}-\mathbf{PY}) + (\mathbf{PY}-\mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{PY}-\mathbf{X}\widetilde{\boldsymbol{\beta}})$$

donde ambos términos son positivos, el primero no depende de $\widetilde{\boldsymbol{\beta}}$ y el segundo se minimiza si es cero, luego $\mathbf{PY} = \mathbf{X}\widehat{\boldsymbol{\beta}}$. ■

En resumen y como ya hemos visto, la solución del problema se basa en la proyección ortogonal sobre el subespacio Ω que garantiza la unicidad del vector de predicciones $\widehat{\mathbf{Y}} = \mathbf{PY}$ y por ende del vector de residuos $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$ y de la suma de cuadrados de los residuos

$$\text{SCR} = \mathbf{e}'\mathbf{e} = (\mathbf{Y}-\mathbf{PY})'(\mathbf{Y}-\mathbf{PY}) = \mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y}$$

ya que $\mathbf{I}-\mathbf{P}$ es idempotente (ver Apéndice B).
La solución para los parámetros $\boldsymbol{\beta}$ debe salir de las ecuaciones normales o de la ecuación $\mathbf{X}\boldsymbol{\beta} = \mathbf{PY}$ y sólo es única cuando el rango de la matriz $\mathbf{X}$ es máximo.

Ejemplo 2.4.1

Consideremos el modelo lineal con $n = 3$, $m = 1$ y $r = 1$

$$\begin{aligned} y_1 &= \theta + \epsilon_1 \\ y_2 &= 2\theta + \epsilon_2 \\ y_3 &= -\theta + \epsilon_3 \end{aligned}$$

que en expresión matricial escribimos

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} \theta + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

de modo que $\mathbf{X}' = (1, 2, -1)$.
Las ecuaciones normales son

$$\begin{pmatrix} 1 & 2 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} \theta = \begin{pmatrix} 1 & 2 & -1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

es decir

$$6\theta = y_1 + 2y_2 - y_3$$

y la estimación MC de θ *es* $\widehat{\theta} = (y_1 + 2y_2 - y_3)/6$.
La suma de cuadrados residual es

$$\text{SCR} = \mathbf{Y}'\mathbf{Y} - \widehat{\theta}'\mathbf{X}'\mathbf{Y} = y_1^2 + y_2^2 + y_3^2 - (y_1 + 2y_2 - y_3)^2/6$$

Ejemplo 2.4.2

Supongamos que se desea pesar tres objetos cuyos pesos exactos son β_1, β_2 *y* β_3. *Se dispone de una balanza de platillos con un error de pesada que podemos considerar con distribución* $N(0, \sigma)$. *Un artificio para mejorar la precisión y ahorrar pesadas consiste en repartir los objetos en uno o en los dos platillos y anotar las sumas o diferencias de pesos:*

$$x_1\beta_1 + x_2\beta_2 + x_3\beta_3 = y$$

donde y *es el peso observado y* $x_i = 0, 1, -1$.
Consideremos las siguientes pesadas:

$$\begin{aligned}
\beta_1 + \beta_2 + \beta_3 &= 5{,}53 \\
\beta_1 - \beta_2 + \beta_3 &= 1{,}72 \\
\beta_1 + \beta_2 - \beta_3 &= 0{,}64 \\
\beta_1 + \beta_2 + \beta_3 &= 5{,}48 \\
\beta_1 - \beta_2 + \beta_3 &= 1{,}70
\end{aligned}$$

A partir de estos datos, las ecuaciones normales son

$$\left\{ \begin{array}{rcr}
5\beta_1 + \beta_2 + 3\beta_3 &=& 15{,}07 \\
\beta_1 + 5\beta_2 - \beta_3 &=& 8{,}23 \\
3\beta_1 - \beta_2 + 5\beta_3 &=& 13{,}79
\end{array} \right.$$

La estimación de los parámetros proporciona

$$\widehat{\beta}_1 = 1{,}175 \qquad \widehat{\beta}_2 = 1{,}898 \qquad \widehat{\beta}_3 = 2{,}433$$

y la suma de cuadrados residual es

$$\text{SCR} = (5{,}53 - (\widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3))^2 + \cdots = 0{,}00145$$

2.5. Estimación de la varianza

La varianza de los errores del modelo lineal

$$\sigma^2 = \text{var}(\epsilon_i) = \text{var}(y_i) \qquad i = 1, \ldots, n$$

es otro parámetro que debe ser estimado a partir de las observaciones de $y_1, \ldots, y_n$.

Teorema 2.5.1

Sea $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ el modelo lineal con las hipótesis impuestas en la sección 2.3. Entonces el estadístico[1]

$$\widehat{\sigma}^2 = \text{ECM} = \text{SCR}/(n-r)$$

es un estimador insesgado de la varianza σ^2. En este estadístico SCR es la suma de cuadrados residual, n el número total de observaciones y r el rango del diseño.

Demostración 1:

Las columnas $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(m)}$ de la matriz de diseño $\mathbf{X}$ generan el subespacio de dimensión r que escribimos

$$\langle \mathbf{X} \rangle = \langle \mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(m)} \rangle$$

Sea ahora $\mathbf{V}$ una matriz ortogonal, es decir, tal que $\mathbf{V}\mathbf{V}' = \mathbf{V}'\mathbf{V} = \mathbf{I}_n$, cuyas columnas $\mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(r)}, \mathbf{v}_{(r+1)}, \ldots, \mathbf{v}_{(n)}$ forman una base ortogonal de $\mathbb{R}^n$. Es posible construir $\mathbf{V}$ de modo que las r primeras columnas generen el subespacio $\langle \mathbf{X} \rangle$

$$\langle \mathbf{X} \rangle = \langle \mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(r)} \rangle$$

Por otra parte, $\mathbf{Y} = (y_1, \ldots, y_n)'$ es un vector aleatorio de $\mathbb{R}^n$ que, mediante $\mathbf{V}$, transformamos en $\mathbf{Z} = (z_1, \ldots, z_n)' = \mathbf{V}'\mathbf{Y}$

$$z_i = v_{1i}y_1 + \cdots + v_{ni}y_n \qquad i = 1, \ldots, n$$

Para las variables transformadas se verifica que

$$E(z_i) = \sum_{h=1}^{n} v_{hi}E(y_h) = \mathbf{v}'_{(i)}\mathbf{X}\boldsymbol{\beta} = \begin{cases} \eta_i & \text{si } i \leq r \\ 0 & \text{si } i > r \end{cases}$$

pues $\mathbf{X}\boldsymbol{\beta} \in \langle \mathbf{X} \rangle$ que es ortogonal a $\mathbf{v}_{(i)}$ para $i > r$.

Sea $\widehat{\boldsymbol{\beta}}$ una estimación MC. Entonces

$$\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{e}$$

donde obviamente $\mathbf{X}\widehat{\boldsymbol{\beta}} \in \langle \mathbf{X} \rangle$ y como sabemos $\mathbf{e} \in \langle \mathbf{X} \rangle^{\perp}$, de manera que la transformación ortogonal $\mathbf{V}'$ aplicada sobre $\mathbf{e}$ proporciona

$$\mathbf{V}'\mathbf{e} = (0, \ldots, 0, z_{r+1}, \ldots, z_n)'$$

Luego, en función de las variables z_i tenemos

$$\text{SCR} = \mathbf{e}'\mathbf{e} = (\mathbf{V}'\mathbf{e})'\mathbf{V}'\mathbf{e} = \sum_{i=r+1}^{n} z_i^2$$

Además, por ser una transformación ortogonal, las variables $z_1, \ldots, z_n$ siguen siendo incorrelacionadas y de varianza σ^2. Así pues

$$E(z_i) = 0 \qquad E(z_i^2) = \text{var}(z_i) = \text{var}(y_i) = \sigma^2$$

[1] En muchos de los libros clásicos escritos en inglés este estadístico se llama MSE, siglas de *mean square error*.

y por lo tanto

$$E(\text{SCR}) = \sum_{i=r+1}^{n} E(z_i^2) = (n-r)\sigma^2$$

La expresión

$$\text{SCR} = z_{r+1}^2 + \cdots + z_n^2 \tag{2.5}$$

se llama *forma canónica* de la suma de cuadrados residual del modelo lineal bajo las hipótesis de Gauss-Markov. ■

Demostración 2:

Se puede hacer una demostración mucho más directa a partir de la propiedad 2 explicada en el Apéndice C1 de Estadística Multivariante:

Para un vector aleatorio $\mathbf{Y}$ con esperanza $E(\mathbf{Y}) = \boldsymbol{\mu}$ y matriz de varianzas y covarianzas $\text{var}(\mathbf{Y}) = \mathbf{V}$, se tiene que

$$E(\mathbf{Y}'\mathbf{A}\mathbf{Y}) = \text{tr}(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$$

donde $\mathbf{A}$ es una matriz constante.

En nuestro caso $E(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ y $\text{var}(\mathbf{Y}) = \mathbf{V} = \sigma^2\mathbf{I}$, de forma que

$$\begin{aligned} E(\text{SCR}) = E(\mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y}) &= \text{tr}(\sigma^2(\mathbf{I}-\mathbf{P})) + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I}-\mathbf{P})\mathbf{X}\boldsymbol{\beta} \\ &= \sigma^2 \,\text{tr}(\mathbf{I}-\mathbf{P}) \\ &= \sigma^2 \,\text{rg}(\mathbf{I}-\mathbf{P}) = \sigma^2(n-r) \end{aligned}$$

gracias a las propiedades de la matriz $\mathbf{I}-\mathbf{P}$. ■

2.6. Distribuciones de los estimadores

Vamos ahora a establecer algunas propiedades de los estimadores MC para un modelo de rango máximo.

Si asumimos que los errores son insesgados $E(\boldsymbol{\epsilon}) = \mathbf{0}$, que es la primera condición de Gauss-Markov, entonces $\widehat{\boldsymbol{\beta}}$ es un estimador insesgado de $\boldsymbol{\beta}$

$$E(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

Si asumimos además que los errores ϵ_i son incorrelacionados y con la misma varianza, es decir $\text{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, resulta que

$$\text{var}(\mathbf{Y}) = \text{var}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \text{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$$

ya que $\mathbf{X}\boldsymbol{\beta}$ no es aleatorio y en consecuencia

$$\begin{aligned} \text{var}(\widehat{\boldsymbol{\beta}}) &= \text{var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{var}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Veamos a continuación algunos resultados acerca de la distribución de $\widehat{\boldsymbol{\beta}}$ y SCR bajo las hipótesis del modelo lineal normal en el caso de rango máximo.

Teorema 2.6.1

Sea $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ con rango $\mathbf{X} = m$. Entonces se verifican las siguientes propiedades:

i) La estimación MC de $\boldsymbol{\beta}$ coincide con la estimación de la máxima verosimilitud. Además es insesgada y de mínima varianza.

ii) $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

iii) $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_m$

iv) $\widehat{\boldsymbol{\beta}}$ es independiente de SCR

v) $\mathrm{SCR}/\sigma^2 \sim \chi^2_{n-m}$

Demostración:

i) La función de verosimilitud es

$$L(\mathbf{Y}; \boldsymbol{\beta}, \sigma^2) = (\sqrt{2\pi\sigma^2})^{-n}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right]$$

de modo que el mínimo de $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ es el máximo de L.

Ya hemos visto que $\widehat{\boldsymbol{\beta}}$ es insesgado y además, cada $\widehat{\beta}_i$ es un estimador lineal de varianza mínima de β_i, ya que es centrado y de máxima verosimilitud, luego suficiente. Se llegará a la misma conclusión como consecuencia del Teorema 3.2.1.

Por otra parte, si sustituimos $\boldsymbol{\beta}$ por $\widehat{\boldsymbol{\beta}}$ en la función de verosimilitud y derivamos respecto a σ^2 resulta que el el estimador de máxima verosimilitud de la varianza es

$$\widehat{\sigma}^2_{MV} = \mathrm{SCR}/n$$

Este estimador es sesgado y en la práctica no se utiliza, ya que disponemos del estimador insesgado propuesto en el apartado anterior. Además, bajo ciertas condiciones generales se puede probar que $\widehat{\sigma}^2 = \mathrm{SCR}/(n-m)$ es un estimador de varianza mínima de σ^2 (véase Seber [65, pág. 52]).

ii) Como $\widehat{\boldsymbol{\beta}} = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}$, $\widehat{\boldsymbol{\beta}}$ es combinación lineal de una normal y, por tanto, tiene distribución normal multivariante con matriz de varianzas-covarianzas

$$(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

iii) Es consecuencia de las propiedades de la normal multivariante del apartado anterior ya que

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathrm{var}(\widehat{\boldsymbol{\beta}})^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_m$$

iv) Si calculamos la matriz de covarianzas entre $\widehat{\boldsymbol{\beta}}$ i $\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ tenemos

$$\begin{aligned}\mathrm{cov}(\widehat{\boldsymbol{\beta}}, \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) &= \mathrm{cov}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, (\mathbf{I} - \mathbf{P})\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{var}(\mathbf{Y})(\mathbf{I} - \mathbf{P})' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{P}) = \mathbf{0}\end{aligned}$$

de modo que efectivamente $\widehat{\boldsymbol{\beta}}$ es independiente de $(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$, ya que la incorrelación entre normales multivariantes implica su independencia.

Este resultado se ampliará en el Teorema 3.4.1.

v) Aplicando la ecuación 2.5

$$\mathrm{SCR}/\sigma^2 = (z_{m+1}/\sigma)^2 + \cdots + (z_n/\sigma)^2$$

obtenemos una suma de cuadrados de $n-m$ variables normales independientes, es decir, una distribución χ^2_{n-m}.

■

Ejemplo 2.6.1

La distribución de $\widehat{\theta}$ del ejemplo 2.4.1 es $N(\theta, \sigma/\sqrt{6})$

$$\begin{aligned} E(\widehat{\theta}) &= E((y_1 + 2y_2 - y_3)/6) = (1/6)(\theta + 4\theta + \theta) = \theta \\ \mathrm{var}(\widehat{\theta}) &= (\sigma^2 + 4\sigma^2 + \sigma^2)/6^2 = \sigma^2/6 \end{aligned}$$

La distribución de SCR/σ^2 es χ^2_2, siendo

$$\mathrm{SCR} = (y_1 - \widehat{\theta})^2 + (y_2 - 2\widehat{\theta})^2 + (y_3 + \widehat{\theta})^2$$

Ejemplo 2.6.2

La estimación de la varianza del error σ^2 en el ejemplo 2.4.2 es

$$\widehat{\sigma}^2 = 0{,}00145/(5-3) = 0{,}725 \times 10^{-3}$$

Observemos que el número de pesadas necesarias para obtener la misma precisión sería mayor si pesáramos cada objeto individualmente.

2.7. Matriz de diseño reducida

Supongamos que varias observaciones y_i han sido obtenidas bajo las mismas condiciones experimentales. Para estas observaciones, el modelo que liga y_i con las $\boldsymbol{\beta}$ es el mismo, lo que se traduce en que las filas de la matriz de diseño correspondientes están repetidas. Para evitar la redundancia que esto supone nos será muy útil, a efectos teóricos y de cálculo, introducir el concepto de matriz de diseño reducida.

Definición 2.7.1

Dado el modelo lineal $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, llamaremos matriz de diseño reducida $\mathbf{X}_R$ a la matriz $k \times m$ obtenida tomando las k filas distintas de la matriz de diseño original $\mathbf{X}$. Diremos entonces que k es el número de condiciones experimentales.

Las matrices de diseño original o ampliada y reducida las indicaremos por $\mathbf{X}$ y $\mathbf{X}_R$ respectivamente, cuando convenga distinguir una de otra.

Si la fila i-ésima de $\mathbf{X}_R$ está repetida n_i veces en $\mathbf{X}$, significa que se han obtenido n_i réplicas de la variable observable bajo la i-ésima condición experimental. Si estos números de réplicas son $n_1, n_2, \ldots, n_k$, entonces

$$n = n_1 + n_2 + \cdots + n_k$$

Además de la matriz reducida $\mathbf{X}_R$, utilizaremos también la matriz diagonal

$$\mathbf{D} = \text{diag}(n_1, n_2, \ldots, n_k)$$

y el vector de medias

$$\bar{\mathbf{Y}} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k)'$$

donde cada $\overline{y}_i$ es la media de las réplicas bajo la condición experimental i.

En una experiencia bajo la cual todas las observaciones han sido tomadas en condiciones experimentales distintas (caso de una sola observación por casilla), entonces

$$\mathbf{X}_R = \mathbf{X} \qquad \bar{\mathbf{Y}} = \mathbf{Y} \qquad \mathbf{D} = \mathbf{I} \qquad n_i = 1$$

Como veremos más adelante (ver sección 10.7), la utilización de $\mathbf{X}_R$, $\mathbf{D}$ e $\bar{\mathbf{Y}}$ nos permitirá abordar diseños no balanceados y el caso de observaciones faltantes.

Teorema 2.7.1

La solución de las ecuaciones normales y la suma de cuadrados residual en términos de la matriz de diseño reducida $\mathbf{X}_R$, de $\mathbf{D}$ e $\bar{\mathbf{Y}}$ es

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}_R'\mathbf{D}\mathbf{X}_R)^{-1}\mathbf{X}_R'\mathbf{D}\bar{\mathbf{Y}}$$
$$\text{SCR} = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}_R'\mathbf{D}\bar{\mathbf{Y}}$$

Demostración:

Sea $\mathbf{M}$ una matriz $n \times k$ de forma que cada columna i es

$$(\underbrace{0, \ldots, 0}_{n'}, \underbrace{1, \ldots, 1}_{n_i}, \underbrace{0, \ldots, 0}_{n''})'$$

donde k es el número de condiciones experimentales (número de filas distintas de $\mathbf{X}$), n_i el número de réplicas bajo la condición i, y además

$$n' = n_1 + \cdots + n_{i-1} \qquad n'' = n_{i+1} + \cdots + n_k$$

Se verifica

$$\mathbf{M}'\mathbf{Y} = \mathbf{D}\bar{\mathbf{Y}} \qquad \mathbf{M}\mathbf{X}_R = \mathbf{X} \qquad \mathbf{M}'\mathbf{M} = \mathbf{D} \qquad \mathbf{X}'\mathbf{Y} = \mathbf{X}_R'\mathbf{M}'\mathbf{Y} = \mathbf{X}_R'\mathbf{D}\bar{\mathbf{Y}}$$

de donde se siguen inmediatamente las fórmulas del teorema. ■

Ejemplo 2.7.1

Con los datos del ejemplo 2.4.2

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \\ 1 & -1 & 1 \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} 5{,}53 \\ 1{,}72 \\ 0{,}64 \\ 5{,}48 \\ 1{,}70 \end{pmatrix}$$

Agrupando las filas $1,4$ *y* $2,5$ *obtenemos*

$$\mathbf{X}_R = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix} \qquad \mathbf{D} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

donde $n_1 = n_2 = 2$, $n_3 = 1$, $k = 3$.

$$\bar{\mathbf{Y}} = \begin{pmatrix} (5{,}53 + 5{,}48)/2 \\ (1{,}72 + 1{,}70)/2 \\ 0{,}64 \end{pmatrix} = \begin{pmatrix} 5{,}505 \\ 1{,}710 \\ 0{,}640 \end{pmatrix}$$

La matriz $\mathbf{M}$ *es*

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Ejemplo 2.7.2

Consideremos el modelo

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

correspondiente al diseño de dos factores sin interacción.

Supongamos que el primer factor tiene 2 *niveles y el segundo tiene* 3 *niveles, y que los números de réplicas son*

$$n_{11} = 2 \quad n_{21} = 1 \quad n_{12} = 3 \quad n_{22} = 3 \quad n_{13} = 5 \quad n_{23} = 4$$

La matriz de diseño reducida es

μ	α_1	α_2	β_1	β_2	β_3
1	1	0	1	0	0
1	0	1	1	0	0
1	1	0	0	1	0
1	0	1	0	1	0
1	1	0	0	0	1
1	0	1	0	0	1

Sin embargo, la matriz de diseño ampliada tiene 6 *columnas y* $\sum n_{ij} = 18$ *filas.*

2.8. Matrices de diseño de rango no máximo

Cuando el modelo lineal corresponde al análisis de los datos de un diseño experimental, la matriz $\mathbf{X}$ tiene todos sus elementos con valores 0 ó 1 y sus columnas acostumbran a ser linealmente dependientes. Ya sabemos que en este caso es posible hallar el estimador MC de $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$ pero, por desgracia, hay múltiples estimaciones de los parámetros $\boldsymbol{\beta}$ que más bien podemos considerar como soluciones $\widehat{\boldsymbol{\beta}}$ de las ecuaciones normales. En todo caso y como veremos en el próximo capítulo estamos interesados en concretar una estimación de los parámetros $\boldsymbol{\beta}$ aunque no sea única. A continuación se comentan algunos métodos para hallar una solución $\widehat{\boldsymbol{\beta}}$ o para hallar la SCR directamente.

2.8.1. Reducción a un modelo de rango máximo

Sea $\mathbf{X}_1$ la matriz $n \times r$ con las $r = \text{rg } \mathbf{X}$ columnas linealmente independientes de la matriz de diseño $\mathbf{X}$, entonces $\mathbf{P} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ de forma que

$$\text{SCR} = \mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\alpha}}'\mathbf{X}_1'\mathbf{Y}$$

donde $\widehat{\boldsymbol{\alpha}} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}$ es la solución del modelo $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\alpha} + \boldsymbol{\epsilon}$ de rango máximo.

Podemos asumir, sin pérdida de generalidad, que $\mathbf{X}_1$ está formada por las r primeras filas de $\mathbf{X}$ de manera que $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$. Entonces $\mathbf{X}_2 = \mathbf{X}_1\mathbf{F}$ ya que las columnas de $\mathbf{X}_2$ son linealmente dependientes de las de $\mathbf{X}_1$ y, por tanto, $\mathbf{X} = \mathbf{X}_1(\mathbf{I}_r, \mathbf{F})$. Así, éste es un caso especial de una factorización más general del tipo

$$\mathbf{X} = \mathbf{KL}$$

donde $\mathbf{K}$ es $n \times r$ de rango r, y $\mathbf{L}$ es $r \times m$ de rango r. Entonces podemos escribir

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{KL}\boldsymbol{\beta} = \mathbf{K}\boldsymbol{\alpha}$$

y estimar $\boldsymbol{\alpha}$.

2.8.2. Imposición de restricciones

Este método consiste en imponer un conjunto de restricciones del tipo $\mathbf{H}\boldsymbol{\beta} = \mathbf{0}$ para evitar la indeterminación de $\boldsymbol{\beta}$. Las restricciones apropiadas, llamadas identificables, son aquellas que, para cada $\boldsymbol{\theta} \in \Omega = \langle \mathbf{X} \rangle$, existe un único $\boldsymbol{\beta}$ que satisface $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$ y $\mathbf{0} = \mathbf{H}\boldsymbol{\beta}$, es decir, que satisface

$$\begin{pmatrix} \boldsymbol{\theta} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{H} \end{pmatrix} \boldsymbol{\beta} = \mathbf{G}\boldsymbol{\beta}$$

La solución es simple. Debemos elegir como filas de $\mathbf{H}$ un conjunto de $m - r$ vectores $m \times 1$ linealmente independientes que sean también linealmente independientes de las filas de $\mathbf{X}$. Entonces la matriz $\mathbf{G}$ de orden $(n + m - r) \times m$ tendrá rango m de modo que $\mathbf{G}'\mathbf{G} = \mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H}$ es $m \times m$ de rango m y en consecuencia tiene inversa. Luego hemos salvado la deficiencia en el rango de $\mathbf{X}'\mathbf{X}$ introduciendo la matriz $\mathbf{H}'\mathbf{H}$.

Así pues, si añadimos $\mathbf{H}'\mathbf{H}\boldsymbol{\beta} = \mathbf{0}$ a las ecuaciones normales tenemos

$$\mathbf{G}'\mathbf{G}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$$

cuya solución es $\widehat{\boldsymbol{\beta}} = (\mathbf{G}'\mathbf{G})^{-1}\mathbf{X}'\mathbf{Y}$. Se puede ver, a partir de $\widehat{\boldsymbol{\theta}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{PY}$, que $\mathbf{P} = \mathbf{X}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{X}'$ ya que $\mathbf{P}$ es única.

La demostración de todos los detalles aquí expuestos puede verse en Seber [65, pág. 74].

Es interesante comprobar que, si $\mathbf{H}\boldsymbol{\beta} = \mathbf{0}$, entonces

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}) &= (\mathbf{G}'\mathbf{G})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= (\mathbf{G}'\mathbf{G})^{-1}(\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H})\boldsymbol{\beta} = \boldsymbol{\beta} \end{aligned}$$

de modo que $\widehat{\boldsymbol{\beta}}$ es un estimador insesgado de $\boldsymbol{\beta}$.

Este método es particularmente útil en los modelos de análisis de la varianza para los que $\mathbf{H}$ se halla con mucha facilidad.

Ejemplo 2.8.1

Consideremos el modelo correspondiente al diseño de un factor con, por ejemplo, 3 niveles

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad i = 1, 2, 3 \quad j = 1, \ldots, n_i$$

entonces, tenemos $m = 4$ *y una matriz de diseño de rango 3. La estimación de los parámetros resulta indeterminada.*
Sin embargo, si añadimos la restricción $\sum \alpha_i = 0$*, es decir, si hacemos* $\mathbf{H} = (0, 1, 1, 1)$*, el sistema conjunto es de rango 4 y podemos determinar una solución o calcular la suma de cuadrados residual.*

2.9. Ejercicios

Ejercicio 2.1

Una variable Y toma los valores y_1, y_2 y y_3 en función de otra variable X con los valores x_1, x_2 y x_3. Determinar cuales de los siguientes modelos son lineales y encontrar, en su caso, la matriz de diseño para $x_1 = 1$, $x_2 = 2$ y $x_3 = 3$.

a) $y_i = \beta_0 + \beta_1 x_i + \beta_2(x_i^2 - 1) + \epsilon_i$

b) $y_i = \beta_0 + \beta_1 x_i + \beta_2 e^{x_i} + \epsilon_i$

c) $y_i = \beta_1 x_i(\beta_2 \text{tang}(x_i)) + \epsilon_i$

Ejercicio 2.2

Dado el modelo lineal

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \theta + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

hallar la estimación MC de θ y la suma de cuadrados residual.

Ejercicio 2.3

Si $\widehat{\boldsymbol{\beta}}$ es una estimación MC, probar que

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

Ejercicio 2.4

Cuatro objetos cuyos pesos exactos son β_1, β_2, β_3 y β_4 han sido pesados en una balanza de platillos de acuerdo con el siguiente esquema:

β_1	β_2	β_3	β_4	peso
1	1	1	1	9,2
1	−1	1	1	8,3
1	0	0	1	5,4
1	0	0	−1	−1,6
1	0	1	1	8,7
1	1	−1	1	3,5

Hallar las estimaciones de cada β_i y de la varianza del error.

Ejercicio 2.5

Sea $\widehat{\boldsymbol{\beta}}$ la estimación MC de $\boldsymbol{\beta}$. Si $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{Y}$, probar que la matriz $\mathbf{P}$ verifica

$$\mathbf{P}^2 = \mathbf{P} \qquad (\mathbf{I} - \mathbf{P})^2 = \mathbf{I} - \mathbf{P}$$

Ejercicio 2.6

La matriz de diseño reducida de un modelo lineal normal es

$$\mathbf{X}_R = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Se sabe además que

$$\overline{y}_1 = 10 \qquad \overline{y}_2 = 12 \qquad \overline{y}_3 = 17 \qquad n_1 = n_2 = n_3 = 10$$

$$s_1^2 = \frac{1}{n_1}\sum (y_{i1} - \overline{y}_1)^2 = 2{,}8 \qquad s_2^2 = 4{,}2 \qquad s_3^2 = 4{,}0$$

Se pide:

a) Hallar la expresión general de las estimaciones MC de los parámetros $\boldsymbol{\beta}$.

b) Calcular SCR. Estimar la varianza del diseño σ^2.

c) Estudiar si la hipótesis nula $H_0 : \sigma^2 = 3$ puede ser aceptada.

Ejercicio 2.7

Consideremos el modelo lineal

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \epsilon_i \qquad i = 1, \ldots, n$$

Sean $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_m$ las estimaciones MC de los parámetros y sea

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_m x_{im} \qquad i = 1, \ldots, n$$

Probar que

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i) = \sum_{i=1}^{n} e_i = 0$$

Capítulo 3

Funciones paramétricas estimables

3.1. Introducción

En los modelos lineales, además de la estimación de los parámetros β_i y de σ^2, interesa también la estimación de ciertas funciones lineales de los parámetros. Como vamos a ver, esto es especialmente necesario cuando los parámetros carecen de una estimación única.

Definición 3.1.1

Llamaremos función paramétrica *a toda función lineal ψ de los parámetros*

$$\psi = a_1\beta_1 + \cdots + a_m\beta_m = \mathbf{a}'\boldsymbol{\beta}$$

y diremos que una función paramétrica ψ es estimable *si existe un estadístico $\widehat{\psi}$, combinación lineal de las observaciones $y_1, \ldots, y_n$*

$$\widehat{\psi} = b_1y_1 + \cdots + b_ny_n = \mathbf{b}'\mathbf{Y}$$

tal que

$$E(\widehat{\psi}) = \psi$$

es decir, $\widehat{\psi}$ es estimador lineal insesgado de ψ.

Estas funciones paramétricas tienen la siguiente caracterización

Teorema 3.1.1

Sea $\psi = \mathbf{a}'\boldsymbol{\beta}$ una función paramétrica estimable asociada al modelo lineal $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Se verifica:

i) ψ es estimable si y sólo si el vector fila $\mathbf{a}'$ es combinación lineal de las filas de $\mathbf{X}$.

ii) Si $\psi_1, \ldots, \psi_q$ son funciones paramétricas estimables, entonces la combinación lineal $\psi = c_1\psi_1 + \cdots + c_q\psi_q$ es también función paramétrica estimable.

iii) El número máximo de funciones paramétricas estimables linealmente independientes es $r = \text{rango}(\mathbf{X})$.

Demostración:

i) Sea $\widehat{\psi} = \mathbf{b}'\mathbf{Y}$ tal que $E(\widehat{\psi}) = \psi$. Entonces

$$\mathbf{a}'\boldsymbol{\beta} = E(\mathbf{b}'\mathbf{Y}) = \mathbf{b}'E(\mathbf{Y}) = \mathbf{b}'\mathbf{X}\boldsymbol{\beta}$$

cualquiera que sea $\boldsymbol{\beta}$, luego

$$\mathbf{a}' = \mathbf{b}'\mathbf{X}$$

lo que nos dice que $\mathbf{a}'$ es combinación lineal de las filas de la matriz de diseño $\mathbf{X}$.

Recíprocamente, si suponemos que $\mathbf{b}'\mathbf{X} = \mathbf{a}'$, entonces basta tomar $\widehat{\psi} = \mathbf{b}'\mathbf{Y}$ como estimador lineal insesgado de ψ.

ii) y iii) para el lector (ver ejercicio 3.4) ■

Observaciones:

1) Si rango $\mathbf{X} = m$, entonces todos los parámetros β_i y todas las funciones paramétricas ψ son estimables, pues el subespacio generado por las filas de $\mathbf{X}$ coincide con $\mathbb{R}^m$.

2) Si rango $\mathbf{X} < m$, pueden construirse funciones paramétricas que no son estimables.

3) Una caracterización algebraica de que $\psi = \mathbf{a}'\boldsymbol{\beta}$ es estimable viene dada por la identidad

$$\mathbf{a}'(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X} = \mathbf{a}'$$

donde $(\mathbf{X}'\mathbf{X})^-$ representa una g-inversa de $\mathbf{X}'\mathbf{X}$.

En efecto, consideremos las matrices

$$\mathbf{S} = \mathbf{X}'\mathbf{X} \qquad \mathbf{S}^- = (\mathbf{X}'\mathbf{X})^- \qquad \mathbf{H} = \mathbf{S}^-\mathbf{S}$$

entonces se comprueba fácilmente que

$$\mathbf{H}^2 = \mathbf{H} \qquad \mathbf{SH} = \mathbf{S}$$

Puesto que $\mathbf{H}$ es idempotente

$$\text{rango } \mathbf{H} = \text{traza } \mathbf{H} = \text{rango } \mathbf{S} = \text{rango } \mathbf{X} = r$$

Por otra parte tenemos

$$\begin{aligned} \mathbf{0} &= \mathbf{S} - \mathbf{SH} = (\mathbf{I}_m - \mathbf{H})'(\mathbf{S} - \mathbf{SH}) = (\mathbf{I}_m - \mathbf{H})'(\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{XH}) \\ &= (\mathbf{I}_m - \mathbf{H})'(\mathbf{X}'(\mathbf{X} - \mathbf{XH})) = (\mathbf{X} - \mathbf{XH})'(\mathbf{X} - \mathbf{XH}) \end{aligned}$$

luego

$$\mathbf{X} = \mathbf{XH}$$

Entonces, si $\psi = \mathbf{a}'\boldsymbol{\beta}$ es estimable, $\mathbf{a}' = \mathbf{b}'\mathbf{X}$ y

$$\mathbf{a}'\mathbf{H} = \mathbf{b}'\mathbf{XH} = \mathbf{b}'\mathbf{X} = \mathbf{a}'$$

Recíprocamente, si $\mathbf{a}'\mathbf{H} = \mathbf{a}'$, resulta que

$$\mathbf{a}' = \mathbf{a}'\mathbf{S}^-\mathbf{S} = (\mathbf{a}'\mathbf{S}^-\mathbf{X}')\mathbf{X} = \mathbf{b}'\mathbf{X}$$

siendo $\mathbf{b}' = \mathbf{a}'\mathbf{S}^-\mathbf{X}'$.

3.2. Teorema de Gauss-Markov

Vamos a ver en primer lugar que, cuando el rango de la matriz de diseño no es máximo y, por tanto, la estimación MC de los parámetros no es única, la estimación de cualquier función paramétrica estimable utilizando cualquiera de los estimadores MC sí es única.

Teorema 3.2.1

Si $\psi = \mathbf{a}'\boldsymbol{\beta}$ una función paramétrica estimable y $\widehat{\boldsymbol{\beta}}$ es un estimador MC de $\boldsymbol{\beta}$, entonces el estimador $\widehat{\psi} = \mathbf{a}'\widehat{\boldsymbol{\beta}}$ de ψ es único.

Demostración:
Si ψ es una función paramétrica estimable, tiene un estimador lineal insesgado $\mathbf{b}'\mathbf{Y}$, donde $\mathbf{b}$ es un vector $n \times 1$. Consideremos el subespacio $\Omega = \langle \mathbf{X} \rangle$ de $\mathbb{R}^n$ generado por las columnas de $\mathbf{X}$. El vector $\mathbf{b}$ se puede descomponer de forma única

$$\mathbf{b} = \widetilde{\mathbf{b}} + \mathbf{c} \qquad \widetilde{\mathbf{b}} \in \Omega \qquad \mathbf{c} \perp \Omega$$

de modo que $\mathbf{c}$ es ortogonal a todo vector de Ω.
Consideremos ahora el estimador lineal $\widetilde{\mathbf{b}}'\mathbf{Y}$ y veamos que es insesgado y que su valor es único. Sabemos que $\mathbf{b}'\mathbf{Y}$ es insesgado

$$\psi = \mathbf{a}'\boldsymbol{\beta} = E(\mathbf{b}'\mathbf{Y}) = E(\widetilde{\mathbf{b}}'\mathbf{Y}) + E(\mathbf{c}'\mathbf{Y}) = E(\widetilde{\mathbf{b}}'\mathbf{Y}) = \widetilde{\mathbf{b}}'\mathbf{X}\boldsymbol{\beta} \tag{3.1}$$

luego $E(\widetilde{\mathbf{b}}'\mathbf{Y}) = \mathbf{a}'\boldsymbol{\beta}$, pues

$$E(\mathbf{c}'\mathbf{Y}) = \mathbf{c}'E(\mathbf{Y}) = \mathbf{c}'\mathbf{X}\boldsymbol{\beta} = \mathbf{0}\boldsymbol{\beta} = \mathbf{0}$$

Supongamos que $\mathbf{b}^{*\prime}\mathbf{Y}$ es otro estimador insesgado para ψ y $\mathbf{b}^* \in \Omega$. Entonces

$$\mathbf{0} = E(\widetilde{\mathbf{b}}'\mathbf{Y}) - E(\mathbf{b}^{*\prime}\mathbf{Y}) = (\widetilde{\mathbf{b}}' - \mathbf{b}^{*\prime})\mathbf{X}\boldsymbol{\beta}$$

luego

$$(\widetilde{\mathbf{b}}' - \mathbf{b}^{*\prime})\mathbf{X} = \mathbf{0}$$

lo que quiere decir que $(\widetilde{\mathbf{b}}' - \mathbf{b}^{*\prime})$ es ortogonal a Ω. Como también pertenece a Ω, debe ser $\widetilde{\mathbf{b}} - \mathbf{b}^* = \mathbf{0}$, es decir, $\widetilde{\mathbf{b}} = \mathbf{b}^*$.
Por último, sabemos que para cualquier estimador MC de $\boldsymbol{\beta}$ $\mathbf{e} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ es ortogonal a Ω, de manera que

$$\mathbf{0} = \widetilde{\mathbf{b}}'\mathbf{e} = \widetilde{\mathbf{b}}'\mathbf{Y} - \widetilde{\mathbf{b}}'\mathbf{X}\widehat{\boldsymbol{\beta}}$$

y así $\widetilde{\mathbf{b}}'\mathbf{Y} = \widetilde{\mathbf{b}}'\mathbf{X}\widehat{\boldsymbol{\beta}}$. Además, por 3.1 sabemos que $\widetilde{\mathbf{b}}'\mathbf{X} = \mathbf{b}'\mathbf{X} = \mathbf{a}'$, luego

$$\widetilde{\mathbf{b}}'\mathbf{Y} = \mathbf{a}'\widehat{\boldsymbol{\beta}}$$

para cualquier $\widehat{\boldsymbol{\beta}}$. ■
A continuación se demuestra la principal ventaja de la utilización de los estimadores MC.

Teorema 3.2.2 (Gauss-Markov)

Si $\psi = \mathbf{a}'\boldsymbol{\beta}$ una función paramétrica estimable y $\widehat{\boldsymbol{\beta}}$ es un estimador MC de $\boldsymbol{\beta}$, entonces $\widehat{\psi} = \mathbf{a}'\widehat{\boldsymbol{\beta}}$ es el estimador de varianza mínima[1] en la clase de los estimadores lineales insesgados de ψ.

[1] BLUE: best linear unbiased estimate

Demostración:
Con la notación

$$\|\mathbf{b}\|^2 = b_1^2 + \cdots + b_n^2$$

tenemos que

$$\text{var}(\mathbf{b}'\mathbf{Y}) = b_1^2\sigma^2 + \cdots + b_n^2\sigma^2 = \|\mathbf{b}\|^2\sigma^2$$

Si consideramos la descomposición de cualquier estimador insesgado de ψ que hemos utilizado en el teorema anterior y dado que

$$\|\mathbf{b}\|^2 = \|\widetilde{\mathbf{b}}\|^2 + \|\mathbf{c}\|^2$$

resulta

$$\text{var}(\mathbf{a}'\widehat{\boldsymbol{\beta}}) = \text{var}(\widetilde{\mathbf{b}}'\mathbf{Y}) = \|\widetilde{\mathbf{b}}\|^2\sigma^2 \leq (\|\widetilde{\mathbf{b}}\|^2 + \|\mathbf{c}\|^2)\sigma^2 = \text{var}(\mathbf{b}'\mathbf{Y})$$

■

Observaciones:

1) Estos resultados son válidos incluso para un modelo lineal sin la hipótesis de normalidad.

2) La estimación con varianza mínima es

$$\widehat{\psi} = \mathbf{a}'(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}$$

3) Como la varianza de $\mathbf{b}'\mathbf{Y}$ es $\mathbf{b}'\mathbf{b}\sigma^2$, resulta que la varianza mínima es

$$\text{var}(\widehat{\psi}) = \text{var}(\mathbf{a}'\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^-\mathbf{a}$$

4) Utilizando la matriz de diseño reducida tenemos

$$\widehat{\psi} = \mathbf{a}'(\mathbf{X}_R'\mathbf{D}\mathbf{X}_R)^-\mathbf{X}_R'\mathbf{D}\bar{\mathbf{Y}}$$

$$\text{var}(\widehat{\psi}) = \sigma^2\mathbf{a}'(\mathbf{X}_R'\mathbf{D}\mathbf{X}_R)^-\mathbf{a}$$

De aquí deducimos que $\widehat{\psi}$ es combinación lineal de las medias de las k condiciones experimentales

$$\widehat{\psi} = c_1\bar{Y}_1 + \cdots + c_k\bar{Y}_k = \mathbf{c}'\bar{\mathbf{Y}}$$

donde $\mathbf{c} = (c_1, \ldots, c_k)'$ es

$$\mathbf{c} = \mathbf{D}\mathbf{X}_R(\mathbf{X}_R'\mathbf{D}\mathbf{X}_R)^-\mathbf{a}$$

Entonces

$$\text{var}(\widehat{\psi}) = \left(\sum_{i=1}^{k} c_i^2/n_i\right)\sigma^2 = \delta^2\sigma^2$$

Por otra parte, todo estimador lineal insesgado $\widehat{\psi} = \mathbf{b}'\mathbf{Y}$ de $\psi = \mathbf{a}'\boldsymbol{\beta}$ se descompone como hemos visto en

$$\mathbf{b}'\mathbf{Y} = \widetilde{\mathbf{b}}'\mathbf{Y} + \mathbf{c}'\mathbf{Y}$$

Diremos que $\widetilde{\mathbf{b}}'\mathbf{Y}$ (donde $\widetilde{\mathbf{b}}$ es único) pertenece al *espacio estimación* y que $\mathbf{c}'\mathbf{Y}$ pertenece al *espacio error*.

Más explícitamente, la descomposición de $\mathbf{b}'$ es

$$\mathbf{b}' = \mathbf{b}'\mathbf{P} + \mathbf{b}'(\mathbf{I} - \mathbf{P})$$

siendo $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ la matriz del operador que proyecta $\mathbf{b}$ en $\Omega = \langle \mathbf{X} \rangle$ (ver Apéndice B). El vector proyectado es $\widetilde{\mathbf{b}}' = \mathbf{b}'\mathbf{P}$. Asimismo, $\mathbf{I} - \mathbf{P}$ es otro operador que proyecta $\mathbf{b}$ en el espacio ortogonal a Ω. La proyección es $\mathbf{c}' = \mathbf{b}'(\mathbf{I} - \mathbf{P})$. Como $\widetilde{\mathbf{b}}'\mathbf{c} = \mathbf{0}$, se verifica

$$\mathrm{cov}(\widetilde{\mathbf{b}}'\mathbf{Y}, \mathbf{c}'\mathbf{Y}) = 0$$

Así pues, todo estimador lineal insesgado $\mathbf{b}'\mathbf{Y}$ se descompone en

$$\mathbf{b}'\mathbf{Y} = \mathbf{b}'\mathbf{P}\mathbf{Y} + \mathbf{b}'(\mathbf{I} - \mathbf{P})\mathbf{Y}$$

donde $\mathbf{b}'\mathbf{P}\mathbf{Y}$ es el estimador de Gauss-Markov, mientras que $\mathbf{b}'(\mathbf{I} - \mathbf{P})\mathbf{Y}$ tiene esperanza cero y provoca un aumento de la varianza mínima del mejor estimador $\widehat{\psi} = \mathbf{b}'\mathbf{P}\mathbf{Y}$.
Finalmente, observemos que

$$\begin{aligned} \widehat{\psi} = \mathbf{b}'\mathbf{P}\mathbf{Y} = \mathbf{b}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} = \mathbf{b}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \\ = \mathbf{b}'\mathbf{X}\mathbf{H}\widehat{\boldsymbol{\beta}} = \mathbf{a}'\widehat{\boldsymbol{\beta}} \end{aligned} \tag{3.2}$$

Siendo $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$, que verifica $\mathbf{X}\mathbf{H} = \mathbf{X}$, y siendo $\mathbf{a}' = \mathbf{b}'\mathbf{X}$.
El aspecto geométrico de las estimaciones se puede resumir en el hecho que el espacio muestral $\mathbb{R}^n$ al que pertenece el vector de observaciones $\mathbf{Y}$, se descompone en

$$\mathbb{R}^n = \Omega + \Omega^{\perp}$$

donde Ω representa el *espacio estimación*. Toda estimación de los parámetros de regresión está ligada a Ω. Toda estimación de la varianza del modelo está ligada al *espacio error* $\Omega^{\perp}$. Ambos espacios son ortogonales y bajo el modelo lineal normal, como veremos más adelante, ambas clases de estimaciones son estocásticamente independientes.

Ejemplo 3.2.1

Sea $y_1, \ldots, y_n$ una muestra aleatoria simple procedente de una población $N(\mu, \sigma)$. El modelo lineal asociado es

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \boldsymbol{\epsilon}$$

El estimador MC de μ es $\widehat{\mu} = (1/n)\sum y_i$ que también es de Gauss-Markov (centrado y de varianza mínima).
En este caso $\mathbb{R}^n = \Omega + \Omega^{\perp}$, siendo

$$\begin{aligned} \Omega &= \langle (1, \ldots, 1)' \rangle \\ \Omega^{\perp} &= \{(x_1, \ldots, x_n)' | \sum x_i = 0\} \end{aligned}$$

Sea $\mathbf{a}'\mathbf{Y} = \sum a_i y_i$ otro estimador centrado de μ. Entonces $E(\mathbf{a}'\mathbf{Y}) = \mu$ implica $\sum a_i = 1$. Luego se verifica $\mathbf{a} = \widetilde{\mathbf{a}} + \mathbf{b}$, es decir,

$$\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} 1/n \\ \vdots \\ 1/n \end{pmatrix} + \begin{pmatrix} a_1 - 1/n \\ \vdots \\ a_n - 1/n \end{pmatrix}$$

con $\widetilde{\mathbf{a}} \in \Omega$, $\mathbf{b} \in \Omega^{\perp}$. *Es fácil ver que* $\widetilde{\mathbf{a}}'\mathbf{b} = 0$. *Además*

$$\sum a_i y_i = (1/n)\sum y_i + \sum (a_i - 1/n) y_i$$

El primer término es estimador centrado y de varianza mínima σ^2/n. *El segundo término verifica*

$$E(\sum (a_i - 1/n) y_i) = 0$$
$$\mathrm{cov}(1/n \sum y_i, \sum (a_i - 1/n) y_i) = 0$$

La matriz del operador que proyecta $\mathbf{a}$ *en* Ω *es*

$$\mathbf{P} = 1/n \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (1, \ldots, 1) = \begin{pmatrix} 1/n & \ldots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \ldots & 1/n \end{pmatrix}$$

siendo fácil ver que

$$\mathbf{a}'\mathbf{P} = (1/n, \ldots, 1/n)$$
$$\mathbf{a}'(\mathbf{I} - \mathbf{P}) = (a_1 - 1/n, \ldots, a_n - 1/n)$$

Ejemplo 3.2.2

Ver especialmente el final del ejemplo 5.3.2.

3.3. Varianza de la estimación y multicolinealidad

Sabemos que $\mathbf{a}'\boldsymbol{\beta}$ se dice estimable si tiene un estimador lineal insesgado $\mathbf{b}'\mathbf{Y}$ o, equivalentemente, cuando $\mathbf{a} = \mathbf{X}'\mathbf{b}$. Es decir, cuando $\mathbf{a}$ es combinación lineal de las filas de la matriz $\mathbf{X}$.

Teorema 3.3.1

La función paramétrica $\mathbf{a}'\boldsymbol{\beta}$ es estimable si y sólo si

$$a \in \langle \mathbf{X}' \rangle = \langle \mathbf{X}'\mathbf{X} \rangle$$

Demostración:
Como sabemos, la función paramétrica $\mathbf{a}'\boldsymbol{\beta}$ es estimable si y sólo si $\mathbf{a}$ es combinación lineal de las filas de $\mathbf{X}$, es decir, cuando $a \in \langle \mathbf{X}' \rangle$. De modo que sólo queda probar que

$$\langle \mathbf{X}' \rangle = \langle \mathbf{X}'\mathbf{X} \rangle$$

Pero $\mathbf{X}'\mathbf{X}\mathbf{c} = \mathbf{X}'\mathbf{d}$ para $\mathbf{d} = \mathbf{X}\mathbf{c}$, de forma que $\langle \mathbf{X}'\mathbf{X} \rangle \subset \langle \mathbf{X}' \rangle$. Además, las dimensiones de ambos subespacios son iguales ya que rg $\mathbf{X}'$ = rg $\mathbf{X}'\mathbf{X}$, de donde se deduce la igualdad.
Los detalles pueden verse en Seber [65, pág. 385]. ■

En el apartado anterior hemos demostrado que para una función paramétrica estimable $\mathbf{a}'\boldsymbol{\beta}$, su estimador MC $\mathbf{a}'\widehat{\boldsymbol{\beta}}$ es el de mínima varianza. Pero, ¿cuanto vale esta varianza?

Supongamos que $\mathbf{X}'\mathbf{X}$ tiene como valores propios $\lambda_1, \ldots, \lambda_r$ todos positivos no nulos asociados a los correspondientes vectores propios ortonormales $\mathbf{v}_1, \ldots, \mathbf{v}_r$, es decir

$$\mathbf{X}'\mathbf{X}\mathbf{v}_i = \lambda_i \mathbf{v}_i \qquad i = 1, \ldots, r$$

y tales que $\mathbf{v}_i'\mathbf{v}_j = \delta_{ij}$.

Si $\mathbf{a}'\boldsymbol{\beta}$ es estimable, entonces $\mathbf{a} \in \langle \mathbf{X}'\mathbf{X} \rangle$ y este subespacio está generado por los vectores propios. Así pues, $\mathbf{a}$ se puede expresar en la forma

$$\mathbf{a} = \sum_{i=1}^{r} c_i \mathbf{v}_i$$

Entonces

$$\begin{aligned} \mathrm{var}(\mathbf{a}'\widehat{\boldsymbol{\beta}}) &= \mathrm{var}\left(\sum_i c_i \mathbf{v}_i'\widehat{\boldsymbol{\beta}} \right) \\ &= \sum_i c_i^2 \mathrm{var}(\mathbf{v}_i'\widehat{\boldsymbol{\beta}}) \\ &= \sigma^2 \sum_i c_i^2 \lambda_i^{-1} \end{aligned}$$

ya que

$$\begin{aligned} \mathrm{cov}(\mathbf{v}_i'\widehat{\boldsymbol{\beta}}, \mathbf{v}_j'\widehat{\boldsymbol{\beta}}) &= \lambda_i^{-1}\lambda_j^{-1}\mathrm{cov}(\mathbf{v}_i'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}, \mathbf{v}_j'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= (\lambda_i\lambda_j)^{-1}\mathrm{cov}(\mathbf{v}_i'\mathbf{X}'\mathbf{Y}, \mathbf{v}_j'\mathbf{X}'\mathbf{Y}) \\ &= (\lambda_i\lambda_j)^{-1}\sigma^2\mathbf{v}_i'\mathbf{X}'\mathbf{X}\mathbf{v}_j \\ &= (\lambda_i\lambda_j)^{-1}\sigma^2\lambda_j\mathbf{v}_i'\mathbf{v}_j \\ &= \sigma^2\lambda_i^{-1}\delta_{ij} \end{aligned}$$

Silvey[67] concluyó que es posible una estimación relativamente precisa en las direcciones de los vectores propios de $\mathbf{X}'\mathbf{X}$ correspondientes a los mayores valores propios, mientras que se obtienen unas estimaciones relativamente imprecisas (poco eficientes) en las direcciones correspondientes a los valores propios más pequeños.

Supongamos que $\mathbf{X}$ tiene rango máximo pero que sus columnas están cerca de ser linealmente dependientes. Entonces $\mathbf{X}'\mathbf{X}$ está cerca de ser singular (no inversible), en el sentido que uno o varios de sus valores propios no nulos son excesivamente pequeños, casi despreciables, y por lo que hemos visto las estimaciones en algunas direcciones serán muy imprecisas.

La presencia de relaciones quasi lineales entre las variables regresoras se conoce en Econometría con el nombre de *multicolinealidad*, cuya forma más extrema se presenta cuando la matriz de datos $\mathbf{X}$ no tiene rango máximo. Este grave problema debe ser detectado previamente a la estimación y se puede corregir de varias formas (ver sección 8.5).

Una solución teórica consiste en minimizar o incluso erradicar la multicolinealidad, mediante la incorporación de nuevas observaciones en las direcciones de los vectores propios con valores propios demasiado pequeños (o cero).

Supongamos que una nueva observación se añade al modelo $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ y resulta

$$\begin{aligned} \begin{pmatrix} \mathbf{Y} \\ Y_{n+1} \end{pmatrix} &= \begin{pmatrix} \mathbf{X} \\ \mathbf{x}_{n+1}' \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\epsilon} \\ \epsilon_{n+1} \end{pmatrix} \\ &= \mathbf{X}_*\boldsymbol{\beta} + \boldsymbol{\epsilon}_* \end{aligned}$$

donde $\mathbf{x}_{n+1} = c\mathbf{v}$, donde $\mathbf{v}$ es un vector propio normalizado de $\mathbf{X}'\mathbf{X}$ correspondiente a un valor propio λ. Entonces se puede probar que $\mathbf{v}$ es también un vector propio de $\mathbf{X}_*'\mathbf{X}_*$ correspondiente al valor propio $\lambda + c^2$. Y de esta forma Sylvey propuso un análisis para la elección de las direcciones en las que es conveniente elegir nuevas observaciones para mejorar la precisión de las estimaciones de un $\mathbf{a}'\boldsymbol{\beta}$ particular.

3.4. Sistemas de funciones paramétricas estimables

Consideremos un sistema de funciones paramétricas estimables

$$\psi_1 = \mathbf{a}_1'\boldsymbol{\beta}, \ldots, \psi_q = \mathbf{a}_q'\boldsymbol{\beta}$$

sobre el mismo modelo lineal normal y donde los vectores $\mathbf{a}_1, \ldots, \mathbf{a}_q$ ($q \leq r = \text{rango } \mathbf{X}$) son linealmente independientes. Para cada una, tenemos las correspondientes estimaciones de Gauss-Markov

$$\widehat{\psi}_i = \mathbf{a}_i'\widehat{\boldsymbol{\beta}} \qquad i = 1, \ldots, q$$

que podemos condensar matricialmente en la forma

$$\widehat{\boldsymbol{\psi}} = (\widehat{\psi}_1, \ldots, \widehat{\psi}_q)' = \mathbf{A}\widehat{\boldsymbol{\beta}}$$

donde

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_q' \end{pmatrix}$$

Con esta matriz, $\widehat{\boldsymbol{\psi}}$ es el conjunto de estimadores MC del sistema de funciones paramétricas $\boldsymbol{\psi} = \mathbf{A}\boldsymbol{\beta}$.

Teorema 3.4.1

Bajo el modelo lineal normal, el conjunto de estimadores $\widehat{\boldsymbol{\psi}} = \mathbf{A}\widehat{\boldsymbol{\beta}}$ del sistema de funciones paramétricas $\boldsymbol{\psi} = \mathbf{A}\boldsymbol{\beta}$ verifica:

i) $\widehat{\boldsymbol{\psi}}$ sigue la distribución normal multivariante

$$\widehat{\boldsymbol{\psi}} \sim N_q(\boldsymbol{\psi}, \boldsymbol{\Sigma}_{\boldsymbol{\psi}})$$

donde $\boldsymbol{\psi} = \mathbf{A}\boldsymbol{\beta}$ es el vector de medias y

$$\boldsymbol{\Sigma}_{\boldsymbol{\psi}} = \sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{A}'$$

es la matriz de varianzas-covarianzas.

ii) La estimación MC de toda función paramétrica estimable es estocásticamente independiente de la suma de cuadrados residual

$$\text{SCR} = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

En particular, $\widehat{\boldsymbol{\psi}} = \mathbf{A}\widehat{\boldsymbol{\beta}}$ es estocásticamente independiente de SCR.

Demostración:

i) Es consecuencia de que $\widehat{\boldsymbol{\psi}}$ es una combinación lineal de variables normales independientes:

$$\widehat{\psi}_i = \mathbf{a}_i'(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}$$

luego si

$$\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{X}' = \mathbf{C}$$

sabemos que $E(\widehat{\boldsymbol{\psi}}) = \boldsymbol{\psi}$ y la matriz de covarianzas de $\mathbf{CY}$ es $\boldsymbol{\Sigma} = \sigma^2\mathbf{CC}'$, de manera que

$$\boldsymbol{\Sigma}_{\boldsymbol{\psi}} = \sigma^2\mathbf{CC}' = \sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{A}' = \sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{A}'$$

ii) Como en el teorema 2.5.1, consideremos la transformación ortogonal

$$\mathbf{Z} = \mathbf{V}'\mathbf{Y}$$

donde las primeras r columnas de la matriz ortogonal $\mathbf{V}$ generan el subespacio $\Omega = \langle\mathbf{X}\rangle$. Entonces las variables $z_1, \ldots, z_n$ son normales e independientes, y toda estimación de Gauss-Markov es una combinación lineal de

$$z_1, \ldots, z_r$$

puesto que pertenece al *espacio estimación*. Sin embargo, la suma de cuadrados residual es

$$\text{SCR} = z_{r+1}^2 + \cdots + z_n^2$$

y, por tanto, será estocásticamente independiente de cualquier estimación $\widehat{\psi}_i = \mathbf{a}_i'\widehat{\boldsymbol{\beta}}$.

Esto mismo se puede deducir de la expresión 3.2 ya que $\widehat{\boldsymbol{\psi}} = \mathbf{BPY}$, mientras que

$$\text{SCR} = \mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y} = ((\mathbf{I}-\mathbf{P})\mathbf{Y})'(\mathbf{I}-\mathbf{P})\mathbf{Y}$$

donde $(\mathbf{I}-\mathbf{P})\mathbf{Y}$ pertenece al espacio ortogonal de Ω.

■

Teorema 3.4.2

La distribución de $U = (\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta})'(\sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta})$ es una χ_q^2.
Además, U es estocásticamente independiente de SCR/σ^2 cuya distribución es χ_{n-r}^2.

Demostración:
Es consecuencia de las propiedades de la distribución normal multivariante y de los teoremas 2.5.1 y 3.4.1. ■

Dos resultados importantes que se deducen de los teoremas anteriores son:

a) Para el modelo lineal normal y el sistema de q funciones paramétricas estimables $\boldsymbol{\psi} = \mathbf{A}\boldsymbol{\beta}$ se verifica que la distribución de

$$F = \frac{(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta})/q}{\text{SCR}/(n-r)} \tag{3.3}$$

es una F con q y $n-r$ grados de libertad, ya que se trata de un cociente de dos χ^2 independientes divididas por sus grados de libertad respectivos. Observemos la desaparición del parámetro σ^2 desconocido.

b) En el caso $q = 1$, si $\widehat{\psi}$ es la estimación de Gauss-Markov de ψ, entonces $\widehat{\psi} \sim N(\psi, \sigma_{\widehat{\psi}})$, siendo

$$\sigma^2_{\widehat{\psi}} = \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{a}\,\sigma^2 = \delta^2\sigma^2$$

luego la distribución de

$$t = \frac{\widehat{\psi} - \psi}{\sqrt{\delta^2 \mathrm{SCR}}}\sqrt{n-r} \tag{3.4}$$

es la de una t de Student con $n - r$ grados de libertad. Este resultado se puede establecer directamente o a partir de 3.3 ya que $F_{1,n-r} = t^2_{n-r}$.

3.5. Intervalos de confianza

Consideremos una función paramétrica estimable $\psi = \mathbf{a}'\boldsymbol{\beta}$, su estimación MC $\widehat{\psi} = \mathbf{a}'\widehat{\boldsymbol{\beta}}$ y sea t_α tal que

$$P(-t_\alpha < t < t_\alpha) = 1 - \alpha$$

para una distribución t de Student con $n - r$ grados de libertad. Entonces, de la distribución 3.4 deducimos que

$$P\left(-t_\alpha < \frac{\widehat{\psi} - \psi}{\sqrt{\delta^2 \mathrm{SCR}}}\sqrt{n-r} < t_\alpha\right) = 1 - \alpha$$

y despejando obtenemos

$$P\left(\widehat{\psi} - t_\alpha\sqrt{\frac{\delta^2\mathrm{SCR}}{n-r}} < \psi < \widehat{\psi} + t_\alpha\sqrt{\frac{\delta^2\mathrm{SCR}}{n-r}}\right) = 1 - \alpha$$

Por lo tanto

$$\widehat{\psi} - t_\alpha\sqrt{\frac{\delta^2\mathrm{SCR}}{n-r}} < \psi < \widehat{\psi} + t_\alpha\sqrt{\frac{\delta^2\mathrm{SCR}}{n-r}}$$

es decir

$$\mathbf{a}'\widehat{\boldsymbol{\beta}} \pm t_\alpha[\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{a}\,\widehat{\sigma}^2]^{1/2} \tag{3.5}$$

es un intervalo de confianza para la función paramétrica estimable $\psi = \mathbf{a}'\boldsymbol{\beta}$, con coeficiente de confianza $1 - \alpha$.

Por otra parte, como SCR/σ^2 sigue una χ^2_{n-r} tenemos

$$P(a < \mathrm{SCR}/\sigma^2 < b) = 1 - \alpha$$

donde a y b son tales que

$$P(\chi^2_{n-r} \le a) = \alpha/2 \qquad\qquad P(\chi^2_{n-r} > b) = \alpha/2$$

Deducimos entonces que

$$P\left(\frac{\mathrm{SCR}}{b} < \sigma^2 < \frac{\mathrm{SCR}}{a}\right) = 1 - \alpha \tag{3.6}$$

define un intervalo de confianza para la varianza σ^2 del modelo lineal normal, con coeficiente de confianza $1 - \alpha$.

3.6. Ejercicios

Ejercicio 3.1

Sea ψ una función paramétrica estimable y $\widehat{\psi}_1$, $\widehat{\psi}_2$ dos estimadores insesgados, estocásticamente independientes, de varianzas σ_1^2 y σ_2^2. Hallar la combinación lineal de $\widehat{\psi}_1$, $\widehat{\psi}_2$ cuya varianza es mínima y además es insesgado.

Ejercicio 3.2

En un modelo lineal, la matriz de diseño es

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Hallar la expresión general de las funciones paramétricas estimables.

Ejercicio 3.3

Probar que

$$\widehat{\psi} = \mathbf{b}'\mathbf{Y} \qquad E(\widehat{\psi}) = \psi = \mathbf{a}'\boldsymbol{\beta}$$

siendo $\mathbf{b}$ combinación lineal de las columnas de $\mathbf{X}$, implica que $\mathbf{a}$ es combinación lineal de las filas de $\mathbf{X}$.

Ejercicio 3.4

Probar que toda combinación lineal de funciones paramétricas estimables es también función paramétrica estimable y que $r = \text{rg } \mathbf{X}$ es el número máximo de funciones linealmente independientes.

Ejercicio 3.5

Si $\widehat{\psi}$ es la estimación de Gauss-Markov, probar que la expresión

$$\widehat{\psi} = c_1\bar{y}_1 + \cdots + c_k\bar{y}_k$$

función de las medias de las condiciones experimentales, es única.

Ejercicio 3.6

La matriz de diseño reducida correspondiente a un modelo lineal normal es

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$$

Se sabe además que

$$\bar{y}_1 = 11 \qquad \bar{y}_2 = 10 \qquad \bar{y}_3 = 15$$
$$n_1 = n_2 = n_3 = 10$$
$$s_1^2 = (1/n_1)\sum_{i=1}^{n_1}(y_i - \bar{y}_1)^2 = 4{,}5$$
$$s_2^2 = 6{,}0 \qquad s_3^2 = 4{,}3$$

Se pide

1) Hallar la expresión general de las estimaciones MC de $\boldsymbol{\beta}$.

2) Calcular SCR. ¿Se ajustan los datos al modelo definido por $\mathbf{X}$? (nivel de significación 0,05)

3) Dada la función paramétrica estimable

$$\psi = \beta_1 + \beta_3$$

contrastar la hipótesis $H_0 : \psi = 3$ en los casos:

a) σ^2 varianza del diseño desconocida

b) $\sigma^2 = 5$ varianza del diseño conocida

(nivel de significación 0,05)

4) Hallar la función paramétrica estimable ψ tal que

$$\widehat{\psi} = c_1\bar{y}_1 + c_2\bar{y}_2 + c_3\bar{y}_3$$

verifica $c_1^2 + c_2^2 + c_3^2 = 1$ y además $\widehat{\psi}$ es máximo.

Ejercicio 3.7

Consideremos el modelo lineal

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 + \epsilon_1 \\ y_2 &= \beta_1 + \beta_3 + \epsilon_2 \\ y_3 &= \beta_1 + \beta_2 + \epsilon_3 \end{aligned}$$

Se pide:

1) ¿Es la función paramétrica

$$\psi = \beta_1 + \beta_2 + \beta_3$$

estimable?

2) Probar que toda función paramétrica

$$\psi = a_1\beta_1 + a_2\beta_2 + a_3\beta_3$$

es estimable si y sólo si $a_1 = a_2 + a_3$.

Ejercicio 3.8

Consideremos el modelo lineal

$$\begin{aligned} y_1 &= \mu + \alpha_1 + \beta_1 + \epsilon_1 \\ y_2 &= \mu + \alpha_1 + \beta_2 + \epsilon_2 \\ y_3 &= \mu + \alpha_2 + \beta_1 + \epsilon_3 \\ y_4 &= \mu + \alpha_2 + \beta_2 + \epsilon_4 \\ y_5 &= \mu + \alpha_3 + \beta_1 + \epsilon_5 \\ y_6 &= \mu + \alpha_3 + \beta_2 + \epsilon_6 \end{aligned}$$

(a) ¿Cuando es $\lambda_0\mu + \lambda_1\alpha_1 + \lambda_2\alpha_2 + \lambda_3\alpha_3 + \lambda_4\beta_1 + \lambda_5\beta_2$ estimable?

(b) ¿Es $\alpha_1 + \alpha_2$ estimable?

(c) ¿Es $\beta_1 - \beta_2$ estimable?

(d) ¿Es $\mu + \alpha_1$ estimable?

(e) ¿Es $6\mu + 2\alpha_1 + 2\alpha_2 + 2\alpha_3 + 3\beta_1 + 3\beta_2$ estimable?

(f) ¿Es $\alpha_1 - 2\alpha_2 + \alpha_3$ estimable?

(g) Hallar la covarianza entre los estimadores lineales MC de las funciones paramétricas $\beta_1 - \beta_2$ y $\alpha_1 - \alpha_2$, si éstas son estimables.

(h) Hallar la dimensión del espacio paramétrico.

(i) Obtener una expresión del espacio de los errores.

Ejercicio 3.9

Cuatro objetos A, B, C, D están involucrados en un experimento de pesado. Todos reunidos pesan y_1 gramos. Cuando A y C se ponen en el plato izquierdo de la balanza y B y D se ponen en el plato derecho, un peso de y_2 gramos es necesario en el plato derecho para equilibrar la balanza. Con A y B en el plato izquierdo y C, D en el plato derecho, y_3 gramos son necesarios en el plato derecho y, finalmente, con A, D en el plato izquierdo y B, C en el plato derecho, y_4 gramos son necesarios en la derecha para equilibrar. Si las observaciones y_1, y_2, y_3, y_4 son todas con errores incorrelacionados y con varianza común σ^2, obtener la estimación BLUE del peso total de los cuatro objetos y su varianza.

Ejercicio 3.10

Con el modelo lineal

$$\begin{aligned}
y_1 &= \theta_1 + \theta_5 + \epsilon_1 \\
y_2 &= \theta_2 + \theta_5 + \epsilon_2 \\
y_3 &= \theta_3 + \theta_6 + \epsilon_3 \\
y_4 &= \theta_4 + \theta_6 + \epsilon_4 \\
y_5 &= \theta_1 + \theta_7 + \epsilon_5 \\
y_6 &= \theta_3 + \theta_7 + \epsilon_6 \\
y_7 &= \theta_2 + \theta_8 + \epsilon_7 \\
y_8 &= \theta_4 + \theta_8 + \epsilon_8
\end{aligned}$$

contestar las siguientes preguntas:

(a) ¿Cuantas funciones paramétricas son estimables? Obtener el conjunto completo de todas ellas.

(b) Probar que $\theta_1 - \theta_2$ es estimable. Calcular su estimador lineal MC y su varianza.

(c) Probar que $\theta_1 + \theta_2$ no es estimable.

(d) Hallar cuatro estimadores insesgados diferentes de $\theta_1 - \theta_2$ y calcular sus varianzas. Compararlas con la varianza del estimador MC.

(e) Hallar un estimador insesgado de la varianza de los errores σ^2.

Ejercicio 3.11

Diremos que el estimador lineal $\mathbf{b}'\mathbf{Y}$ pertenece al espacio error si $E(\mathbf{b}'\mathbf{Y}) = \mathbf{0}$. Probar que la covarianza entre $\mathbf{b}'\mathbf{Y}$ y todo estimador de Gauss-Markov $\widehat{\psi} = \mathbf{a}'\boldsymbol{\beta}$ es siempre cero.

Ejercicio 3.12

Consideremos el modelo lineal normal $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, siendo rg $\mathbf{X} = r$. Sea $\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}'$ una descomposición en valores singulares de $\mathbf{X}$. Se pide:

1) Expresar la estimación MC de $\boldsymbol{\beta}$ en términos de $\mathbf{U}$, $\boldsymbol{\Delta}$, $\mathbf{V}$ y $\mathbf{Y}$.

2) Sea $\psi = \mathbf{a}'\boldsymbol{\beta}$ una función paramétrica. Probar que ψ es estimable si y sólo si se verifica

$$\mathbf{a}' = \mathbf{b}'\mathbf{V}'$$

para algún vector $\mathbf{b}$.

Capítulo 4

Complementos de estimación

En este capítulo se presentan algunas extensiones del método de los mínimos cuadrados. Estos complementos no son estrictamente necesarios para continuar con el desarrollo de la teoría de los modelos lineales y, en particular, para el contraste de hipótesis que se explica en el capítulo 5. En una primera lectura de este libro se puede pasar directamente a ese capítulo.

4.1. Ampliar un modelo con más variables regresoras

4.1.1. Una variable extra

Supongamos que después de ajustar el modelo lineal

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \qquad \text{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$$

decidimos introducir una nueva variable regresora con las mismas observaciones que ya teníamos.

Sean $\mathbf{x}_{(i)}$, $i = 1, \ldots, m$ las columnas de la matriz $\mathbf{X}$ $n \times m$ de rango m de modo que

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(m)})\boldsymbol{\beta} = \mathbf{x}_{(1)}\beta_1 + \cdots + \mathbf{x}_{(m)}\beta_m$$

La inclusión de la nueva variable regresora $\mathbf{x}_{(m+1)}$ proporciona un modelo ampliado

$$G : E(\mathbf{Y}) = \mathbf{x}_{(1)}\beta_1 + \cdots + \mathbf{x}_{(m)}\beta_m + \mathbf{x}_{(m+1)}\beta_{m+1} = \mathbf{X}\boldsymbol{\beta} + \mathbf{x}_{(m+1)}\beta_{m+1} = \mathbf{G}\boldsymbol{\gamma}$$

donde la matriz $\mathbf{G} = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(m)}, \mathbf{x}_{(m+1)})$ es $n \times (m+1)$ de rango $m+1$.

Para hallar la estimación de los $m+1$ parámetros $\boldsymbol{\gamma} = (\beta_1, \ldots, \beta_m, \beta_{m+1})'$ podemos hacerlo directamente como

$$\widehat{\boldsymbol{\gamma}}_G = (\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{Y} \qquad \text{var}(\widehat{\boldsymbol{\gamma}}_G) = \sigma^2(\mathbf{G}'\mathbf{G})^{-1}$$

o a partir del modelo original que ya hemos resuelto. Vamos a ver el desarrollo de esta segunda opción que proporciona unos cálculos más simples.

Partimos de las ecuaciones normales del modelo ampliado $\mathbf{G}'\mathbf{G}\widehat{\boldsymbol{\gamma}}_G = \mathbf{G}'\mathbf{Y}$ que podemos descomponer así

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}_G + \mathbf{X}'\mathbf{x}_{(m+1)}\widehat{\beta}_{m+1} = \mathbf{X}'\mathbf{Y}$$
$$\mathbf{x}'_{(m+1)}\mathbf{X}\widehat{\boldsymbol{\beta}}_G + \mathbf{x}'_{(m+1)}\mathbf{x}_{(m+1)}\widehat{\beta}_{m+1} = \mathbf{x}'_{(m+1)}\mathbf{Y}$$

De la primera ecuación tenemos

$$\widehat{\boldsymbol{\beta}}_G = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{x}_{(m+1)}\widehat{\beta}_{m+1}) = \widehat{\boldsymbol{\beta}} - \mathbf{f}\widehat{\beta}_{m+1} \tag{4.1}$$

donde $\mathbf{f} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{x}_{(m+1)}$, y sustituyendo en la segunda

$$\mathbf{x}'_{(m+1)}\mathbf{x}_{(m+1)}\widehat{\beta}_{m+1} = \mathbf{x}'_{(m+1)}\mathbf{Y} - \mathbf{x}'_{(m+1)}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{x}_{(m+1)}\widehat{\beta}_{m+1})$$

es decir

$$\mathbf{x}'_{(m+1)}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{x}_{(m+1)}\widehat{\beta}_{m+1} = \mathbf{x}'_{(m+1)}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}$$

de manera que

$$\widehat{\beta}_{m+1} = [\mathbf{x}'_{(m+1)}(\mathbf{I} - \mathbf{P})\mathbf{x}_{(m+1)}]^{-1}\mathbf{x}'_{(m+1)}(\mathbf{I} - \mathbf{P})\mathbf{Y} = g\mathbf{x}'_{(m+1)}(\mathbf{I} - \mathbf{P})\mathbf{Y} \tag{4.2}$$

donde $g = [\mathbf{x}'_{(m+1)}(\mathbf{I} - \mathbf{P})\mathbf{x}_{(m+1)}]^{-1}$ es un escalar.
Observemos que ahora este resultado se puede sustituir en la ecuación 4.1 de modo que $\widehat{\boldsymbol{\beta}}_G$ queda determinado.
Por otra parte

$$\begin{aligned}\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_G - \mathbf{x}_{(m+1)}\widehat{\beta}_{m+1} &= \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{x}_{(m+1)}\widehat{\beta}_{m+1}) - \mathbf{x}_{(m+1)}\widehat{\beta}_{m+1} \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{Y} - \mathbf{x}_{(m+1)}\widehat{\beta}_{m+1}) \\ &= (\mathbf{I} - \mathbf{P})(\mathbf{Y} - \mathbf{x}_{(m+1)}\widehat{\beta}_{m+1})\end{aligned}$$

de manera que la suma de cuadrados de los residuos para el modelo ampliado es

$$\begin{aligned}\mathrm{SCR}_G &= (\mathbf{Y} - \mathbf{G}\widehat{\boldsymbol{\gamma}}_G)'(\mathbf{Y} - \mathbf{G}\widehat{\boldsymbol{\gamma}}_G) \\ &= (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_G - \mathbf{x}_{(m+1)}\widehat{\beta}_{m+1})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_G - \mathbf{x}_{(m+1)}\widehat{\beta}_{m+1}) \\ &= (\mathbf{Y} - \mathbf{x}_{(m+1)}\widehat{\beta}_{m+1})'(\mathbf{I} - \mathbf{P})(\mathbf{Y} - \mathbf{x}_{(m+1)}\widehat{\beta}_{m+1})\end{aligned}$$

ya que $\mathbf{I} - \mathbf{P}$ es simétrica e idempotente.
Si desarrollamos esta expresión se obtiene

$$\begin{aligned}\mathrm{SCR}_G &= \mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y} - \mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{x}_{(m+1)}\widehat{\beta}_{m+1} \\ &\quad - \mathbf{x}'_{(m+1)}(\mathbf{I} - \mathbf{P})\mathbf{Y}\widehat{\beta}_{m+1} + \mathbf{x}'_{(m+1)}(\mathbf{I} - \mathbf{P})\mathbf{x}_{(m+1)}\widehat{\beta}^2_{m+1} \\ &= \mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y} - \mathbf{x}'_{(m+1)}(\mathbf{I} - \mathbf{P})\mathbf{Y}\widehat{\beta}_{m+1} \\ &\quad - [\mathbf{x}'_{(m+1)}(\mathbf{I} - \mathbf{P})\mathbf{Y} - \mathbf{x}'_{(m+1)}(\mathbf{I} - \mathbf{P})\mathbf{x}_{(m+1)}\widehat{\beta}_{m+1}]\widehat{\beta}_{m+1}\end{aligned}$$

y por 4.2 resulta

$$\mathrm{SCR}_G = \mathrm{SCR} - \mathbf{x}'_{(m+1)}(\mathbf{I} - \mathbf{P})\mathbf{Y}\widehat{\beta}_{m+1} \tag{4.3}$$

En cuanto a las varianzas y covarianzas de los estimadores se tiene lo siguiente: A partir de la ecuación 4.2 tenemos

$$\mathrm{var}(\widehat{\beta}_{m+1}) = \sigma^2(\mathbf{x}'_{(m+1)}(\mathbf{I} - \mathbf{P})\mathbf{x}_{(m+1)})^{-1} = \sigma^2 g$$

Además

$$\begin{aligned}\mathrm{cov}(\widehat{\boldsymbol{\beta}}, \widehat{\beta}_{m+1}) &= \mathrm{cov}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, g\mathbf{x}'_{(m+1)}(\mathbf{I} - \mathbf{P})\mathbf{Y}] \\ &= \sigma^2 g(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{P})\mathbf{x}_{(m+1)} = \mathbf{0}\end{aligned}$$

ya que $\mathbf{X}'(\mathbf{I}-\mathbf{P})=\mathbf{0}$. Esto permite calcular la covarianza entre $\widehat{\boldsymbol{\beta}}_G$ y $\widehat{\beta}_{m+1}$

$$\begin{aligned}\mathrm{cov}(\widehat{\boldsymbol{\beta}}_G,\widehat{\beta}_{m+1}) &= \mathrm{cov}[\widehat{\boldsymbol{\beta}}-\mathbf{f}\widehat{\beta}_{m+1},\widehat{\beta}_{m+1}]\\ &= \mathrm{cov}(\widehat{\boldsymbol{\beta}},\widehat{\beta}_{m+1})-\mathbf{f}\mathrm{var}(\widehat{\beta}_{m+1})\\ &= \mathbf{0}-\mathbf{f}\sigma^2 g\end{aligned}$$

Finalmente

$$\begin{aligned}\mathrm{var}(\widehat{\boldsymbol{\beta}}_G) &= \mathrm{var}(\widehat{\boldsymbol{\beta}}-\mathbf{f}\widehat{\beta}_{m+1})\\ &= \mathrm{var}(\widehat{\boldsymbol{\beta}})-2\mathrm{cov}(\widehat{\boldsymbol{\beta}},\mathbf{f}\widehat{\beta}_{m+1})+\mathrm{var}(\mathbf{f}\widehat{\beta}_{m+1})\\ &= \mathrm{var}(\widehat{\boldsymbol{\beta}})-2\mathrm{cov}(\widehat{\boldsymbol{\beta}},\widehat{\beta}_{m+1})\mathbf{f}'+\mathbf{f}\mathrm{var}(\widehat{\beta}_{m+1})\mathbf{f}'\\ &= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}+g\mathbf{f}\mathbf{f}']\end{aligned}$$

En resumen

$$\mathrm{var}(\widehat{\boldsymbol{\gamma}}_G)=\sigma^2\begin{pmatrix}(\mathbf{X}'\mathbf{X})^{-1}+g\mathbf{f}\mathbf{f}' & -g\mathbf{f}\\ -g\mathbf{f}' & g\end{pmatrix} \tag{4.4}$$

donde $g=[\mathbf{x}'_{(m+1)}(\mathbf{I}-\mathbf{P})\mathbf{x}_{(m+1)}]^{-1}$ y $\mathbf{f}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{x}_{(m+1)}$.
En consecuencia, las fórmulas 4.1, 4.2, 4.3 y 4.4 demuestran que es posible calcular todos los elementos del modelo ampliado a partir del modelo original, mediante productos de matrices en los que interviene únicamente la matriz $(\mathbf{X}'\mathbf{X})^{-1}$ original.

