# Regresión y Análisis de Varianza

V. Núñez Antón
y
F. Tusell[1]

18 de septiembre de 2007

# Índice general

# Índice de figuras

# Índice de cuadros

# Introducción

Lo que sigue contiene una introducción muy concisa al análisis de regresión, concebida como apoyo de las clases. Evita por eso largas explicaciones y se concentra en cuestiones conceptuales y formales.

Hay varios niveles de lectura: en un primer nivel, las Observaciones que jalonan el texto pueden en su mayoría omitirse, sin pérdida de continuidad. Ello proporciona una lectura bastante lineal.

Si se desea una lectura más detallada, con digresiones que, no siendo imprescindibles, pueden mejorar la comprensión del conjunto, conviene leer tanto las observaciones como los complementos y ejercicios: las secciones de CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER al fin de cada capitulo son parte integrante del texto a este segundo nivel y completan muchos detalles. Creemos que se trata de unas notas a las que será factible volver bastantes veces, aclarándose progresivamente cosas que distaron de ser obvias en una primera lectura.

A lo largo del texto, tanto en demostraciones como en ejercicios o complementos se ha hecho uso abundante del símbolo de "giro peligroso" representado en el margen, popularizado por la obra clásica Knuth (1986). Se trata de fragmentos que corresponderían a un tercer nivel, con detalles de interés, extensiones de alguna idea, referencias a la literatura o ejercicios y demostraciones de mayor dificultad.

Hay un mundo de diferencia entre saber *cómo se hacen* las cosas y *saber hacerlas*. Querríamos que los alumnos supieran *hacerlas*. La experiencia sugiere que lo que resulta de más ayuda al lector es ver ejemplos de aplicación detallados, que pueda reproducir o modificar para resolver sus propios problemas. Intercalados entre las exposiciones teóricas hay programas en R, que el lector puede ejecutar o tomar como modelo. Todos se han ejecutado con R version 1.9.1. No se ha tratado de exhibir el código más terso ni la forma más rápida o elegante de hacer las cosas, sino la que ilustra mejor la teoría.

Bilbao, 18 de septiembre de 2007

# Parte I

# Regresión Lineal

# Capítulo 1

# El modelo de regresión lineal.

## 1.1. Planteamiento del problema.

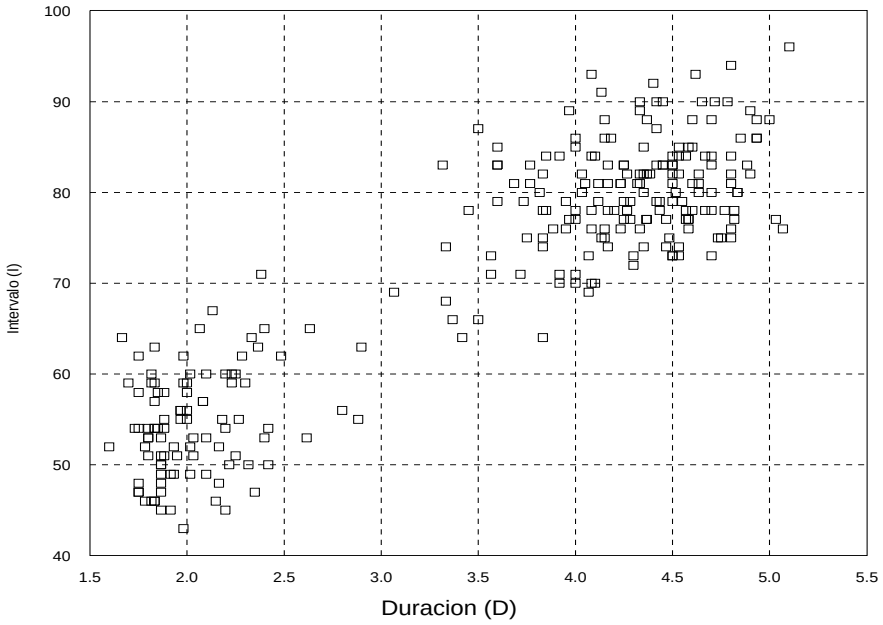
Son frecuentes en la práctica situaciones en las que se cuenta con observaciones de diversas variables, y es razonable pensar en una relación entre ellas. El poder determinar si existe esta relación —y, en su caso, una forma funcional para la misma— es de sumo interés. Por una parte, ello permitiría, conocidos los valores de algunas variables, efectuar predicciones sobre los valores previsibles de otra. Podríamos también responder con criterio estadístico a cuestiones acerca de la relación de una variable sobre otra.

**Ejemplo 1.1** La Figura 1.1 (pág. 4), muestra una gráfica recogiendo datos correspondientes a 272 erupciones del *geyser* Old Faithfull, en el Parque Nacional de Yellowstone (los datos proceden de Cook and Weisberg (1982)). En abscisas se representa la duración de las erupciones. En ordenadas, el intervalo de tiempo transcurrido hasta la siguiente erupción.

A la vista del gráfico, parece evidente que existe una relación entre ambas variables —erupciones de duración $D$ corta son seguidas de otras tras un intervalo de tiempo $I$ más reducido que erupciones largas—. Podría interesarnos contrastar con criterio estadístico si tal relación existe (en el caso presente, la relación es tan nítida que el plantearse el contraste de hipótesis correspondiente no tendría demasiado sentido). Más interesante, en el caso presente, sería llegar a una expresión del tipo $I = f(D)$ relacionando el intervalo con la duración (ello nos permitiría anticipar en qué momento se presentará la siguiente erupción, conocida la duración $D$ que se ha observado en la anterior).

A la vista de los datos, es claro que la relación $I = f(D)$ no puede ser exacta —es difícil pensar en una función que pase precisamente por cada uno de los 272 puntos en la Figura 1.1—. Habremos de considerar más bien funciones del tipo $I = f(D) + \epsilon$, en que el valor de $I$ es una cierta función (desconocida) de $D$ *más* una cantidad aleatoria inobservable $\epsilon$. Decimos que $f(D)$ es una *función de regresión* de $I$ sobre $D$, y nuestro objetivo es especificar su forma. Habitualmente realizamos para ello supuestos simplificadores, como el de que $f(D)$ es una función lineal.

Figura 1.1: Old Faithful Geyser: datos de 272 erupciones.




Es de interés señalar que el ajuste de un modelo de regresión no se limita a analizar la relación entre dos variables; en general, buscaremos relaciones del tipo

$$Y = f(X_0, X_1, \ldots, X_{p-1}) + \epsilon,$$

relacionando de manera aproximada los valores de $Y$ con los que toman otras variables, $X_0, \ldots, X_{p-1}$. Por simplicidad, limitaremos por el momento nuestra atención a funciones $f(X_0, \ldots, X_{p-1})$ lineales; el modelo resultante es el modelo de regresión lineal, que se examina en la Sección a continuación.