4.1.2. Una interpretación

Partimos del modelo

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}\qquad E(\boldsymbol{\epsilon})=\mathbf{0},\mathrm{var}(\boldsymbol{\epsilon})=\sigma^2\mathbf{I} \tag{4.5}$$

donde $\mathbf{X}=(\mathbf{x}_{(1)},\ldots,\mathbf{x}_{(m)})$ y $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_m)'$, y queremos ampliar el modelo con una nueva variable regresora para llegar al modelo

$$G:\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{x}_{(m+1)}\beta_{m+1}+\boldsymbol{\epsilon}_G=\mathbf{G}\boldsymbol{\gamma}+\boldsymbol{\epsilon}_G \tag{4.6}$$

donde $\mathbf{G}=(\mathbf{x}_{(1)},\ldots,\mathbf{x}_{(m)},\mathbf{x}_{(m+1)})$ y $\boldsymbol{\gamma}=(\beta_1,\ldots,\beta_m,\beta_{m+1})'$.
Consideremos $\widehat{\boldsymbol{\beta}}$ la estimación MC en el modelo original, de forma que

$$\mathbf{Y}=\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{e} \tag{4.7}$$

donde $\mathbf{e}$ es el vector de residuos o parte de $\mathbf{Y}$ no explicada linealmente por $\mathbf{X}$.
Sea $\widehat{\mathbf{c}}$ la estimación MC en el modelo lineal $\mathbf{x}_{(m+1)}=\mathbf{X}\mathbf{c}+\boldsymbol{\epsilon}_{m+1}$, de forma que

$$\mathbf{x}_{(m+1)}=\mathbf{X}\widehat{\mathbf{c}}+\mathbf{e}_{m+1} \tag{4.8}$$

donde el vector de residuos $\mathbf{e}_{m+1}$ representa la parte de $\mathbf{x}_{(m+1)}$ no explicada linealmente por las variables anteriores.
Consideremos ahora la regresión lineal simple de $\boldsymbol{\epsilon}$ (parte de $\mathbf{Y}$ no explicada por $\mathbf{X}$) con $\boldsymbol{\epsilon}_{m+1}$ (parte de $\mathbf{x}_{(m+1)}$ independiente de $\mathbf{X}$)

$$\mathbf{e}=\mathbf{e}_{m+1}\widehat{d}+\mathbf{e}^* \tag{4.9}$$

Teorema 4.1.1 Si consideramos las estimaciones MC que se han calculado en las ecuaciones 4.7, 4.8 y 4.9, resulta que la estimación MC de β_{m+1} en el modelo ampliado 4.6 es $\widehat{\beta}_{m+1} = \widehat{d}$.

Demostración:
Si sustituimos 4.9 en la ecuación 4.7, se obtiene

$$\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{e}_{m+1}\widehat{d} + \mathbf{e}^* = \mathbf{X}\widehat{\boldsymbol{\beta}} + (\mathbf{x}_{(m+1)} - \mathbf{X}\widehat{\mathbf{c}})\widehat{d} + \mathbf{e}^*$$

La solución MC del modelo ampliado es

$$\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}}_G + \mathbf{x}_{(m+1)}\widehat{\beta}_{m+1} + \mathbf{e}_G$$

donde $\widehat{\boldsymbol{\beta}}_G = \widehat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{x}_{(m+1)}\widehat{\beta}_{m+1}$ como hemos visto en 4.1. De forma que

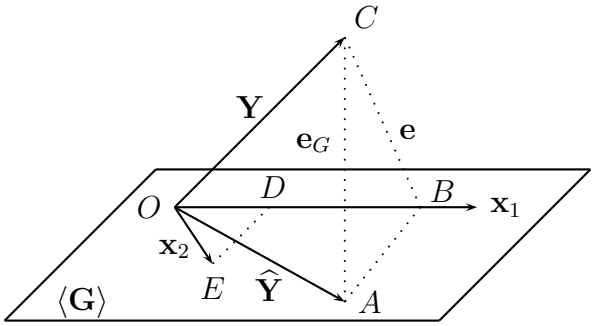
$$\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + (\mathbf{x}_{(m+1)} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{x}_{(m+1)})\widehat{\beta}_{m+1} + \mathbf{e}_G$$

Pero por 4.8 sabemos que $\widehat{\mathbf{c}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{x}_{(m+1)}$, de manera que

$$\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + (\mathbf{x}_{(m+1)} - \mathbf{X}\widehat{\mathbf{c}})\widehat{\beta}_{m+1} + \mathbf{e}_G$$

y entonces $\widehat{\beta}_{m+1} = \widehat{d}$ y $\mathbf{e}_G = \mathbf{e}^*$. ■

En el gráfico se dibuja la consecuencia de añadir a un modelo con una variable regresora x_1 una nueva variable x_2.



En este gráfico tenemos los siguientes datos:

$$ED = \mathbf{e}_{m+1} \quad OD = \mathbf{x}_1\widehat{c} \quad AB = \mathbf{e}_{m+1}\widehat{d} \quad OB = \mathbf{x}_1\widehat{\beta}$$

de forma que

$$ED||AB \quad BC \perp OB \quad ED \perp OD \quad AB \perp OB \quad AC \perp OA$$

y en especial

$$\widehat{\mathbf{Y}} = \overrightarrow{OB} + \overrightarrow{AB}$$

Como conclusión podemos decir que cualquier coeficiente estimado $\widehat{\beta}_i$ puede interpretarse como la pendiente de la recta que relaciona los residuos de la regresión de Y respecto a todas las otras variables, es decir, la parte de Y no explicada por el resto de las variables regresoras, con la aportación diferencial de x_i o parte de x_i no común con las demás

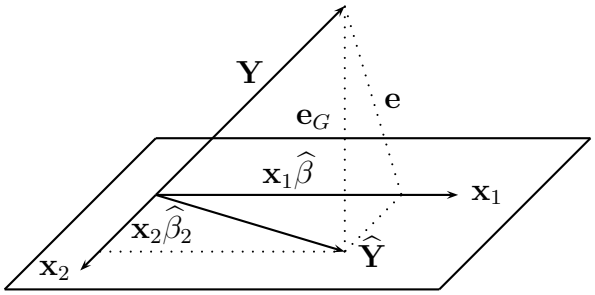
variables regresoras que se obtiene tomando el residuo de la regresión de x_i sobre las restantes x.

Observemos que cuando $\mathbf{x}_{(m+1)}$ es independiente de $\mathbf{X}$ el paso 4.8 no es posible. En esta situación

$$\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{e}$$
$$\mathbf{e} = \mathbf{x}_{(m+1)}\widehat{\beta}_{m+1} + \mathbf{e}_G$$

de modo que la solución del modelo ampliado es

$$\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{x}_{(m+1)}\widehat{\beta}_{m+1} + \mathbf{e}_G$$



Esto significa que si excluimos del modelo variables regresoras independientes, esto no afecta a la estimación de los parámetros β_i, pero si excluimos variables relevantes esto afecta considerablemente a las estimaciones.

4.1.3. Más variables

Supongamos que después de ajustar el modelo lineal

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \qquad \text{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$$

decidimos introducir un grupo de variables regresoras. El modelo es ahora

$$G : E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\delta} \end{pmatrix} = \mathbf{W}\boldsymbol{\gamma}$$

y vamos a suponer que las matrices son de rango máximo, de forma que $\mathbf{X}$ es $n \times m$ de rango m, $\mathbf{Z}$ es $n \times t$ de rango t, y las columnas de $\mathbf{Z}$ son linealmente independientes de las columnas de $\mathbf{X}$, de forma que $\mathbf{W}$ es $n \times (m+t)$ de rango $m+t$.

Si queremos hallar el estimador mínimo cuadrático $\widehat{\boldsymbol{\gamma}}_G$ de $\boldsymbol{\gamma}$, podemos hacerlo a partir del modelo completo G

$$\widehat{\boldsymbol{\gamma}}_G = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y} \qquad \text{var}(\widehat{\boldsymbol{\gamma}}_G) = \sigma^2(\mathbf{W}'\mathbf{W})^{-1}$$

o reducir los cálculos utilizando los resultados del modelo inicial. El siguiente teorema resume las principales propiedades de esta segunda propuesta.

Teorema 4.1.2

Consideremos las matrices $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{P}_G = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$, $\mathbf{L} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$, $\mathbf{M} = (\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Z})^{-1}$ y el vector

$$\widehat{\boldsymbol{\gamma}}_G = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_G \\ \widehat{\boldsymbol{\delta}}_G \end{pmatrix}$$

Entonces,

(i) $\widehat{\boldsymbol{\beta}}_G = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\delta}}_G) = \widehat{\boldsymbol{\beta}} - \mathbf{L}\widehat{\boldsymbol{\delta}}_G$

(ii) $\widehat{\boldsymbol{\delta}}_G = (\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Y}$

(iii) $\mathrm{SCR}_G = \mathbf{Y}'(\mathbf{I}-\mathbf{P}_G)\mathbf{Y} = (\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\delta}}_G)'(\mathbf{I}-\mathbf{P})(\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\delta}}_G)$

(iv) $\mathrm{SCR}_G = \mathrm{SCR} - \widehat{\boldsymbol{\delta}}_G'\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Y}$

(v)

$$\mathrm{var}(\widehat{\boldsymbol{\gamma}}_G) = \sigma^2 \begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} + \mathbf{L}\mathbf{M}\mathbf{L}' & -\mathbf{L}\mathbf{M} \\ -\mathbf{M}\mathbf{L}' & \mathbf{M} \end{pmatrix}$$

Demostración:
Se puede reseguir sin mayor dificultad todos los cálculos que hemos realizado en el apartado anterior. El único detalle importante es que debe demostrarse que la matriz $\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Z}$ es inversible. Este resultado y los detalles de la demostración pueden verse en Seber [65, pág. 65]. ■

A partir de estas fórmulas se deduce que, una vez invertida la matriz $\mathbf{X}'\mathbf{X}$, podemos hallar $\widehat{\boldsymbol{\gamma}}_G$ y su matriz de varianzas-covarianzas $\mathrm{var}(\widehat{\boldsymbol{\gamma}}_G)$ simplemente invirtiendo $\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Z}$ $t \times t$ y no se necesita calcular la inversa de la matriz $\mathbf{W}'\mathbf{W}$ $(m+t)\times(m+t)$.

Estos resultados se pueden utilizar de diversas formas en modelos de Análisis de la Varianza y de Análisis de la Covarianza. Para introducir un grupo de variables en un modelo de regresión es mejor hacerlo de una en una, lo que se llama regresión paso a paso.

4.2. Mínimos cuadrados generalizados

Hasta este momento se ha presentado la teoría de los modelos lineales $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ con la asunción de las hipótesis $E(\boldsymbol{\epsilon}) = \mathbf{0}$ y $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. Vamos ahora a estudiar lo que ocurre cuando permitimos a los ϵ_i ser correlacionados. En particular, vamos a considerar el modelo lineal más general

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad E(\boldsymbol{\epsilon}) = \mathbf{0}, \mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V} \tag{4.10}$$

donde $\mathbf{V}$ es una matriz $n \times n$ definida positiva con valores plenamente conocidos.

Dado que $\mathbf{V}$ es definida positiva, existe una matriz $n \times n$ $\mathbf{K}$ no singular tal que $\mathbf{V} = \mathbf{K}\mathbf{K}'$ y con la que podemos transformar el modelo anterior

$$\begin{array}{ccccc} \mathbf{K}^{-1}\mathbf{Y} & = & \mathbf{K}^{-1}\mathbf{X}\boldsymbol{\beta} & + & \mathbf{K}^{-1}\boldsymbol{\epsilon} \\ \mathbf{Z} & = & \mathbf{B}\boldsymbol{\beta} & + & \boldsymbol{\eta} \end{array} \tag{4.11}$$

donde $\mathbf{B}$ es $n \times r$, $\mathrm{rg}\mathbf{B} = \mathrm{rg}\mathbf{X}$ y además

$$E(\boldsymbol{\eta}) = \mathbf{K}^{-1}E(\boldsymbol{\epsilon}) = \mathbf{0}$$
$$\mathrm{var}(\boldsymbol{\eta}) = \sigma^2\mathbf{K}^{-1}\mathbf{V}(\mathbf{K}^{-1})' = \sigma^2\mathbf{I}$$

de forma que el modelo 4.11 verifica las condiciones del modelo lineal ordinario. Así es posible calcular el estimador MC de $\boldsymbol{\beta}$ que minimiza $\boldsymbol{\eta}'\boldsymbol{\eta}$.

Definición 4.2.1

Un estimador $\boldsymbol{\beta}^*$ es un estimador MCG de $\boldsymbol{\beta}$ para el modelo 4.10 si y sólo si $\boldsymbol{\beta}^*$ es un estimador MC ordinario para el modelo 4.11. En el caso particular de que la matriz $\mathbf{V}$ sea diagonal se llama MC ponderado.

En consecuencia, un estimador MCG $\boldsymbol{\beta}^*$ de $\boldsymbol{\beta}$ satisface la ecuación

$$\begin{aligned}&\mathbf{B}(\mathbf{B}'\mathbf{B})^{-}\mathbf{B}'\mathbf{Z} = \mathbf{B}\boldsymbol{\beta}^*\\ &\mathbf{K}^{-1}\mathbf{X}((\mathbf{K}^{-1}\mathbf{X})'\mathbf{K}^{-1}\mathbf{X})^{-}(\mathbf{K}^{-1}\mathbf{X})'\mathbf{K}^{-1}\mathbf{Y} = \mathbf{K}^{-1}\mathbf{X}\boldsymbol{\beta}^*\\ &\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^*\end{aligned}$$

Como un estimador MCG es simplemente un estimador MC ordinario del modelo transformado, es de esperar que tenga las mismas propiedades óptimas.

Propiedades

(a) Si $\mathbf{X}$ es de rango máximo, la estimación MC se puede obtener de las ecuaciones normales

$$\boldsymbol{\beta}^* = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{Z} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

con las siguientes propiedades

$$\begin{aligned}E(\boldsymbol{\beta}^*) &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}\\ \text{var}(\boldsymbol{\beta}^*) &= \sigma^2(\mathbf{B}'\mathbf{B})^{-1} = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\\ \text{SCR} &= (\mathbf{Z}-\mathbf{B}\boldsymbol{\beta}^*)'(\mathbf{Z}-\mathbf{B}\boldsymbol{\beta}^*) = (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^*)'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^*)\end{aligned}$$

(b) Una función paramétrica $\mathbf{a}'\boldsymbol{\beta}$ es estimable en el modelo 4.10 si y sólo si es estimable en el modelo 4.11.

En efecto, si $\mathbf{a}'\boldsymbol{\beta}$ es estimable en el modelo 4.10 podemos escribir

$$\mathbf{a}' = \mathbf{b}'\mathbf{X} = (\mathbf{b}'\mathbf{K})\mathbf{K}^{-1}\mathbf{X} = \mathbf{c}'\mathbf{B}$$

luego también es estimable en el modelo 4.11.

Si $\mathbf{a}'\boldsymbol{\beta}$ es estimable en el modelo 4.11, entonces

$$\mathbf{a}' = \mathbf{c}'\mathbf{B} = \mathbf{c}'\mathbf{K}^{-1}\mathbf{X} = (\mathbf{c}'\mathbf{K}^{-1})\mathbf{X} = \mathbf{b}'\mathbf{X}$$

luego es estimable en el modelo 4.10.

(c) Para una f.p.e. $\mathbf{a}'\boldsymbol{\beta}$, el estimador MCG es el mejor estimador lineal, en el sentido de insesgado y de varianza mínima, y además es único.

Aplicando el teorema 3.2.1 de Gauss-Markov al modelo 4.11, sabemos que $\mathbf{a}'\boldsymbol{\beta}^*$ es el estimador lineal insesgado y de mínima varianza entre todas las combinaciones lineales del vector $\mathbf{K}^{-1}\mathbf{Y}$. Sin embargo, cualquier combinación lineal de $\mathbf{Y}$ se puede obtener de $\mathbf{K}^{-1}\mathbf{Y}$ porque $\mathbf{K}^{-1}$ es inversible. Luego el estimador MCG es el mejor. También por una propiedad anterior sabemos que es único.

Para un modelo de rango no máximo y en el caso ordinario hemos visto que un estimador debe verificar la ecuación $\mathbf{PY} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, donde $\mathbf{P}$ es el operador proyección ortogonal sobre el subespacio $\langle \mathbf{X} \rangle$. Veamos una propiedad similar en el caso generalizado.

Teorema 4.2.1

Un estimador MCG $\boldsymbol{\beta}^*$ en el modelo 4.10 verifica la ecuación $\mathbf{AY} = \mathbf{X}\boldsymbol{\beta}^*$ donde $\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}$ es una matriz idempotente pero no, en general, simétrica.

Demostración:
Se trata de probar que $\mathbf{A}$ es una especie de operador proyección sobre $\langle \mathbf{X} \rangle$ aunque no necesariamente ortogonal.
Por la definición de estimador MCG ya hemos visto que

$$\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} = \mathbf{AY} = \mathbf{X}\boldsymbol{\beta}^*$$

Es fácil ver que $\mathbf{AA} = \mathbf{A}$ de manera que $\mathbf{A}$ es idempotente y no necesariamente simétrica, veamos ahora que $\mathbf{A}$ es un operador proyección sobre $\langle \mathbf{X} \rangle$, en el sentido de que $\langle \mathbf{A} \rangle = \langle \mathbf{X} \rangle$ de modo que $\mathbf{AY} \in \langle \mathbf{X} \rangle$.
La proyección ortogonal sobre $\langle \mathbf{K}^{-1}\mathbf{X} \rangle$ es

$$\mathbf{K}^{-1}\mathbf{X}[(\mathbf{K}^{-1}\mathbf{X})'(\mathbf{K}^{-1}\mathbf{X})]^-(\mathbf{K}^{-1}\mathbf{X})'$$

Por la definición de proyección se verifica

$$\begin{aligned}
&\mathbf{K}^{-1}\mathbf{X}[(\mathbf{K}^{-1}\mathbf{X})'(\mathbf{K}^{-1}\mathbf{X})]^-(\mathbf{K}^{-1}\mathbf{X})'\mathbf{K}^{-1}\mathbf{X} = \mathbf{K}^{-1}\mathbf{X} \\
&\mathbf{K}^{-1}\mathbf{AX} = \mathbf{K}^{-1}\mathbf{X} \\
&\mathbf{AX} = \mathbf{X}
\end{aligned}$$

y en consecuencia $\langle \mathbf{X} \rangle \subset \langle \mathbf{A} \rangle$. Pero también tenemos que

$$\mathbf{A} = \mathbf{X}[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}]$$

y por tanto $\langle \mathbf{A} \rangle \subset \langle \mathbf{X} \rangle$. ■
Para una función paramétrica estimable $\mathbf{a}'\boldsymbol{\beta}$ con $\mathbf{a}' = \mathbf{b}'\mathbf{X}$, el estimador MCG es $\mathbf{a}'\boldsymbol{\beta}^* = \mathbf{b}'\mathbf{AY}$. Vamos a calcular su varianza.
En primer lugar

$$\begin{aligned}
\operatorname{var}(\mathbf{X}\boldsymbol{\beta}^*) = \operatorname{var}(\mathbf{AY}) &= \sigma^2\mathbf{AVA}' \\
&= \sigma^2\mathbf{AV} \\
&= \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'
\end{aligned}$$

de forma que si $\mathbf{a}'\boldsymbol{\beta}$ es estimable

$$\operatorname{var}(\mathbf{a}'\boldsymbol{\beta}^*) = \sigma^2\mathbf{a}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{a}$$

También es necesario obtener un estimador para σ^2.
A partir del modelo 4.11

$$\begin{aligned}
\text{SCR} &= (\mathbf{K}^{-1}\mathbf{Y})'[\mathbf{I} - \mathbf{K}^{-1}\mathbf{X}((\mathbf{K}^{-1}\mathbf{X})'(\mathbf{K}^{-1}\mathbf{X}))^-]\mathbf{K}^{-1}\mathbf{Y} \\
&= \mathbf{Y}'(\mathbf{I} - \mathbf{A})'\mathbf{V}^{-1}(\mathbf{I} - \mathbf{A})\mathbf{Y}
\end{aligned}$$

y como $\mathrm{rg}(\mathbf{K}^{-1}\mathbf{X}) = \mathrm{rg}(\mathbf{X})$, tenemos

$$\widehat{\sigma^2} = \mathbf{Y}'(\mathbf{I}-\mathbf{A})'\mathbf{V}^{-1}(\mathbf{I}-\mathbf{A})\mathbf{Y}/(n-r)$$

Además, cuando asumimos la hipótesis de normalidad $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{V})$ se verifican otras propiedades también heredadas del caso ordinario. En especial, cualquier estimador MCG de $\boldsymbol{\beta}$ es de máxima verosimilitud. También, para cualquier función estimable $\mathbf{a}'\boldsymbol{\beta}$ el estimador MCG es insesgado de varianza mínima.

En cuanto a las distribuciones asociadas, si $\boldsymbol{\epsilon}$ tiene distribución normal, la SCR es independiente de $\mathbf{K}^{-1}\mathbf{X}\boldsymbol{\beta}^*$, ya que $\mathrm{cov}(\mathbf{B}\boldsymbol{\beta}^*, \mathbf{Z}-\mathbf{B}\boldsymbol{\beta}^*) = \mathbf{0}$, y en consecuencia independiente de $\mathbf{X}\boldsymbol{\beta}^*$.

Es evidente que $\mathbf{X}\boldsymbol{\beta}^*$ se distribuye normalmente y se demuestra que $\mathrm{SCR}/\sigma^2 \sim \chi^2$.

Así pues, para una función paramétrica estimable $\mathbf{a}'\boldsymbol{\beta}$

$$\frac{\mathbf{a}'\boldsymbol{\beta}^* - \mathbf{a}'\boldsymbol{\beta}}{[\widehat{\sigma^2}\ \mathbf{a}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{a}]^{1/2}} \sim t_{n-r}$$

lo que se puede utilizar para el cálculo de intervalos de confianza de $\mathbf{a}'\boldsymbol{\beta}$ o en contrastes de hipótesis.

Por último nos podemos preguntar si la estimación generalizada $\boldsymbol{\beta}^*$ puede coincidir con la ordinaria $\widehat{\boldsymbol{\beta}}$ y en qué circunstancias. La respuesta es que ambas estimaciones coinciden si y sólo si $\langle \mathbf{V}^{-1}\mathbf{X}\rangle = \langle \mathbf{X}\rangle$ que es equivalente a $\langle \mathbf{V}\mathbf{X}\rangle = \langle \mathbf{X}\rangle$. La demostración puede verse en [65, pág. 63].

4.3. Otros métodos de estimación

4.3.1. Estimación sesgada

Dado el modelo lineal ordinario $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, donde $E(\boldsymbol{\epsilon}) = \mathbf{0}$ y $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, sabemos que el estimador MC $\mathbf{a}'\widehat{\boldsymbol{\beta}}$ es el estimador insesgado de varianza mínima para una f.p.e. $\mathbf{a}'\boldsymbol{\beta}$ cuando $\boldsymbol{\epsilon}$ tiene distribución normal, y el estimador *lineal* insesgado de varianza mínima sin la hipótesis de normalidad. Pero el hecho de ser un estimador de varianza mínima no garantiza que ésta sea realmente pequeña. Ya hemos visto en el apartado 3.3 cómo se calcula dicha varianza en función de los valores propios de la matriz $\mathbf{X}'\mathbf{X}$ y una posible solución propuesta por Silvey. Veamos ahora otra propuesta cuando en un modelo de rango máximo, $\mathbf{X}'\mathbf{X}$ está cerca de la singularidad, es decir, cuando uno o más de sus valores propios son casi cero.

Consideremos la llamada *varianza total* de los estimadores de los parámetros en un modelo

$$\sum_{i=1}^{m} \mathrm{var}(\widehat{\beta}_i) = \sigma^2\,\mathrm{tr}[(\mathbf{X}'\mathbf{X})^{-1}] = \sigma^2 \sum_{i=1}^{m} \lambda_i^{-1} > \sigma^2\lambda_m^{-1}$$

donde $\lambda_m > 0$ es el más pequeño de los valores propios de $\mathbf{X}'\mathbf{X}$. En la práctica, aunque la matriz $\mathbf{X}$ sea de rango máximo, puede ocurrir que λ_m sea muy pequeño y en consecuencia provocar que la varianza total sea muy grande.

Para solucionar este problema Hoerl y Kennard (1970) introducen los *ridge estimators*

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_{(k)} &= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= (\mathbf{I} + k(\mathbf{X}'\mathbf{X})^{-1})^{-1}\widehat{\boldsymbol{\beta}} \\ &= \mathbf{K}\widehat{\boldsymbol{\beta}}\end{aligned}$$

donde $k \geq 0$ es un escalar a elegir de forma que, si no es cero, $\widetilde{\boldsymbol{\beta}}_{(k)}$ es un estimador sesgado de $\boldsymbol{\beta}$.
Las principales razones para la utilización de estos estimadores son:

- Los gráficos de los componentes de $\widetilde{\boldsymbol{\beta}}_{(k)}$ y de sus correspondientes SCR al variar k permiten estudiar la enfermedad de $\mathbf{X}$.
- Es posible elegir un valor de k tal que los coeficientes de regresión tengan valores razonables y la SCR no sea muy grande.
- Se ha demostrado que es posible hallar un k que, por un pequeño incremento del sesgo, reduce la varianza total y, en consecuencia, el error cuadrático medio total.

El estudio de generalizaciones de estos estimadores y sus propiedades ha tenido bastante éxito.

4.3.2. Estimación robusta

En el capítulo anterior se ha demostrado que, mientras se verifique la hipótesis de normalidad para las observaciones, los estimadores obtenidos por el método de los mínimos cuadrados gozan de muy buenas propiedades. Sin embargo, también se han estudiado los resultados cuando las observaciones siguen distribuciones distintas de la normal y se ha constatado que el método de los mínimos cuadrados falla en muchos aspectos. En especial, cuando la distribución de los errores tiene una alta curtosis los estimadores mínimo-cuadráticos son muy poco eficientes, comparados con estimadores robustos de localización (ver Andrews et al.[4, cap. 7]). Puede probarse (ver Peña [54, pág. 465]) que en estas situaciones la estimación de máxima verosimilitud es equivalente a minimizar una función ponderada de los errores, que proporcione menos pesos a los residuos más grandes. Se trata de calcular estimadores que minimicen

$$\sum \omega_i(\epsilon_i)\epsilon_i^2$$

donde $\omega_i(\epsilon_i)$ es una función para reducir el efecto de los datos con un residuo muy alto. Los métodos de estimación robusta que utilicen esta idea requieren la definición de la función de ponderación ω y un procedimiento iterativo para acercarnos a los valores $\omega_i(\epsilon_i)$, ya que los errores ϵ_i son, en principio, desconocidos. Entre las propuestas más interesantes destaca la función de ponderación de Huber (1981)

$$\omega_i = \begin{cases} \dfrac{1}{2} & \text{si } |r_i| < c \\ \left|\dfrac{c}{r_i}\right| - \dfrac{1}{2}\left|\dfrac{c}{r_i}\right|^2 & \text{si } |r_i| \geq c \end{cases}$$

donde los r_i son los residuos estudentizados y c una constante entre 1,5 y 2 que establece el grado de "protección". Para calcular la estimación de los parámetros se toma inicialmente la mínimo cuadrática ordinaria, se calculan los residuos y con ellos las ponderaciones para la siguiente estimación, y así sucesivamente.

Otra alternativa es minimizar $\sum_i |\epsilon_i|$ con respecto a $\boldsymbol{\beta}$. Este es un problema de minimización de una norma L1 que se puede reducir a un problema de programación lineal y a un procedimiento similar al método del simplex, aunque la solución no siempre es única y algunos de los algoritmos proporcionan estimadores sesgados. Otros procedimientos iterativos propuestos no tienen resuelta la cuestión de la convergencia y el sesgo (ver Seber [65, pág. 91]).

4.3.3. Más posibilidades

También se ha estudiado el problema de la estimación mínimo cuadrática sujeta a las restricciones $\beta_i \geq 0$, $i = 1, \ldots, m$.

Por otra parte, en algunos problemas de regresión, los datos de la variable respuesta pueden ser censurados, es decir, los valores de algunas observaciones sólo se conocen si son superiores (o inferiores) a algún valor dado. Esto se suele producir en problemas donde la variable observada es el tiempo de vida. En estos casos el método clásico de los mínimos cuadrados no sirve y se han estudiado otros procedimientos (ver Seber [65, pág. 90]).

4.4. Ejercicios

Ejercicio 4.1

Sea el modelo lineal

$$\begin{aligned} y_1 &= \theta_1 + \theta_2 + \epsilon_1 \\ y_2 &= \theta_1 - 2\theta_2 + \epsilon_2 \\ y_3 &= 2\theta_1 - \theta_2 + \epsilon_3 \end{aligned}$$

Hallar las estimaciones MC de θ_1 y θ_2. Utilizando el método mínimo-cuadrático en dos pasos, hallar la estimación MC de θ_3, cuando el modelo se amplía en la forma

$$\begin{aligned} y_1 &= \theta_1 + \theta_2 + \theta_3 + \epsilon_1 \\ y_2 &= \theta_1 - 2\theta_2 + \theta_3 + \epsilon_2 \\ y_3 &= 2\theta_1 - \theta_2 + \theta_3 + \epsilon_3 \end{aligned}$$

Ejercicio 4.2

Un experimentador desea estimar la densidad d de un líquido mediante el pesado de algunos volúmenes del líquido. Sean y_i los pesos para los volúmenes x_i, $i = 1, \ldots, n$ y sean $E(y_i) = dx_i$ y $\text{var}(y_i) = \sigma^2 f(x_i)$. Hallar el estimador MC de d en los siguientes casos:

$$\text{(a) } f(x_i) \equiv 1 \qquad \text{(b) } f(x_i) = x_i \qquad \text{(c) } f(x_i) = x_i^2$$

Capítulo 5

Contraste de hipótesis lineales

5.1. Hipótesis lineales contrastables

Consideremos el modelo lineal $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, donde $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ y $\text{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$.

Una hipótesis lineal consiste en una o varias restricciones lineales planteadas sobre los parámetros $\boldsymbol{\beta}$. En un diseño de rango máximo $\text{rg}\,\mathbf{X} = m$ vamos a ver que cualquier hipótesis lineal es contrastable (testable o *demostrable*), es decir, es posible encontrar un estadístico (el test F del teorema 5.3.1) mediante el cual podemos decidir si se rechaza o acepta la hipótesis. Si $\text{rg}\,\mathbf{X} = r < m$, entonces pueden existir hipótesis estadísticamente no contrastables.

Definición 5.1.1

Una hipótesis lineal de rango q sobre los parámetros $\boldsymbol{\beta}$ es un conjunto de restricciones lineales

$$a_{i1}\beta_1 + \cdots + a_{im}\beta_m = 0 \qquad i = 1, \ldots, q$$

Si escribimos la matriz de la hipótesis como

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{q1} & \cdots & a_{qm} \end{pmatrix} \qquad rg\ \mathbf{A} = q$$

entonces las restricciones se resumen en

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{0}$$

Una hipótesis se dice que es contrastable o demostrable *si el conjunto $\mathbf{A}\boldsymbol{\beta}$ es un sistema de funciones paramétricas estimables. Entonces, las filas de $\mathbf{A}$ son combinación lineal de las filas de la matriz de diseño $\mathbf{X}$, es decir, que existe una matriz $\mathbf{B}$ de tamaño $q \times n$ tal que*

$$\mathbf{A} = \mathbf{B}\mathbf{X}$$

También $\mathbf{B}$ puede ser $q \times k$ si consideramos la matriz de diseño reducida $\mathbf{X}_R$ $k \times m$.

Cuando $\mathbf{X}$ no es de rango máximo, un conjunto de restricciones $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$ donde las filas de $\mathbf{A}$ son linealmente independientes de las filas de $\mathbf{X}$ no forman una alternativa al modelo general, en el sentido de un modelo más sencillo. En realidad son restricciones que permiten identificar mejor las estimaciones indeterminadas que resultan de las ecuaciones

normales. Por ello exigimos que las filas de $\mathbf{A}$ sean linealmente dependientes de las filas de $\mathbf{X}$ y que el rango de la matriz $\mathbf{A}$ $q \times m$ sea q. De hecho, cualquier ecuación $\mathbf{a}_i'\boldsymbol{\beta} = 0$ para la que $\mathbf{a}_i'$ sea linealmente independiente de las filas de $\mathbf{X}$ puede ignorarse y la hipótesis contrastable estará formada por el resto de las ecuaciones.

Una caracterización para saber si una hipótesis lineal es contrastable es

$$\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{A}$$

Este resultado es una generalización del que se ha demostrado en la página 42 para una función paramétrica estimable (ver ejercicio 5.3).

5.2. El modelo lineal de la hipótesis

El modelo lineal inicial $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, que se supone válido, constituye la hipótesis alternativa

$$H_1 : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \text{rg } \mathbf{X} = r$$

Por otra parte, el modelo lineal junto con la restricción lineal contrastable forman la hipótesis nula

$$H_0 : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \mathbf{A}\boldsymbol{\beta} = \mathbf{0} \qquad \text{rg } \mathbf{A} = q$$

Pero esta restricción lineal transforma los parámetros $\boldsymbol{\beta}$ y la matriz de diseño $\mathbf{X}$ en un nuevo modelo llamado el modelo lineal de la hipótesis

$$H_0 : \mathbf{Y} = \widetilde{\mathbf{X}}\boldsymbol{\theta} + \boldsymbol{\epsilon} \qquad \text{rg } \widetilde{\mathbf{X}} = r - q > 0$$

que es otra forma de plantear la hipótesis nula.

Existen varios procedimientos para estimar $\boldsymbol{\beta}$ o $\boldsymbol{\theta}$ bajo la hipótesis nula y calcular la suma de cuadrados residual.

Método 1

Si la hipótesis es contrastable, las filas de $\mathbf{A}$ son combinación lineal de las filas de $\mathbf{X}$. El subespacio $\langle \mathbf{A}' \rangle$ generado por las filas de $\mathbf{A}$ está incluido en el subespacio $\langle \mathbf{X}' \rangle$ generado por las filas de $\mathbf{X}$. Existe entonces una base ortogonal

$$\mathbf{v}_1, \ldots, \mathbf{v}_q, \mathbf{v}_{q+1}, \ldots, \mathbf{v}_r, \mathbf{v}_{r+1} \ldots, \mathbf{v}_m$$

tal que

$$\langle \mathbf{A}' \rangle = \langle \mathbf{v}_1, \ldots, \mathbf{v}_q \rangle \subset \langle \mathbf{v}_1, \ldots, \mathbf{v}_q, \mathbf{v}_{q+1}, \ldots, \mathbf{v}_r \rangle = \langle \mathbf{X}' \rangle \subset \mathbb{R}^m$$

Sea entonces $\mathbf{C}$ una matriz $m \times r'$, con $r' = r - q$, construida tomando los vectores columna $\mathbf{v}_{q+1}, \ldots, \mathbf{v}_r$

$$\mathbf{C} = (\mathbf{v}_{q+1}, \ldots, \mathbf{v}_r)$$

y definamos el vector paramétrico $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{r'})'$ tal que

$$\boldsymbol{\beta} = \mathbf{C}\boldsymbol{\theta}$$

Los parámetros $\boldsymbol{\theta}$ constituyen la reparametrización inducida por la hipótesis H_0, pues

$$\mathbf{A}\boldsymbol{\beta} = \mathbf{A}\mathbf{C}\boldsymbol{\theta} = \mathbf{0}\boldsymbol{\theta} = \mathbf{0}$$

El modelo $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ bajo la restricción $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$, se convierte en

$$E(\mathbf{Y}) = \mathbf{X}\mathbf{C}\boldsymbol{\theta} = \widetilde{\mathbf{X}}\boldsymbol{\theta}$$

y la matriz de diseño se transforma en

$$\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{C}$$

relación también válida para la matriz de diseño reducida

$$\widetilde{\mathbf{X}}_R = \mathbf{X}_R\mathbf{C}$$

La estimación MC de los parámetros $\boldsymbol{\theta}$ es

$$\widehat{\boldsymbol{\theta}} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\mathbf{Y}$$

La suma de cuadrados residual bajo la restricción $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$ es

$$\begin{aligned}\mathrm{SCR}_H &= \min_{\mathbf{A}\boldsymbol{\beta}=\mathbf{0}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{Y} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}})'(\mathbf{Y} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}) \\ &= \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\theta}}'\widetilde{\mathbf{X}}'\mathbf{Y}\end{aligned}$$

Método 2

Introduzcamos q multiplicadores de Lagrange

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)'$$

uno para cada restricción lineal. El mínimo restringido de $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ se halla igualando a cero las derivadas respecto a cada β_i de

$$\sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - \cdots - x_{im}\beta_m)^2 + \sum_{i=1}^{q}\lambda_i(a_{i1}\beta_1 + \cdots + a_{im}\beta_m)$$

En notación matricial, donde ahora $\mathbf{X}$ es la matriz ampliada, escribiremos

$$\begin{aligned}f(\boldsymbol{\beta}, \boldsymbol{\lambda}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}'\mathbf{A}')\boldsymbol{\lambda} \\ \partial f/\partial\boldsymbol{\beta} &= -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'\boldsymbol{\lambda} = \mathbf{0} \\ \mathbf{X}'\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}'\mathbf{Y} - \frac{1}{2}\mathbf{A}'\boldsymbol{\lambda} \qquad (5.1)\end{aligned}$$

La solución es

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_H &= (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} - \frac{1}{2}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}'\widehat{\boldsymbol{\lambda}}_H \\ &= \widehat{\boldsymbol{\beta}} - \frac{1}{2}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}'\widehat{\boldsymbol{\lambda}}_H\end{aligned}$$

y como $\mathbf{A}\widehat{\boldsymbol{\beta}}_H = \mathbf{0}$, resulta

$$\mathbf{0} = \mathbf{A}\widehat{\boldsymbol{\beta}} - \frac{1}{2}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}'\widehat{\boldsymbol{\lambda}}_H$$

La matriz $\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{A}'$ posee inversa, puesto que es de rango q, así

$$\frac{1}{2}\widehat{\boldsymbol{\lambda}}_H = (\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}})$$

y finalmente tenemos que la estimación MC restringida es

$$\widehat{\boldsymbol{\beta}}_H = \widehat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^-\mathbf{A}'(\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{A}')^{-1}\mathbf{A}\widehat{\boldsymbol{\beta}} \tag{5.2}$$

La suma de cuadrados residual es

$$\mathrm{SCR}_H = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_H)'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_H)$$

Hemos visto (teorema 2.5.1) que la forma canónica de la suma de cuadrados residual bajo el modelo sin restricciones es

$$\mathrm{SCR} = z_{r+1}^2 + \cdots + z_n^2$$

La hipótesis $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{0}$, que implica $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{C}$, significa que las columnas de $\widetilde{\mathbf{X}}$ son combinación lineal de las de $\mathbf{X}$. Luego los subespacios generados por dichas columnas verifican

$$\langle\widetilde{\mathbf{X}}\rangle \subset \langle\mathbf{X}\rangle \subset \mathbb{R}^n \tag{5.3}$$

Podemos entonces construir una base ortogonal

$$\mathbf{u}_1, \ldots, \mathbf{u}_{r'}, \mathbf{u}_{r'+1}, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}, \ldots, \mathbf{u}_n$$

tal que

$$\langle\widetilde{\mathbf{X}}\rangle = \langle\mathbf{u}_1, \ldots, \mathbf{u}_{r'}\rangle \subset \langle\mathbf{X}\rangle = \langle\mathbf{u}_1, \ldots, \mathbf{u}_r\rangle$$

Entonces, si se cumple la hipótesis, por idéntico razonamiento al seguido en el teorema 2.5.1 tendremos que la forma canónica de la suma de cuadrados residual bajo el modelo H_0 es

$$\mathrm{SCR}_H = z_{r'+1}^2 + \cdots + z_n^2$$

Además, siempre se verificará que $\mathrm{SCR}_H > \mathrm{SCR}$ pues

$$\mathrm{SCR}_H - \mathrm{SCR} = \sum_{r'+1}^{r} z_i^2$$

Ejemplo 5.2.1

Consideremos el siguiente modelo lineal normal

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 + \epsilon_1 \\ y_2 &= 2\beta_2 + \epsilon_2 \\ y_3 &= -\beta_1 + \beta_2 + \epsilon_3 \end{aligned}$$

y la hipótesis lineal

$$H_0 : \beta_1 = 2\beta_2$$

Las matrices de diseño y de la hipótesis son

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \\ -1 & 1 \end{pmatrix} \qquad \mathbf{A} = (1 \quad -2) \qquad \mathrm{rg}\ \mathbf{X} = 2 \qquad \mathrm{rg}\ \mathbf{A} = 1$$

Como $\mathbf{A}$ *es combinación lineal de las filas de* $\mathbf{X}$*,* H_0 *es una hipótesis contrastable. Además, en este caso particular el rango de la matriz de diseño es máximo, de modo que toda hipótesis lineal es contrastable.*

Con unos sencillos cálculos, tenemos:

Ecuaciones normales

$$2\beta_1 + 0\beta_2 = y_1 - y_3 \qquad 0\beta_1 + 6\beta_2 = y_1 + 2y_2 + y_3$$

Estimaciones MC

$$\widehat{\beta}_1 = (y_1 - y_3)/2 \qquad \widehat{\beta}_2 = (y_1 + 2y_2 + y_3)/6$$

Suma de cuadrados residual

$$\mathrm{SCR} = y_1^2 + y_2^2 + y_3^2 - 2\widehat{\beta}_1^2 - 6\widehat{\beta}_2^2$$

Si consideramos los vectores columna

$$\mathbf{v}_1 = (1, -2)' \qquad \mathbf{v}_2 = (2, 1)'$$

que constituyen una base ortogonal de $\mathbb{R}^2$*, se verifica*

$$\langle \mathbf{A}' \rangle = \langle \mathbf{v}_1 \rangle \subset \langle \mathbf{X}' \rangle = \langle \mathbf{v}_1, \mathbf{v}_2 \rangle$$

Podemos entonces tomar la matriz

$$\mathbf{C} = (2, 1)'$$

que verifica $\mathbf{AC} = \mathbf{0}$*. La reparametrización* $\boldsymbol{\beta} = \mathbf{C}\boldsymbol{\theta}$ *es*

$$\beta_1 = 2\theta \qquad \beta_2 = \theta$$

El modelo bajo la hipótesis es ahora

$$\begin{aligned} y_1 &= 3\theta + \epsilon_1 \\ y_2 &= 2\theta + \epsilon_2 \\ y_3 &= -\theta + \epsilon_3 \end{aligned}$$

Finalmente

$$\widehat{\theta} = (3y_1 + 2y_2 - y_3)/14$$
$$\mathrm{SCR}_H = y_1^2 + y_2^2 + y_3^2 - 14\,\widehat{\theta}^2$$

5.3. Teorema fundamental del Análisis de la Varianza

En esta sección vamos a deducir el test F que nos permite decidir sobre la aceptación de una hipótesis lineal contrastable.

Teorema 5.3.1

Sea $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$ un modelo lineal normal, de manera que $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Consideremos una hipótesis lineal contrastable

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{0} \qquad \text{rango } \mathbf{A} = q$$

entonces, los estadísticos

$$\begin{aligned} \text{SCR} &= (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ \text{SCR}_H &= (\mathbf{Y} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}})'(\mathbf{Y} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}) \end{aligned}$$

verifican:

(i) $\text{SCR}/\sigma^2 \sim \chi^2_{n-r}$

(ii) Si H_0 es cierta

$$\begin{aligned} \text{SCR}_H/\sigma^2 &\sim \chi^2_{n-r'} \qquad (r' = r - q) \\ (\text{SCR}_H - \text{SCR})/\sigma^2 &\sim \chi^2_q \end{aligned}$$

(iii) Si H_0 es cierta, los estadísticos $\text{SCR}_H - \text{SCR}$ y SCR son estocásticamente independientes.

(iv) Si H_0 es cierta, el estadístico

$$F = \frac{(\text{SCR}_H - \text{SCR})/q}{\text{SCR}/(n-r)} \tag{5.4}$$

sigue la distribución F de Fisher-Snedecor con q y $n - r$ grados de libertad.

Demostración:

(i) Aunque este resultado ya se ha establecido en el teorema 3.4.2, nos interesa ahora su demostración explícita. En el teorema 2.5.1 se ha visto que

$$\text{SCR} = z^2_{r+1} + \cdots + z^2_n$$

donde las z_i son normales, independientes y además $E(z_i) = 0$, $\text{var}(z_i) = \sigma^2$. Luego SCR/σ^2 es suma de los cuadrados de $n - r$ variables $N(0, 1)$ independientes.

(ii) La forma canónica de la suma de cuadrados residual bajo la restricción $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$ es

$$\text{SCR}_H = z^2_{r'+1} + \cdots + z^2_n$$

luego análogamente tenemos que $\text{SCR}_H/\sigma^2 \sim \chi^2_{n-r'}$, donde $r' = r - q$. Además

$$\text{SCR}_H - \text{SCR} = z^2_{r'+1} + \cdots + z^2_r$$

es también una suma de cuadrados en las mismas condiciones.

(iii) Las variables $z_{r'+1}, \ldots, z_n$ son normales e independientes. $\text{SCR}_H - \text{SCR}$ depende de las q primeras, mientras que SCR depende de las $n - r$ últimas y no hay términos comunes. Luego son estocásticamente independientes.

(iv) Es una consecuencia evidente de los apartados anteriores de este teorema. Si H_0 es cierta, el estadístico

$$F = \frac{[(\mathrm{SCR}_H - \mathrm{SCR})/\sigma^2]/q}{(\mathrm{SCR}/\sigma^2)/(n-r)} = \frac{(\mathrm{SCR}_H - \mathrm{SCR})/q}{\mathrm{SCR}/(n-r)}$$

sigue la distribución F de Fisher-Snedecor con q y $n-r$ grados de libertad.

■

Obsérvese que F no depende del parámetro desconocido σ^2 y se puede calcular exclusivamente en función de las observaciones $\mathbf{Y}$.
La expresión de SCR es

$$\mathrm{SCR} = \mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$$

Veamos que, del mismo modo, la expresión de SCR_H es

$$\mathrm{SCR}_H = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}_H'\mathbf{X}'\mathbf{Y}$$

donde $\widehat{\boldsymbol{\beta}}_H$ es la estimación MC de $\boldsymbol{\beta}$ restringida a $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$.
En efecto,

$$\mathrm{SCR}_H = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_H)'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_H) = \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\widehat{\boldsymbol{\beta}}_H + \widehat{\boldsymbol{\beta}}_H'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}_H$$

Además (ver página 69), se verifica

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}_H = \mathbf{X}'\mathbf{Y} - \frac{1}{2}\mathbf{A}'\widehat{\boldsymbol{\lambda}}_H$$

luego

$$\begin{aligned}\mathrm{SCR}_H &= \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\widehat{\boldsymbol{\beta}}_H + \widehat{\boldsymbol{\beta}}_H'(\mathbf{X}'\mathbf{Y} - \frac{1}{2}\mathbf{A}'\widehat{\boldsymbol{\lambda}}_H)\\ &= \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\widehat{\boldsymbol{\beta}}_H + \mathbf{Y}'\mathbf{X}\widehat{\boldsymbol{\beta}}_H - \frac{1}{2}\widehat{\boldsymbol{\beta}}_H'\mathbf{A}'\widehat{\boldsymbol{\lambda}}_H\end{aligned}$$

Pero como $\mathbf{A}\widehat{\boldsymbol{\beta}}_H = \mathbf{0}$, nos queda

$$\mathrm{SCR}_H = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\widehat{\boldsymbol{\beta}}_H$$

Calculemos ahora $\mathrm{SCR}_H - \mathrm{SCR}$. Considerando 5.2 tenemos

$$\widehat{\boldsymbol{\beta}}' - \widehat{\boldsymbol{\beta}}_H' = (\mathbf{A}\widehat{\boldsymbol{\beta}})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{A}')^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^-$$

luego

$$\begin{aligned}\mathrm{SCR}_H - \mathrm{SCR} &= (\widehat{\boldsymbol{\beta}}' - \widehat{\boldsymbol{\beta}}_H')\mathbf{X}'\mathbf{Y}\\ &= (\mathbf{A}\widehat{\boldsymbol{\beta}})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{A}')^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y} \qquad (5.5)\\ &= (\mathbf{A}\widehat{\boldsymbol{\beta}})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}})\end{aligned}$$

El estadístico F puede escribirse entonces

$$F = \frac{(\mathbf{A}\widehat{\boldsymbol{\beta}})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}})}{q\,\widehat{\sigma}^2} \tag{5.6}$$

donde $\widehat{\sigma}^2 = \text{SCR}/(n-r)$.

Cuando $q > 2$ es mejor obtener SCR y SCR_H directamente por minimización de $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$ sin restricciones y con restricciones, respectivamente. Sin embargo, si $q \leq 2$ se puede utilizar la fórmula 5.6, ya que la matriz a invertir $\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}'$ es sólo de orden uno o dos.

Obsérvese que si $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$ es cierta, entonces $\mathbf{A}\widehat{\boldsymbol{\beta}} \approx \mathbf{0}$. Luego es probable que F no sea significativa.

Cuando sea posible, también se puede utilizar la matriz de diseño reducida $\mathbf{X}_R$, junto con las matrices $\mathbf{D}$ y $\bar{\mathbf{Y}}$. Las expresiones son entonces

$$\begin{aligned} \text{SCR} &= \mathbf{Y}'\mathbf{Y} - \bar{\mathbf{Y}}'\mathbf{D}\mathbf{X}_R(\mathbf{X}_R'\mathbf{D}\mathbf{X}_R)^{-}\mathbf{X}_R'\mathbf{D}\bar{\mathbf{Y}} \\ \text{SCR}_H - \text{SCR} &= (\mathbf{A}\widehat{\boldsymbol{\beta}})'(\mathbf{A}(\mathbf{X}_R'\mathbf{D}\mathbf{X}_R)^{-}\mathbf{A}')^{-}(\mathbf{A}\widehat{\boldsymbol{\beta}}) \end{aligned}$$

El cálculo de ambas cantidades se suele expresar en forma de tabla general del análisis de la varianza (ver tabla 5.1).

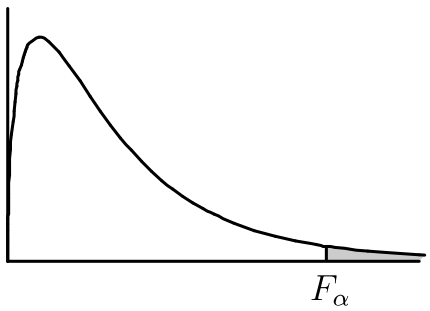
	grados de libertad	suma de cuadrados	cuadrados medios	cociente
Desviación hipótesis	q	$\text{SCR}_H - \text{SCR}$	$(\text{SCR}_H - \text{SCR})/q$	F
Residuo	$n-r$	SCR	$\text{SCR}/(n-r)$	

Cuadro 5.1: Tabla general del análisis de la varianza

Criterio de decisión

Si $F > F_\alpha$ se rechaza H_0; si $F \leq F_\alpha$ se acepta H_0.

Donde, para un nivel de significación α, F_α se elige de forma que $P(F_{q,n-r} > F_\alpha) = \alpha$.



Del teorema 5.3.1 deducimos que, si H_0 es cierta, entonces

$$E[(\text{SCR}_H - \text{SCR})/q] = \sigma^2$$

Luego $(\text{SCR}_H - \text{SCR})/q$ y $\text{SCR}/(n-r)$ son dos estimaciones independientes de la varianza σ^2. El test F nos indica hasta que punto coinciden. Un valor grande de F indica que la primera estimación difiere demasiado de la varianza σ^2 y entonces H_0 debe ser rechazada. Se puede demostrar además (ver ejercicio 5.7) que en general

$$E(\text{SCR}_H - \text{SCR}) = q\sigma^2 + (\mathbf{A}\boldsymbol{\beta})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}')^{-}(\mathbf{A}\boldsymbol{\beta}) \tag{5.7}$$

Ejemplo 5.3.1

Para decidir sobre la hipótesis $H_0 : \beta_1 = 2\beta_2$ en el ejemplo 5.2.1 calcularemos

$$F = \frac{(\mathrm{SCR}_H - \mathrm{SCR})/1}{\mathrm{SCR}/(3-2)} = \frac{-14\widehat{\theta}^2 + 2\widehat{\beta}_1^2 + 6\widehat{\beta}_2^2}{y_1^2 + y_2^2 + y_3^2 - 2\widehat{\beta}_1^2 - 6\widehat{\beta}_2^2}$$

Si utilizamos 5.6, se obtiene una expresión más sencilla

$$F = \frac{(\widehat{\beta}_1 - 2\widehat{\beta}_2)^2}{(\mathrm{SCR}/1)(7/6)}$$

En cualquier caso, se decide por la significación en una distribución $F_{1,1}$ con 1 y 1 grados de libertad.

Ejemplo 5.3.2 ***Diseño "cross-over" simplificado***

Supongamos una experiencia clínica en la que se desean comparar dos fármacos **a** *y* **b**, *para combatir una determinada enfermedad. El estado de los pacientes se valora mediante una cierta variable cuantitativa Y.*

En el diseño "cross-over" la experiencia se organiza asignando a $N_{\mathbf{a}}$ pacientes el tratamiento **a** *y a $N_{\mathbf{b}}$ pacientes el tratamiento* **b**, *en un primer periodo. En un segundo periodo, los que tomaban* **a** *pasan a tomar* **b** *y recíprocamente. En este diseño los datos son de la forma:*

Grupo 1					***media***	***varianza***
a *(primera vez)*	y_{11}	y_{12}	$\ldots$	$y_{1N_{\mathbf{a}}}$	$\bar{y}_{1\cdot}$	$s_1^2 = \frac{1}{N_{\mathbf{a}}}\sum_{i=1}^{N_{\mathbf{a}}}(y_{1i} - \bar{y}_{1\cdot})^2$
b *(después de* **a**$)$	y_{21}	y_{22}	$\ldots$	$y_{2N_{\mathbf{a}}}$	$\bar{y}_{2\cdot}$	$s_2^2 = \frac{1}{N_{\mathbf{a}}}\sum_{i=1}^{N_{\mathbf{a}}}(y_{2i} - \bar{y}_{2\cdot})^2$
Grupo 2						
b *(primera vez)*	y_{31}	y_{32}	$\ldots$	$y_{3N_{\mathbf{b}}}$	$\bar{y}_{3\cdot}$	$s_3^2 = \frac{1}{N_{\mathbf{b}}}\sum_{i=1}^{N_{\mathbf{b}}}(y_{3i} - \bar{y}_{3\cdot})^2$
a *(después de* **b**$)$	y_{41}	y_{42}	$\ldots$	$y_{4N_{\mathbf{b}}}$	$\bar{y}_{4\cdot}$	$s_4^2 = \frac{1}{N_{\mathbf{b}}}\sum_{i=1}^{N_{\mathbf{b}}}(y_{4i} - \bar{y}_{4\cdot})^2$

Indicando

$$\begin{aligned}
\mu &= \textit{media general} \\
\alpha &= \textit{efecto fármaco } \mathbf{a} \\
\beta &= \textit{efecto fármaco } \mathbf{b} \\
\gamma &= \textit{efecto recíproco entre } \mathbf{a} \textit{ y } \mathbf{b}
\end{aligned}$$

se propone el siguiente modelo:

$$\begin{array}{lll}
\mathbf{a}\ \textit{(primera vez)} & y_{1i} = \mu + \alpha + \epsilon_{1i} & i = 1, \ldots, N_{\mathbf{a}} \\
\mathbf{b}\ \textit{(después de } \mathbf{a}\textit{)} & y_{2i} = \mu + \beta + \gamma + \epsilon_{2i} & i = 1, \ldots, N_{\mathbf{a}} \\
\mathbf{b}\ \textit{(primera vez)} & y_{3i} = \mu + \beta + \epsilon_{3i} & i = 1, \ldots, N_{\mathbf{b}} \\
\mathbf{a}\ \textit{(después de } \mathbf{b}\textit{)} & y_{4i} = \mu + \alpha + \gamma + \epsilon_{4i} & i = 1, \ldots, N_{\mathbf{b}}
\end{array}$$

Es decir, cuando sólo se ha tomado un fármaco actúa un solo efecto, pero cuando se ha tomado uno después del otro actúa entonces un efecto aditivo γ que recoge la mejoría del enfermo que ya ha tomado el primer medicamento.

Tenemos $k = 4$ condiciones experimentales, que en el "cross-over" simplificado se consideran independientes, y $N_1 = N_2 = N_{\mathbf{a}}$, $N_3 = N_4 = N_{\mathbf{b}}$. El vector de observaciones $\mathbf{Y}$ y la matriz de diseño reducida $\mathbf{X}_R$ son

$$\mathbf{Y} = (y_{11}, \ldots, y_{1N_{\mathbf{a}}}, y_{21}, \ldots, y_{2N_{\mathbf{a}}}, y_{31}, \ldots, y_{3N_{\mathbf{b}}}, y_{41}, \ldots, y_{4N_{\mathbf{b}}})'$$

$$\mathbf{X}_R = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \qquad \text{rg } \mathbf{X}_R = 3$$

La hipótesis nula de mayor interés es

$$H_0 : \alpha = \beta \qquad \mathbf{a} \text{ y } \mathbf{b} \text{ tienen la misma efectividad}$$

que expresada en forma de hipótesis lineal es

$$H_0 : \begin{pmatrix} 0 & 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \\ \beta \\ \gamma \end{pmatrix} = 0$$

Como el vector $\begin{pmatrix} 0 & 1 & -1 & 0 \end{pmatrix}$ es combinación lineal de las filas de $\mathbf{X}_R$, se trata de una hipótesis contrastable. Para reparametrizar el diseño bajo H_0 tomaremos como matriz ortogonal a $\mathbf{A}$

$$\mathbf{C} = \begin{pmatrix} 2/3 & 0 \\ 1/3 & 0 \\ 1/3 & 0 \\ 0 & 1 \end{pmatrix}$$

Obsérvese que las columnas de $\mathbf{C}$ son también combinación lineal de las filas de $\mathbf{X}_R$.

Al establecer la relación $\boldsymbol{\beta} = \mathbf{C}\boldsymbol{\theta}$ tendremos

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

siendo $\theta_1 = \mu + \alpha = \mu + \beta$ y $\theta_2 = \gamma$.

Es decir, bajo H_0 el diseño reparametrizado depende de dos parámetros:

θ_1 : *efecto debido a la medicación (común a $\mathbf{a}$ y $\mathbf{b}$ bajo H_0)*

θ_2 : *efecto recíproco entre $\mathbf{a}$ y $\mathbf{b}$*

y la nueva matriz de diseño es

$$\widetilde{\mathbf{X}}_R = \mathbf{X}_R\mathbf{C} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$$

siendo rg $\widetilde{\mathbf{X}}_R = r - t = 3 - 1 = 2$.

Si el diseño es balanceado ($N_{\mathbf{a}} = N_{\mathbf{b}}$), entonces $N = 4N_{\mathbf{a}} = 4N_{\mathbf{b}}$ y se puede calcular que

$$\mathrm{SCR} = \frac{N_{\mathbf{a}}}{4}(y_{1\cdot} + y_{2\cdot} - y_{3\cdot} - y_{4\cdot})^2 + N_{\mathbf{a}}\left(\sum_{i=1}^{4} s_i^2\right)$$

con $N - 3$ grados de libertad

$$\mathrm{SCR}_H = \frac{N_{\mathbf{a}}}{4}[(y_{1\cdot} + y_{2\cdot} - y_{3\cdot} - y_{4\cdot})^2 + (y_{1\cdot} - y_{2\cdot} - y_{3\cdot} + y_{4\cdot})^2] + N_{\mathbf{a}}\left(\sum_{i=1}^{4} s_i^2\right)$$

con $N - 2$ grados de libertad.

Luego, si H_0 es cierta, bajo el modelo lineal normal, el estadístico

$$F = \frac{(y_{1\cdot} - y_{2\cdot} - y_{3\cdot} + y_{4\cdot})^2}{4\,\mathrm{SCR}} N_{\mathbf{a}}(4N_{\mathbf{a}} - 3)$$

sigue la distribución F con 1 y $N - 3$ g.l..

La tabla 5.2 contiene los datos de dos grupos de 10 y 10 enfermos reumáticos a los que se valoró la variación del dolor respecto del estado inicial, mediante una escala convencional, con el deseo de comparar dos fármacos antirreumáticos **a** *y* **b**, *administrados a lo largo de dos meses. Se incluye además la tabla del análisis de la varianza para contrastar H_0.*

Grupo 1		Grupo 2	
a (mes 1)	**b** (mes 2)	**b** (mes 1)	**a** (mes 2)
17	17	21	10
34	41	20	24
26	26	11	32
10	3	26	26
19	-6	42	52
17	-4	28	28
8	11	3	27
16	16	3	28
13	16	16	21
11	4	-10	42

Cuadro 5.2: Datos de los enfermos reumáticos

	g.l.	suma de cuadrados	cuadrados medios	F
Entre fármacos	1	783.2	783.2	4.71 ($p < 0{,}05$)
Residuo	37	6147.9	166.2	

Cuadro 5.3: Tabla del análisis de la varianza para $H_0 : \alpha = \beta$

Con estos datos se han detectado diferencias significativas entre los dos fármacos **a** *y* **b**. *Para estimar la eficacia de cada fármaco, pasaremos a considerar las funciones paramétricas*

$$\psi_{\mathbf{a}} = \mu + \alpha \qquad \psi_{\mathbf{b}} = \mu + \beta$$

que son ambas estimables.
Para estimar $\psi_{\mathbf{a}}, \psi_{\mathbf{b}}$ hallaremos primeramente "una" estimación MC de los parámetros:

$$\widehat{\mu} = 0 \qquad \widehat{\alpha} = 20{,}975 \qquad \widehat{\beta} = 12{,}125$$

Aplicando el teorema de Gauss-Markov, las estimaciones óptimas de $\psi_{\mathbf{a}}, \psi_{\mathbf{b}}$ se obtienen sustituyendo parámetros por estimaciones MC, es decir

$$\widehat{\psi_{\mathbf{a}}} = \widehat{\mu} + \widehat{\alpha} = 20{,}975 \qquad \widehat{\psi_{\mathbf{b}}} = \widehat{\mu} + \widehat{\beta} = 12{,}125$$

Por otra parte, las expresiones en función de las medias y las varianzas mínimas correspondientes son:

$$\begin{aligned} \widehat{\psi_{\mathbf{a}}} &= 3/4\bar{y}_1 - 1/4\bar{y}_2 + 1/4\bar{y}_3 + 1/4\bar{y}_4 \qquad \operatorname{var}(\widehat{\psi_{\mathbf{a}}}) = 0{,}075\sigma^2 \\ \widehat{\psi_{\mathbf{b}}} &= 1/4\bar{y}_1 + 1/4\bar{y}_2 + 3/4\bar{y}_3 - 1/4\bar{y}_4 \qquad \operatorname{var}(\widehat{\psi_{\mathbf{b}}}) = 0{,}075\sigma^2 \end{aligned}$$

5.3.1. Un contraste más general

Consideremos la hipótesis nula

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{c} \qquad \mathbf{A} \text{ es } q \times m, \text{ rg } \mathbf{A} = q$$

donde $\mathbf{c}$ es un vector columna que lógicamente debe ser combinación lineal de las columnas de $\mathbf{A}$. También suponemos que las filas de $\mathbf{A}$ son combinación lineal de las filas de $\mathbf{X}$, de manera que $\mathbf{A}\boldsymbol{\beta}$ es un conjunto de funciones paramétricas estimables.
Sea $\boldsymbol{\beta}_0$ tal que $\mathbf{A}\boldsymbol{\beta}_0 = \mathbf{c}$ y consideremos $\boldsymbol{\gamma} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$. Entonces, si en el modelo lineal

$$\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0 = \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \boldsymbol{\epsilon}$$

ponemos $\widetilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0$, obtenemos el modelo transformado

$$\widetilde{\mathbf{Y}} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{5.8}$$

y en este modelo la hipótesis planteada adopta la expresión

$$H_0 : \mathbf{A}\boldsymbol{\gamma} = \mathbf{0}$$

La estimación MC del conjunto de funciones paramétricas estimables $\mathbf{A}\boldsymbol{\gamma}$ en este modelo transformado es

$$\begin{aligned} \mathbf{A}\widehat{\boldsymbol{\gamma}} =& \mathbf{B}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\widetilde{\mathbf{Y}} \\ &= \mathbf{B}\mathbf{P}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0) = \mathbf{B}\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{B}\mathbf{X}\boldsymbol{\beta}_0 \\ &= \mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta}_0 = \mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{c} \end{aligned}$$

En consecuencia, de la ecuación 5.5 se deduce

$$\begin{aligned} \text{SCR}_H - \text{SCR} &= (\mathbf{A}\widehat{\boldsymbol{\gamma}})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\gamma}}) \\ &= (\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{c}) \end{aligned}$$

donde $\widehat{\boldsymbol{\beta}}$ es tal que $\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$. Se verifica también

$$E(\text{SCR}_H - \text{SCR}) = q\,\sigma^2 + (\mathbf{A}\boldsymbol{\beta} - \mathbf{c})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}')^{-1}(\mathbf{A}\boldsymbol{\beta} - \mathbf{c})$$

Finalmente, a partir de la fórmula 5.6 el test para contrastar la hipótesis es

$$F = \frac{(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{c})/q}{\mathrm{SCR}/(n-r)} \tag{5.9}$$

donde, si es cierta la hipótesis nula, el estadístico F sigue una distribución $F_{q,n-r}$.
En el caso particular $q = 1$, donde la hipótesis es $H_0 : \mathbf{a}'\boldsymbol{\beta} = c$, el test F se puede simplificar en un test t con

$$t = \frac{\mathbf{a}'\widehat{\boldsymbol{\beta}} - c}{(\widehat{\sigma}^2(\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{a}))^{1/2}} \tag{5.10}$$

que sigue una distribución t_{n-r}, si H_0 es cierta.

Ejemplo 5.3.3

Contraste de medias en poblaciones normales con igual varianza

Sean $u_1, u_2, \ldots, u_{n_1}$ y $v_1, v_2, \ldots, v_{n_2}$ dos muestras aleatorias simples de dos poblaciones normales $N(\mu_1, \sigma^2)$ y $N(\mu_2, \sigma^2)$, respectivamente.

Vamos a contrastar la hipótesis lineal $H_0 : \mu_1 - \mu_2 = d$ con la ayuda de la teoría de los modelos lineales.

Podemos pensar que las observaciones son de la forma

$$\begin{array}{ll} u_i = \mu_1 + \epsilon_i & i = 1, \ldots, n_1 \\ v_j = \mu_2 + \epsilon_{n_1+j} & j = 1, \ldots, n_2 \end{array}$$

o en notación matricial

$$\begin{pmatrix} u_1 \\ \vdots \\ u_{n_1} \\ v_1 \\ \vdots \\ v_{n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{n_1} \\ \epsilon_{n_1+1} \\ \vdots \\ \epsilon_n \end{pmatrix}$$

donde $n = n_1 + n_2$. Observemos que, gracias a la igualdad de varianzas en las dos poblaciones, se trata de un modelo lineal y se verifican las condiciones de Gauss-Markov.

En este modelo, la matriz de diseño reducida es 2×2 de rango máximo

$$\mathbf{X}_R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad y \qquad \mathbf{D} = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}$$

Así pues, la hipótesis nula es lineal y contrastable

$$H_0 : \mu_1 - \mu_2 = d \qquad \Leftrightarrow \qquad H_0 : \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = d \qquad q = 1$$

Con unos sencillos cálculos se obtiene

$$\widehat{\boldsymbol{\beta}} = (\hat{\mu}_1, \hat{\mu}_2)' = (\mathbf{X}_R'\mathbf{D}\mathbf{X}_R)^{-1}\mathbf{X}_R'\mathbf{D}\bar{\mathbf{Y}} = \bar{\mathbf{Y}} = (\bar{u}, \bar{v})'$$

$$\mathbf{A}\widehat{\boldsymbol{\beta}} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{u} - \bar{v}$$

$$\begin{aligned}\mathrm{SCR} &= \mathbf{Y}'\mathbf{Y} - \bar{\mathbf{Y}}'\mathbf{D}\mathbf{X}_R(\mathbf{X}_R'\mathbf{D}\mathbf{X}_R)^{-}\mathbf{X}_R'\mathbf{D}\bar{\mathbf{Y}} \\ &= \sum_i u_i^2 + \sum_j v_j^2 - n_1\bar{u}^2 - n_2\bar{v}^2 \\ &= \sum_i (u_i - \bar{u})^2 + \sum_j (v_j - \bar{v})^2\end{aligned}$$

$$\mathbf{A}(\mathbf{X}_R'\mathbf{D}\mathbf{X}_R)^{-1}\mathbf{A}' = \frac{1}{n_1} + \frac{1}{n_2}$$

de modo que

$$F = \frac{(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{A}(\mathbf{X}_R'\mathbf{D}\mathbf{X}_R)^{-1}\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{c})}{q\,\widehat{\sigma}^2} = \frac{(\bar{u} - \bar{v} - d)^2}{\widehat{\sigma}^2(1/n_1 + 1/n_2)}$$

donde $\widehat{\sigma}^2 = \mathrm{SCR}/(n_1 + n_2 - 2)$ *y cuya distribución, bajo* H_0*, es una* F_{1,n_1+n_2-2}.
Pero cuando $q = 1$*, tenemos que* $F_{1,n_1+n_2-2} \equiv t^2_{n_1+n_2-2}$ *y se deduce que el contraste es equivalente al test t usual, en especial el caso* $d = 0$.

5.3.2. Test de la razón de verosimilitud

Para simplificar, consideremos un modelo de rango máximo. Bajo la hipótesis de normalidad de las observaciones, ya sabemos (ver pág. 33) que las estimaciones de máxima verosimilitud de los parámetros son

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad\qquad \widehat{\sigma}^2_{MV} = \mathrm{SCR}/n$$

y el valor máximo de la función de verosimilitud es

$$L(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2_{MV}) = (2\pi\widehat{\sigma}^2_{MV})^{-n/2}e^{-n/2}$$

Del mismo modo, los estimadores de máxima verosimilitud de los parámetros con las restricciones $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ son

$$\widehat{\boldsymbol{\beta}}_H \qquad\qquad \widehat{\sigma}^2_H = \mathrm{SCR}_H/n$$

y el valor máximo de la función de verosimilitud, bajo la hipótesis nula, es

$$L(\widehat{\boldsymbol{\beta}}_H, \widehat{\sigma}^2_H) = (2\pi\widehat{\sigma}^2_H)^{-n/2}e^{-n/2}$$

De modo que el estadístico de la razón de verosimilitud es

$$\Lambda = \frac{L(\widehat{\boldsymbol{\beta}}_H, \widehat{\sigma}^2_H)}{L(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2_{MV})} = \left[\frac{\widehat{\sigma}^2_{MV}}{\widehat{\sigma}^2_H}\right]^{n/2}$$

Es fácil ver que

$$F = \frac{n - m}{q}(\Lambda^{-2/n} - 1)$$

luego son contrastes equivalentes.

5.4. Cuando el test es significativo

Si el estadístico F para $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ es significativo, podemos investigar la causa del rechazo de dicha hipótesis. Una posibilidad consiste en contrastar cada una de las restricciones $\mathbf{a}_i'\boldsymbol{\beta} = c_i$, $i = 1, \ldots, q$ por separado, utilizando un test t para ver cual es la responsable.

Hemos visto de varias formas que, bajo la hipótesis lineal $H_i : \mathbf{a}_i'\boldsymbol{\beta} = c_i$, el estadístico t_i verifica

$$t_i = \frac{\mathbf{a}_i'\widehat{\boldsymbol{\beta}} - c_i}{[\widehat{\sigma}^2\mathbf{a}_i'(\mathbf{X}'\mathbf{X})^-\mathbf{a}_i]^{1/2}} \sim t_{n-r}$$

de modo que podemos rechazar $H_i : \mathbf{a}_i'\boldsymbol{\beta} = c_i$ con un nivel de significación α si

$$|t_i| \geq t_{n-r}(\alpha)$$

donde $t_{n-r}(\alpha)$ es el valor de la tabla tal que $P(|t_{n-r}| \geq t_{n-r}(\alpha)) = \alpha$.

También podemos construir intervalos de confianza para cada $\mathbf{a}_i'\boldsymbol{\beta}$

$$\mathbf{a}_i'\widehat{\boldsymbol{\beta}} \pm t_{n-r}(\alpha) \cdot \widehat{\sigma}(\mathbf{a}_i'(\mathbf{X}'\mathbf{X})^-\mathbf{a}_i)^{1/2}$$

Este procedimiento en dos etapas para el contraste de $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$, es decir, un contraste global F seguido de una serie de test t cuando F es significativo, se conoce con el nombre de MDS[1] o *mínima diferencia significativa*. El valor significativo mínimo es $t_{n-r}(\alpha)$ y la palabra "diferencia" se refiere a que este método se utiliza con frecuencia para comparar parámetros tales como *medias* dos a dos.

Este método es simple y versátil, sin embargo tiene sus debilidades: es posible rechazar H_0 y no rechazar ninguna de las H_i. Este problema, otras dificultades y, en general, otros métodos de inferencia simultánea se estudian de forma más completa en lo que se llama *Métodos de comparación múltiple*.

5.5. Contraste de hipótesis sobre funciones paramétricas estimables

Sea $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_q)' = \mathbf{A}\boldsymbol{\beta}$ un sistema de funciones paramétricas estimables, de modo que las filas de la matriz $\mathbf{A}$ sean linealmente independientes. La distribución F que sigue la expresión 3.3 permite construir diferentes contrastes de hipótesis bajo el modelo lineal normal.

Sea $\mathbf{c} = (c_1, \ldots, c_q)'$ un vector de constantes, con la condición de que $\mathbf{c}$ sea combinación lineal de las columnas de $\mathbf{A}$. Planteamos la hipótesis nula

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{c} \tag{5.11}$$

Para decidir la aceptación de H_0, como una consecuencia de 3.3, podemos utilizar el estadístico

$$F = \frac{(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{c})/q}{\mathrm{SCR}/(n-r)} \tag{5.12}$$

[1] en inglés: LSD o *least significant difference*

con distribución $F_{q,n-r}$. Pero es evidente que 5.11 es una hipótesis lineal contrastable, de modo que podemos utilizar el test F que resulta ser idéntico al anterior. Es otra forma de demostrar 5.9 y también que

$$\mathrm{SCR}_H - \mathrm{SCR} = (\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{c})$$

Además, podemos plantear otras hipótesis sobre las funciones paramétricas estimables $\boldsymbol{\psi}$, siempre que sean lineales. Por ejemplo, consideremos ahora la hipótesis lineal planteada sobre las q funciones linealmente independientes

$$H_0 : \psi_1 = \psi_2 = \cdots = \psi_q \tag{5.13}$$

es decir, bajo H_0 las q funciones son iguales. Si consideramos las nuevas funciones

$$\widetilde{\psi}_i = \psi_1 - \psi_{i+1} \qquad i = 1, \ldots, q-1$$

entonces 5.13 se reduce a 5.11 tomando $\widetilde{\boldsymbol{\psi}} = (\widetilde{\psi}_1, \ldots, \widetilde{\psi}_{q-1})'$, $\mathbf{c} = \mathbf{0}$ y sustituyendo q por $q-1$. Dicho de otra manera, sea la matriz

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{q1} & a_{q2} & \ldots & a_{qm} \end{pmatrix}$$

Entonces 5.13 es equivalente a la hipótesis lineal

$$H_0 : \mathbf{A}^*\boldsymbol{\beta} = \mathbf{0}$$

tomando como matriz de hipótesis

$$\mathbf{A}^* = \begin{pmatrix} a_{11} - a_{21} & a_{12} - a_{22} & \ldots & a_{1m} - a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{11} - a_{q1} & a_{12} - a_{q2} & \ldots & a_{1m} - a_{qm} \end{pmatrix}$$

Luego podemos utilizar el estadístico F de 5.6, con $\mathbf{A}^*$ y $q-1$, que bajo H_0 tiene distribución $F_{q-1,n-r}$, para decidir si 5.13 debe ser aceptada.

5.6. Elección entre dos modelos lineales

5.6.1. Sobre los modelos

Para la estimación en el modelo lineal

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad\qquad E(\boldsymbol{\epsilon}) = \mathbf{0}, \mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$$

hemos establecido (ver pág. 28) que el punto crucial es la utilización de la matriz $\mathbf{P}$, proyección ortogonal sobre el espacio de las estimaciones $\Omega = \langle \mathbf{X} \rangle$. Así, dos modelos son iguales si tienen el mismo espacio de las estimaciones. Dos de estos modelos darán las mismas predicciones y el mismo estimador de σ^2.

Sean $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$ y $\mathbf{Y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2$ dos modelos lineales tales que $\langle \mathbf{X}_1 \rangle = \langle \mathbf{X}_2 \rangle$. La matriz proyección no depende de $\mathbf{X}_1$ o $\mathbf{X}_2$ sino sólo de $\Omega (= \langle \mathbf{X}_1 \rangle = \langle \mathbf{X}_2 \rangle)$. La estimación de σ^2 es la misma $\widehat{\sigma}^2 = \mathrm{SCR}/(n-r)$ y las predicciones también

$$\widehat{\mathbf{Y}} = \mathbf{P}\mathbf{Y} = \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 = \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2$$

En cuanto a las funciones paramétricas estimables, hemos visto que la estimabilidad se restringe a las combinaciones lineales de las filas $\mathbf{X}_1$, es decir, $\mathbf{a}_1'\boldsymbol{\beta}_1$ es estimable si se escribe como $\mathbf{b}'\mathbf{X}_1\boldsymbol{\beta}_1$. Pero $\mathbf{X}_1\boldsymbol{\beta}_1$ pertenece a Ω de forma que $\mathbf{X}_1\boldsymbol{\beta}_1 = \mathbf{X}_2\boldsymbol{\beta}_2$ para algún $\boldsymbol{\beta}_2$ y así

$$\mathbf{a}_1'\boldsymbol{\beta}_1 = \mathbf{b}'\mathbf{X}_1\boldsymbol{\beta}_1 = \mathbf{b}'\mathbf{X}_2\boldsymbol{\beta}_2 = \mathbf{a}_2'\boldsymbol{\beta}_2$$

Las funciones paramétricas estimables son las mismas pero están escritas con diferentes parámetros. Su estimador $\mathbf{b}'\mathbf{P}\mathbf{Y}$ también es único.

Ejemplo 5.6.1

El ANOVA de un factor se puede escribir de dos formas:

$$\begin{array}{ll} y_{ij} = \mu + \alpha_i + \epsilon_{ij} & i = 1, \ldots, I,\ j = 1, \ldots, n_i \\ y_{ij} = \mu_i + \epsilon_{ij} & i = 1, \ldots, I,\ j = 1, \ldots, n_i \end{array}$$

pero son equivalentes puesto que $\langle \mathbf{X}_1 \rangle = \langle \mathbf{X}_2 \rangle$.
En este modelo las relaciones entre los dos conjuntos de parámetros son sencillas

$$\mu_i = \mu + \alpha_i \qquad \mu_1 - \mu_2 = \alpha_1 - \alpha_2 \qquad \textit{etc.}$$

Ejemplo 5.6.2

La regresión lineal simple admite dos modelos:

$$\begin{array}{ll} y_i = \beta_0 + \beta_1 x_i + \epsilon_i & i = 1, \ldots, n \\ y_i = \gamma_0 + \gamma_1 (x_i - \bar{x}) + \epsilon_i & i = 1, \ldots, n \end{array}$$

pero son equivalentes ya que

$$\begin{aligned} \gamma_0 &= \beta_0 + \beta_1 \bar{x} \\ \gamma_1 &= \beta_1 \end{aligned}$$

En resumen, en un modelo lineal $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ la esencia es el subespacio $\Omega = \langle \mathbf{X} \rangle$. Si conservamos Ω, podemos cambiar $\mathbf{X}$ a nuestra conveniencia.

5.6.2. Contraste de modelos

El contraste de hipótesis en modelos lineales se reduce esencialmente a restringir el espacio de las estimaciones.
Si partimos de un modelo que sabemos o suponemos válido

$$\text{Modelo inicial:} \qquad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \mathrm{rg}\ \mathbf{X} = r$$

debemos intentar reducir este modelo, es decir, ver si algún modelo más simple se ajusta aceptablemente a los datos, como

$$\text{Modelo restringido:} \qquad \mathbf{Y} = \widetilde{\mathbf{X}}\boldsymbol{\theta} + \boldsymbol{\epsilon} \qquad \mathrm{rg}\ \widetilde{\mathbf{X}} = \widetilde{r}$$

Dado que la esencia de un modelo está en el subespacio generado por las columnas de la matriz de diseño o espacio de las estimaciones, es absolutamente necesario que el modelo restringido verifique

$$\Omega_0 = \langle \widetilde{\mathbf{X}} \rangle \subset \langle \mathbf{X} \rangle = \Omega$$

Sólo en este caso se puede plantear la elección entre dos modelos alternativos como un contraste de hipótesis

$$\begin{array}{l} H_0 : \mathbf{Y} = \widetilde{\mathbf{X}}\boldsymbol{\theta} + \boldsymbol{\epsilon} \\ H_1 : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \end{array} \Leftrightarrow \begin{array}{l} H_0 : E(\mathbf{Y}) \in \Omega_0 = \langle \widetilde{\mathbf{X}} \rangle \\ H_1 : E(\mathbf{Y}) \in \Omega = \langle \mathbf{X} \rangle \end{array} \tag{5.14}$$

donde $E(\mathbf{Y}) = \widetilde{\mathbf{X}}\boldsymbol{\theta}$ y $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, respectivamente.

Sean $\mathbf{P}_\Omega$ y $\mathbf{P}_{\Omega_0}$ las proyecciones ortogonales sobre $\Omega = \langle \mathbf{X} \rangle$ y $\Omega_0 = \langle \widetilde{\mathbf{X}} \rangle$ respectivamente. Bajo el modelo inicial el estimador de $E(\mathbf{Y})$ es $\mathbf{P}_\Omega \mathbf{Y}$, mientras que bajo el modelo restringido el estimador es $\mathbf{P}_{\Omega_0}\mathbf{Y}$. Si la hipótesis H_0 es cierta, ambas estimaciones deben estar próximas.

Teorema 5.6.1

La condición necesaria y suficiente para que 5.14 sea contrastable es que se verifique

$$\Omega_0 = \langle \widetilde{\mathbf{X}} \rangle \subset \langle \mathbf{X} \rangle = \Omega \tag{5.15}$$

El test F se basa entonces en el estadístico

$$F = \frac{(\mathrm{SCR}_H - \mathrm{SCR})/(r - \widetilde{r})}{\mathrm{SCR}/(n - r)}$$

cuya distribución, bajo H_0, es $F_{r-\widetilde{r},n-r}$ y donde

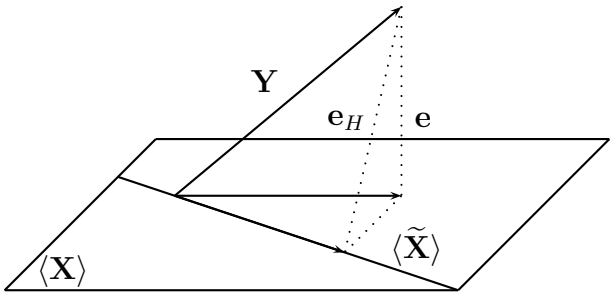
$$\mathrm{SCR}_H = \mathbf{Y}'(\mathbf{I} - \mathbf{P}_{\Omega_0})\mathbf{Y} \qquad \mathrm{SCR} = \mathbf{Y}'(\mathbf{I} - \mathbf{P}_\Omega)\mathbf{Y}$$

Demostración:

La expresión 5.15 implica la relación $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{C}$ para una cierta matriz $\mathbf{C}$. Entonces H_0 significa formular una hipótesis lineal contrastable al modelo $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, que lo reduce a $E(\mathbf{Y}) = \widetilde{\mathbf{X}}\boldsymbol{\theta}$. El resto es consecuencia del *Método 1* explicado en la sección 5.2 y el teorema 5.3.1. ■

Observemos que si $\Omega_0 \not\subseteq \Omega$, entonces estamos ante modelos de naturaleza diferente. No podemos decidir entre ambos modelos mediante ningún criterio estadístico conocido. Si se verifica $\Omega_0 = \Omega$, entonces tenemos dos versiones paramétricas del mismo modelo, pudiendo pasar del uno al otro por una reparametrización. Un modelo $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$ determina el espacio $\Omega = \langle \mathbf{X} \rangle$, y recíprocamente el espacio Ω determina el modelo (salvo reparametrizaciones que no disminuyan el rango).

Como ya hemos visto, la interpretación geométrica de la solución al modelo lineal $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ es considerar la proyección del vector $\mathbf{Y}$ sobre el subespacio $\Omega = \langle \mathbf{X} \rangle$ de $\mathbb{R}^n$. La relación 5.15 indica que las columnas de $\widetilde{\mathbf{X}}$ generan un subespacio de $\langle \mathbf{X} \rangle$. Entonces SCR y SCR_H son distancias de la observación $\mathbf{Y}$ a los subespacios $\langle \mathbf{X} \rangle$ y $\langle \widetilde{\mathbf{X}} \rangle$, respectivamente. El test F nos dice hasta que punto la diferencia $\mathrm{SCR}_H - \mathrm{SCR}$ es pequeña (comparada con SCR) para poder afirmar que el modelo se ajusta al subespacio $\langle \widetilde{\mathbf{X}} \rangle$ en lugar de $\langle \mathbf{X} \rangle$ (ver figura).



La longitud al cuadrado de la diferencia $\mathbf{P}_\Omega\mathbf{Y} - \mathbf{P}_{\Omega_0}\mathbf{Y}$ es

$$((\mathbf{P}_\Omega - \mathbf{P}_{\Omega_0})\mathbf{Y})'((\mathbf{P}_\Omega - \mathbf{P}_{\Omega_0})\mathbf{Y}) = \mathbf{Y}'(\mathbf{P}_\Omega - \mathbf{P}_{\Omega_0})\mathbf{Y}$$

ya que $\mathbf{P}_\Omega - \mathbf{P}_{\Omega_0} = \mathbf{P}_{\Omega_0^\perp \cap \Omega}$ es una matriz proyección (ver Apéndice). Pero además

$$\mathbf{Y}'(\mathbf{P}_\Omega - \mathbf{P}_{\Omega_0})\mathbf{Y} = \mathbf{Y}'(\mathbf{I} - \mathbf{P}_{\Omega_0})\mathbf{Y} - \mathbf{Y}'(\mathbf{I} - \mathbf{P}_\Omega)\mathbf{Y} = \mathrm{SCR}_H - \mathrm{SCR}$$

Cuando la hipótesis nula se plantea en términos de un grupo de funciones paramétricas estimables del tipo $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{0}$, sabemos que existe una matriz $\mathbf{B} = \mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$ tal que $\mathbf{A} = \mathbf{B}\mathbf{X}$. De modo que

$$\mathbf{0} = \mathbf{A}\boldsymbol{\beta} = \mathbf{B}\mathbf{X}\boldsymbol{\beta} = \mathbf{B}E(\mathbf{Y}) \Leftrightarrow E(\mathbf{Y}) \in \ker(\mathbf{B})$$

y el subespacio que define la hipótesis nula es $\Omega_0 = \ker(\mathbf{B}) \cap \Omega$. En este caso se puede demostrar (ver Apéndice) que $\Omega_0^\perp \cap \Omega = \langle \mathbf{P}_\Omega \mathbf{B}' \rangle$ y reencontrar así el test 5.6.

Ejemplo 5.6.3

Consideremos de nuevo el diseño cross-over explicado en el ejemplo 5.3.2. Supongamos ahora que la influencia γ de un fármaco sobre el otro no es recíproca. El efecto aditivo no es necesariamente el mismo cuando se administra **a** *después de* **b**, *que cuando se administra* **b** *después de* **a**. *Entonces debemos introducir los parámetros*

$$\begin{aligned} \gamma_1 &: \textit{influencia de } \mathbf{a} \textit{ sobre } \mathbf{b} \\ \gamma_2 &: \textit{influencia de } \mathbf{b} \textit{ sobre } \mathbf{a} \end{aligned}$$

y admitir que la matriz de diseño reducida, para los parámetros $\mu, \alpha, \beta, \gamma_1, \gamma_2$ es

$$\mathbf{X}_R = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix} \qquad \mathrm{rg}\ \mathbf{X}_R = 4$$

que representa una alternativa a la propuesta inicialmente para los parámetros $\mu, \alpha, \beta, \gamma$

$$\widetilde{\mathbf{X}}_R = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \qquad \mathrm{rg}\ \widetilde{\mathbf{X}}_R = 3$$

Es fácil ver que se verifica 5.15. El análisis de la varianza para decidir entre $\widetilde{\mathbf{X}}_R$ y $\mathbf{X}_R$, sobre los datos de la tabla 5.2, se encuentra en la tabla 5.4. Como F no es significativo se admite como válido el modelo más simple representado por $\widetilde{\mathbf{X}}_R$.

	grados de libertad	suma de cuadrados	cuadrados medios	F
Desviación hipótesis	1	600,6	600,6	3,898
Residuo	36	5547,3	154,1	

Cuadro 5.4: Tabla del análisis de la varianza para contrastar dos modelos de *cross-over*

5.7. Ejemplos con R

En esta sección vamos a ver como se contrastan las hipótesis que hemos planteado en el ejemplo 5.3.2 sobre el diseño cross-over simplificado.

En primer lugar procedemos a introducir los datos en el vector de observaciones.

```
> y<-c(17,34,26,10,19,17,8,16,13,11,
+ 17,41,26,3,-6,-4,11,16,16,4,
+ 21,20,11,26,42,28,3,3,16,-10,
+ 10,24,32,26,52,28,27,28,21,42)
```

A continuación construimos las columnas de la matriz de diseño que corresponden a los parámetros α, β, γ con las funciones de repetición.

```
> alpha<-c(rep(1,10),rep(0,10),rep(0,10),rep(1,10))
> beta<-c(rep(0,10),rep(1,10),rep(1,10),rep(0,10))
> gamma<-c(rep(0,10),rep(1,10),rep(0,10),rep(1,10))
```

Los modelos lineales se definen en R con la función `lm`. Así, el modelo general y el modelo bajo la hipótesis nula se definen como

```
> crossover.lm<-lm(y~alpha+beta+gamma)
> crossover.lm0<-lm(y~gamma)
```

La columna de unos que corresponde al parámetro μ no es necesario escribirla, ya que por defecto está incluida en cualquier modelo lineal de R así definido. Observemos además que bajo la hipótesis nula $H_0 : \alpha = \beta$, el modelo a considerar sólo tiene dos parámetros μ, γ. En este caso, el efecto del fármaco (común) se puede incluir en la media general.

La tabla del análisis de la varianza para el contraste de la hipótesis nula considerada se realiza mediante la función `anova`(*modelo* H_0,*modelo general*).

```
> anova(crossover.lm0,crossover.lm)
Analysis of Variance Table

Model 1: y ~ gamma
Model 2: y ~ alpha + beta + gamma
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     38 6931.1
2     37 6147.9  1     783.2 4.7137 0.03641 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Del mismo modo también se puede realizar el contraste de modelos propuesto en el ejemplo 5.6.3. En este caso, el modelo más general necesita las columnas correspondientes a los parámetros γ_1, γ_2.

```
> gamma1<-c(rep(0,10),rep(1,10),rep(0,10),rep(0,10))
> gamma2<-c(rep(0,10),rep(0,10),rep(0,10),rep(1,10))
> crossover.lm1<-lm(y~alpha+beta+gamma1+gamma2)
> anova(crossover.lm,crossover.lm1)
Analysis of Variance Table

Model 1: y ~ alpha + beta + gamma
Model 2: y ~ alpha + beta + gamma1 + gamma2
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     37 6147.9
2     36 5547.3  1     600.6 3.8978 0.05606 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5.8. Ejercicios

Ejercicio 5.1

Sean $X \sim N(\mu_1, \sigma)$, $Y \sim N(\mu_2, \sigma)$ variables independientes. En muestras de extensión n_1 de X, n_2 de Y, plantear la hipótesis nula

$$H_0 : \mu_1 = \mu_2$$

mediante el concepto de hipótesis lineal contrastable y deducir el test t de Student de comparación de medias como una consecuencia del test F.

Ejercicio 5.2

Una variable Y depende de otra x (variable control no aleatoria) que toma los valores $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$ de acuerdo con el modelo lineal normal

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

Encontrar la expresión del estadístico F para la hipótesis

$$H_0 : \beta_2 = 0$$

Ejercicio 5.3

Probar que una hipótesis lineal de matriz $\mathbf{A}$ es contrastable si y sólo si

$$\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{A}$$

Ejercicio 5.4

Con el modelo del ejercicio 3.10:

(a) ¿Podemos contrastar la hipótesis $H_0 : \theta_1 + \theta_8 = 0$?

(b) Contrastar la hipótesis $H_0 : \theta_1 = \theta_2$.

Ejercicio 5.5

Dado el siguiente modelo lineal normal

$$\begin{aligned} \beta_1 + \beta_2 &= 6{,}6 \\ 2\beta_1 + \beta_2 &= 7{,}8 \\ -\beta_1 + \beta_2 &= 2{,}1 \\ 2\beta_1 - \beta_2 &= 0{,}4 \end{aligned}$$

estudiar si se puede aceptar la hipótesis $H_0 : \beta_2 = 2\beta_1$.

Ejercicio 5.6

Consideremos el modelo lineal normal $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Probar que para la hipótesis lineal

$$H_0 : \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

se verifica $\mathrm{SCR}_H - \mathrm{SCR} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$. Hallar el estadístico F correspondiente.

Ejercicio 5.7

Demostrar que para una hipótesis lineal contrastable se verifica

$$E(\mathrm{SCR}_H - \mathrm{SCR}) = q\sigma^2 + (\mathbf{A}\boldsymbol{\beta})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^-\mathbf{A}')^-(\mathbf{A}\boldsymbol{\beta})$$

Indicación: Utilizar la propiedad 2 del Apéndice de Estadística Multivariante con la expresión 5.5.

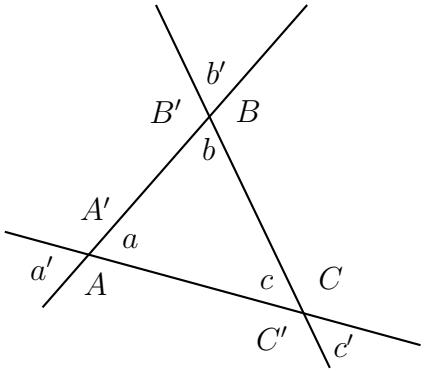
Ejercicio 5.8

Demostrar que para una hipótesis lineal contrastable se verifica la siguiente descomposición en suma de cuadrados

$$\|\mathbf{Y} - \widehat{\mathbf{Y}}_H\|^2 = \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 + \|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_H\|^2$$

Ejercicio 5.9

Supongamos que cada uno de los valores $x_1, x_2, \ldots, x_{12}$ son las observaciones de los ángulos $a, a', A, A', b, b', B, B', c, c', C, C'$ del triángulo del gráfico adjunto. Los errores de las observaciones $\epsilon_1, \ldots, \epsilon_{12}$ se asume que son independientes y con distribución $N(0, \sigma)$.



Antes de escribir el modelo asociado a estos datos observemos que, aunque aparentemente hay 12 parámetros $a, a', \ldots$, éstos están ligados por las conocidas propiedades de un triángulo, es decir

$$a = a' \qquad A = A' \qquad a + A = 180 \qquad a + b + c = 180$$

y de forma similar para b, b', B, B' y c, c', C, C'. El conjunto de estas relaciones nos conduce a que, realmente, sólo hay dos parámetros independientes, les llamaremos α y β. Si trasladamos a la izquierda las cantidades 180 y con estos parámetros, el modelo es

$$\begin{array}{llll} y_1 = \alpha + \epsilon_1 & y_2 = \alpha + \epsilon_2 & y_3 = -\alpha + \epsilon_3 & y_4 = -\alpha + \epsilon_4 \\ y_5 = \beta + \epsilon_5 & y_6 = \beta + \epsilon_6 & y_7 = -\beta + \epsilon_7 & y_8 = -\beta + \epsilon_8 \\ y_9 = -\alpha - \beta + \epsilon_9 & y_{10} = -\alpha - \beta + \epsilon_{10} & y_{11} = \alpha + \beta + \epsilon_{11} & y_{12} = \alpha + \beta + \epsilon_{12} \end{array}$$

donde

$$\begin{array}{llll} y_1 = x_1 & y_2 = x_2 & y_3 = x_3 - 180 & y_4 = x_4 - 180 \\ y_5 = x_5 & y_6 = x_6 & y_7 = x_7 - 180 & y_8 = x_8 - 180 \\ y_9 = x_9 - 180 & y_{10} = x_{10} - 180 & y_{11} = x_{11} & y_{12} = x_{12} \end{array}$$

Deseamos contrastar la hipótesis de que el triángulo es equilátero, es decir, que $a = b = c = 60$. Pero si $a = 60, b = 60$, c es automáticamente 60, luego la hipótesis es

$$H_0 : \alpha = \beta = 60$$

con 2 grados de libertad, no 3. Resolver el contraste.

Ejercicio 5.10

Con el modelo cross-over expuesto en el ejemplo 5.3.2 calcular los siguientes elementos:

(a) Una estimación de los parámetros mediante la fórmula $(\mathbf{X}_R'\mathbf{D}\mathbf{X}_R)^-\mathbf{X}_R'\mathbf{D}\bar{\mathbf{Y}}$.

(b) La suma de cuadrados residual

$$\begin{aligned}\text{SCR} = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{P}\mathbf{Y} &= \sum y_{ij}^2 - \mathbf{Y}'\mathbf{P}\mathbf{Y} \\ &= N_{\mathbf{a}}\left(\sum_{i=1}^4 \bar{y}_{i\cdot}^2 + \sum_{i=1}^4 s_i^2\right) - \mathbf{Y}'\mathbf{P}\mathbf{Y} \\ &= N_{\mathbf{a}}\left(\sum_{i=1}^4 \bar{y}_{i\cdot}^2 + \sum_{i=1}^4 s_i^2\right) - \bar{\mathbf{Y}}'\mathbf{D}\mathbf{X}_R(\mathbf{X}_R'\mathbf{D}\mathbf{X}_R)^-\mathbf{X}_R'\mathbf{D}\bar{\mathbf{Y}}\end{aligned}$$

(c) La estimación de la función paramétrica $\alpha - \beta$ y su varianza.

(d) El estadístico con distribución t de Student para contrastar la hipótesis $H_0 : \alpha = \beta$

$$t = \frac{\hat{\alpha} - \hat{\beta}}{\text{ee}(\hat{\alpha} - \hat{\beta})}$$

cuyo cuadrado coincide con el estadístico F del ejemplo.

Capítulo 6

Regresión lineal simple

Sea Y una variable aleatoria y x una variable controlable, es decir, los valores que toma x son fijados por el experimentador. Supongamos que calculamos Y para diferentes valores de x de acuerdo con el siguiente modelo

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n \tag{6.1}$$

donde $E(\epsilon_i) = 0$, $\text{var}(\epsilon_i) = \sigma^2 \quad i = 1, \ldots, n$.
Este modelo es la formulación lineal del problema de hallar la recta de regresión de Y sobre x. Los parámetros β_0, β_1 reciben el nombre de coeficientes de regresión. El parámetro β_0 es la ordenada en el origen, *intercept* en inglés, y β_1 es la pendiente de la recta, *slope* en inglés. La expresión matricial de 6.1 es

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \qquad \text{rg } \mathbf{X} = 2$$

Ahora podemos aplicar toda la teoría general desarrollada en los capítulos anteriores para un modelo lineal cualquiera, al caso particular de la regresión lineal simple.

6.1. Estimación de los coeficientes de regresión

Con los datos observados se pueden calcular los siguientes estadísticos

$$\begin{array}{ll} \bar{x} = (1/n)\sum x_i & s_x^2 = (1/n)\sum (x_i - \bar{x})^2 \\ \bar{y} = (1/n)\sum y_i & s_y^2 = (1/n)\sum (y_i - \bar{y})^2 \end{array}$$

$$s_{xy} = (1/n)\sum (x_i - \bar{x})(y_i - \bar{y})$$

donde $\bar{x}, \bar{y}, s_x^2, s_y^2, s_{xy}$ son las medias, varianzas y covarianzas muestrales, aunque el significado de s_x^2 y s_{xy} es *convencional* pues x no es variable aleatoria. Con esta notación las ecuaciones normales son:

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y} \qquad \Leftrightarrow \qquad \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} n\bar{y} \\ \sum x_i y_i \end{pmatrix}$$

y como

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n s_x^2} \begin{pmatrix} (1/n)\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

la solución es

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{S_{xy}}{S_x} = \frac{s_{xy}}{s_x^2}$$

donde

$$S_{xy} = \sum x_i y_i - (1/n) \sum x_i \sum y_i = \sum (x_i - \bar{x})(y_i - \bar{y}) = n\, s_{xy}$$
$$S_x = \sum x_i^2 - (1/n)(\sum x_i)^2 = \sum (x_i - \bar{x})^2 = n\, s_x^2$$

En el ejercicio 6.2 se ven otras formas de expresar $\hat{\beta}_1$.

La recta de regresión es

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

que se expresa también en la forma

$$y - \bar{y} = \hat{\beta}_1 (x - \bar{x})$$

lo que deja claro que la recta pasa por el punto $(\bar{x}, \bar{y})$ y que el modelo es válido en el rango de las x_i, centrado en $\bar{x}$. Ésta es también la recta que se obtiene a partir del modelo equivalente con los datos x_i centrados (ver ejemplo 5.6.2 y ejercicio 6.3).

Recordemos que por lo que hemos estudiado, estas estimaciones son insesgadas y de varianza mínima entre todos los estimadores lineales (teorema de Gauss-Markov). Las varianzas y covarianza de los estimadores son

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \begin{pmatrix} \mathrm{var}(\hat{\beta}_0) & \mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) & \mathrm{var}(\hat{\beta}_1) \end{pmatrix} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \tag{6.2}$$

Es decir

$$E(\hat{\beta}_0) = \beta_0 \qquad \mathrm{var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_x} \right) \tag{6.3}$$

$$E(\hat{\beta}_1) = \beta_1 \qquad \mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_x} \tag{6.4}$$

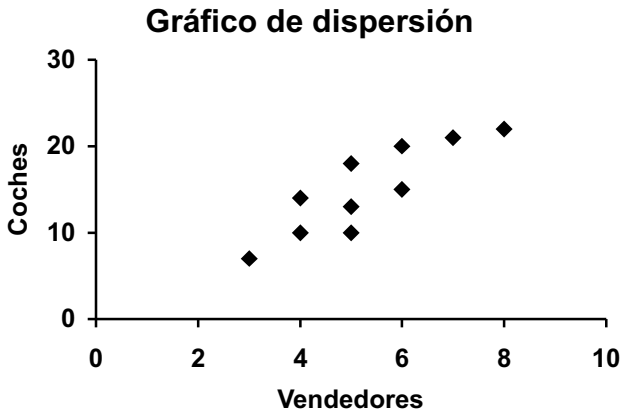
$$\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{S_x} \tag{6.5}$$

Ejemplo 6.1.1

Vamos a ilustrar el cálculo "manual" de las estimaciones de los parámetros con un ejemplo muy sencillo de muy pocos datos.

Supongamos que una empresa de compra-venta de automóviles organiza exposiciones los fines de semana i contrata un número variable de vendedores que oscila entre 3 *y* 8*. El gerente de esta empresa quiere estudiar la relación entre el número de vendedores y el número de coches vendidos ya que, si es posible, podría prever las ventas a partir del número de vendedores que contrata. Para aclararlo, el gerente examina el registro de ventas de los últimos cuatro meses y localiza un período de* 10 *semanas durante las cuales no hubo ningún incentivo especial ni a la venta ni a la compra. El número de*

Semana	Vendedores	Coches
1	5	10
2	6	20
3	5	18
4	4	10
5	3	7
6	4	14
7	7	21
8	6	15
9	5	13
10	8	22



Cuadro 6.1: Datos de las ventas en 10 semanas y gráfico de dispersión

coches vendidos durante este período y el número de vendedores empleados en cada caso se muestra en la tabla adjunta.

Para examinar esta relación es muy útil empezar por dibujar un diagrama de dispersión. Este gráfico muestra una relación bastante evidente entre el número de vendedores y las ventas, relación que se podía esperar. Vamos a cuantificarla con la ayuda de la recta de regresión MC.

En la siguiente tabla tenemos los cálculos necesarios para obtener los coeficientes de regresión, las predicciones, los residuos y la suma de cuadrados de los errores para los datos de las 10 *semanas. Esta tabla se ha calculado con una hoja de cálculo, lo que permite una mayor precisión en los cálculos sucesivos.*

i	x_i	y_i	x_i^2	$x_i y_i$	$\hat{y}_i$	e_i	e_i^2
1	5	10	25	50	14,10	−4,10	16,85
2	6	20	36	120	17,09	2,91	8,47
3	5	18	25	90	14,10	3,90	15,18
4	4	10	16	40	11,12	−1,12	1,25
5	3	7	9	21	8,13	−1,13	1,29
6	4	14	16	56	11,12	2,88	8,30
7	7	21	49	147	20,07	0,93	0,86
8	6	15	36	90	17,09	−2,09	4,37
9	5	13	25	65	14,10	−1,10	1,22
10	8	22	64	176	23,06	−1,06	1,12
Suma	53	150	301	855		0	58,90
Media	5,3	15					

Cuadro 6.2: Cálculos de regresión simple para los datos de ventas

Con estos cálculos, las estimaciones de los coeficientes de regresión son

$$\hat{\beta}_1 = \frac{855 - \frac{1}{10}\,53 \cdot 150}{301 - \frac{1}{10}\,(53)^2} = 2{,}9850746$$

$$\hat{\beta}_0 = 15 - \hat{\beta}_1 \cdot 5{,}3 = -0{,}820896$$

La ecuación de la recta de regresión es

$$y = -0{,}821 + 2{,}985x$$

o también

$$y - 15 = 2{,}985\,(x - 5{,}3)$$

Para calcular la precisión de estas estimaciones, primero debemos estimar la varianza del modelo.

Nota: Una aplicación de hojas de cálculo como *Microsoft Excel* tiene la función ESTIMACION.LINEAL que calcula de forma directa los coeficientes de regresión y algunos estadísticos más. Otra función matricial es TENDENCIA que permite calcular directamente las predicciones. Además, *Excel* lleva un conjunto de macros opcionales llamadas "Herramientas para análisis" que, entre otras cosas, calculan una regresión lineal completa.

En el ejemplo anterior, se comprueba que las suma de los residuos es cero, salvo problemas de redondeo. Esto no es una casualidad. Vamos a ver algunas propiedades adicionales para las predicciones $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ y para los residuos $e_i = y_i - \hat{y}_i$, cuya demostración se deja para el lector (ver ejercicio 6.4).

(i) La suma de los residuos es cero: $\sum e_i = 0$.

(ii) La suma de los residuos ponderada por los valores de la variable regresora es cero: $\sum x_i e_i = 0$.

(iii) $\sum y_i = \sum \hat{y}_i$

(iv) La suma de los residuos ponderada por las predicciones de los valores observados es cero: $\sum \hat{y}_i e_i = 0$.

6.2. Medidas de ajuste

La evaluación global del ajuste de la regresión se puede hacer con la SCR o, mejor, con la varianza muestral de los residuos $(1/n)\sum e_i^2$. Pero los residuos no son todos independientes, si no que están ligados por dos ecuaciones (la (i) y la (ii) de arriba), de forma que utilizaremos la llamada *varianza residual* o estimación MC de σ^2:

$$\hat{\sigma}^2 = \mathrm{SCR}/(n-2)$$

Su raíz cuadrada $\hat{\sigma}$, que tiene las mismas unidades que Y, es el llamado *error estándar de la regresión*. La varianza residual o el error estándar son índices de la precisión del modelo, pero dependen de las unidades de la variable respuesta y no son útiles para comparar rectas de regresión de variables diferentes. Otra medida de ajuste requiere una adecuada descomposición de la variabilidad de la variable respuesta.

Teorema 6.2.1

Consideremos el coeficiente de correlación muestral, cuyo significado es convencional,

$$r = \frac{s_{xy}}{s_x s_y} = \frac{S_{xy}}{(S_x S_y)^{1/2}}$$

Entonces se verifican las siguientes relaciones

(i) $\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2$

(ii) $\text{SCR} = \sum(y_i - \hat{y}_i)^2 = (1 - r^2)\sum(y_i - \bar{y})^2 = (1 - r^2)S_y$

(iii) $\hat{\sigma}^2 = (\sum e_i^2)/(n-2) = (1 - r^2)S_y/(n-2)$

Demostración:

$$\begin{aligned}\sum(y_i - \bar{y})^2 &= \sum(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2 + 2\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})\end{aligned}$$

pero $\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum(y_i - \hat{y}_i)\hat{y}_i - \bar{y}\sum(y_i - \hat{y}_i) = 0$ por las propiedades del apartado anterior. También podemos recordar la ortogonalidad de los subespacios de los errores y de las estimaciones. Queda así demostrada la relación (i).

Por otra parte, es fácil ver que

$$\sum(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum(x_i - \bar{x})^2 = r^2 \sum(y_i - \bar{y})^2$$

de forma que finalmente

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + r^2 \sum(y_i - \bar{y})^2$$

Luego

$$\sum(y_i - \hat{y}_i)^2 = (1 - r^2)\sum(y_i - \bar{y})^2$$

Como consecuencia tenemos que el estimador centrado de la varianza σ^2 del modelo 6.1 es

$$\hat{\sigma}^2 = \text{SCR}/(n-2) = (1 - r^2)S_y/(n-2) \tag{6.6}$$

■

La descomposición de la suma de cuadrados de las observaciones en dos términos independientes se interpreta así: la variabilidad de la variable Y se descompone en un primer término que refleja la variabilidad no explicada por la regresión, que es debida al azar, y el segundo término que contiene la variabilidad explicada o eliminada por la regresión y puede interpretarse como la parte determinista de la variabilidad de la respuesta.

Podemos definir:

$$\begin{aligned}&\text{Variación total} = \text{VT} = \sum(y_i - \bar{y})^2 = S_y \\ &\text{Variación no explicada} = \text{VNE} = \sum(y_i - \hat{y}_i)^2 = \text{SCR} \\ &\text{Variación explicada} = \text{VE} = \sum(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 S_x\end{aligned}$$

de forma que

$$\text{VT} = \text{VNE} + \text{VE} \tag{6.7}$$

Definición 6.2.1

Una medida del ajuste de la recta de regresión a los datos es la proporción de variabilidad explicada que definimos con el nombre de coeficiente de determinación *así:*

$$R^2 = \frac{\text{VE}}{\text{VT}} = 1 - \frac{\text{SCR}}{S_y}$$

Esta medida se puede utilizar en cualquier tipo de regresión, pero en el caso particular de la regresión lineal simple con una recta tenemos

$$R^2 = 1 - \frac{(1-r^2)S_y}{S_y} = r^2$$

que es el cuadrado del coeficiente de correlación lineal entre las dos variables.

El coeficiente de determinación R^2 es una medida de la bondad del ajuste, $0 \leq R^2 \leq 1$, mientras que el coeficiente de correlación es una medida de la dependencia lineal entre las dos variables, cuando son aleatorias y sólo hay una variable regresora.

Ejemplo 6.2.1

Continuando con el ejemplo de los datos de ventas tenemos:

$$\begin{aligned} \text{SCR} &= 58{,}896 \\ \hat{\sigma}^2 &= 58{,}896/8 = 7{,}362 \qquad \hat{\sigma} = 2{,}713 \\ \text{VT} &= S_y = 238 \\ R^2 &= 1 - \frac{58{,}896}{238} = 0{,}7525 \end{aligned}$$

6.3. Inferencia sobre los parámetros de regresión

Supongamos que el modelo 6.1 es un modelo lineal normal. Entonces (ver teorema 2.6.1) se verifica que

$$\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)' \sim N_2(\boldsymbol{\beta}, \text{var}(\widehat{\boldsymbol{\beta}}))$$

donde

$$\text{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} 1/n + \bar{x}/S_x & -\bar{x}/S_x \\ -\bar{x}/S_x & 1/S_x \end{pmatrix}$$

como hemos visto en 6.2–6.5. Además sabemos que $\widehat{\boldsymbol{\beta}}$ es independiente de SCR.

Como consecuencia de estas distribuciones hemos demostrado (ver 3.4 o 5.10) que para contrastar una hipótesis del tipo $H_0 : \mathbf{a}'\boldsymbol{\beta} = c$ se utiliza el estadístico

$$t = \frac{\mathbf{a}'\widehat{\boldsymbol{\beta}} - c}{(\hat{\sigma}^2(\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}))^{1/2}} \tag{6.8}$$

que seguirá una distribución t_{n-2}, si H_0 es cierta.

6.3.1. Hipótesis sobre la pendiente

El contraste de la hipótesis $H_0 : \beta_1 = b_1$ frente a $H_1 : \beta_1 \neq b_1$ se resuelve rechazando H_0 si

$$\left| \frac{\hat{\beta}_1 - b_1}{(\hat{\sigma}^2/S_x)^{1/2}} \right| > t_{n-2}(\alpha)$$

donde $P[|t_{n-2}| > t_{n-2}(\alpha)] = \alpha$.

En particular, estamos interesados en contrastar si la pendiente es cero, es decir, la hipótesis $H_0 : \beta_1 = 0$. Vamos a deducir este contraste directamente.

Si $H_0 : \beta_1 = 0$ es cierta, el modelo 6.1 se simplifica y se convierte en

$$y_i = \beta_0 + \epsilon_i$$

de donde

$$\mathrm{SCR}_H = \sum (y_i - \hat{\beta}_{0|H})^2 = \sum (y_i - \bar{y})^2 = S_y \tag{6.9}$$

dado que $\hat{\beta}_{0|H} = \bar{y}$.

Por el teorema 6.2.1 sabemos que $\mathrm{SCR} = (1 - r^2)S_y$, de manera que

$$F = \frac{\mathrm{SCR}_H - \mathrm{SCR}}{\mathrm{SCR}/(n-2)} = \frac{S_y - (1-r^2)S_y}{(1-r^2)S_y/(n-2)} = (n-2)\frac{r^2}{1-r^2} \sim F_{1,n-2}$$

Finalmente,

$$t = \sqrt{F} = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}} \tag{6.10}$$

sigue la distribución t de Student con $n-2$ grados de libertad.

Este contraste $H_0 : \beta_1 = 0$ se llama *contraste para la significación de la regresión* y se formaliza en una tabla de análisis de la varianza donde se explicita la descomposición de la suma de cuadrados 6.7.

Fuente de variación	grados de libertad	suma de cuadrados	cuadrados medios	F
Regresión	1	$\hat{\beta}_1 S_{xy}$	$\mathrm{CM_R}$	$\mathrm{CM_R}/\mathrm{ECM}$
Error	$n-2$	SCR	ECM	
Total	$n-1$	S_y		

Cuadro 6.3: Tabla del análisis de la varianza para contrastar la significación de la regresión

El hecho de aceptar $H_0 : \beta_1 = 0$ puede implicar que la mejor predicción para todas las observaciones es $\bar{y}$, ya que la variable x no influye, y la regresión es inútil. Pero también podría pasar que la relación no fuera de tipo recta.

Rechazar la hipótesis $H_0 : \beta_1 = 0$ puede implicar que el modelo lineal 6.1 es adecuado. Pero también podría ocurrir que no lo sea. En todo caso, es muy importante no confundir la significación de la regresión con una prueba de causalidad. Los modelos de regresión únicamente cuantifican la relación lineal entre la variable respuesta y las variables explicativas, una en el caso simple, pero no justifican que éstas sean la causa de aquella.

Tanto la adecuación del modelo 6.1, como la hipótesis de normalidad han de estudiarse a través del análisis de los residuos.

6.3.2. Hipótesis sobre el punto de intercepción

Para el contraste de hipótesis $H_0 : \beta_0 = b_0$, se utiliza el estadístico

$$t = \frac{\hat{\beta}_0 - b_0}{(\hat{\sigma}^2(1/n + \bar{x}^2/S_x))^{1/2}}$$

que, si la hipótesis es cierta, sigue una distribución t de Student con $n-2$ grados de libertad.

6.3.3. Intervalos de confianza para los parámetros

Además de los estimadores puntuales de β_0, β_1 y σ^2, con las distribuciones estudiadas podemos proporcionar intervalos de confianza para estos parámetros. El ancho de estos intervalos estará en función de la calidad de la recta de regresión.

Con la hipótesis de normalidad y teniendo en cuenta las distribuciones de $\hat{\beta}_0$ y $\hat{\beta}_1$ estudiadas, un intervalo de confianza para la pendiente β_1 con nivel de confianza $100(1-\alpha)\,\%$ es

$$\hat{\beta}_1 \pm t_{n-2}(\alpha) \cdot (\hat{\sigma}^2/S_x)^{1/2}$$

donde $t_{n-2}(\alpha)$ es tal que $P[|t_{n-2}| < t_{n-2}(\alpha)] = 1-\alpha$.

Análogamente, para β_0 es

$$\hat{\beta}_0 \pm t_{n-2}(\alpha) \cdot (\hat{\sigma}^2(1/n + \bar{x}^2/S_x))^{1/2}$$

Las cantidades

$$\text{ee}(\hat{\beta}_1) = (\hat{\sigma}^2/S_x)^{1/2} \qquad \text{ee}(\hat{\beta}_0) = (\hat{\sigma}^2(1/n + \bar{x}^2/S_x))^{1/2}$$

son los *errores estándar* de la pendiente $\hat{\beta}_1$ y la intercepción $\hat{\beta}_0$, respectivamente. Se trata de estimaciones de la desviación típica de los estimadores. Son medidas de la precisión de la estimación de los parámetros.

Como sabemos

$$\hat{\sigma}^2 = \frac{\text{SCR}}{n-2} = \frac{1}{n-2} S_y(1-r^2)$$

es el estimador insesgado de σ^2 y la distribución de SCR/σ^2 es $\sim \chi^2_{n-2}$. Así, el intervalo de confianza al $100(1-\alpha)\,\%$ de σ^2 es

$$\frac{\text{SCR}}{\chi^2_{n-2}(\alpha/2)} \leq \sigma^2 \leq \frac{\text{SCR}}{\chi^2_{n-2}(1-\alpha/2)}$$

donde $\chi^2_{n-2}(\alpha/2)$ y $\chi^2_{n-2}(1-\alpha/2)$ son los valores de una χ^2_{n-2} para que la suma de las probabilidades de las colas sea α.

6.3.4. Intervalo para la respuesta media

Uno de los usos principales de los modelos de regresión es la estimación de la respuesta media $E[Y|x_0]$ para un valor particular x_0 de la variable regresora. Asumiremos que x_0 es un valor dentro del recorrido de los datos originales de x. Un estimador puntual insesgado de $E[Y|x_0]$ se obtiene con la predicción

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y} + \hat{\beta}_1(x_0 - \bar{x})$$

Podemos interpretar $\beta_0 + \beta_1 x_0$ como una función paramétrica estimable

$$\beta_0 + \beta_1 x_0 = (1, x_0)\boldsymbol{\beta} = \mathbf{x}_0'\boldsymbol{\beta}$$

cuyo estimador es $\hat{y}_0 = \mathbf{x}_0'\widehat{\boldsymbol{\beta}}$, de manera que

$$\text{var}(\mathbf{x}_0'\widehat{\boldsymbol{\beta}}) = \sigma^2\, \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$$

y el error estándar de $\mathbf{x}_0'\widehat{\boldsymbol{\beta}}$ es

$$\text{ee}(\mathbf{x}_0'\widehat{\boldsymbol{\beta}}) = [\hat{\sigma}^2(1/n + (x_0 - \bar{x})^2/S_x)]^{1/2}$$

Entonces, el intervalo de confianza para la respuesta media $E[Y|x_0]$ es

$$\hat{y}_0 \pm t_{n-2}(\alpha) \cdot \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_x}}$$

Destacaremos el hecho de que evidentemente el ancho del intervalo depende de x_0, es mínimo para $x_0 = \bar{x}$ y crece cuando $|x_0 - \bar{x}|$ crece. Esto es intuitivamente razonable.

6.3.5. Predicción de nuevas observaciones

Otra de las importantes aplicaciones de los modelos de regresión es la predicción de nuevas observaciones para un valor x_0 de la variable regresora. El intervalo definido en el apartado anterior es adecuado para el valor esperado de la respuesta, ahora queremos un intervalo de predicción para una respuesta individual concreta. Estos intervalos reciben el nombre de intervalos de predicción en lugar de intervalos de confianza, ya que se reserva el nombre de intervalo de confianza para los que se construyen como estimación de un parámetro. Los intervalos de predicción tienen en cuenta la variabilidad en la predicción del valor medio y la variabilidad al exigir una respuesta individual.

Si x_0 es el valor de nuestro interés, entonces

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

es el estimador puntual de un nuevo valor de la respuesta $Y_0 = Y|x_0$.

Si consideramos la obtención de un intervalo de confianza para esta futura observación Y_0, el intervalo de confianza para la respuesta media en $x = x_0$ es inapropiado ya que es un intervalo sobre la *media* de Y_0 (un parámetro), no sobre futuras observaciones de la distribución.

Se puede hallar un intervalo de predicción para una respuesta concreta de Y_0 del siguiente modo:

Consideremos la variable aleatoria $Y_0 - \hat{y}_0 \sim N(0, \text{var}(Y_0 - \hat{y}_0))$ donde

$$\text{var}(Y_0 - \hat{y}_0) = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_x}\right)$$

ya que Y_0, una futura observación, es independiente de $\hat{y}_0$.

Si utilizamos el valor muestral de $\hat{y}_0$ para predecir Y_0, obtenemos un intervalo de predicción al $100(1-\alpha)\,\%$ para Y_0

$$\hat{y}_0 \pm t_{n-2}(\alpha) \cdot \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_x}}$$

Este resultado se puede generalizar al caso de un intervalo de predicción al $100(1-\alpha)\,\%$ para la media de k futuras observaciones de la variable respuesta cuando $x = x_0$. Si $\bar{y}_0$ es la media de k futuras observaciones para $x = x_0$, un estimador de $\bar{y}_0$ es $\hat{y}_0$ de forma que el intervalo es

$$\hat{y}_0 \pm t_{n-2}(\alpha) \cdot \hat{\sigma}\sqrt{\frac{1}{k} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_x}}$$

6.3.6. Región de confianza y intervalos de confianza simultáneos

Habitualmente, los intervalos de confianza se dan de forma conjunta para los dos parámetros β_0, β_1 de la regresión simple. Sin embargo, la confianza conjunta de ambos intervalos no es $100(1-\alpha)\,\%$, aunque los dos se hayan construido para verificar ese nivel de confianza. Si deseamos que el nivel de confianza conjunta sea el $100(1-\alpha)\,\%$ debemos construir una región de confianza o, alternativamente, los llamados intervalos de confianza simultáneos.
A partir de la distribución de la ecuación 5.9 sabemos que, en general,

$$F = \frac{(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta})/q}{\mathrm{SCR}/(n-r)} \sim F_{q,n-r}$$

donde, en este caso, $\mathbf{A}\widehat{\boldsymbol{\beta}} = \mathbf{I}\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)'$ y $q = 2$. Así pues

$$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{2\mathrm{ECM}} \sim F_{2,n-2}$$

y

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix}$$

Con esta distribución se puede construir una región de confianza al $100(1-\alpha)\,\%$ para β_0, β_1 conjuntamente que viene dada por la elipse

$$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{2\mathrm{ECM}} \leq F_{2,n-2}(\alpha)$$

Con el mismo objetivo, se pueden utilizar diversos métodos de obtención de intervalos simultáneos del tipo

$$\hat{\beta}_j \pm \Delta \cdot \mathrm{ee}(\hat{\beta}_j) \qquad j = 0, 1$$

Por ejemplo, el método de Scheffé proporciona los intervalos simultáneos

$$\hat{\beta}_j \pm (2F_{2,n-2}(\alpha))^{1/2} \cdot \mathrm{ee}(\hat{\beta}_j) \qquad j = 0, 1$$

6.4. Regresión pasando por el origen

Supongamos que, por alguna razón justificada, el experimentador decide proponer el modelo de regresión simple

$$y_i = \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n$$

que carece del término β_0.
El estimador MC del parámetro β_1 es

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

y su varianza es

$$\mathrm{var}(\hat{\beta}_1) = \frac{1}{(\sum x_i^2)^2} \sum x_i^2 \mathrm{var}(y_i) = \sigma^2 \frac{1}{\sum x_i^2}$$

El estimador de σ^2 es

$$\hat{\sigma}^2 = \text{SCR}/(n-1) = \frac{1}{n-1}\left(\sum y_i^2 - \hat{\beta}_1 \sum x_i y_i\right) \tag{6.11}$$

Con la hipótesis de normalidad se pueden construir intervalos de confianza al $100(1-\alpha)\,\%$ para β_1

$$\hat{\beta}_1 \pm t_{n-1}(\alpha) \cdot \hat{\sigma} \sqrt{\frac{1}{\sum x_i^2}}$$

para $E[Y|x_0]$

$$\hat{y}_0 \pm t_{n-1}(\alpha) \cdot \hat{\sigma} \sqrt{\frac{x_0^2}{\sum x_i^2}}$$

y para predecir una futura observación

$$\hat{y}_0 \pm t_{n-1}(\alpha) \cdot \hat{\sigma} \sqrt{1 + \frac{x_0^2}{\sum x_i^2}}$$

Es preciso estar muy seguros para utilizar este modelo. Frecuentemente la relación entre la variable respuesta Y y la variable regresora x varía cerca del origen. Hay ejemplos en química y en otras ciencias. El diagrama de dispersión nos puede ayudar a decidir el mejor modelo. Si no estamos seguros, es mejor utilizar el modelo completo y contrastar la hipótesis $H_0 : \beta_0 = 0$.

Una medida del ajuste del modelo a los datos es el error cuadrático medio 6.11 que se puede comparar con el del modelo completo 6.6. El coeficiente de determinación R^2 no es un buen índice para comparar los dos tipos de modelos.

Para el modelo sin β_0, la descomposición

$$\sum y_i^2 = \sum (y_i - \hat{y}_i)^2 + \sum \hat{y}_i^2$$

justifica que la definición del coeficiente de determinación sea

$$R_0^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2}$$

que no es comparable con el R^2 de la definición 6.2.1. De hecho puede ocurrir que $R_0^2 > R^2$, aunque $\text{ECM}_0 < \text{ECM}$.

6.5. Correlación

Consideremos la situación en la que las dos variables son aleatorias, tanto la la variable respuesta como la variable explicativa o regresora. De modo que tomamos una muestra aleatoria simple de tamaño n formada por las parejas $(x_1, y_1), \ldots, (x_n, y_n)$ de dos variables aleatorias (X, Y) con distribución conjunta normal bivariante

$$(X,Y)' \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{\mu} = (\mu_1, \mu_2)' \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$$

donde $\text{cov}(X, Y) = \sigma_1\sigma_2\rho$ y ρ es el coeficiente de correlación entre Y y X.

La distribución condicionada de Y dado un valor de $X = x$ es

$$Y|X = x \sim N(\beta_0 + \beta_1 x, \sigma^2_{2\cdot 1})$$

donde

$$\beta_0 = \mu_1 - \mu_2 \frac{\sigma_2}{\sigma_1}\rho$$
$$\beta_1 = \frac{\sigma_2}{\sigma_1}\rho$$
$$\sigma^2_{2\cdot 1} = \sigma^2_2(1 - \rho^2)$$

De modo que la esperanza de $Y|X = x$ es el modelo de regresión lineal simple

$$E[Y|X = x] = \beta_0 + \beta_1 x$$

Además, hay una clara relación entre β_1 y ρ, $\rho = 0 \Leftrightarrow \beta_1 = 0$, en cuyo caso no hay regresión lineal, es decir, el conocimiento de $X = x$ no nos ayuda a predecir Y.

El método de la máxima verosimilitud proporciona estimadores de β_0 y β_1 que coinciden con los estimadores MC.

Ahora también es posible plantearse inferencias sobre el parámetro ρ. En primer lugar, el estimador natural de ρ es

$$r = \frac{S_{xy}}{(S_x S_y)^{1/2}}$$

y

$$\hat{\beta}_1 = \left(\frac{S_y}{S_x}\right)^{1/2} r$$

Así, $\hat{\beta}_1$ y r están relacionados, pero mientras r representa una medida de la asociación entre X e Y, $\hat{\beta}_1$ mide el grado de predicción en Y por unidad de X.

Nota: Ya hemos advertido de que cuando X es una variable controlada, r tiene un significado convencional, porque su magnitud depende de la elección del espaciado de los valores x_i. En este caso, ρ no existe y r no es un estimador.

También sabemos que $r^2 = R^2$, de modo que el coeficiente de determinación es el cuadrado de la correlación.

Finalmente, el principal contraste sobre ρ es el de incorrelación $H_0 : \rho = 0$ que es equivalente a $H_0 : \beta_1 = 0$ y se resuelve con el estadístico

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

que, si H_0 es cierta, sigue una distribución t_{n-2}.

6.6. Carácter lineal de la regresión simple

Supongamos ahora que estamos interesados en decidir si la regresión de Y sobre x es realmente lineal. Consideremos las hipótesis

$$H_0 : Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$H_1 : Y_i = g(x_i) + \epsilon_i$$

donde $g(x)$ es una función no lineal desconocida de x. Sin embargo, vamos a ver que podemos reconducir el contraste a la situación prevista en la sección 5.6.2 para la elección entre dos modelos lineales.

Necesitamos n_i valores de Y para cada x_i. Con un cambio de notación, para cada $i = 1, \ldots, k$, sean

$$\begin{array}{lll} x_i : y_{i1}, \ldots, y_{in_i} & \bar{y}_i = (1/n_i)\sum_j y_{ij} & s^2_{y_i} = (1/n_i)\sum_j (y_{ij} - \bar{y}_i)^2 \\ \bar{y} = (1/n)\sum_{i,j} y_{ij} & s^2_y = (1/n)\sum_{i,j}(y_{ij} - \bar{y})^2 & n = n_1 + \cdots + n_k \end{array}$$

Introducimos a continuación el coeficiente

$$\hat{\eta}^2 = 1 - \frac{1}{n}\sum_{i=1}^{k} n_i \frac{s^2_{y_i}}{s^2_y} \tag{6.12}$$

que verifica $0 \leq \hat{\eta}^2 \leq 1$, y mide el grado de concentración de los puntos (x_i, y_{ij}) a lo largo de la curva $y = g(x)$ (ver figura 6.1).

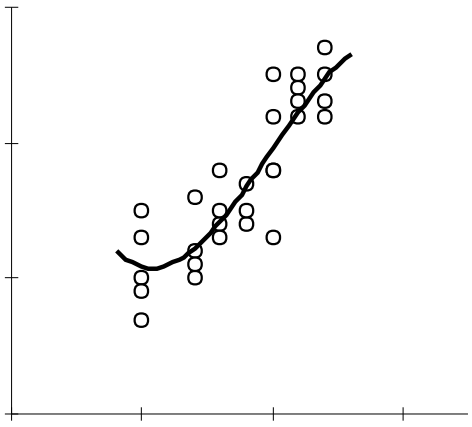


Figura 6.1: Curva que mejor se ajusta a los datos

Si indicamos $\delta_i = g(x_i)$ $i = 1, \ldots, k$ convertimos la hipótesis H_1 en una hipótesis lineal con k parámetros. Cuando H_1 es cierta, la estimación de δ_i es $\hat{\delta}_i = \bar{y}_i$. La identidad

$$\mathrm{SCR}_H = \mathrm{SCR} + (\mathrm{SCR}_H - \mathrm{SCR})$$

es entonces

$$\sum_{i,j}(y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i,j}(y_{ij} - \bar{y}_i)^2 + \sum_i n_i(\bar{y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Dividiendo por n tenemos

$$s^2_y(1 - r^2) = s^2_y(1 - \hat{\eta}^2) + s^2_y(\hat{\eta}^2 - r^2)$$

y el contraste para decidir si la regresión es lineal se resuelve a través del estadístico

$$F = \frac{(\hat{\eta}^2 - r^2)/(k-2)}{(1 - \hat{\eta}^2)/(n-k)} \tag{6.13}$$

que tiene $(k-2)$ y $(n-k)$ grados de libertad. Si F resulta significativa, rechazaremos el carácter lineal de la regresión.

Observaciones:

1) Solamente se puede aplicar este test si se tienen $n_i > 1$ observaciones de Y para cada x_i $(i = 1, \ldots, k)$.

2) $\hat{\eta}^2$ es una versión muestral de la llamada *razón de correlación* entre dos variables aleatorias X, Y

$$\eta^2 = \frac{E[(g(X) - E(Y))^2]}{\text{var}(Y)}$$

siendo

$$y = g(x) = E(Y|X = x)$$

la curva de regresión de la media de Y sobre X. Este coeficiente η^2 verifica:

a) $0 \leq \eta^2 \leq 1$

b) $\eta^2 = 0 \Longrightarrow y = E(Y)$ (la curva es la recta $y =$ constante).

c) $\eta^2 = 1 \Longrightarrow y = g(X)$ (Y es función de X)

3) Análogamente, podemos también plantear la hipótesis de que Y es alguna función (no lineal) de x frente a la hipótesis nula de que no hay ningún tipo de relación. Las hipótesis son:

$$H_0 : y_i = \mu + \epsilon_i$$
$$H_1 : y_i = g(x_i) + \epsilon_i$$

siendo μ constante. Entonces, con las mismas notaciones de antes,

$$\text{SCR}_H = \sum_{i,j} (y_{ij} - \bar{y})^2 \qquad \text{con } n - 1 \text{ g.l.}$$
$$\text{SCR} = \sum_{i,j} (y_{ij} - \bar{y}_i)^2 \qquad \text{con } n - k \text{ g.l.}$$

Operando, se llega al estadístico

$$F = \frac{\hat{\eta}^2/(k-1)}{(1 - \hat{\eta}^2)/(n-k)} \tag{6.14}$$

Comparando 6.14 con 6.10, podemos interpretar 6.14 como una prueba de significación de la razón de correlación.

Ejemplo 6.6.1

Se mide la luminosidad (en lúmenes) de un cierto tipo de lámparas después de un tiempo determinado de funcionamiento (en horas). Los resultados para una serie de $3, 2, 3, 2$ *y* 2 *lámparas fueron:*

Tiempo (x)	*Luminosidad (Y)*			
250	*5460*	*5475*	*5400*	$(n_1 = 3)$
500	*4800*	*4700*		$(n_2 = 2)$
750	*4580*	*4600*	*4520*	$(n_3 = 3)$
1000	*4320*	*4300*		$(n_4 = 2)$
1250	*4000*	*4010*		$(n_5 = 2)$

Con estos datos podemos ilustrar algunos aspectos de la regresión lineal de la luminosidad sobre el tiempo de funcionamiento.

- *Recta de regresión y coeficiente de correlación:*

$$\bar{x} = 708{,}33 \quad \bar{y} = 4680{,}42 \quad n{=}12$$
$$s_x = 351{,}09 \quad s_y = 500{,}08 \quad s_{xy} = -170190{,}97$$
$$r = -0{,}969 \qquad \hat{\beta}_1 = -1{,}381$$
$$y - 4680{,}42 = -1{,}381(x - 708{,}33)$$

La hipótesis $H_0 : \beta_1 = 0$ debe ser rechazada pues (ver 6.10) obtenemos $t = 12{,}403$ (10 g.l.) que es muy significativo.

- *Razón de correlación y carácter lineal de la regresión:*

$$\bar{y}_1 = 5445 \quad \bar{y}_2 = 4750 \quad \bar{y}_3 = 4566{,}7 \quad \bar{y}_4 = 4310 \quad \bar{y}_5 = 4005$$
$$s^2_{y_1} = 1050 \quad s^2_{y_2} = 2500 \quad s^2_{y_3} = 1155{,}5 \quad s^2_{y_4} = 100 \quad s^2_{y_5} = 25$$
$$\bar{y} = 4680{,}42 \qquad s^2_y = 250077 \qquad n = 12 \quad k = 5$$

$$\hat{\eta}^2 = 1 - \frac{1}{n}\sum_{i=1}^{k} n_i \frac{s^2_{y_i}}{s^2_y} = 0{,}996$$

Aplicando 6.13

$$F = \frac{(0{,}996 - 0{,}939)/3}{(1 - 0{,}996)/7} = 33{,}3$$

con 3 y 7 g.l. Se puede rechazar que la regresión es lineal.

Aplicando ahora 6.14

$$F = \frac{0{,}996/4}{(1 - 0{,}996)/7} = 435{,}7$$

vemos que la razón de correlación es muy significativa.

6.7. Comparación de rectas

En primer lugar, vamos a estudiar detalladamente la comparación de dos rectas, ya que en este caso las fórmulas son un poco más sencillas. A continuación presentaremos el caso general cuyos detalles pueden verse en Seber[65] pág. 197-205.

6.7.1. Dos rectas

Consideremos dos muestras independientes de tamaños n_1 y n_2

$$(x_{11}, y_{11}), (x_{12}, y_{12}), \ldots, (x_{1n_1}, y_{1n_1})$$
$$(x_{21}, y_{21}), (x_{22}, y_{22}), \ldots, (x_{2n_1}, y_{2n_1})$$

sobre la misma variable regresora x y la misma variable respuesta Y con distribución normal, pero para dos poblaciones distintas.

Los dos modelos de regresión simple para las dos poblaciones por separado son

$$y_{1i} = \alpha_1 + \beta_1 x_{1i} + \epsilon_{1i} \qquad i = 1, \ldots, n_1$$
$$y_{2i} = \alpha_2 + \beta_2 x_{2i} + \epsilon_{2i} \qquad i = 1, \ldots, n_2$$

y sus estimadores MC son

$$\hat{\alpha}_h = \bar{y}_h - \hat{\beta}_h \bar{x}_h \qquad \hat{\beta}_h = r_h \left(\frac{S_{yh}}{S_{xh}} \right)^{1/2} \qquad h = 1, 2$$

donde $\bar{x}_h, S_{xh}, \bar{y}_h, S_{yh}, r_h$ son las medias, sumas de cuadrados de las desviaciones y coeficiente de correlación para cada una de las muestras $h = 1, 2$ respectivamente.

También deberemos considerar $\bar{x}, S_x, \bar{y}, S_y, r$ las medias, sumas de cuadrados de las desviaciones y coeficiente de correlación de las dos muestras conjuntamente.

Vamos a considerar las dos regresiones simples como un único modelo lineal. Para ello hacemos

$$\mathbf{Y} = (y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2})'$$

y

$$\mathbf{X}\boldsymbol{\gamma} = \begin{pmatrix} 1 & 0 & x_{11} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_{1n_1} & 0 \\ 0 & 1 & 0 & x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_{2n_2} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

donde $\mathbf{X}$ es $(n_1 + n_2) \times 4$ de $\text{rg}(\mathbf{X}) = 4$.

Así pues, el modelo que presenta a las dos regresiones simples conjuntamente $\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ es un modelo lineal siempre que los errores verifiquen las condiciones de Gauss-Markov. Entonces es necesario suponer que las varianzas de los errores para las dos poblaciones son iguales $\sigma_1^2 = \sigma_2^2$.

Para este modelo lineal, las estimaciones MC de los parámetros $\alpha_1, \alpha_2, \beta_1, \beta_2$ coinciden con las estimaciones MC de las rectas por separado $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2$ y la suma de cuadrados residual es

$$\begin{aligned} \text{SCR} &= \sum_{i=1}^{n_1} (y_{1i} - \hat{\alpha}_1 - \hat{\beta}_1 x_{1i})^2 + \sum_{i=1}^{n_2} (y_{2i} - \hat{\alpha}_2 - \hat{\beta}_2 x_{2i})^2 \\ &= \text{SCR}_1 + \text{SCR}_2 = S_{y1}(1 - r_1^2) + S_{y2}(1 - r_2^2) \\ &= S_{y1} - \hat{\beta}_1^2 S_{x1} + S_{y2} - \hat{\beta}_2^2 S_{x2} \end{aligned} \tag{6.15}$$

Para contrastar la hipótesis de homogeneidad de varianzas $H_0 : \sigma_2^2 = \sigma_2^2$ podemos utilizar el estadístico

$$F = \frac{\text{SCR}_1/(n_1 - 2)}{\text{SCR}_2/(n_2 - 2)} \sim F_{n_1 - 2, n_2 - 2}$$

y la estimación de la varianza común es

$$\text{ECM} = \text{SCR}/(n_1 + n_2 - 4)$$

También se pueden utilizar los contrastes que se explican en la sección 6.7.3.

Test de coincidencia

Se trata de investigar si las dos rectas se pueden considerar iguales, es decir, vamos a contrastar la hipótesis

$$H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2$$

Ésta es una hipótesis lineal contrastable (el modelo es de rango máximo) del tipo $H_0 : \mathbf{A}\boldsymbol{\gamma} = \mathbf{0}$ con

$$\mathbf{A}\boldsymbol{\gamma} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

donde $\mathbf{A}$ es 2×4 y $q = \text{rg } \mathbf{A} = 2$. Luego podríamos utilizar las fórmulas obtenidas para el contraste. Sin embargo, en este caso es mucho más fácil calcular directamente la suma de cuadrados bajo la hipótesis.

Bajo H_0 la estimación MC de los parámetros comunes $\alpha = \alpha_1 = \alpha_2$ y $\beta = \beta_1 = \beta_2$ es sencillamente la que se obtiene del modelo lineal conjunto, es decir, una única recta de regresión con todos los datos juntos:

$$\alpha^* = \bar{y} - \beta^* \bar{x}$$
$$\beta^* = r \left(\frac{S_y}{S_x} \right)^{1/2}$$

Luego

$$\begin{aligned} \text{SCR}_H &= \sum_{i=1}^{n_1} (y_{1i} - \alpha^* - \beta^* x_{1i})^2 + \sum_{i=1}^{n_2} (y_{2i} - \alpha^* - \beta^* x_{2i})^2 \\ &= S_y(1 - r^2) \end{aligned}$$

De modo que el estadístico F es

$$F = \frac{(\text{SCR}_H - \text{SCR})/2}{\text{SCR}/(n_1 + n_2 - 4)} = \frac{(S_y(1 - r^2) - \text{SCR})/2}{\text{ECM}} \tag{6.16}$$

con distribución F_{2,n_1+n_2-4}, si H_0 es cierta.

Test de paralelismo

Ahora queremos comprobar la hipótesis

$$H_0' : \beta_1 = \beta_2$$

para la que $\mathbf{A}$ es 1×4 y $q = \text{rg } \mathbf{A} = 1$.

Bajo H_0', la estimación MC de los parámetros α_1, α_2 y $\beta = \beta_1 = \beta_2$ se obtiene de la minimización de

$$\xi = \sum_{i=1}^{n_1} (y_{1i} - \alpha_1 - \beta x_{1i})^2 + \sum_{i=1}^{n_2} (y_{2i} - \alpha_2 - \beta x_{2i})^2$$

Las derivadas parciales son

$$\frac{\partial \xi}{\partial \alpha_1} = \sum_{i=1}^{n_1} 2(y_{1i} - \alpha_1 - \beta x_{1i})(-1)$$
$$\frac{\partial \xi}{\partial \alpha_2} = \sum_{i=1}^{n_2} 2(y_{2i} - \alpha_2 - \beta x_{2i})(-1)$$
$$\frac{\partial \xi}{\partial \beta} = \sum_{i=1}^{n_1} 2(y_{1i} - \alpha_1 - \beta x_{1i})(-x_{1i}) + \sum_{i=1}^{n_2} 2(y_{2i} - \alpha_2 - \beta x_{2i})(-x_{2i})$$

Al igualar a cero, de las dos primeras ecuaciones tenemos

$$\tilde{\alpha}_1 = \bar{y}_1 - \tilde{\beta}\bar{x}_1 \qquad \tilde{\alpha}_2 = \bar{y}_2 - \tilde{\beta}\bar{x}_2$$

y si sustituimos en la tercera ecuación

$$\begin{aligned}\tilde{\beta} &= \frac{\sum_{i=1}^{n_1} x_{1i}(y_{1i} - \bar{y}_1) + \sum_{i=1}^{n_2} x_{2i}(y_{2i} - \bar{y}_2)}{\sum_{i=1}^{n_1} x_{1i}(x_{1i} - \bar{x}_1) + \sum_{i=1}^{n_2} x_{2i}(x_{2i} - \bar{x}_2)} \\ &= \frac{\sum_{h=1}^{2}\sum_{i=1}^{n_h}(x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h)}{\sum_{h=1}^{2}\sum_{i=1}^{n_h}(x_{hi} - \bar{x}_h)^2} \\ &= \frac{r_1(S_{x1}S_{y1})^{1/2} + r_2(S_{x2}S_{y2})^{1/2}}{S_{x1} + S_{x2}}\end{aligned}$$

De modo que la suma de cuadrados es

$$\begin{aligned}\mathrm{SCR}_{H'} &= \sum_{i=1}^{n_1}(y_{1i} - \tilde{\alpha}_1 - \tilde{\beta}x_{1i})^2 + \sum_{i=1}^{n_2}(y_{2i} - \tilde{\alpha}_2 - \tilde{\beta}x_{2i})^2 \\ &= \sum_{h=1}^{2}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h - \tilde{\beta}(x_{hi} - \bar{x}_h))^2 \\ &= \sum_{h=1}^{2}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2 - \tilde{\beta}^2\sum_{h=1}^{2}\sum_{i=1}^{n_h}(x_{hi} - \bar{x}_h)^2\end{aligned}$$

y el numerador del test F es

$$\mathrm{SCR}_{H'} - \mathrm{SCR} = \sum_{h=1}^{2}\hat{\beta}_h^2\sum_{i=1}^{n_h}(x_{hi} - \bar{x}_h)^2 - \tilde{\beta}^2\sum_{h=1}^{2}\sum_{i=1}^{n_h}(x_{hi} - \bar{x}_h)^2$$

Finalmente el estadístico F se puede escribir

$$F = \frac{\hat{\beta}_1^2 S_{x1} + \hat{\beta}_2^2 S_{x2} - \tilde{\beta}^2(S_{x1} + S_{x2})}{\mathrm{ECM}}$$

que bajo la hipótesis sigue una distribución F_{1,n_1+n_2-4}.

En la práctica, primero se realiza un test de paralelismo y, si se acepta, se realiza el test cuyo estadístico es

$$F = \frac{\mathrm{SCR}_{H'} - \mathrm{SCR}_H}{\mathrm{SCR}_H/(n_1 + n_2 - 3)}$$

Finalmente, y si este último ha sido no significativo, procederemos con el contraste de coincidencia.

Test de concurrencia

Se trata de comprobar la igualdad de los términos independientes de las dos rectas, es decir

$$H_0'' : \alpha_1 = \alpha_2$$

Como en el apartado anterior, se puede ver que el mínimo de la función

$$\xi^* = \sum_{i=1}^{n_1}(y_{1i} - \alpha - \beta_1 x_{1i})^2 + \sum_{i=1}^{n_2}(y_{2i} - \alpha - \beta_2 x_{2i})^2$$

se alcanza cuando

$$\breve{\alpha} = \left(n_1 + n_2 - \frac{x_{1\cdot}^2}{\sum_{i=1}^{n_1} x_{1i}^2} - \frac{x_{2\cdot}^2}{\sum_{i=1}^{n_2} x_{2i}^2}\right)^{-1} \left(y_{\cdot\cdot} - \frac{x_{1\cdot}\sum_{i=1}^{n_1} x_{1i}y_{1i}}{\sum_{i=1}^{n_1} x_{1i}^2} - \frac{x_{2\cdot}\sum_{i=1}^{n_2} x_{2i}y_{2i}}{\sum_{i=1}^{n_2} x_{2i}^2}\right)$$

$$\breve{\beta}_1 = \frac{\sum_{i=1}^{n_1}(y_{1i} - \breve{\alpha})x_{1i}}{\sum_{i=1}^{n_1} x_{1i}^2} \qquad \breve{\beta}_2 = \frac{\sum_{i=1}^{n_2}(y_{2i} - \breve{\alpha})x_{2i}}{\sum_{i=1}^{n_2} x_{2i}^2}$$

donde $y_{\cdot\cdot} = \sum_{h=1}^{2}\sum_{i=1}^{n_h} y_{hi}$, $x_{1\cdot} = \sum_{i=1}^{n_1} x_{1i}$ y $x_{2\cdot} = \sum_{i=1}^{n_2} x_{2i}$.
Con estos resultados se puede calcular la suma de cuadrados

$$\mathrm{SCR}_{H''} = \sum_{h=1}^{2}\sum_{i=1}^{n_h}(y_{hi} - \breve{\alpha} - \breve{\beta}_h x_{hi})^2$$

y el estadístico

$$F = \frac{\mathrm{SCR}_{H''} - \mathrm{SCR}}{\mathrm{ECM}}$$

que, bajo H_0'', sigue una distribución F_{1,n_1+n_2-4}.
El test que acabamos de ver contrasta la concurrencia de las dos rectas en $x = 0$. Si deseamos comprobar la concurrencia en un punto $x = c$, bastará aplicar este mismo test sustituyendo los datos x_{hi} por $x_{hi} - c$. Si lo que queremos es saber simplemente si las rectas se cortan (en algún punto), es suficiente con rechazar la hipótesis de paralelismo.

6.7.2. Varias rectas

Supongamos que tenemos la intención de comparar H rectas de regresión

$$Y = \alpha_h + \beta_h x_h + \epsilon \qquad h = 1, \ldots, H$$

donde $E(\epsilon) = 0$ y $\mathrm{var}(\epsilon) = \sigma^2$ es la misma para cada recta. Esta última condición es absolutamente imprescindible para poder aplicar los contrastes estudiados al modelo lineal conjunto que ahora describiremos.
Para cada h, consideremos los n_h pares (x_{hi}, y_{hi}) $i = 1, \ldots, n_h$ de modo que

$$y_{hi} = \alpha_h + \beta_h x_{hi} + \epsilon_{hi} \qquad i = 1, \ldots, n_h$$

con ϵ_{hi} independientes e idénticamente distribuidos como $N(0, \sigma^2)$.

Sea $\mathbf{Y} = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{H1}, \ldots, y_{Hn_2})'$ y

$$\mathbf{X\gamma} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{x}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{x}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_H \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_H \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_H \end{pmatrix}$$

donde $\mathbf{x}_h = (x_{h1}, \ldots, x_{hn_h})'$, para cada $h = 1, \ldots, H$.
Con todo ello disponemos del modelo lineal

$$\mathbf{Y} = \mathbf{X\gamma} + \boldsymbol{\epsilon}$$

donde $\mathbf{X}$ es $N \times 2H$, con $\mathrm{rg}(\mathbf{X}) = 2H$ y $N = \sum_{h=1}^{H} n_h$.
De esta forma podemos contrastar cualquier hipótesis lineal de la forma $H_0 : \mathbf{A\gamma} = \mathbf{c}$.
La estimación MC de los parámetros α_h, β_h de este modelo se obtiene de cada recta particular

$$\hat{\beta}_h = \frac{\sum_i (y_{hi} - \bar{y}_{h\cdot})(x_{hi} - \bar{x}_{h\cdot})}{\sum_i (x_{hi} - \bar{x}_{h\cdot})^2} = r_h \left(\frac{S_{yh}}{S_{xh}}\right)^{1/2}$$

$$\hat{\alpha}_h = \bar{y}_{h\cdot} - \hat{\beta}_h \bar{x}_{h\cdot}$$

donde $\bar{x}_{h\cdot}, S_{xh}, \bar{y}_{h\cdot}, S_{yh}, r_h$ son las medias, sumas de cuadrados de las desviaciones y coeficiente de correlación para cada una de las muestras $h = 1, \ldots, H$ respectivamente.
También la suma de cuadrados general SCR es simplemente la suma de las sumas de cuadrados de los residuos de cada recta de regresión por separado

$$\begin{aligned} \mathrm{SCR} &= \sum_{h=1}^{H} \left(\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_{h\cdot})^2 - \hat{\beta}_h^2 \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_{h\cdot})^2 \right) \\ &= \sum_{h=1}^{H} \mathrm{SCR}_h = \sum_{h=1}^{H} S_{yh}(1 - r_h^2) \\ &= \sum_{h=1}^{H} S_{yh} - \hat{\beta}_h^2 S_{xh} \end{aligned}$$

Test de coincidencia

Se trata de investigar si las rectas son iguales, es decir, si

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_H (= \alpha) \ ; \ \beta_1 = \beta_2 = \cdots = \beta_H (= \beta)$$

que podemos escribir matricialmente con una matriz $\mathbf{A}$ de tamaño $(2H - 2) \times 2H$ de rango $2H - 2$.

A partir de las estimaciones MC de los parámetros α, β que se obtienen de la recta ajustada con todos los puntos reunidos en una única muestra, la suma de cuadrados residual es

$$\begin{aligned}\mathrm{SCR}_H &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_{..} - \beta^*(x_{hi} - \bar{x}_{..}))^2 \\ &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_{..})^2 - (\beta^*)^2\sum_{h=1}^{H}\sum_{i=1}^{n_h}(x_{hi} - \bar{x}_{..})^2 \\ &= S_y(1 - r^2)\end{aligned}$$

donde

$$\beta^* = \frac{\sum_h \sum_i (y_{hi} - \bar{y}_{..})(x_{hi} - \bar{x}_{..})}{\sum_h \sum_i (x_{hi} - \bar{x}_{..})^2} = r\left(\frac{S_y}{S_x}\right)^{1/2}$$

y los estadísticos $\bar{x}_{..}, S_x, \bar{y}_{..}, S_y, r$ son las medias, sumas de cuadrados de las desviaciones y coeficiente de correlación de la muestra conjunta.
Entonces el estadístico F para el contraste de esta hipótesis es

$$F = \frac{(\mathrm{SCR}_H - \mathrm{SCR})/(2H-2)}{\mathrm{SCR}/(N-2H)} \tag{6.17}$$

Contraste de paralelismo

Ahora se trata de investigar si las pendientes de las rectas son iguales, es decir, si

$$H_0' : \beta_1 = \beta_2 = \cdots = \beta_H$$

que matricialmente es equivalente a

$$H_0' : \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & -1 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \mathbf{0}$$

En este caso, la matriz $\mathbf{A}$ que representa las restricciones de los parámetros es $(H-1)\times 2H$ y su rango es $H-1$. De modo que tomando, en el contraste F, los valores $q = H-1$, $n = N$ y $k = 2H$, el estadístico especificado para este contraste es

$$F = \frac{(\mathrm{SCR}_{H'} - \mathrm{SCR})/(H-1)}{\mathrm{SCR}/(N-2H)}$$

Para calcular el numerador de este estadístico podemos proceder con las fórmulas generales estudiadas u observar las peculiaridades de este modelo que permiten obtener $\mathrm{SCR}_{H'}$.
Primero hay que minimizar $\sum_h \sum_i (y_{hi} - \alpha_h - \beta x_{hi})^2$, de donde se obtienen los estimadores

$$\tilde{\alpha}_h = \bar{y}_{h\cdot} - \tilde{\beta}\bar{x}_{h\cdot} \qquad h = 1, \ldots, H$$

$$\begin{aligned}\tilde{\beta} &= \frac{\sum_h \sum_i x_{hi}(y_{hi} - \bar{y}_{h\cdot})}{\sum_h \sum_i x_{hi}(x_{hi} - \bar{x}_{h\cdot})} \\ &= \frac{\sum_h \sum_i (y_{hi} - \bar{y}_{h\cdot})(x_{hi} - \bar{x}_{h\cdot})}{\sum_h \sum_i (x_{hi} - \bar{x}_{h\cdot})^2} \\ &= \frac{\sum_h r_h (S_{xh} S_{yh})^{1/2}}{\sum_h S_{xh}}\end{aligned}$$

Este último estimador es un estimador conjunto (*pooled*) de la pendiente común.
Con estas estimaciones se procede a calcular la suma de cuadrados

$$\mathrm{SCR}_{H'} = \sum_{h=1}^{H} S_{yh} - \tilde{\beta}^2 \sum_{h=1}^{H} S_{xh}$$

y el estadístico F es

$$F = \frac{(\sum_h \hat{\beta}_h^2 S_{xh} - \tilde{\beta}^2 \sum_h S_{xh})/(H-1)}{\mathrm{SCR}/(N-2H)}$$

que bajo H_0' sigue una distribución $F_{H-1,N-2H}$.
En la práctica, es aconsejable comenzar por un contraste de paralelismo y, si se acepta, continuar con el contraste cuyo estadístico es

$$F = \frac{(\mathrm{SCR}_{H'} - \mathrm{SCR}_H)/(H-1)}{\mathrm{SCR}_H/(N-H-1)}$$

Finalmente, y si este último ha sido no significativo, procederemos con el contraste 6.17.

Test de concurrencia

Deseamos contrastar la hipótesis de que todas las rectas se cortan en un punto del eje de las Y, es decir, para $x = 0$:

$$H_0'' : \alpha_1 = \alpha_2 = \cdots = \alpha_H (= \alpha)$$

En este caso, las estimaciones de los parámetros bajo la hipótesis son

$$\breve{\alpha} = \left(N - \frac{x_{1\cdot}^2}{\sum_i x_{1i}^2} - \cdots - \frac{x_{H\cdot}^2}{\sum_i x_{Hi}^2} \right)^{-1} \left(y_{\cdot\cdot} - \frac{x_{1\cdot} \sum_i x_{1i} y_{1i}}{\sum_i x_{1i}^2} - \cdots - \frac{x_{H\cdot} \sum_i x_{Hi} y_{Hi}}{\sum_i x_{Hi}^2} - \right)$$

$$\breve{\beta}_h = \frac{\sum_i (y_{hi} - \breve{\alpha}) x_{hi}}{\sum_i x_{hi}^2} \qquad h = 1, 2, \ldots, H$$

donde $x_{h\cdot} = \sum_i x_{hi}$ y $y_{\cdot\cdot} = \sum_h \sum_i y_{hi}$.
La suma de cuadrados residual es

$$\mathrm{SCR}_{H''} = \sum_h \sum_i (y_{hi} - \breve{\alpha} - \breve{\beta}_h x_{hi})^2$$

y con ella se puede calcular el estadístico F para el contraste

$$F = \frac{(\mathrm{SCR}_{H''} - \mathrm{SCR})/(H-1)}{\mathrm{SCR}/(N-2H)}$$

Cuando los valores de las x son los mismos para todas las rectas, tenemos que $n_h = n$ y $x_{hi} = x_i$ para toda $h = 1, \ldots, H$ y así las fórmulas son más simples

$$\begin{aligned} \breve{\alpha} &= \left(Hn - \frac{Hx_{\cdot}^2}{\sum_i x_i^2} \right)^{-1} \left(y_{\cdot\cdot} - \frac{x_{\cdot} \sum_i x_i y_{\cdot i}}{\sum_i x_i^2} \right) \\ &= \bar{y}_{\cdot\cdot} - \frac{\bar{x} \sum_h \sum_i y_{hi}(x_i - \bar{x})}{H \sum_i (x_i - \bar{x})^2} = \bar{y}_{\cdot\cdot} - \bar{x} \frac{\sum_h \hat{\beta}_h}{H} \end{aligned}$$

donde cada $\hat{\beta}_h$ es la estimación de la pendiente de la h-ésima recta, mientras que $\breve{\alpha}$ es el corte de la recta de regresión *media*.