Señalemos, finalmente, que el hecho de aislar una variable $Y$ al lado izquierdo y escribirla como función de otras más una perturbación aleatoria $\epsilon$ no prejuzga ninguna relación de causalidad en ningún sentido; sólo postulamos la existencia de una relación cuya forma y alcance queremos investigar. En el Ejemplo 1.1, el ajuste de un modelo del tipo $I = f(D) + \epsilon$ no implica que consideremos que la duración $D$ *causa* el subsiguiente intervalo $I$ hasta la próxima erupción.

## 1.2. Notación

Consideramos una variable aleatoria $Y$ (*regresando*, *respuesta*, o *variable endógena*) de la que suponemos que se genera así:

$$Y = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \epsilon, \tag{1.1}$$

siendo:

1. $\beta_0, \ldots, \beta_{p-1}$, parámetros fijos desconocidos.

2. $X_0, \ldots, X_{p-1}$, variables explicativas no estocásticas, *regresores*, cuyos valores son fijados por el experimentador. Frecuentemente $X_0$ toma el valor constante "uno".

3. $\epsilon$ una variable aleatoria inobservable.

La ecuación (1.1) indica que la variable aleatoria $Y$ se genera como combinación lineal de las variables explicativas, salvo en una perturbación aleatoria $\epsilon$. En el Ejemplo 1.1, $Y$ sería la variable $I$, y el único regresor sería la variable $D$. Si decidimos ajustar un modelo con término constante $\beta_0$, tendríamos como regresores $D$ y $X_0 =$"uno". La función que aparece en (1.1) sería entonces $f(D) = \beta_0 + \beta_1 D$.

El problema que abordamos es el de estimar los parámetros desconocidos $\beta_0, \ldots, \beta_{p-1}$. Para ello contamos con una muestra de $N$ observaciones de la variable aleatoria $Y$, y de los correspondientes valores de las variables explicativas $X$. Como se ha dicho, $\epsilon$ es inobservable. La muestra nos permitirá escribir $N$ igualdades similares a (1.1):

$$\begin{aligned} y_1 &= \beta_0 X_{1,0} + \beta_1 X_{1,1} + \cdots + \beta_{p-1} X_{1,p-1} + \epsilon_1 \\ y_2 &= \beta_0 X_{2,0} + \beta_1 X_{2,1} + \cdots + \beta_{p-1} X_{2,p-1} + \epsilon_2 \\ &\vdots \\ y_N &= \beta_0 X_{N,0} + \beta_1 X_{N,1} + \cdots + \beta_{p-1} X_{N,p-1} + \epsilon_N. \end{aligned}$$

En forma matricial, escribiremos dichas $N$ igualdades así:

$$\vec{y} = X\vec{\beta} + \vec{\epsilon}, \tag{1.2}$$

siendo:

- $\vec{y}$ el vector $N \times 1$ de observaciones de la variable aleatoria Y,
- $X$ la matriz $N \times p$ de valores de las variables explicativas. Su elemento $X_{ij}$ denota el valor que la $j$–ésima variable explicativa toma en la $i$–ésima observación,
- $\vec{\beta}$ el vector de parámetros $(\beta_0, \ldots, \beta_{p-1})'$,
- $\vec{\epsilon}$ el vector $N \times 1$ de valores de la perturbacion aleatoria $\epsilon$.

Denotaremos mediante $\hat{\beta}$ al vector de estimadores de los parámetros, y por $\hat{\epsilon}$ al vector $N \times 1$ de residuos, definido por $\hat{\epsilon} = \vec{y} - X\hat{\beta}$; es decir, los residuos recogen la diferencia entre los valores muestrales observados y ajustados de la variable aleatoria $Y$.

Utilizamos minúsculas para designar valores muestrales y mayúsculas para las correspondientes variables aleatorias (así por ejemplo, $\vec{y}$ denota el vector de valores observados de la variable aleatoria $Y$ en una determinada experimentación). El contexto aclarará, por otra parte, cuando $\hat{\beta}$ y $\hat{\epsilon}$ son variables aleatorias o valores muestrales.

Adoptaremos como criterio de estimación el mínimo cuadrático. Por consiguiente, diremos que $\hat{\beta}$ es óptimo si $\| \vec{y} - X\hat{\beta} \|^2$ es mínimo, denotando $\| \cdot \|$ la norma euclídea ordinaria, $\| \vec{y} \| = \sqrt{\sum_i y_i^2}$ (ver Definición A.2, pág. 179).

**Observación 1.1** El suponer que los valores de los regresores pueden ser fijados por el analista (Supuesto 2) nos coloca en una situación de *diseño experimental*. De ahí que a la matriz $X$ se la denomine *matriz de diseño*.

Muchas veces (notablemente en Ciencias Sociales) no es posible fijar los valores de $X$, sino tan solo recolectar una muestra. Decimos entonces que estamos ante una *situación observacional* (en oposición a un diseño experimental). Ello no afecta a la teoría que sigue; la inferencia sobre los parámetros $\vec{\beta}$, etc. es entonces condicional a los valores observados de $X$.

## 1.3. Supuestos.

Además de suponer que $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$ y que la matriz $X$ es no aleatoria, requeriremos lo siguiente:

1. $E\vec{\epsilon} = \vec{0}$.
2. $E\vec{\epsilon}\ \vec{\epsilon}\ ' = \sigma^2 I$.
3. $\text{rango}(X) = p < N$.

Nos referiremos a 1)–3) en lo sucesivo como los *supuestos habituales*.

El supuesto 1) no implica pérdida de generalidad ni supone ninguna restricción, al menos en el caso en que $X$ tiene entre sus columnas una cuyos valores sean constantes (y ésto suele suceder; típicamente, la primera columna está formada por "unos"). En efecto, es claro que si:

$$\vec{Y} \quad = \quad \beta_0\vec{1} + \beta_1\vec{X}_{\ 1} + \cdots + \beta_{p-1}\vec{X}_{\ p-1} + \vec{\epsilon} \tag{1.3}$$

y el vector de perturbaciones verifica $E\vec{\epsilon} = \vec{\mu}$, entonces (1.3) puede reescribirse equivalentemente como:

$$\vec{Y} \quad = \quad (\beta_0\vec{1} + \vec{\mu}) + \beta_1\vec{X}_1 + \cdots + \beta_{p-1}\vec{X}_{p-1} + (\vec{\epsilon} - \vec{\mu}), \tag{1.4}$$

y (1.4) incorpora un vector de perturbaciones $(\vec{\epsilon} - \vec{\mu})$ verificando el primero de nuestros supuestos.

El supuesto 2), bastante más restrictivo, requiere que las perturbaciones sean incorreladas (covarianzas cero) y homoscedásticas (de idéntica varianza).

El supuesto 3) simplemente fuerza la independencia lineal entre las $(p)$ columnas de $X$. El requerimiento $N > p$ excluye de nuestra consideración el caso $N = p$, pues entonces $\vec{y} = X\hat{\beta}$ es un sistema de ecuaciones lineales determinado, y tiene siempre solución para algún vector $\hat{\beta}$ que hace los residuos nulos. Las estimaciones del vector $\vec{\beta}$ se obtendrían entonces resolviendo dicho sistema. Veremos en lo que sigue que este caso particular carece de interés (se dice que no tiene "grados de libertad").