En este caso

$$\mathrm{SCR}_{H''} = \sum_h \sum_i y_{hi}^2 - \frac{\left(\sum_h \sum_i x_i y_{hi} \right)^2}{\sum_i x_i^2} - \breve{\alpha}^2 Hn \frac{\sum_i (x_i - \bar{x})^2}{\sum_i x_i^2}$$

Además, como $\bar{y}_{\cdot\cdot}$ y $\hat{\beta}_h$ están incorrelacionados

$$\begin{aligned} \mathrm{var}(\breve{\alpha}) &= \mathrm{var}(\bar{y}_{\cdot\cdot}) + H\bar{x}^2 \frac{\mathrm{var}(\hat{\beta}_h)}{H^2} \\ &= \frac{\sigma^2}{H} \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right) = \frac{\sigma^2 \sum_i x_i^2}{nH \sum_i (x_i - \bar{x})^2} \end{aligned}$$

de modo que tenemos la posibilidad de construir un intervalo de confianza para α ya que

$$(\breve{\alpha} - \alpha) \left(\frac{nH \sum_i (x_i - \bar{x})^2}{\mathrm{ECM} \sum_i x_i^2} \right)^{1/2} \sim t_{H(n-2)}$$

donde $\mathrm{ECM} = \mathrm{SCR}/(nH - 2H)$.

Por otra parte, también podemos estudiar si las rectas se cortan en un punto $x = c$ distinto del cero. Simplemente reemplazaremos x_{hi} por $x_{hi} - c$ en todas las fórmulas anteriores. La coordenada y del punto de corte sigue siendo estimada por $\breve{\alpha}$.

Sin embargo, si el punto de corte es desconocido $x = \phi$, la hipótesis a contrastar es mucho más complicada

$$H_0''' : \alpha_h + \beta_h \phi = \mathrm{cte.} = \bar{\alpha} + \bar{\beta}\phi \qquad h = 1, 2, \ldots, h$$

o también

$$H_0''' : \frac{\alpha_1 - \bar{\alpha}}{\beta_1 - \bar{\beta}} = \cdots = \frac{\alpha_H - \bar{\alpha}}{\beta_H - \bar{\beta}}$$

y desgraciadamente no es lineal.

6.7.3. Contraste para la igualdad de varianzas

En los contrastes de comparación de rectas se hace la suposición de la igualdad de las varianzas σ_h^2 de los modelos lineales simples $h = 1, \ldots, H$.

Los estimadores de dichas varianzas son los errores cuadráticos medios particulares

$$S_h^2 = \frac{\sum_i (y_{hi} - \bar{y}_{h\cdot} - \hat{\beta}_h (x_{hi} - \bar{x}_{h\cdot}))^2}{n_h - 2}$$

y sabemos que

$$(n_h - 2)S_h^2/\sigma_h^2 \sim \chi^2_{n_h-2} \qquad h = 1, \ldots, H \qquad \text{indep.}$$

Para contrastar la hipótesis

$$H_0 : \sigma_1^2 = \cdots = \sigma_H^2$$

hay varios métodos, desde los más clásicos de Bartlett(1937) o Hartley(1950), muy sensibles a la no normalidad de los datos, hasta los más robustos entre los que destaca el de Levene con sus variantes.

Si hacemos $f_h = n_h - 2$, el test de Bartlett es

$$T = \frac{(\sum f_h)\log S^2 - \sum(f_h \log S_h^2)}{C}$$

donde

$$S^2 = \frac{\sum f_h S_h^2}{\sum f_h} \qquad C = 1 + \frac{\sum f_h^{-1} - (\sum f_h)^{-1}}{3(H-1)}$$

Si H_0 es cierta, aproximadamente $T \sim \chi^2_{H-1}$.

Cuando los f_h son todos iguales, Hartley propone el estadístico

$$F = \frac{\text{máx}\{S_1^2, \ldots, S_H^2\}}{\text{mín}\{S_1^2, \ldots, S_H^2\}}$$

Sin embargo, como se trata de comparar las varianzas a partir de las observaciones o *réplicas* de H poblaciones, es mejor considerar el problema como un análisis de la varianza de un factor. La prueba robusta de Levene sobre la homogeneidad de varianzas se basa en el análisis de la varianza de un factor con los datos $z_{hi} = |y_{hi} - \bar{y}_{h\cdot}|$. Para reforzar la resistencia del método se puede utilizar como medida de localización la mediana.

Finalmente podemos añadir que, cuando la heterogeneidad de las varianzas es evidente, siempre es posible estudiar alguna transformación potencia de los datos originales y_{hi} que mejore la situación.

6.8. Un ejemplo para la reflexión

La siguiente tabla presenta cinco conjuntos de datos para cinco modelos de regresión simple diferentes: los datos bajo el encabezamiento x_1(a-d) son los valores de una variable regresora que es común en las cuatro regresiones con las variables respuesta y(a), y(b), y(c) y y(d). Las series de datos x(e) y y(e) definen otra regresión.

Se puede comprobar que, en los cinco casos, la regresión de y sobre x conduce exactamente a la misma recta

$$y = 0{,}520 + 0{,}809x$$

La varianza explicada, la no explicada i la varianza residual son idénticas en todas las regresiones, así como también el coeficiente de determinación.

Por lo tanto, las cinco regresiones parecen ser formalmente idénticas. A pesar de ello, si dibujamos en cada caso los diagramas de dispersión y la recta de regresión, observaremos que nuestra impresión se modifica radicalmente: en la página 116 tenemos los gráficos para los cinco conjuntos de datos.

obs.	x_1(a-d)	y(a)	y(b)	y(c)	y(d)	x(e)	y(e)
1	7	5,535	0,103	7,399	3,864	13,715	5,654
2	8	9,942	3,770	8,546	4,942	13,715	7,072
3	9	4,249	7,426	8,468	7,504	13,715	8,496
4	10	8,656	8,792	9,616	8,581	13,715	9,909
5	12	10,737	12,688	10,685	12,221	13,715	9,909
6	13	15,144	12,889	10,607	8,842	13,715	9,909
7	14	13,939	14,253	10,529	9,919	13,715	11,327
8	14	9,450	16,545	11,754	15,860	13,715	11,327
9	15	7,124	15,620	11,676	13,967	13,715	12,746
10	17	13,693	17,206	12,745	19,092	13,715	12,746
11	18	18,100	16,281	13,893	17,198	13,715	12,746
12	19	11,285	17,647	12,590	12,334	13,715	14,164
13	19	21,385	14,211	15,040	19,761	13,715	15,582
14	20	15,692	15,577	13,737	16,382	13,715	15,582
15	21	18,977	14,652	14,884	18,945	13,715	17,001
16	23	17,690	13,947	29,431	12,187	33,281	27,435

Cuadro 6.4: Datos de cinco regresiones simples

número de obs.	$n = 16$	$\hat{\beta}_1 = 0{,}809$	ee($\hat{\beta}_1$)=0,170
media de las x_1	$\bar{x}_1 = 14{,}938$	$\hat{\beta}_0 = 0{,}520$	ee($\hat{\beta}_0$)=2,668
media de las y	$\bar{y} = 12{,}600$	$R^2 = 0{,}617$	

$\sum(y_i - \bar{y})^2 = 380{,}403$ con 15 g.l. $\sum(y_i - \hat{y}_i)^2 = 145{,}66$ con 14 g.l.

$\hat{\sigma} = 3{,}226$

Cuadro 6.5: Principales resultados de la regresión simple

- La figura **a** es la que representan todos los manuales que explican la regresión simple. El modelo de la regresión lineal simple parece correcto y adaptado a los datos que permite describir correctamente. El modelo parece válido.

- La figura **b** sugiere que el modelo lineal simple no está absolutamente adaptado a los datos que pretende describir. Más bien, la forma adecuada es la cuadrática con una débil variabilidad. El modelo lineal simple es incorrecto; en particular, las predicciones que él proporciona son sesgadas: subestimaciones para los valores próximos a la media de x y sobreestimaciones para el resto.

- La figura **c** sugiere todavía que el modelo lineal simple no se adapta a los datos, pero una única observación parece ser la causa. Por contra, las otras observaciones están bien alineadas pero respecto a otra recta de ecuación $y = 4{,}242 + 0{,}503x_1$. Hay pues, un dato verdaderamente sospechoso. La reacción natural del experimentador será la de investigar con detalle la razón de esta desviación. ¿No será un error de transcripción? ¿Hay alguna causa que justifique la desviación y que no tiene en cuenta el modelo lineal simple?

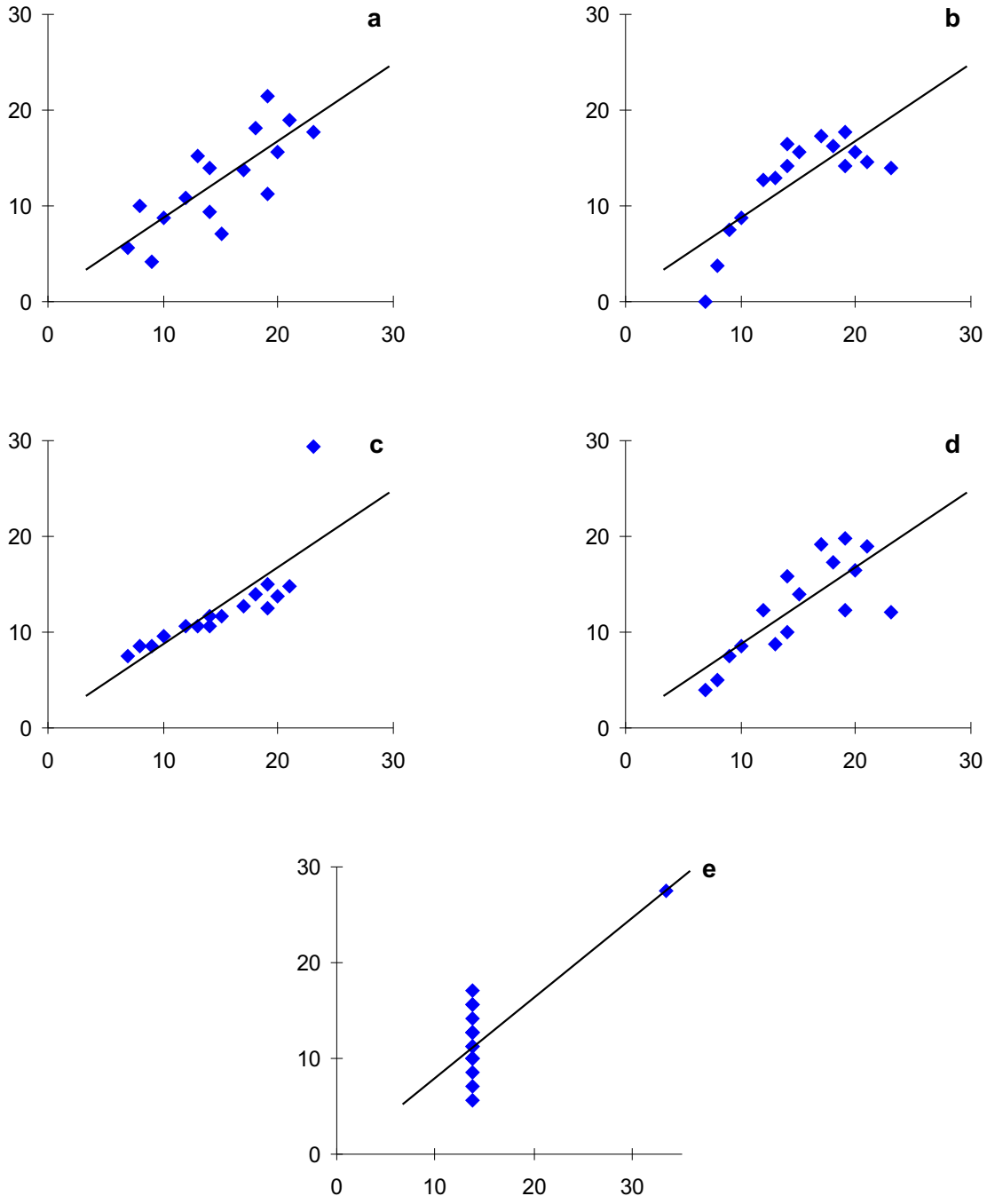


Figura 6.2: Gráficos de los cinco conjuntos de datos con la recta de regresión

- La figura **d** tiene un análisis más sutil: los puntos rodean la recta, pero aumentan las desviaciones a medida que crecen los valores de la variable regresora. Se hace evidente que la suposición de una varianza común de los residuos no se verifica.

- Finalmente, la figura **e** es más contundente: el modelo parece correcto. Si la calidad de los datos no puede ponerse en duda, este conjunto es tan válido como el primero y los resultados numéricos de la regresión son correctos. Pero nosotros intuimos que este resultado no es lo suficientemente satisfactorio: todo depende de la presencia

de un único punto, si lo suprimimos, incluso no será posible calcular la pendiente de la recta, ya que la suma de los cuadrados de las desviaciones de las x es cero. Éste no es el caso del primer conjunto de datos, donde la supresión de un punto no conduce más que a una ligera modificación de los resultados. Así pues, deberíamos ser extremadamente cautos con las posibles utilizaciones de este modelo. Además, debemos indicar que el experimento definido por los valores de x es muy malo.

Naturalmente, los conjuntos de datos **b**, **c**, **d** y **e** muestran casos extremos que, en la práctica, no hallaremos de forma tan clara. Ésta es una razón suplementaria para dotar al experimentador de medios para detectarlos. Cuando las desviaciones de las suposiciones del modelo son débiles, los resultados no serán erróneos, pero si las suposiciones son groseramente falsas, las conclusiones pueden incluso no tener sentido. La herramienta fundamental para la validación de las hipótesis del modelo es el análisis de los residuos del modelo estimado.

El **análisis de los residuos** (ver capítulo 9) tiene como objetivo contrastar a posteriori las hipótesis del modelo lineal. Es especialmente importante cuando, si tenemos un único valor de y para cada x, los contrastes de homocedasticidad, normalidad e independencia no se pueden hacer a priori. Analizaremos los residuos para comprobar:

a) si la distribución es aproximadamente normal;

b) si su variabilidad es constante y no depende de x o de otra causa;

c) si presentan evidencia de una relación no lineal entre las variables;

d) si existen observaciones atípicas o heterogéneas respecto a la variable x, la y o ambas.

6.9. Ejemplos con R

Vamos a recuperar el ejemplo de la sección 1.8 donde se calculan algunas regresiones a partir del ejemplo inicial con los datos de la tabla 1.1. En esa sección, el cálculo de la regresión simple se realiza con la función `lsfit(x,y)` que asignamos al objeto `recta.ls`

```
> recta.ls<-lsfit(dens,rvel)
```

Ahora utilizaremos la función `lm` que define el modelo de regresión simple.

```
> recta<-lm(rvel~dens)
> recta
Call:
lm(formula = rvel ~ dens)

Coefficients:
 (Intercept)         dens
    8.089813 -0.05662558

Degrees of freedom: 24 total; 22 residual
Residual standard error: 0.2689388
```

También se pueden obtener otros datos importantes con la función `summary`:

```
> recta.resumen<-summary(recta)
> recta.resumen

Call: lm(formula = rvel ~ dens)
Residuals:
     Min      1Q   Median     3Q    Max
 -0.3534 -0.2272 -0.03566 0.1894 0.5335

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   8.0898   0.1306    61.9295   0.0000
       dens  -0.0566   0.0022   -26.0076   0.0000

Residual standard error: 0.2689 on 22 degrees of freedom
Multiple R-Squared: 0.9685
F-statistic: 676.4 on 1 and 22 degrees of freedom, the p-value is 0

Correlation of Coefficients:
     (Intercept)
dens -0.9074
```

Además se puede acceder a muchos valores de los objetos `recta` y `recta.resumen` de forma directa.

```
> recta$coef
 (Intercept)        dens
    8.089813 -0.05662558
> recta.resumen$sigma
[1] 0.2689388
```

En general, podemos saber los diferentes resultados que se obtienen con el comando `lm` si escribimos `names(recta)` o `names(summary(recta))`.

```
> names(recta)
 [1] "coefficients" "residuals" "fitted.values" "effects" "R" "rank"
 [7] "assign" "df.residual" "contrasts" "terms" "call"
> names(summary(recta))
 [1] "call" "terms" "residuals" "coefficients" "sigma" "df"
 [7] "r.squared" "fstatistic" "cov.unscaled" "correlation"
```

De modo que podemos utilizar estos datos para nuevos cálculos. Por ejemplo podemos calcular la matriz estimada de covarianzas entre los estimadores de los parámetros $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ así:

```
> cov.beta<-round(recta.resumen$sigma^2*recta.resumen$cov.unscaled,6)
> cov.beta
            (Intercept)      dens
(Intercept)    0.017064 -0.000258
       dens   -0.000258  0.000005
```

Por otra parte, y aunque el resumen proporcionado por la función `summary(recta)` incluye el test F de significación de la regresión, la tabla del Análisis de la Varianza se puede calcular con la función `aov`.

```
> summary(aov(recta))
            Df Sum of Sq  Mean Sq  F Value Pr(F)
     dens    1  48.92231 48.92231 676.3944     0
Residuals 22   1.59122  0.07233
```

También se pueden calcular intervalos de confianza al 95 % para los parámetros β_0, β_1.

```
> coef(recta)
 (Intercept)        dens
    8.089813 -0.05662558
> coef.recta<-coef(recta)
> names(coef.recta)
[1] "(Intercept)" "dens"
> names(coef.recta)<-NULL # Truco para utilizar mejor los coeficientes
> coef.recta
        1           2
 8.089813 -0.05662558
> ee0<-sqrt(cov.beta[1,1])
> ee1<-sqrt(cov.beta[2,2])
> c(coef.recta[1]+qt(0.025,22)*ee0,coef.recta[1]+qt(0.975,22)*ee0)
[1] 7.818905 8.360721
> c(coef.recta[2]+qt(0.025,22)*ee1,coef.recta[2]+qt(0.975,22)*ee1)
[1] -0.06126290 -0.05198826
```

Cabe señalar que si el modelo de regresión simple debe pasar por el origen, es decir, no tiene término de intercepción, podemos utilizar la función `lsfit(x,y,int=F)` o la función `lm(y ~ x - 1)`.

La predicción puntual o por intervalo de nuevos valores de la variable respuesta se puede hacer con la función `predict` del modelo lineal. Atención, porque los argumentos en R y S-PLUS difieren.

Por último, podemos añadir que en R existe un conjunto de datos similares a los explicados en la sección 6.8:

```
> data(anscombe)
> summary(anscombe)
```

6.10. Ejercicios

Ejercicio 6.1

Probar que bajo el modelo lineal normal $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ las estimaciones MC $\hat{\beta}_0, \hat{\beta}_1$ son estocásticamente independientes si y sólo si $\sum x_i = 0$.

Ejercicio 6.2

Comprobar que la pendiente de la recta de regresión es

$$\hat{\beta}_1 = r\frac{S_y^{1/2}}{S_x^{1/2}} = r\frac{s_y}{s_x}$$

donde r es el coeficiente de correlación

$$r = \frac{S_{xy}}{(S_x S_y)^{1/2}} = \frac{s_{xy}}{s_x s_y}$$

Ejercicio 6.3

Consideremos el modelo de regresión simple alternativo

$$y_i = \gamma_0 + \gamma_1(x_i - \bar{x}) + \epsilon_i \qquad i = 1, \ldots, n$$

La matriz de diseño asociada es $\mathbf{X}_* = (\mathbf{1}, \mathbf{x} - \bar{x}\mathbf{1})$ donde $\mathbf{1} = (1, \ldots, 1)'$ y $\mathbf{x} = (x_1, \ldots, x_n)'$. Este modelo es equivalente al modelo 6.1 ya que $\langle \mathbf{X}_* \rangle = \langle \mathbf{X} \rangle$.
Calcular las estimaciones $\widehat{\boldsymbol{\gamma}} = (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{Y}$ para comprobar que

$$\hat{\gamma}_0 = \bar{y}$$
$$\hat{\gamma}_1 = \hat{\beta}_1 = \sum \frac{x_i - \bar{x}}{S_x}\, y_i$$

Calcular la matriz de varianzas-covarianzas $\text{var}(\widehat{\boldsymbol{\gamma}}) = \sigma^2(\mathbf{X}_*'\mathbf{X}_*)^{-1}$ y comprobar que $\hat{\gamma}_0 = \bar{y}$ está incorrelacionado con $\hat{\gamma}_1 = \hat{\beta}_1$. A partir de este resultado, calcular $\text{var}(\hat{\beta}_1) = \text{var}(\hat{\gamma}_1)$ y $\text{var}(\hat{\beta}_0) = \text{var}(\bar{y} - \hat{\beta}_1\bar{x})$.
Calcular también la matriz proyección $\mathbf{P} = \mathbf{X}_*(\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Ejercicio 6.4

En un modelo de regresión simple, con β_0, demostrar que se verifican las siguientes propiedades para las predicciones $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ y los residuos $e_i = y_i - \hat{y}_i$:

(i) La suma de los residuos es cero: $\sum e_i = 0$.

(ii) $\sum y_i = \sum \hat{y}_i$

(iii) La suma de los residuos ponderada por los valores de la variable regresora es cero: $\sum x_i e_i = 0$.

(iv) La suma de los residuos ponderada por las predicciones de los valores observados es cero: $\sum \hat{y}_i e_i = 0$.

Ejercicio 6.5 Modelo de regresión simple estandarizado

A partir de los datos observados de una variable respuesta y_i y de una variable regresora x_i se definen unas nuevas variables estandarizadas como

$$u_i = \frac{x_i - \bar{x}}{S_x^{1/2}} \qquad v_i = \frac{y_i - \bar{y}}{S_y^{1/2}} \qquad i = 1, \ldots, n$$

La estandarización significa que los datos transformados están centrados y los vectores $\mathbf{u} = (u_1, \ldots, u_n)'$, $\mathbf{v} = (v_1, \ldots, v_n)'$ son de longitud uno, es decir, $||\mathbf{u}|| = 1$ y $||\mathbf{v}|| = 1$.
Se define el modelo de regresión simple estandarizado como

$$v_i = b_1 u_i + \epsilon_i \qquad i = 1, \ldots, n$$

En este modelo desaparece de manera natural la ordenada en el origen al realizar el cambio de variables.
Comprobar que

$$\hat{\beta}_1 = \hat{b}_1 \sqrt{\frac{S_y}{S_x}}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Además, la "matriz" $\mathbf{u}'\mathbf{u} = ||\mathbf{u}||^2 = 1$ y la estimación de b_1 es muy sencilla $\hat{b}_1 = r$.

Ejercicio 6.6

En el caso de una regresión lineal simple pasando por el origen y con la hipótesis de normalidad, escribir el contraste de la hipótesis $H_0 : \beta_1 = b_1$, donde b_1 es una constante conocida.

Ejercicio 6.7

Para el modelo lineal simple consideremos la hipótesis

$$H_0 : y_0 = \beta_0 + \beta_1 x_0$$

donde (x_0, y_0) es un punto dado. Esta hipótesis significa que la recta de regresión pasa por el punto (x_0, y_0). Construir un test para esta hipótesis.

Ejercicio 6.8

Hallar la recta de regresión simple de la variable respuesta *raíz cuadrada de la velocidad* sobre la variable regresora *densidad* con los datos de la tabla 1.1 del capítulo 1.

Comprobar las propiedades del ejercicio 6.4 para estos datos.

Calcular la estimación de σ^2 y, a partir de ella, las estimaciones de las desviaciones estándar de los estimadores de los parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$.

Escribir los intervalos de confianza para los parámetros con un nivel de confianza del 95 %.

Construir la tabla para la significación de la regresión y realizar dicho contraste.

Hallar el intervalo de la predicción de la respuesta media cuando la densidad es de 50 vehículos por km. Nivel de confianza: 90 %.

Ejercicio 6.9

Comparar las rectas de regresión de hombres y mujeres con los logaritmos de los datos del ejercicio 1.4.

Ejercicio 6.10

Se admite que una persona es *proporcionada* si su altura en cm es igual a su peso en kg más 100. En términos estadísticos si la recta de regresión de Y (altura) sobre X (peso) es

$$Y = 100 + X$$

Contrastar, con un nivel de significación $\alpha = 0{,}05$, si se puede considerar válida esta hipótesis a partir de los siguientes datos que corresponden a una muestra de mujeres jóvenes:

X :	55	52	65	54	46	60	54	52	56	65	52	53	60
Y :	164	164	173	163	157	168	171	158	169	172	168	160	172

Razonar la bondad de la regresión y todos los detalles del contraste.

Ejercicio 6.11

El período de oscilación de un péndulo es $2\pi\sqrt{\frac{l}{g}}$, donde l es la longitud y g es la constante de gravitación. En un experimento observamos t_{ij} $(j = 1, \ldots, n_i)$ períodos correspondientes a l_i $(i = 1, \ldots, k)$ longitudes.

(a) Proponer un modelo, con las hipótesis que se necesiten, para estimar la constante $\frac{2\pi}{\sqrt{g}}$ por el método de los mínimos cuadrados.

(b) En un experimento se observan los siguientes datos:

longitud	período
18,3	8,58 7,9 8,2 7,8
20	8,4 9,2
21,5	9,7 8,95 9,2
15	7,5 8

Contrastar la hipótesis $H_0 : \frac{2\pi}{\sqrt{g}} = 2$.

Capítulo 7

Una recta resistente

Para ajustar una linea recta de la forma

$$y = a + bx$$

a un conjunto de datos $(x_i, y_i), i = 1, \ldots, n$ se han desarrollado varios métodos a lo largo de la historia. La regresión por mínimos cuadrados que hemos explicado es el método más conocido y más ampliamente utilizado. Es un método que involucra cálculos algebraicamente simples, utiliza la inferencia deducida para la distribución normal y requiere únicamente una derivación matemática sencilla. Desgraciadamente, la recta de regresión mínimo-cuadrática no es resistente. Un dato "salvaje" puede tomar fácilmente el control de la recta ajustada y conducirnos a conclusiones engañosas sobre la relación entre y y x. La llamada *recta resistente de los tres grupos* evita esta dificultad. Así, esta recta es muy útil en la exploración de los datos y-versus-x.

A continuación exponemos las principales ideas en este tema del clásico libro *Understanding Robust and Exploratory Data Analysis* de Hoaglin, Mosteller y Tukey [39].

7.1. Recta resistente de los tres grupos

7.1.1. Formación de los tres grupos

Empezaremos por ordenar los valores x de manera que $x_1 \leq x_2 \leq \cdots \leq x_n$. Entonces, sobre la base de estos valores ordenados, dividiremos los n puntos (x_i, y_i) en tres grupos: un grupo izquierdo, un grupo central y un grupo derecho, de tamaño tan igual como sea posible. Cuando no hay repeticiones en les x_i, el número de puntos en cada uno de los tres grupos depende del residuo de la división de n por 3:

Grupo	$n = 3k$	$n = 3k + 1$	$n = 3k + 2$
Izquierdo	k	k	$k + 1$
Central	k	$k + 1$	k
Derecho	k	k	$k + 1$

Repeticiones de los x_i nos harán estar alerta para formar tres conjuntos que no separen los puntos con igual x en conjuntos diferentes. Un examen detallado del tratamiento de las repeticiones nos puede llevar incluso a formar únicamente dos grupos. Cuando cada uno de los tercios ha sido definitivamente formado, determinaremos las dos coordenadas de unos puntos centrales, uno para cada grupo, con la mediana de los valores de las x y

la mediana de los valores de las y, por separado. Etiquetaremos las coordenadas de estos tres puntos centrales con las letras I de izquierda, C de centro i D de derecha:

$$(x_I, y_I), (x_C, y_C), (x_D, y_D)$$

La figura 7.1 muestra los puntos observados y los puntos centrales de un ejemplo hipotético con 9 puntos. Como se ve en este gráfico, ninguno de los puntos centrales coincide con un punto de los datos, ya que las medianas de les x y de las y se han calculado separadamente. A pesar de ello, los tres podrían ser puntos observados, como ocurre a menudo, cuando las x y las y siguen el mismo orden.

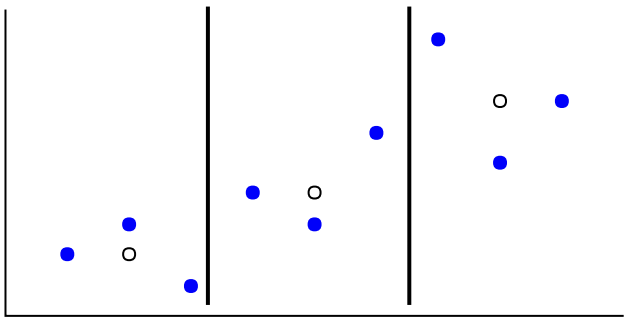


Figura 7.1: Puntos observados y puntos centrales en un ejemplo hipotético.

Este sistema de determinación de los puntos centrales de cada grupo es el que da a la recta que calcularemos su resistencia. Cuanto mayor es el número de puntos observados en cada grupo, la mediana proporciona la resistencia a los valores influyentes de x, y o ambos.

7.1.2. Pendiente e intercepción

Ahora utilizaremos los puntos centrales para calcular la pendiente b y la ordenada en el origen o intercepción a de la recta $y = a + bx$ que ajusta los valores observados y permite la predicción de los valores x_i observados y cualquier otro valor apropiado de x. En este sentido, la pendiente b nos dice cuantas unidades de y cambian por una unidad de x. Es razonable obtener esta información de los datos, en concreto de los puntos centrales de los grupos izquierdo y derecho:

$$b_0 = \frac{y_D - y_I}{x_D - x_I}$$

La utilización de los dos puntos centrales de los grupos extremos nos da la ventaja de medir el cambio de y sobre un intervalo bastante ancho de x, siempre que hayan suficientes puntos observados en estos grupos para asegurar la resistencia.

Cuando tomamos la pendiente b_0 para ajustar el valor y de cada punto central, la diferencia es el valor de la intercepción de una linea con pendiente b_0 que pasa exactamente por este punto. La intercepción ajustada es la media de estos tres valores:

$$a_0 = \frac{1}{3}[(y_I - b_0 x_I) + (y_C - b_0 x_C) + (y_D - b_0 x_D)]$$

De nuevo, como los puntos centrales están basados en la mediana, a_0 es resistente.

El ajuste de una recta en términos de pendiente e intercepción es convencional, pero usualmente artificial. La intercepción, que da el valor de y cuando $x = 0$, puede ser determinada de forma imprecisa, especialmente cuando los valores de x están todos muy alejados del cero y el cero es un valor sin sentido en el rango de las x. Ajustar la recta en términos de pendiente y un valor central de las x, como la media, la mediana o x_C, es mucho más útil. Nosotros escogeremos x_C por conveniencia y entonces la recta inicial es

$$y = a_0^* + b_0(x - x_C)$$

donde b_0 es la de antes y el valor central (también llamado *nivel*) es

$$a_0^* = \frac{1}{3}[(y_I - b_0(x_I - x_C)) + y_C + (y_D - b_0(x_D - x_C))]$$

Como ahora explicaremos, esta recta se toma como punto de partida para ajustar una mejor con iteraciones sucesivas.

7.1.3. Ajuste de los residuos e iteraciones

Una vez que hemos obtenido la pendiente y el nivel de la recta ajustada, el siguiente paso es calcular los residuos para cada punto

$$r_i = y_i - [a^* + b(x_i - x_C)]$$

Los gráficos de los residuos son muy útiles en la evaluación del ajuste y para descubrir patrones de comportamiento inesperados. Pero ahora, de momento, resaltaremos una propiedad general de todo conjunto de residuos, en nuestro problema actual o en situaciones más complejas:

> Si substituimos los valores originales de y por los residuos, es decir, si utilizamos (x_i, r_i) en lugar de (x_i, y_i), $i = 1, \ldots, n$ y repetimos el proceso de ajuste, llegaremos a un ajuste cero.

Para una linea recta esto significa que, con los puntos $(x_i, r_i), i = 1, \ldots, n$ como datos, obtendremos una pendiente cero y un nivel cero. En otras palabras, los residuos no contienen más aportación a la recta ajustada.

Una importante característica de los procedimientos resistentes es que habitualmente requieren iteraciones. Es el caso de la recta resistente de los tres grupos. Los residuos de la recta con la pendiente b_0 y el nivel a_0^* no tienen pendiente y nivel cero cuando hacemos el ajuste de la recta con las mismas x_i, aunque los nuevos valores de pendiente y nivel son substancialmente menores (en magnitud) que b_0 y a_0^*. Por esta razón, pensaremos en b_0 y a_0^* como los valores iniciales de una iteración.

El ajuste a una recta de los residuos obtenidos con la recta inicial da unos valores δ_1 y γ_1 a la pendiente y el nivel, respectivamente. En concreto, utilizaremos los residuos iniciales

$$r_i^{(0)} = y_i - [a_0^* + b_0(x_i - x_C)], \qquad i = 1, \ldots, n$$

en lugar de los y_i y repetiremos los pasos del proceso de ajuste. Como el conjunto de los x_i no ha cambiado, los tres grupos y las medianas de los x en los puntos centrales serán los mismos.

Cuadro 7.1: Edad y altura de unos niños en una escuela privada.

Niño	Edad (meses)	Altura (cm)
1	109	137,6
2	113	147,8
3	115	136,8
4	116	140,7
5	119	132,7
6	120	145,4
7	121	135,0
8	124	133,0
9	126	148,5
10	129	148,3
11	130	147,5
12	133	148,8
13	134	133,2
14	135	148,7
15	137	152,0
16	139	150,6
17	141	165,3
18	142	149,9

Fuente: B.G. Greenberg (1953). "The use of analysis of covariance and balancing in analytical studies", American Journal of Public Health, 43, 692-699 (datos de la tabla 1, pág. 694).

La pendiente y el nivel ajustados son $b_0 + \delta_1$ y $a_0^* + \gamma_1$ y los nuevos residuos

$$r_i^{(1)} = r_i^{(0)} - [\gamma_1 + \delta_1(x_i - x_C)], \qquad i = 1, \ldots, n$$

Ahora podemos avanzar con otra iteración. En general no sabremos si hemos conseguido un conjunto apropiado de residuos, hasta que verifiquemos el ajuste cero. En la práctica continuaremos las iteraciones hasta que el ajuste de la pendiente sea suficientemente pequeño en magnitud, del orden del 1 % o del 0,01 % del tamaño de b_0. Cada iteración añade su pendiente y su nivel a los valores previos

$$b_1 = b_0 + \delta_1, b_2 = b_1 + \delta_2, \ldots$$

y

$$a_1^* = a_0^* + \gamma_1, a_2^* = a_1^* + \gamma_2, \ldots$$

Las iteraciones son normalmente pocas y los cálculos no muy largos.

Ejemplo 7.1.1

En una discusión en 1953, Greenberg consideró los datos de edad y altura de dos muestras de niños, una de una escuela privada urbana y la otra de una escuela pública rural. En la tabla 7.1 se reproducen los datos de los 18 niños de la escuela privada.

Aunque los datos no siguen claramente una linea recta, su patrón no es notablemente curvado y el ajuste a una linea puede resumir cómo la altura y crece con la edad x en

este grupo de niños. Sólo los niños 13 *y* 17 *tienen puntos muy separados y veremos cómo influyen en el conjunto. Dado que* 18 *es divisible por* 3 *y los datos* x *no tienen repeticiones, cada grupo contiene seis puntos. Los puntos centrales de cada grupo son*

$$(x_I, y_I) = (115{,}50, 139{,}15)$$
$$(x_C, y_C) = (127{,}50, 147{,}90)$$
$$(x_D, y_D) = (138{,}00, 150{,}25)$$

de forma que el valor inicial de la pendiente es

$$b_0 = \frac{150{,}25 - 139{,}15}{138{,}00 - 115{,}50} = 0{,}4933$$

y el valor inicial del nivel

$$a_0^* = \frac{1}{3}[(139{,}15 - 0{,}4933(115{,}5 - 127{,}5)) + 147{,}9 + (150{,}25 - 0{,}4933(138 - 127{,}5))] = 146{,}0133$$

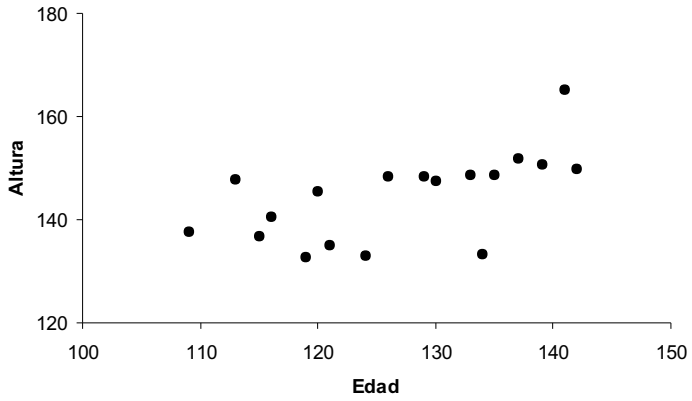


Figura 7.2: Altura versus edad para los niños de una escuela privada.

Los datos de la tabla 7.2 están ya ordenados en función de los valores de x = *Edad y se han calculado los residuos de la recta inicial.*

Para ver cómo van las iteraciones, calcularemos los primeros ajustes de la pendiente y del nivel

$$\delta_1 = \frac{-1{,}0500 - 0{,}5367}{138{,}00 - 115{,}50} = -0{,}0705$$
$$\gamma_1 = -0{,}1519$$

Notemos que δ_1 *es sustancialmente menor en magnitud que* b_0*, pero todavía no es negligible. Dos iteraciones más nos proporcionan unos valores para los que el proceso puede parar:* $\delta_3 = -0{,}0006$ *es menor que un* 1 % *de la pendiente acumulada.*

La recta ajustada es

$$y = 145{,}8643 + 0{,}4285(x - 127{,}5)$$

La figura 7.3 representa los residuos de este ajuste. En general, el aspecto global es bastante satisfactorio. Sólo los dos puntos destacados, el del niño 13 *y el del niño* 17*, se separan mucho y son atípicos. También hay tres residuos demasiado negativos para niños*

Cuadro 7.2: Edad y altura de los niños en los tres grupos y residuos de la recta inicial

Niño	Edad (meses)	Altura (cm)	Residuo
1	109	137,6	0,7133
2	113	147,8	8,9400
3	115	136,8	−3,0467
4	116	140,7	0,3600
5	119	132,7	−9,1200
6	120	145,4	3,0867
7	121	135,0	−7,8067
8	124	133,0	−11,2867
9	126	148,5	3,2267
10	129	148,3	1,5467
11	130	147,5	0,2533
12	133	148,8	0,0733
13	134	133,2	−16,0200
14	135	148,7	−1,0133
15	137	152,0	1,3000
16	139	150,6	−1,0867
17	141	165,3	12,6267
18	142	149,9	−3,2667

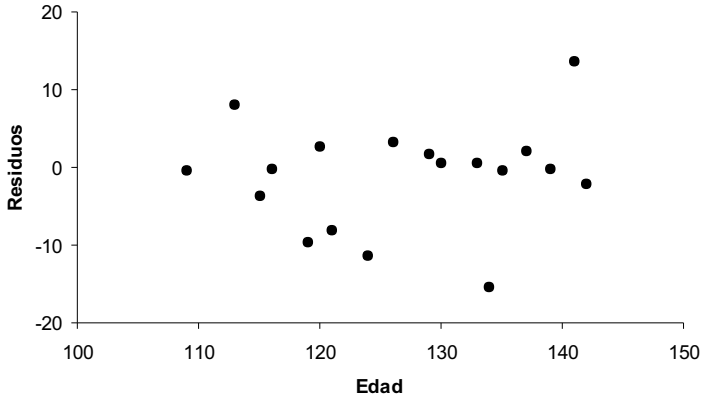


Figura 7.3: Residuos de la altura versus edad, después del ajuste por la recta resistente.

que tienen alrededor de 120 meses. Si tuviéramos más información, podríamos estudiar porqué estos niños son demasiado altos o demasiado bajos para su edad. Por ejemplo, podríamos separar los niños de las niñas.

En este ejemplo hemos visto cómo dos puntos, hasta cierto punto inusuales, han tenido muy poco efecto, si han tenido alguno, en el ajuste general de los datos. Una recta ajustada por el método de los mínimos cuadrados corre mucho más riesgo de dejarse influenciar por estos puntos. Para estos datos la recta de regresión mínimo-cuadrática es

$$y = 79{,}6962 + 0{,}5113x$$

o

$$y = 144{,}8853 + 0{,}5113(x - 127{,}5)$$

donde observamos cómo los puntos $5, 7, 8$ *y* 17 *han torcido la recta. Además, si el valor de* y *del punto* 13 *no fuera tan bajo, la recta mínimo-cuadrática podría ser más empinada. En todo caso, como la evaluación del ajuste se hace con los residuos, la figura 7.4 nos muestra los residuos mínimo-cuadráticos con la edad. Aunque es bastante similar al anterior, este gráfico nos da la sensación de una ligera tendencia a la baja. Es decir, los residuos mínimo-cuadráticos resultarían más horizontales si elimin——áramos de ellos una recta con una pendiente ligeramente negativa.*

En este ejemplo la variabilidad de los residuos merece más atención que la diferencia entre las pendientes de la recta de regresión mínimo-cuadrática y la recta resistente. Por ejemplo, la desviación estándar de los residuos mínimo-cuadráticos es $6{,}8188$ *y el error estándar de la pendiente es* $0{,}1621$*, sobre dos veces la diferencia entre las pendientes.*

Así hemos visto, cualitativamente, cómo algunos datos pueden afectar a la recta mínimo-cuadrática mucho más que la recta resistente. En todo caso, cuando los datos están razonablemente bien dispuestos las dos líneas son parecidas.

7.1.4. Mejora del método de ajuste

Para algunos conjuntos de datos, el procedimiento iterativo explicado para ajustar la recta resistente encuentra dificultades. Los ajustes de la pendiente pueden decrecer muy lentamente o, después de unos pocos pasos, dejar de decrecer y oscilar entre dos valores.

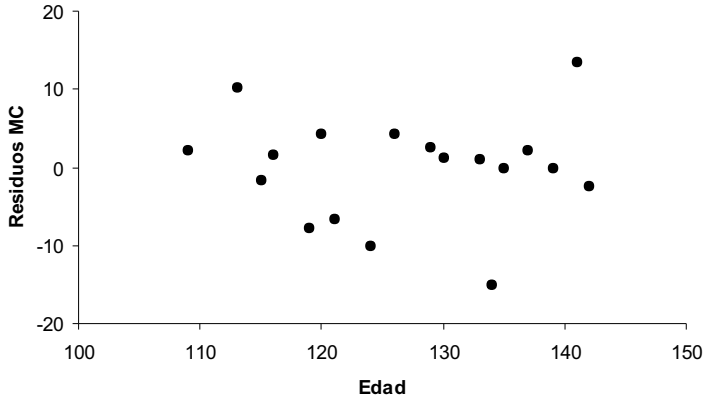


Figura 7.4: Residuos mínimo-cuadráticos versus edad.

Afortunadamente, una modificación elimina completamente estos problemas y permite que el número de iteraciones sea mucho más limitado.

La solución propuesta por Johnstone y Velleman (1982) es un procedimiento iterativo para el cálculo de la pendiente que asegura la convergencia hacia un valor único.

En el cálculo de la pendiente en la $j+1$ iteración tenemos

$$\delta_{j+1} = \frac{r_D^{(j)} - r_I^{(j)}}{x_D - x_I}$$

y esto será 0 justamente cuando el numerador $r_D^{(j)} - r_I^{(j)} = 0$. Es decir, lo que debemos hacer es hallar el valor de b que proporciona la misma mediana a los residuos del grupo derecho y del grupo izquierdo. Más formalmente

$$\Delta r(b) = r_D(b) - r_I(b)$$

muestra la dependencia funcional de b y prescinde del número de la iteración. Buscamos el valor de b que hace $\Delta r(b) = 0$. Notemos que centraremos el proceso iterativo en b y dejaremos a para el final.

Empezaremos por calcular b_0 como antes y calcularemos $\Delta r(b_0)$ y δ_1 como ya sabemos. A continuación calcularemos $\Delta r(b_0 + \delta_1)$. Generalmente, $\Delta r(b_0)$ y $\Delta r(b_0 + \delta_1)$ tendrán signos opuestos, indicando que el valor deseado de b cae entre b_0 y $b_1 = b_0 + \delta_1$. Si pasa lo contrario, cuando $\Delta r(b_0)$ y $\Delta r(b_0 + \delta_1)$ tienen el mismo signo, hace falta seguir los pasos desde b_0 y $b_1 = b_0 + \delta_1$ hasta que hallamos un b_1 tal que $\Delta r(b_1)$ tiene el signo contrario a $\Delta r(b_0)$.

En este punto tenemos un b_0 con $\Delta r(b_0)$ y un b_1 con $\Delta r(b_1)$ y sabemos que Δr ha de ser 0 para algún valor b entre b_0 y b_1. (Este hecho y que la solución es única requieren una demostración formal que aquí no reproducimos.) Así que podemos continuar por interpolación lineal

$$b_2 = b_1 - \Delta r(b_1)\frac{b_1 - b_0}{\Delta r(b_1) - \Delta r(b_0)}$$

Cuando $\Delta r(b_2)$ no es todavía 0 (o suficientemente cerca de cero), hace falta repetir la interpolación con otro paso. Para hacer esto, consideraremos el intervalo que contiene b utilizando b_2 en lugar de b_1 o de b_0, el que tenga Δr con el mismo signo que $\Delta r(b_2)$. Y así los pasos necesarios.

7.2. Métodos que dividen los datos en grupos

Otras técnicas anteriores al método resistente de los tres grupos fueron propuestas e involucran la división de los datos en grupos. Algunos de estos métodos no pretenden ser una alternativa al método de los mínimos cuadrados y fueron desarrollados para ajustar una recta "cuando ambas variables están sujetas a error".

Método de Wald

Wald (1940) propuso dividir los datos en dos grupos de igual tamaño. Idealmente, los valores teóricos X_i del primer grupo son menores que los del segundo. En la práctica, porque los valores de X_i son desconocidos, agruparemos los puntos en base a los x_i observados.

Supongamos que n es par y sea $m = n/2$. Entonces, si asumimos que los valores de x están ordenados en orden creciente, la pendiente propuesta es

$$b_W = \frac{(y_{m+1} + \cdots + y_n) - (y_1 + \cdots + y_m)}{(x_{m+1} + \cdots + x_n) - (x_1 + \cdots + x_m)}$$

Si $x_{m+1} = x_m$, el método descarta los puntos con repetición en el centro.

El punto de intercepción es

$$a_W = \bar{y} - b_W \bar{x}$$

donde $\bar{y}$ y $\bar{x}$ son las medias totales, de la misma forma que en la recta mínimo-cuadrática.

Método de Nair y Shrivastava

Como una alternativa computacionalmente atractiva respecto al método de los mínimos cuadrados, Nair y Shrivastava (1942) introdujeron el método de las medias por grupo. Si ordenamos las x, podemos considerar un primer grupo con n_I puntos, un segundo grupo con n_D puntos y descartamos los $n - n_I - n_D$ restantes. Los puntos resumen de cada grupo son las medias

$$\bar{x}_I = \frac{x_1 + \cdots + x_{n_I}}{n_I} \qquad \bar{y}_I = \frac{y_1 + \cdots + y_{n_I}}{n_I}$$
$$\bar{x}_D = \frac{x_{n-n_D+1} + \cdots + x_n}{n_D} \quad \bar{y}_D = \frac{y_{n-n_D+1} + \cdots + y_n}{n_D}$$

y la pendiente y el punto de intercepción resultan de la recta que pasa por $(\bar{x}_I, \bar{y}_I)$ y $(\bar{x}_D, \bar{y}_D)$

$$b_{NS} = \frac{\bar{y}_D - \bar{y}_I}{\bar{x}_D - \bar{x}_I}$$
$$a_{NS} = \bar{y}_I - b_{NS}\bar{x}_I = \bar{y}_D - b_{NS}\bar{x}_D$$

Para formar los grupos se puede tomar $n_I = n_D$ como el entero más próximo a $n/3$.

Método de Bartlett

Bartlett (1949) modificó los dos métodos anteriores con la propuesta

$$b_B = \frac{\bar{y}_D - \bar{y}_I}{\bar{x}_D - \bar{x}_I}$$
$$a_B = \bar{y} - b_B \bar{x}$$

de forma que la recta pasa por el punto $(\bar{x}, \bar{y})$.

Recta de Brown-Mood

La propuesta de Brown y Mood (1951) es un método diferente que utiliza la mediana de dos grupos. La pendiente b_{BM} y el punto de intercepción a_{BM} se calculan de forma que la mediana de los residuos en cada uno de los dos grupos sea cero:

$$\underset{x_i \leq M_x}{\text{mediana}}\{y_i - a_{BM} - b_{BM}x_i\} = 0$$
$$\underset{x_i > M_x}{\text{mediana}}\{y_i - a_{BM} - b_{BM}x_i\} = 0$$

La inclusión de la mediana M_x en el primer grupo es arbitraria: el objetivo es que los dos grupos sean muy parecidos en su tamaño.

Para hallar los valores efectivos se propone un método iterativo similar al de las secciones anteriores.

7.3. Métodos que ofrecen resistencia

En la sección anterior hemos visto que la recta resistente de los tres grupos no fue la primera alternativa a la de los mínimos cuadrados. Incluso la última de las rectas propuestas, la recta de Brown-Mood, ofrece también resistencia. Ahora acabaremos esta breve descripción de técnicas con algunas que proporcionan como mínimo un cierto grado de resistencia. Pero primero debemos definir una medida de resistencia.

Una de las atractivas características de la recta resistente de los tres grupos es su habilidad para tolerar puntos "salvajes", es decir, puntos que son inusuales en su valor x o en su valor y o en ambos. Para medir esta resistencia aplicaremos el concepto de colapso (*breakdown*) introducido por Hampel (1971).

Definición 7.3.1

El punto de colapso *(breakdown bound) de un procedimiento para ajustar una recta a n parejas de datos y-versus-x es la proporción k/n, donde k es el mayor número de puntos que pueden ser reemplazados arbitrariamente mientras dejen la pendiente y el punto de intercepción delimitados.*

En la práctica, podemos pensar en enviar puntos al infinito al azar o en direcciones problemáticas hasta que la pendiente y el punto de intercepción no lo puedan tolerar más y se colapsen yendo también ellos hacia el infinito. Nos preguntamos cuan grande debe ser una parte de los datos para que un cambio drástico no afecte de forma considerable la recta ajustada.

Está claro que la recta mínimo-cuadrática tiene punto de colapso cero.

Dado que la recta resistente de los tres grupos usa la mediana dentro de cada grupo, hallaremos su punto de colapso en 1/3 veces el punto de colapso de la mediana de una muestra ordinaria. La mediana es el valor central, entonces su punto de colapso es 1/2, de manera que el punto de colapso de la recta resistente es 1/6. A pesar de las diversas posibilidades de construcción de los tres grupos y el hecho que los puntos salvajes pueden estar repartidos en los tres grupos, la idea es que 1/6 es lo mejor que podemos garantizar en la más desfavorable de las circunstancias.

Residuos mínimo-absolutos

Minimizar la suma de los residuos en valor absoluto tiene una historia casi tan larga como la del método de los mínimos cuadrados. Para ajustar una recta hace falta hallar b_{MA} y a_{MA} que minimicen

$$\sum_{i=1}^{n} |y_i - a_{MA} - b_{MA}x_i|$$

Al contrario que para los mínimos cuadrados, no hay una fórmula para calcular b_{MA} y a_{MA}. De hecho, la pendiente y el punto de intercepción pueden no ser únicos.
Como la mediana es la medida que minimiza

$$\sum_{i=1}^{n} |y_i - t|$$

hace falta esperar que este procedimiento tenga un alto punto de colapso. Desgraciadamente, este colapso es 0. La suma que se minimiza involucra tanto los valores x_i como los y_i y así es posible pensar en un punto (x_i, y_i) que tome el control de la recta.

Mediana de las pendientes por parejas

Otra forma de aplicar la mediana al ajuste de una recta consiste en determinar, para cada pareja de puntos, la pendiente y entonces calcular la mediana de estas pendientes. Con más cuidado, supongamos que los x_i son todos diferentes, definimos

$$b_{ij} = \frac{y_j - y_i}{x_j - x_i} \qquad 1 \leq i < j \leq n$$

que son $n(n-1)/2$ valores. La pendiente ajustada es

$$b_T = \text{Med}\{b_{ij}\}$$

Este método es una propuesta de Theil (1950), mejorada por Sen (1968), para manejar las repeticiones de los x_i.
Para deducir el punto de colapso, supongamos que exactamente k de los n puntos son salvajes. Entonces el número de pendientes salvajes es

$$\frac{k(k-1}{2} + k(n-k)$$

Si este número es suficientemente grande, b_T quedará descontrolada. Para valores de n grandes, podemos multiplicar el número de pendientes $n(n-1)/2$ por $1/2$, el punto de colapso de la mediana, y igualar con la expresión anterior. Si resolvemos la ecuación planteada para k, obtenemos un valor de k/n aproximadamente de 0,29. Esto quiere decir que el punto de colapso de b_T es 0,29.

Recta con medianas repetidas

Para conseguir un alto punto de colapso, Siegel (1982) ideó el método de las medianas repetidas.

Empezamos con las pendientes por parejas del método anterior, pero ahora tomaremos las medianas en dos pasos, primero en cada punto y después para todos

$$b_{MR} = \underset{i}{\text{Med}}\{\underset{j\neq i}{\text{Med}}\{b_{ij}\}\}$$

En el primer paso se toma la mediana de las pendientes de $n-1$ rectas que pasan por el punto (x_i, y_i) y en el segundo paso se toma la mediana de estas n pendientes.

Para el punto de intercepción calcularemos $a_i = y_i - b_{MR}x_i$ y entonces

$$a_{MR} = \underset{i}{\text{Med}}\{a_i\}$$

Siegel probó que el punto de colapso de la recta con medianas repetidas es esencialmente $1/2$.

7.3.1. Discusión

Ahora que tenemos diversos métodos con diferentes puntos de colapso, ¿cómo podemos elegir uno?

Una consideración es el grado de resistencia que una particular aplicación pide. Otro asunto es la precisión relativa de las pendientes estimadas, especialmente en muestras pequeñas. También es evidente que el tiempo de computación es otro de los factores a tener en cuenta.

Finalmente, podemos decir que la recta resistente de los tres grupos tiene un comportamiento suficientemente bueno en los tres aspectos considerados y, por ello, es el método resistente que hemos destacado.

Capítulo 8

Regresión lineal múltiple

8.1. El modelo

De forma análoga al caso de la regresión lineal simple, podemos considerar el modelo lineal entre una variable aleatoria respuesta Y y un grupo de k variables no aleatorias $x_1, \ldots, x_k$ explicativas o regresoras.

Si $y_1, \ldots, y_n$ son n observaciones independientes de Y, el modelo lineal de la regresión múltiple se define como

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \qquad i = 1, \ldots, n \tag{8.1}$$

donde $(x_{i1}, \ldots, x_{ik})$ son los valores observados correspondientes a y_i y se asumen las consabidas hipótesis de Gauss-Markov sobre los errores.

En notación matricial, el modelo se escribe

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

donde $\mathbf{Y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ y la matriz de diseño es

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}$$

Se supone además que $\mathrm{rg}(\mathbf{X}) = k + 1 = m$ coincide con el número de parámetros.

Se trata de calcular el ajuste MC a un hiperplano k dimensional, donde β_0 es el punto de intersección del hiperplano con el eje y cuando $x_1 = x_2 = \cdots = x_k = 0$.

Las ecuaciones normales son $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$ donde

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum x_{i1} & \sum x_{i2} & \ldots & \sum x_{ik} \\ & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \ldots & \sum x_{i1}x_{ik} \\ & & \sum x_{i2}^2 & \ldots & \sum x_{i2}x_{ik} \\ & & & \ddots & \vdots \\ & & & & \sum x_{ik}^2 \end{pmatrix} \qquad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \vdots \\ \sum x_{ik}y_i \end{pmatrix}$$

y cuya solución son las estimaciones $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, sin ningún problema de estimabilidad ya que el modelo es de rango máximo. Además, estas estimaciones son insesgadas y de varianza mínima.

Las predicciones de los valores de Y dadas las observaciones de las variables regresoras $x_1, \ldots, x_k$ son

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{Y}$$

es decir

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} \qquad i = 1, \ldots, n \tag{8.2}$$

También podemos considerar el modelo con las variables regresoras centradas

$$\mathbf{Y} = (\mathbf{1}, \mathbf{Z}) \begin{pmatrix} \gamma \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \boldsymbol{\epsilon}$$

donde las columnas de $\mathbf{Z}$ tienen media cero, es decir, $\mathbf{z}_{(j)} = \mathbf{x}_{(j)} - \bar{x}_j \mathbf{1}$ o

$$z_{ij} = x_{ij} - \bar{x}_j \qquad i = 1, \ldots, n \quad j = 1, \ldots, k$$

Este modelo es equivalente al anterior con $\gamma = \beta_0 + \sum_j \bar{x}_j \beta_j$, pero su estimación es más sencilla porque

$$[(\mathbf{1}, \mathbf{Z})'(\mathbf{1}, \mathbf{Z})]^{-1} = \begin{pmatrix} 1/n & \mathbf{0} \\ \mathbf{0} & (\mathbf{Z}'\mathbf{Z})^{-1} \end{pmatrix}$$

ya que $\mathbf{Z}'\mathbf{1} = \mathbf{0}$.

Entonces

$$\hat{\gamma} = \bar{y} \qquad (\hat{\beta}_1, \ldots, \hat{\beta}_k)' = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{Y} - \mathbf{1}\bar{y})$$

Si definimos la matriz simétrica de varianzas-covarianzas, aunque de forma convencional, entre las variables $Y, x_1, \ldots, x_k$

$$\mathbf{S} = \begin{pmatrix} s_y^2 & \mathbf{S}_{\mathbf{x}y} \\ \mathbf{S}_{y\mathbf{x}} & \mathbf{S}_{\mathbf{xx}} \end{pmatrix} = n^{-1}(\mathbf{Y} - \mathbf{1}\bar{y}, \mathbf{Z})'(\mathbf{Y} - \mathbf{1}\bar{y}, \mathbf{Z})$$

resulta

$$(\hat{\beta}_1, \ldots, \hat{\beta}_k)' = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{S}_{y\mathbf{x}}$$

Por todo ello, si consideramos las medias de los datos

$$\bar{y} = (1/n)\sum_i y_i \qquad \bar{x}_j = (1/n)\sum_i x_{ij} \qquad j = 1, \ldots, k$$

8.2 se expresa también en la forma

$$\hat{y}_i - \bar{y} = \hat{\beta}_1(x_{i1} - \bar{x}_1) + \cdots + \hat{\beta}_k(x_{ik} - \bar{x}_k)$$

Finalmente, observemos que el parámetro β_j, $j = 1, \ldots, k$, indica el incremento en Y cuando x_j aumenta en una unidad manteniéndose constantes el resto de variables regresoras. A veces se les llama coeficientes de regresión parcial porque reflejan el efecto de una variable regresora dada la presencia del resto que permanece constante.

Los residuos de la regresión son

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$$

que verifican las propiedades que se han explicado para la regresión simple en la página 94 (ver ejercicio 6.4).

8.2. Medidas de ajuste

Como en la regresión simple, la evaluación del ajuste del hiperplano de regresión a los datos se puede hacer con la *varianza residual* o estimación MC de σ^2.

La suma de cuadrados residual es

$$\text{SCR} = \mathbf{e}'\mathbf{e} = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2 = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

que tiene $n - m$ grados de libertad. Así, la estimación centrada de la varianza del diseño es el llamado *error cuadrático medio*

$$\hat{\sigma}^2 = \text{SCR}/(n-m) = \text{ECM}$$

Su raíz cuadrada $\hat{\sigma}$, que tiene las mismas unidades que Y, es el *error estándar de la regresión múltiple*. También aquí, la varianza residual y el error estándar dependen de las unidades de la variable respuesta y no son útiles para comparar diversas regresiones.

En primer lugar, vamos a introducir el *coeficiente de correlación múltiple* de Y sobre $x_1, \ldots, x_k$. El uso del término *correlación* es convencional puesto que las variables regresoras no son aleatorias. El coeficiente se define como la correlación muestral entre Y e $\hat{Y}$

$$r_{y\mathbf{x}} = \text{corr}(Y, \hat{Y}) = \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{y})}{[\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{y})^2]^{1/2}}$$

ya que $(1/n)\sum \hat{y}_i = \bar{y}$.

El coeficiente de correlación múltiple $r_{y\mathbf{x}}$ verifica $0 \leq r_{y\mathbf{x}} \leq 1$ y es una buena medida del ajuste de Y al modelo $\mathbf{X}\boldsymbol{\beta}$, pues

$$r_{y\mathbf{x}} = 1 \Longrightarrow \|\mathbf{Y} - \widehat{\mathbf{Y}}\| = 0$$

El siguiente teorema, idéntico al teorema 6.2.1, justifica la definición del coeficiente de determinación como medida de ajuste.

Teorema 8.2.1

Las sumas de cuadrados asociadas a la regresión múltiple verifican:

(i) $\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$

(ii) $r^2_{y\mathbf{x}} = \dfrac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$

(iii) $\text{SCR} = \sum (y_i - \hat{y}_i)^2 = (1 - r^2_{y\mathbf{x}}) S_y$

Demostración:

La descomposición en suma de cuadrados (i) se justifica de la misma forma que se ha visto en el teorema 6.2.1. También se puede ver el ejercicio 5.8.

El hecho fundamental es la ortogonalidad

$$(\mathbf{Y} - \widehat{\mathbf{Y}})'\widehat{\mathbf{Y}} = 0$$

pues el vector $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ es ortogonal a $\Omega = \langle \mathbf{X} \rangle$, mientras que $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} \in \Omega$ (ver teorema 2.4.2 y su interpretación geométrica).

Luego

$$\begin{aligned}\sum(y_i-\bar{y})(\hat{y}_i-\bar{y}) &= \sum(y_i-\hat{y}_i+\hat{y}_i-\bar{y})(\hat{y}_i-\bar{y})\\ &= \sum(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})+\sum(\hat{y}_i-\bar{y})^2\\ &= \sum(\hat{y}_i-\bar{y})^2\end{aligned}$$

puesto que el primer sumando es nulo. Teniendo en cuenta la definición de $r_{y\mathbf{x}}$, es fácil deducir (ii).

Finalmente, combinando (i) y (ii) obtenemos (iii). ■

Como en 6.7, la descomposición (i) del teorema anterior justifica la definición del coeficiente de determinación

$$R^2 = \frac{\text{VE}}{\text{VT}} = 1 - \frac{\text{SCR}}{S_y}$$

También aquí, esta medida del ajuste verifica $0 \leq R^2 \leq 1$ y coincide con el cuadrado del coeficiente de correlación múltiple

$$R^2 = 1 - \frac{(1-r_{y\mathbf{x}}^2)S_y}{S_y} = r_{y\mathbf{x}}^2$$

Sin embargo, el coeficiente de correlación múltiple $r_{y\mathbf{x}}$ es una medida de la asociación lineal entre la variable respuesta Y y las regresoras $\mathbf{x} = (x_1, \ldots, x_k)$ que, en este caso, es convencional.

Como R^2 es la proporción de variabilidad explicada por las variables regresoras, resulta que si $R^2 \approx 1$, entonces la mayor parte de la variabilidad es explicada por dichas variables. Pero R^2 es la proporción de la variabilidad total explicada por el modelo con todas las variables frente al modelo $y = \beta_0$, de manera que un R^2 alto muestra que el modelo mejora el modelo nulo y por tanto sólo tiene sentido comparar coeficientes de determinación entre modelos anidados (casos particulares).

Además un valor grande de R^2 no necesariamente implica que el modelo lineal es bueno. El coeficiente R^2 no mide si el modelo lineal es apropiado. Es posible que un modelo con un valor alto de R^2 proporcione estimaciones y predicciones pobres, poco precisas. El análisis de los residuos es imprescindible.

Tampoco está claro lo que significa un valor "grande", ya que problemas en diversas ciencias (física, ingeniería, sociología,...) tienen razonablemente criterios diferentes.

Por otra parte, cuando se añaden variables regresoras R^2 crece, pero eso no significa que el nuevo modelo sea superior:

$$R^2_{\text{nuevo}} = 1 - \frac{\text{SCR}_{\text{nuevo}}}{S_y} \geq R^2 = 1 - \frac{\text{SCR}}{S_y} \qquad \Rightarrow \qquad \text{SCR}_{\text{nuevo}} \leq \text{SCR}$$

pero es posible que

$$\text{ECM}_{\text{nuevo}} = \frac{\text{SCR}_{\text{nuevo}}}{n-(m+p)} \geq \text{ECM} = \frac{\text{SCR}}{n-m}$$

luego, en esta situación, el nuevo modelo será peor. Así, como R^2 crece al añadir nuevas variables regresoras, se corre el peligro de sobreajustar el modelo añadiendo términos innecesarios. El coeficiente de determinación ajustado penaliza esto.

Definición 8.2.1
Una medida del ajuste de la regresión múltiple a los datos es el coeficiente de determinación *o proporción de variabilidad explicada*

$$R^2 = \frac{\text{VE}}{\text{VT}} = 1 - \frac{\text{SCR}}{S_y}$$

Sin embargo, para corregir el peligro de sobreajuste se define el coeficiente de determinación ajustado *como*

$$\bar{R}^2 = 1 - \frac{\text{SCR}/(n-m)}{S_y/(n-1)} = 1 - \frac{n-1}{n-m}(1-R^2)$$

Cuando $\bar{R}^2$ y R^2 son muy distintos, el modelo ha sido sobreajustado y debemos eliminar variables o términos.

8.3. Inferencia sobre los coeficientes de regresión

Cuando asumimos la hipótesis de normalidad sobre la distribución de los errores $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, se deduce la normalidad de la variable respuesta

$$\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

lo que nos permite utilizar las distribuciones asociadas a los estimadores de los parámetros que hemos estudiado.
En el capítulo de contraste de hipótesis se ha visto de varias formas (ver 5.10) que para una función paramétrica estimable $\mathbf{a}'\boldsymbol{\beta}$

$$\frac{\mathbf{a}'\widehat{\boldsymbol{\beta}} - \mathbf{a}'\boldsymbol{\beta}}{(\hat{\sigma}^2 \cdot \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{a})^{1/2}} \sim t_{n-r}$$

En nuestro caso, todas las funciones paramétricas son estimables ya que $r = k+1 = m$. De modo que el estimador $\hat{\beta}_j$ verifica

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{ECM}\, c_{jj}}} \sim t_{n-m} \tag{8.3}$$

donde c_{jj} es el j-ésimo elemento de la diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$ y $\hat{\sigma}^2 = \text{SCR}/(n-m) = \text{ECM}$.
En consecuencia, los intervalos de confianza de los coeficientes de regresión β_j con un nivel de confianza $100(1-\alpha)\,\%$ son

$$\hat{\beta}_j \pm t_{n-m}(\alpha) \cdot \text{ee}(\hat{\beta}_j)$$

donde $\text{ee}(\hat{\beta}_j) = \sqrt{\text{ECM}\, c_{jj}}$.
En cuanto a los intervalos de confianza para la respuesta media o los intervalos de predicción para una respuesta concreta, su deducción es similar al caso de la regresión simple.
Si $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0k})'$ recoge una observación particular del conjunto de variables regresoras, el intervalo de confianza con nivel $100(1-\alpha)\,\%$ para la respuesta media $E[Y|\mathbf{x}_0]$ está centrado en su estimación $\hat{y}_0 = \mathbf{x}_0'\widehat{\boldsymbol{\beta}}$

$$\hat{y}_0 \pm t_{n-m}(\alpha) \cdot (\text{ECM}\, \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)^{1/2}$$

ya que $E(\hat{y}_0) = \mathbf{x}_0'\boldsymbol{\beta} = E[Y|\mathbf{x}_0]$ y $\text{var}(\hat{y}_0) = \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$.

Extrapolación oculta

En la estimación de la respuesta media o la predicción de nuevas respuestas en un punto $(x_{01}, \ldots, x_{0k})$ debemos ser muy cuidadosos con la extrapolación. Si únicamente tenemos en cuenta el producto cartesiano de los recorridos de las variables regresoras, es fácil considerar la predicción para un punto que puede estar fuera de la nube de puntos con la que hemos calculado la regresión. Para evitar este problema deberemos ceñirnos al menor conjunto convexo que contiene los n puntos originales y que recibe el nombre de casco (*hull*) de las variables regresoras (ver figura 8.1).

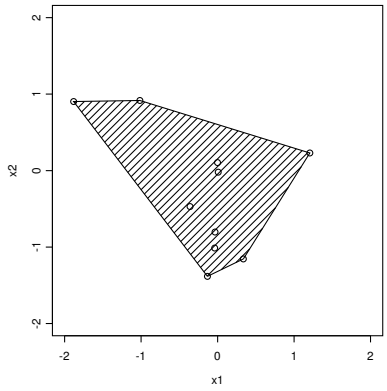


Figura 8.1: Conjunto convexo para los puntos de dos variables regresoras

Si consideramos los elementos h_{ii} de la diagonal de la matriz proyección $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, podemos definir $h_{\text{máx}} = \text{máx}\{h_{11}, \ldots, h_{nn}\}$ y se puede comprobar que

$$\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x} \leq h_{\text{máx}}$$

es un elipsoide que contiene al casco. No es el menor elipsoide, pero es el más fácil de calcular.

Así pues, para evitar en lo posible la extrapolación, podemos comprobar en el punto $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0k})'$ si

$$\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 < h_{\text{máx}}$$

Contraste de significación de la regresión

La hipótesis de mayor interés es la afirmación de que Y es independiente de las variables $x_1, \ldots, x_k$, es decir

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \tag{8.4}$$

El Análisis de la Varianza del teorema 5.3.1 se puede aplicar al contraste de la significación conjunta de los coeficientes de regresión puesto que se trata de una hipótesis contrastable del tipo $H_0 : \mathbf{A}\boldsymbol{\beta} = 0$, donde

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \qquad \text{rango } \mathbf{A} = k$$

Si H_0 es cierta, al igual que en 6.9, la estimación del único parámetro que queda en el modelo es $\hat{\beta}_{0|H} = \bar{y}$ y la suma de cuadrados residual es

$$\mathrm{SCR}_H = \sum (y_i - \bar{y})^2 = S_y$$

que tiene $n-1$ grados de libertad.
La descomposición en suma de cuadrados es

$$S_y = \mathrm{SCR} + (\mathrm{SCR}_H - \mathrm{SCR})$$

es decir

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

La tabla siguiente recoge esta descomposición y realiza el contraste de la hipótesis. La hipótesis se rechaza si $F > F_{k,n-k-1}(\alpha)$.

Fuente de variación	grados de libertad	suma de cuadrados	cuadrados medios	F
Regresión	k	$\mathrm{SC}_R = \mathrm{SCR}_H - \mathrm{SCR}$	CM_R	$\mathrm{CM}_R/\mathrm{ECM}$
Error	$n-k-1$	SCR	ECM	
Total	$n-1$	S_y		

Cuadro 8.1: Tabla del análisis de la varianza para contrastar la significación de la regresión múltiple

Teniendo en cuenta las fórmulas del teorema 8.2.1

$$\mathrm{SCR}_H - \mathrm{SCR} = r^2_{y\mathbf{x}} S_y$$

y deducimos una expresión equivalente al estadístico F

$$F = \frac{r^2_{y\mathbf{x}}}{1 - r^2_{y\mathbf{x}}} \cdot \frac{n-k-1}{k}$$

que también se presenta en forma de tabla.

Fuente de variación	Grados de libertad	Suma de cuadrados	F
Regresión	k	$r^2_{y\mathbf{x}} S_y$	$\dfrac{r^2_{y\mathbf{x}}}{1 - r^2_{y\mathbf{x}}} \cdot \dfrac{n-k-1}{k}$
Residuo	$n-k-1$	$(1 - r^2_{y\mathbf{x}}) S_y$	
Total	$n-1$	S_y	

Cuadro 8.2: Tabla del análisis de la varianza en regresión múltiple

Del mismo modo que en la sección 6.5 la hipótesis 8.4 equivale a afirmar que el coeficiente de correlación múltiple poblacional es cero y se resuelve con el contraste asociado a la tabla anterior.

Significación parcial

El contraste de significación de un coeficiente de regresión particular $H_0 : \beta_j = 0$, para un j fijo, se resuelve con el estadístico 8.3 y la región crítica

$$\left| \frac{\hat{\beta}_j}{(\mathrm{ECM}\, c_{jj})^{1/2}} \right| > t_{n-k-1}(\alpha) \tag{8.5}$$

donde c_{jj} es el j-ésimo elemento de la diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$.

Aceptar esta hipótesis significa que la variable regresora x_j se puede eliminar del modelo. Sin embargo, es preciso actuar con cuidado ya que se trata de un contraste *parcial* porque el coeficiente $\hat{\beta}_j$ depende de todas las otras variables regresoras x_i $(i \neq j)$. Es un contraste de la contribución de x_j dada la presencia de las otras variables regresoras en el modelo.

De forma general podemos estudiar la contribución al modelo de un subconjunto de las variables regresoras. Esto se puede hacer mediante la descomposición de la suma de cuadrados asociada a un contraste de modelos.

Consideremos el modelo lineal completo, dividido en dos grupos de variables regresoras,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\epsilon}$$

donde $\mathbf{X}_1$ es $n \times (m-p)$ y $\mathbf{X}_2$ es $n \times p$.

Para este modelo, la estimación de los parámetros es $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ y la suma de cuadrados de la regresión es

$$\mathrm{SC}_R(\boldsymbol{\beta}) = \mathrm{SCR}_H - \mathrm{SCR} = \mathbf{Y}'\mathbf{Y} - (\mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}) = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$$

con m grados de libertad. Esto es así porque la hipótesis considerada es $H_0 : \boldsymbol{\beta} = \mathbf{0}$ y, bajo esta hipótesis, $\mathrm{SCR}_H = \mathbf{Y}'\mathbf{Y}$.

Para hallar la contribución de los términos de $\boldsymbol{\beta}_2$ en la regresión, podemos considerar la hipótesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ que es equivalente al modelo reducido $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$. Bajo esta hipótesis, la estimación de los parámetros es $\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}$ y la suma de cuadrados de la regresión

$$\mathrm{SC}_R(\boldsymbol{\beta}_1) = \widehat{\boldsymbol{\beta}}_1'\mathbf{X}_1'\mathbf{Y}$$

con $m-p$ grados de libertad.

Luego la suma de cuadrados de la regresión debida a $\boldsymbol{\beta}_2$, dado que $\boldsymbol{\beta}_1$ está ya en el modelo, es

$$\mathrm{SC}_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) = \mathrm{SC}_R(\boldsymbol{\beta}) - \mathrm{SC}_R(\boldsymbol{\beta}_1)$$

con $m - (m-p) = p$ grados de libertad.

Como $\mathrm{SC}_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)$ es independiente de SCR, la hipótesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ se puede contrastar con el estadístico

$$\frac{\mathrm{SC}_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)/p}{\mathrm{ECM}} \sim F_{p,n-m}$$

que se puede llamar una F parcial, pues mide la contribución de $\mathbf{X}_2$ considerando que $\mathbf{X}_1$ está en el modelo.

Por ejemplo, la suma de cuadrados de la regresión $\mathrm{SC}_R(\beta_j|\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k)$ para $1 \leq j \leq k$ es el crecimiento en la suma de cuadrados debido a añadir x_j al modelo

que ya contiene todas las otras variables, como si fuera la última variable añadida al modelo. El contraste es equivalente al contraste 8.5.

Estos contrastes F parciales juegan un papel muy importante en la búsqueda del mejor conjunto de variables regresoras a utilizar en un modelo. Por ejemplo, en el modelo parabólico $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ estaremos interesados en $\mathrm{SC}_R(\beta_1|\beta_0)$ y luego en $\mathrm{SC}_R(\beta_2|\beta_0, \beta_1)$ que es la contribución cuadrática al modelo lineal simple.

En el modelo $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$, la descomposición en suma de cuadrados es

$$S_y = \mathrm{SC}_R(\beta_1, \beta_2, \beta_3|\beta_0) + \mathrm{SCR}$$

pero

$$\begin{aligned}\mathrm{SC}_R(\beta_1, \beta_2, \beta_3|\beta_0) &= \mathrm{SC}_R(\beta_1|\beta_0) + \mathrm{SC}_R(\beta_2|\beta_0, \beta_1) + \mathrm{SC}_R(\beta_3|\beta_0, \beta_1, \beta_2) \\ &= \mathrm{SC}_R(\beta_2|\beta_0) + \mathrm{SC}_R(\beta_1|\beta_0, \beta_2) + \mathrm{SC}_R(\beta_3|\beta_0, \beta_1, \beta_2) \\ &= \dots\end{aligned}$$

Sin embargo, hay que ir con cuidado porque este método no siempre produce una partición de la suma de cuadrados de la regresión y, por ejemplo,

$$\mathrm{SC}_R(\beta_1, \beta_2, \beta_3|\beta_0) \neq \mathrm{SC}_R(\beta_1|\beta_2, \beta_3, \beta_0) + \mathrm{SC}_R(\beta_2|\beta_1, \beta_3, \beta_0) + \mathrm{SC}_R(\beta_3|\beta_1, \beta_2, \beta_0)$$

Un resultado interesante se tiene cuando las columnas de $\mathbf{X}_1$ y $\mathbf{X}_2$ son ortogonales, ya que entonces

$$\mathrm{SC}_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) = \mathrm{SC}_R(\boldsymbol{\beta}_2) \qquad \mathrm{SC}_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) = \mathrm{SC}_R(\boldsymbol{\beta}_1)$$

Región de confianza y intervalos simultáneos

Del mismo modo que hemos explicado en 6.3.6, en regresión múltiple la región con una confianza conjunta del $100(1-\alpha)\,\%$ es

$$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{m\mathrm{ECM}} \leq F_{m,n-m}(\alpha)$$

Los intervalos simultáneos para los coeficientes de la regresión son del tipo

$$\hat{\beta}_j \pm \Delta \cdot \mathrm{ee}(\hat{\beta}_j)$$

para un conjunto de s coeficientes entre los $k+1$. Por ejemplo, el método de Scheffé proporciona los intervalos simultáneos

$$\hat{\beta}_j \pm (sF_{s,n-k-1}(\alpha))^{1/2} \cdot \mathrm{ee}(\hat{\beta}_j)$$

Los intervalos simultáneos para un conjunto de s respuestas medias a los puntos $\mathbf{x}_{01}, \dots, \mathbf{x}_{0s}$ son

$$\hat{y}_{\mathbf{x}_{0j}} \pm \Delta(\mathrm{ECM}\,\mathbf{x}_{0j}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{0j})^{1/2}$$

donde $\Delta = (sF_{s,n-k-1}(\alpha))^{1/2}$ por el método de Scheffé.

8.4. Coeficientes de regresión estandarizados

Es difícil comparar coeficientes de regresión porque la magnitud de $\hat{\beta}_j$ refleja las unidades de medida de la variable regresora. Por ejemplo, en el modelo

$$Y = 5 + x_1 + 1000x_2$$

donde x_1 se mide en litros y x_2 en mililitros, aunque $\hat{\beta}_2 = 1000$ es mucho mayor que $\hat{\beta}_1 = 1$, el efecto sobre Y es el mismo.
Generalmente, las unidades de los coeficientes de regresión son

$$\text{unidades } \hat{\beta}_j = \frac{\text{unidades } Y}{\text{unidades } x_j}$$

Por todo ello, frecuentemente es de gran ayuda trabajar con variables estandarizadas que producen coeficientes de regresión sin dimensión. Básicamente hay dos técnicas:
Escala normal unidad

$$\begin{aligned} z_{ij} &= \frac{x_{ij} - \bar{x}_j}{\hat{s}_j} \qquad i = 1, \ldots, n; j = 1, \ldots, k \\ y_i^* &= \frac{y_i - \bar{y}}{\hat{s}_y} \qquad i = 1, \ldots, n \end{aligned}$$

donde

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad \hat{s}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \qquad \hat{s}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

El modelo es

$$y_i^* = b_0 + b_1 z_{i1} + b_2 z_{i2} + \cdots + b_k z_{ik} + \eta_i \qquad i = 1, \ldots, n$$

donde las variables regresoras y la variable respuesta tienen media cero y varianza muestral uno. La estimación del modelo es $\widehat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_k)' = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}^*$ y $\hat{b}_0 = \bar{y}^* = 0$.
Escala longitud unidad

$$\begin{aligned} w_{ij} &= \frac{x_{ij} - \bar{x}_j}{S_j^{1/2}} \qquad i = 1, \ldots, n; j = 1, \ldots, k \\ y_i^0 &= \frac{y_i - \bar{y}}{S_y^{1/2}} \qquad i = 1, \ldots, n \end{aligned}$$

donde

$$S_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \qquad S_y = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

El modelo es

$$y_i^0 = b_1 w_{i1} + b_2 w_{i2} + \cdots + b_k w_{ik} + \eta_i \qquad i = 1, \ldots, n$$

donde las variables regresoras y la variable respuesta tienen media cero y longitud

$$\sqrt{\sum_{i=1}^{n}(w_{ij} - \bar{w}_j)^2} = 1$$

y la estimación de los parámetros es $\widehat{\mathbf{b}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}^0$.
Pero en este modelo tenemos

$$\mathbf{W}'\mathbf{W} = \mathbf{R}_{\mathbf{xx}} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \dots & 1 \end{pmatrix}$$

donde $\mathbf{R}_{\mathbf{xx}}$ es la matriz de correlaciones de las variables regresoras ya que

$$r_{ij} = \frac{\sum_{s=1}^{n}(x_{si} - \bar{x}_i)(x_{sj} - \bar{x}_j)}{(S_i S_j)^{1/2}}$$

También podemos considerar que $\mathbf{W}'\mathbf{Y}^0 = \mathbf{R}_{\mathbf{x}y}$ es el vector de correlaciones de las variables regresoras con la variable respuesta. También aquí el término correlación es convencional.
En todo caso, como

$$\mathbf{Z}'\mathbf{Z} = (n-1)\mathbf{W}'\mathbf{W}$$
$$\mathbf{Z}'\mathbf{Y}^* = (n-1)\mathbf{W}'\mathbf{Y}^0$$

las estimaciones de $\mathbf{b} = (b_1, \dots, b_k)'$ por ambos métodos son idénticas.

Definición 8.4.1

Se llaman coeficientes de regresión estandarizados *los que se obtienen como solución del sistema de ecuaciones*

$$\begin{array}{rcccl} b_1 & +r_{12}b_2 & +\cdots+ & r_{1k}b_k & = r_{1y} \\ r_{21}b_1 + & b_2 & +\cdots+ & r_{2k}b_k & = r_{2y} \\ \vdots & \vdots & & \vdots & \vdots \\ r_{k1}b_1 + & r_{k2}b_2 & +\cdots+ & b_k & = r_{ky} \end{array}$$

es decir

$$\mathbf{R}_{\mathbf{xx}}\mathbf{b} = \mathbf{R}_{\mathbf{x}y}$$

donde $\mathbf{R}_{\mathbf{xx}}$ *es la matriz de coeficientes de correlación entre las variables regresoras y* $\mathbf{R}_{\mathbf{x}y} = (r_{1y}, \dots, r_{ky})'$ *el vector columna con los coeficientes de correlación entre las variables regresoras y la respuesta.*

Los coeficientes de regresión ordinarios se deducen de las ecuaciones

$$\hat{\beta}_j = \hat{b}_j \left(\frac{S_y}{S_j}\right)^{1/2} = \hat{b}_j \frac{s_y}{s_j} \qquad j = 1, \dots, k$$
$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{k} \hat{\beta}_j \bar{x}_j$$

Además, el coeficiente de determinación es

$$R^2 = r^2_{y\mathbf{x}} = \hat{b}_1 r_{1y} + \hat{b}_2 r_{2y} + \cdots + \hat{b}_k r_{ky}$$

Algunos paquetes estadísticos calculan ambos conjuntos de coeficientes de regresión. En algún caso, a los coeficientes de regresión estandarizados les llaman "beta coeficientes" lo que para nosotros es confuso.

Finalmente señalaremos que debemos cuidar las interpretaciones puesto que los coeficientes estandarizados todavía son parciales, es decir, miden el efecto de x_j dada la presencia de las otras variables regresoras. También $\hat{b}_j$ está afectado por el recorrido de los valores de las variables regresoras, de modo que es peligroso utilizar $\hat{b}_j$ para medir la importancia relativa de la variable regresora x_j.

Ejemplo 8.4.1

En un estudio sobre la incidencia que puede tener sobre el rendimiento en lenguaje Y, la comprensión lectora x_1 y la capacidad intelectual x_2, se obtuvieron datos sobre 10 *estudiantes tomados al azar de un curso de básica (ver tabla 8.3).*

Y	x_1	x_2
3	1	3
2	1	4
4	3	7
9	7	9
6	8	7
7	7	6
2	4	5
6	6	8
5	6	5
8	9	7

Cuadro 8.3: Tabla de datos del rendimiento en lenguaje

La matriz de correlaciones, las medias y las desviaciones típicas son:

	x_1	x_2	Y		
x_1	1	0,6973	0,8491	$\bar{x}_1 = 5{,}2$	$s_1 = 2{,}82$
x_2		1	0,7814	$\bar{x}_2 = 6{,}1$	$s_2 = 1{,}86$
Y			1	$\bar{y} = 5{,}2$	$s_y = 2{,}44$

Empezaremos planteando el sistema

$$b_1 + 0{,}6973 \cdot b_2 = 0{,}8491$$
$$0{,}6973 \cdot b_1 + b_2 = 0{,}7814$$

cuya solución es

$$\hat{b}_1 = 0{,}592 \qquad \hat{b}_2 = 0{,}368$$

Entonces

$$\hat{\beta}_1 = \hat{b}_1 \frac{s_y}{s_1} = 0{,}512 \qquad \hat{\beta}_2 = \hat{b}_2 \frac{s_y}{s_2} = 0{,}485$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 = -0{,}424$$

La ecuación de regresión es

$$y = -0{,}424 + 0{,}512x_1 + 0{,}485x_2$$

El coeficiente de determinación es

$$R^2 = r^2_{y\mathbf{x}} = \hat{b}_1 \cdot 0{,}849 + \hat{b}_2 \cdot 0{,}781 = 0{,}791$$

y puede afirmarse que hay una buena relación entre el rendimiento en lenguaje y la comprensión lectora y la capacidad intelectual.

Finalmente, para decidir sobre la hipótesis $H_0 : \beta_1 = \beta_2 = 0$ *calcularemos*

$$F = \frac{r^2_{y\mathbf{x}}}{1 - r^2_{y\mathbf{x}}} \cdot \frac{10-3}{3-1} = 13{,}22$$

con 2 *y* 7 *grados de libertad. Así* H_0 *puede ser rechazada, es decir, la relación anterior es significativa.*

8.5. Multicolinealidad

Cuando la matriz $\mathbf{X}$ no es de rango máximo, sabemos que $\mathbf{X}'\mathbf{X}$ es singular y no podemos calcular su inversa. Ya sabemos que la solución puede ser la utilización de alguna g-inversa, aunque ello implica que la solución de las ecuaciones normales no es única. En el caso de la regresión múltiple es difícil, aunque no imposible, que alguna columna sea linealmente dependiente de las demás. Si ocurriera esto diríamos que existe colinealidad entre las columnas de $\mathbf{X}$. Sin embargo, el término colinealidad o multicolinealidad se refiere al caso, mucho más frecuente, de que la dependencia entre las columnas no es exacta sino aproximada, es decir, a la quasi-dependencia lineal entre las variables regresoras. Esto puede provocar problemas de computación de los parámetros y en el cálculo de la precisión de los mismos (ver Apéndice A.4).

Entre las múltiples formas de detección de la multicolinealidad vamos a destacar el cálculo de los factores de inflación de la varianza. Nosotros hemos visto que la matriz de varianzas-covarianzas de los estimadores de los parámetros de un modelo lineal es

$$\text{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Si consideramos el modelo de regresión estandarizado por la escala de longitud unidad, la matriz de varianzas-covarianzas de los coeficientes de regresión estandarizados es

$$\text{var}(\widehat{\mathbf{b}}) = \tilde{\sigma}^2\mathbf{R}^{-1}_{\mathbf{xx}}$$

donde $\tilde{\sigma}^2$ es la varianza del error del modelo transformado. En particular, la varianza de uno de los coeficientes es

$$\text{var}(\hat{b}_j) = \tilde{\sigma}^2[\mathbf{R}^{-1}_{\mathbf{xx}}]_{jj}$$

donde $[\mathbf{R}^{-1}_{\mathbf{xx}}]_{jj}$ es el j-ésimo elemento de la diagonal de la matriz. Estas varianzas pueden estar "infladas" a causa de la multicolinealidad que puede ser evidente a partir de la observación de los elementos no nulos fuera de la diagonal de $\mathbf{R}_{\mathbf{xx}}$, es decir, de las correlaciones simples entre las variables regresoras.

Definición 8.5.1

Los elementos de la diagonal de la matriz $\mathbf{R}^{-1}_{\mathbf{xx}}$ *se llaman* FIV *o factores de inflación de la varianza ya que*

$$\text{var}(\hat{b}_j) = \tilde{\sigma}^2\text{FIV}_j$$

Se demuestra que

$$\text{FIV}_j = (1 - R_j^2)^{-1}$$

donde R_j^2 es el coeficiente de determinación múltiple de la variable regresora x_j con todas las demás variables regresoras.

El factor de inflación de la varianza $\text{FIV}_j = 1$ cuando $R_j^2 = 0$, es decir, cuando x_j no depende linealmente del resto de las variables. Cuando $R_j^2 \neq 0$, entonces $\text{FIV}_j > 1$ y si $R_j^2 \approx 1$, entonces FIV_j es grande. Así pues, el factor de inflación de la varianza mide el incremento que se produce en la varianza de los estimadores de los coeficientes de regresión al comparar dicha varianza con la que deberían tener si las variables regresoras fuesen incorrelacionadas.

Cuando $\text{FIV}_j > 10$ tenemos un grave problema de multicolinealidad. Algunos autores prefieren calcular la media de los FIV_j y alertar sobre la multicolinealidad cuando dicha media supera el número 10.

Una de las posibles soluciones tras la detección de multicolinealidad es la estimación por la regresión *ridge* (ver 4.3.1).

Ejemplo 8.5.1

Con los datos del ejemplo 8.4.1, la matriz de correlaciones $\mathbf{R_{xx}}$ *y su inversa son*

$$\mathbf{R_{xx}} = \begin{pmatrix} 1{,}0000 & 0{,}6973 \\ 0{,}6973 & 1{,}0000 \end{pmatrix} \qquad \mathbf{R_{xx}^{-1}} = \begin{pmatrix} 1{,}9465 & -1{,}3574 \\ -1{,}3574 & 1{,}9465 \end{pmatrix}$$

y los factores de inflación de la varianza son $\text{FIV}_1 = 1{,}9465, \text{FIV}_2 = 1{,}9465$*, que coinciden naturalmente cuando* $k = 2$*.*

8.6. Regresión polinómica

Supongamos que una variable aleatoria Y se ajusta a una variable de control x según un modelo polinómico de grado m

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_m x_i^m + \epsilon_i \tag{8.6}$$

Observemos que se trata de un modelo de regresión lineal múltiple de Y sobre las variables $x_1 = x, x_2 = x^2, \ldots, x_m = x^m$. Para una regresión polinómica de grado m, la matriz de diseño es

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{pmatrix}$$

Estos modelos se pueden aplicar cuando el analista sabe que efectos curvilíneos están presentes en la función respuesta. También se pueden utilizar como aproximaciones a desconocidas, y posiblemente muy complejas, relaciones no lineales. Así, los polinomios se pueden considerar los desarrollos de Taylor de la función desconocida.

La regresión polinómica se justifica por el teorema de Weierstrass, el cual dice que toda función continua $f(x)$ se puede aproximar por un polinomio $P_m(x)$ de grado m adecuado. Se puede probar esta propiedad desde el punto de vista probabilístico:

Sea $f(x)$ una función continua en el intervalo $(0,1)$ y consideremos

$$P_n(x) = \sum_{k=0}^{n} f(k/n) x^k (1-x)^{n-k}$$

llamados polinomios de Bernstein. Entonces $P_n(x)$ converge a $f(x)$ cuando $n \to \infty$, uniformemente en x.

Como en cualquier modelo lineal, la estimación de los parámetros de regresión se hace con las ecuaciones normales. Sin embargo, hay varios problemas especiales que se presentan en este caso.

1) Es muy importante que el orden del polinomio sea tan bajo como sea posible. Para utilizar polinomio de grado $m > 2$ se debe justificar con razones externas a los datos. Existen transformaciones de las variables, en particular de la respuesta, que hacen que el modelo sea de primer orden. Un modelo de orden bajo con una variable transformada es casi siempre preferible a un modelo de orden superior con la métrica original. Se trata de mantener el principio de parsimonia o simplicidad de los modelos.

2) Hay varias estrategias para elegir el grado del polinomio.

   Selección hacia adelante (forward selection): Se trata de ir ajustando modelos en orden creciente hasta que el test t para el término de mayor orden es no significativo ($\alpha = 0{,}1$).

   Selección hacia atrás (backward selection): Se trata de ajustar un modelo de alto orden e ir eliminando términos si no son significativos para el test t ($\alpha = 0{,}1$).

   Ambos métodos no necesariamente conducen al mismo modelo. En todo caso, hay que recordar el consejo anterior y tratar con modelos de orden dos o muy bajo.

3) Debemos ser muy cuidadosos con la extrapolación (ver página 140), ya que las consecuencias pueden ser ruinosas.

4) Cuando el orden del polinomio es alto, la matriz $\mathbf{X}'\mathbf{X}$ está mal condicionada (ver apéndice A.4 y sección 8.5). Esto provoca problemas graves para el cálculo de los coeficientes de regresión y deficiencias en la precisión de los mismos. En Seber [65] pág. 214 se ve un ejemplo en el que variaciones del orden de 10^{-10} en $\mathbf{X}'\mathbf{Y}$ producen variaciones del orden de 3 en los elementos de $\widehat{\boldsymbol{\beta}}$.

   De hecho, los modelos de regresión polinómicos están notablemente mal condicionados cuando el grado es mayor que 5 o 6, particularmente si los valores de x están igualmente espaciados.

5) Si los valores de x tienen un recorrido muy estrecho, esto puede conducir a la multicolinealidad entre las columnas de $\mathbf{X}$. Por ejemplo, si x varía entre 1 y 2, x^2 varía entre 1 y 4, lo que puede provocar una fuerte dependencia entre los datos de x y x^2.

Para reducir el efecto no esencial de la mala condición de los modelos de regresión polinómicos se deben centrar las variables regresoras. Además se pueden utilizar polinomios de Tchebychev o, mejor, polinomios ortogonales.

La utilización de polinomios de Tchebychev consiste en considerar el modelo

$$y_i = \gamma_0 T_0(x_i) + \gamma_1 T_1(x_i) + \cdots + \gamma_m T_m(x_i) + \epsilon_i$$

donde $T_j(x)$ es un polinomio de Tchebychev de grado j. Estos polinomios se generan mediante la relación de recurrencia

$$T_{j+1}(x) = 2xT_j(x) - T_{j-1}(x)$$

Tomando inicialmente

$$T_0(x) = 1 \qquad T_1(x) = x$$

se obtienen

$$\begin{aligned} T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1 \\ &\vdots \end{aligned}$$

El campo de variación de x debe "normalizarse" adecuadamente entre -1 y 1 mediante un cambio de variable. Esto se hace en favor de la estabilidad numérica.

Los polinomios de Tchebychev tienen propiedades muy interesantes que sugieren que, para valores de x razonablemente espaciados, la matriz del modelo $\widetilde{\mathbf{X}}$ tiene columnas que son aproximadamente ortogonales, de forma que la matriz $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$ tiene los elementos de fuera de la diagonal bastante pequeños y generalmente está bien condicionada. Así pues, un procedimiento de cálculo de regresión polinómica consiste en usar polinomios de Tchebychev junto con un método de descomposición ortogonal de la matriz de diseño, como el algoritmo QR.

8.6.1. Polinomios ortogonales

El replanteamiento del modelo 8.6 mediante polinomios ortogonales permite una solución sencilla de los problemas numéricos mencionados.

Consideremos ahora el modelo

$$y_i = \gamma_0 \phi_0(x_i) + \gamma_1 \phi_1(x_i) + \cdots + \gamma_m \phi_m(x_i) + \epsilon_i \tag{8.7}$$

donde $\phi_j(x_i)$ es un polinomio de grado j en x_i ($j = 0, 1, \ldots, m$). Supongamos que los m polinomios son ortogonales, es decir,

$$\sum_{i=1}^{n} \phi_j(x_i)\phi_{j'}(x_i) = 0 \qquad \forall j \neq j' \tag{8.8}$$

El modelo lineal es entonces

$$\mathbf{Y} = \widetilde{\mathbf{X}}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

donde

$$\widetilde{\mathbf{X}} = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \ldots & \phi_m(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \ldots & \phi_m(x_2) \\ \vdots & \vdots & & \vdots \\ \phi_0(x_n) & \phi_1(x_n) & \ldots & \phi_m(x_n) \end{pmatrix}$$

Entonces, debido a la ortogonalidad, tenemos que

$$\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} = \begin{pmatrix} \sum \phi_0^2(x_i) & 0 & \dots & 0 \\ 0 & \sum \phi_1^2(x_i) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sum \phi_m^2(x_i) \end{pmatrix}$$

y la solución de las ecuaciones normales es

$$\hat{\gamma}_j = \frac{\sum_i \phi_j(x_i) y_i}{\sum_i \phi_j^2(x_i)} \qquad j = 0, 1, \dots, m$$

lo que es cierto para toda m. La estructura ortogonal de $\widetilde{\mathbf{X}}$ implica que el estimador MC de γ_j $(j \leq m)$ es independiente del grado m del polinomio, lo que es una propiedad muy deseable.

Como $\phi_0(x)$ es un polinomio de grado cero, si tomamos $\phi_0(x) = 1$ tendremos $\hat{\gamma}_0 = \bar{y}$.

La suma de cuadrados residual es entonces

$$\text{SCR}(m) = \sum (y_i - \bar{y})^2 - \sum_{j=1}^{m} (\sum_i \phi_j^2(x_i)) \hat{\gamma}_j^2 \tag{8.9}$$

cantidad que indicaremos por $Q(m)$.

En efecto:

$$\hat{y}_i = \sum_{j=0}^{m} \phi_j(x_i) \hat{\gamma}_j \qquad \text{siendo} \qquad \bar{y} = \phi_0(x_i) \hat{\gamma}_0$$

Aplicando (i) de 8.2.1 tenemos

$$\text{SCR}(m) = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \bar{y})^2 - \sum_i (\hat{y}_i - \bar{y})^2$$

siendo ahora

$$\sum_i (\hat{y}_i - \bar{y})^2 = \sum_i (\sum_{j=1}^{m} \phi_j(x_i) \hat{\gamma}_j)^2$$

Por otra parte

$$(\sum_{j=1}^{m} \phi_j(x_i) \hat{\gamma}_j)^2 = \sum_j \sum_{j'} \phi_j(x_i) \hat{\gamma}_j \cdot \phi_{j'}(x_i) \hat{\gamma}_{j'}$$

y sumando respecto de i tenemos, considerando 8.8,

$$\begin{aligned} \sum_i (\hat{y}_i - \bar{y})^2 &= \sum_j \sum_{j'} \hat{\gamma}_j \hat{\gamma}_{j'} (\sum_i \phi_j(x_i) \phi_{j'}(x_i)) \\ &= \sum_{j=1}^{m} \hat{\gamma}_j^2 (\sum_{i=1}^{n} \phi_j^2(x_i)) \end{aligned}$$

lo que demuestra 8.9.

Existen diversos procedimientos para generar polinomios ortogonales (Fisher, Forsythe, Hayes, etc.).

En el caso particular que los valores de x sean igualmente espaciados podemos transformarlos de manera que

$$x_i = i - \tfrac{1}{2}(n+1) \qquad i = 1, 2, \ldots, n$$

Entonces se puede considerar el siguiente sistema de polinomios ortogonales

$$\begin{aligned}
\phi_0(x) &= 1 \\
\phi_1(x) &= \lambda_1 x \\
\phi_2(x) &= \lambda_2(x^2 - \tfrac{1}{12}(n^2-1)) \\
\phi_3(x) &= \lambda_3(x^3 - \tfrac{1}{20}(3n^2-7)x) \\
&\vdots
\end{aligned}$$

donde las λ_j se eligen de forma que los valores de $\phi_j(x_i)$ sean enteros. Estos polinomios se encuentran tabulados para varios valores de n.

8.6.2. Elección del grado

Un aspecto importante de la regresión polinómica es la elección del grado m adecuado. El contraste de hipótesis

$$\begin{aligned}
H_0 &: m = m_0 \\
H_1 &: m = m_1 > m_0
\end{aligned} \tag{8.10}$$

equivale a plantear una regresión polinómica de grado m y entonces establecer la hipótesis lineal

$$H_0 : \beta_{m_0+1} = \cdots = \beta_{m_1} = 0$$

sobre el modelo 8.6, o bien, utilizando el modelo equivalente 8.7 en términos de polinomios ortogonales

$$H_0 : \gamma_{m_0+1} = \cdots = \gamma_{m_1} = 0$$

Las sumas de cuadrados residuales son

$$\mathrm{SCR} = Q(m_1) \qquad \mathrm{SCR}_H = Q(m_0)$$

Teniendo en cuenta 8.9 resulta

$$\mathrm{SCR}_H - \mathrm{SCR} = Q(m_0) - Q(m_1) = \sum_{j=m_0+1}^{m_1} (\sum_{i=1}^{n} \phi_j^2(x_i))\hat{\gamma}_j^2$$

Entonces, para contrastar $H_0 : m = m_0$ frente $H_1 : m = m_1$, calcularemos el estadístico

$$F = \frac{(Q(m_0) - Q(m_1))/(m_1 - m_0)}{Q(m_1)/(n - m_1 - 1)} \tag{8.11}$$

cuya distribución, bajo H_0, es una F con $m_1 - m_0$ y $n - m_1 - 1$ grados de libertad.

La estrategia para elegir el grado puede ser mediante elección descendente o elección ascendente. En el primer caso empezamos por el grado que se supone máximo. Supongamos, por ejemplo, que $m = 5$. Entonces se contrasta $m = 4$ frente a $m = 5$. Si el test F no es

significativo, se contrasta $m = 3$ con $m = 4$, y así sucesivamente. El proceso es el inverso en el caso de elección ascendente.

También es útil tener en cuenta que un descenso importante de la suma de cuadrados residual $Q(m)$ al pasar de grado k a grado m, es un indicio de que el grado es m.

Finalmente, si disponemos de n_i observaciones $y_{i1}, \ldots, y_{in_i}$ para cada valor de la variable de control x_i $i = 1, \ldots, p$, una vez elegido el grado m, podemos analizar la validez del modelo planteando el contraste

$$H_0 : y_{ih} = P_m(x_i) + \epsilon_{ih}$$
$$H_1 : y_{ih} = g(x_i) + \epsilon_{ih}$$

donde $g(x)$ es una función desconocida de x. La hipótesis nula significa afirmar que $g(x) = P_m(x)$ es un polinomio de grado m en x. Tenemos entonces (véase 6.12):

$$\text{SCR} = \sum_{i,h}(y_{ih} - \bar{y}_i)^2 = ns_y^2(1 - \hat{\eta}^2) \qquad n - p \qquad \text{g.l.}$$
$$\text{SCR}_H = Q(m) = ns_y^2(1 - r_{y\mathbf{x}}^2) \qquad n - m - 1 \quad \text{g.l.}$$

donde $r_{y\mathbf{x}}$ es la correlación múltiple de Y sobre $x, x^2, \ldots, x^m$ (ver teorema 8.2.1). Calcularemos entonces el estadístico

$$F = \frac{(\hat{\eta}^2 - r_{y\mathbf{x}}^2)/(p - m - 1)}{(1 - \hat{\eta}^2)/(n - p)}$$

y aceptaremos el ajuste polinómico de grado m si esta F no es significativa.

Ejemplo 8.6.1

Se dispone de la respuesta a un test de conducta de dos grupos de ratas, uno control y otro experimental, para diez observaciones realizadas cada tres días desde el día 47 *al día* 74 *de vida (ver tabla 8.4).*

dia	grupo control	grupo experimental
47	25,7	34,1
50	20,1	24,9
53	16,2	21,2
56	14,0	23,3
59	21,3	22,0
62	20,3	30,9
65	28,4	31,4
68	23,5	26,5
71	16,8	23,0
74	9,9	17,2

Cuadro 8.4: Datos del test de conducta a dos grupos de ratas

El modelo considerado hace depender la variable conducta (medida mediante el test) del tiempo t según una función polinómica

$$\textit{var. obs.} = \textit{polinomio de grado } m \textit{ en } t + \textit{error} \qquad \Leftrightarrow \qquad y = P_m(t) + \epsilon$$

Para determinar el grado del polinomio al cual se ajustan los valores experimentales se plantea la hipótesis 8.10 que se resuelve mediante el test F 8.11.
Los resultados, obtenidos según el método de los polinomios ortogonales, son los siguientes

grupo control	*g.l.*	*grupo experimental*	*g.l.*
$Q(0) = 273{,}87$	9	$Q(0) = 249{,}99$	9
$Q(1) = 249{,}22$	8	$Q(1) = 216{,}12$	8
$Q(2) = 233{,}52$	7	$Q(2) = 213{,}15$	7
$Q(3) = 41{,}61$	6	$Q(3) = 37{,}80$	6
$Q(4) = 41{,}52$	5	$Q(4) = 27{,}10$	5

Observemos que hay un fuerte descenso de la suma de cuadrados residual $Q(m)$ al pasar de grado 2 a grado 3, indicio de que los datos experimentales se ajustan a un polinomio de grado 3.
Las F obtenidas son:

contraste	*grupo control*	*grupo experimental*
0 v.s. 1	$F = 0{,}79$ *(n.s.)*	$F = 1{,}25$ *(n.s.)*
0 v.s. 2	$F = 0{,}60$ *(n.s.)*	$F = 0{,}60$ *(n.s.)*
0 v.s. 3	$F = 11{,}16$ *($p < 0{,}01$)*	$F = 11{,}23$ *($p < 0{,}01$)*
1 v.s. 3	$F = 14{,}97$ *($p < 0{,}01$)*	$F = 14{,}25$ *($p < 0{,}01$)*
2 v.s. 3	$F = 27{,}67$ *($p < 0{,}01$)*	$F = 27{,}83$ *($p < 0{,}01$)*
3 v.s. 4	$F = 0{,}01$ *(n.s.)*	$F = 1{,}98$ *(n.s.)*

Efectivamente, tanto los datos del grupo control como los del grupo experimental se ajustan a un polinomio de grado 3 (ver Figura 8.2).

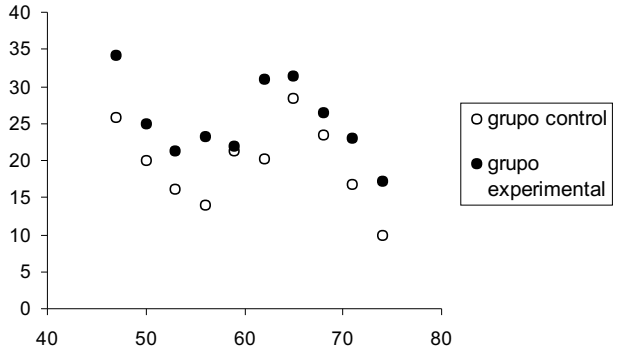


Figura 8.2: Gráfico de los dos grupos de ratas

El modelo es:
grupo control (○)

$$y_i = 1929{,}24 - 97{,}86t_i + 1{,}654t_i^2 - 0{,}0092t_i^3 + \epsilon_i$$

grupo experimental (●)

$$y_i = 1892{,}28 - 94{,}94t_i + 1{,}593t_i^2 - 0{,}0088t_i^3 + \epsilon_i$$

8.7. Comparación de curvas experimentales

8.7.1. Comparación global

Si dos curvas experimentales se ajustan bien a modelos de formulación matemática diferente (por ejemplo, dos polinomios de distinto grado) hay que aceptar que las curvas experimentales son distintas.

Si las dos curvas son polinomios del mismo grado

$$\begin{aligned} y_1 &= P_m(x) + \epsilon \\ y_2 &= \bar{P}_m(x) + \epsilon \end{aligned}$$

la comparación se expresa planteando el siguiente contraste de hipótesis

$$\begin{aligned} H_0 &: P_m(x) = \bar{P}_m(x) \\ H_1 &: P_m(x) \neq \bar{P}_m(x) \end{aligned} \tag{8.12}$$

que implica la hipótesis lineal

$$H_0 : \beta_i = \bar{\beta}_i \qquad i = 0, 1, \ldots, m$$

análoga a

$$H_0 : \gamma_i = \bar{\gamma}_i \qquad i = 0, 1, \ldots, m \tag{8.13}$$

si utilizamos el modelo planteado mediante polinomios ortogonales (ver 8.7).

Sean $\mathrm{SCR}_1 = Q_1(m)$, $\mathrm{SCR}_2 = Q_2(m)$ las sumas de cuadrados residuales para cada curva y $\mathrm{SCR} = \mathrm{SCR}_1 + \mathrm{SCR}_2$ la suma de cuadrados residual del modelo conjunto construido mediante la unión de los dos modelos.

La construcción del modelo conjunto es sólo posible si los dos modelos poseen varianzas iguales. Por este motivo, es necesario plantear previamente el test de homogeneidad de varianzas

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_1 &: \sigma_1^2 \neq \sigma_2^2 \end{aligned}$$

que se resuelve mediante el estadístico

$$F = \frac{\mathrm{SCR}_1/(n_1 - m - 1)}{\mathrm{SCR}_2/(n_2 - m - 1)} \tag{8.14}$$

cuya distribución si H_0 es cierta es una F con $n_1 - m - 1$ y $n_2 - m - 1$ g.l..

Si aceptamos la igualdad de varianzas, podemos resolver 8.13 mediante el estadístico

$$F = \frac{(\mathrm{SCR}_H - \mathrm{SCR}_1 - \mathrm{SCR}_2)/(m+1)}{(\mathrm{SCR}_1 + \mathrm{SCR}_2)/(n_1 + n_2 - 2m - 2)} \tag{8.15}$$

que bajo H_0 sigue una F con $m + 1$ y $n_1 + n_2 - 2m - 2$ g.l.. La suma de cuadrados $\mathrm{SCR}_H = Q_{12}(m)$ es la suma de cuadrados residual bajo H_0, es decir, considerando que las dos curvas son iguales y que en consecuencia todos los datos se ajustan a un mismo polinomio de grado m.

8.7.2. Test de paralelismo

La hipótesis lineal de que las curvas son paralelas se plantea de la siguiente forma

$$H_0 : \beta_i = \bar{\beta}_i \qquad i = 1, \ldots, m$$

o bien, si nos referimos a 8.7

$$H_0 : \gamma_i = \bar{\gamma}_i \qquad i = 1, \ldots, m \tag{8.16}$$

Es decir, las curvas difieren únicamente respecto a la ordenada en el origen.

Esta hipótesis tiene generalmente interés cuando se rechaza H_0 de 8.12. Se resuelve mediante el estadístico

$$F = \frac{(\mathrm{SCR}^*_H - \mathrm{SCR}_1 - \mathrm{SCR}_2)/m}{(\mathrm{SCR}_1 + \mathrm{SCR}_2)/(n_1 + n_2 - 2m - 2)} \tag{8.17}$$

cuya distribución sigue una F con m y $n_1 + n_2 - 2m - 2$ g.l. cuando H_0 es cierta. La suma de cuadrados SCR^*_H es la suma de cuadrados residual bajo H_0 que supone aceptar la existencia de dos curvas distintas pero paralelas.

Ejemplo 8.7.1

En el ejemplo 8.6.1 hemos ajustado los datos del grupo control y del grupo experimental a dos polinomios de grado 3.

¿Podemos aceptar que en realidad los dos polinomios son iguales? Esta pregunta equivale a plantear la hipótesis lineal 8.13. Para resolverla es necesario realizar previamente el test de homogeneidad de varianzas utilizando 8.14

$$F = \frac{41{,}61/(10 - 3 - 1)}{37{,}80/(10 - 3 - 1)} = 1{,}10$$

con 6 *y* 6 *g.l. (no significativa).*

Pasamos pues a contrastar 8.13 mediante el estadístico 8.15. La suma de cuadrados residual bajo H_0 *es* $\mathrm{SCR}_H = Q_{12}(3) = 249{,}06$

$$F = \frac{(249{,}06 - 41{,}61 - 37{,}80)/(3 + 1)}{(41{,}61 + 37{,}80)/(10 + 10 - 6 - 2)} = 6{,}41$$

con 4 *y* 12 *g.l. que es significativa* ($p < 0{,}01$*). Debemos aceptar en consecuencia que las dos curvas son diferentes (la conducta de los individuos del grupo control es diferente de la conducta de los individuos del grupo experimental).*

No obstante, podemos preguntarnos si las dos curvas son paralelas y plantear la hipótesis lineal 8.16 que resolveremos utilizando el estadístico 8.17. La suma de cuadrados residual bajo H_0 *es ahora* $\mathrm{SCR}^*_H = Q^*_{12} = 82{,}59$

$$F = \frac{(82{,}59 - 41{,}61 - 37{,}80)/3}{(41{,}61 + 37{,}80)/(10 + 10 - 6 - 2)} = 0{,}16$$

con 3 *y* 12 *g.l. (no significativa). Podemos entonces aceptar que las dos curvas experimentales son paralelas. La interpretación en términos de la conducta podría realizarse conociendo con más precisión el planteamiento del problema.*

8.8. Ejemplos con R

Vamos a utilizar los datos del ejemplo 8.4.1 sobre el lenguaje. Las siguientes instrucciones permiten introducir los datos y dibujar los diagramas de dispersión dos a dos de las variables del ejemplo (ver figura 8.3).

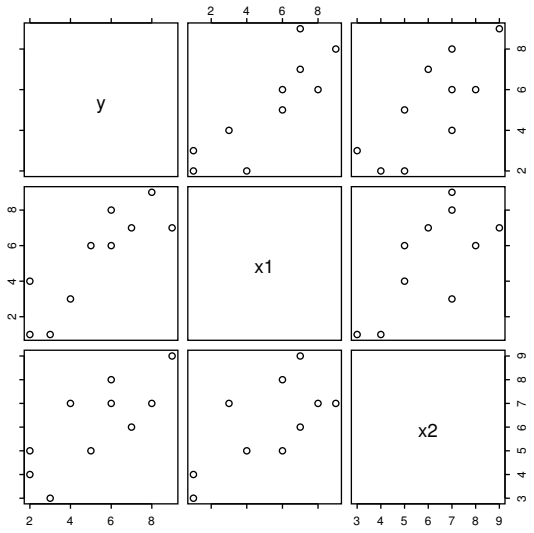


Figura 8.3: Diagramas de dispersión dos a dos entre la variable respuesta y las variables explicativas del ejemplo 8.4.1

```
> y<-c(3,2,4,9,6,7,2,6,5,8)
> x1<-c(1,1,3,7,8,7,4,6,6,9)
> x2<-c(3,4,7,9,7,6,5,8,5,7)
> exp<-cbind(x1,x2)
> lenguaje.datos<-data.frame(y,exp)
> par(pty="s")
> pairs(lenguaje.datos)
```

El siguiente paso es calcular el modelo de regresión lineal múltiple que permita predecir los valores de Y en función de las variables explicativas x_1 y x_2.

```
> regrem<-lm(y~x1+x2)
> summary(regrem)

Call: lm(formula = y ~ x1 + x2)
Residuals:
    Min      1Q   Median     3Q  Max
 -2.051 -0.5264 -0.05257 0.7989 1.47

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) -0.4244  1.4701    -0.2887  0.7812
         x1  0.5123  0.2087     2.4543  0.0438
```

```
         x2  0.4853  0.3178     1.5273  0.1705

Residual standard error: 1.266 on 7 degrees of freedom
Multiple R-Squared: 0.7907
F-statistic: 13.22 on 2 and 7 degrees of freedom, the p-value is 0.004196

Correlation of Coefficients:
   (Intercept)      x1
x1  0.1811
x2 -0.8036     -0.6973
```

El plano estimado es $\hat{y} = -0{,}4244 + 0{,}5123x_1 + 0{,}4853x_2$ con un coeficiente de determinación $R^2 = 0{,}7907$ y el estadístico F nos dice que el modelo es útil, si un estudio más profundo decide finalmente que es realmente válido.

Resulta curioso que en S-PLUS se puede obtener el coeficiente de determinación R^2 a partir de la función `summary.lm` en la forma

```
> summary(regrem)$r.squared
[1] 0.790684
```

pero no hay nombre para el coeficiente ajustado. Mientras que en R sí es posible.

También se pueden obtener los coeficientes a partir de la matriz $\mathbf{X}'\mathbf{X}$:

```
> XtX<-t(regrem$R)%*%regrem$R
> XtX
            (Intercept)  x1  x2
(Intercept)          10  52  61
         x1          52 342 350
         x2          61 350 403
> XtX.inv<-solve(XtX)
> XtX.inv
            (Intercept)          x1         x2
(Intercept)  1.34840753  0.03466479 -0.2342073
         x1  0.03466479  0.02718635 -0.0288580
         x2 -0.23420728 -0.02885800  0.0629949
> XtX.inv%*%t(cbind(1,exp))%*%y
                  [,1]
(Intercept) -0.4244237
         x1  0.5123174
         x2  0.4853071
```

La matriz `XtX.inv` se puede obtener de forma directa así:

```
> summary(regrem)$cov.unscaled
            (Intercept)          x1         x2
(Intercept)  1.34840753  0.03466479 -0.2342073
         x1  0.03466479  0.02718635 -0.0288580
         x2 -0.23420728 -0.02885800  0.0629949
```

También se obtiene más fácilmente con los elementos que proporciona la función `lsfit`:

```
> regrem.ls<-lsfit(exp,y)
> regrem.diag<-ls.diag(regre.ls)
> regrem.diag$cov.unscaled
```

La matriz $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ de varianzas y covarianzas entre los estimadores MC de los coeficientes se obtiene de forma sencilla:

```
> summary(regrem)$sigma^2*summary(regrem)$cov.unscaled
            (Intercept)          x1          x2
(Intercept)  2.16117719  0.05555943 -0.37537868
         x1  0.05555943  0.04357326 -0.04625252
         x2 -0.37537868 -0.04625252  0.10096587
```

o también

```
> regrem.diag$std.dev^2*regrem.diag$cov.unscaled
```

Para calcular intervalos de confianza sobre los coeficientes de regresión hacemos

```
> beta.est<-cbind(regrem.ls$coef);beta.est
                [,1]
Intercept -0.4244237
       x1  0.5123174
       x2  0.4853071
> cbind(beta.est+qt(0.025,7)*regrem.diag$std.err,
+ beta.est+qt(0.975,7)*regrem.diag$std.err)
                    [,1]     [,2]
(Intercept) -3.90064431 3.051797
         x1  0.01872084 1.005914
         x2 -0.26605529 1.236669
```

Observamos que los intervalos correspondientes a β_0 y β_2 contienen al cero, en coherencia con los test t parciales. Pero también nos puede interesar reproducir la tabla ANOVA sobre la significación de la regresión, aunque el test F ya se ha obtenido con la función `summary(regrem)`. Las funciones `anova.lm` o `summary.aov` nos pueden ayudar.

```
> summary.aov(regrem)
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
       x1  1  38.64190 38.64190 24.10956 0.0017330
       x2  1   3.73876  3.73876  2.33270 0.1705213
Residuals  7  11.21934  1.60276
```

Sin embargo, los resultados se refieren a contrastes F secuenciales y parciales. Exactamente $\mathrm{SC}_R(\beta_0, \beta_1) = 38{,}64190$ y $\mathrm{SC}_R(\beta_2|\beta_0, \beta_1) = 3{,}73876$, de manera que

$$\mathrm{SC}_R = \mathrm{SC}_R(\beta_1, \beta_0) + \mathrm{SC}_R(\beta_2|\beta_0, \beta_1) = 42{,}38066$$

Por otra parte, se observa directamente que $\mathrm{SCR} = 11{,}21934$. Con estos datos, completar la tabla 8.1 es relativamente sencillo. Sin embargo se puede conseguir dicha tabla, aunque con otra organización, mediante un contraste de modelos:

```
> regrem0<-lm(y~1)
> anova(regrem0,regrem)
Analysis of Variance Table

Response: y

    Terms Resid. Df      RSS Test Df Sum of Sq  F Value      Pr(F)
1       1         9 53.60000
2 x1 + x2         7 11.21934       2  42.38066 13.22113 0.00419574
```

Otro aspecto que también hemos visto ha sido el cálculo de los coeficientes de regresión estandarizados, que con R se obtienen así:

```
> cor(exp)
          x1        x2
x1 1.0000000 0.6973296
x2 0.6973296 1.0000000
> cor(exp,y)
        [,1]
x1 0.8490765
x2 0.7813857
> solve(cor(exp),cor(exp,y))
        [,1]
x1 0.5921248
x2 0.3684796
```

Si queremos más detalles sobre los coeficientes de regresión estandarizados, podemos utilizar el siguiente modelo sin coeficiente de intercepción:

```
> x1.est<-(x1-mean(x1))/stdev(x1)
> x2.est<-(x2-mean(x2))/stdev(x2)
> y.est<-(y-mean(y))/stdev(y)
> regrem.est<-lm(y.est~-1+x1.est+x2.est)
> summary(regrem.est)
```

Por último, podemos estudiar la multicolinealidad calculando los FIV

```
> diag(solve(cor(exp)))
[1] 1.946542 1.946542
```

que en este caso no existe.

El cálculo de predicciones puntuales o por intervalo se obtiene mediante la función `predict.lm` del modelo lineal.

8.9. Ejercicios

Ejercicio 8.1

Consideremos el modelo de la regresión lineal múltiple

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} \qquad i = 1, \ldots, n$$

Sean $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ las estimaciones MC de los parámetros. Explicar en qué condiciones podemos afirmar que $E(\hat{\beta}_j) = \beta_j$, $j = 0, 1, \ldots, m$.
Por otra parte, ¿es siempre válido afirmar que

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_m x_{im}$$

es una estimación centrada de

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} \ ?$$

Ejercicio 8.2

En la regresión múltiple de una variable Y sobre tres variables control x_1, x_2, x_3

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i \qquad i = 1, \ldots, n$$

donde $\epsilon_i \sim N(0, \sigma^2)$, se desea contrastar la hipótesis nula

$$H_0 : \beta_2 = \beta_3 = 0$$

Sea $r_{y\mathbf{x}}$ el coeficiente de correlación múltiple de Y sobre x_1, x_2, x_3 y sea r_{y1} el coeficiente de correlación simple entre Y y x_1. Deducir un test F para contrastar H_0 que sea función de $r_{y\mathbf{x}}$ y r_{y1}.

Ejercicio 8.3

En una gran ciudad, queremos relacionar el número de muertos diarios por enfermedades cardio-respiratorias con la media de humos (mg/m^3) i la media de dióxido de azufre (partes/millón) medidas por los equipos del Ayuntamiento en diversas zonas de la ciudad.
Consideremos un modelo de regresión lineal no centrado con los siguientes datos:

$$\mathbf{X'X} = \begin{pmatrix} 15 & 6{,}87 & 21{,}09 \\ & 5{,}6569 & 18{,}7243 \\ & & 63{,}2157 \end{pmatrix} \quad (\mathbf{X'X})^{-1} = \begin{pmatrix} 0{,}2243 & -1{,}2611 & 0{,}2987 \\ & 16{,}1158 & -4{,}3527 \\ & & 1{,}2054 \end{pmatrix}$$

$$\mathbf{X'Y} = \begin{pmatrix} 3922 \\ 2439{,}54 \\ 7654{,}35 \end{pmatrix} \qquad \mathbf{Y'Y} = 1264224$$

Se pide:

1) Calcular la estimación MC de todos los coeficientes de regresión del modelo.

2) Obtener una estimación insesgada de la varianza del modelo.

3) Contrastar la significación del modelo propuesto con $\alpha = 0{,}1$.

4) Calcular el intervalo de confianza al 95 % para la media del valor respuesta para una media de humos de 1 mg/m^3 y una media de SO_2 de 1.

Ejercicio 8.4

Se dispone de los siguientes datos sobre diez empresas fabricantes de productos de limpieza doméstica:

Empresa	V	IP	PU
1	60	100	1,8
2	48	110	2,4
3	42	130	3,6
4	36	100	0,6
5	78	80	1,8
6	36	80	0,6
7	72	90	3,6
8	42	120	1,2
9	54	120	2,4
10	90	90	4,2

En el cuadro anterior, V son las ventas anuales, expresadas en millones de euros, IP es un índice de precios relativos (Precios de la empresa/Precios de la competencia) y PU son los gastos anuales realizados en publicidad y campañas de promoción y difusión, expresados también en millones de euros.

Tomando como base la anterior información:

1) Estimar el vector de coeficientes $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ del modelo

$$V_i = \beta_0 + \beta_1 IP_i + \beta_2 PU_i + \epsilon_i$$

2) Estimar la matriz de varianzas-covarianzas del vector $\widehat{\boldsymbol{\beta}}$.

3) Calcular el coeficiente de determinación.

Ejercicio 8.5

Dado el modelo

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$$

y los siguientes datos

Y_t	X_{1t}	X_{2t}
10	1	0
25	3	−1
32	4	0
43	5	1
58	7	−1
62	8	0
67	10	−1
71	10	2

obtener:

(a) La estimación MC de $\beta_0, \beta_1, \beta_2$ utilizando los valores originales.

(b) La estimación MC de $\beta_0, \beta_1, \beta_2$ utilizando los datos expresados en desviaciones respecto a la media.

(c) La estimación insesgada de σ^2.

(d) El coeficiente de determinación.

(e) El coeficiente de determinación corregido.

(f) El contraste de la hipótesis nula $H_0 : \beta_0 = \beta_1 = \beta_2 = 0$.

(g) El contraste de la hipótesis nula $H_0 : \beta_1 = \beta_2 = 0$ utilizando datos originales.

(h) El contraste de la hipótesis nula $H_0 : \beta_1 = \beta_2 = 0$ utilizando datos en desviaciones respecto a la media.

(i) La representación gráfica de una región de confianza del 95 % para β_1 y β_2.

(j) El contraste individual de los parámetros β_0, β_1 y β_2.

(k) El contraste de la hipótesis nula $H_0 : \beta_1 = 10\beta_2$.

(l) El contraste de la hipótesis nula $H_0 : 2\beta_0 + 2\beta_1 + 7\beta_2 = 50$.

(m) El contraste de la hipótesis nula conjunta $H_0 : \beta_1 = 10\beta_2, 2\beta_0 + 2\beta_1 + 7\beta_2 = 50$.

Ejercicio 8.6

Supongamos que hemos estimado la siguiente ecuación utilizando MC (con las variables medidas en logaritmos)

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} \qquad t = 1, \ldots, 17$$

y las estimaciones de los parámetros son:

$$\hat{\beta}_0 = 1{,}37 \qquad \hat{\beta}_1 = 1{,}14 \qquad \hat{\beta}_2 = -0{,}83$$

También hemos obtenido la siguiente expresión escalar:

$$\mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = 0{,}0028$$

y los elementos triangulares de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$ son:

$$\begin{pmatrix} 510{,}89 & -254{,}35 & 0{,}42 \\ & 132{,}70 & -6{,}82 \\ & & 7{,}11 \end{pmatrix}$$

Se pide:

1. Calcular las varianzas de los estimadores MC de $\beta_0, \beta_1, \beta_2$.
2. Si X_{1t} aumenta en un 1 por 100 y X_{2t} en un 2 por 100, ¿cuál sería el efecto estimado en Y_t?

3. Efectuar un test estadístico para verificar la hipótesis de que $\beta_1 = 1$ y $\beta_2 = -1$ y dar el valor de dicho estadístico. ¿Cuáles son las tablas que necesitaremos para realizar el test y cuántos son los grados de libertad?

Ejercicio 8.7

Una variable Y depende de otra variable control x que toma los valores $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$ de acuerdo con el modelo lineal normal

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \qquad i = 1, 2, 3, 4$$

Estudiar la expresión del estadístico F para contrastar la hipótesis $H_0 : \beta_1 = \beta_2$.

Ejercicio 8.8

La puntuación del test *open-field* para un grupo de 10 ratas control (C) y otro grupo de 10 ratas experimentales (E) a lo largo de los días $47, 50, \ldots, 74$ contados desde el instante del nacimiento fue

Día	47	50	53	56	59	62	65	68	71	74
grupo C	34	24	21	23	23	30	31	26	23	17
grupo E	25	20	16	15	21	20	28	23	18	9

Se ajustaron al grupo control polinomios de grado $0, 1, 2$ y 3 respecto la variable "edad en días" y se obtuvieron las siguientes sumas de cuadrados residuales:

$$\begin{aligned} Q(0) &= 235{,}6 \\ Q(1) &= 202{,}8 \\ Q(2) &= 199{,}4 \\ Q(3) &= \ 29{,}7 \end{aligned}$$

Se pide:

1) Comprobar que se puede aceptar como válido el polinomio de grado 3 como polinomio de regresión de Y (puntuación) sobre x (edad en días).

2) El polinomio de grado 3 que ajusta Y a x es

$$y = 318{,}8 - 93{,}3x + 1{,}56x^2 - 0{,}0086x^3$$

El coeficiente de correlación múltiple de Y sobre x, x^2, x^3 es $r_{y\mathbf{x}} = 0{,}8734$. Estudiar si es significativo.

3) Para el grupo experimental es también adecuado un ajuste polinómico de grado 3 con suma de cuadrados residual $Q(3) = 29{,}2$. Además, juntando todos los datos referentes a Y, es decir, juntando los dos grupos y en consecuencia las 20 observaciones y realizando un ajuste polinómico de grado 3, se obtiene

$$\mathrm{SCR}_H = 225{,}8$$

Contrastar las hipótesis

H_0 : los dos polinomios (C y E) son idénticos

H_1 : hay diferencias significativas entre ambos polinomios

Capítulo 9

Diagnosis del modelo

En este capítulo se investiga la detección de posibles deficiencias en el modelo por incumplimiento de las hipótesis fijadas en 2.3. Para ello la principal herramienta es el análisis de los residuos que nos permite detectar los siguientes problemas:

1. Algunas de las variables explicativas del modelo tienen una relación no lineal con la variable respuesta.
2. No hay homocedasticidad, es decir, los errores no tienen varianza constante.
3. Los errores no son independientes.
4. Muchas observaciones atípicas.
5. Hay observaciones demasiado influyentes.
6. Los errores no tienen distribución normal

También estudiaremos la consecución del mejor grupo reducido de variables regresoras.

9.1. Residuos

9.1.1. Estandarización interna

Los residuos de un modelo lineal se obtienen como diferencia entre los valores observados de la variable respuesta y las predicciones obtenidas para los mismos datos:

$$\mathbf{e} = (e_1, \ldots, e_n)' = \mathbf{Y} - \widehat{\mathbf{Y}}$$

La media de los residuos es cero

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

y una estimación aproximada de la varianza es

$$\frac{1}{n-k-1}\sum_{i=1}^{n}(e_i - \bar{e})^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2 = \mathrm{SCR}/(n-k-1) = \mathrm{ECM}$$

que tiene sólo $n-k-1$ grados de libertad, donde k es el número de variables regresoras, ya que los n residuos no son independientes,

Se llaman *residuos estandarizados* a

$$d_i = \frac{e_i}{\sqrt{\text{ECM}}} \qquad i = 1, \ldots, n$$

que tienen media cero y varianza aproximada uno.

Ahora bien, como el vector de residuos aleatorios es $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P})\mathbf{Y} = (\mathbf{I} - \mathbf{P})\boldsymbol{\epsilon}$, donde $\mathbf{P}$ es la matriz proyección, la matriz de varianzas-covarianzas de los residuos es $\text{var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{P})$ de manera que

$$\text{var}(e_i) = \sigma^2(1 - h_{ii})$$

donde h_{ii} es el i-ésimo elemento[1] de la diagonal de $\mathbf{P}$.

La utilización de los residuos $\mathbf{e}$ como estimaciones de los errores $\boldsymbol{\epsilon}$ requiere que mejoremos la estandarización. Como $0 \leq h_{ii} \leq 1$, utilizar ECM para estimar la varianza $\text{var}(e_i)$ es una sobreestimación:

$$\begin{array}{rcl} 0 \leq & \text{var}(e_i) & \leq \quad \sigma^2 \\ 0 \leq & \text{ECM}(1 - h_{ii}) & \leq \text{ECM} \end{array}$$

De modo que muchos autores recomiendan trabajar con los *residuos studentizados*

$$r_i = \frac{e_i}{[\text{ECM}(1 - h_{ii})]^{1/2}} \qquad i = 1, \ldots, n$$

Además, h_{ii} es una medida de la localización del i-ésimo punto $\mathbf{x}_i$ respecto al punto medio.

En la regresión lineal simple

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{9.1}$$

En el modelo de regresión múltiple

$$h_{ii} = \frac{1}{n}[1 + (\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}_{\mathbf{xx}}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})] = \frac{1}{n}(1 + D_i^2) \tag{9.2}$$

donde D_i es la llamada distancia de Mahalanobis.

Así, la varianza de un error e_i depende de la posición del punto $\mathbf{x}_i$. Puntos cercanos al punto central $\bar{\mathbf{x}}$ tienen mayor varianza (pobre ajuste MC) que los puntos alejados.

Como las violaciones de las hipótesis del modelo son más probables en los puntos remotos, pero más difíciles de detectar con los residuos e_i (o d_i), porque los residuos son menores, es mejor trabajar con los residuos r_i ya que $\text{var}(r_i) = 1$ constante, desde el punto de vista de la localización de los $\mathbf{x}_i$.

Para n grande se puede trabajar con los d_i o con los r_i. Pero como valores altos de e_i y de h_{ii} pueden indicar un punto de alta influencia en el ajuste MC, se recomienda la utilización de los residuos estudentizados r_i. Estos residuos se utilizarán en el diagnóstico de valores atípicos.

[1] En muchos libros escritos en inglés la matriz proyección se llama *hat* y se escribe $\mathbf{H}$.

Ejemplo 9.1.1

Si recuperamos el ejemplo de regresión simple propuesto en la sección 1.2 con los datos de tráfico, podemos calcular los residuos studentizados de ese modelo.
Primero calculamos los elementos de la diagonal de la matriz $\mathbf{P}$*, por ejemplo*

$$h_{11} = \frac{1}{24} + \frac{(12{,}7 - 54{,}44167)^2}{15257{,}4383} = 0{,}155865$$

y con este valor se obtiene el residuo

$$r_1 = \frac{0{,}528699}{0{,}2689388(1 - 0{,}155865)^{1/2}} = 2{,}13968$$

Los otros residuos se calculan de forma similar, mejor con la ayuda de una hoja de cálculo o con un programa estadístico (ver sección 9.4).

9.1.2. Estandarización externa

Para calcular los residuos estudentizados r_i en el apartado anterior hemos utilizado ECM como estimador de la varianza σ^2. Nos referiremos a esto como una estimación *interna* puesto que para calcularla se utilizan los n puntos. Otra aproximación consiste en estimar σ^2 con el conjunto de datos sin la i-ésima observación.
Si $s^2_{(i)}$ es la estimación de σ^2 así obtenida, se demuestra que

$$s^2_{(i)} = \frac{(n-k-1)\text{ECM} - e_i^2/(1-h_{ii})}{n-k-2} = \text{ECM}\left(\frac{n-k-1-r_i^2}{n-k-2}\right)$$

Si utilizamos estos estimadores de σ^2 en lugar de ECM, producimos los llamados residuos studentizados externamente o R-Student

$$t_i = \frac{e_i}{[s^2_{(i)}(1-h_{ii})]^{1/2}} \qquad i = 1, \ldots, n \tag{9.3}$$

En la mayoría de situaciones los residuos t_i no diferirán de los residuos studentizados r_i. Sin embargo, si la i-ésima observación es influyente, entonces $s^2_{(i)}$ puede diferir significativamente de ECM y el estadístico t_i será más sensible para este punto. Además, bajo las hipótesis estándar $t_i \sim t_{n-k-2}$, de modo que podemos considerar un procedimiento formal para la detección de valores atípicos mediante el contraste de hipótesis y utilizando algún método múltiple. En la práctica, un diagnóstico "a ojo" es más útil y rápido. En general, se considera que un residuo es atípico o *outlier* si $|t_i| > 2$. Además, la detección de los valores atípicos está ligada a la detección de puntos influyentes.

Ejemplo 9.1.2

Vamos a calcular el residuo studentizado externamente t_1 *para la primera observación de la regresión simple continuación del ejemplo 9.1.1. Para ello necesitamos el valor del error* $\text{ECM} = (0{,}2689388)^2 = 0{,}072328$ *con el que calculamos*

$$s^2_{(i)} = 0{,}072328\frac{24 - 1 - 1 - 2{,}13968^2}{24 - 1 - 2} = 0{,}060004$$

y con esta estimación externa

$$t_1 = \frac{0{,}528699}{\sqrt{0{,}060004(1 - 0{,}155865)}} = 2{,}349159$$

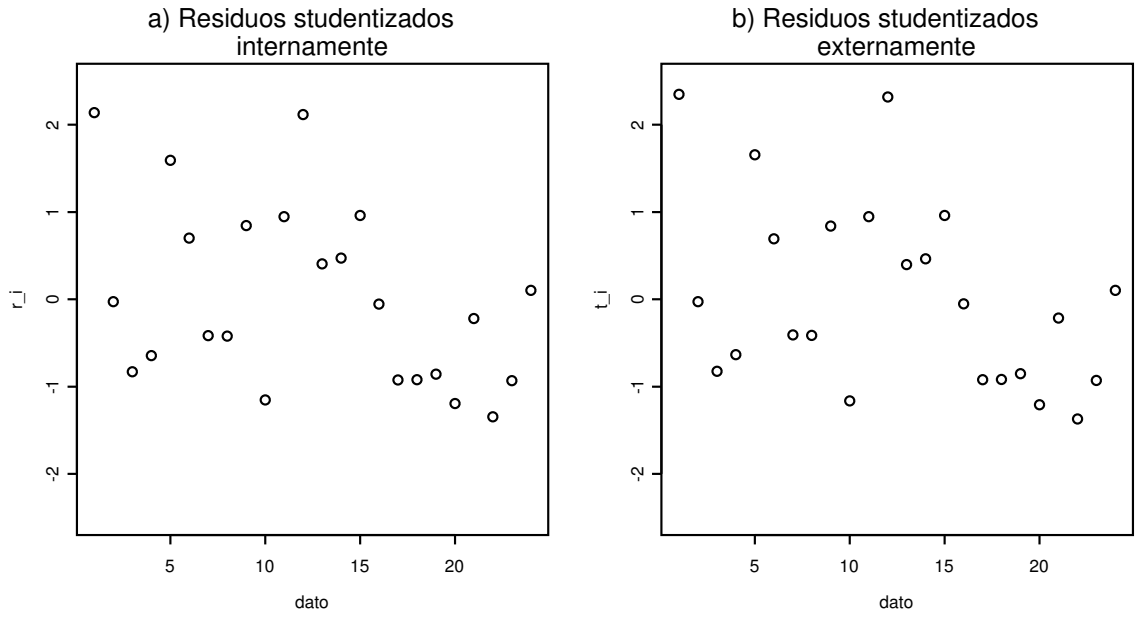


Figura 9.1: Gráficos de los residuos studentizados del ejemplo 9.1.1.

Siguiendo con la misma idea, también podemos calcular los residuos en función de las predicciones $\hat{y}_{i(i)}$ calculadas con el modelo de regresión sin la i-ésima observación. Sean $e_{(i)} = y_i - \hat{y}_{i(i)}$ los residuos así obtenidos y

$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)}^2 \tag{9.4}$$

su suma de cuadrados[2]. También algunos autores llaman *error cuadrático de validación* a esta suma de cuadrados por ser una medida externa de precisión del modelo.
Se demuestra que

$$e_{(i)} = \frac{e_i}{1 - h_{ii}} \qquad \text{var}(e_{(i)}) = \frac{\sigma^2}{1 - h_{ii}} \tag{9.5}$$

de modo que la estandarización de estos residuos

$$\frac{e_{(i)}}{[\text{var}(e_{(i)})]^{1/2}} = \frac{e_i}{[\sigma^2(1 - h_{ii})]^{1/2}}$$

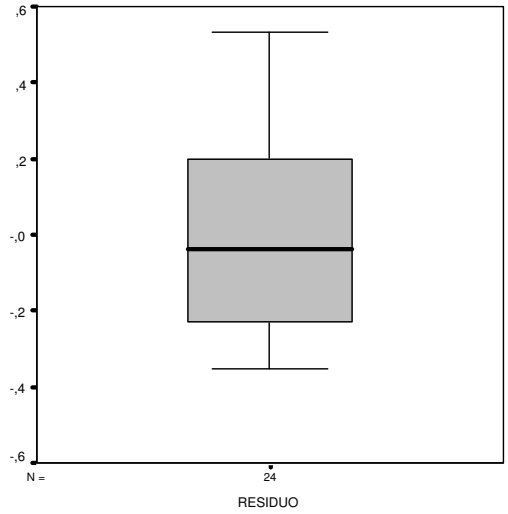
también depende del estimador que utilicemos para estimar σ^2. Si utilizamos el estimador interno ECM, recuperamos los residuos studentizados r_i y si utilizamos el estimador externo $s_{(i)}^2$ obtenemos los residuos studentizados externamente t_i.
Los residuos asociados con puntos para los que h_{ii} sea grande, tendrán residuos $e_{(i)}$ grandes. Estos puntos serán puntos de alta influencia. Una gran diferencia entre el residuo ordinario e_i y el residuo $e_{(i)}$ indicará un punto en el que el modelo, con ese punto, se ajusta bien a los datos, pero un modelo construido sin ese punto "predice" pobremente.

9.1.3. Gráficos

Algunos gráficos de los residuos nos van a ayudar en el diagnóstico del modelo aplicado.

[2] *prediction error sum of squares*

En primer lugar, el análisis de datos univariante de los residuos y, en particular, los gráficos como histogramas, diagramas de caja, diagramas de tallo y hojas, etc. nos mostrarán algunos detalles. Por ejemplo, en el diagrama de caja podemos estudiar la centralidad, la simetría y la presencia de valores atípicos.




```
RESIDUO Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

    14,00       -0 .  00011122222333
     8,00        0 .  01112224
     2,00        0 .  55

 Stem width:  1,000000
 Each leaf:       1 case(s)
```

Figura 9.2: Boxplot y diagrama de tallo y hojas de los residuos en la regresión simple del ejemplo 9.1.3.

Ejemplo 9.1.3

También con los datos de tráfico del ejemplo de regresión simple propuesto en la sección 1.2 podemos representar algunos gráficos de los residuos sin estandarizar. En la figura 9.2 se muestran dos de los gráficos obtenidos con el programa SPSS. En ellos se observa una cierta asimetría de los residuos, aunque no hay ningún valor atípico.

Otros gráficos adecuados para el análisis de la regresión son:

- Gráfico de dispersión de los residuos respecto al índice $i = 1, \ldots, n$.

  Este diagrama puede indicar algún tipo de correlación no deseada entre los residuos o alguna agrupación contraria a la supuesta aleatoriedad (figura 9.3 a).

- Gráfico de los residuos versus los datos de la variable respuesta.

  Permite observar los residuos desde los valores observados de la variable respuesta.

- Gráfico de los residuos versus los valores ajustados.

  Este gráfico es muy importante porque debe mostrar una total aleatoriedad. La dispersión horizontal no debe presentar ninguna tendencia. Una curvatura indica la violación del supuesto de linealidad del modelo en el caso de regresión lineal simple (figura 9.3 b). Una forma triangular indica una posible heterogeneidad o violación de la hipótesis de varianza constante de los errores.

- Gráficos de los residuos versus las observaciones de la variable o variables regresoras.

  Sirven para detectar si las variables regresoras o explicativas han de incluirse en el modelo con alguna transformación no lineal.

- Gráfico de los valores observados versus los valores ajustados.

  La proximidad de los puntos a la bisectriz muestra el ajuste de la recta de regresión (figura 9.3 c).

- Gráfico de los cuantiles de la distribución normal o QQ-plot y gráfico de las probabilidades acumuladas de la distribución normal o PP-plot.

  Con estos gráficos se pretende visualizar el ajuste de la distribución muestral de los residuos a la ley normal. En el QQ-plot se dibujan los puntos asociados a los cuantiles de la distribución normal (estándar en R o sin estandarizar como en SPSS). En el PP-plot se dibujan las probabilidades acumuladas estimadas y teóricas para la distribución normal. En ambos casos se dibuja también una recta que representa el ajuste perfecto a la distribución normal. Los desvíos exagerados de dichas rectas indican una posible violación de la hipótesis de normalidad (figura 9.3 d).

  El estudio de la normalidad de los residuos se debe completar con algún contraste de ajuste como la prueba ji-cuadrado o el test de Kolmogorov (ver sección 9.4).

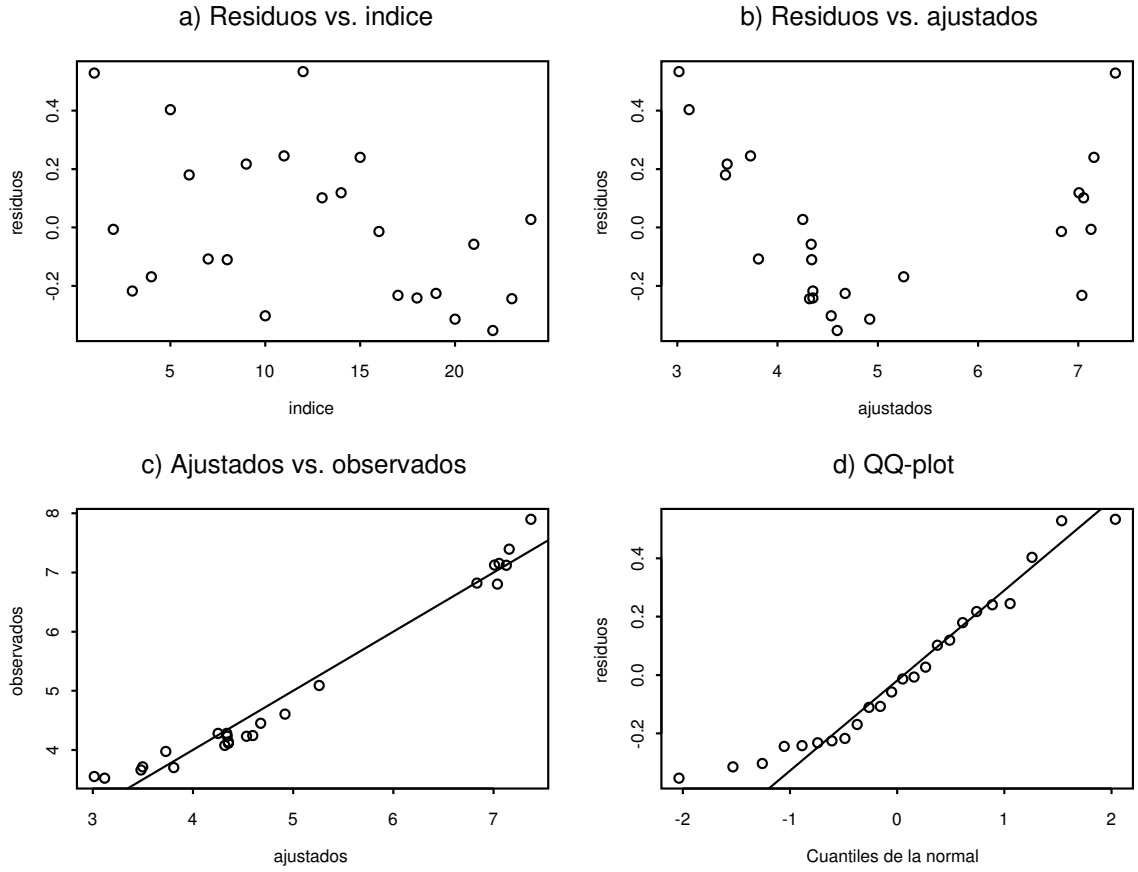


Figura 9.3: Gráficos en el análisis de la regresión simple del ejemplo 9.1.4.

Ejemplo 9.1.4

Como continuación del ejemplo de regresión simple 9.1.3 con los datos de tráfico, podemos representar algunos gráficos como los de la figura 9.3. Entre esos gráficos podemos

destacar la no aleatoriedad manifiesta del gráfico (b) que indica un ajuste no lineal entre las variables. Ello justifica la introducción del modelo parabólico (ejercicio 9.1).

9.2. Diagnóstico de la influencia

Ocasionalmente hallamos que algún dato o un pequeño subconjunto de datos ejerce una desproporcionada influencia en el ajuste del modelo de regresión. Esto es, los estimadores de los parámetros o las predicciones pueden depender más del subconjunto influyente que de la mayoría de los datos. Queremos localizar estos puntos influyentes y medir su impacto en el modelo. Si por alguna razón concreta son puntos "malos" los eliminaremos, pero si no ocurre nada extraño, su estudio puede darnos algunas claves del modelo.

9.2.1. Nivel de un punto

Casi siempre los puntos definidos por las variables regresoras o explicativas forman una nube y están razonablemente repartidos alrededor del punto medio. Sin embargo, alguno de ellos o un pequeño grupo puede aparecer muy alejado del resto. Estos valores son potencialmente peligrosos, puesto que pueden afectar excesivamente al ajuste del modelo. Vamos a definir el concepto de nivel[3] de un punto y señalaremos los que tengan un nivel muy alto (*leverage points*).

El nivel de un punto es una medida de la distancia del punto al centroide del conjunto de datos. Existen varias propuestas pero la más extendida se basa en los elementos h_{ii} de la diagonal de la matriz proyección $\mathbf{P}$. Estos elementos se calculan con las fórmulas 9.1 en el caso de la regresión simple y 9.2 para la regresión múltiple.

Como

$$\sum_{i=1}^{n} h_{ii} = \text{traza}(\mathbf{P}) = \text{rango}(\mathbf{P}) = k+1$$

el tamaño medio de cada h_{ii} es $(k+1)/n$. Así, cuando un punto verifique $h_{ii} > 2(k+1)/n$ diremos que dicha observación es un punto de alto nivel. Estos puntos se deben marcar para su posterior estudio ya que son potencialmente influyentes.

Ejemplo 9.2.1

Siguiendo con el ejemplo 9.1.1 los datos con mayor nivel son

dato	nivel
1	0,15586452
15	0,13601868
2	0,13354830

Dado que $2(k+1)/n = (2 \cdot 2)/24 = 0{,}1666$*, no hay ningún punto de alto nivel.*

[3] *leverage*

9.2.2. Influencia en los coeficientes de regresión

Entre las medidas de influencia sobre los coeficientes de regresión la más empleada es la distancia de Cook (1977,1979)

$$C_i = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{(k+1)\text{ECM}} \qquad i = 1, \ldots, n \tag{9.6}$$

donde $\widehat{\boldsymbol{\beta}}$ son las estimaciones MC en el modelo con todos los puntos, mientras que $\widehat{\boldsymbol{\beta}}_{(i)}$ son las estimaciones sin el i-ésimo punto. Esta medida calcula la distancia cuadrática entre $\widehat{\boldsymbol{\beta}}$ y $\widehat{\boldsymbol{\beta}}_{(i)}$, relativa a la geometría fija de $\mathbf{X}'\mathbf{X}$.

Otra versión equivalente de esta distancia es

$$C_i = \frac{(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{(i)})'(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{(i)})}{(k+1)\text{ECM}}$$

ya que $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ y $\widehat{\mathbf{Y}}_{(i)} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{(i)}$.

Sin embargo para el cálculo de esta distancia es mejor utilizar la fórmula

$$C_i = \frac{r_i^2}{k+1} \cdot \frac{h_{ii}}{1 - h_{ii}}$$

donde la primera parte depende del ajuste al modelo de la i-ésima predicción, mientras que el segundo factor es una función de la distancia del punto $\mathbf{x}_i$ al centroide del conjunto de observaciones de las variables explicativas. Una demostración de esta fórmula puede verse en el ejercicio 9.19 del libro de Ugarte y Militino[69].

La búsqueda de puntos influyentes se puede iniciar con la identificación de puntos con distancia de Cook elevada. Sin embargo se desconoce la distribución exacta de este estadístico y no hay reglas fijas para la determinación de los puntos con valor de C_i grande. Los puntos con distancias de Cook grandes pueden ser influyentes y podemos extraerlos del análisis para ver si los cambios son apreciables.

Ejemplo 9.2.2

Con el ejemplo de regresión simple que estamos estudiando desde el ejemplo 9.1.1 se observa que los datos con mayor distancia de Cook son:

dato	h_{ii}	r_i	C_i
1	0,1559	2,1397	0,4227
12	0,1227	2,1178	0,3136

Estos datos son los de mayor influencia debida al gran residuo studentizado (los dos mayores) y a su alto nivel, especialmente el dato 1.

Otra medida de influencia sobre cada coeficiente de regresión por separado fue propuesta por Belsley et al.[6] y consiste en la diferencia estandarizada entre la estimación MC de dicho parámetro con todas las observaciones y la estimación MC del mismo sin la i-ésima:

$$\text{Dfbetas}_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{s^2_{(i)} c_{jj}}}$$

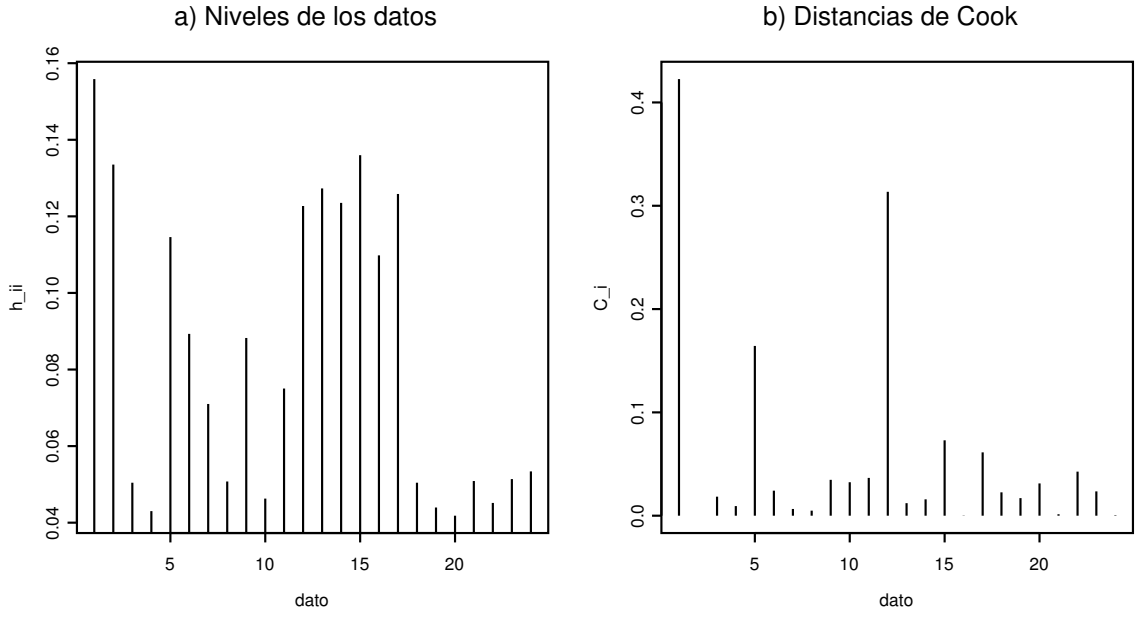


Figura 9.4: Gráficos de los niveles y distancias de Cook de los datos del ejemplo 9.2.2.

para $j = 0, 1, \ldots, k$ y $i = 1, \ldots, n$, donde c_{jj} es el j-ésimo elemento de la diagonal de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$ y $s^2_{(i)}$ la estimación MC de la varianza σ^2 sin la i-ésima observación. Observemos que $s^2_{(i)}c_{jj}$ es una estimación de la varianza $\text{var}(\hat{\beta}_j) = \sigma^2 c_{jj}$.

Un valor absoluto desmesurado de esta medida indica una gran influencia de la observación i-ésima sobre la estimación del coeficiente β_j. En la práctica se considera una observación influyente cuando $|\text{Dfbetas}| > 1$ para un pequeño conjunto de datos y $|\text{Dfbetas}| > 2/\sqrt{n}$ en general.

9.2.3. Influencia en las predicciones

Como hemos visto, la distancia de Cook es también una medida de la influencia de un punto sobre el conjunto de predicciones.

Otra medida de influencia de la i-ésima observación sobre la predicción de la propia observación i es el estadístico

$$\text{Dffits}_i = \frac{|\hat{y}_i - \hat{y}_{i(i)}|}{\sqrt{s^2_{(i)}h_{ii}}}$$

donde se estandariza la diferencia entre las predicciones de la i-ésima observación con y sin ella misma.

A partir de las ecuaciones 9.3 y 9.5 se demuestra que (ejercicio 9.3)

$$\text{Dffits}_i = |t_i|\sqrt{\frac{h_{ii}}{1 - h_{ii}}} \tag{9.7}$$

donde t_i son los residuos studentizados externamente.

En general se considera que la influencia es notable si el Dffits es superior a $2\sqrt{(k+1)/n}$, mientras que para un conjunto de datos reducido basta que sea mayor que uno.