Algunos de los supuestos anteriores serán relajados, y las consecuencias que de ello se derivan estudiadas.

**Observación 1.2** Nada impide que los regresores sean transformaciones adecuadas de las variables originales. Por ejemplo, si pensamos que la variable aleatoria $Y$ depende del cuadrado de $X_k$ y de otras variables, podríamos especificar un modelo de regresión así:

$$Y \quad = \quad \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k^2 + \cdots + \beta_{p-1}X_{p-1} + \epsilon.$$

Análogamente, si pensáramos que la variable aleatoria $W$ se genera del siguiente modo:

$$W = kZ_1{}^{\beta_1} Z_2{}^{\beta_2} \nu,$$

siendo $\nu$ una perturbación aleatoria no negativa (por ejemplo, con distribución logarítmico normal), nada impediría que tomáramos logaritmos para obtener:

$$Y = \log(W) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

en que $X_i = \log(Z_i)$, $\beta_0 = \log(k)$ y $\epsilon = \log(\nu)$.

Lo que realmente se requiere es que la expresión de la variable endógena o regresando $Y$ sea lineal *en los parámetros*.

## 1.4. La estimación mínimo cuadrática como problema de aproximación vectorial.

La ecuación matricial $\vec{y} = X\hat{\beta} + \hat{\epsilon}$ puede reescribirse asi:

$$\vec{y} = \hat{\beta}_0 \vec{X}_0 + \cdots + \hat{\beta}_{p-1} \vec{X}_{p-1} + \hat{\epsilon}, \tag{1.5}$$

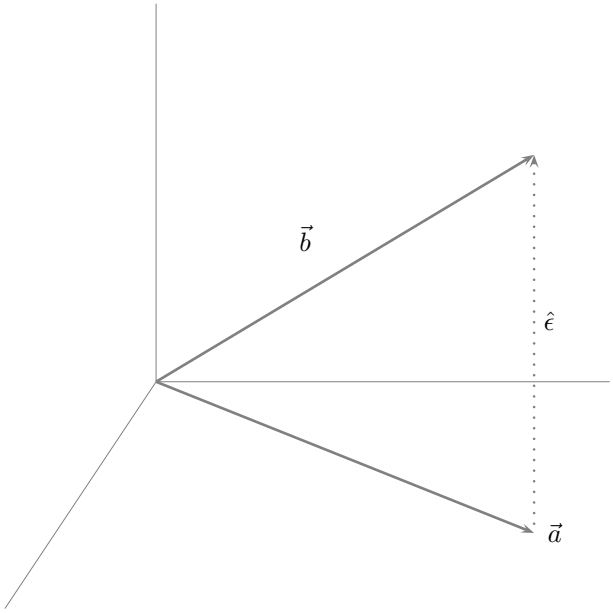
donde $\vec{X}_0, \ldots, \vec{X}_{p-1}$ denotan los vectores columna de la matriz $X$ ($\vec{X}_0$ será en general una columna de "unos", como se ha indicado). Hay diferentes posibilidades en cuanto a criterio de estimación de los $\beta$. Uno de ellos, el mínimo cuadrático ordinario (MCO), consiste en minimizar $\| \hat{\epsilon} \|^2$ , problema que, según muestra la ecuación (1.5), puede reformularse así: ¿Cuales son los coeficientes $\hat{\beta}_0, \ldots, \hat{\beta}_{p-1}$ que hacen que la combinación lineal $\hat{\beta}_0 \vec{X}_0 + \cdots + \hat{\beta}_{p-1} \vec{X}_{p-1}$ aproxime óptimamente (en sentido mínimo cuadrático) al vector $\vec{y}$ ? Veremos inmediatamente que esta combinación lineal es precisamente lo que llamaremos *proyección* de $\vec{y}$ sobre el subespacio generado por las columnas $\vec{X}_0 \ldots, \vec{X}_{p-1}$.

## 1.5. Proyecciones.

Aunque en lo que sigue se hace un tratamiento generalizable, implícitamente consideramos productos internos (véase Definición A.1) real-valorados, lo que simplifica algunas fórmulas. Hacemos también un uso bastante tosco del lenguaje y notación, identificando vectores con matrices columna, operadores lineales y matrices asociadas a ellos, etc. Lo inadecuado del formalismo puede ser fácilmente suplido por el lector, y evita notación que podría hacerse agobiante.

**Definición 1.1** *Sea $\{\vec{v}_n\}$ una sucesión de vectores en $H$, espacio vectorial sobre el cuerpo de los números reales $R$ con las operaciones "suma" de vectores y "producto" por números reales, definidas ambas del modo usual. Supongamos definido sobre $H$ un producto interno $< \cdot, \cdot >$ y correspondiente norma $\| \vec{v} \|^2 = < \vec{v}, \vec{v} >$. Decimos que $\{\vec{v}_n\}$ es una* sucesión de Cauchy *si para cualquier $\delta > 0$ hay un $N(\delta)$ tal que $\forall m, n \geq N(\delta), \quad \| \vec{v}_n - \vec{v}_m \| < \delta$; es decir, si prefijado un $\delta$ arbitrariamente pequeño, existe siempre un $N(\delta)$ tal que cualesquiera vectores $\vec{v}_m, \vec{v}_n$ que aparezcan en la sucesión en lugar posterior al $N(\delta)$ distan entre sí menos de $\delta$.*

Figura 1.2: El vector $\vec{a}$ es la proyección de $\vec{b}$ sobre $M$.




**Definición 1.2** *Sea $H$ un espacio vectorial como en la Definición 1.1. Decimos que tiene estructura de espacio de Hilbert si es completo, es decir, si contiene los límites de todas las sucesiones de Cauchy de vectores en $H$. Cualquier subespacio vectorial de un espacio de Hilbert, es a su vez espacio de Hilbert.*

**Definición 1.3** *Sea $H$ un espacio vectorial. Sea $M \subseteq H$ un subespacio del mismo, e $\vec{y} \in H$ un vector cualquiera. Decimos que $\vec{u}$ es la proyección de $\vec{y}$ sobre $M$ (y lo denotamos por $\vec{u} = P_M \vec{y}$) si:*

1. $\vec{u} \in M$,
2. $\vec{u} = \vec{y}$ si $\vec{y} \in M$,
3. $(\vec{y} - \vec{u}) \perp M$ si $\vec{y} \notin M$.

La Fig. 1.2 muestra en tres dimensiones la noción de proyección, y hace intuitivamente evidente el Teorema 1.1.

**Teorema 1.1** *Sea $H$ un espacio de Hilbert, y $M$ un subespacio del mismo. Para cualquier vector $\vec{y} \in H$ existe siempre un único vector $\vec{v} = P_M \vec{y}$ , proyección de $\vec{y}$ sobre $M$. Se verifica que:*

$$\| \vec{y} - \vec{v} \|^2 = \min_{\vec{z} \in M} \| \vec{y} - \vec{z} \|^2. \tag{1.6}$$

**Demostración.** Veamos[1] primero la existencia. Sea $d = \text{mín}_{\vec{z}\in M} \parallel \vec{y} - \vec{z} \parallel^2$. Entonces, necesariamente existirá en $M$ algún vector $\vec{v}_1$ tal que: $\parallel \vec{y} - \vec{v}_1 \parallel^2 \leq d + 1$; de no haberlo, $\text{mín} \parallel \vec{y} - \vec{z} \parallel^2$ tendría que ser mayor que $d + 1$, contra la hipótesis. Análogamente, para cualquier número natural $n$ existirá $\vec{v}_n$ verificando: $\parallel \vec{y} - \vec{v}_n \parallel^2 \leq d + 1/n$. Mostraremos que la sucesión $\{\vec{v}_n\}$ es de Cauchy. Mostraremos también que su límite –único– verifica las condiciones definitorias de proyección de $\vec{y}$ sobre $M$. Probaremos, en fin, que ningún otro vector en $M$ distinto del límite anterior verifica las mismas condiciones, así como la propiedad de mínima distancia en el enunciado.

Sea:

$$D = \parallel (\vec{y} - \vec{v}_n) - (\vec{y} - \vec{v}_m) \parallel^2 + \parallel (\vec{y} - \vec{v}_n) + (\vec{y} - \vec{v}_m) \parallel^2 \tag{1.7}$$

Podemos escribir:

$$\begin{aligned} D &= \parallel (\vec{y} - \vec{v}_n) \parallel^2 + \parallel (\vec{y} - \vec{v}_m) \parallel^2 - 2 < (\vec{y} - \vec{v}_m), (\vec{y} - \vec{v}_n) > \\ &\quad + \parallel (\vec{y} - \vec{v}_n) \parallel^2 + \parallel (\vec{y} - \vec{v}_m) \parallel^2 + 2 < (\vec{y} - \vec{v}_m), (\vec{y} - \vec{v}_n) > \\ &= 2\parallel (\vec{y} - \vec{v}_n) \parallel^2 + 2\parallel (\vec{y} - \vec{v}_m) \parallel^2. \end{aligned} \tag{1.8}$$

Por otra parte, tenemos:

$$\begin{aligned} D &= \parallel (\vec{v}_m - \vec{v}_n) \parallel^2 + \parallel 2\vec{y} - 2\,(\tfrac{1}{2})\,(\vec{v}_n + \vec{v}_m) \parallel^2 \\ &= \parallel (\vec{v}_m - \vec{v}_n) \parallel^2 + 4\parallel \vec{y} - (\tfrac{1}{2})\,(\vec{v}_n + \vec{v}_m) \parallel^2. \end{aligned} \tag{1.9}$$

Igualando (1.8) y (1.9) obtenemos:

$$\begin{aligned} \parallel \vec{v}_m - \vec{v}_n \parallel^2 &= 2\parallel \vec{y} - \vec{v}_n \parallel^2 + 2\parallel \vec{y} - \vec{v}_m \parallel^2 \\ &\quad -4\parallel \vec{y} - (\tfrac{1}{2})\,(\vec{v}_n + \vec{v}_m) \parallel^2. \end{aligned} \tag{1.10}$$

Como la norma al cuadrado del último término de (1.10) es al menos $d$, tenemos:

$$\parallel \vec{v}_m - \vec{v}_n \parallel^2 \;\leq\; 2\parallel (\vec{y} - \vec{v}_n) \parallel^2 + 2\parallel (\vec{y} - \vec{v}_m) \parallel^2 - 4d \tag{1.11}$$

Sea $\delta > 0$. Para $m, n$ mayores que $N(\delta/4)$, tenemos:

$$\parallel (\vec{y} - \vec{v}_n) \parallel^2 \;\leq\; d + \delta/4 \tag{1.12}$$

$$\parallel (\vec{y} - \vec{v}_m) \parallel^2 \;\leq\; d + \delta/4. \tag{1.13}$$

Sustituyendo ésto en (1.10) obtenemos:

$$\parallel (\vec{v}_m - \vec{v}_n) \parallel^2 \leq 2(d + \delta/4) + 2(d + \delta/4) - 4d = \delta, \tag{1.14}$$

luego la sucesión $\{\vec{v}_n\}$ es de Cauchy. Tendrá por tanto un límite único $\vec{v}$ en $M$ ($M$ es completo), y fácilmente se deduce que $\parallel \vec{y} - \vec{v} \parallel^2 = d$.

Por otra parte, para cualquier $\vec{z} \in M$ y para cualquier $\alpha$ real se tiene:

$$\parallel \vec{y} - \vec{v} - \alpha\vec{z} \parallel^2 = \parallel \vec{y} - \vec{v} \parallel^2 + \alpha^2\parallel \vec{z} \parallel^2 - 2\alpha < \vec{y} - \vec{v}, \vec{z} > \tag{1.15}$$

$$= d + \alpha^2\parallel \vec{z} \parallel^2 - 2\alpha < \vec{y} - \vec{v}, \vec{z} > \tag{1.16}$$

$$\geq d. \tag{1.17}$$

[1] Demostración tomada de Anderson (1971). Es más general de lo que estrictamente necesitamos, pero merece la pena enunciar este Teorema así para poderlo emplear inalterado en otros contextos (por ejemplo, en predicción lineal de procesos estocásticos). Una demostración más simple y menos general puede encontrarse en Arnold (1981), pág. 34.

Por tanto:

$$\alpha^2 \| \vec{z} \|^2 - 2\alpha < \vec{y} - \vec{v}, \vec{z} > \quad \geq \quad 0, \tag{1.18}$$

$$\alpha^2 \| \vec{z} \|^2 \quad \geq \quad 2\alpha < \vec{y} - \vec{v}, \vec{z} > . \tag{1.19}$$

Como (1.19) se ha de cumplir para cualquier posible valor de $\alpha$, ha de suceder que $< \vec{y} - \vec{v}, \vec{z} >= 0$, y como $\vec{z}$ es arbitrario en $M$, se deduce que $(\vec{y} - \vec{v}) \perp M$. Como además hemos visto que $\vec{v} \in M$, tenemos que $\vec{v}$ es proyección de $\vec{y}$ en $M$ (Definición 1.3). El desarrollo anterior muestra también que $\vec{v}$ es la mejor aproximación de $\vec{y}$ por un vector de $M$ (en términos de la norma definida).

Veamos, en fin, que ningún otro vector $\vec{u} \in M,\ \vec{u} \neq \vec{v}$ puede ser proyección de $\vec{y}$ en $M$, ni verificar $\| \vec{y} - \vec{u} \|^2 = d$. Supongamos que hubiera un tal $\vec{u}$. Entonces, $(\vec{y} - \vec{u}) = (\vec{y} - \vec{v}) + (\vec{v} - \vec{u})$. Además, $(\vec{y} - \vec{v}) \perp M$, y $(\vec{v} - \vec{u}) \in M$. Por tanto,

$$\begin{aligned} \| \vec{y} - \vec{u} \|^2 &= < \vec{y} - \vec{u}, \vec{y} - \vec{u} > \\ &= < (\vec{y} - \vec{v}) + (\vec{v} - \vec{u}), (\vec{y} - \vec{v}) + (\vec{v} - \vec{u}) > \\ &= \| \vec{y} - \vec{v} \|^2 + \| \vec{v} - \vec{u} \|^2 + 2 < \vec{y} - \vec{v}, \vec{v} - \vec{u} > \\ &\geq \| \vec{y} - \vec{v} \|^2, \end{aligned}$$

ya que $2 < \vec{y} - \vec{v}, \vec{v} - \vec{u} >= 0$, $\| \vec{v} - \vec{u} \|^2 \geq 0$, y $\| \vec{v} - \vec{u} \|^2 = 0$ implicaría $\vec{u} = \vec{v}$.

## CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

Algunos de los ejercicios que siguen requieren hacer uso de un ordenador y un programa especializado, tal como S-PLUS o R. Puede consultarse Becker et al. (1988). Es también de utilidad Venables and Ripley (1999a) (con sus complementos *on line*, Venables and Ripley (1999b)). El lector que realice los ejercicios en orden irá no obstante adquiriendo, si examina las soluciones de aquéllos que requieren alguna programación, familiaridad con cualquiera de los paquetes, sin necesidad de estudiar previamente el manual correspondiente. Un breve resumen de las funciones específicamente destinadas a regresión aparece en los Apéndices.

Donde quiera que se menciona S-PLUS puede utilizarse R, un programa similar en organización y sintaxis y que a efectos del presente curso es casi 100 % compatible; de hecho, los ejemplos de código que figuran intercalados en el texto se han procesado en su totalidad con R 1.9.1. Puede consultarse Venables et al. (1997) como referencia. Hay traducción castellana, Venables et al. (2000), un poco desfasada, y mucha documentación adicional en `http://cran.r-project.org/`.

**1.1** ¿Qué trascendencia tiene en la Sección 1.5 que $H$ (y, en consecuencia, su subespacio $M$) tengan estructura de espacio de Hilbert? Examinando la demostración del Teorema 1.1, vemos que se da por supuesta la existencia en $M$ del límite de la sucesión $\{v_n\}$ construida. Si $M$ no fuera espacio de Hilbert, tal límite podría no existir en $M$.

**1.2** En S-PLUS o R para asignar un valor a una variable podemos colocarla a la izquierda del operador `<-`. Por ejemplo,

```
x <- 5
```

El valor de la variable puede ser utilizado en cálculos subsiguientes; tecleando

```
x + 5
```

obtendríamos “10”.

**1.3** En S-PLUS o R para crear un vector y asignarlo a la variable `x` haremos:

```
x <- c(1,3,4)
```

**1.4** Para efectuar multitud de cálculos en R o S-PLUS empleamos funciones. Por ejemplo, para sumar varios números y asignar el resultado a `x` podríamos escribir:

```
x <- 5 + 7 + 12
```

o también

```
x <- sum(c(5,7,12))
```

que hace uso de la función `sum`.

**1.5** El producto interno euclídeo de dos vectores `x` e `y` viene dado por:

```
sum(x * y)
```

o alternativamente:

```
x %*% y
```

**1.6** En S-PLUS o R opera la “regla del reciclado”, que permite operar con operando disimilares. Por ejemplo, si:

```
a <- c(1,2,3) b <- 5
```

entonces, tecleando

```
a + b
```

obtendríamos el vector $(6 \quad 7 \quad 8)'$. El argumento más corto, `b`, se ha usado repetidamente para construir un operando que pueda sumarse a `a`.

**1.7** En S-PLUS o R es muy fácil acceder a elementos aislados de un vector. Por ejemplo, si:

```
a <- c(6,7,8)
```

entonces, tecleando las expresiones que aparece a la izquierda obtendríamos los resultados que se indican a la derecha:

```
a                produce:    6 7 8
a[1]             produce:    6
a[1:2]           produce:    6 7
a[c(1,2)]        produce:    6 7
a[-1]            produce:    7 8
a[-(1:2)]        produce:    8
a[c(F,F,T)]      produce:    8
a[a>6]           produce:    7 8
```

Los subíndices se ponen entre corchetes, []. Un subíndice negativo se interpreta como omitir el correspondiente valor. Además de subíndices numéricos, podemos emplear subíndices lógicos: `F` (falso) y `T` (cierto). Podemos incluso, como en la última línea, emplear expresiones que den como valor un vector lógico: `a >6` produce el vector `F T T`, que empleado como subíndices retorna los elementos de `a` mayores que 6.

**1.8** La función `help` permite interrogar a S-PLUS o R sobre el mode de empleo de cualquier función. Por ejemplo, para obtener la descripción de `sum` podríamos teclear:

```
help(sum)
```

Empléese la función `help` para averiguar el cometido de las siguientes funciones de R: `t`, `cbind`, `rbind`, `solve`, `scan`, `read.table`, `list`, `nrow`, `ncol`. Obsérvese que tecleando

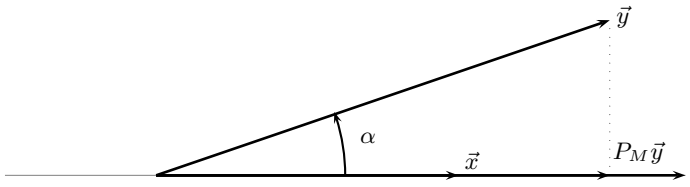
```
example(scan)
```

podemos ejecutar los ejemplos que aparecen en la documentación *on line* sin necesidad de reteclearlos[2]

**1.9** Recordemos que el producto euclídeo (o *escalar*) de dos vectores $\vec{x}\,, \vec{y}$ en $R^3$ verifica:

$$< \vec{x}\,, \vec{y}\, >= |\vec{x}\,||\vec{y}\,| \cos(\alpha)$$

siendo $\alpha$ el ángulo que ambos vectores forman. Esta igualdad se extiende a $R^N$ definiendo $\cos(\alpha)$ convenientemente (véase Definición A.3, pág. 179). Sea $P_M \vec{y}$ la proyección de $\vec{y}$ sobre el subespacio $M$. Si $|\vec{x}\,| = 1$, del esquema a continuación inmediatamente se deduce que $< \vec{x}\,, \vec{y}\, >= |P_M \vec{y}\,|$, siendo $M$ el subespacio generado por $\vec{x}\,$.




Dedúzcase que, en el caso general en que $|\vec{x}\,| \neq 1$, se verifica:

$$P_M \vec{x} = \frac{< \vec{x}\,, \vec{y}\, >}{< \vec{y}\,, \vec{y}\, >} \vec{x}$$

**1.10** Cuando escribimos expresiones como

[2] Sólo en R. S-PLUSno dispone de análoga facilidad.

```
sum(x * y)
```

estamos empleando funciones predefinidas (en este caso, `sum`). En S-PLUS o R no necesitamos limitarnos a ellas; el lenguaje es extensible por el usuario. Podríamos definir una función `eucl` para realizar el producto interno así:

```
eucl <- function(x,y) { sum(x*y) }
```

que asigna a `eucl` la función especificada en el lado derecho. Para invocarla con los vectores `u` y `v`, teclearíamos: `eucl(u,v)`.