Ejemplo 9.2.3

Como continuación del ejemplo 9.2.2 podemos calcular el Dffits_1 *para la primera observación:*

$$\text{Dffits}_1 = |2{,}349159|\sqrt{\frac{0{,}155865}{1-0{,}155865}} = 1{,}009439$$

que supera el valor frontera $2\sqrt{2/24} = 0{,}577$ *y muestra la alta influencia de esta observación.*

9.3. Selección de variables

Con el objetivo de considerar el mejor modelo de regresión posible, el experimentador debe seleccionar un conjunto de variables regresoras entre las observadas y, si es necesario, entre potencias y productos de las mismas. Una primera decisión fijará el tipo de relación funcional con la variable respuesta pero, en todo caso, la selección de un conjunto reducido de variables explicativas es un problema complicado. Si consideramos un número demasiado pequeño de variables es posible que la potencia del modelo se vea reducida y que las estimaciones obtenidas sean sesgadas, tanto de los coeficientes de regresión, como de las predicciones. Este sesgo se origina ya que los errores calculados con los datos observados pueden contener efectos no aleatorios de las variables desechadas. Por otra parte, un número muy grande de variables explicativas complica la utilidad práctica del modelo y, aunque mejora el ajuste aparente, aumenta la varianza de los estimadores de los parámetros.

Decidir el mejor conjunto de variables es prácticamente un arte, en el que algunas técnicas sirven de apoyo: test t de Student de los coeficientes de regresión, test F de significación de la regresión, estudio de la multicolinealidad, etc. Sin embargo, ya hemos alertado sobre la utilización ciega de los test t parciales para medir la importancia de las variables. Así pues, es preciso añadir algunas técnicas específicas para comparar modelos de regresión que pasamos a detallar.

9.3.1. Coeficiente de determinación ajustado

Esta técnica consiste en calcular los coeficientes de determinación de todos los modelos posibles con la combinación de cualquier número de variables explicativas. Para evitar los problemas que justifican la definición 8.2.1 resulta obvio utilizar el coeficiente ajustado cuando hay muchas variables en juego. El objetivo es reconocer el modelo con mayor coeficiente. Sin embargo, si el número de variables es considerable esta técnica puede tener dificultades de cálculo.

9.3.2. Criterio C_P de Mallows

Con este criterio se debe fijar en primera instancia un número P de parámetros, incluido el término independiente, aunque con posterioridad se podrá variar. Se trata de hallar el mejor modelo con P variables explicativas, incluida la constante, utilizando el estadístico de Mallows

$$C_P = \frac{\text{SCR}_P}{\hat{\sigma}^2} - (n - 2P)$$

donde SCR_P es la suma de cuadrados residual del modelo particular y $\hat{\sigma}^2$ un estimador de la varianza del modelo que acostumbra a ser el ECM del modelo completo.
Para el modelo completo $P = k + 1$, el estadístico de Mallows es

$$C_{k+1} = \frac{\mathrm{SCR}}{\mathrm{ECM}} - (n - 2(k+1)) = n - (k+1) - (n - 2(k+1)) = k + 1$$

También para todo modelo no completo se puede demostrar que aproximadamente $E(C_P) = P$, si el modelo es adecuado. En consecuencia parece recomendable elegir los conjuntos para los que C_P sea aproximadamente P.

9.3.3. Selección paso a paso

El procedimiento se puede realizar hacia adelante (forward stepwise) o hacia atrás (backward stepwise), seleccionando las variables una a una e incorporándolas desde el modelo inicial o eliminándolas desde el modelo completo en función de su contribución al modelo. Aunque es el método más utilizado por su facilidad de computación, este sistema tiene el inconveniente de que puede conducir a modelos distintos y no necesariamente óptimos.
En la selección hacia adelante se incorpora como primera variable la de mayor F de significación de la regresión simple. La segunda variable se selecciona por su mayor contribución al modelo que ya contiene la primera variable del paso anterior y así sucesivamente.

9.4. Ejemplos con R

Con los datos de tráfico de la sección 1.2 se calcula la regresión como se explica en la sección 6.9 mediante la instrucción

```
> recta<-lm(rvel~dens)
```

Para el análisis de los residuos, la función `summary` nos ofrece un resumen de cinco números

```
Call: lm(formula = rvel ~ dens)
Residuals:
     Min      1Q   Median     3Q    Max
 -0.3534 -0.2272 -0.03566 0.1894 0.5335
```

También podemos obtener algunos gráficos univariantes como los de la figura 9.5 con las siguientes instrucciones:

```
> par(mfrow=c(1,2))
> par(pty="s")
> hist(residuals(recta),xlab="residuos")
> title("a) Histograma")
> boxplot(residuals(recta))
> title("b) Diagrama de caja")
> stem(residuals(recta))

N = 24   Median = -0.0356607
Quartiles = -0.228869, 0.1987335
```

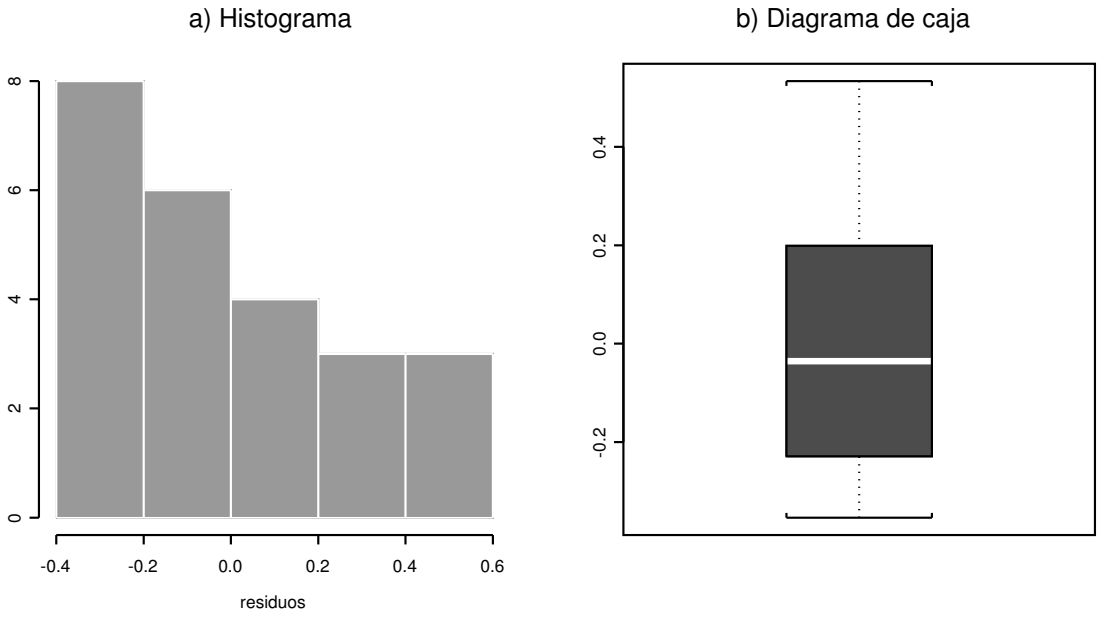


Figura 9.5: Gráficos de los residuos de la regresión simple del ejemplo de la sección 1.2.

```
Decimal point is 1 place to the left of the colon

  -3 : 510
  -2 : 44332
  -1 : 711
  -0 : 611
   0 : 3
   1 : 028
   2 : 245
   3 :
   4 : 0
   5 : 33
```

Para obtener los gráficos de la figura 9.3 se requieren las siguientes instrucciones:

```
> par(mfrow=c(2,2))
> plot(residuals(recta),xlab="indice",ylab="residuos")
> title("a) Residuos vs. indice")
> plot(fitted(recta),residuals(recta),xlab="ajustados",ylab="residuos")
> title("b) Residuos vs. ajustados")
> plot(fitted(recta),rvel,xlab="ajustados",ylab="observados")
> abline(0,1)
> title("c) Ajustados vs. observados")
> qqnorm(residuals(recta),xlab="Cuantiles de la normal",ylab="residuos")
> qqline(residuals(recta))
> title("d) QQ-plot")
```

R también permite obtener 6 gráficos para el análisis de un modelo de regresión lineal de una forma directa, mediante las instrucciones

```
> par(mfrow=c(2,3))
> plot(recta)
```

En cuanto a los contrastes de ajuste a la distribución normal, podemos optar entre el test de Kolmogorov-Smirnov `ks.gof` y la prueba ji-cuadrado `chisq.gof`. En nuestro caso:

```
> ks.gof(residuals(recta), distribution = "normal")

    One sample Kolmogorov-Smirnov Test of Composite Normality

data:  residuals(recta)
ks = 0.129, p-value = 0.5 alternative
hypothesis: True cdf is not the normal distn. with estimated parameters
sample estimates:
     mean of x standard deviation of x
 2.298509e-017               0.2630273
```

También se puede calcular la regresión con la instrucción

```
recta.ls<-lsfit(dens,rvel)
```

que nos proporciona muchos de los elementos para el diagnóstico en la forma:

```
> recta.diag<-ls.diag(recta.ls)
> recta.diag$hat # nivel
...
> recta.diag$std.res # residuos studentizados
...
> recta.diag$stud.res # residuos studentizados externamente
...
> recta.diag$cooks # distancias de Cook
...
> recta.diag$dfits # medidas Dffits
...
```

Los gráficos ...

```
> par(mfrow=c(1,2))
> par(pty="s")
> plot(recta.diag$hat,type="h",xlab="dato",ylab="h_ii")
> title("a) Niveles de los datos")
> plot(recta.diag$cooks,type="h",xlab="dato",ylab="C_i")
> title("b) Distancias de Cook")

> par(mfrow=c(1,2))
> par(pty="s")
> plot(recta.diag$std.res,xlab="dato",ylab="r_i",ylim=c(-2.5,2.5))
> title("a) Residuos studentizados \n internamente")
> plot(recta.diag$stud.res,xlab="dato",ylab="t_i",ylim=c(-2.5,2.5))
> title("b) Residuos studentizados \n externamente")
```

9.5. Ejercicios

Ejercicio 9.1

Realizar el análisis completo de los residuos del modelo de regresión parabólico propuesto en la sección 1.2 con los datos de tráfico.

Ejercicio 9.2

Realizar el análisis completo de los residuos de los modelos de regresión simple y parabólico propuestos en la sección 1.2 con los datos de tráfico, pero tomando como variable respuesta la velocidad (sin raíz cuadrada). Este análisis debe justificar la utilización de la raíz cuadrada de la velocidad como variable dependiente.

Ejercicio 9.3

Probar la relación 9.7 a partir de las ecuaciones 9.3 y 9.5.

Ejercicio 9.4

Se define el coeficiente de robustez como

$$B^2 = \frac{\text{SCR}}{\text{PRESS}}$$

donde PRESS es la suma de cuadrados 9.4. Este coeficiente está entre 0 y 1 y representa una medida de la robustez del modelo.

Calcular el coeficiente de robustez para los cinco conjuntos de datos de la sección 6.8.

Capítulo 10

Análisis de la Varianza

10.1. Introducción

El Análisis de la Varianza es un conjunto de técnicas estadístico-matemáticas que permiten analizar cómo operan sobre una variable respuesta diversos factores considerados simultáneamente según un determinado *diseño factorial*. Normalmente interesa estudiar cómo se diferencian los *niveles* de un cierto *factor*, llamado factor tratamiento, teniendo en cuenta la incidencia de otros factores cualitativos o cuantitativos (factores ambientales), cuya influencia es eliminada mediante una adecuada descomposición de la variabilidad de la variable observada. También se pretende detectar la relevancia en el resultado de las variables o factores influyentes, es decir, estudiar la causalidad.

La variable respuesta se considera del tipo continuo, mientras que las variables experimentales o factores son variables categóricas o categorizadas en niveles. Un experimento de este tipo consiste en tomar una *unidad experimental* o elemento muestral, fijar los valores de los factores a distintos niveles y observar el valor de la variable respuesta en cada caso. Ahora bien, para llegar a conclusiones estadísticas correctas es preciso, en la mayoría de los problemas, observar el resultado tras la repetición del experimento en varias unidades experimentales para cada una de las diversas condiciones que indica el diseño pero lo más homogéneas posibles dentro de cada una. Esto redundará en la reducción de la variabilidad y, por tanto, aumentará la capacidad estadística de detectar cambios o identificar variables influyentes. Con una variabilidad muy grande respecto al error experimental no se pueden detectar diferencias entre tratamientos.

Como ocurre con la varianza de la media muestral, para reducir la variabilidad es posible tomar un pequeño número de observaciones llamadas *réplicas* en condiciones totalmente homogéneas o aumentar el número de observaciones. Esto último es preciso cuando tomamos observaciones fuera del laboratorio o con variables influyentes que escapan a nuestro control.

Es muy importante que las réplicas sean exactamente eso, es decir, repeticiones del experimento en las mismas condiciones y no repeticiones de la observación que pueden dar lugar a observaciones dependientes. Así pues, debemos repetir todo el experimento desde el principio para cada una de las observaciones.

Como ya hemos dicho, para investigar el efecto del factor principal o tratamiento es posible que debamos considerar y eliminar los efectos de muchas variables que influyen en el resultado. Para eliminar el efecto de una variable sobre el resultado del experimento tenemos tres opciones: a) fijar el valor de la variable para toda la investigación y restringir la validez de nuestras conclusiones a ese dato; b) diseñar el experimento de manera

que dicha variable aparezca como factor con unos determinados valores o niveles y c) aleatorizar su aparición en cada condición experimental. Las dos primeras opciones son propias del laboratorio y dependen del experimentador. La tercera resulta útil cuando queremos eliminar el efecto de una variable no directamente controlable y de poca influencia esperada, así la parte de la variabilidad que le corresponde se incluirá en el error experimental.

Para diseñar correctamente un experimento es preciso trabajar bajo el *principio de aleatorización*. Este principio consiste en tomar las observaciones de las réplicas asignando al azar todos los factores no directamente controlados por el experimentador y que pueden influir en el resultado. En el ejemplo 10.2.1 la comparación entre tres tratamientos se hace con pacientes con ciertas condiciones de homogeneidad pero asignando los pacientes al azar a cada tratamiento. Con la aleatorización se consigue prevenir sesgos, evitar la dependencia entre observaciones y validar estadísticamente los resultados. En particular, debemos aleatorizar el orden de realización de los experimentos.

En resumen, es necesario que el experimento esté bien diseñado mediante el control físico, fijando niveles, o estadístico, mediante la aleatorización, de todas las variables o factores relevantes. Así se garantizará que las diferencias se deben a las condiciones experimentales fijadas el diseño y se podrá concluir estadísticamente una relación causal.

Además, en Peña[54, pág. 82] se muestra cómo la aleatorización permite la comparación de medias mediante los llamados tests de permutaciones que no requieren ningún tipo de hipótesis sobre la distribución del error. Por otra parte, puede demostrarse (ver Scheffé[63]) que los contrastes F son una buena aproximación a los contrastes de permutaciones, de manera que la aleatorización justifica la utilización de la teoría de los modelos lineales bajo hipótesis de normalidad, aunque dicha hipótesis no esté plenamente validada.

Para comparar tratamientos es necesario hacerlo en condiciones homogéneas y para ello se deben introducir en el diseño todas las variables que pueden influir, para luego promediar la respuesta en situaciones homogéneas. Una vez fijados los factores, la idea básica de los diseños factoriales es cruzar los niveles de los factores y considerar todas las situaciones. También cuando los efectos de los factores no son puramente aditivos se puede introducir el efecto de las llamadas interacciones.

En general, en todo Análisis de la Varianza es necesario considerar tres etapas:

a) Diseño del experimento a fin de obtener observaciones de una variable Y, combinando adecuadamente los factores incidentes.

b) Planteo de hipótesis, cálculo de sumas de cuadrados (residuales, de desviación de la hipótesis, etc.) y obtención de los cocientes F. Esta parte del análisis se formula mediante la teoría de los modelos lineales.

c) Toma de decisiones e interpretación de los resultados. Planteamiento "a posteriori" de nuevas hipótesis.

En Ugarte[69, sec. 8.2] puede consultarse un buen resumen de las estructuras básicas de un diseño de experimentos.

10.2. Diseño de un factor

10.2.1. Comparación de medias

Supongamos que una variable Y ha sido observada bajo k condiciones experimentales distintas. Puede ser que las observaciones provengan de k poblaciones, o bien tratarse de réplicas para cada uno de los k niveles de un factor.

Indiquemos por y_{ih} la réplica h ($h = 1, \ldots, n_i$) en la población o nivel i ($i = 1, \ldots, k$), donde n_i es el número de réplicas en la población i. El conjunto de datos es:

$$\begin{array}{ll} \text{Nivel } 1 & y_{11}, y_{12}, \ldots, y_{1n_1} \\ \text{Nivel } 2 & y_{21}, y_{22}, \ldots, y_{2n_2} \\ \vdots & \\ \text{Nivel } k & y_{k1}, y_{k2}, \ldots, y_{kn_k} \end{array}$$

Con estos datos podemos calcular algunas medias que indicaremos de la siguiente forma:

$$\text{Media en la población } i \text{ o nivel } i: \quad y_{i\cdot} = \frac{1}{n_i} \sum_{h=1}^{n_i} y_{ih}$$

$$\text{Media general:} \quad \bar{y} = y_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{k} \sum_{h=1}^{n_i} y_{ih}$$

donde $n = \sum_{i=1}^{k} n_i$ es el número total de observaciones.

El modelo lineal que se adapta a este diseño es

$$y_{ih} = \mu_i + \epsilon_{ih} \qquad i = 1, \ldots, k \ ; \ h = 1, \ldots, n_i \tag{10.1}$$

siendo $(\mu_1, \mu_2, \ldots, \mu_k)'$ el vector de parámetros y

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} \qquad \text{rango } \mathbf{X} = k$$

la matriz de diseño (reducida).

Recordemos en este momento que asumir un modelo lineal significa aceptar las condiciones de Gauss-Markov (ver sección 1.5) y además, en este caso y en todo el capítulo, aceptar la distribución normal de los errores $N(0, \sigma)$. Entonces, se comprueba fácilmente que la estimación MC de los parámetros es

$$\hat{\mu}_i = y_{i\cdot} \qquad i = 1, \ldots, k$$

Luego los residuos de este modelo son

$$e_{ih} = \text{observación} - \text{predicción} = y_{ih} - \hat{\mu}_i$$

de modo que la suma de cuadrados residual resulta

$$\text{SCR} = \sum_{i=1}^{k} \sum_{h=1}^{n_i} (y_{ih} - y_{i\cdot})^2$$

Esta suma se indica por SC_D y se denomina *suma de cuadrados dentro de grupos* o también *intragrupos*.
Consideremos la identidad

$$y_{ih} - \bar{y} = (y_{i\cdot} - \bar{y}) + (y_{ih} - y_{i\cdot})$$

Elevando al cuadrado y sumando tenemos

$$\sum_{i,h}(y_{ih} - \bar{y})^2 = \sum_{i,h}(y_{i\cdot} - \bar{y})^2 + \sum_{i,h}(y_{ih} - y_{i\cdot})^2 + 2\sum_{i,h}(y_{i\cdot} - \bar{y})(y_{ih} - y_{i\cdot})$$

pero

$$\sum_{i,h}(y_{i\cdot} - \bar{y})(y_{ih} - y_{i\cdot}) = \sum_{i,h}(y_{ih} - y_{i\cdot})y_{i\cdot} - \sum_{i,h}(y_{ih} - y_{i\cdot})\bar{y} = 0$$

En efecto, el vector $\{y_{ih} - y_{i\cdot}\}$ pertenece al espacio error y por tanto es ortogonal al vector $\{y_{i\cdot}\}$ que pertenece al espacio estimación como hemos visto en 2.4.2; por otra parte

$$\sum_{i,h}(y_{ih} - y_{i\cdot}) = 0$$

Así pues, con la siguiente notación

$$\mathrm{SC}_T = \sum_{i,h}(y_{ih} - \bar{y})^2 \qquad \text{suma de cuadrados total}$$
$$\mathrm{SC}_E = \sum_{i} n_i(y_{i\cdot} - \bar{y})^2 \quad \text{suma de cuadrados entre grupos}$$

hemos probado que se verifica la identidad

$$\mathrm{SC}_T = \mathrm{SC}_E + \mathrm{SC}_D \tag{10.2}$$

Esta igualdad muestra la descomposición de la variabilidad total que también se puede expresar en términos de variabilidad explicada y no explicada como en la ecuación 6.7.
La hipótesis nula de mayor interés es

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

Si H_0 es cierta, las medias de las k poblaciones son iguales o, en términos de diseño factorial, los niveles del factor no son significativos para la variable observable. Entonces, el modelo 10.1 se reduce a la forma

$$y_{ih} = \mu + \epsilon_{ih} \qquad i = 1, \ldots, k \ ; \ h = 1, \ldots, n_i$$

La estimación MC de μ es $\hat{\mu} = \bar{y}$ y la suma de cuadrados residual es

$$\mathrm{SCR}_H = \sum_{i,h}(y_{ih} - \bar{y})^2 = \mathrm{SC}_T$$

Considerando la relación 10.2 deducimos que la suma de cuadrados debida a la desviación de la hipótesis es

$$\mathrm{SCR}_H - \mathrm{SCR} = \sum_i n_i (y_{i\cdot} - \bar{y})^2 = \mathrm{SC}_E$$

Obsérvese que SC_E mide la variabilidad entre las medias $y_{1\cdot}, y_{2\cdot}, \ldots, y_{k\cdot}$.

Por otra parte y según el teorema 2.5.1, una estimación insesgada del error experimental σ^2 es

$$\hat{\sigma}^2 = \mathrm{SC}_D/(n-k)$$

Además, gracias a la hipótesis de normalidad $\epsilon_{ih} \sim N(0, \sigma)$ se verifica (ver teorema 5.3.1):

a) $\mathrm{SC}_D/\sigma^2 \sim \chi^2_{n-k}$

b) Si H_0 es cierta, entonces $\mathrm{SC}_E/(k-1)$ es otra estimación insesgada de σ^2 y además

$$\mathrm{SC}_E/\sigma^2 \sim \chi^2_{k-1}$$

c) Si H_0 es cierta, el estadístico

$$F = \frac{\mathrm{SC}_E/(k-1)}{\mathrm{SC}_D/(n-k)} \tag{10.3}$$

sigue la distribución F con $k-1$ y $n-k$ grados de libertad.

La hipótesis H_0 de igualdad de medias se rechaza si 10.3 es significativo. En todo caso es recomendable disponer los cálculos de la forma indicada en la tabla 10.1.

Fuente de variación	suma de cuadrados	g.l.	cuadrados medios	F
Entre grupos	$\mathrm{SC}_E = \sum_i n_i (y_{i\cdot} - \bar{y})^2$	$k-1$	$\mathrm{SC}_E/(k-1)$	$\frac{\mathrm{SC}_E/(k-1)}{\mathrm{SC}_D/(n-k)}$
Dentro grupos	$\mathrm{SC}_D = \sum_{i,h} (y_{ih} - y_{i\cdot})^2$	$n-k$	$\mathrm{SC}_D/(n-k)$	
Total	$\mathrm{SC}_T = \sum_{i,h} (y_{ih} - \bar{y})^2$	$n-1$		

Cuadro 10.1: Tabla del Análisis de la Varianza para diseños de un factor

También se puede calcular el coeficiente de determinación como medida de la proporción de la variabilidad explicada por los grupos

$$R^2 = \frac{\mathrm{SC}_E}{\mathrm{SC}_T}$$

10.2.2. Un modelo equivalente

El modelo 10.1 no se puede extender al caso de varios factores. Sin embargo, se puede reparametrizar en la forma

$$y_{ih} = \mu + \alpha_i + \epsilon_{ih} \qquad i = 1, \ldots, k \ ; \ h = 1, \ldots, n_i \tag{10.4}$$

con la restricción

$$\sum_{i=1}^{k} \alpha_i = 0$$

Esta restricción es necesaria para determinar el cálculo de los $k+1$ parámetros en un modelo de rango k.

El modelo 10.4 también representa el diseño de un factor a k niveles, pero con la siguiente interpretación de los parámetros

$$\mu = \text{media general}$$
$$\alpha_i = \text{efecto del nivel } i$$

La hipótesis H_0 de igualdad entre niveles o tratamientos, antes igualdad de medias, se expresa ahora así

$$H_0 : \alpha_1 = \cdots = \alpha_k = 0$$

Las estimaciones de μ y α_i son

$$\hat{\mu} = \bar{y} \qquad\qquad \hat{\alpha}_i = y_{i\cdot} - \bar{y}$$

Se verifica entonces

$$\mathrm{SCR}_H - \mathrm{SCR} = \mathrm{SC}_E = \sum_i n_i \hat{\alpha}_i^2$$

de modo que SC_E refleja bien la variabilidad entre los diferentes niveles del factor estudiado.

La formulación matricial de H_0 es

$$\mathbf{A}\boldsymbol{\beta} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{k-1} \\ \alpha_k \end{pmatrix} = 0$$

Aplicando entonces 5.7, tenemos que

$$E(\mathrm{SCR}_H - \mathrm{SCR}) = E(\mathrm{SC}_E) = (k-1)\sigma^2 + \sum_i n_i \alpha_i^2 \tag{10.5}$$

y como ya sabíamos, si es cierta la hipótesis H_0, el estadístico $\mathrm{SC}_E/(k-1)$ es otro estimador insesgado de σ^2.

En todo caso, como se trata de una reparametrización, el contraste de H_0 se realiza exactamente con la misma tabla 10.1 y el mismo estadístico F de 10.3.

Finalmente, si se desean comparar dos niveles, es decir, plantear la hipótesis parcial

$$H_0^{(ij)} : \alpha_i = \alpha_j$$

utilizaremos el estadístico

$$t = \frac{y_{i\cdot} - y_{j\cdot}}{\sqrt{\mathrm{SC}_D/(n-k)}} \sqrt{\frac{n_i n_j}{n_i + n_j}} \tag{10.6}$$

que bajo $H_0^{(ij)}$ sigue una t de Student con $n-k$ grados de libertad. Con más generalidad, si se desea estudiar si la función paramétrica estimable, tal que $c_1 + \cdots + c_k = 0$,

$$\psi = c_1\alpha_1 + \cdots + c_k\alpha_k$$

se aparta significativamente de 0, utilizaremos

$$t = \frac{\sum_i c_i y_{i\cdot}}{\sqrt{\sum_i c_i^2/n_i}\sqrt{\mathrm{SC}_D/(n-k)}} \tag{10.7}$$

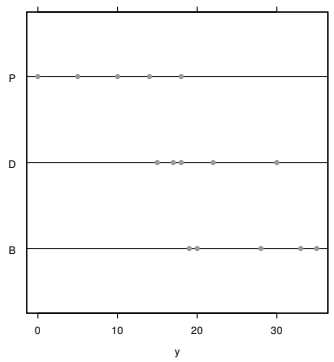
también con $n-k$ grados de libertad (ver fórmula 3.4). Del mismo modo se pueden construir intervalos de confianza para las funciones paramétricas estimables $\psi = c_1\alpha_1 + \cdots + c_k\alpha_k$ y en particular para $\alpha_i - \alpha_j$.

Otro aspecto mucho más complejo es la consideración de varias de estas hipótesis de forma conjunta. Es lo que se llama el problema de las comparaciones múltiples o intervalos simultáneos como en la sección 6.3.6.

Ejemplo 10.2.1

Se desean comparar dos medicamentos D (diurético), B (betabloqueante) con un producto inocuo P (placebo). Se tomó una muestra de 15 *individuos hipertensos cuyas condiciones iniciales eran suficientemente homogéneas y se asignaron los tres tratamientos al azar. El objetivo del estudio es ver cómo actúan los tres tratamientos frente a la hipertensión, concretamente si disminuyen la misma. A tal fin se ha elegido la variable observable "porcentaje de descenso de la presión arterial media". Los datos obtenidos se presentan en la tabla 10.2.*

D	B	P
22	20	10
18	28	5
30	35	0
15	19	14
17	33	18



Cuadro 10.2: Datos de los pacientes según el tratamiento

Vamos a estudiar si hay diferencias significativas entre los tres fármacos y la significación de la función paramétrica

$$\psi = \frac{1}{2}(D+B) - P$$

que se puede interpretar como una medida de la diferencia entre los productos activos respecto al placebo.

Las medias son:

$$y_{1\cdot} = 20{,}40 \qquad y_{2\cdot} = 27{,}00 \qquad y_{3\cdot} = 9{,}40 \qquad \bar{y} = 18{,}93$$

Las sumas de cuadrados son:

$$\mathrm{SC}_T = 1349{,}93 \qquad \mathrm{SC}_E = 790{,}53 \qquad \mathrm{SC}_D = 558{,}40$$

de manera que podemos disponer las estimaciones en forma de tabla del Análisis de la Varianza como se muestra en la tabla 10.3.

Fuente de variación	suma de cuadrados	g.l.	cuadrados medios	F
Entre fármacos	790,53	2	395,27	8,49
Dentro fármacos	558,40	12	46,53	
Total	1349,93	14		

Cuadro 10.3: Ejemplo de Análisis de la Varianza para un diseño de un factor

Con $2, 12$ *grados de libertad y un nivel de significación del* 0,01 *leemos en la tabla de la distribución* F *el valor* 6,93*. Luego la diferencia entre los tres fármacos es claramente significativa.*
La estimación de Gauss-Markov de la función paramétrica es

$$\hat{\psi} = \frac{1}{2}(20{,}40 + 27{,}00) - 9{,}40 = 14{,}30$$

Además

$$\sum_i c_i^2/n_i = \frac{1}{5}(\frac{1}{4} + \frac{1}{4} + 1) = 0{,}3$$
$$\mathrm{SC}_D/(n-k) = 46{,}53$$

Aplicando 10.7 obtenemos

$$t = \frac{14{,}30}{\sqrt{0{,}3}\sqrt{46{,}53}} = 3{,}827$$

Contrastando con la tabla de la t *de Student, para* 12 *grados de libertad, vemos que* ψ *es significativa al nivel* 0,01*. Finalmente, para analizar si hay diferencias significativas entre* D *y* B*, utilizaremos 10.6*

$$t = \frac{20{,}40 - 27{,}00}{\sqrt{46{,}53}}\sqrt{\frac{5 \times 5}{5+5}} = -1{,}530$$

que no es significativa.
Conclusión: Hay variabilidad significativa entre los tres fármacos. La variabilidad reside principalmente en la diferencia entre los dos fármacos activos frente al placebo.

10.3. Diseño de dos factores sin interacción

Una variable o factor cuyo efecto sobre la respuesta no es directamente de interés pero que se introduce en el experimento para obtener comparaciones homogéneas se denomina una variable *bloque*. Por ejemplo, en una investigación para comparar la efectividad de varios fertilizantes (tratamientos) se puede considerar las fincas donde se prueban como

un factor bloque (ver ejemplo 10.3.1). El efecto de la finca sobre la producción no es de interés y el objetivo es comparar los fertilizantes eliminando el efecto de la pertenencia de una cosecha a una finca. Además, en general, se supone que no hay interacción entre la variable bloque y los factores de interés.

En este tipo de diseños los tratamientos se asignan aleatoriamente a un grupo de unidades experimentales en cada bloque o nivel de la variable bloque. Para poder detectar diferencias entre los tratamientos es importante que haya diferencias entre los bloques, mientras que las unidades experimentales dentro de cada bloque han de ser muy homogéneas. Un buen resumen de las características más importantes del diseño en bloques puede verse en Ugarte[69, pág. 405].

En este apartado vamos a tratar el caso más simple del llamado *diseño completamente aleatorizado por bloques* o, más brevemente, diseño en bloques aleatorizados con un factor principal y una variable bloque.

Supongamos que la variable respuesta está afectada por dos causas de variabilidad, es decir, por dos variables o factores cualitativos A y B, con a y b niveles respectivamente. El factor A es el factor principal, mientras que el factor B es una variable bloque. Supongamos también que tenemos únicamente una observación por casilla o combinación de niveles. Eso significa tener tantas unidades experimentales por bloque como tratamientos o niveles del factor principal y que la asignación del tratamiento se hace al azar en cada bloque (ver ejemplo 10.3.1). Entonces, podemos disponer las observaciones del siguiente modo

	B_1	B_2	$\dots$	B_b	
A_1	y_{11}	y_{12}	$\dots$	y_{1b}	$y_{1\cdot}$
A_2	y_{21}	y_{22}	$\dots$	y_{2b}	$y_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_a	y_{a1}	y_{a2}	$\dots$	y_{ab}	$y_{a\cdot}$
	$y_{\cdot 1}$	$y_{\cdot 2}$	$\dots$	$y_{\cdot b}$	$y_{\cdot\cdot}$

siendo

$$y_{i\cdot} = \frac{1}{b}\sum_j y_{ij} \qquad y_{\cdot j} = \frac{1}{a}\sum_i y_{ij} \qquad y_{\cdot\cdot} = \bar{y} = \frac{1}{ab}\sum_{i,j} y_{ij}$$

En relación a la tabla de datos anterior, diremos que A es el factor fila y B el factor columna con $A_1, A_2, \dots, A_a$ y $B_1, B_2, \dots, B_b$ niveles respectivamente.

Modelo aditivo

Si suponemos que tanto el efecto fila como el efecto columna son aditivos, admitiremos el modelo lineal

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \qquad i = 1, \dots, a \ ; \ j = 1, \dots, b \tag{10.8}$$

siendo

$$\begin{aligned} \mu &= \text{media general} \\ \alpha_i &= \text{efecto del nivel } A_i \text{ del factor } A \\ \beta_j &= \text{efecto del nivel } B_j \text{ del factor } B \end{aligned}$$

Como 10.8 no es un diseño de rango máximo, impondremos las siguientes restricciones naturales

$$\sum_i \alpha_i = \sum_j \beta_j = 0 \tag{10.9}$$

Entonces, el modelo depende de los parámetros

$$\mu, \alpha_1, \ldots, \alpha_{a-1}, \beta_1, \ldots, \beta_{b-1}$$

siendo

$$\alpha_a = -\alpha_1 - \cdots - \alpha_{a-1} \qquad \beta_b = -\beta_1 - \cdots - \beta_{b-1}$$

Por ejemplo, la matriz de diseño reducida $\mathbf{X}$ para el caso $a = 3$, $b = 2$ es

μ	α_1	α_2	β_1
1	1	0	1
1	0	1	1
1	−1	−1	1
1	1	0	−1
1	0	1	−1
1	−1	−1	−1

Como las columnas de $\mathbf{X}$ correspondientes a parámetros con distinta letra son ortogonales, mientras que las correspondientes a parámetros con la misma letra son linealmente independientes, deducimos que el rango de $\mathbf{X}$ es igual al número de parámetros resultantes después de imponer las restricciones 10.9, es decir,

$$\text{rango } \mathbf{X} = 1 + (a-1) + (b-1) = a + b - 1 \tag{10.10}$$

Estimación de parámetros

Consideremos la identidad

$$\begin{aligned} y_{ij} - \mu - \alpha_i - \beta_j &= (\bar{y} - \mu) + (y_{i\cdot} - \bar{y} - \alpha_i) + (y_{\cdot j} - \bar{y} - \beta_j) \\ &\quad + (y_{ij} - y_{i\cdot} - y_{\cdot j} + \bar{y}) \end{aligned}$$

Elevando al cuadrado, sumando para todo i, j y teniendo en cuenta 10.9, como los productos cruzados se anulan (puede probarse con algo de esfuerzo), obtenemos

$$\begin{aligned} \sum (y_{ij} - \mu - \alpha_i - \beta_j)^2 &= \sum (\bar{y} - \mu)^2 + \sum (y_{i\cdot} - \bar{y} - \alpha_i)^2 \\ &\quad + \sum (y_{\cdot j} - \bar{y} - \beta_j)^2 \\ &\quad + \sum (y_{ij} - y_{i\cdot} - y_{\cdot j} + \bar{y})^2 \end{aligned} \tag{10.11}$$

Entonces 10.11, con las restricciones 10.9, alcanza su mínimo para

$$\hat{\mu} = \bar{y} \qquad \hat{\alpha}_i = y_{i\cdot} - \bar{y} \qquad \hat{\beta}_j = y_{\cdot j} - \bar{y} \tag{10.12}$$

de modo que la suma de cuadrados residual es

$$\text{SCR} = \sum_{i,j} (y_{ij} - y_{i\cdot} - y_{\cdot j} + \bar{y})^2 \tag{10.13}$$

Obsérvese que

$$y_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + e_{ij}$$

siendo e_{ij} la *estimación* del término de error

$$e_{ij} = y_{ij} - y_{i\cdot} - y_{\cdot j} + \bar{y}$$

Finalmente, SCR tiene $n - r = ab - (a + b - 1) = (a - 1)(b - 1)$ grados de libertad, luego

$$\hat{\sigma}^2 = \mathrm{SCR}/[(a - 1)(b - 1)]$$

es un estimador centrado de la varianza del diseño.

Hipótesis lineales

La hipótesis de que el factor principal A no es significativo (no hay efecto fila) es

$$H_0^A : \alpha_1 = \cdots = \alpha_a = 0 \tag{10.14}$$

Análogamente, la hipótesis para B (no hay efecto columna), es

$$H_0^B : \beta_1 = \cdots = \beta_b = 0 \tag{10.15}$$

El rango de H_0^A es $a - 1$, mientras que el de H_0^B es $b - 1$.
Vamos a obtener el test F adecuado para contrastar la hipótesis 10.15. Consideremos la siguiente descomposición fundamental de la suma de cuadrados (que demostraremos más adelante)

$$\sum_{i,j}(y_{ij} - \bar{y})^2 = b\sum_i (y_{i\cdot} - \bar{y})^2 + a\sum_j (y_{\cdot j} - \bar{y})^2 + \sum_{i,j}(y_{ij} - y_{i\cdot} - y_{\cdot j} + \bar{y})^2$$

$$\mathrm{SC}_T = \mathrm{SC}_F + \mathrm{SC}_C + \mathrm{SCR} \tag{10.16}$$

donde SC_T es la suma de cuadrados total, SC_F la suma de cuadrados entre filas, SC_C la suma de cuadrados entre columnas y SCR la suma de cuadrados residual (ver cuadro 10.4). La suma de cuadrados residual bajo el modelo 10.8 es 10.13.
Ahora bien, si la hipótesis 10.15 es cierta, el modelo se reduce a la forma

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

que corresponde al modelo de un solo factor. La suma de cuadrados residual (ver sección 10.2) será entonces

$$\mathrm{SCR}_H = \sum_{i,j}(y_{ij} - y_{i\cdot})^2$$

puesto que para cada i, las observaciones $y_{i1}, \ldots, y_{ib}$ hacen el papel de réplicas. Pero de la identidad

$$y_{ij} - y_{i\cdot} = (y_{\cdot j} - \bar{y}) + (y_{ij} - y_{i\cdot} - y_{\cdot j} + \bar{y})$$

elevando al cuadrado y teniendo en cuenta que los productos cruzados también se anulan, deducimos

$$\mathrm{SCR}_H = \mathrm{SC}_C + \mathrm{SCR}$$

Luego podemos decidir si puede aceptarse o no la hipótesis 10.15 utilizando el estadístico

$$F = \frac{\mathrm{SC}_C/(b-1)}{\mathrm{SCR}/[(a-1)(b-1)]} \tag{10.17}$$

cuya distribución bajo H_0 es F con $b-1$ y $(a-1)(b-1)$ grados de libertad.

Análogamente se procede para estudiar el efecto fila. Así pues, gracias a la descomposición fundamental 10.16 es posible contrastar las dos hipótesis 10.14 y 10.15 con los mismos cálculos que deben disponerse en forma de tabla (ver tabla 10.4).

Fuente de variación	suma de cuadrados	g.l.	cuadrados medios	F
Entre filas	$\mathrm{SC}_F = b\sum_i (y_{i\cdot} - \bar{y})^2$	$a-1$	$\mathrm{SC}_F/(a-1)$	$\frac{\mathrm{SC}_F/(a-1)}{\mathrm{SCR}/[(a-1)(b-1)]}$
Entre col.	$\mathrm{SC}_C = a\sum_j (y_{\cdot j} - \bar{y})^2$	$b-1$	$\mathrm{SC}_C/(b-1)$	$\frac{\mathrm{SC}_C/(b-1)}{\mathrm{SCR}/[(a-1)(b-1)]}$
Residuo	$\mathrm{SCR} = \sum_{i,j}(y_{ij} - y_{i\cdot} - y_{\cdot j} + \bar{y})^2$	$(a-1)(b-1)$	$\frac{\mathrm{SCR}}{(a-1)(b-1)}$	
Total	$\mathrm{SC}_T = \sum_{i,j}(y_{ij} - \bar{y})^2$	$ab-1$		

Cuadro 10.4: Tabla del Análisis de la Varianza para diseños de dos factores sin interacción

Cuando el efecto de la variable bloque no es significativo se puede considerar el modelo más simple con un solo factor, prescindiendo de los bloques. Sin embargo, si hay diferencias entre los bloques, el modelo en bloques aleatorizados es mucho más eficaz en la detección de diferencias entre tratamientos.

Finalmente, si se desea comparar dos niveles de un mismo factor, plantearemos la hipótesis parcial

$$H_0^{A(ij)} : \alpha_i = \alpha_j \qquad \text{o bien} \qquad H_0^{B(ij)} : \beta_i = \beta_j$$

según se trate de factor fila o columna. El estadístico utilizado en el primer caso será

$$t = \frac{y_{i\cdot} - y_{j\cdot}}{\sqrt{\mathrm{SCR}/[(a-1)(b-1)]}}\sqrt{b/2}$$

cuya distribución bajo la hipótesis es una t de Student con $(a-1)(b-1)$ grados de libertad. Análogamente, para comparar dos niveles del factor columna, utilizaremos

$$t = \frac{y_{\cdot i} - y_{\cdot j}}{\sqrt{\mathrm{SCR}/[(a-1)(b-1)]}}\sqrt{a/2}$$

con la misma distribución que el estadístico anterior si la hipótesis es cierta.

Por otra parte, en Ugarte[69, sec. 8.8] pueden verse algunos ejemplos de comparaciones múltiples para este modelo.

Coeficientes de determinación parcial

El coeficiente de determinación se define como

$$R^2 = 1 - \frac{\mathrm{SCR}}{\mathrm{SC}_T} = \frac{\mathrm{SC}_F + \mathrm{SC}_C}{\mathrm{SC}_T}$$

De modo que los coeficientes de determinación parcial

$$R_F^2 = \frac{\mathrm{SC}_F}{\mathrm{SC}_T} \qquad R_C^2 = \frac{\mathrm{SC}_C}{\mathrm{SC}_T}$$

indican el porcentaje de la variabilidad total explicada por el factor principal y por el factor bloque, respectivamente.

Descomposición fundamental de la suma de cuadrados

Vamos a probar la descomposición aditiva 10.16 en sumas de cuadrados. Para ello expresaremos el modelo 10.8 en notación vectorial

$$\mathbf{Y} = \mu\mathbf{1} + \sum_i \alpha_i \mathbf{u}_i + \sum_j \beta_j \mathbf{v}_j + \boldsymbol{\epsilon} \tag{10.18}$$

siendo

$$\begin{aligned}
\mathbf{1} &= (1,1,\ldots,1;1,1,\ldots,1;\ldots;1,1,\ldots,1)' \\
\mathbf{u}_1 &= (1,0,\ldots,0;1,0,\ldots,0;\ldots;1,0,\ldots,0)' \\
&\vdots \\
\mathbf{u}_a &= (0,\ldots,0,1;0,\ldots,0,1;\ldots;0,\ldots,0,1)' \\
\mathbf{v}_1 &= (1,1,\ldots,1;0,0,\ldots,0;\ldots;0,0,\ldots,0)' \\
&\vdots \\
\mathbf{v}_b &= (0,0,\ldots,0;0,0,\ldots,0;\ldots;1,1,\ldots,1)'
\end{aligned}$$

La matriz de diseño ampliada es

$$\mathbf{X} = (\mathbf{1}, \mathbf{u}_1, \ldots, \mathbf{u}_a, \mathbf{v}_1, \ldots, \mathbf{v}_b)$$

y es evidente que 10.18 es equivalente a

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

siendo $\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_a, \beta_1, \ldots, \beta_b)'$.
Se verifica

$$\begin{aligned}
&\mathbf{u}_{i_1}'\mathbf{u}_{i_2} = 0 \quad i_1 \neq i_2, && \mathbf{u}_i'\mathbf{u}_i = b \\
& && \mathbf{u}_i'\mathbf{v}_j = 1 \\
&\mathbf{v}_{j_1}'\mathbf{v}_{j_2} = 0 \quad j_1 \neq j_2, && \mathbf{v}_j'\mathbf{v}_j = a
\end{aligned}$$

Sustituyendo en 10.18 los parámetros por sus estimaciones MC obtenemos

$$\mathbf{Y} - \hat{\mu}\mathbf{1} = \sum_i \hat{\alpha}_i \mathbf{u}_i + \sum_j \hat{\beta}_j \mathbf{v}_j + \mathbf{e}$$

Como $\mathbf{e}$ es ortogonal al subespacio generado por las columnas de $\mathbf{X}$ (teorema 2.4.2), tendremos

$$\mathbf{u}_i'\mathbf{e} = \mathbf{v}_j'\mathbf{e} = 0$$

Entonces

$$\|\mathbf{Y} - \hat{\mu}\mathbf{1}\|^2 = \sum_i \hat{\alpha}_i^2\|\mathbf{u}_i\|^2 + \sum_j \hat{\beta}_j^2\|\mathbf{v}_j\|^2 + \sum_{i,j}\hat{\alpha}_i\hat{\beta}_j\mathbf{u}_i'\mathbf{v}_j + \|\mathbf{e}\|^2$$

Pero

$$\begin{aligned}\sum_{i,j}\hat{\alpha}_i\hat{\beta}_j &= \sum_{i,j}(y_{i\cdot} - \bar{y})(y_{\cdot j} - \bar{y}) \\ &= \sum_{i,j}(y_{i\cdot} - \bar{y})y_{\cdot j} - \bar{y}\sum_{i,j}(y_{i\cdot} - \bar{y}) \\ &= \sum_j y_{\cdot j}\sum_i(y_{i\cdot} - \bar{y}) - \bar{y}\sum_j\sum_i(y_{i\cdot} - \bar{y}) = 0\end{aligned}$$

pues $\sum_i(y_{i\cdot} - \bar{y}) = 0$.
Luego

$$\|\mathbf{Y} - \hat{\mu}\mathbf{1}\|^2 = \sum_i \hat{\alpha}_i^2\|\mathbf{u}_i\|^2 + \sum_j \hat{\beta}_j^2\|\mathbf{v}_j\|^2 + \|\mathbf{e}\|^2$$

lo que demuestra la descomposición fundamental 10.16.

Ejemplo 10.3.1

Para estudiar las diferencias entre los efectos de 4 fertilizantes sobre la producción de patatas, se dispuso de 5 fincas, cada una de las cuales se dividió en 4 parcelas del mismo tamaño y tipo. Los fertilizantes fueron asignados al azar en las parcelas de cada finca. El rendimiento en toneladas fue

	Finca				
Fert.	1	2	3	4	5
1	2,1	2,2	1,8	2,0	1,9
2	2,2	2,6	2,7	2,5	2,8
3	1,8	1,9	1,6	2,0	1,9
4	2,1	2,0	2,2	2,4	2,1

Algunos gráficos exploratorios pueden verse en la figura 10.4 de la página 213.

Se trata de un diseño en bloques aleatorizados. Este diseño utiliza el modelo 10.8 y es especialmente utilizado en experimentación agrícola. El objetivo es comparar $a = 4$ tratamientos (fertilizantes en este caso) utilizando $b = 5$ bloques (fincas) y repartiendo aleatoriamente los a tratamientos en cada uno de los bloques (los fertilizantes son asignados al azar en las parcelas de cada finca). Para una correcta aplicación de este diseño debe haber máxima homogeneidad dentro de cada bloque, de modo que el efecto bloque sea el mismo para todos los tratamientos.

Interesa pues saber si hay diferencias significativas entre los tratamientos α_i y entre los bloques β_j estableciendo con este fin las hipótesis lineales 10.14 y 10.15 respectivamente. Los resultados obtenidos son

$$\begin{array}{lllll}\bar{y} = 2{,}14 & y_{1\cdot} = 2{,}00 & y_{2\cdot} = 2{,}56 & y_{3\cdot} = 1{,}84 & y_{4\cdot} = 2{,}16 \\ y_{\cdot 1} = 2{,}050 & y_{\cdot 2} = 2{,}175 & y_{\cdot 3} = 2{,}075 & y_{\cdot 4} = 2{,}225 & y_{\cdot 5} = 2{,}175\end{array}$$

Bloques				
1	1	2	4	3
2	4	3	2	1
3	2	1	4	3
4	3	1	4	2
5	2	4	3	1

Cuadro 10.5: Formación correcta de bloques y asignación al azar de los tratamientos

La tabla del Análisis de la Varianza (ver tabla 10.4) es

Fuente variación	*suma cuadrados*	*g.l.*	*cuadrados medios*
Entre fertiliz.	1,432	3	0,477
Entre fincas	0,088	4	0,022
Residuo	0,408	12	0,034
Total	1,928	19	

El estadístico F para comparar las fincas es

$$F = \frac{0{,}022}{0{,}034} = 0{,}65$$

con 4 *y* 12 *grados de libertad. Como no es significativo, admitimos que no hay diferencias entre las fincas. Asimismo, para comparar los fertilizantes, el estadístico F es*

$$F = \frac{0{,}477}{0{,}034} = 14{,}04$$

con 3 *y* 12 *grados de libertad. Dado que es muy significativo podemos admitir que hay diferencias entre los fertilizantes.*

Como el efecto del factor bloque no es significativo podemos considerar el modelo de un factor, añadiendo la suma de cuadrados entre fincas al residuo.

Fuente variación	*suma cuadrados*	*g.l.*	*cuadrados medios*
Entre fertiliz.	1,432	3	0,477
Residuo	0,496	16	0,031
Total	1,928	19	

El estadístico F vale 15,39 *lo que muestra una significativa diferencia entre fertilizantes.*

10.4. Diseño de dos factores con interacción

Supongamos que la variable observable está influida por dos causas de variabilidad A y B, con a y b niveles respectivamente. Pero ahora, a diferencia del diseño de la sección anterior, los dos factores tienen a priori la misma importancia y aceptamos añadir un nuevo efecto denominado interacción entre factores. Entonces es preciso disponer de r observaciones por casilla, porque con una sola unidad experimental para cada combinación de niveles, el modelo tendría más parámetros que observaciones y la varianza del modelo no sería estimable.

Podemos disponer los datos de la siguiente manera

	B_1	B_2	$\ldots$	B_b
A_1	y_{111}	y_{121}	$\ldots$	y_{1b1}
	y_{112}	y_{122}		y_{1b2}
	$\vdots$	$\vdots$		$\vdots$
	y_{11r}	y_{12r}		y_{1br}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
A_a	y_{a11}	y_{a21}	$\ldots$	y_{ab1}
	y_{a12}	y_{a22}		y_{ab2}
	$\vdots$	$\vdots$		$\vdots$
	y_{a1r}	y_{a2r}		y_{abr}

Indicaremos las medias con la siguiente notación

$$y_{i\cdot\cdot} = \frac{1}{br}\sum_{j,k} y_{ijk} \qquad y_{\cdot j\cdot} = \frac{1}{ar}\sum_{i,k} y_{ijk}$$
$$y_{ij\cdot} = \frac{1}{r}\sum_{k} y_{ijk} \qquad y_{\cdots} = \bar{y} = \frac{1}{abr}\sum_{i,j,k} y_{ijk}$$

Modelo aditivo con interacción

En este modelo suponemos que el efecto fila (efecto debido al factor A) y el efecto columna (efecto debido al factor B) son aditivos, pero aceptamos también la presencia del efecto interacción. En otras palabras, el modelo lineal es

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \tag{10.19}$$

para todo $i = 1, \ldots, a; j = 1, \ldots, b; k = 1, \ldots, r$ y donde

$$\begin{aligned} \mu &= \text{media general} \\ \alpha_i &= \text{efecto del nivel } A_i \text{ de } A \\ \beta_j &= \text{efecto del nivel } B_j \text{ de } B \\ \gamma_{ij} &= \text{interacción entre los niveles } A_i \text{ y } B_j \end{aligned}$$

Para determinar todos los parámetros, se imponen también las restricciones naturales

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0 \tag{10.20}$$

con lo cual el modelo depende de

$$1 + (a-1) + (b-1) + (a-1)(b-1) = ab \tag{10.21}$$

parámetros.

La interacción γ_{ij} debe añadirse para prever el caso de que no se verifique la aditividad supuesta en 10.8. Indicando $\eta_{ij} = E(y_{ijk})$, la interacción mide la desviación respecto a un modelo totalmente aditivo

$$\gamma_{ij} = \eta_{ij} - \mu - \alpha_i - \beta_j \tag{10.22}$$

Por otra parte, diremos que un diseño es de *rango completo* si el número de parámetros es igual al número de condiciones experimentales, es decir, al número de filas distintas de la matriz de diseño. En un diseño que no es de rango completo hay menos parámetros que condiciones experimentales, por lo que en realidad "admitimos" que los datos se ajustan al modelo propuesto. Por ejemplo, en el diseño sin interacción tenemos (ver 10.10) $a+b-1<ab$, luego admitimos de partida el modelo 10.8. Sin embargo, este modelo puede no ser cierto y de hecho existe la llamada prueba de Tukey para comprobarlo (ver Peña[54, pág. 104] y Hoaglin et al.[39, págs. 268-273]). En cambio, por 10.21, el modelo 10.19 posee tantos parámetros como condiciones experimentales de variabilidad, de modo que es válido por construcción. En general, un modelo de rango completo se ajusta intrínsecamente a los datos sin problemas. No obstante, para poder estimar todos los parámetros es necesario disponer de más de una réplica por condición experimental. Esta es la razón por la cual la interacción no puede ser incluida en 10.8.

El modelo 10.19 puede ser reparamentrizado en la forma

$$y_{ijk}=\eta_{ij}+\epsilon_{ijk} \tag{10.23}$$

Pasamos del modelo 10.23 al 10.19 mediante las transformaciones

$$\begin{aligned} \mu &= \frac{1}{ab}\sum_{i,j}\eta_{ij} & \alpha_i &= \frac{1}{b}\left(\sum_j \eta_{ij}\right)-\mu \\ \beta_j &= \frac{1}{a}\left(\sum_i \eta_{ij}\right)-\mu & \gamma_{ij} &= \eta_{ij}-\mu-\alpha_i-\beta_j \end{aligned} \tag{10.24}$$

Estimación de los parámetros

Consideremos la identidad

$$\begin{aligned} y_{ijk}-\mu-\alpha_i-\beta_j-\gamma_{ij} \;=\; & (\bar{y}-\mu)+(y_{i\cdot\cdot}-\bar{y}-\alpha_i) \\ & +(y_{\cdot j\cdot}-\bar{y}-\beta_j) \\ & +(y_{ij\cdot}-y_{i\cdot\cdot}-y_{\cdot j\cdot}+\bar{y}-\gamma_{ij}) \\ & +(y_{ijk}-y_{ij\cdot}) \end{aligned}$$

Elevando al cuadrado y teniendo en cuenta las restricciones 10.20, los productos cruzados se anulan y queda

$$\begin{aligned} \sum_{i,j,k}(y_{ijk}-\mu-\alpha_i-\beta_j-\gamma_{ij})^2 = & \sum_{i,j,k}(\bar{y}-\mu)^2+\sum_{i,j,k}(y_{i\cdot\cdot}-\bar{y}-\alpha_i)^2 \\ & +\sum_{i,j,k}(y_{\cdot j\cdot}-\bar{y}-\beta_j)^2 \\ & +\sum_{i,j,k}(y_{ij\cdot}-y_{i\cdot\cdot}-y_{\cdot j\cdot}+\bar{y}-\gamma_{ij})^2 \\ & +\sum_{i,j,k}(y_{ijk}-y_{ij\cdot})^2 \end{aligned} \tag{10.25}$$

Como el último término de esta expresión no depende de los parámetros, es fácil ver que las estimaciones MC son

$$\hat{\mu}=\bar{y}\quad \hat{\alpha}_i=y_{i\cdot\cdot}-\bar{y}\quad \hat{\beta}_j=y_{\cdot j\cdot}-\bar{y}\quad \hat{\gamma}_{ij}=y_{ij\cdot}-y_{i\cdot\cdot}-y_{\cdot j\cdot}+\bar{y} \tag{10.26}$$

mientras que la suma de cuadrados residual es

$$\text{SCR} = \sum_{i,j,k} (y_{ijk} - y_{ij\cdot})^2$$

que tiene $ab(r-1)$ grados de libertad. Luego la estimación de la varianza es

$$\hat{\sigma}^2 = \text{SCR}/[ab(r-1)]$$

Por otra parte, considerando 10.23 y 10.24 podemos obtener las estimaciones 10.26 por otro camino. Es obvio que las estimaciones de η_{ij} son

$$\hat{\eta}_{ij} = y_{ij\cdot}$$

Interpretando $\mu, \alpha_i, \beta_j, \gamma_{ij}$ como funciones paramétricas sobre el modelo 10.23, por el teorema de Gauss-Markov, sus estimaciones se obtendrán sustituyendo η_{ij} por $y_{ij\cdot}$ en 10.24, lo que nos dará 10.26.

Hipótesis lineales

En el diseño de dos factores con interacción, las hipótesis de mayor interés son

$$\begin{array}{lcll} H_0^A & : & \alpha_1 = \cdots = \alpha_a = 0 & \text{(no hay efecto fila)} \\ H_0^B & : & \beta_1 = \cdots = \beta_b = 0 & \text{(no hay efecto columna)} \\ H_0^{AB} & : & \gamma_{ij} = 0 \quad \forall i,j & \text{(no hay interacción)} \end{array}$$

Los rangos son $a-1$, $b-1$ y $(a-1)(b-1)$ respectivamente.
A fin de deducir el test F correspondiente, consideremos la siguiente descomposición fundamental de la suma de cuadrados

$$\begin{aligned} \sum_{i,j,k} (y_{ijk} - \bar{y})^2 &= br \sum_i (y_{i\cdot\cdot} - \bar{y})^2 + ar \sum_j (y_{\cdot j\cdot} - \bar{y})^2 \\ &\quad + r \sum_{i,j} (y_{ij\cdot} - y_{i\cdot\cdot} - y_{\cdot j\cdot} + \bar{y})^2 \\ &\quad + \sum_{i,j,k} (y_{ijk} - y_{ij\cdot})^2 \end{aligned}$$

Esta relación, que se puede probar con algo de esfuerzo, la expresaremos brevemente como

$$\text{SC}_T = \text{SC}_F + \text{SC}_C + \text{SC}_I + \text{SCR}$$

donde SC_T es la suma de cuadrados total, SC_I es la suma de cuadrados correspondiente a la interacción, etc.
Consideremos ahora la hipótesis H_0^A. La suma de cuadrados residual es SCR. Supongamos la hipótesis cierta, entonces el modelo 10.19 se convierte en

$$y_{ijk} = \mu + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

Además, como no hay α_i, el mínimo de 10.25, es decir, la suma de cuadrados residual bajo H_0^A es

$$\text{SCR}_H = \sum (y_{i\cdot\cdot} - \bar{y})^2 + \sum (y_{ijk} - y_{ij\cdot})^2 = \text{SC}_F + \text{SCR}$$

Luego si H_0^A es cierta (teorema 5.3.1) tendremos que

$$F = \frac{(\mathrm{SCR}_H - \mathrm{SCR})/(a-1)}{\mathrm{SCR}/[ab(r-1)]} = \frac{\mathrm{SC}_F/(a-1)}{\mathrm{SCR}/[ab(r-1)]}$$

sigue la distribución $F(a-1, ab(r-1))$.

La obtención del test F para decidir sobre H_0^B y H_0^{AB} es análoga. En la práctica, los cálculos suelen disponerse en forma de tabla (ver tabla 10.6).

Fuente de variación	suma de cuadrados	g.l.	cuadrados medios	F
Entre filas	$\mathrm{SC}_F = br\sum_i(y_{i\cdot\cdot} - \bar{y})^2$	$a-1$	$\mathrm{SC}_F/(a-1)$	$\frac{\mathrm{SC}_F/(a-1)}{\mathrm{SCR}/[ab(r-1)]}$
Entre col.	$\mathrm{SC}_C = ar\sum_j(y_{\cdot j\cdot} - \bar{y})^2$	$b-1$	$\mathrm{SC}_C/(b-1)$	$\frac{\mathrm{SC}_C/(b-1)}{\mathrm{SCR}/[ab(r-1)]}$
Interacción	$\mathrm{SC}_I = r\sum_{i,j}(y_{ij\cdot} - y_{i\cdot\cdot} - y_{\cdot j\cdot} + \bar{y})^2$	$(a-1)(b-1)$	$\frac{\mathrm{SC}_I}{(a-1)(b-1)}$	$\frac{\mathrm{SC}_I/[(a-1)(b-1)]}{\mathrm{SCR}/[ab(r-1)]}$
Residuo	$\mathrm{SCR} = \sum_{i,j,h}(y_{ijh} - y_{ij\cdot})^2$	$ab(r-1)$	$\frac{\mathrm{SCR}}{ab(r-1)}$	
Total	$\mathrm{SC}_T = \sum_{i,j,h}(y_{ijh} - \bar{y})^2$	$abr-1$		

Cuadro 10.6: Tabla del Análisis de la Varianza para diseños de dos factores con interacción

Ejemplo 10.4.1

Se desean comparar tres genotipos distintos de Drosophila melanogaster*, observando si existen diferencias de viabilidad sembrando* 100 *y* 800 *huevos. De este modo, para cada una de las* 6 *casillas del experimento (*3 *genotipos* $\times$ 2 *siembras) se dispusieron* 6 *preparados (*6 *réplicas) y al cabo del tiempo suficiente de ser sembrados los huevos, se obtuvo el porcentaje de huevos que habían eclosionado. Los resultados fueron:*

Huevos sembrados	*Genotipo* ++			+−			−−		
100	93	94	93	95,5	83,5	92	92	91	90
	90	93	86	92,5	82	82,5	95	84	78
800	83,3	87,6	81,9	84	84,4	77	85,3	89,4	85,4
	80,1	79,6	49,4	67	69,1	88,4	87,4	52	77

El número X *de huevos eclosionados por casilla sigue la distribución binomial con* $n = 100$ *ó* $n = 800$*. Para normalizar la muestra aplicaremos la transformación*

$$Y = \mathrm{arcsen}\sqrt{\frac{X}{n}} = \mathrm{arcsen}\sqrt{\frac{porcentaje}{100}}$$

Los datos transformados son:

Huevos sembrados	*Genotipo*								
	++			+−			−−		
100	74,7	75,8	74,7	77,8	66	73,6	73,6	72,5	71,6
	71,6	74,7	68	74,1	64,9	65,3	77,1	66,4	62
800	65,9	69,4	64,8	66,4	66,7	61,3	67,5	71	67,5
	63,5	63,1	44,7	54,9	56,2	70,1	69,2	46,1	61,3

Con estos datos se dibujan los gráficos de la figura 10.1 que avanzan el resultado final.

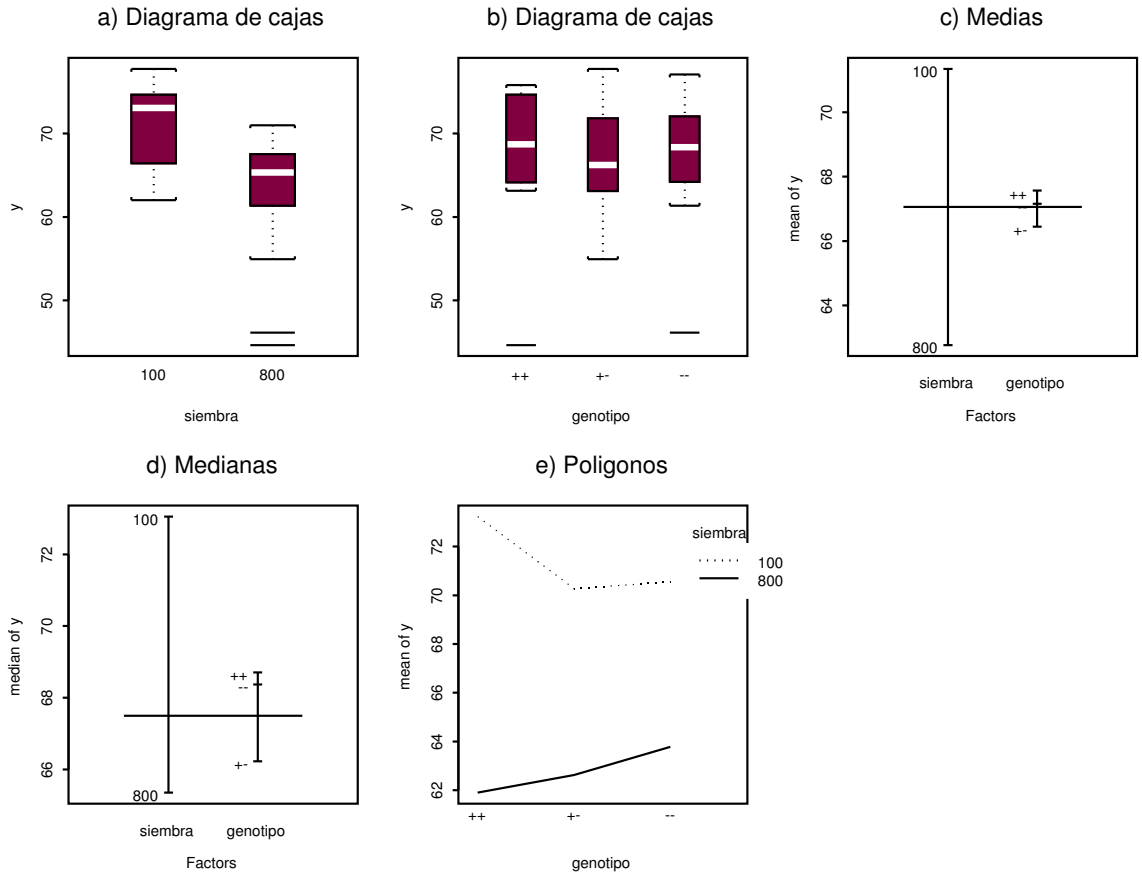


Figura 10.1: Gráficos del análisis exploratorio de los datos del ejemplo 10.4.1

A continuación se calculan las siguientes medias:

$$y_{11\cdot} = 73{,}231 \quad y_{12\cdot} = 70{,}271 \quad y_{13\cdot} = 70{,}534 \quad y_{21\cdot} = 61{,}899$$
$$y_{22\cdot} = 62{,}626 \quad y_{23\cdot} = 63{,}781 \quad y_{1\cdot\cdot} = 71{,}346 \quad y_{2\cdot\cdot} = 62{,}769$$
$$y_{\cdot 1\cdot} = 67{,}565 \quad y_{\cdot 2\cdot} = 66{,}449 \quad y_{\cdot 3\cdot} = 67{,}158 \quad \bar{y} = 67{,}057$$

Con ellas podemos obtener entonces la tabla del Análisis de la Varianza para un diseño de dos factores con interacción:

Fuente variación	*suma cuadrados*	*g.l.*	*cuadrados medios*	*F*
Entre siembras	662,086	1	662,086	14,833
Entre genotipos	7,665	2	3,832	0,086
Interacción	35,354	2	17,677	0,396
Residuo	1339,094	30	44,636	
Total	2044,199	35		

A la vista de los valores F obtenidos, se concluye que no es significativa la diferencia entre genotipos ni la interacción, pero sí existen diferencias significativas sembrando 100 *o* 800 *huevos, siendo el porcentaje de eclosiones mayor en el primer caso, ya que según parece al haber menos huevos, las larvas disponen de más alimento.*

Observación: cuando un factor no es significativo, la interacción generalmente tampoco lo es.

10.5. Descomposición ortogonal de la variabilidad

En las secciones anteriores han sido tratados los diseños de uno y dos factores y se ha estudiado cómo descomponer adecuadamente la variabilidad. Los diseños en los que intervienen tres o más factores pueden estudiarse también descomponiendo adecuadamente la variabilidad total

$$\mathrm{SC}_T = \sum (y_{ij\ldots m} - \bar{y})^2$$

en diferentes sumas de cuadrados, más una suma de cuadrados residual. Veamos cómo debe procederse para un diseño de cuatro factores que indicaremos A, B, C y D, con a, b, c y d niveles respectivamente. Distinguiremos dos casos:

(a) D es el factor réplica, es decir, d es el número de réplicas para cada condición experimental o combinación de los niveles de los factores A, B, C. El modelo lineal es

$$y_{ijkr} = \mu + \alpha_i^A + \alpha_j^B + \alpha_k^C + \alpha_{ij}^{AB} + \alpha_{ik}^{AC} + \alpha_{jk}^{BC} + \alpha_{ijk}^{ABC} + \epsilon_{ijkr}$$

para $i = 1, \ldots, a$; $j = 1, \ldots, b$; $k = 1, \ldots, c$; $r = 1, \ldots, d$ y siendo

y_{ijkr} = réplica r para los niveles i, j, k de A, B, C
μ = media general
$\alpha_i^A, \alpha_j^B, \alpha_k^C$ = efectos principales de A, B, C
$\alpha_{ij}^{AB}, \alpha_{ik}^{AC}, \alpha_{jk}^{BC}$ = interacciones entre los factores A y B, A y C, B y C
α_{ijk}^{ABC} = interacción entre los tres factores
ϵ_{ijkr} = desviación aleatoria $N(0, \sigma)$

Debe imponerse la restricción de que la suma (respecto a uno o dos subíndices) de los parámetros α sea igual a cero.

(b) D es un verdadero factor con d niveles, de modo que el diseño depende de cuatro factores con una sola observación por casilla. El modelo es

$$\begin{aligned} y_{ijkm} \;=\; & \mu + \alpha_i^A + \alpha_j^B + \alpha_k^C + \alpha_m^D + \alpha_{ij}^{AB} + \alpha_{ik}^{AC} + \alpha_{im}^{AD} + \alpha_{jk}^{BC} + \alpha_{jm}^{BD} + \alpha_{km}^{CD} \\ & + \alpha_{ijk}^{ABC} + \alpha_{ijm}^{ABD} + \alpha_{ikm}^{ACD} + \alpha_{jkm}^{BCD} + \epsilon_{ijkm} \end{aligned}$$

La interpretación de los parámetros es análoga.

La tabla 10.7 contiene la descomposición de la variabilidad. Los sumatorios deben desarrollarse para todos los subíndices i, j, k, m, verificándose por lo tanto

$$\begin{aligned} \mathrm{SC}_A &= \sum_{i,j,k,m} (y_{i\cdots} - \bar{y})^2 = bcd \sum_i (y_{i\cdots} - \bar{y})^2 \\ \mathrm{SC}_B &= \sum_{i,j,k,m} (y_{\cdot j\cdot\cdot} - \bar{y})^2 = acd \sum_j (y_{\cdot j\cdot\cdot} - \bar{y})^2 \\ \mathrm{SC}_{BC} &= ad \sum_{j,k} (y_{\cdot jk\cdot} - y_{\cdot j\cdot\cdot} - y_{\cdot\cdot k\cdot} + \bar{y})^2 \\ & \text{etcétera.} \end{aligned}$$

Cuadro 10.7: Descomposición ortogonal de la suma de cuadrados correspondiente a un diseño de cuatro factores

Fuente de variación	suma de cuadrados	grados de libertad
A	$\sum (y_{i\cdots} - \bar{y})^2$	$a-1$
B	$\sum (y_{\cdot j\cdot\cdot} - \bar{y})^2$	$b-1$
C	$\sum (y_{\cdot\cdot k\cdot} - \bar{y})^2$	$c-1$
D	$\sum (y_{\cdots m} - \bar{y})^2$	$d-1$
AB	$\sum (y_{ij\cdot\cdot} - y_{i\cdots} - y_{\cdot j\cdot\cdot} + \bar{y})^2$	$(a-1)(b-1)$
AC	$\sum (y_{i\cdot k\cdot} - y_{i\cdots} - y_{\cdot\cdot k\cdot} + \bar{y})^2$	$(a-1)(c-1)$
AD	$\sum (y_{i\cdot\cdot m} - y_{i\cdots} - y_{\cdots m} + \bar{y})^2$	$(a-1)(d-1)$
BC	$\sum (y_{\cdot jk\cdot} - y_{\cdot j\cdot\cdot} - y_{\cdot\cdot k\cdot} + \bar{y})^2$	$(b-1)(c-1)$
BD	$\sum (y_{\cdot j\cdot m} - y_{\cdot j\cdot\cdot} - y_{\cdots m} + \bar{y})^2$	$(b-1)(d-1)$
CD	$\sum (y_{\cdot\cdot km} - y_{\cdot\cdot k\cdot} - y_{\cdots m} + \bar{y})^2$	$(c-1)(d-1)$
ABC	$\sum (y_{ijk\cdot} - y_{ij\cdot\cdot} - y_{i\cdot k\cdot} - y_{\cdot jk\cdot} + y_{i\cdots} + y_{\cdot j\cdot\cdot} + y_{\cdot\cdot k\cdot} - \bar{y})^2$	$(a-1)(b-1)(c-1)$
ABD	$\sum (y_{ij\cdot m} - y_{ij\cdot\cdot} - y_{i\cdot\cdot m} - y_{\cdot j\cdot m} + y_{i\cdots} + y_{\cdot j\cdot\cdot} + y_{\cdots m} - \bar{y})^2$	$(a-1)(b-1)(d-1)$
ACD	$\sum (y_{i\cdot km} - y_{i\cdot k\cdot} - y_{i\cdot\cdot m} - y_{\cdot\cdot km} + y_{i\cdots} + y_{\cdot\cdot k\cdot} + y_{\cdots m} - \bar{y})^2$	$(a-1)(c-1)(d-1)$
BCD	$\sum (y_{\cdot jkm} - y_{\cdot jk\cdot} - y_{\cdot j\cdot m} - y_{\cdot\cdot km} + y_{\cdot j\cdot\cdot} + y_{\cdot\cdot k\cdot} + y_{\cdots m} - \bar{y})^2$	$(b-1)(c-1)(d-1)$
$ABCD$	$\sum (y_{ijkm} - y_{ijk\cdot} - y_{ij\cdot m} - y_{i\cdot km} - y_{\cdot jkm} + y_{ij\cdot\cdot} + y_{i\cdot k\cdot} + y_{\cdot jk\cdot} + y_{i\cdot\cdot m} + y_{\cdot j\cdot m} + y_{\cdot\cdot km} - y_{i\cdots} - y_{\cdot j\cdot\cdot} - y_{\cdot\cdot k\cdot} - y_{\cdots m} + \bar{y})^2$	$(a-1)(b-1)(c-1)(d-1)$
Total	$\sum (y_{ijkm} - \bar{y})^2$	$abcd-1$

Estas sumas de cuadrados pueden reunirse convenientemente, sumando también los grados de libertad, según el tipo de diseño factorial para obtener la suma de cuadrados residual. Veamos tres casos:

1) Supongamos que se trata de un diseño de tres factores y réplicas, como el descrito en (a). Entonces:

$$\mathrm{SC}_T = \mathrm{SC}_A + \mathrm{SC}_B + \mathrm{SC}_C + \mathrm{SC}_{AB} + \mathrm{SC}_{AC} + \mathrm{SC}_{BC} + \mathrm{SC}_{ABC} + \mathrm{SCR}$$

siendo la suma de cuadrados residual

$$\begin{aligned} \text{SCR} &= \text{SC}_D + \text{SC}_{AD} + \text{SC}_{BD} + \text{SC}_{CD} + \text{SC}_{ABD} + \text{SC}_{ACD} + \text{SC}_{BCD} + \text{SC}_{ABCD} \\ &= \sum (y_{ijkm} - y_{ijk\cdot})^2 \end{aligned}$$

con $(d-1) + \cdots + [(a-1)(b-1)(c-1)(d-1)] = abc(d-1)$ grados de libertad.

Para estudiar, por ejemplo, si la interacción entre A y B es significativa, calcularemos

$$F = \frac{\text{SC}_{AB}/[(a-1)(b-1)]}{\text{SCR}/[abc(d-1)]}$$

y consultaremos la tabla F con $(a-1)(b-1)$ y $abc(d-1)$ grados de libertad.

2) Supongamos que se trata de un diseño de 4 factores con una sola observación por casilla, como el descrito en (b). Entonces:

$$\text{SC}_T = \text{SC}_A + \text{SC}_B + \text{SC}_C + \text{SC}_D + \text{SC}_{AB} + \cdots + \text{SC}_{CD} + + \text{SC}_{ABC} + \cdots + \text{SC}_{BCD} + \text{SCR}$$

siendo $\text{SCR} = \text{SC}_{ABCD}$ la suma de cuadrados residual. La significación de los efectos principales o las interacciones deberá efectuarse dividiendo por SC_{ABCD}.

3) Supongamos que C es un factor (por ejemplo, un factor bloque) que no interacciona con A, B y que D es un "factor réplica". Entonces

$$\text{SC}_T = \text{SC}_A + \text{SC}_B + \text{SC}_C + \text{SC}_{AB} + \text{SCR}$$

siendo

$$\text{SCR} = \text{SC}_D + \text{SC}_{AC} + \text{SC}_{AD} + \cdots + \text{SC}_{CD} + \text{SC}_{ABC} + \text{SC}_{ABD} + \text{SC}_{BCD} + \text{SC}_{ABCD}$$

la suma de cuadrados residual.

La formulación general de esta descomposición de la suma de cuadrados permite abordar muchos tipos de diseños que resulten de la combinación de varios factores, con una sola réplica por casilla, o con el mismo número de réplicas por casilla (diseños balanceados). En este caso, las réplicas se consideran como un factor formal y el residuo estará formado por todas las sumas de cuadrados en los que interviene el factor réplica. Las interacciones no presentes en un determinado modelo (por condiciones experimentales o por cocientes F claramente no significativos) se añaden al residuo. Esta formulación general no permite tratar ciertos diseños como cuadrados latinos, bloques incompletos balanceados, etc.

Esta descomposición ortogonal, para un número cualquiera de factores, puede programarse por ordenador siguiendo el algoritmo propuesto por Hartley[36].

La principal dificultad de estos diseños es la gran cantidad de observaciones necesarias, de modo que en la práctica no se consideran diseños con más de cuatro factores. En algunos casos se puede suponer que las interacciones altas son nulas y estimar el resto de parámetros. Ésta es la propuesta de los diseños en cuadrados latinos y greco-latinos que permiten estimar los efectos principales con el mínimo de observaciones (ver Peña[54, pág. 116-128] y Cuadras[20, pág. 261-262]).

10.5.1. Descomposición de la variabilidad en algunos diseños

Indicando simbólicamente por $A, B, AB, \ldots, T$ las sumas de cuadrados SC_A,SC_B, SC_{AB}, $\ldots, \mathrm{SC}_T$, exponemos seguidamente diferentes diseños del Análisis de la Varianza, presentando la descomposición de la variabilidad. Algunos diseños han sido tratados en las secciones anteriores de este capítulo.

1. Un factor y réplicas

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$T = A + R + AR$

Entre grupos	A	$a-1$
Residuo	$R + AR$	$ar - a$

2. Dos factores con una observación por casilla

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

$T = A + B + AB$

Entre filas	A	$a-1$
Entre columnas	B	$b-1$
Residuo	AB	$(a-1)(b-1)$

3. Dos factores con interacción

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

$T = A + B + R + AB + AR + BR + ABR$

Efecto fila	A	$a-1$
Efecto columna	B	$b-1$
Interacción	AB	$(a-1)(b-1)$
Residuo	$R + AR + BR + ABR$	$ab(r-1)$

4. Dos factores con interacción en bloques aleatorizados

$$y_{ijk} = \mu + \alpha_i + \beta_j + b_k + \gamma_{ij} + \epsilon_{ijk}$$

$T = A + B + R + AB + AR + BR + ABR$

Efecto fila	A	$a-1$
Efecto columna	B	$b-1$
Efecto bloque	R	$r-1$
Interacción	AB	$(a-1)(b-1)$
Residuo	$AR + BR + ABR$	$(ab-1)(r-1)$

Este modelo se utiliza cuando se combinan dos factores A, B y se obtienen réplicas organizadas en bloques. El factor bloque tiene un efecto principal, pero no interacciona con A, B.