Una función puede emplearse como bloque constructivo de otras, y esto hasta el nivel de complejidad que se desee. La norma euclídea podría calcularse mediante una función definida así:

```
norma.eucl <- function(x) {
                  sqrt(eucl(x,x)) }
```

que hace uso de `eucl` definida anteriormente. Tras esta definición, podemos calcular la norma euclídea de un vector `x` tecleando simplemente:

```
norma.eucl(x))
```

En realidad, la definición de una función como `eucl` es innecesaria: en S-PLUS o R podemos emplear `x%*% x` que cumple su cometido.

**1.11** Escríbase una función que, dados dos vectores arbitrarios $\vec{x}$ e $\vec{y}$, obtenga el vector proyección del primero sobre el espacio (unidimensional) generado por el segundo. Compruébese que el vector $\vec{z}$ resultante es efectivamente la proyección buscada (para lo cual es preciso ver: i) Que $\vec{z}$ es colineal con $\vec{y}$, y ii) Que $(\vec{x} - \vec{z}) \perp \vec{y}$).

**1.12** Demuéstrese que los siguientes cuatro vectores de $R^3$ son un sistema generador de dicho espacio, pero no base.

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

**1.13** (↑ 1.12) Selecciónese, de entre los cuatro vectores indicados en el Problema 1.12, tres que formen base de $R^3$.

**1.14** (↑ 1.11) Los siguientes dos vectores generan un subespacio 2-dimensional de $R^3$. Encuentrese —por ejemplo, mediante el procedimiento de Gram-Schmidt— una base ortonormal de dicho subespacio.

$$\begin{pmatrix}2\\0\\1\end{pmatrix}, \begin{pmatrix}1\\3\\0\end{pmatrix}$$

**1.15** Demuéstrese que la correspondencia $P_M : \vec{x} \longrightarrow \vec{y} = P_M\vec{x}$ es una aplicación lineal.

**1.16**  La estimación de un modelo de regresión lineal realiza una aproximación del vector respuesta $\vec{Y}$ similar a la que llevaría a cabo una red neuronal compuesta por una única neurona. "Similar" porque en el caso de una red neuronal la "estimación" (*entrenamiento* o *aprendizaje*) se realiza de ordinario mediante un proceso iterativo, cuyo resultado no necesariamente ha de coincidir exactamente con la estimación MCO. Un excelente manual sobre redes neuronales es Haykin (1998). Textos que tratan redes neuronales desde una perspectiva estadística son Ripley (1996) y Bishop (1996).

**1.17**  Hay alternativas a la regresión lineal: regresión no lineal y regresión no paramétrica (en que se considera una relación entre regresores y regresando que no está constreñida a ser lineal ni de ninguna otra forma funcional prefijada). En regresión no paramétrica se emplean principalmente tres métodos: *kernels*, vecinos más próximos y *splines*. Pueden consultarse, por ejemplo, Hastie et al. (2001) y Eubank (1988).

# Capítulo 2

# Estimación mínimo cuadrática.

## 2.1. Estimación de los parámetros.

Si $\vec{y}$ es un vector $N \times 1$, consideremos $H = R^N$ y $M =$ subespacio generado por las columnas de $X$. Si dotamos a $H$ del producto interno euclídeo $< \vec{v}, \vec{w} > = \vec{v}\,'\vec{w}$, de las Secciones 1.4 y 1.5 inmediatamente se deduce que el vector en $M$ más próximo a $\vec{y}$ (en el sentido de minimizar la norma al cuadrado del vector de residuos $\hat{\epsilon}$ ) es la proyección de $\vec{y}$ sobre $M$. Por consiguiente, ha de verificarse que $(\vec{y} - X\hat{\beta}) \perp M$. Como $M$ es el subespacio generado por las columnas de $X$, ello equivale a: $(\vec{y} - X\hat{\beta})'X = \vec{0}$, y de aquí se deduce que:

$$X'X\hat{\beta} = X'\vec{y}. \tag{2.1}$$

La igualdad matricial anterior recoge las *e*cuaciones normales. Si, como suponemos, rango$(X) = p$, $(X'X)$ es de rango completo, y posee inversa. Por tanto, el vector de estimadores de los parámetros será:

$$\hat{\beta} = (X'X)^{-1}X'\vec{y}. \tag{2.2}$$

Obsérvese que el supuesto de rango total de la matriz $X$ —y consiguientemente de $(X'X)$— es requerido exclusivamente para pasar de (2.1) a (2.2). Las ecuaciones normales se verifican en todo caso, y la proyección de $\vec{y}$ sobre $M$ es también única (Teorema 1.1, pág. 8). El defecto de rango en $X$ tiene tan solo por consecuencia que el vector $\hat{\beta}$ deja de estar unívocamente determinado. Volveremos sobre esta cuestión al hablar de multicolinealidad.

De (2.2) se deduce también que, en el caso de rango total, la proyección de $\vec{y}$ sobre $M$ viene dada por

$$P_M\vec{y} = X(X'X)^{-1}X'\vec{y}, \tag{2.3}$$

y el vector de residuos por

$$\begin{aligned} \hat{\epsilon} &= \vec{y} - X\hat{\beta} && (2.4)\\ &= \vec{y} - X(X'X)^{-1}X'\vec{y} && (2.5)\\ &= (I - X(X'X)^{-1}X')\vec{y} && (2.6)\\ &= (I - P_M)\vec{y}\,. && (2.7) \end{aligned}$$

Podemos ver $X\hat{\beta}$ y $\hat{\epsilon}$ como las proyecciones de $\vec{y}$ sobre dos espacios mutuamente ortogonales: $M$ y $M^{\perp}$. Las matrices $P_M$ e $(I - P_M)$ que, para aligerar la notación, denominaremos en lo sucesivo $P$ e $(I-P)$, sobreentendiendo el subespacio $M$, tienen algunas propiedades que detallamos a continuación.

**Teorema 2.1** *Sean $P$ e $(I - P)$ las matrices de proyección definidas en el párrafo anterior. Se verifica lo siguiente:*

1. *Las matrices $P$ e $(I - P)$ son simétricas e idempotentes.*
2. *rango$(I - P) = N - p$.*
3. *Se verifica que $(I - P)X = 0$.*

DEMOSTRACION:

El apartado 1) es inmediato. En cuanto a 2), siendo $(I - P)$ idempotente, su rango coincide con su traza (véase Apéndice A.1, Teorema A.1). Por tanto:

$$\begin{aligned} \text{rango}(I - P) &= \text{traza}(I - P) && (2.8)\\ &= \text{traza}(I) - \text{traza}(P) && (2.9)\\ &= N - \text{traza}[X(X'X)^{-1}X'] && (2.10)\\ &= N - \text{traza}[(X'X)^{-1}X'X] && (2.11)\\ &= N - p. && (2.12) \end{aligned}$$

El apartado 3), por último, se prueba sin más que efectuar el producto matricial indicado. Es además inmediato si reparamos en que la matriz $(I - P)$ proyecta sobre el subespacio $M^{\perp}$, por lo que su producto por cualquiera de los vectores columna de $X$ (pertenecientes a $M$) da el vector $\vec{0}$. El ejemplo siguiente ilustra el modo de realizar algunos de los cálculos en R.

**R: Ejemplo 2.1** *(uso de la función `lsfit`)*

El siguiente listado crea artificialmente una matriz $X$, el vector respuesta $\vec{y}$; a continuación realiza la regresión de dos formas. En la primera, se realizan los cálculos de modo explícito. En la segunda, se recurre a la función predefinida en R `lsfit`, que simplifica considerablemente el trabajo. Existen funciones alternativas más avanzadas que se introducen más adelante.

Al margen de la comodidad, `lsfit` realiza los cálculos de un modo mucho más eficiente en tiempo y estable numéricamente que el sugerido por la teoría: no se invierte la matriz $(X'X)$ sino que se emplea la descomposición QR (ver Sección D.2 o la obra monográfica Lawson and Hanson (1974)). Se trata de detalles que no necesitan preocuparnos por el momento.

```
--- Obtenido mediante R BATCH demo1.R
> options(digits=5)
> #
```

```
> #  Generamos los datos y realizamos la estimación
> #  aplicando la teoría de modo más directo.
> #
> X <- matrix(c(1,1,1,1,1,1,1,4,12,1,4,
+               13,0,6,7,0,2,2),6,3)     # matriz de diseño
> X
     [,1] [,2] [,3]
[1,]    1    1    0
[2,]    1    4    6
[3,]    1   12    7
[4,]    1    1    0
[5,]    1    4    2
[6,]    1   13    2
> beta  <- c(2,3,4)                      # parámetros
> y <- X %*% beta + rnorm(6)             # variable respuesta
>
> b <- solve(t(X)%*%X) %*% t(X) %*% y    # estimadores betas
>                                        # "solve" calcula la inversa.
> b
       [,1]
[1,] 2.6502
[2,] 2.9404
[3,] 3.8585
> e <- y - X %*% b                       # residuos
> e
          [,1]
[1,]  0.096439
[2,]  0.758922
[3,] -0.697449
[4,] -0.322218
[5,] -0.380389
[6,]  0.544695
> t(e) %*% (X %*% b)                     # comprobación ortogonalidad
            [,1]
[1,] 1.4724e-12
> s2 <- sum(e*e) / (nrow(X) - ncol(X))  # estimador varianza
> s2
[1] 0.53897
> #
> #  Lo mismo puede hacerse con mucha mayor comodidad mediante
> #  funciones de regresión especializadas.
> #
> ajuste <- lsfit(X,y,intercept=FALSE)
> #
> #  Los estimadores de los parámetros y los residuos coinciden
> #  con los obtenidos más arriba. El argumento "intercept=FALSE"
> #  es necesario porque la matriz tiene ya columna de unos: sin
> #  dicho argumento, lsfit añadiría una columna de unos
> #  adicional, creando una matriz de rango deficiente.
> #
> ajuste
$coefficients
    X1     X2     X3
2.6502 2.9404 3.8585
```

```
$residuals
[1]  0.096439  0.758922 -0.697449 -0.322218 -0.380389  0.544695

$intercept
[1] FALSE

$qr
$qt
[1] -75.28468  48.97029 -21.82294  -0.60712  -0.43452   1.02932

$qr
           X1        X2        X3
[1,] -2.44949 -14.28869 -6.940221
[2,]  0.40825  11.95129  3.583992
[3,]  0.40825  -0.63322 -5.655823
[4,]  0.40825   0.28718 -0.375532
[5,]  0.40825   0.03616 -0.004607
[6,]  0.40825  -0.71690  0.047314

$qraux
[1] 1.4082 1.0362 1.9256

$rank
[1] 3

$pivot
[1] 1 2 3

$tol
[1] 1e-07

attr(,"class")
[1] "qr"

> #
> #  Podemos obtener (y asignar) elementos individuales del
> #  ajuste obtenido. Por ejemplo, los residuos:
> #
> resid <- ajuste$residuals
> resid
[1]  0.096439  0.758922 -0.697449 -0.322218 -0.380389  0.544695
>
>
```

## 2.2. Propiedades del estimador mínimo cuadrático.

El siguiente resultado recoge algunas propiedades de $\hat{\beta}$.

**Teorema 2.2** *Se verifica que:*

1. $\hat{\beta}$ *es un estimador lineal insesgado de* $\vec{\beta}$.
2. *La matriz de covarianzas de* $\hat{\beta}$ *es* $\Sigma_{\hat{\beta}} = \sigma^2(X'X)^{-1}$.

3. *(Gauss-Markov). Si $\hat{\beta}$ es el estimador mínimo cuadrático ordinario de $\vec{\beta}$, cualquier otro estimador $\hat{\beta}_*$ de $\vec{\beta}$ que sea lineal e insesgado tiene matriz de covarianzas con elementos diagonales no menores que los de $\Sigma_{\hat{\beta}}$.*

DEMOSTRACION:

Tomando valor medio en (2.2):

$$\begin{aligned} E[\hat{\beta}] &= E[(X'X)^{-1}X'\vec{y}\,] \\ &= E[(X'X)^{-1}X'(X\vec{\beta}+\vec{\epsilon}\,)] \\ &= \vec{\beta} + E[(X'X)^{-1}X'\vec{\epsilon}\,] \\ &= \vec{\beta}. \end{aligned}$$

luego $\hat{\beta}$ es insesgado. Por consiguiente, la matriz de covarianzas $\Sigma_{\hat{\beta}}$ tendrá por expresión:

$$\begin{aligned} \Sigma_{\hat{\beta}} &= E(\hat{\beta}-\vec{\beta})(\hat{\beta}-\vec{\beta})' && (2.13)\\ &= E[(X'X)^{-1}X'(X\vec{\beta}+\vec{\epsilon}\,)-\vec{\beta}][(X'X)^{-1}X'(X\vec{\beta}+\vec{\epsilon}\,)-\vec{\beta}]' && (2.14)\\ &= E[(X'X)^{-1}X'\vec{\epsilon}\,][(X'X)^{-1}X'\vec{\epsilon}\,]' && (2.15)\\ &= E[(X'X)^{-1}X'\vec{\epsilon}\,\vec{\epsilon}\,'X(X'X)^{-1}] && (2.16)\\ &= (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1} && (2.17)\\ &= \sigma^2(X'X)^{-1}. && (2.18) \end{aligned}$$

Para demostrar 3), dado que restringimos nuestra atención a estimadores lineales, podemos escribir $\hat{\beta}_* = C\vec{Y}$, siendo $C$ una matriz de orden adecuado. Siempre podremos expresar $C$ así:

$$C = (X'X)^{-1}X' + D. \tag{2.19}$$

Puesto que nos limitamos a considerar estimadores insesgados, ha de verificarse: $E\hat{\beta}_* = EC\vec{Y} = \vec{\beta}$, y por tanto: $E[(X'X)^{-1}X' + D]\vec{Y} = \vec{\beta}$. De aquí se deduce:

$$\begin{aligned} E[(X'X)^{-1}X'(X\vec{\beta}+\vec{\epsilon}\,)+D(X\vec{\beta}+\vec{\epsilon}\,)] &= \vec{\beta}, && (2.20)\\ \vec{\beta} + DX\vec{\beta} &= \vec{\beta}, && (2.21) \end{aligned}$$

dado que $E\vec{\epsilon} = \vec{0}$. Como (2.21) se ha de verificar sea cual fuere $\vec{\beta}$, la insesgadez de $\hat{\beta}_*$ implica $DX = 0$.