5. Tres factores con una observación por casilla

$$y_{ijk} = \mu + \alpha_i + \beta_j + \delta_k + (\alpha\beta)_{ij} + (\alpha\delta)_{ik} + (\beta\delta)_{jk} + \epsilon_{ijk}$$

$$T = A + B + C + AB + AC + BC + ABC$$

Efecto A	A	$a-1$
Efecto B	B	$b-1$
Efecto C	C	$c-1$
Interacción $A \times B$	AB	$(a-1)(b-1)$
Interacción $A \times C$	AC	$(a-1)(c-1)$
Interacción $B \times C$	BC	$(b-1)(c-1)$
Residuo	ABC	$(a-1)(b-1)(c-1)$

6. Tres factores con r observaciones por casilla

$$y_{ijkm} = \mu + \alpha_i + \beta_j + \delta_k + (\alpha\beta)_{ij} + (\alpha\delta)_{ik} + (\beta\delta)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkm}$$

$$T = A + B + C + R + AB + AC + AR + BC + BR + CR + ABC + ABR + ACR + BCR + ABCR$$

Efecto A	A	$a-1$
Efecto B	B	$b-1$
Efecto C	C	$c-1$
Interacción $A \times B$	AB	$(a-1)(b-1)$
Interacción $A \times C$	AC	$(a-1)(c-1)$
Interacción $B \times C$	BC	$(b-1)(c-1)$
Interacción $A \times B \times C$	ABC	$(a-1)(b-1)(c-1)$
Residuo	$R + AR + BR + CR + ABR + ACR + BCR + ABCR$	$abc(r-1)$

7. Diseño de parcela dividida

$$y_{ijk} = \mu + \alpha_i + \gamma_j + b_k + (\alpha\gamma)_{ij} + (\alpha b)_{ik} + + \epsilon_{ijk}$$

$$T = A + C + B + AC + AB + CB + ACB$$

Tratamiento principal	A	$a-1$
Subtratamiento	C	$c-1$
Bloque	B	$b-1$
Interacción $A \times C$	AC	$(a-1)(c-1)$
Interacción $A \times B$	AB	$(a-1)(b-1)$
Residuo	$CB + ACB$	$a(b-1)(c-1)$

B_1	A_2		A_1		A_3		A_4	
	C_1	C_2	C_2	C_1	C_2	C_1	C_1	C_2
B_2	A_1		A_3		A_4		A_2	
	C_2	C_1	C_2	C_1	C_1	C_2	C_1	C_2
B_3	A_3		A_4		A_2		A_1	
	C_1	C_2	C_1	C_2	C_2	C_1	C_2	C_1

Este diseño se utiliza en investigación agrícola, también en otras ciencias experimentales, para comparar a tratamientos (factor A) que se asignan aleatoriamente en b bloques o fincas (factor B), a razón de a tratamientos por bloque. Se divide cada una de las

ab parcelas y se asignan al azar c subtratamientos ($factor$C), tal como se ilustra en el esquema para el caso $a = 4$, $b = 3$, $c = 2$. Se supone que actúan los efectos principales A, B y C, la interacción $A \times C$ y la interacción $A \times B$. La interacción entre A y los bloques es debida a que estos no pueden considerarse completamente homogéneos. Sin embargo, se supone que cada una de las ab parcelas dentro de los bloques son homogéneas, de modo que los subtratamientos C no interaccionan con los bloques.

Para la significación de C y la interacción $A \times C$ debe calcularse

$$F_C = \frac{C/(c-1)}{(CB + ABC)/[a(b-1)(c-1)]} \qquad F_{AC} = \frac{AC/[(a-1)(c-1)]}{(CB + ABC)/[a(b-1)(c-1)]}$$

Para estudiar la significación del factor A y del factor bloque debe calcularse

$$F_A = \frac{A/(a-1)}{AB/[(a-1)(b-1)]} \qquad F_B = \frac{B/(b-1)}{AB/[(a-1)(b-1)]}$$

10.5.2. Estimación de parámetros y cálculo del residuo

La estimación de los efectos principales y las interacciones se obtienen utilizando los términos que intervienen en las correspondientes sumas de cuadrados (ver tabla 10.7).

Por ejemplo, en un estudio de dos factores con interacción en bloques aleatorizados, las estimaciones son:

$$\hat{\mu} = \bar{y} \qquad \hat{\alpha}_i = y_{i\cdot\cdot} - \bar{y} \qquad \hat{\beta}_j = y_{\cdot j\cdot} - \bar{y}$$
$$\hat{b}_k = y_{\cdot\cdot k} - \bar{y} \qquad \hat{\gamma}_{ij} = y_{ij\cdot} - y_{i\cdot\cdot} - y_{\cdot j\cdot} + \bar{y}$$

Se puede aplicar una regla sencilla para encontrar la expresión algebraica del residuo. En el diseño citado, cuyo modelo es

$$y_{ijk} = \mu + \alpha_i + \beta_j + b_k + \gamma_{ij} + \epsilon_{ijk}$$

sustituiremos los parámetros por sus estimaciones

$$\begin{aligned} y_{ijk} &= \bar{y} + (y_{i\cdot\cdot} - \bar{y}) + (y_{\cdot j\cdot} - \bar{y}) + (y_{\cdot\cdot k} - \bar{y}) \\ &\quad + (y_{ij\cdot} - y_{i\cdot\cdot} - y_{\cdot j\cdot} + \bar{y}) + e_{ijk} \end{aligned}$$

Para que exista identidad entre y_{ijk} y el término de la derecha, la estimación de la desviación aleatoria e_{ijk} debe ser

$$e_{ijk} = y_{ijk} - y_{ij\cdot} - y_{\cdot\cdot k} + \bar{y}$$

El residuo correspondiente al diseño de dos factores con interacción en bloques aleatorizados es entonces

$$\sum_{i,j,k} e_{ijk}^2 = \sum_{i,j,k} (y_{ijk} - y_{ij\cdot} - y_{\cdot\cdot k} + \bar{y})^2$$

fórmula que coincide con $AR + BR + ABR$.

Esta regla sirve para todos los diseños que admiten descomposición ortogonal de la suma de cuadrados. Por poner otro ejemplo, para el diseño de parcela dividida se comprueba de este modo que la estimación de la desviación aleatoria es

$$e_{ijk} = y_{ijk} - y_{i\cdot k} - y_{ij\cdot} + y_{i\cdot\cdot}$$

Ejemplo 10.5.1

Con el fin de valorar la acción de los hongos xilófagos sobre la madera, se han tomado 240 muestras de madera procedente de tocones de Pinus silvestris*, clasificados atendiendo simultáneamente a* 4 *factores (edad, orientación, altura y profundidad). La descripción de los factores es:*

Edad (E)**:** *Años transcurridos desde la fecha de tala (*1,4,7,10 *o* 13 *años).*

Orientación (O)**:** *N,S,E,O según la ubicación de la muestra en el tocón.*

Altura (A)**:** $0, 2, 5, 15$ *expresada en cm contados a partir de la superficie de corte.*

Profundidad (P)**:** $0, 2, 5$ *expresada en cm contados radialmente a partir de la superficie lateral.*

Cada una de las $5 \times 4 \times 4 \times 3 = 240$ *muestras era en realidad la homogeneización de* 3 *muestras procedentes de* 3 *tocones distintos pero de las mismas características en cuanto a la edad, orientación, altura y profundidad.*

Se estudiaron 8 *variables químicas. Para la variable que medía la cantidad de hemicelulosa, se obtuvo la siguiente descomposición ortogonal de la suma de cuadrados:*

Fuente de variación	*Suma de cuadrados*	*Grados de libertad*	*Cuadrados medios*	F
E	1227,53	4	306,88	59,21
O	51,94	3	17,31	3,34
A	58,59	3	19,53	3,76
P	18,04	2	9,02	1,74
EO	152,70	12	12,72	2,45
EA	137,13	12	11,42	2,20
EP	72,22	8	9,03	1,74
OA	54,60	9	6,06	1,17
OP	37,26	6	6,21	1,20
AP	21,04	6	3,50	0,68
EOA	189,89	36	5,27	1,01
EOP	145,12	24	6,04	1,16
EAP	132,22	24	5,50	1,06
OAP	60,70	18	3,37	0,65
$EOAP$	373,19	72	5,18	
Total	2732,64	239		

Los datos se adaptan a un diseño de 4 *factores con una observación por casilla. El residuo es la suma de cuadrados indicada simbólicamente por EOAP y su valor es* 373,19 *con* 72 *grados de libertad. Un examen inicial de los cocientes F de la tabla, obtenidos dividiendo los cuadrados medios por* $373{,}19/72 = 5{,}18$*, para un nivel de significación de* 0,05 *nos lleva a las siguientes conclusiones:*

a) Son significativos los efectos principales E,O,A. No es significativo el efecto principal P.

b) Son significativas las interacciones EA y EO. No son significativas el resto de las interacciones.

Prescindiendo de los efectos no significativos, resulta un diseño de tres factores (E,O,A), de los cuales interaccionan E con A y E con O (edad con altura y edad con orientación). Añadiendo las correspondientes sumas de cuadrados al residuo, obtenemos la siguiente tabla:

Fuente de variación	*Suma de cuadrados*	*Grados de libertad*	*Cuadrados medios*	*F*
E	1227,53	4	306,88	56,97
O	51,94	3	17,31	3,21
A	58,59	3	19,53	3,63
EO	152,70	12	12,72	2,36
EA	137,13	12	11,42	2,12
Residuo	1104,26	205	5,39	
Total	2732,64	239		

Se observa que sigue existiendo variabilidad significativa respecto E,O y A. También son significativas las interacciones EO y EA. Por lo tanto, se confirman las conclusiones iniciales. Una estimación insesgada de la varianza σ^2 *es* $\hat{\sigma}^2 = 5{,}39$.

10.6. Diagnosis del modelo

Una vez decidido el modelo, calculados los parámetros y contrastadas las hipótesis sobre los parámetros, es necesario comprobar si las hipótesis esenciales del modelo lineal se cumplen. En caso contrario debemos analizar las consecuencias y si es preciso un cambio de modelo. Para ello y como en el caso de los modelos de regresión (ver capítulo 9) realizaremos un completo análisis de los residuos. Como ya se ha explicado, dicho análisis debe comprobar la normalidad, la independencia y la aleatoriedad, la no existencia de valores atípicos y la homocedasticidad. Así pues, en esta sección vamos a comentar los aspectos específicos de este tema en el caso de los modelos de Análisis de la Varianza.

Podemos empezar por una exploración previa de las observaciones, especialmente gráfica, como puede ser un diagrama de cajas múltiple o, cuando el número de datos sea muy pequeño, un gráfico de puntos como el de la tabla 10.2.

Una vez resuelto el modelo, podemos realizar el estudio descriptivo y gráfico de la distribución de los residuos. En este sentido los gráficos propuestos en 9.1.3, como diagramas de dispersión frente a las previsiones (medias en cada grupo), QQ-plots, etc., nos proporcionarán mucha información sobre la veracidad de las hipótesis básicas de un modelo lineal normal.

Por otra parte, como la mayoría de diseños responden a una situación experimental, siempre conviene representar los residuos respecto a su índice temporal de observación. Con ello podemos detectar posibles cambios en las condiciones experimentales que provocarían una correlación indeseable o una alteración en la variabilidad experimental. Por ejemplo, esto último ocurre cuando se manifiesta el llamado efecto de aprendizaje.

La falta de normalidad no es un problema grave. Aunque las observaciones no sigan la ley normal, los contrastes son esencialmente válidos y se dice que el Análisis de la Varianza es en este caso una *técnica robusta*. Sin embargo, la no normalidad sí afecta a la precisión de la estimación de la varianza del modelo y su estimación por intervalos.

El efecto de varianzas desiguales en los grupos afecta al contraste F si el número de observaciones en cada grupo es diferente (máx n_i/mín $n_i > 2$). En caso contrario, cuando el número de réplicas por casilla es el mismo, el contraste F es bastante robusto incluso cuando las varianzas son fuertemente distintas (por ejemplo en una relación 1 a 30). Por supuesto, una fuerte desigualdad de la varianza del error en los grupos sí influye marcadamente en la estimación de σ^2.

En algunos casos se puede aplicar alguna transformación para conseguir homocedasticidad (ver Peña[pág. 59][54]).

En cuanto a efectuar un contraste formal de igualdad de varianzas antes del test F, es mejor utilizar un test robusto frente a la falta de normalidad como el test de Levene (ver 6.7.3 y Ugarte[69, pág. 375]). Si los datos se desvían ligeramente de la normalidad, esto va a afectar poco al test F, pero mucho a un test de varianzas no robusto, muy dependiente de la normalidad.

Finalmente, el efecto de dependencia entre observaciones puede ser muy grave, ya que las fórmulas para las varianzas de las distribuciones muestrales de las medias son inválidas en este caso, por lo que todos los cálculos sobre la precisión de los estimadores serán erróneos. El procedimiento más eficaz para prevenir la dependencia es la aleatorización.

Ejemplo 10.6.1

Con el modelo lineal propuesto en el ejemplo 10.2.1 se pueden realizar los gráficos de los residuos que se presentan en la figura 10.2. No se observan patologías que hagan dudar de las hipótesis básicas del modelo.

Ejemplo 10.6.2

Con los datos del ejemplo 10.3.1 y el modelo más simple tras el contraste se calculan (ver página 214) los residuos estandarizados que tenemos en la tabla 10.8. Sólo hay un residuo

	prod	fert	finca	ajustado	resid	resid.std	atipico
1	2.1	A	1	2.00	0.10	0.57	
2	2.2	A	2	2.00	0.20	1.14	
3	1.8	A	3	2.00	-0.20	-1.14	
4	2.0	A	4	2.00	0.00	0.00	
5	1.9	A	5	2.00	-0.10	-0.57	
6	2.2	B	1	2.56	-0.36	-2.04	*
7	2.6	B	2	2.56	0.04	0.23	
8	2.7	B	3	2.56	0.14	0.80	
9	2.5	B	4	2.56	-0.06	-0.34	
10	2.8	B	5	2.56	0.24	1.36	
11	1.8	C	1	1.84	-0.04	-0.23	
12	1.9	C	2	1.84	0.06	0.34	
13	1.6	C	3	1.84	-0.24	-1.36	
14	2.0	C	4	1.84	0.16	0.91	
15	1.9	C	5	1.84	0.06	0.34	
16	2.1	D	1	2.16	-0.06	-0.34	
17	2.0	D	2	2.16	-0.16	-0.91	
18	2.2	D	3	2.16	0.04	0.23	
19	2.4	D	4	2.16	0.24	1.36	
20	2.1	D	5	2.16	-0.06	-0.34	

Cuadro 10.8: Residuos estandarizados del ejemplo 10.3.1

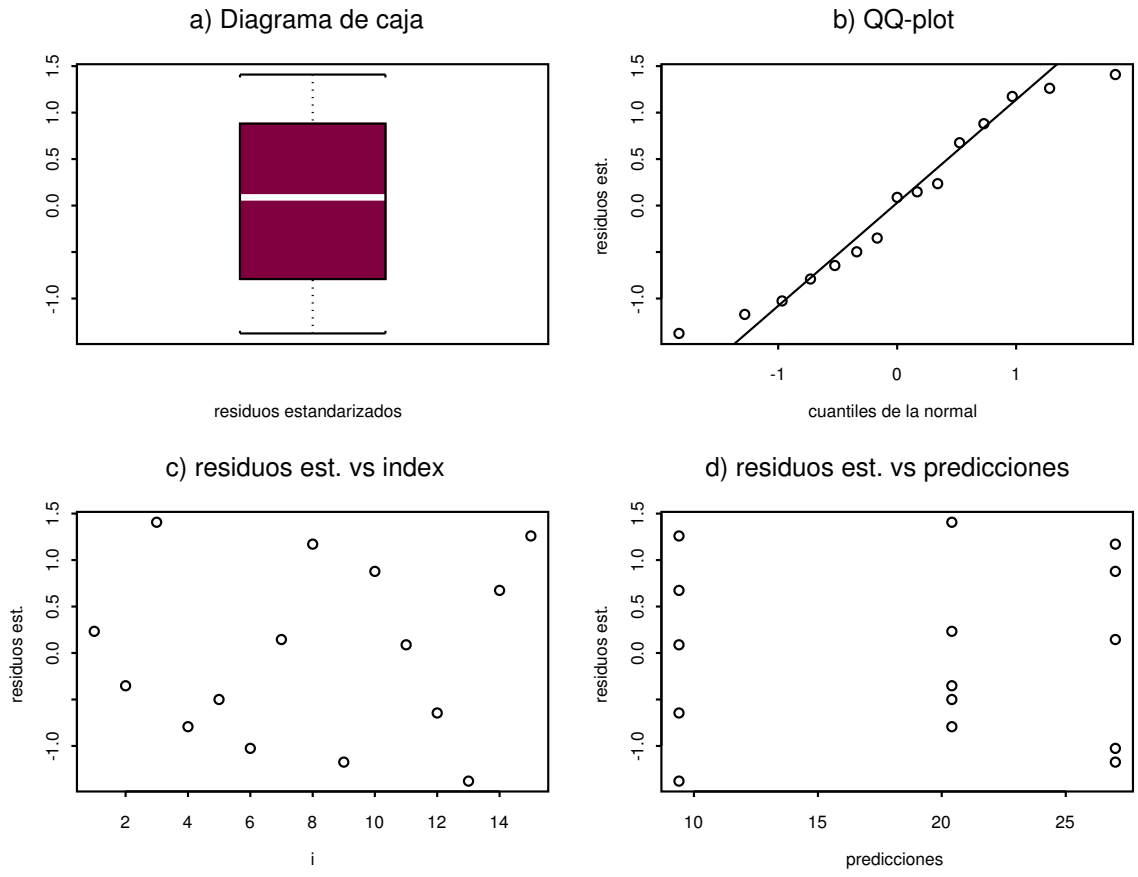


Figura 10.2: Gráficos para el análisis de los residuos del ejemplo 10.2.1

ligeramente atípico, situado en las colas del 5 %, y marcado con un asterisco.
Los gráficos de estos residuos, como en el ejemplo anterior, tampoco muestran ningún indicio de apartarse de las hipótesis básicas del modelo lineal.

Ejemplo 10.6.3

En el análisis de los residuos del ejemplo 10.4.1 con el modelo simplificado a un factor, el factor significativo siembra, se observan dos residuos estandarizados con valores atípicos: $-2{,}84$ *(*$p < 0{,}007$*) y* $-2{,}61$ *(*$p < 0{,}01$*). Habría que estudiar estos dos resultados.*
Los gráficos pueden verse en la figura 10.3.

10.7. Diseños no balanceados y observaciones faltantes

Un diseño experimental (observaciones y modelo del experimento) puede describirse mediante el modelo lineal $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, donde $\mathbf{X}$ es la matriz de diseño ampliada. Sean $n_1, \ldots, n_k$ los números de réplicas para cada una de las condiciones experimentales (ver sección 2.7). Excepto el diseño de un factor, los demás diseños deben tener el mismo número de réplicas por condición experimental. Sin embargo, en las aplicaciones no siempre es posible mantener tal restricción. Además, las réplicas de alguna condición experimental

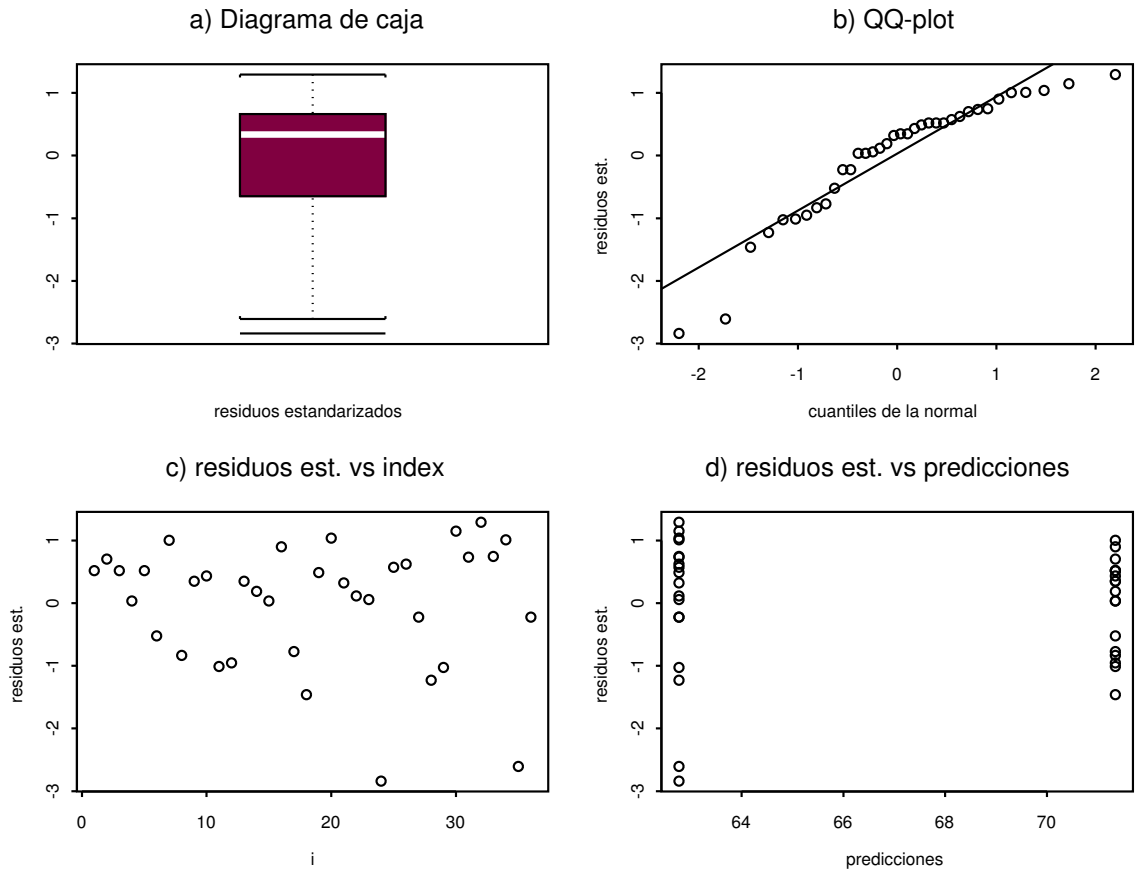


Figura 10.3: Gráficos para el análisis de los residuos del ejemplo 10.4.1

pueden perderse (un tubo de ensayo que se rompe, unos datos que se extravían, etc.). Veamos como pueden ser tratados ambos problemas.
Dado el modelo lineal $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, diremos que corresponde a:

1) Un diseño balanceado si $n_1 = n_2 = \cdots = n_k \neq 0$.

2) Un diseño no balanceado si $n_i \neq n_j$ para algún i, j.

3) Un diseño con observaciones faltantes si $n_i = 0$ para algún i.

Supongamos que $\mathbf{X}_R$ es la matriz de diseño reducida "estándar" para un diseño experimental determinado. Los diseños no balanceados y con observaciones faltantes se pueden manejar, sin modificar $\mathbf{X}_R$, utilizando

$$\mathbf{D} = \text{diag}(n_1, n_2, \ldots, n_k)$$

Adoptemos el convenio de que si $n_i = 0$ para algún i, la correspondiente observación contenida en $\mathbf{Y}$ se sustituye por 0 y en el vector de medias $\bar{\mathbf{Y}} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k)'$ se toma $\bar{y}_i = 0$. Entonces se verifica

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}_R'\mathbf{D}\mathbf{X}_R)^{-}\mathbf{X}_R'\mathbf{D}\bar{\mathbf{Y}}$$

$$\text{SCR} = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}_R'\mathbf{D}\bar{\mathbf{Y}}$$

$$\text{SCR}_H - \text{SCR} = (\mathbf{A}\widehat{\boldsymbol{\beta}})'(\mathbf{A}(\mathbf{X}_R'\mathbf{D}\mathbf{X}_R)^{-}\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}})$$

siendo $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{0}$ una hipótesis contrastable. La matriz $\mathbf{M}$ que relaciona $\mathbf{X}$ con $\mathbf{X}_R$ mediante $\mathbf{X} = \mathbf{M}\mathbf{X}_R$ se define como en la sección 2.7, pero añadiendo una fila de ceros en el lugar correspondiente a una casilla con observaciones faltantes (ver Cuadras[20]). Para otros tratamientos del caso no balanceado y de las observaciones faltantes véase Seber[65, pág. 259,290-300].

Ejemplo 10.7.1

Consideremos un diseño de dos factores A, B sin interacción, con $a = 2$, $b = 3$, $n_{11} = 1$, $n_{12} = 2$, $n_{13} = 0$, $n_{21} = 3$, $n_{22} = 0$, $n_{23} = 1$; es decir, no balanceado y con observaciones faltantes en los niveles A_1B_3 y A_2B_2. Entonces, para los parámetros $\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3$, tenemos:

$$\mathbf{X}_R = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \qquad \mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{D} = (1, 2, 0, 3, 1, 0)$$

$$\mathbf{X} = \mathbf{M}\mathbf{X}_R = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

10.8. Ejemplos con R

Empezamos por reproducir el ejemplo 10.2.1 con el diseño de un factor. En primer lugar introducimos los datos en una tabla. Observemos en especial la definición del vector `tratam` como factor.

```
> y<-c(22,18,30,15,17,20,28,35,19,33,10,5,0,14,18)
> tratam<-factor(c(rep("D",5),rep("B",5),rep("P",5)))
> pacientes<-data.frame(y,tratam)
> design.table(pacientes)
   B  D  P
1 20 22 10
2 28 18  5
3 35 30  0
```

```
4 19 15 14
5 33 17 18
```

A continuación podemos presentar algún gráfico de los datos como un diagrama de cajas o un gráfico de puntos (ver figura de la tabla 10.2).

```
> par(pty="s")
> boxplot(split(y,tratam))
> par(pty="s")
> dotplot(formula=tratam~y,data=pacientes)
```

El Análisis de la Varianza se realiza con la función `aov`.

```
> aov(y~tratam,data=pacientes)
Call:
   aov(formula = y ~ tratam, data = pacientes)

Terms:
                  tratam Residuals
 Sum of Squares 790.5333  558.4000
Deg. of Freedom        2        12

Residual standard error: 6.821535 Estimated effects are balanced
```

Aunque para obtener la tabla 10.3 del Análisis de la Varianza es mejor asignar el resultado en la forma

```
> pacientes.aov<-aov(y~tratam,data=pacientes)
> summary(pacientes.aov)
          Df Sum of Sq  Mean Sq  F Value       Pr(F)
   tratam  2  790.5333 395.2667 8.494269 0.005031871
Residuals 12  558.4000  46.5333
```

Observemos que el estadístico F es significativo, de forma que se rechaza la hipótesis nula de igualdad de tratamientos.
Las estimaciones de los parámetros α_i son $\hat{\alpha}_i = y_{i\cdot} - \bar{y}$

```
> model.tables(pacientes.aov)

Tables of effects

 tratam
       B       D       P
  8.0667  1.4667 -9.5333
```

y las estimaciones de las medias $\hat{\mu} = \bar{y}, \hat{\mu} + \hat{\alpha}_i = y_{i\cdot}$ son

```
> model.tables(pacientes.aov,type="mean")

Tables of means Grand mean
```

```
 18.933

 tratam
    B    D    P
 27.0 20.4  9.4
```

Los residuos estandarizados de este modelo se calculan fácilmente:

```
> ECM<-deviance(pacientes.aov)/pacientes.aov$df.residual;ECM
[1] 46.53333
> resstd<-residuals(pacientes.aov)/sqrt(ECM)
```

Nos interesa además señalar los residuos atípicos, si los hay. Para ello definimos una escala de tres estrellas que corresponde a las colas de probabilidad 0,007, 0,01 y 0,05, respectivamente.

```
> outlier<-as.character(ifelse(abs(resstd)>2.698,"***",
+ ifelse(abs(resstd)>2.576,"**",
+ ifelse(abs(resstd)>1.96,"*"," ))))
```

La siguiente tabla muestra los resultados:

```
> cbind(pacientes,ajustado=fitted(pacientes.aov),
+ resid=round(residuals(pacientes.aov),2),
+ resid.std=round(resstd,2),atipico=outlier)
    y tratam ajustado resid resid.std atipico
 1 22      D     20.4   1.6      0.23
 2 18      D     20.4  -2.4     -0.35
 3 30      D     20.4   9.6      1.41
 4 15      D     20.4  -5.4     -0.79
 5 17      D     20.4  -3.4     -0.50
 6 20      B     27.0  -7.0     -1.03
 7 28      B     27.0   1.0      0.15
 8 35      B     27.0   8.0      1.17
 9 19      B     27.0  -8.0     -1.17
10 33      B     27.0   6.0      0.88
11 10      P      9.4   0.6      0.09
12  5      P      9.4  -4.4     -0.65
13  0      P      9.4  -9.4     -1.38
14 14      P      9.4   4.6      0.67
15 18      P      9.4   8.6      1.26
```

Los gráficos de la figura 10.2 se han obtenido con las siguientes instrucciones:

```
> par(mfrow=c(2,2))
> boxplot(resstd,xlab="residuos estandarizados")
> title("a) Diagrama de caja")
> qqnorm(resstd,xlab="cuantiles de la normal",ylab="residuos est.")
> qqline(resstd)
> title("b) QQ-plot")
```

```
> plot((1:length(resstd)),resstd,type="p",xlab="i",ylab="residuos est.")
> title("c) residuos est. vs index")
> plot(fitted(pacientes.aov),resstd,xlab="predicciones",ylab="residuos est.")
> title("d) residuos est. vs predicciones")
```

Veamos ahora la reproducción del ejemplo 10.3.1. Primero la introducción de los datos:

```
> prod<-c(2.1,2.2,1.8,2.0,1.9,2.2,2.6,2.7,2.5,2.8,1.8,
+ 1.9,1.6,2.0,1.9,2.1,2.0,2.2,2.4,2.1)
> fert<-c(rep(1,5),rep(2,5),rep(3,5),rep(4,5))
> fert<-factor(fert,labels=c("A","B","C","D"))
> finca<-rep(c(1,2,3,4,5),4)
> finca<-factor(finca)
> problema<-data.frame(prod,fert,finca)
> rm(prod,fert,finca)
> problema
```

Con la última instrucción veremos la tabla con todos los datos.
Ahora podemos presentar algunos gráficos descriptivos como los de la figura 10.4.

```
> par(mfrow=c(1,3),pty="s")
> plot.factor(prod~fert,data=problema)
> title("a) Diagrama de cajas")
> plot.factor(prod~finca,data=problema)
> title("b) Diagrama de cajas")
> interaction.plot(finca,fert,prod)
> title("c) Poligonos")
```

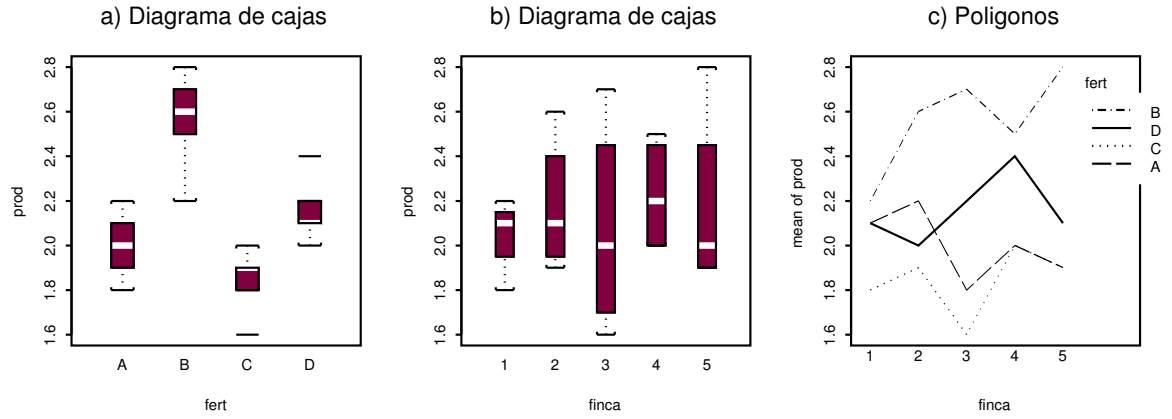


Figura 10.4: Gráficos para la exploración de los datos del ejemplo 10.3.1

También se pueden obtener otros gráficos con las instrucciones:

```
> plot.design(prod~fert,fun="mean")    # Medias de prod por fert
> plot.design(problema,fun="median")   # Medianas de prod
> dotplot(fert ~ prod | finca, data = problema)
```

Con la última se obtiene un conjunto de gráficos de puntos, uno para cada nivel del factor bloque, en este caso las fincas.

A continuación se calcula el Análisis de la Varianza:

```
> attach(problema)
> problema.aov<-aov(prod~fert+finca,data=problema)
> summary(problema.aov)
          Df Sum of Sq   Mean Sq  F Value     Pr(F)
     fert  3     1.432 0.4773333 14.03922 0.0003137
    finca  4     0.088 0.0220000  0.64706 0.6395716
Residuals 12     0.408 0.0340000
```

Ahora se pueden calcular las estimaciones de los efectos y las medias.

```
> efectos<-model.tables(problema.aov);efectos

Tables of effects

 fert
     A     B     C     D
 -0.14  0.42 -0.30  0.02

 finca
      1      2      3      4      5
 -0.090  0.035 -0.065  0.085  0.035
> medias<-model.tables(problema.aov,type="means");medias

Tables of means
Grand mean

 2.14

 fert
    A    B    C    D
 2.00 2.56 1.84 2.16

 finca
     1     2     3     4     5
 2.050 2.175 2.075 2.225 2.175
```

Como el efecto del factor bloque no es significativo, podemos evaluar la tabla del Análisis de la Varianza del modelo con el factor principal:

```
> simple.aov<-aov(prod~fert,data=problema)
> summary(simple.aov)
          Df Sum of Sq   Mean Sq  F Value         Pr(F)
     fert  3     1.432 0.4773333 15.39785 0.00005624767
Residuals 16     0.496 0.0310000
```

El análisis de los residuos debe hacerse con `simple.aov`.

```
> ECM<-deviance(simple.aov)/simple.aov$df.residual;ECM
[1] 0.031
> resstd<-residuals(simple.aov)/sqrt(ECM)
> outlier<-as.character(ifelse(abs(resstd)>2.698,"***",
+ ifelse(abs(resstd)>2.576,"**",
+ ifelse(abs(resstd)>1.96,"*"," ")))
> cbind(problema,ajustado=fitted(simple.aov),
+ resid=round(residuals(simple.aov),2),
+ resid.std=round(resstd,2),atipico=outlier)
```

El resultado puede verse en la tabla 10.8 de la página 207. Los gráficos del análisis de estos residuos se pueden realizar con las mismas instrucciones que en el ejemplo anterior y son muy parecidos a los de la figura 10.2.

Veamos ahora cómo se procede con los datos del ejemplo 10.4.1.

```
> huevos<-c(93,94,93,90,93,86,
+ 95.5,83.5,92,92.5,82,82.5,
+ 92,91,90,95,84,78,
+ 83.3,87.6,81.9,80.1,79.6,49.4,
+ 84,84.4,77,67,69.1,88.4,
+ 85.3,89.4,85.4,87.4,52,77)
> genotipo<-c(rep(1,6),rep(2,6),rep(3,6),rep(1,6),rep(2,6),rep(3,6))
> siembra<-c(rep(1,18),rep(2,18))
> genotipo<-factor(genotipo,labels=c("++","+-","--"))
> siembra<-factor(siembra,labels=c("100","800"))
> y<-asin(sqrt(huevos/100))
> y<-y*180/pi
> split(round(y,2),genotipo)
...
> problema<-data.frame(y,siembra,genotipo)
> rm(y,siembra,genotipo)
> attach(problema)
> par(mfrow=c(2,3))
> plot.factor(y~siembra,data=problema)
> title("a) Diagrama de cajas")
> plot.factor(y~genotipo,data=problema)
> title("b) Diagrama de cajas")
> plot.design(problema,fun="mean")
> title("c) Medias")
> plot.design(problema,fun="median")
> title("d) Medianas")
> interaction.plot(genotipo,siembra,y)
> title("e) Poligonos")
```

Este conjunto de gráficos puede verse en la figura 10.1. Se intuye la falta de diferencias significativas entre los genotipos, mientras hay una clara diferencia entre las dos siembras. También es evidente la no interacción entre los dos factores.

A continuación resolvemos el Análisis de la Varianza con los dos factores y su interacción. Observemos que la fórmula que se introduce en la función `aov` es

```
        siembra + genotipo + siembra:genotipo == siembra*genotipo
```

```
> problema.aov<-aov(y~siembra*genotipo,data=problema)
> summary(problema.aov)
                Df Sum of Sq  Mean Sq  F Value     Pr(F)
         siembra  1   662.086 662.0865 14.83286 0.0005736
        genotipo  2     7.665   3.8323  0.08585 0.9179521
siembra:genotipo  2    35.354  17.6772  0.39603 0.6764562
       Residuals 30  1339.094  44.6365

> medias<-model.tables(problema.aov,type="means");medias
```

El análisis de los residuos se hace con el modelo simple. Ver figura 10.3.

```
> simple.aov<-aov(y~siembra,data=problema)
> summary(simple.aov)
          Df Sum of Sq  Mean Sq  F Value      Pr(F)
  siembra  1   662.086 662.0865 16.28734 0.00029224
Residuals 34  1382.113  40.6504
> ECM<-deviance(simple.aov)/simple.aov$df.residual;ECM
[1] 40.65038
> resstd<-residuals(simple.aov)/sqrt(ECM)
> par(mfrow=c(2,2))
> boxplot(resstd,xlab="residuos estandarizados")
> title("a) Diagrama de caja")
> qqnorm(resstd,xlab="cuantiles de la normal",ylab="residuos est.")
> qqline(resstd)
> title("b) QQ-plot")
> plot((1:length(resstd)),resstd,type="p",xlab="i",ylab="residuos est.")
> title("c) residuos est. vs index")
> plot(fitted(simple.aov),resstd,xlab="predicciones",ylab="residuos est.")
> title("d) residuos est. vs predicciones")
> outlier<-as.character(ifelse(abs(resstd)>2.698,"***",
+ ifelse(abs(resstd)>2.576,"**",
+ ifelse(abs(resstd)>1.96,"*",""))))
> cbind(problema,ajustado=fitted(simple.aov),
+ resid=round(residuals(simple.aov),2),
+ resid.std=round(resstd,2),atipico=outlier)
```

10.9. Ejercicios

Ejercicio 10.1

Los siguientes datos corresponden a los índices de mortalidad, en un período de 10 años, clasificados por estaciones. Determinar si hay diferencias significativas entre las diferentes estaciones al nivel 0,01.

Invierno	Primavera	Verano	Otoño
9,8	9,0	8,8	9,4
9,9	9,3		9,4
9,8	9,3	8,7	10,3
10,6	9,2	8,8	9,8
9,9	9,4	8,6	9,4
10,7	9,1	8,3	9,6
9,7	9,2	8,8	9,5
10,2	8,9	8,7	9,6
10,9	9,3	8,9	9,5
10,0	9,3		9,4

Por otra parte, ¿difiere significativamente de 10,0 el índice medio registrado en invierno?

Ejercicio 10.2

Para el diseño de un factor con k niveles

$$y_{ih} = \mu + \alpha_i + \epsilon_{ih} \qquad i = 1, \ldots, k;\ h = 1, \ldots, n_i$$

con $\sum \alpha_i = 0$, demostrar:

a) La relación entre el contraste de la razón de verosimilitud Λ y el contraste F para la hipótesis $H_0 : \alpha_1 = \cdots = \alpha_k = 0$ es

$$\Lambda = \left(1 + \frac{k-1}{n-k} F\right)^{-n/2}$$

b) El valor esperado de los cuadrados medios entre grupos es

$$E(\mathrm{CM}_E) = \sigma^2 + \frac{1}{k-1} \sum n_i \alpha_i^2$$

c) Cuando H_0 es cierta y mín$\{n_1, \ldots, n_k\} \to \infty$, entonces $F \xrightarrow{P} 1$.

d) Si $k = 2$, el contraste F para la hipótesis

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

es equivalente al contraste t de Student para comparar las medias $\mu + \alpha_1$, $\mu + \alpha_2$ de dos poblaciones normales suponiendo que las varianzas son iguales.

Ejercicio 10.3

La siguiente tabla registra las producciones de 4 variedades de maíz, plantadas según un diseño en bloques aleatorizados

		Variedad			
		1	2	3	4
Bloque	a	7	6	6	7
	b	10	8	7	9
	c	6	3	5	7
	d	4	3	3	3
	e	8	5	5	6

Al nivel 0,05 estudiar si hay diferencias entre variedades y entre bloques. Comparar la variedad 1 con la variedad 3.

Ejercicio 10.4

En una experiencia agrícola en la que se combina año con genotipo, se admite el siguiente modelo

$$y_{ikr} = \mu + \alpha_i + \beta_k + \gamma_{ik} + \omega_{ir} + \epsilon_{ikr} \tag{10.27}$$

donde y_{ikr} es la longitud de la planta, α_i $i = 1, \ldots, 5$ es el efecto principal del año, β_k $k = 1, 2, 3$ es el efecto principal del genotipo, γ_{ik} es la interacción genotipo $\times$ año, ω_{ir} es una interacción de las réplicas con los años y ϵ_{ikr} es el término de error con distribución $N(0, \sigma^2)$. La tabla 10.9 presenta la descomposición ortogonal de la suma de cuadrados.

	g.l.	SC Y	$Y \times T$	T
A (año)	4	742	412	630
B (genotipo)	2	118	105	110
C (bloque)	3	74	87	97
AB	8	647	630	521
AC	12	454	478	372
BC	6	87	63	79
ABC	24	345	247	270

Cuadro 10.9: Tabla con las sumas de cuadrados para el diseño 10.27

Se pide:

a) Hallar la expresión algebraica del residuo y encontrar tres estimaciones independientes de σ^2.

b) Estudiar si los efectos principales y las interacciones son significativas (nivel 0,05).

Observación: La variable T es una variable concomitante y su utilidad será estudiada en el siguiente capítulo. Por este motivo, las columnas correspondientes a $Y \times T$ y T no tienen interés ahora.

Ejercicio 10.5

En un estudio sobre viabilidad de Drosophila melanogaster se tienen en cuenta los siguientes factores:

Genotipo (G):	se estudian 3 genotipos distintos
Generación (N):	el experimento se repite durante 4 generaciones sucesivas
Temperatura (T):	incubación a 17 y 25 grados centígrados

Se obtuvieron 5 réplicas para cada una de las combinaciones de los 3 factores. El experimento se realizó sembrando 100 huevos y anotando el número de huevos eclosionados (esto constituye una réplica). Después de transformar adecuadamente los datos originales (ver ejemplo 10.5.1), se obtuvo la siguiente descomposición ortogonal de la suma de cuadrados (R es el *factor réplica*)

	SC	g.l.
G	621	2
N	450	3
T	925	1
R	347	4
GN	35	6
GT	210	2
GR	48	8
NT	23	3
NR	34	12
TR	110	4
GNT	75	6
GNR	17	24
GTR	22	8
NTR	11	12
$GNTR$	107	24

Se pide:

a) Sabiendo que las interacciones entre 2 o 3 factores en las que intervenga el factor N no forman parte del modelo lineal asociado al diseño, estudiar la significación de los efectos principales y de las interacciones (nivel de significación: 0,01).

b) Hallar tres estimaciones insesgadas de la varianza σ^2 del diseño estocásticamente independientes.

Capítulo 11

Análisis de Componentes de la Varianza

11.1. Introducción

En los diseños del capítulo anterior hemos supuesto que los efectos de los factores son fijos, elegidos por el experimentador, y por este motivo se denominan modelos de efectos fijos. Se trataba de investigar el efecto que producen algunos niveles fijados sobre la variable respuesta. Sin embargo, en ciertas situaciones es necesario interpretar los efectos de los factores como aleatorios. En estos diseños los niveles no se eligen, sino que se consideran una muestra al azar. A los modelos relacionados con los efectos aleatorios se les denomina modelos de *efectos aleatorios*.

En el caso de diseños con efectos fijos estamos interesados en estimar el efecto de los diversos niveles sobre la respuesta. Por el contrario, en los diseños de efectos aleatorios dicha estimación no tiene sentido y buscaremos saber si el efecto existe y, si así es, conocer su efecto sobre la variabilidad con la estimación de las varianzas asociadas. Por ello este estudio se conoce con el nombre de *Análisis de Componentes de la Varianza*.

También pueden presentarse efectos de ambos tipos en un mismo modelo: son los llamados *modelos mixtos*. Veamos como distinguirlos mediante ejemplos.

Un modelo de efectos fijos

Una experiencia agrícola consistió en comparar la producción de cuatro variedades de maíz. Para ello, se plantaron las cuatro variedades en 40 parcelas idénticas, 10 por variedad. Transcurrido el tiempo necesario se recolectó, estudiándose la variable "peso de maíz por parcela".

Un modelo adecuado para analizar esta experiencia es el de un factor

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad i = 1, 2, 3, 4;\ j = 1, 2, \ldots, 10$$

y_{ij}	es la observación j del nivel i, es decir, la producción de la parcela j de la variedad i
μ	es la media general
α_i	es un parámetro fijo y representa el efecto de la variedad i
ϵ_{ij}	es el error aleatorio con distribución $N(0, \sigma)$

La hipótesis de interés en este estudio es

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$$

es decir, no hay efecto variedad y las cuatro pueden considerarse homogéneas en cuanto a la productividad.

Un modelo de efectos aleatorios

Para determinar el contenido en DNA de los hepatocitos de rata hemos tomado al azar cinco ratas. De cada hígado realizamos tres preparaciones y evaluamos con las técnicas adecuadas la cantidad de DNA por célula.

Un modelo apropiado para estos datos sería también el de un factor

$$y_{ij} = \mu + A_i + \epsilon_{ij} \qquad i = 1, 2, \ldots, 5;\ j = 1, 2, 3$$

pero la diferencia respecto al anterior estriba en que A_i no es un parámetro fijo sino el efecto aleatorio de la rata i que procede de una población de ratas en la cual se supone que la variable (cantidad DNA / célula hepática) sigue una distribución $N(\mu, \sigma_y)$. La distribución de los A_i es $N(0, \sigma_A)$ que se supone independiente de los errores ϵ_{ij} con distribución $N(0, \sigma)$.

La hipótesis de interés en este caso es

$$H_0 : \sigma_A^2 = 0$$

lo que equivale a afirmar que no hay variabilidad entre las distintas ratas de la población respecto la variable estudiada.

Un modelo mixto

Para un estudio sobre la ecología de un lago se han elegido al azar cuatro tardes de verano y se ha medido la variable temperatura a diferentes profundidades (0,1,2,3,4 y 5 metros). Nuestro objetivo es examinar mediante los datos obtenidos si hay diferencias significativas entre profundidades y días.

El modelo adecuado en este caso es el de dos factores sin interacción

$$y_{ij} = \mu + \alpha_i + B_j + \epsilon_{ij} \qquad i = 1, 2, \ldots, 6;\ j = 1, 2, 3, 4$$

y_{ij}	es la temperatura a la profundidad i en el día j
μ	es la media general
α_i	es un parámetro fijo y representa el efecto de la profundidad i
B_j	es el efecto aleatorio del día j y sigue una distribución $N(0, \sigma_B)$
ϵ_{ij}	es el error aleatorio con distribución $N(0, \sigma)$

La hipótesis de que la temperatura no varía con la profundidad es

$$H_0 : \alpha_1 = \cdots = \alpha_6 = 0$$

mientras que la hipótesis de que existe homogeneidad entre los diferentes días del verano es

$$H_0 : \sigma_B^2 = 0$$

11.2. Contraste de hipótesis

El tratamiento mediante Análisis de la Varianza de diseños con efectos aleatorios es, en general, muy similar al caso de efectos fijos en diseños balanceados, existiendo diferencias solamente cuando existen interacciones. En diseños no balanceados el análisis es mucho más complejo.

11.2.1. Los test F

Para realizar los contrastes principales utilizaremos los test F adaptados a cada situación y que justificaremos en la sección 11.3. La tabla 11.1 muestra los cuadrados medios esperados y el cociente a efectuar para obtener la F en diseños de uno y dos factores con efectos fijos, aleatorios o mixtos. Por ejemplo, en el diseño de dos factores sin interacción se verifica

$$E[\mathrm{SCR}_B/(b-1)] = E(\mathrm{CM}_B) = \sigma^2 + \frac{a}{b-1}\sum_j \beta_j^2$$

si los efectos son fijos y

$$E(\mathrm{CM}_B) = \sigma^2 + a\sigma_B^2$$

si los efectos son aleatorios. Observemos que para este diseño y el de un factor, los cocientes F son iguales tanto si se trata de efectos aleatorios como de efectos fijos.

Sin embargo, en el diseño de dos factores con interacción, los cocientes F difieren según el modelo sea de efectos fijos, aleatorios o mixto:

a) El modelo de efectos fijos ya ha sido ampliamente tratado en la sección 10.4.

b) Si los dos factores son aleatorios, los cocientes F que deben calcularse para las distintas hipótesis son

$$H_0 : \sigma_A^2 = 0 \qquad F = \frac{\mathrm{SCR}_A/(a-1)}{\mathrm{SCR}_I/[(a-1)(b-1)]}$$

$$H_0' : \sigma_B^2 = 0 \qquad F = \frac{\mathrm{SCR}_B/(b-1)}{\mathrm{SCR}_I/[(a-1)(b-1)]}$$

$$H_0'' : \sigma_{AB}^2 = 0 \qquad F = \frac{\mathrm{SCR}_I/[(a-1)(b-1)]}{\mathrm{SCR}/[ab(r-1)]}$$

En los dos primeros casos es necesario dividir por la interacción para hallar la F. En efecto, si H_0 es cierta $\sigma_A^2 = 0$ y entonces $\mathrm{SCR}_A/(\sigma^2 + r\sigma_{AB}^2)$ y $\mathrm{SCR}_I/(\sigma^2 + r\sigma_{AB}^2)$ siguen distribuciones ji-cuadrado independientes con $a-1$ y $(a-1)(b-1)$ grados de libertad respectivamente. Luego

$$F = \frac{\mathrm{CM}_A}{\mathrm{CM}_I}$$

sigue la distribución F con $a-1$ y $(a-1)(b-1)$ grados de libertad. Observemos que el término desconocido $\sigma^2 + r\sigma_{AB}^2$ desaparece. Podemos realizar consideraciones análogas para H_0' y H_0''.

Cuadro 11.1: Tabla de los cuadrados medios esperados y el cociente a efectuar para obtener la F en diseños de uno y dos factores con efectos fijos, aleatorios o mixtos

		EFECTOS FIJOS		EFECTOS ALEATORIOS		MIXTOS (A fijo,B aleatorio)	
	suma de cuadrados	cuadrados medios esperados	F	cuadrados medios esperados	F	cuadrados medios esperados	F
un factor	SCR_A	$\sigma^2 + \frac{1}{k-1}\sum n_i\alpha_i^2$	$\mathrm{CM}_A/\mathrm{CM}_R$	$\sigma^2 + n_0\sigma_A^2$ $(n_0 = n_1 = \ldots = n_k)$	$\mathrm{CM}_A/\mathrm{CM}_R$		
	SCR	σ^2		σ^2			
dos factores	SCR_A	$\sigma^2 + \frac{b}{a-1}\sum \alpha_i^2$	$\mathrm{CM}_A/\mathrm{CM}_R$	$\sigma^2 + b\sigma_A^2$	$\mathrm{CM}_A/\mathrm{CM}_R$	$\sigma^2 + \frac{b}{a-1}\sum \alpha_i^2$	$\mathrm{CM}_A/\mathrm{CM}_R$
	SCR_B	$\sigma^2 + \frac{a}{b-1}\sum \beta_j^2$	$\mathrm{CM}_B/\mathrm{CM}_R$	$\sigma^2 + b\sigma_B^2$	$\mathrm{CM}_B/\mathrm{CM}_R$	$\sigma^2 + a\sigma_B^2$	$\mathrm{CM}_B/\mathrm{CM}_R$
	SCR	σ^2		σ^2		σ^2	
dos factores con interacción	SCR_A	$\sigma^2 + \frac{br}{a-1}\sum \alpha_i^2$	$\mathrm{CM}_A/\mathrm{CM}_R$	$\sigma^2 + r\sigma_{AB}^2 + br\sigma_A^2$	$\mathrm{CM}_A/\mathrm{CM}_I$	$\sigma^2 + r\sigma_{AB}^2 + \frac{br\sum \alpha_i^2}{a-1}$	$\mathrm{CM}_A/\mathrm{CM}_I$
	SCR_B	$\sigma^2 + \frac{ar}{b-1}\sum \beta_j^2$	$\mathrm{CM}_B/\mathrm{CM}_R$	$\sigma^2 + r\sigma_{AB}^2 + ar\sigma_B^2$	$\mathrm{CM}_B/\mathrm{CM}_I$	$\sigma^2 + ar\sigma_B^2$	$\mathrm{CM}_B/\mathrm{CM}_R$
	SCR_I	$\sigma^2 + \frac{r\sum \gamma_{ij}^2}{(a-1)(b-1)}$	$\mathrm{CM}_I/\mathrm{CM}_R$	$\sigma^2 + r\sigma_{AB}^2$	$\mathrm{CM}_I/\mathrm{CM}_R$	$\sigma^2 + r\sigma_{AB}^2$	$\mathrm{CM}_I/\mathrm{CM}_R$
	SCR	σ^2		σ^2		σ^2	

c) Si A es fijo y B es aleatorio, los cocientes F a efectuar son

$$H_0 : \alpha_1 = \cdots = \alpha_a = 0 \qquad F = \frac{\mathrm{SCR}_A/(a-1)}{\mathrm{SCR}_I/[(a-1)(b-1)]}$$

$$H_0' : \sigma_B^2 = 0 \qquad F = \frac{\mathrm{SCR}_B/(b-1)}{\mathrm{SCR}/[ab(r-1)]}$$

$$H_0'' : \sigma_{AB}^2 = 0 \qquad F = \frac{\mathrm{SCR}_I/[(a-1)(b-1)]}{\mathrm{SCR}/[ab(r-1)]}$$

En este caso solamente el efecto principal de A debe ser dividido por la interacción. En efecto, si H_0 es cierta $\alpha_i = 0$ $i = 1, \ldots, a$ y entonces $\mathrm{SCR}_A/(\sigma^2 + r\sigma_{AB}^2)$ y $\mathrm{SCR}_I/(\sigma^2 + r\sigma_{AB}^2)$ siguen distribuciones ji-cuadrado independientes. Al realizar el cociente para obtener la F desaparece el término $\sigma^2 + r\sigma_{AB}^2$.

En cambio, para $\sigma_B^2 = 0$ (H_0' cierta), tenemos que

$$\mathrm{SCR}_B/\sigma^2 \qquad \mathrm{SCR}_I/(\sigma^2 + \sigma_{AB}^2) \qquad \mathrm{SCR}/\sigma^2$$

siguen distribuciones ji-cuadrado independientes entre sí con $b-1$, $(a-1)(b-1)$ y $ab(r-1)$ g.l. respectivamente. Luego es necesario para obtener la F realizar el cociente entre CM_B/σ^2 y CM_R/σ^2 de modo que el término desconocido σ^2 desaparezca. Observemos que dividiendo por la interacción los términos σ^2 y $\sigma^2 + \sigma_{AB}^2$ no se anulan, imposibilitando el cálculo de la F.

Ejemplo 11.2.1

Se desea estudiar y comparar la acción de tres fármacos tranquilizantes A, B C en la conducción de automóviles. La variable que sirvió de referencia fue el tiempo que un individuo tarda en iniciar la frenada ante la puesta repentina en rojo de un semáforo. Se eligieron 8 *hombres al azar y se sometió a cada hombre a los* 3 *tratamientos, en períodos sucesivos y secuencias al azar, mediante el procedimiento del doble ciego (ni el médico ni el paciente saben cual es el fármaco suministrado en un determinado momento). Los resultados fueron, en milésimas de segundo (cada dato es el promedio de varias observaciones):*

		1	2	3	4	5	6	7	8
	A	548	619	641	846	517	876	602	628
Tratamiento	B	519	776	678	858	493	741	719	595
	C	637	818	701	855	618	849	731	687

Como hay tres tratamientos fijos y ocho individuos elegidos al azar de la población, nos encontramos ante un diseño mixto, donde el efecto individuo (efecto bloque) es aleatorio. Las hipótesis a contemplar son

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 \qquad \textit{(no hay efecto tratamiento)}$$
$$H_0' : \sigma_B^2 = 0 \qquad \textit{(no hay homogeneidad entre individuos)}$$

donde σ_B^2 es la varianza del efecto individuo. La tabla del Análisis de la Varianza es

Fuente de variación	*suma de cuadrados*	*g.l.*	*cuadrados medios*	F
Entre tratam.	27535,6	2	13767,79	5,15
Entre individuos	258040,7	7	36862,95	13,78
Residuo	37451,1	14	2675,08	
Total	323027,4	23		

Para 2 *y* 14 *g.l.* $F = 5{,}15$ *es significativa al nivel* 0,025*, aceptamos pues que hay diferencias entre fármacos. Para* 7 *y* 14 *g.l.* $F = 13{,}78$ *es significativa al nivel* 0,005*, aceptamos que hay variabilidad entre individuos.*

11.2.2. Estimación de los componentes de la varianza

Una estimación aproximada de las varianzas σ^2, σ_A^2, σ_B^2, σ_{AB}^2 se puede obtener igualando los cuadrados medios con los cuadrados medios esperados y resolviendo el sistema resultante. Por ejemplo, en el diseño de un factor tenemos

$$\begin{array}{rcl} \widehat{\sigma}^2 + n_0\widehat{\sigma}_A^2 & = & \mathrm{CM}_A \\ \widehat{\sigma}^2 & = & \mathrm{CM}_R \end{array}$$

y para el diseño de dos factores con interacción

$$\begin{array}{rcl} \widehat{\sigma}^2 + r\widehat{\sigma}_{AB}^2 + br\widehat{\sigma}_A^2 & = & \mathrm{CM}_A \\ \widehat{\sigma}^2 + r\widehat{\sigma}_{AB}^2 + ar\widehat{\sigma}_B^2 & = & \mathrm{CM}_B \\ \widehat{\sigma}^2 + r\widehat{\sigma}_{AB}^2 & = & \mathrm{CM}_I \\ \widehat{\sigma}^2 & = & \mathrm{CM}_R \end{array}$$

Puede ocurrir que la estimación puntual de un componente de la varianza resulte negativa. En este caso aceptaremos que su valor es cero dado que la varianza es un parámetro estrictamente positivo.

Ejemplo 11.2.2

Para estimar la variabilidad entre individuos del ejemplo anterior, igualaremos los cuadrados medios a sus valores esperados

$$\begin{array}{rcl} 36862{,}95 & = & \widehat{\sigma}^2 + 3\widehat{\sigma}_B^2 \\ 2675{,}08 & = & \widehat{\sigma}^2 \end{array}$$

de donde

$$\widehat{\sigma}_B^2 = (36862{,}95 - 2675{,}08)/3 = 11395{,}96$$

El tiempo de frenado entre los individuos varía con una desviación típica estimada $\widehat{\sigma}_B = 106{,}75$ *milésimas de segundo.*

11.3. Comparación entre modelos de efectos fijos y modelos de efectos aleatorios

En esta sección vamos a probar los principales resultados teóricos que desembocan en los test F prescritos en cada caso para modelos sencillos, tanto de efectos fijos como de efectos aleatorios . A los modelos de efectos fijos los denominaremos también modelos de tipo I y a los de efectos aleatorios modelos de tipo II.

11.3.1. Diseño de un factor con efectos fijos

Tal como se ha visto en la sección 10.2, el modelo lineal que se adapta a este diseño es

$$y_{ij} = \mu_i + \epsilon_{ij}$$

o, reparametrizado,

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad i = 1, \ldots, k;\ j = 1, \ldots, n_i$$

con la restricción $\sum_{i=1}^{k} \alpha_i = 0$. Las y_{ij} son independientes y normales $N(\mu_i, \sigma)$. Las ϵ_{ij} son independientes y normales $N(0, \sigma)$.
La descomposición de la variabilidad viene dada por

$$\sum_{i,j}(y_{ij} - \bar{y})^2 = \sum_{i,j}(y_{i\cdot} - \bar{y})^2 + \sum_{i,j}(y_{ij} - y_{i\cdot})^2$$

es decir

$$\mathrm{SC}_T = \mathrm{SC}_e + \mathrm{SC}_d$$

o también

$$\mathrm{SCR}_H = (\mathrm{SCR}_H - \mathrm{SCR}) + \mathrm{SCR}$$

con $n-1$, $k-1$ y $n-k$ grados de libertad respectivamente, siendo $n_1 + \cdots + n_k = n$.

Teorema 11.3.1

El valor esperado de la suma de cuadrados entre grupos es

$$E(\mathrm{SC}_e) = (k-1)\sigma^2 + \sum_{i=1}^{k} n_i \alpha_i^2$$

luego

$$E(\mathrm{CM}_e) = E\left(\frac{\mathrm{SC}_e}{k-1}\right) = \sigma^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i \alpha_i^2$$

Demostración:
Por definición $\mathrm{SC}_e = \sum_{i=1}^{k} n_i (y_{i\cdot} - \bar{y})^2$.
Del modelo $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ se obtiene

$$y_{i\cdot} = \mu + \alpha_i + \epsilon_{i\cdot}$$
$$\bar{y} = \mu + \epsilon_{\cdot\cdot}$$

ya que $\sum_{i=1}^{k} \alpha_i = 0$ y en consecuencia $\alpha_\cdot = (1/k)\sum_{i=1}^{k} \alpha_i = 0$.
Entonces

$$\begin{aligned}
\mathrm{SC}_e &= \sum_{i=1}^{k} n_i(\alpha_i + \epsilon_{i\cdot} - \epsilon_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{k} n_i \alpha_i^2 + \sum_{i=1}^{k} n_i \epsilon_{i\cdot}^2 + n\epsilon_{\cdot\cdot}^2 + 2\sum_{i=1}^{k} n_i \alpha_i \epsilon_{i\cdot} \\
&\quad - 2\epsilon_{\cdot\cdot}\sum_{i=1}^{k} n_i \alpha_i - 2\epsilon_{\cdot\cdot}\sum_{i=1}^{k} n_i \epsilon_{i\cdot}
\end{aligned}$$

pero

$$\epsilon_{\cdot\cdot}\sum_{i=1}^{k} n_i\epsilon_{i\cdot} = \epsilon_{\cdot\cdot}\sum_{i=1}^{k} n_i\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\epsilon_{ij}\right) = \epsilon_{\cdot\cdot}\sum_{i,j}\epsilon_{ij} = n\epsilon_{\cdot\cdot}^2$$

luego

$$\begin{aligned} E(\mathrm{SC}_e) = &\sum_{i=1}^{k} n_i\alpha_i^2 + \sum_{i=1}^{k} n_i E(\epsilon_{i\cdot}^2) + n\,E(\epsilon_{\cdot\cdot}^2) \\ &+ 2\sum_{i=1}^{k} n_i\alpha_i E(\epsilon_{i\cdot}) - 2\left(\sum_{i=1}^{k} n_i\alpha_i\right)E(\epsilon_{\cdot\cdot}) \\ &- 2n\,E(\epsilon_{\cdot\cdot}^2) \end{aligned}$$

Recordando que las v.a. ϵ_{ij} son independientes y normales $N(0,\sigma)$ se verifica

$$\epsilon_{i\cdot} \sim N(0,\sigma/\sqrt{n_i}) \qquad \epsilon_{\cdot\cdot} \sim N(0,\sigma/\sqrt{n})$$

Por ser centradas, la esperanza de su cuadrado coincide con la varianza, es decir

$$E(\epsilon_{i\cdot}^2) = \mathrm{var}(\epsilon_{i\cdot}) = \frac{\sigma^2}{n_i}$$
$$E(\epsilon_{\cdot\cdot}^2) = \mathrm{var}(\epsilon_{\cdot\cdot}) = \frac{\sigma^2}{n}$$

y por tanto

$$\begin{aligned} E(\mathrm{SC}_e) &= \sum_{i=1}^{k} n_i\alpha_i^2 + \sum_{i=1}^{k} n_i\frac{\sigma^2}{n_i} + n\frac{\sigma^2}{n} - 2n\frac{\sigma^2}{n} \\ &= \sum_{i=1}^{k} n_i\alpha_i^2 + k\sigma^2 + \sigma^2 - 2\sigma^2 \\ &= (k-1)\sigma^2 + \sum_{i=1}^{k} n_i\alpha_i^2 \end{aligned}$$

■

Teorema 11.3.2

El valor esperado de la suma de cuadrados dentro de los grupos es

$$E(\mathrm{SC}_d) = (n-k)\sigma^2$$

y por lo tanto

$$E(\mathrm{CM}_d) = E\left(\frac{\mathrm{SC}_d}{n-k}\right) = \sigma^2$$

Demostración:
Teniendo en cuenta que $\mathrm{SC}_d = \mathrm{SCR}$, este resultado es evidente y ya se probó en el teorema 2.5.1 para un modelo lineal general. También se puede demostrar siguiendo un proceso parecido al del teorema anterior. ■

Caso particular

Si el diseño es balanceado, es decir, igual número de réplicas por condición experimental ($n_1 = \cdots = n_k = n_0$), entonces de los teoremas 11.3.1 y 11.3.2 se deducen las fórmulas

$$E(\mathrm{CM}_e) = \sigma^2 + \frac{n_0}{k-1}\sum_{i=1}^{k}\alpha_i^2$$
$$E(\mathrm{CM}_d) = E\left(\frac{\mathrm{SC}_d}{k(n_0-1)}\right) = \sigma^2$$

Inferencia en el modelo de un factor con efectos fijos

La hipótesis nula de mayor interés es

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

o, utilizando el modelo alternativo,

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$

Por el teorema 11.3.1, CM_e es un estimador insesgado de σ^2 si H_0 es cierta. Por el teorema 11.3.2 es siempre un estimador insesgado de σ^2, sea H_0 cierta o no. Además, suponiendo que $\epsilon_{ij} \sim N(0, \sigma)$, se verifica el teorema 5.3.1 de la teoría general del modelo lineal normal (Teorema fundamental del Análisis de la Varianza) como hemos visto en 10.3:

a) $\mathrm{SC}_d/\sigma^2 \sim \chi^2_{n-k}$

b) Si H_0 es cierta, entonces $\mathrm{CM}_e = \mathrm{SC}_e/(k-1)$ es otra estimación insesgada de σ^2 y además
$$\mathrm{SC}_e/\sigma^2 \sim \chi^2_{k-1}$$

c) Si H_0 es cierta, el estadístico
$$F = \frac{\mathrm{SC}_e/[\sigma^2(k-1)]}{\mathrm{SC}_d/[\sigma^2(n-k)]} = \frac{\mathrm{CM}_e}{\mathrm{CM}_d}$$
sigue la distribución F con $k-1$ y $n-k$ grados de libertad. La hipótesis H_0 se rechaza si el estadístico es significativo.

11.3.2. Diseño de un factor con efectos aleatorios

El modelo lineal que se adapta a este diseño es

$$y_{ij} = \mu + A_i + \epsilon_{ij} \qquad i = 1, \ldots, k;\ j = 1, \ldots, n_i$$

con las siguientes particularidades

1) $E(A_i) = 0$, $\mathrm{var}(A_i) = \sigma_A^2 \quad i = 1, \ldots, k$

2) $E(A_i \cdot A_{i'}) = 0 \quad \forall i \neq i'$

3) $E(A_i \cdot \epsilon_{ij}) = 0 \quad \forall i, j$

es decir, $\{A_i\}$ son variables aleatorias de media cero y varianza σ_A^2, independientes entre sí y de los errores $\{\epsilon_{ij}\}$. Luego

$$\begin{aligned} \text{var}(y_{ij}) &= \text{var}(A_i) + \text{var}(\epsilon_{ij}) \\ \sigma_y^2 &= \sigma_A^2 + \sigma^2 \end{aligned}$$

y por este motivo es apropiado denominar *componentes de la varianza* a σ_A^2 y σ^2.
Para su tratamiento clásico mediante Análisis de la Varianza de un factor es necesario además que

4) $A_i \sim N(0, \sigma_A)$, $\epsilon_{ij} \sim N(0, \sigma)$ y por lo tanto $y_{ij} \sim N(\mu, \sigma_y)$

5) el diseño sea balanceado $n_1 = n_2 = \cdots = n_k = n_0$

Este modelo de efectos aleatorios que hemos formulado y en general cualquier modelo de efectos aleatorios, difiere de un modelo de efectos fijos en que bajo las asunciones realizadas

a) Para un i dado, todas las observaciones tienen igual esperanza

$$E(y_{ij}) = \mu + A_i \qquad \forall j$$

b) Para un i dado, las observaciones no son estocásticamente independientes entre sí.

c) La variable $\sum_{i=1}^{k} A_i$ es aleatoria y puede tomar un valor distinto de cero.

Teorema 11.3.3

Para el diseño de un factor con efectos aleatorios el valor esperado de la suma de cuadrados entre grupos es

$$E(\text{SC}_e) = (k-1)\sigma^2 + n_0(k-1)\sigma_A^2$$

luego

$$E(\text{CM}_e) = E\left(\frac{\text{SC}_e}{k-1}\right) = \sigma^2 + n_0\sigma_A^2$$

Demostración:
Por definición $\text{SC}_e = n_0 \sum_{i=1}^{k} (y_{i\cdot} - \bar{y})^2$.
Del modelo se obtiene

$$\begin{aligned} y_{i\cdot} &= \mu + A_i + \epsilon_{i\cdot} \\ \bar{y} &= \mu + A_\cdot + \epsilon_{\cdot\cdot} \end{aligned}$$

de donde

$$\begin{aligned} \text{SC}_e &= n_0 \sum_{i=1}^{k} [(A_i - A_\cdot) + (\epsilon_{i\cdot} - \epsilon_{\cdot\cdot})]^2 \\ &= n_0 \left[\sum_{i=1}^{k} A_i^2 + \sum_{i=1}^{k} A_\cdot^2 - 2A_\cdot \sum_{i=1}^{k} A_i + \sum_{i=1}^{k} \epsilon_{i\cdot}^2 \right. \\ &\quad \left. + k\epsilon_{\cdot\cdot}^2 - 2\epsilon_{\cdot\cdot} \sum_{i=1}^{k} \epsilon_{i\cdot} + 2 \sum_{i=1}^{k} (A_i - A_\cdot)(\epsilon_{i\cdot} - \epsilon_{\cdot\cdot}) \right] \end{aligned}$$

pero

$$\sum_{i=1}^{k} \epsilon_{i\cdot} = \sum_{i=1}^{k} \frac{1}{n_0} \sum_{j=1}^{n_0} \epsilon_{ij} = \frac{1}{n_0} \sum_{i=1}^{k} \sum_{j=1}^{n_0} \epsilon_{ij} = \frac{1}{n_0} k n_0 \epsilon_{\cdot\cdot} = k\epsilon_{\cdot\cdot}$$

ya que

$$\epsilon_{\cdot\cdot} = \frac{1}{kn_0} \sum_{i=1}^{k} \sum_{j=1}^{n_0} \epsilon_{ij}$$

Entonces

$$\mathrm{SC}_e = n_0 \left[\sum_{i=1}^{k} A_i^2 + kA_{\cdot}^2 + \sum_{i=1}^{k} \epsilon_{i\cdot}^2 - k\epsilon_{\cdot\cdot}^2 + 2 \sum_{i=1}^{k} (A_i - A_{\cdot})(\epsilon_{i\cdot} - \epsilon_{\cdot\cdot}) \right]$$

$$\begin{aligned} E(\mathrm{SC}_e) &= n_0 \sum_{i=1}^{k} E(A_i^2) - n_0 k E(A_{\cdot}^2) + n_0 \sum_{i=1}^{k} E(\epsilon_{i\cdot}^2) \\ &\quad - n_0 k E(\epsilon_{\cdot\cdot}^2) + 2n_0 \sum_{i=1}^{k} E[(A_i - A_{\cdot})(\epsilon_{i\cdot} - \epsilon_{\cdot\cdot})] \end{aligned}$$

Por las hipótesis del modelo se verifica

$$A_{\cdot} \sim N(0, \sigma_A/\sqrt{k}) \qquad \epsilon_{i\cdot} \sim N(0, \sigma/\sqrt{n_0}) \qquad \epsilon_{\cdot\cdot} \sim N(0, \sigma/\sqrt{kn_0})$$

Debido a que las variables aleatorias A_i, $A_{\cdot}$, $\epsilon_{i\cdot}$, $\epsilon_{\cdot\cdot}$ son centradas, la esperanza de su cuadrado coincide con su varianza, es decir,

$$\begin{aligned} E(A_i^2) &= \mathrm{var}(A_i) = \sigma_A^2 \\ E(A_{\cdot}^2) &= \mathrm{var}(A_{\cdot}) = \sigma_A^2/k \\ E(\epsilon_{i\cdot}^2) &= \mathrm{var}(\epsilon_{i\cdot}) = \sigma^2/n_0 \\ E(\epsilon_{\cdot\cdot}^2) &= \mathrm{var}(\epsilon_{\cdot\cdot}) = \sigma^2/(kn_0) \end{aligned}$$

Además, al ser independientes las variables A_i con las ϵ_{ij}

$$E[(A_i - A_{\cdot})(\epsilon_{i\cdot} - \epsilon_{\cdot\cdot})] = E(A_i - A_{\cdot}) \cdot E(\epsilon_{i\cdot} - \epsilon_{\cdot\cdot}) = 0 \cdot 0 = 0$$

Por lo tanto

$$\begin{aligned} E(\mathrm{SC}_e) &= n_0 k \sigma_A^2 - n_0 k \frac{\sigma_A^2}{k} + n_0 k \frac{\sigma^2}{n_0} - n_0 k \frac{\sigma^2}{kn_0} \\ &= n_0 k \sigma_A^2 - n_0 \sigma_A^2 + k\sigma^2 - \sigma^2 \\ &= (k-1)\sigma^2 + n_0(k-1)\sigma_A^2 \end{aligned}$$

■

Teorema 11.3.4

El valor esperado de la suma de cuadrados dentro de los grupos es

$$E(\mathrm{SC}_d) = k(n_0 - 1)\sigma$$

es decir

$$E(\mathrm{CM}_d) = E\left(\frac{\mathrm{SC}_d}{k(n_0 - 1)} \right) = \sigma^2$$

Demostración:
Por definición $\mathrm{SC}_e = \sum_{i=1}^{k}\sum_{j=1}^{n_0}(y_{ij} - y_{i\cdot})^2$.
Del modelo se obtiene

$$y_{i\cdot} = \mu + A_i + \epsilon_{i\cdot}$$

Entonces

$$\begin{aligned}
\mathrm{SC}_d &= \sum_{i=1}^{k}\sum_{j=1}^{n_0}(\epsilon_{ij} - \epsilon_{i\cdot})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_0}\epsilon_{ij}^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_0}\epsilon_{i\cdot}^2 - 2\sum_{i=1}^{k}\sum_{j=1}^{n_0}\epsilon_{i\cdot}\epsilon_{ij} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_0}\epsilon_{ij}^2 + n_0\sum_{i=1}^{k}\epsilon_{i\cdot}^2 - 2\sum_{i=1}^{k}\epsilon_{i\cdot}\sum_{j=1}^{n_0}\epsilon_{ij} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_0}\epsilon_{ij}^2 + n_0\sum_{i=1}^{k}\epsilon_{i\cdot}^2 - 2\sum_{i=1}^{k}\epsilon_{i\cdot}n_0\epsilon_{i\cdot} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_0}\epsilon_{ij}^2 - n_0\sum_{i=1}^{k}\epsilon_{i\cdot}^2
\end{aligned}$$

de manera que

$$\begin{aligned}
E(\mathrm{SC}_d) &= \sum_{i=1}^{k}\sum_{j=1}^{n_0}E(\epsilon_{ij}^2) - n_0\sum_{i=1}^{k}E(\epsilon_{i\cdot}^2) \\
&= kn_0\sigma^2 - n_0k\frac{\sigma^2}{n_0} \\
&= kn_0\sigma^2 - k\sigma^2 \\
&= k(n_0 - 1)\sigma^2 \qquad \blacksquare
\end{aligned}$$

Inferencia en el modelo de un factor con efectos aleatorios

La hipótesis de interés en este modelo es

$$H_0 : \sigma_A^2 = 0$$

Recordemos que

$$\begin{aligned}
\mathrm{SC}_A &= n_0\sum_{i=1}^{k}(y_{i\cdot} - \bar{y})^2 = n_0\sum_{i=1}^{k}(A_i + \epsilon_{i\cdot} - A_{\cdot} - \epsilon_{\cdot\cdot})^2 \\
\mathrm{SCR} &= \sum_{i,j}(y_{ij} - y_{i\cdot})^2 = \sum_{i,j}(\epsilon_{ij} - \epsilon_{i\cdot})^2
\end{aligned}$$

siendo SC_A la suma de cuadrados entre grupos o suma de cuadrados del factor y SCR la suma de cuadrados dentro de los grupos o suma de cuadrados residual, representadas hasta ahora por SC_e y SC_d respectivamente. Recuérdese también que $A_{\cdot}$ es una variable aleatoria y en consecuencia susceptible de tomar un valor distinto de cero.

Realizando el cambio $g_i = A_i + \epsilon_{i\cdot}$ obtenemos k v.a. independientes con distribución normal de media cero y varianza

$$\mathrm{var}(g_i) = \mathrm{var}(A_i) + \mathrm{var}(\epsilon_{i\cdot}) = \sigma_A^2 + \frac{\sigma^2}{n_0}$$

Por el teorema de Fisher, la variable aleatoria

$$ks_g^2/\sigma_g^2$$

se distribuye según una ji-cuadrado con $k-1$ g.l., es decir,

$$\frac{\sum_{i=1}^k (g_i - \bar{g})^2}{\sigma_A^2 + \frac{\sigma^2}{n_0}} = \frac{n_0 \sum_{i=1}^k (g_i - \bar{g})^2}{n_0\sigma_A^2 + \sigma^2} = \frac{\mathrm{SC}_A}{n_0\sigma_A^2 + \sigma^2} \sim \chi_k^2$$

Entonces

$$\begin{aligned} \mathrm{SC}_A &= (n_0\sigma_A^2 + \sigma^2) \cdot \chi_{k-1}^2 \\ E(\mathrm{CM}_A) = E\left(\frac{\mathrm{SC}_A}{k-1}\right) &= n_0\sigma_A^2 + \sigma^2 \end{aligned}$$

A este resultado habíamos llegado también anteriormente por el teorema 11.3.3.

Por otra parte, SCR está distribuida de idéntica forma que en los modelos de efectos fijos. Los ϵ_{ij} desempeñan el papel de las observaciones, con media cero y varianza σ^2. Luego

$$\begin{aligned} \mathrm{SCR} &= \sigma^2 \cdot \chi_{k(n_0-1)}^2 \\ E(\mathrm{CM}_R) = E\left(\frac{\mathrm{SCR}}{k(n_0-1)}\right) &= \sigma^2 \end{aligned}$$

Para efectuar comparaciones falta demostrar que SC_A y SCR son independientes. Para ello, basta probar la independencia entre $A_i + \epsilon_{i\cdot} - A_\cdot - \epsilon_{\cdot\cdot}$ y $\epsilon_{ij} - \epsilon_{i\cdot}$. Tenemos que $A_i - A_\cdot$ y $\epsilon_{ij} - \epsilon_{i\cdot}$ son obviamente independientes. Si expresamos $\epsilon_{ij} = \epsilon_{\cdot\cdot} + (\epsilon_{i\cdot} - \epsilon_{\cdot\cdot}) + (\epsilon_{ij} - \epsilon_{i\cdot})$, utilizando otra vez la analogía con los modelos de efectos fijos, $\epsilon_{i\cdot} - \epsilon_{\cdot\cdot}$ pertenece al espacio de las estimaciones y $\epsilon_{ij} - \epsilon_{i\cdot}$ pertenece al espacio error, espacios que son ortogonales entre sí. Debido a la normalidad del modelo, sus vectores son independientes, luego SC_A y SCR son independientes. Entonces, si H_0 es cierta, el estadístico

$$F = \frac{\mathrm{SC}_A/[\sigma^2(k-1)]}{\mathrm{SCR}/[\sigma^2 k(n_0-1)]} = \frac{\mathrm{SC}_A/(k-1)}{\mathrm{SCR}/[k(n_0-1)]} = \frac{\mathrm{CM}_A}{\mathrm{CM}_R}$$

sigue la distribución F con $k-1$ y $k(n_0-1)$ g.l.. La hipótesis H_0 se rechaza si el estadístico es significativo.

Como resumen de lo expuesto en los apartados anteriores véase el cuadro 11.2. Obsérvese que, si bien la hipótesis a contrastar del modelo I es formalmente distinta de la hipótesis del modelo II, se utiliza el mismo estadístico de contraste

$$F = \frac{\mathrm{CM}_A}{\mathrm{CM}_R} \sim F_{k-1,k(n_0-1)}$$

Una estimación de los componentes de la varianza es

$$\widehat{\sigma}^2 = \mathrm{CM}_R \qquad \widehat{\sigma}_A^2 = \frac{\mathrm{CM}_A - \mathrm{CM}_R}{n_0}$$

solución obtenida resolviendo el sistema resultante de igualar los cuadrados medios con los cuadrados medios esperados (ver sección 11.2.2). Obsérvese que los estimadores $\widehat{\sigma}^2$ y $\widehat{\sigma}_A^2$ son siempre estimadores insesgados de los parámetros σ^2 y σ_A^2 respectivamente.

Fuente de variación	g.l.	cuadrados medios	Esperanza del cuadrado medio	
			Modelo I	Modelo II
Tratamientos	$k-1$	$\mathrm{CM}_A = \mathrm{SC}_A/(k-1)$	$\sigma^2 + \dfrac{n_0 \sum \alpha_i^2}{k-1}$	$\sigma^2 + n_0\sigma_A^2$
Error	$k(n_0-1)$	$\mathrm{CM}_R = \mathrm{SCR}/[k(n_0-1)]$	σ^2	σ^2
Total	n_0k-1			

Cuadro 11.2: Tabla comparativa para diseños de un factor con efectos fijos y efectos aleatorios

11.3.3. Diseño de dos factores sin interacción con efectos fijos o diseño en bloques al azar completos

Este diseño recibe también el nombre de *bloques aleatorizados*. Un desarrollo típico para este diseño, utilizando tres tratamientos en cuatro bloques, es el siguiente

Bloque 1	Bloque 2	Bloque 3	Bloque 4
t_3	t_2	t_1	t_1
t_1	t_1	t_2	t_3
t_2	t_3	t_3	t_2

Las letras t indican la asignación aleatoria de los tratamientos en los bloques. Como ejemplo véase el ejemplo 10.3.1.

Generalizando, consideremos el caso de a tratamientos en b bloques. La observación y_{ij} indica la respuesta del i-ésimo tratamiento aplicado al j-ésimo bloque. Se supondrá que y_{ij} $(i = 1, \ldots, a;\ j = 1, \ldots, b)$ son valores de v.a. independientes con distribución normal de media μ_{ij} y varianza común σ^2. Serán de utilidad también

$$\begin{array}{lcl} y_{i\cdot} & = & \text{media del } i\text{-ésimo tratamiento} \\ y_{\cdot j} & = & \text{media del } j\text{-ésimo bloque} \\ y_{\cdot\cdot} & = & \text{media general} \end{array}$$

El promedio de las medias poblacionales para el i-ésimo tratamiento está definido por

$$\mu_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b} \mu_{ij}$$

Asimismo, el promedio de las medias poblacionales para el j-ésimo bloque está definido por

$$\mu_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a} \mu_{ij}$$

y el promedio de las ab medias poblacionales es

$$\mu_{\cdot\cdot} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}$$

Si representamos por A al factor tratamiento y por B al factor bloque, las hipótesis lineales de interés son

$$\begin{aligned} H_0^A &: \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{a\cdot} = \mu \\ H_0^B &: \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot b} = \mu \end{aligned}$$

Si se cumple la primera hipótesis, el factor A no es significativo o, equivalentemente, no existen diferencias significativas entre los tratamientos. También se dice que no hay efecto fila. En el caso de que se cumpla la segunda hipótesis, el factor B no es significativo, es decir, no existen diferencias significativas entre los bloques; no hay efecto columna.

Cada observación puede descomponerse en

$$y_{ij} = \mu_{ij} + \epsilon_{ij}$$

donde ϵ_{ij} mide la desviación del valor observado y_{ij} frente la media poblacional μ_{ij}. La forma más común de expresar esta ecuación se obtiene al sustituir

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

donde α_i es el efecto del i-ésimo tratamiento y β_j el efecto del j-ésimo bloque. Se supone que los efectos del tratamiento y del bloque son aditivos. Así, el modelo es

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

Obsérvese que se asemeja al modelo de un criterio de clasificación, pero con la adición del efecto bloque. Ahora la variación se controla sistemáticamente en dos direcciones.

Si se imponen las restricciones naturales

$$\sum_{i=1}^{a}\alpha_i = 0 \qquad \sum_{j=1}^{b}\beta_j = 0$$

entonces

$$\begin{aligned} \mu_{i\cdot} &= \frac{1}{b}\sum_{j=1}^{b}(\mu + \alpha_i + \beta_j) = \mu + \alpha_i \\ \mu_{\cdot j} &= \frac{1}{a}\sum_{i=1}^{a}(\mu + \alpha_i + \beta_j) = \mu + \beta_j \end{aligned}$$

Las hipótesis pueden ahora plantearse del siguiente modo

$$\begin{aligned} H_0^A &: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \\ H_0^B &: \beta_1 = \beta_2 = \cdots = \beta_b = 0 \end{aligned}$$

En la sección 10.3 se vio que la descomposición fundamental de la suma de cuadrados (descomposición de la variabilidad) viene dada por

$$\begin{aligned}\sum_{i,j}(y_{ij}-\bar{y})^2 &= b\sum_i(y_{i\cdot}-\bar{y})^2 + a\sum_j(y_{\cdot j}-\bar{y})^2 \\ &\quad + \sum_{i,j}(y_{ij}-y_{i\cdot}-y_{\cdot j}+\bar{y})^2\end{aligned}$$

es decir

$$\mathrm{SC}_T = \mathrm{SC}_F + \mathrm{SC}_C + \mathrm{SCR}$$

donde SC_T es la suma de cuadrados total, SC_F la suma de cuadrados entre filas, SC_C la suma de cuadrados entre columnas y SCR la suma de cuadrados residual.

Teorema 11.3.5

El valor esperado de la suma de cuadrados entre filas es

$$E(\mathrm{SC}_F) = (a-1)\sigma^2 + b\sum_{i=1}^{a}\alpha_i^2$$

luego

$$E(\mathrm{CM}_F) = E(\mathrm{SC}_F/(a-1)) = \sigma^2 + \frac{b}{a-1}\sum_{i=1}^{a}\alpha_i^2$$

Demostración:
Es análoga a la del teorema 11.3.1.

Teorema 11.3.6

El valor esperado de la suma de cuadrados entre columnas es

$$E(\mathrm{SC}_C) = (b-1)\sigma^2 + a\sum_{j=1}^{b}\beta_j^2$$

luego

$$E(\mathrm{CM}_C) = E(\mathrm{SC}_C/(b-1)) = \sigma^2 + \frac{a}{b-1}\sum_{j=1}^{b}\beta_j^2$$

Demostración:
Es análoga a la del teorema 11.3.1.

Teorema 11.3.7

El valor esperado de la suma de cuadrados residual es

$$E(\mathrm{SCR}) = (a-1)(b-1)\sigma^2$$

luego

$$E(\mathrm{CM}_R) = E(\mathrm{SCR}/[(a-1)(b-1)]) = \sigma^2$$

Demostración:
Es análoga a la del teorema 11.3.2.

Inferencia en el diseño de dos factores sin interacción con efectos fijos

Una de las hipótesis a contrastar es

$$H_0^A : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$

Por el teorema 11.3.5, CM_F es un estimador insesgado de σ^2 si H_0^A es cierta. Por el teorema 11.3.7, SCR es siempre un estimador insesgado de σ^2, tanto si H_0^A es cierta como si no lo es. Además, suponiendo que $\epsilon_{ij} \sim N(0, \sigma)$, se verifica el teorema 5.3.1 de la teoría general del modelo lineal formal:

a) $\mathrm{SCR}/\sigma^2 \sim \chi^2_{(a-1)(b-1)}$

b) Si H_0^A es cierta, entonces $\mathrm{CM}_F = \mathrm{SC}_F/(a-1)$ es otra estimación insesgada de σ^2 y además
$$\mathrm{SC}_F/\sigma^2 \sim \chi^2_{a-1}$$

c) Si H_0^A es cierta, el estadístico
$$F = \frac{\mathrm{SC}_F/[\sigma^2(a-1)]}{\mathrm{SCR}/[\sigma^2(a-1)(b-1)]} = \frac{\mathrm{CM}_F}{\mathrm{CM}_R}$$
sigue la distribución F con $a-1$ y $(a-1)(b-1)$ g.l.. La hipótesis H_0^A se rechaza si el estadístico es significativo.

Otra hipótesis a contrastar es

$$H_0^B : \beta_1 = \beta_2 = \cdots = \beta_b = 0$$

Análogamente al caso anterior, el estadístico

$$F = \frac{\mathrm{SC}_C/[\sigma^2(b-1)]}{\mathrm{SCR}/[\sigma^2(a-1)(b-1)]} = \frac{\mathrm{CM}_C}{\mathrm{CM}_R}$$

sigue la distribución F con $b-1$ y $(a-1)(b-1)$ g.l.. La hipótesis H_0^B se rechaza si el estadístico es significativo.

11.3.4. Diseño de dos factores sin interacción con efectos aleatorios

El modelo lineal que se adapta a este diseño es

$$y_{ij} = \mu + A_i + B_j + \epsilon_{ij} \qquad i = 1, \ldots, a;\ j = 1, \ldots, b$$

siendo A_i, B_j, ϵ_{ij} variables aleatorias normales independientes con media cero y varianzas $\sigma_A^2, \sigma_B^2, \sigma^2$ respectivamente. La descomposición fundamental de la suma de cuadrados (descomposición de la variabilidad) viene dada por

$$\begin{aligned}\sum_{i,j}(y_{ij} - \bar{y})^2 &= b\sum_i (y_{i\cdot} - \bar{y})^2 + a\sum_j (y_{\cdot j} - \bar{y})^2 \\ &\quad + \sum_{i,j}(y_{ij} - y_{i\cdot} - y_{\cdot j} + \bar{y})^2\end{aligned}$$

es decir

$$\mathrm{SC}_T = \mathrm{SC}_F + \mathrm{SC}_C + \mathrm{SCR}$$

Teorema 11.3.8

El valor esperado de la suma de cuadrados entre filas es

$$E(\mathrm{SC}_F) = (a-1)\sigma^2 + b(a-1)\sigma_A^2$$

luego

$$E(\mathrm{CM}_F) = E(\mathrm{SC}_F/(a-1)) = \sigma^2 + b\sigma_A^2$$

Demostración:
Es análoga a la del teorema 11.3.3.

Teorema 11.3.9

El valor esperado de la suma de cuadrados entre columnas es

$$E(\mathrm{SC}_C) = (b-1)\sigma^2 + a(b-1)\sigma_B^2$$

luego

$$E(\mathrm{CM}_C) = E(\mathrm{SC}_C/(b-1)) = \sigma^2 + a\sigma_B^2$$

Demostración:
Es análoga a la del teorema 11.3.3.

Teorema 11.3.10

El valor esperado de la suma de cuadrados residual es

$$E(\mathrm{SCR}) = (a-1)(b-1)\sigma^2$$

luego

$$E(\mathrm{CM}_R) = E(\mathrm{SCR}/[(a-1)(b-1)]) = \sigma^2$$

Demostración:
Es análoga a la del teorema 11.3.4.

Inferencia en el diseño de dos factores sin interacción con efectos aleatorios

Las hipótesis de interés en este modelo son

$$H_0 : \sigma_A^2 = 0 \qquad H_0' : \sigma_B^2 = 0$$

Para contrastar la primera se utiliza el estadístico

$$F = \frac{\mathrm{SC}_F/[\sigma^2(a-1)]}{\mathrm{SCR}/[\sigma^2(a-1)(b-1)]} = \frac{\mathrm{CM}_F}{\mathrm{CM}_R}$$

que sigue bajo H_0 la distribución F con $a-1$ y $(a-1)(b-1)$ g.l.. La hipótesis H_0 se rechaza si el estadístico es significativo.
De manera análoga, para contrastar la segunda hipótesis se utiliza el estadístico

$$F = \frac{\mathrm{SC}_C/[\sigma^2(b-1)]}{\mathrm{SCR}/[\sigma^2(a-1)(b-1)]} = \frac{\mathrm{CM}_C}{\mathrm{CM}_R}$$

Fuente de variación	g.l.	cuadrados medios	Esperanza del cuadrado medio: Modelo I	Esperanza del cuadrado medio: Modelo II
Entre filas	$a-1$	$\mathrm{CM}_F = \mathrm{SC}_F/(a-1)$	$\sigma^2 + \frac{b}{a-1}\sum \alpha_i^2$	$\sigma^2 + b\sigma_A^2$
Entre col.	$b-1$	$\mathrm{CM}_C = \mathrm{SC}_C/(b-1)$	$\sigma^2 + \frac{a}{b-1}\sum \beta_j^2$	$\sigma^2 + a\sigma_B^2$
Error	$(a-1)(b-1)$	$\mathrm{CM}_R = \frac{\mathrm{SCR}}{(a-1)(b-1)}$	σ^2	σ^2
Total	$ab-1$			

Cuadro 11.3: Tabla comparativa para diseños de dos factores con efectos aleatorios y sin interacción

que sigue bajo H_0' la distribución F con $b-1$ y $(a-1)(b-1)$ g.l.. La hipótesis H_0' se rechaza si el estadístico es significativo.

A modo de resumen de lo expuesto en los apartados anteriores, véase el cuadro 11.3.

Las estimaciones *insesgadas* de las componentes de la varianza se obtienen igualando los cuadrados medios a los cuadrados medios esperados y resolviendo el sistema de ecuaciones resultante (ver sección 11.2.2). Las soluciones en este caso son

$$\widehat{\sigma}^2 = \mathrm{CM}_R \qquad \widehat{\sigma}_A^2 = (\mathrm{CM}_F - \mathrm{CM}_R)/b \qquad \widehat{\sigma}_B^2 = (\mathrm{CM}_C - \mathrm{CM}_R)/a$$

verificándose

$$E(\widehat{\sigma}^2) = \sigma^2 \qquad E(\widehat{\sigma}_A^2) = \sigma_A^2 \qquad E(\widehat{\sigma}_B^2) = \sigma_B^2$$

11.3.5. Diseño de dos factores aleatorios con interacción

El modelo lineal que se adapta a este diseño es

$$y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk}$$
$$i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, r$$

siendo A_i, B_j, $(AB)_{ij}$ y ϵ_{ijk} variables aleatorias normales independientes con media cero y varianza σ_A^2, σ_B^2, σ_{AB}^2 y σ^2 respectivamente.

En el cuadro 11.4 figuran las esperanzas de los cuadrados medios tanto para el modelo I como para el modelo II, indicando por modelo I cuando los dos factores son fijos y por modelo II cuando los dos factores son aleatorios. La demostración de las fórmulas de estas esperanzas se hace de forma análoga a la de los teoremas 11.3.5, 11.3.6 y 11.3.7 para el modelo I, y 11.3.8, 11.3.9 y 11.3.10 para el modelo II.

Las hipótesis a contrastar en el modelo II son

$$H_0^A : \sigma_A^2 = 0 \qquad H_0^B : \sigma_B^2 = 0 \qquad H_0^{AB} : \sigma_{AB}^2 = 0$$

Para contrastar la primera se utiliza el estadístico

$$F = \frac{\mathrm{SC}_A/[(a-1)(\sigma^2 + r\sigma^2_{AB})]}{\mathrm{SC}_{AB}/[(a-1)(b-1)(\sigma^2 + r\sigma^2_{AB})]} = \frac{\mathrm{SC}_A/(a-1)}{\mathrm{SC}_{AB}/(a-1)(b-1)} = \frac{\mathrm{CM}_A}{\mathrm{CM}_{AB}}$$

que sigue bajo H_0^A la distribución F con $a-1$ y $(a-1)(b-1)$ g.l.. La hipótesis H_0^A se rechaza si el estadístico es significativo.

De manera análoga para contrastar la segunda hipótesis se utiliza el estadístico

$$F = \frac{\mathrm{SC}_B/[(b-1)(\sigma^2 + r\sigma^2_{AB})]}{\mathrm{SC}_{AB}/[(a-1)(b-1)(\sigma^2 + r\sigma^2_{AB})]} = \frac{\mathrm{SC}_B/(b-1)}{\mathrm{SC}_{AB}/(a-1)(b-1)} = \frac{\mathrm{CM}_B}{\mathrm{CM}_{AB}}$$

que sigue bajo H_0^B la distribución F con $b-1$ y $(a-1)(b-1)$ g.l..

En el contraste de las dos hipótesis anteriores se divide por el cuadrado medio de la interacción; en cambio, para contrastar la tercera hipótesis se divide por el cuadrado medio del error, es decir, se utiliza el estadístico

$$F = \frac{\mathrm{SC}_{AB}/[(a-1)(b-1)\sigma^2]}{\mathrm{SCR}/[ab(r-1)\sigma^2]} = \frac{\mathrm{SC}_{AB}/[(a-1)(b-1)]}{\mathrm{SCR}/[ab(r-1)]} = \frac{\mathrm{CM}_{AB}}{\mathrm{CM}_R}$$

que sigue bajo H_0^{AB} la distribución F con $(a-1)(b-1)$ y $ab(r-1)$ g.l.. La hipótesis H_0^{AB} se rechaza si el estadístico es significativo.

Las estimaciones *insesgadas* de las componentes de la varianza (ver sección 11.2.2) son

$$\widehat{\sigma}^2 = \mathrm{CM}_R \qquad E(\widehat{\sigma}^2) = \sigma^2$$

$$\widehat{\sigma}^2_A = (\mathrm{CM}_A - \mathrm{CM}_{AB})/(br) \qquad E(\widehat{\sigma}^2_A) = \sigma^2_A$$

$$\widehat{\sigma}^2_B = (\mathrm{CM}_B - \mathrm{CM}_{AB})/(ar) \qquad E(\widehat{\sigma}^2_B) = \sigma^2_B$$

$$\widehat{\sigma}^2_{AB} = (\mathrm{CM}_{AB} - \mathrm{CM}_R)/r \qquad E(\widehat{\sigma}^2_{AB}) = \sigma^2_{AB}$$

11.3.6. Diseño de tres factores aleatorios y réplicas

La esperanza de los cuadrados medios se muestra en el cuadro 11.5. De tales esperanzas se deduce que se pueden formar las razones F apropiadas para contrastar las hipótesis relativas a los componentes de la varianza de las interacciones. Sin embargo, para contrastar las hipótesis relativas a los efectos principales, es decir,

$$H_0^A : \sigma^2_A = 0 \qquad H_0^B : \sigma^2_B = 0 \qquad H_0^C : \sigma^2_C = 0$$

no hay una razón F apropiada a menos que uno o más de los componentes de la varianza de la interacción de dos factores no sean significativos. Por ejemplo, supongamos que se ha comprobado previamente la hipótesis $H_0 : \sigma^2_{AC} = 0$ y ha resultado no significativa. Se puede afirmar entonces que el término σ^2_{AC} puede excluirse de todas las esperanzas de los cuadrados medios en las que intervenga. Si deseamos ahora contrastar la hipótesis $H_0^A : \sigma^2_A = 0$ es posible utilizar el estadístico $F = \mathrm{CM}_A/\mathrm{CM}_{AB}$.

En definitiva, si se desea contrastar las hipótesis relativas a los efectos principales, habrá que estudiar primero la significación de los componentes de la varianza relativos a las interacciones.

Fuente de variación	g.l.	cuadrados medios	Esperanza del cuadrado medio Modelo I	Esperanza del cuadrado medio Modelo II
Entre filas	$a-1$	$\mathrm{CM}_A = \frac{\mathrm{SC}_A}{a-1}$	$\sigma^2 + \frac{rb}{a-1}\sum \alpha_i^2$	$\sigma^2 + r\sigma^2_{AB} + br\sigma^2_A$
Entre col.	$b-1$	$\mathrm{CM}_B = \frac{\mathrm{SC}_B}{b-1}$	$\sigma^2 + \frac{ra}{b-1}\sum \beta_j^2$	$\sigma^2 + r\sigma^2_{AB} + ar\sigma^2_B$
Interac.	g^*	$\mathrm{CM}_{AB} = \frac{\mathrm{SC}_{AB}}{g}$	$\sigma^2 + \frac{r}{g}\sum \tau_{ij}$	$\sigma^2 + r\sigma^2_{AB}$
Residuo	$ab(r-1)$	$\mathrm{CM}_R = \frac{\mathrm{SCR}}{ab(r-1)}$	σ^2	σ^2
Total	$abr-1$			$* \; g = (a-1)(b-1)$

Cuadro 11.4: Tabla comparativa para diseños de dos factores con efectos aleatorios y con interacción

Fuente de variación	g.l.	cuadrados medios	Esperanza del cuadrado medio Modelo II
A	$a-1$	CM_A	$\sigma^2 + r\sigma^2_{ABC} + cr\sigma^2_{AB} + br\sigma^2_{AC} + bcr\sigma^2_A$
B	$b-1$	CM_B	$\sigma^2 + r\sigma^2_{ABC} + cr\sigma^2_{AB} + ar\sigma^2_{BC} + acr\sigma^2_B$
C	$c-1$	CM_C	$\sigma^2 + r\sigma^2_{ABC} + br\sigma^2_{AC} + ar\sigma^2_{BC} + abr\sigma^2_C$
AB	$(a-1)(b-1)$	CM_{AB}	$\sigma^2 + r\sigma^2_{ABC} + cr\sigma^2_{AB}$
AC	$(a-1)(c-1)$	CM_{AC}	$\sigma^2 + r\sigma^2_{ABC} + br\sigma^2_{AC}$
BC	$(b-1)(c-1)$	CM_{BC}	$\sigma^2 + r\sigma^2_{ABC} + ar\sigma^2_{BC}$
ABC	$(a-1)(b-1)(c-1)$	CM_{ABC}	$\sigma^2 + r\sigma^2_{ABC}$
Residuo	$abc(r-1)$	CM_R	σ^2
Total	$abcr-1$		

Cuadro 11.5: Tabla para diseños de tres factores con efectos aleatorios

11.3.7. Diseño anidado de dos factores aleatorios

En muchas situaciones experimentales los niveles o elementos observados de un factor aleatorio no pueden ser los mismos para cada nivel del factor aleatorio principal. Por ejemplo, cuando queremos estudiar algún resultado académico y el factor principal son las diversas universidades, pero los resultados se observan en estudiantes de dichas uni-

versidades como segundo factor aleatorio, entonces los estudiantes son necesariamente distintos (ver figura 11.1). En estos casos no se pueden cruzar los factores y se debe trabajar con los llamados *diseños jerarquizados* o *diseños anidados*.

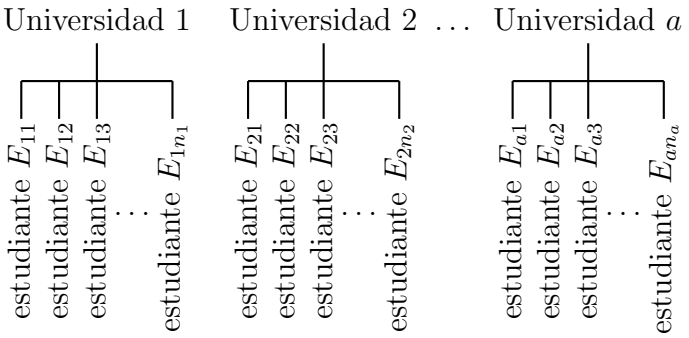


Figura 11.1: Niveles en un ejemplo de diseño de clasificación jerárquica

Como no hay cruces entre niveles de los factores, no puede haber interacciones y el modelo es el siguiente

$$\begin{aligned} &y_{ijk} = \mu + A_i + B_{j(i)} + \epsilon_{ijk} \\ &i = 1, \ldots, a;\ j = 1, \ldots, n_i;\ k = 1, \ldots, r \end{aligned}$$

siendo A_i, $B_{j(i)}$, y ϵ_{ijk} variables aleatorias normales independientes con media cero y varianza σ_A^2, σ_B^2 y σ^2 respectivamente. Así pues, la variabilidad de las observaciones es

$$\operatorname{var}(y_{ijk}) = \sigma_y^2 = \sigma_A^2 + \sigma_B^2 + \sigma^2$$

Observemos que hemos señalado los diversos niveles del segundo factor con el subíndice $j(i)$ para mostrar la jerarquía del primer factor. Como siempre se ha tomado el mismo número de réplicas r para cada situación $j(i)$. También podemos simplificar mucho el modelo si decidimos tomar el mismo número de niveles del segundo factor de forma que $n_1 = \cdots = n_a = b$. En este caso, incluso podemos disponer los datos en una tabla de doble entrada. Pero no nos dejemos engañar, los elementos del segundo factor son distintos.
La descomposición fundamental de la suma de cuadrados es

$$\begin{aligned} \sum_{i,j,k}(y_{ijk} - \bar{y})^2 &= \sum_{i,j,k}(y_{i\cdot\cdot} - \bar{y})^2 + \sum_{i,j,k}(y_{ij\cdot} - y_{i\cdot\cdot})^2 + \sum_{i,j,k}(y_{ijk} - y_{ij\cdot})^2 \\ &= br\sum_i (y_{ijk} - \bar{y})^2 + r\sum_i\sum_j (y_{ij\cdot} - y_{i\cdot\cdot})^2 + \sum_{i,j,k}(y_{ijk} - y_{ij\cdot})^2 \end{aligned}$$

lo que podemos escribir en la forma

$$\mathrm{SC}_T = \mathrm{SC}_A + \mathrm{SC}_{B|A} + \mathrm{SCR}$$

Ahora debemos hallar el valor esperado de cada una de las sumas de cuadrados. Tomando las medias de las observaciones, tenemos

$$\begin{aligned} y_{i\cdot\cdot} &= \mu + A_i + B_{\cdot(i)} + \epsilon_{i\cdot\cdot} \\ \bar{y} &= \mu + A_\cdot + B_{\cdot(\cdot)} + \epsilon_{\cdots} \end{aligned}$$

de modo que

$$y_{i\cdot\cdot} - \bar{y} = A_i - A_\cdot + B_{\cdot(i)} - B_{\cdot(\cdot)} + \epsilon_{i\cdot\cdot} - \epsilon_{\cdot\cdot\cdot}$$

Si elevamos al cuadrado, sumamos para todos los $n = abr$ datos y tomamos esperanzas, resulta

$$E(\mathrm{SC}_A) = brE\left(\sum_i (A_i - A_\cdot)^2\right) + brE\left(\sum_i (B_{\cdot(i)} - B_{\cdot(\cdot)})^2\right) + brE\left(\sum_i (\epsilon_{i\cdot\cdot} - \epsilon_{\cdot\cdot\cdot})^2\right)$$

ya que las esperanzas de los dobles productos son cero porque las variables que intervienen son independientes de media cero.
De manera que si dividimos por los grados de libertad $a-1$ obtenemos

$$E(\mathrm{CM}_A) = E(\mathrm{SC}_A/(a-1)) = br\sigma_A^2 + r\sigma_B^2 + \sigma^2$$

ya que

$$E\left(\frac{1}{a-1}\sum_i (A_i - A_\cdot)^2\right) = \mathrm{var}(A_i) = \sigma_A^2$$
$$E\left(\frac{1}{a-1}\sum_i (B_{\cdot(i)} - B_{\cdot(\cdot)})^2\right) = \mathrm{var}(B_{\cdot(i)}) = \frac{\sigma_B^2}{b}$$
$$E\left(\frac{1}{a-1}\sum_i (\epsilon_{i\cdot\cdot} - \epsilon_{\cdot\cdot\cdot})^2\right) = \mathrm{var}(\epsilon_{i\cdot\cdot}) = \frac{\sigma^2}{br}$$

Del mismo modo, podemos calcular la esperanza de la suma de cuadrados del factor jerarquizado, ya que

$$y_{ij\cdot} - y_{i\cdot\cdot} = B_{j(i)} - B_{\cdot(i)} + \epsilon_{ij\cdot} - \epsilon_{i\cdot\cdot}$$

y resulta

$$E(\mathrm{CM}_{B|A}) = r\sigma_B^2 + \sigma^2$$

En la tabla 11.6 se resume la información relevante para el Análisis de los Componentes de la Varianza de este modelo.

Fuente de variación	g.l.	cuadrados medios	Esperanza del cuadrado medio
A	$a-1$	CM_A	$br\sigma_A^2 + r\sigma_B^2 + \sigma^2$
$B\|A$	$a(b-1)$	$\mathrm{CM}_{B\|A}$	$r\sigma_B^2 + \sigma^2$
Residuo	$ab(r-1)$	CM_R	σ^2
Total	$abr-1$		

Cuadro 11.6: Tabla para el diseño anidado de dos factores con efectos aleatorios

A la vista de estos resultados, la hipótesis $H_0^A : \sigma_A^2 = 0$ se puede contrastar con el estadístico $\mathrm{CM}_A/\mathrm{CM}_{B|A}$, ya que bajo la hipótesis, numerador y denominador tienen el

mismo valor esperado. Del mismo modo, la hipótesis $H_0^B : \sigma_B^2 = 0$ se puede contrastar con el estadístico $\mathrm{CM}_{B|A}/\mathrm{CM}_R$.

En general estas hipótesis son rechazadas y las varianzas no son nulas. Entonces podemos estimar su valor con la resolución del sistema

$$\begin{aligned} \mathrm{CM}_A &= br\sigma_A^2 + r\sigma_B^2 + \sigma^2 \\ \mathrm{CM}_{B|A} &= r\sigma_B^2 + \sigma^2 \\ \mathrm{CM}_R &= \sigma^2 \end{aligned}$$

de donde los estimadores insesgados propuestos son

$$\begin{aligned} \hat{\sigma}_A^2 &= \frac{1}{br}(\mathrm{CM}_A - \mathrm{CM}_{B|A}) \\ \hat{\sigma}_B^2 &= \frac{1}{r}(\mathrm{CM}_{B|A} - \mathrm{CM}_R) \end{aligned}$$

11.3.8. Resumen

Como hemos visto con los diversos modelos estudiados en esta sección, el análisis de ambos tipos de modelos, de efectos fijos o de efectos aleatorios, presenta una base común con planteamientos y interpretaciones diferentes.

En resumen podemos destacar:

1. La formulación del modelo es muy parecida, pero los efectos fijos que representan la respuesta media son parámetros a estimar, mientras que los efectos aleatorios son variables aleatorias normales de media cero y varianzas a estimar. Esto significa que:

   a) Los efectos fijos son constantes y los efectos aleatorios son variables aleatorias.

   b) Los efectos fijos influyen en la respuesta media, mientras que los efectos aleatorios influyen en la variabilidad.

   c) En los efectos aleatorios no tiene sentido imponer restricciones del tipo $\sum \alpha_i = 0$.

   d) Los niveles de un factor de efectos fijos se fijan arbitrariamente por el experimentador, mientras que los niveles de un factor aleatorio son una muestra al azar de una población.

   e) Para un factor de efectos fijos nuestro interés es estimar los parámetros y contrastar su nulidad. Para un factor de efectos aleatorios nos proponemos estimar su varianza y contrastar si es nula.

2. La descomposición de la varianza en sus fuentes de variabilidad, la tabla del Análisis de la Varianza y los contrastes son muy similares para ambos tipos de modelos. Especialmente cuando no hay interacción, en cuyo caso los contrastes son idénticos.

   Cuando hay interacción, en lugar de comparar los cuadrados medios del factor con los cuadrados medios del error, se compara con los cuadrados medios de la interacción.

   Observemos que la tabla del Análisis de la Varianza tiene formalmente la misma apariencia, pero las esperanzas de los cuadrados medios son distintas (ver tabla 11.4).

3. En un modelo de efectos aleatorios el objetivo es contrastar si los efectos existen, a través de sus varianzas, y estimar dichas varianzas. No tiene sentido plantearse contrastes múltiples como en el caso de los efectos fijos.

Por último, debemos advertir que en todos los problemas se precisa una diagnosis del modelo a través de los residuos como se ha visto en la sección 10.6.

11.4. Correlación intraclásica

Sea el modelo de un factor con efectos aleatorios

$$y_{ij} = \mu + A_i + \epsilon_{ij} \qquad i = 1, \ldots, k;\ j = 1, \ldots, n_0$$

donde $\text{var}(A_i) = \sigma_A^2$, $\text{var}(\epsilon_{ij}) = \sigma^2$. Se llama correlación intraclásica al coeficiente de correlación entre dos observaciones $y_{ij}, y_{ij'}$ de un mismo grupo i.
El coeficiente de correlación intraclásica viene dado por

$$\rho_I = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2} \qquad 0 \leq \rho_I \leq 1$$

En efecto

$$\begin{aligned}
\rho_I(y_{ij}, y_{ij'}) &= \frac{\text{cov}(y_{ij}, y_{ij'})}{\sqrt{\text{var}(y_{ij})}\sqrt{\text{var}(y_{ij'})}} \\
&= \frac{E[(y_{ij} - \mu)(y_{ij'} - \mu)]}{\sigma_A^2 + \sigma^2} \\
&= \frac{E(A_i^2 + A_i\epsilon_{ij} + A_i\epsilon_{ij'} + \epsilon_{ij}\epsilon_{ij'})}{\sigma_A^2 + \sigma^2} \\
&= \frac{E(A_i^2)}{\sigma_A^2 + \sigma^2} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2}
\end{aligned}$$

La correlación intraclásica nos expresa el porcentaje de la variabilidad entre grupos respecto la variabilidad total y se utiliza para estudiar la dependencia entre los individuos de un mismo grupo respecto a una variable observable Y. Por ejemplo, es utilizado en Genética descomponiendo la variabilidad total σ_y^2 (varianza de la componente genética) y σ^2 (varianza de la componente ambiental).

Estimación y contraste de significación

Una estimación adecuada de ρ_I es

$$\widehat{\rho}_I = \max\{0, r_I\}$$

siendo

$$r_I = \frac{\widehat{\sigma}_A^2}{\widehat{\sigma}_A^2 + \widehat{\sigma}^2} = \frac{F - 1}{F + n_0 - 1}$$

donde $F = \text{CM}_A/\text{CM}_R$.
Para ver si r_I es significativo hemos de plantear el contraste de la hipótesis $H_0 : \rho_I = 0$ equivalente a $H_0 : \sigma_A^2 = 0$ que se resuelve mediante Análisis de la Varianza.

Ejemplo 11.4.1

En un estudio sobre los guisantes se tomaron 5 *vainas, cada una de las cuales contenía* 8 *guisantes. Los pesos en centigramos fueron*

vaina	1	44	41	42	40	48	46	46	42
	2	43	46	48	42	50	45	45	49
	3	33	34	37	39	32	35	37	41
	4	56	52	50	51	54	52	49	52
	5	36	37	38	40	40	41	44	44

Los datos se asimilan a un diseño de un factor de efectos aleatorios. Las sumas de cuadrados son ($n_0 = 8$)

$$\begin{array}{ll} \mathrm{SC}_A = 1176{,}1 & \textit{con } 4 \textit{ g.l.} \\ \mathrm{SCR} = 273{,}9 & \textit{con } 35 \textit{ g.l.} \end{array}$$

y entonces

$$F = \frac{\mathrm{CM}_A}{\mathrm{CM}_R} = 37{,}57$$

El coeficiente de correlación intraclásica es

$$\widehat{\rho}_I = \max\{0, 0{,}8205\} = 0{,}8205$$

ya que

$$r_I = \frac{F-1}{F+n_0-1} = \frac{36{,}57}{44{,}57} = 0{,}8205$$

Realicemos el contraste de hipótesis para comprobar que es significativo. La hipótesis $H_0 : \rho_I = 0$ *equivale a plantear el contraste* $H_0 : \sigma_A^2 = 0$*, que se resuelve mediante Análisis de la Varianza. Como* $F = 37{,}57$ *con* 4 *y* 35 *g.l. es muy significativa, aceptamos que es distinto de cero. La interpretación en este caso es la siguiente: aproximadamente el* 80 % *de la variabilidad se explica por la componente genética, el resto es debido a factores ambientales.*

11.5. Ejemplos con R

Empecemos con los datos del ejemplo 11.2.1. Observemos la definición del factor aleatorio con la instrucción `is.random` que se aplica a factores o `data.frame`.

```
> tiempo<-c(548,519,637,619,776,818,641,678,701,
+ 846,858,855,517,493,618,876,741,849, + 602,719,731,628,595,687)
> farmaco<-factor(rep(c("A","B","C"),8))
> indiv <- factor(rep(1:8, each=3))
> is.random(indiv) <- T
> frenada<-data.frame(tiempo,farmaco,indiv)
> rm(tiempo,farmaco,indiv)
> attach(frenada)
```

En este caso no hay interacción entre el efecto individuo (bloque) y el efecto fármaco (tratamiento). Por ello los estadísticos F se calculan como en el Análisis de la Varianza con efectos fijos.

```
> frenada.aov<-aov(tiempo~farmaco+indiv,data=frenada)
> summary(frenada.aov)
          Df Sum of Sq  Mean Sq  F Value      Pr(F)
  farmaco  2   27535.6 13767.79  5.14669 0.02110964
    indiv  7  258040.7 36862.95 13.78014 0.00002651
Residuals 14   37451.1  2675.08
```

La estimación de las varianzas y los coeficientes del modelo se consigue con `varcom`:

```
> frenada.var<-varcomp(tiempo~farmaco+indiv,data=frenada)
> summary(frenada.var)
Call: varcomp(formula = tiempo ~ farmaco + indiv, data = frenada)
Variance Estimates:
           Variance
    indiv 11395.958
Residuals  2675.077 Method:  minque0

Coefficients:
 (Intercept) farmaco1 farmaco2
    689.6667    6.375 23.66667
Approximate Covariance Matrix of Coefficients:
            (Intercept) farmaco1 farmaco2
(Intercept)    1535.956    0.000    0.000
   farmaco1       0.000  167.192    0.000
   farmaco2       0.000    0.000   55.731
```

Con el ejemplo 11.4.1 procederemos de otra forma. En primer lugar, introducimos los datos:

```
> peso<-c(44,41,42,40,48,46,46,42,
...
+ 36,37,38,40,40,41,44,44)
> vaina<-factor(rep(1:5, each=8))
> estudio<-data.frame(peso,vaina)
> rm(peso,vaina)
> attach(estudio)
```

Pero ahora no hace falta definir el factor como de efectos aleatorios, ya que vamos a utilizar la función `raov`, que supone que todos los factores son aleatorios.

```
> estudio.raov<-raov(peso~vaina,data=estudio)
> summary(estudio.raov)
          Df Sum of Sq Mean Sq Est. Var.
    vaina  4  1176.100 294.025    35.775
Residuals 35   273.875   7.825     7.825
```

Para validar estos modelos realizaremos los cálculos y gráficos de los residuos de forma idéntica al caso de los factores fijos que hemos visto en el capítulo anterior.

11.6. Ejercicios

Ejercicio 11.1

Eligiendo 4 tardes al azar del verano, se midió la temperatura de un lago a diferentes profundidades con los siguientes resultados

	Fecha			
Profundidad (m)	1	2	3	4
0	23,8	24,0	34,6	24,8
1	22,6	22,4	22,9	23,2
2	22,2	22,1	22,1	22,2
3	21,2	21,8	21,0	21,2
4	18,4	19,3	19,0	18,8
5	13,5	14,4	14,2	13,8

Determinar si son factores de efectos fijos o de efectos aleatorios y si hay diferencias entre profundidades y entre fechas.

Ejercicio 11.2

Para valorar la variabilidad del contenido de zumo de una cierta variedad de limón, se tomaron 4 árboles al azar y se midió el contenido de zumo de 3 limones de cada árbol. Esta observación se hizo durante 5 días, eligiendo fechas al azar. Los resultados fueron (en cm^3):

	Árbol											
Día	1			2			3			4		
1	24	26	26	28	20	27	28	18	21	27	24	20
2	18	25	19	21	24	23	27	19	17	25	23	22
3	16	21	15	24	20	21	22	25	24	29	27	27
4	21	24	22	23	20	26	24	24	23	20	21	27
5	23	24	28	27	21	28	26	25	27	25	27	28

Estudiar si existe variabilidad entre árboles, entre días y entre las interacciones árboles $\times$ días.

Ejercicio 11.3

En una población, de entre las mujeres que habían concebido tres hijos varones, se seleccionaron 5 al azar y se anotó el peso que registró cada hijo al nacer:

1	3,250	3,125	3,400
2	2,800	3,100	2,900
3	3,400	3,500	3,350
4	4,100	4,200	4,150
5	2,900	2,750	2,800

Calcular la correlación intraclásica y estudiar si es significativa.

Ejercicio 11.4

Se han obtenido réplicas de una variable observable y combinado dos factores A, B. El número de réplicas ("factor" R) por casilla es de tres. La descomposición de la suma de cuadrados es la siguiente:

Fuente variación	g.l.	Suma cuadrados
A	3	420
B	1	143
AB	3	32
R	2	109
AR	6	197
BR	2	39
ABR	6	155

Utilizando el nivel de significación 0,01, se pide:

a) Suponiendo A, B factores de efectos fijos, estudiar si son significativos. Hallar tres estimaciones independientes de la varianza del diseño.

b) Suponiendo A, B factores de efectos aleatorios, estudiar si A y la interacción $A \times B$ son significativos.

Ejercicio 11.5

Consideremos de nuevo el enunciado del problema 10.4. Supongamos ahora que en el modelo 10.27 las interacciones ω_{ir} son nulas, A (año) es de efectos aleatorios y B (genotipo) es de efectos fijos. Estudiar si los efectos principales y las interacciones son significativas.

Ejercicio 11.6

Los resultados y_{ijh} de un cierto experimento, donde $i = 1, \ldots, p$; $j = 1, \ldots, q$; $h = 1, \ldots, b$ combinan dos factores X, Y, junto con un factor bloque B que no interacciona con X, Y. En este experimento las réplicas son bloques y el modelo es

$$y_{ijk} = \mu + X_i + Y_j + I_{ij} + B_h + \epsilon_{ijh}$$

La tabla de suma de cuadrados es:

Fuente variación	g.l.	Suma cuadrados
X	2	625
Y	3	1340
B	4	402
XY	6	227
XB	8	289
YB	12	310
XYB	24	528

Se pide:

a) Suponiendo los efectos fijos, estudiar la significación de los efectos principales y la interacción (nivel 0,05). Hallar dos estimadores insesgados de la varianza del modelo.

b) Suponiendo todos los efectos aleatorios, y sabiendo que los valores esperados de los cuadrados medios son:

$$E(\mathrm{CM}_X) = rq\sigma_X^2 + r\sigma_I^2 + \sigma^2 \qquad E(\mathrm{CM}_Y) = rp\sigma_Y^2 + r\sigma_I^2 + \sigma^2$$
$$E(\mathrm{CM}_I) = r\sigma_I^2 + \sigma^2 \qquad E(\mathrm{CM}_B) = pq\sigma_B^2 + \sigma^2 \qquad E(\mathrm{CM}_R) = \sigma^2$$

Apéndice A

Matrices

A.1. Inversa generalizada

Para una matriz $\mathbf{A}$ $(n \times p)$, $\mathbf{A}^-$ se llama una g-inversa o inversa generalizada de $\mathbf{A}$ si

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$$

Una inversa generalizada siempre existe aunque en general no es única.

Métodos de construcción

(1) Utilizando la descomposición en valores singulares de la matriz $\mathbf{A}$ $(n \times p)$, tenemos $\mathbf{A} = \mathbf{U}\mathbf{L}\mathbf{V}'$. Luego es sencillo comprobar que

$$\mathbf{A}^- = \mathbf{V}\mathbf{L}^{-1}\mathbf{U}'$$

define una g-inversa.

(2) Si $\mathrm{rg}(\mathbf{A}) = r$, una permutación de las filas y columnas de $\mathbf{A}$ $(n \times p)$ nos permite hallar una submatriz no singular $\mathbf{A}_r$ $(r \times r)$. Entonces resulta que

$$\mathbf{A}^- = \begin{pmatrix} \mathbf{A}_r^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

es una g-inversa.

(3) Si $\mathbf{A}$ $(p \times p)$ es no singular, entonces $\mathbf{A}^- = \mathbf{A}^{-1}$ y es única.

(4) Si $\mathbf{A}$ $(p \times p)$ es simétrica de $\mathrm{rg}(\mathbf{A}) = r$, podemos escribir $\mathbf{A} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}'$, donde $\mathbf{\Gamma}$ $(p \times r)$ es la matriz cuyas columnas son los vectores propios ortonormales correspondientes a los vectores propios no nulos $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$ de $\mathbf{A}$. Entonces se comprueba que

$$\mathbf{A}^- = \mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{\Gamma}'$$

Un caso especial de g-inversa es la llamada inversa de Moore-Penrose $\mathbf{A}^+$ de $\mathbf{A}$ $(n \times p)$ que verifica

$$\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A} \quad \mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+ \quad \mathbf{A}^+\mathbf{A} = (\mathbf{A}^+\mathbf{A})' \quad \mathbf{A}\mathbf{A}^+ = (\mathbf{A}\mathbf{A}^+)'$$

La inversa de Moore-Penrose es única.

A.2. Derivación matricial

Definimos la derivada de $f(\mathbf{X})$ con respecto a $\mathbf{X}$ $n \times p$ como la matriz

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \left(\frac{\partial f(\mathbf{X})}{\partial x_{ij}} \right).$$

El cálculo matricial de derivadas tiene, entre otras, las siguientes propiedades:

1. $\dfrac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}, \quad \dfrac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}'$
2. $\dfrac{\partial \mathbf{x}'\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}, \quad \dfrac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A}' + \mathbf{A})\mathbf{x}, \quad \dfrac{\partial \mathbf{x}'\mathbf{A}\mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}\mathbf{y}$

A.3. Matrices idempotentes

Una matriz $\mathbf{P}$ es idempotente si $\mathbf{P}^2 = \mathbf{P}$. Una matriz simétrica e idempotente se llama matriz proyección.

1. Si $\mathbf{P}$ es simétrica, entonces $\mathbf{P}$ es idempotente y $\text{rg}(\mathbf{P}) = r$ si y sólo si $\mathbf{P}$ tiene r valores propios iguales a 1 y el resto son cero.

   Demostración:

   Como $\mathbf{P}^2 = \mathbf{P}$, entonces $\mathbf{P}\mathbf{x} = \lambda\mathbf{x}$ con $\mathbf{x} \neq \mathbf{0}$ implica que

   $$\lambda\mathbf{x} = \mathbf{P}\mathbf{x} = \mathbf{P}^2\mathbf{x} = \mathbf{P}(\mathbf{P}\mathbf{x}) = \mathbf{P}(\lambda\mathbf{x}) = \lambda(\mathbf{P}\mathbf{x}) = \lambda(\lambda\mathbf{x}) = \lambda^2\mathbf{x}$$

   de manera que $\lambda^2 - \lambda = 0$ ó $\lambda(\lambda - 1) = 0$.

   Luego los valores propios de $\mathbf{P}$ son la unidad tantas veces como indica el rango y el resto son cero, ya que la suma de los valores propios es el rango.

   Recíprocamente, si los valores propios son 0 y 1, entonces podemos pensar sin pérdida de generalidad que los primeros r son unos.

   Así, debe existir una matriz ortogonal $\mathbf{T}$ tal que $\mathbf{P} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}'$ donde

   $$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

   Luego

   $$\mathbf{P}^2 = \mathbf{T}\mathbf{\Lambda}\mathbf{T}'\mathbf{T}\mathbf{\Lambda}\mathbf{T}' = \mathbf{T}\mathbf{\Lambda}^2\mathbf{T}' = \mathbf{T}\mathbf{\Lambda}\mathbf{T}' = \mathbf{P}$$

   y $\text{rg}(\mathbf{P}) = r$.

2. Si $\mathbf{P}$ es una matriz proyección, entonces $\text{tr}(\mathbf{P}) = \text{rg}(\mathbf{P})$.

   Demostración:

   Si $\text{rg}(\mathbf{P}) = r$, entonces por el apartado anterior, $\mathbf{P}$ tiene r valores propios 1 y el resto son cero. De aquí que $\text{tr}(\mathbf{P}) = r$.

3. Si $\mathbf{P}$ es idempotente, también $\mathbf{I}-\mathbf{P}$ lo es.

Demostración:

$$(\mathbf{I}-\mathbf{P})^2 = \mathbf{I}-2\mathbf{P}+\mathbf{P}^2 = \mathbf{I}-2\mathbf{P}+\mathbf{P} = \mathbf{I}-\mathbf{P}.$$

4. Las matrices proyección son semidefinidas positivas.

Demostración:

$$\mathbf{x}'\mathbf{P}\mathbf{x} = \mathbf{x}'\mathbf{P}^2\mathbf{x} = (\mathbf{P}\mathbf{x})'(\mathbf{P}\mathbf{x}) \geq 0.$$

A.4. Matrices mal condicionadas

Un sistema determinado de ecuaciones lineales $\mathbf{B}\mathbf{x} = \mathbf{c}$ se dice que está mal condicionado (*ill-conditioned*) si pequeños errores o variaciones en los elementos de $\mathbf{B}$ y $\mathbf{c}$ tienen un gran efecto en la solución exacta de $\mathbf{x}$. Por ejemplo, la solución exacta del sistema es $\mathbf{x} = \mathbf{B}^{-1}\mathbf{c}$, pero si $\mathbf{B}$ está cerca de la singularidad, es decir, pequeños cambios en sus elementos pueden causar la singularidad, entonces el cálculo de la inversa de $\mathbf{B}$ puede provocar una gran diferencia con la solución exacta.

En el caso de las ecuaciones normales la matriz $\mathbf{B} = \mathbf{X}'\mathbf{X}$ y el vector $\mathbf{c} = \mathbf{X}'\mathbf{Y}$ contienen errores de redondeo, fruto del cálculo a partir de las matrices $\mathbf{X}$ y $\mathbf{Y}$. Además, su almacenamiento en el ordenador también puede tener inconvenientes de precisión. Esto significa que si la matriz $\mathbf{X}$ está mal condicionada, es decir, pequeños cambios en los elementos de $\mathbf{X}$ pueden causar grandes cambios en $(\mathbf{X}'\mathbf{X})^{-1}$ y en $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, entonces cualquier error en la formación de $\mathbf{X}'\mathbf{X}$ puede tener un efecto muy serio en la precisión y la estabilidad de la solución, que en este caso es la estimación de los parámetros. El problema de la mala condición es especialmente preocupante en la regresión polinómica (ver sección 8.6).

Una medida de la mala condición de una matriz de datos $\mathbf{X}$ es el *número de condición* $\kappa[\mathbf{X}]$ que se define como la razón entre el mayor y el menor valor singular no nulo de $\mathbf{X}$. Los valores singulares de $\mathbf{X}$ son las raíces cuadradas positivas de los valores propios de la matriz $\mathbf{X}'\mathbf{X}$. Entre las propiedades más notorias de $\kappa[\mathbf{X}]$ tenemos que

$$\kappa[\mathbf{X}'\mathbf{X}] = (\kappa[\mathbf{X}])^2$$

Por la definición $\kappa > 1$, por tanto $\mathbf{X}'\mathbf{X}$ siempre está peor condicionada que $\mathbf{X}$. Luego, a no ser que $\kappa[\mathbf{X}]$ sea un valor moderado, es mejor no calcular $\mathbf{X}'\mathbf{X}$ en los métodos de computación de las soluciones (ver capítulo 11 de Seber[65]).

En la práctica, es muy común que una variable regresora esté altamente correlacionada con una combinación lineal de las otras variables regresoras, de forma que las columnas de $\mathbf{X}$ estarán muy próximas a ser linealmente dependientes. Así $\mathbf{X}'\mathbf{X}$ estará cerca de la singularidad (o será singular), el menor valor propio será pequeño y $\kappa[\mathbf{X}]$ será grande (ver sección 8.5).

Apéndice B

Proyecciones ortogonales

B.1. Descomposición ortogonal de vectores

1. Dado Ω, un subespacio vectorial de E_n (un espacio euclídeo n-dimensional), todo vector $\mathbf{y}$ puede expresarse de forma única como $\mathbf{y} = \mathbf{u} + \mathbf{v}$, donde $\mathbf{u} \in \Omega$ y $\mathbf{v} \in \Omega^{\perp}$.

Demostración:

Supongamos que hubiera dos descomposiciones $\mathbf{y} = \mathbf{u}_1 + \mathbf{v}_1 = \mathbf{u}_2 + \mathbf{v}_2$, entonces $(\mathbf{u}_1 - \mathbf{u}_2) + (\mathbf{v}_1 - \mathbf{v}_2) = \mathbf{0}$. Como $\mathbf{u}_1 - \mathbf{u}_2 \in \Omega$, $\mathbf{v}_1 - \mathbf{v}_2 \in \Omega^{\perp}$ y $\Omega \cap \Omega^{\perp} = \{\mathbf{0}\}$, resulta que $\mathbf{u}_1 - \mathbf{u}_2 = \mathbf{0}$ y $\mathbf{v}_1 - \mathbf{v}_2 = \mathbf{0}$, es decir, $\mathbf{u}_1 = \mathbf{u}_2$ y $\mathbf{v}_1 = \mathbf{v}_2$.

2. Si la descomposición adopta la forma $\mathbf{y} = \mathbf{P}_{\Omega}\mathbf{y} + (\mathbf{I} - \mathbf{P}_{\Omega})\mathbf{y}$, la matriz $\mathbf{P}_{\Omega}$ es única.

Demostración:

Si fueran dos las matrices $\mathbf{P}_i \quad i = 1, 2$, entonces, como $\mathbf{u}$ es único para cada $\mathbf{y}$, resulta que $(\mathbf{P}_1 - \mathbf{P}_2)\mathbf{y} = \mathbf{0}$ para todo $\mathbf{y}$. Luego $\mathbf{P}_1 - \mathbf{P}_2 = \mathbf{O}$.

3. La matriz $\mathbf{P}_{\Omega}$ puede expresarse en la forma $\mathbf{P}_{\Omega} = \mathbf{T}\mathbf{T}'$, donde las columnas de $\mathbf{T}$ forman una base ortonormal de Ω.

Demostración:

Sea $\mathbf{T} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r)$, donde $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r$ es una base ortonormal de Ω y r es su dimensión. Podemos extender esta base hasta obtener una base ortonormal de todo E_n, digamos $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r, \boldsymbol{\alpha}_{r+1}, \ldots, \boldsymbol{\alpha}_n$. Entonces

$$\mathbf{y} = \sum_{i=1}^{n} c_i \boldsymbol{\alpha}_i = \sum_{i=1}^{r} c_i \boldsymbol{\alpha}_i + \sum_{i=r+1}^{n} c_i \boldsymbol{\alpha}_i = \mathbf{u} + \mathbf{v}$$

donde $\mathbf{u} \in \Omega$ y $\mathbf{v} \in \Omega^{\perp}$. Pero $\boldsymbol{\alpha}_i' \boldsymbol{\alpha}_i = \delta_{ij}$ de forma que $\boldsymbol{\alpha}_i' \mathbf{y} = c_i$ y podemos escribir

$$\mathbf{u} = \sum_{i=1}^{r} c_i \boldsymbol{\alpha}_i = \sum_{i=1}^{r} (\boldsymbol{\alpha}_i' \mathbf{y}) \boldsymbol{\alpha}_i = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r)(\boldsymbol{\alpha}_1' \mathbf{y}, \ldots, \boldsymbol{\alpha}_r' \mathbf{y})' = \mathbf{T}\mathbf{T}'\mathbf{y}$$

y por el apartado anterior $\mathbf{P}_{\Omega} = \mathbf{T}\mathbf{T}'$.

4. $\mathbf{P}_{\Omega}$ es simétrica e idempotente.

Demostración:

Dado que $\mathbf{P}_{\Omega} = \mathbf{T}\mathbf{T}'$ es obviamente simétrica y

$$\mathbf{P}_{\Omega}^2 = \mathbf{T}\mathbf{T}'\mathbf{T}\mathbf{T}' = \mathbf{T}\mathbf{I}_r\mathbf{T}' = \mathbf{T}\mathbf{T}' = \mathbf{P}_{\Omega}$$

5. El subespacio generado por las columnas de $\mathbf{P}_\Omega$ es $\langle \mathbf{P}_\Omega \rangle = \Omega$.

Demostración:

Es evidente que $\langle \mathbf{P}_\Omega \rangle \subset \Omega$, ya que $\mathbf{P}_\Omega$ es la proyección sobre Ω. Recíprocamente si $\mathbf{x} \in \Omega$, entonces $\mathbf{x} = \mathbf{P}_\Omega \mathbf{x} \in \langle \mathbf{P}_\Omega \rangle$. Luego los dos subespacios son el mismo.

6. $\mathbf{I}_n - \mathbf{P}_\Omega$ representa la proyección ortogonal sobre $\Omega^\perp$.

Demostración:

A partir de la igualdad $\mathbf{y} = \mathbf{P}_\Omega \mathbf{y} + (\mathbf{I}_n - \mathbf{P}_\Omega)\mathbf{y}$ tenemos que $\mathbf{v} = (\mathbf{I}_n - \mathbf{P}_\Omega)\mathbf{y}$. Los resultados anteriores se obtienen intercambiando los papeles de Ω y $\Omega^\perp$.

7. Si $\mathbf{P}$ es una matriz cuadrada e idempotente, entonces $\mathbf{P}$ representa la proyección ortogonal sobre $\langle \mathbf{P} \rangle$.

Demostración:

Sea $\mathbf{y} = \mathbf{P}\mathbf{y} + (\mathbf{I}_n - \mathbf{P})\mathbf{y}$. Entonces $(\mathbf{P}\mathbf{y})'(\mathbf{I}_n - \mathbf{P})\mathbf{y} = \mathbf{y}'(\mathbf{P} - \mathbf{P}^2)\mathbf{y} = 0$, de manera que la descomposición da las componentes ortogonales de $\mathbf{y}$. El resultado se obtiene al aplicar la propiedad B.1.5.

8. Si $\Omega = \langle \mathbf{X} \rangle$, entonces

$$\boxed{\mathbf{P}_\Omega = \mathbf{X}(\mathbf{X}'\mathbf{X})^- \mathbf{X}'}$$

donde $(\mathbf{X}'\mathbf{X})^-$ es una inversa generalizada de $\mathbf{X}'\mathbf{X}$, es decir, si $\mathbf{B} = \mathbf{X}'\mathbf{X}$, entonces $\mathbf{B}\mathbf{B}^-\mathbf{B} = \mathbf{B}$.

Demostración:

Las ecuaciones normales $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$ se pueden escribir como $\mathbf{B}\boldsymbol{\beta} = \mathbf{c}$, si $\mathbf{c} = \mathbf{X}'\mathbf{Y}$. Entonces $\widehat{\boldsymbol{\beta}} = \mathbf{B}^-\mathbf{c}$ es una solución de dichas ecuaciones normales ya que

$$\mathbf{B}\widehat{\boldsymbol{\beta}} = \mathbf{B}(\mathbf{B}^-\mathbf{c}) = \mathbf{B}\mathbf{B}^-\mathbf{B}\boldsymbol{\beta} = \mathbf{B}\boldsymbol{\beta} = \mathbf{c}.$$

Por otra parte, si escribimos $\widehat{\boldsymbol{\theta}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, tenemos $\mathbf{Y} = \widehat{\boldsymbol{\theta}} + (\mathbf{Y} - \widehat{\boldsymbol{\theta}})$ donde

$$\begin{aligned} \widehat{\boldsymbol{\theta}}'(\mathbf{Y} - \widehat{\boldsymbol{\theta}}) &= \widehat{\boldsymbol{\beta}}'\mathbf{X}'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= \widehat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}) = 0 \end{aligned}$$

Luego $\mathbf{Y} = \widehat{\boldsymbol{\theta}} + (\mathbf{Y} - \widehat{\boldsymbol{\theta}})$ es una descomposición ortogonal de $\mathbf{Y}$ tal que $\widehat{\boldsymbol{\theta}} \in \langle \mathbf{X} \rangle$ y $(\mathbf{Y} - \widehat{\boldsymbol{\theta}}) \perp \langle \mathbf{X} \rangle$. Como $\widehat{\boldsymbol{\theta}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\mathbf{B}^-\mathbf{c} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}$ tenemos que $\mathbf{P}_\Omega = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$ por la unicidad demostrada en (2).

9. Cuando las columnas de la matriz $\mathbf{X}$ son linealmente independientes y el $\mathrm{rg}(\mathbf{X})$ es máximo, resulta que $\mathbf{P}_\Omega = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Demostración:

Cuando el $\mathrm{rg}(\mathbf{X})$ es máximo, la matriz cuadrada $\mathbf{X}'\mathbf{X}$ es inversible.

B.2. Proyecciones en subespacios

1. Dado $\omega \subset \Omega$, entonces $\mathbf{P}_\Omega \mathbf{P}_\omega = \mathbf{P}_\omega \mathbf{P}_\Omega = \mathbf{P}_\omega$.

Demostración:

Como $\omega \subset \Omega$ y $\omega = \langle \mathbf{P}_\omega \rangle$ (por el punto B.1.5), tenemos que la proyección sobre Ω de las columnas de $\mathbf{P}_\omega$ son las propias columnas, es decir, $\mathbf{P}_\Omega \mathbf{P}_\omega = \mathbf{P}_\omega$. El resultado completo se deduce porque $\mathbf{P}_\Omega$ y $\mathbf{P}_\omega$ son matrices simétricas.

2. $\mathbf{P}_\Omega - \mathbf{P}_\omega = \mathbf{P}_{\omega^\perp \cap \Omega}$.

Demostración:

Consideremos la descomposición $\mathbf{P}_\Omega \mathbf{y} = \mathbf{P}_\omega \mathbf{y} + (\mathbf{P}_\Omega - \mathbf{P}_\omega)\mathbf{y}$. Como $\mathbf{P}_\Omega \mathbf{y}$ y $\mathbf{P}_\omega \mathbf{y}$ pertenecen a Ω resulta que $(\mathbf{P}_\Omega - \mathbf{P}_\omega)\mathbf{y} \in \Omega$. Así la ecuación anterior presenta la descomposición ortogonal de Ω en ω y $\omega^\perp \cap \Omega$ ya que $\mathbf{P}_\omega(\mathbf{P}_\Omega - \mathbf{P}_\omega) = \mathbf{O}$ (por B.2.1).

3. Si $\mathbf{A}_*$ es una matriz tal que $\omega = \ker(\mathbf{A}_*) \cap \Omega$, entonces $\omega^\perp \cap \Omega = \langle \mathbf{P}_\Omega \mathbf{A}_*' \rangle$.

Demostración:

En primer lugar, observamos que

$$\begin{aligned} \omega^\perp \cap \Omega &= \{\Omega \cap \ker(\mathbf{A}_*)\}^\perp \cap \Omega \\ &= \{\Omega^\perp + \langle \mathbf{A}_*' \rangle\} \cap \Omega \end{aligned}$$

ya que $(\Omega_1 \cap \Omega_2)^\perp = \Omega_1^\perp + \Omega_2^\perp$ y $[\ker(\mathbf{A}_*)]^\perp = \langle \mathbf{A}_*' \rangle$.

Si $\mathbf{x} \in \{\Omega^\perp + \langle \mathbf{A}_*' \rangle\} \cap \Omega$, entonces

$$\mathbf{x} = \mathbf{P}_\Omega \mathbf{x} = \mathbf{P}_\Omega \{(\mathbf{I}_n - \mathbf{P}_\Omega)\boldsymbol{\alpha} + \mathbf{A}_*' \boldsymbol{\beta}\} = \mathbf{P}_\Omega \mathbf{A}_*' \boldsymbol{\beta} \in \langle \mathbf{P}_\Omega \mathbf{A}_*' \rangle.$$

Recíprocamente, si $\mathbf{x} \in \langle \mathbf{P}_\Omega \mathbf{A}_*' \rangle$, entonces $\mathbf{x} \in \langle \mathbf{P}_\Omega \rangle = \Omega$. También para cualquier $\mathbf{z} \in \omega$, resulta $\mathbf{x}'\mathbf{z} = \boldsymbol{\beta}' \mathbf{A}_* \mathbf{P}_\Omega \mathbf{z} = \boldsymbol{\beta}' \mathbf{A}_* \mathbf{z} = 0$, es decir, $\mathbf{x} \in \omega^\perp$. Luego $\mathbf{x} \in \omega^\perp \cap \Omega$.

4. Si $\mathbf{A}_*$ $(q \times n)$ tiene $\mathrm{rg}(\mathbf{A}_*) = q$, entonces $\mathrm{rg}(\mathbf{P}_\Omega \mathbf{A}_*') = q$ si y sólo si $\langle \mathbf{A}_*' \rangle \cap \Omega^\perp = \{\mathbf{0}\}$.

Apéndice C

Estadística multivariante

C.1. Esperanza, varianza y covarianza

1. Sean $\mathbf{X}$ e $\mathbf{Y}$ vectores aleatorios no necesariamente de la misma longitud.

Definimos la matriz

$$\mathrm{cov}(\mathbf{X},\mathbf{Y}) = (\mathrm{cov}(X_i, Y_j))$$

y si $\mathbf{X} = \mathbf{Y}$ escribimos $\mathrm{var}(\mathbf{X}) = \mathrm{cov}(\mathbf{X},\mathbf{X})$. Entonces se verifican las siguientes propiedades:

(a) Si $\mathbf{a}$ es un vector constante de la misma dimensión que $\mathbf{X}$, $\mathrm{var}(\mathbf{a}+\mathbf{X}) = \mathrm{var}(\mathbf{X})$.

(b) Si $\lambda \in \mathbb{R}$, entonces $\mathrm{var}(\lambda\mathbf{X}) = \lambda^2\mathrm{var}(\mathbf{X})$.

(c) Si $\mathbf{A}$ y $\mathbf{B}$ son matrices de constantes,

$$\mathrm{cov}(\mathbf{AX},\mathbf{BY}) = \mathbf{A}\mathrm{cov}(\mathbf{X},\mathbf{Y})\mathbf{B}'$$

(d) Para cualquier vector aleatorio $\mathbf{X},\mathbf{Y},\mathbf{U},\mathbf{V}$ y todo escalar $a,b,c,d \in \mathbb{R}$,

$$\begin{aligned}&\mathrm{cov}(a\mathbf{X}+b\mathbf{Y}, c\mathbf{U}+d\mathbf{V}) = \\ &\quad ac\,\mathrm{cov}(\mathbf{X},\mathbf{U}) + ad\,\mathrm{cov}(\mathbf{X},\mathbf{V}) + bc\,\mathrm{cov}(\mathbf{Y},\mathbf{U}) + bd\,\mathrm{cov}(\mathbf{Y},\mathbf{V})\end{aligned}$$

2. Sea $\mathbf{Y}$ un vector aleatorio con esperanza $E(\mathbf{Y}) = \boldsymbol{\mu}$ y matriz de varianzas y covarianzas $\mathrm{var}(\mathbf{Y}) = \mathbf{V}$, entonces

$$E(\mathbf{Y}'\mathbf{AY}) = \mathrm{tr}(\mathbf{AV}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$$

donde $\mathbf{A}$ es una matriz constante.

Demostración:

Es evidente que

$$(\mathbf{Y}-\boldsymbol{\mu})'\mathbf{A}(\mathbf{Y}-\boldsymbol{\mu}) = \mathbf{Y}'\mathbf{AY} - \boldsymbol{\mu}'\mathbf{AY} - \mathbf{Y}'\mathbf{A}\boldsymbol{\mu} + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$$

de modo que

$$\begin{aligned}E((\mathbf{Y}-\boldsymbol{\mu})'\mathbf{A}(\mathbf{Y}-\boldsymbol{\mu})) &= E(\mathbf{Y}'\mathbf{AY}) - \boldsymbol{\mu}'\mathbf{A}E(\mathbf{Y}) - E(\mathbf{Y}')\mathbf{A}\boldsymbol{\mu} + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} \\ &= E(\mathbf{Y}'\mathbf{AY}) - \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}\end{aligned}$$

Por otra parte, sabemos que, para dos matrices $\mathbf{C}$ y $\mathbf{D}$, la traza del producto verifica

$$\mathrm{tr}(\mathbf{CD}) = \mathrm{tr}(\mathbf{DC}) = \sum_{i,j} c_{ij} d_{ji}$$

y por eso

$$\begin{aligned} \mathrm{tr}(\mathbf{AV}) &= \sum_{i,j} a_{ij} \mathrm{cov}(Y_j, Y_i) = \sum_{i,j} a_{ij} E((Y_j - \mu_j)(Y_i - \mu_i)) \\ &= E\left(\sum_{i,j} (Y_i - \mu_i) a_{ij} (Y_j - \mu_j) \right) = E((\mathbf{Y} - \boldsymbol{\mu})' \mathbf{A} (\mathbf{Y} - \boldsymbol{\mu})) \end{aligned}$$

con lo que obtenemos el resultado enunciado.

C.2. Normal multivariante

1. Cuando $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, se verifica:

(a) $(\mathbf{Y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \boldsymbol{\mu}) \sim \chi^2_n$

(b) Para cualquier matriz $\mathbf{C}$ constante, $\mathbf{CY} \sim N_n(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')$.

(c) Si consideramos una partición del vector $\mathbf{Y}$ en dos vectores $\mathbf{Y}_1$ y $\mathbf{Y}_2$, éstos son independientes ssi $\mathrm{cov}(\mathbf{Y}_1, \mathbf{Y}_2) = \mathbf{O}$.

2. Sea $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$. Sean $\mathbf{U} = \mathbf{AY}$, $\mathbf{V} = \mathbf{BY}$ dos vectores aleatorios combinación de $\mathbf{Y}$ y sea $\mathbf{A}_*$ la matriz formada por las filas de $\mathbf{A}$ linealmente independientes. Si $\mathrm{cov}(\mathbf{U}, \mathbf{V}) = \mathbf{O}$, entonces

(a) $\mathbf{A}_* \mathbf{Y}$ es independiente de $\mathbf{V}'\mathbf{V}$.

(b) $\mathbf{U}'\mathbf{U}$ y $\mathbf{V}'\mathbf{V}$ son independientes.

3. Supongamos que $Q_1 \sim \chi^2_r$ y $Q_2 \sim \chi^2_s$, con $r > s$. Si $Q = Q_1 - Q_2$ y Q_2 son independientes, entonces $Q \sim \chi^2_{r-s}$.

Bibliografía

[1] J. Alegre y J. Arcarons, *Aplicaciones de Econometría*. Textos Docents, Universitat de Barcelona, 1991.

[2] D.A. Allen and F.B. Cady, *Analyzing Experimental Data by Regression*. Wadsworth, 1982.

[3] V.L. Anderson and R.A. McLean, *Design of Experiments*. Marcel Dekker, 1974.

[4] D.F. Andrews et al., *Robust Estimates of Location: Survey and Advances*. Princeton University Press, 1972.

[5] S.F. Arnold, *The Theory of Linear Models and Multivariate Observations*. Wiley, 1981.

[6] D.A. Belsley, E. Kuh and R.E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, 1980.

[7] J. Bibby and H. Toutenberg, *Prediction and Improved Estimation in Linear Models*. Wiley, 1977.

[8] D. Birkes and Y. Dodge, *Alternative Methods of Regression*. Wiley, 1993.

[9] G.E.P. Box and N. Draper, *Empirical Model Building and Response Surfaces*. Wiley, 1987.

[10] G.E.P. Box, W. Hunter and J.S. Hunter, *Estadística para Investigadores*. Reverté, 1988.

[11] R.J. Brook and G.C. Arnold, *Applied Regression Analysis and Experimental Design*. Marcel Dekker, 1985.

[12] R. Christensen, *Plane Answers to Complex Questions*. Springer-Verlag, 1987.

[13] W.G. Cochran and G.M. Cox, *Experimental Designs*. Wiley, 2nd Edition, 1992.

[14] R.D. Cook and S. Weisberg, *Residuals and Influence in Regression*. Chapman and Hall, 1982.

[15] R.D. Cook and S. Weisberg, *Applied Regression Including Computing and Graphics*. Wiley, 1999.

[16] J.A. Cornell, *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data*. Wiley, 3rd Edition, 2002.

[17] D.R. Cox, *Planning of Experiments*. Wiley, 1958.

[18] C.M. Cuadras, "Sobre la comparació estadística de corbes experimentals". Qüestiió, **3(1)**, 1-10, 1979.

[19] C.M. Cuadras, "Diseños no balanceados y con observaciones faltantes en MANOVA", *Actas XIII Reun.Anual S.Esp.Est.I.O.Inf.*, Valladolid, 1982.

[20] C.M. Cuadras, *Problemas de Probabilidades y Estadística. VOL. 2 Inferencia estadística*. EUB, Barcelona 2000.

[21] S. Chatterjee and B. Price, *Regression Analysis by Example*. Wiley, 3rd Edition, 1999.

[22] C. Daniel, *Applications of Statistics to industrial experimentation*. Wiley, 1976.

[23] C. Daniel and F.S. Wood, *Fitting Equations to Data*. Wiley, 1980.

[24] P.J. Dhrymes, *Econometría*. Editorial AC, Madrid, 1984.

[25] Y. Dodge, *Analysis of Experiments with missing data*. Wiley, 1985.

[26] H.E. Doran, *Applied Regression Analysis in Econometrics*. Marcel Dekker, 1989.

[27] N.R. Draper and H. Smith, *Applied Regression Analysis*. Wiley, 3rd Edition, 1998.

[28] R.A. Fisher, *The Design of Experiments*. Oliver Boyd, Edimburgo, 1953.

[29] J. Fox, *Linear Statistical Models & Related Methods*. Wiley, 1984.

[30] A.R. Gallant, *Nonlinear Statistical Models*. Wiley, 1987.

[31] A.S. Goldberger, *A Course in Econometrics*. Harvard University Press, 1991.

[32] F.A. Graybill, *Theory and Application of the Linear Model*. Wadsworth, 1976.

[33] R.F. Gunst and R.L. Mason, *Regression Analysis and its Aplication*. Marcel Dekker, 1980.

[34] I. Guttman, *Linear Models*. Wiley, 1982.

[35] W. Härdle, *Applied Nonparametric Regression*. Cambridge University Press, 1990.

[36] H.O. Hartley, "Analysis of Variance". *Mathematical Methods for Digital Computers*. A. Ralston and H. Wilf eds., Wiley, cap. 20, 1962.

[37] C.R. Hicks, *Fundamental Conceps in the Design of Experiments*. Holt, Renehart and Winston, 1982.

[38] K. Hinkelmann and O. Kempthorne, *Design and Analysis of Experiments , Volume 1, Introduction to Experimental Design*. Wiley, 1994.

[39] D.C. Hoaglin, F. Mosteller, and J.W. Tukey, *Understanding Robust and Exploratory Data Analysis*. Wiley, 1983.

[40] R.R. Hocking, *Methods and Applications of Linear Models: Regression and the Analysis of Variance*. Wiley, 2nd Edition, 2003.

[41] P.W.M. John, *Statistical Design and Analysis of Experiments*. Mc Millan, 1971.

[42] J.A. John and M.H. Quenouille, *Experiments: Design and Analysis*. Charles Griffin, 1977.

[43] O. Kempthorne, *The Design and Analysis of Experiments*. Wiley, 1952.

[44] M. Kendall, A. Stuart and J.K. Ord, *The Avanced Teory of Statistics (vol. 3, Design and Analysis, and Time Series)*. Charles Griffin, 1983.

[45] A. Kshirsagar, *A Course on Linear Models*. Marcel Dekker, 1983.

[46] T.J. Lorenzen and V. L. Anderson, *Design of Experiments*. Marcel Dekker, 1993.

[47] R.L. Mason, R.F. Gunst and J.L. Hess, *Statistical Design and Analysis of Experiments: With Applications to Engineering and Science*. Wiley, 2nd Edition, 2003.

[48] P. McCullagh and J.A. Nelder, *Generalized Linear Models*. Chapman and Hall, 1989.

[49] D.C. Montgomery, *Design and Analysis of Experiments*. Wiley, 1984.

[50] D.C. Montgomery, E.A. Peck and G.G. Vining *Introduction to Linear Regression Analysis*. Wiley, 3rd Edition, 2001.

[51] F. Mosteller and J.W. Tukey, *Data Analysis and Regression*. Addison-Wesley, 1977.

[52] R.H. Myers, *Clasical and Modern Regression with Application*. Duxbury Press, 1986.

[53] J. Neter, W. Wasserman and M.H. Kutner, *Applied Linear Statistical Models*. Richard D. Irwin, 1990.

[54] D. Peña, *Estadística: Modelos y métodos. 2. Modelos Lineales y Series Temporales*. Alianza, 1993.

[55] B.L. Raktoe et al., *Factorial Designs*. Wiley, 1981.

[56] C.R. Rao, *Linear Statistical Inference and its Applications*. Wiley, 1973

[57] C.R. Rao and H. Toutenburg, *Linear Models*. Springer Series in Statistics, 1995.

[58] D.A. Ratkowsky, *Non Linear Regression Modeling*. Marcel Dekker, 1983.

[59] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*. Wiley, 2003.

[60] L. Ruiz-Maya, *Métodos Estadísticos de Investigación*. INE, Madrid, 1972.

[61] T.P. Ryan, *Modern Regression Methods*. Wiley, 1996.

[62] S.R. Searle, *Linear Models*. Wiley, 1971.

[63] H. Scheffé, *The Analysis of Variance*. Wiley, 1959.

[64] G.A.F. Seber and C.J. Wild, *Nonlinear Regression*, Wiley, 2003.

[65] G.A.F. Seber and A.J. Lee, *Linear Regression Analysis*. Wiley, 2nd. Edition, 2003.

[66] A. Sen and M. Srivastava, *Regression Analysis*. Springer-Verlag, 1990.

[67] S.D. Silvey, "Multicollinearity and imprecise estimation", *J. R. Stat. Soc. B*, **31**, 539-552, 1969.

[68] S.D. Silvey, *Optimal Design: An Introduction to the Teory for Parameter Estimation*. Chapman and Hall, 1980.

[69] M.D. Ugarte y A.F. Militino. *Estadística Aplicada con S-Plus*. Universidad Pública de Navarra, 2001.

[70] H.D. Vinod and A. Ullah, *Recent Advances in Regression Methods*. Marcel Dekker, 1981.

[71] S. Weisber, *Applied Linear Regression*. Wiley, 2nd Edition, 1985.

[72] B.J. Winer, *Statistical Principes in Experimental Design*. McGraw-Hill, 1970.

[73] T.H. Wonnacott and R.J. Wonnacott, *Regression: a second course in statistics*. Wiley, 1981

Índice alfabético