La matriz de covarianzas de $\hat{\beta}_*$ es:

$$\Sigma_{\hat{\beta}_*} = E[(\hat{\beta}_*-\vec{\beta})(\hat{\beta}_*-\vec{\beta})']. \tag{2.22}$$

Pero:

$$\begin{aligned} (\hat{\beta}_*-\vec{\beta}) &= [(X'X)^{-1}X'+D]\vec{Y} - \vec{\beta} && (2.23)\\ &= [(X'X)^{-1}X'+D](X\vec{\beta}+\vec{\epsilon}\,)-\vec{\beta} && (2.24)\\ &= [(X'X)^{-1}X'+D]\vec{\epsilon}. && (2.25) \end{aligned}$$

donde (2.25) se ha obtenido haciendo uso de $DX = 0$. Llevando (2.25) a (2.22), obtenemos:

$$\Sigma_{\hat{\beta}_*} = E\{[(X'X)^{-1}X'+D]\vec{\epsilon}\,\vec{\epsilon}\,'[(X'X)^{-1}X'+D]'\ \} \tag{2.26}$$

que, de nuevo haciendo uso de que $DX = 0$, se transforma en:

$$\begin{aligned} \Sigma_{\hat{\beta}_*} &= (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} + \sigma^2 DID' && (2.27)\\ &= \sigma^2(X'X)^{-1} + \sigma^2 DD' && (2.28)\\ &= \Sigma_{\hat{\beta}} + \sigma^2 DD'. && (2.29) \end{aligned}$$

La matriz $DD'$ tiene necesariamente elementos no negativos en la diagonal principal (sumas de cuadrados), lo que concluye la demostración de 3). De forma completamente similar se puede demostrar una versión ligeramente más general: la estimación lineal insesgada con varianza mínima de cualquier forma lineal $\vec{c}\,'\vec{\beta}$ es $\vec{c}\,'\hat{\beta}$, siendo $\hat{\beta}$ el vector de estimadores mínimo cuadráticos.

## 2.3. Estimación de la varianza de la perturbación.

Una estimación de la varianza de la perturbación —y, consiguientemente, de la varianza de Y— puede obtenerse a partir de la suma de cuadrados de los residuos. Sea,

$$SSE = \| \hat{\epsilon} \|^2 = \hat{\epsilon}\,'\hat{\epsilon} = (\vec{Y} - X\hat{\beta})'(\vec{Y} - X\hat{\beta}). \quad (2.30)$$

Como

$$X\hat{\beta} = P\vec{Y} = X(X'X)^{-1}X'\vec{Y}\,, \quad (2.31)$$

tenemos que

$$\begin{aligned} (\vec{Y} - X\hat{\beta}) &= (I-P)\vec{Y} && (2.32)\\ &= (I-P)(X\vec{\beta} + \vec{\epsilon}\,) && (2.33)\\ &= (I-P)\vec{\epsilon}, && (2.34) \end{aligned}$$

y por tanto

$$SSE = \vec{Y}\,'(I-P)'(I-P)\,\vec{Y} = \vec{\epsilon}\,'(I-P)'(I-P)\,\vec{\epsilon}.$$

En virtud de la simetría e idempotencia de $(I-P)$,

$$\begin{aligned} SSE &= \vec{\epsilon}\,'(I-P)\vec{\epsilon} && (2.35)\\ &= \text{traza}\,\vec{\epsilon}\,'(I-P)\vec{\epsilon} && (2.36)\\ &= \text{traza}\,(I-P)\vec{\epsilon}\,\vec{\epsilon}\,'. && (2.37) \end{aligned}$$

Tomando valor medio en (2.37) tenemos:

$$E(SSE) = \text{traza}(I-P)(\sigma^2 I) = \sigma^2(N-p) \quad (2.38)$$

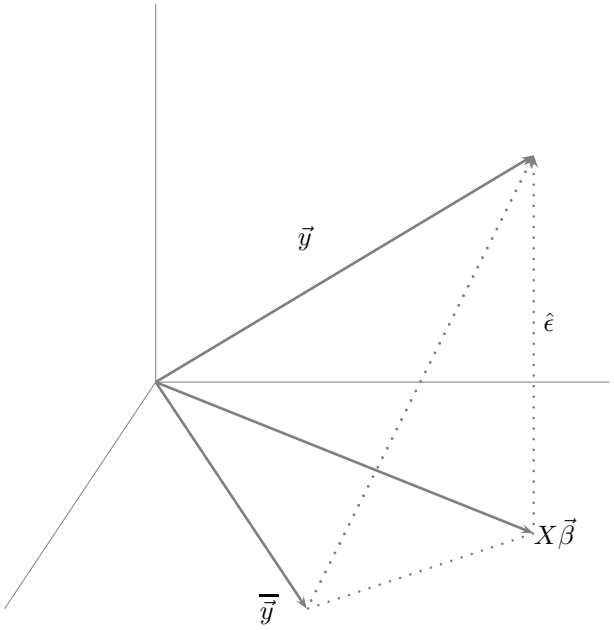
El último paso ha hecho uso de la propiedad

$$\text{traza}(I-P) = \text{rango}(I-P) = N-p$$

(Teorema 2.1, pág. 16). De (2.38) se deduce que $\hat{\sigma}^2 = \frac{SSE}{(N-p)}$ es un estimador insesgado de $\sigma^2$.

**Observación 2.1** En lo que sigue, $SSE$ denotará tanto la variable aleatoria definida más arriba como su valor en una experimentación concreta, contra la convención habitual con otras variables en que se emplean minúsculas para denotar sus valores en una experimentación. El contexto aclarará si nos estamos refiriendo a una variable aleatoria o a un valor experimental de la misma.

Figura 2.1: $X\hat{\beta}$ es la proyección de $\vec{y}$ sobre $M$. $R^2 = \cos^2 \alpha$




## 2.4. El coeficiente $R^2$

Hay una relación interesante entre $SSE$ y otras dos sumas de cuadrados que definimos a continuación. Sea $\vec{\overline{y}}$ el vector $N \times 1$ siguiente:

$$\vec{\overline{y}} = \begin{pmatrix} \overline{y} \\ \overline{y} \\ \vdots \\ \overline{y} \end{pmatrix}$$

en que $\overline{y}$ denota la media aritmética de las observaciones en $\vec{y}$. Definamos:

$$\begin{aligned} SST &= \| \vec{y} - \vec{\overline{y}} \|^2 \\ SSR &= \| X\hat{\beta} - \vec{\overline{y}} \|^2 \end{aligned}$$

Se verifica entonces el Teorema a continuación.

**Teorema 2.3** *Si $\vec{\overline{y}}$ pertenece al subespacio $M$ generado por las columnas de la matriz $X$ —lo que acontece, por ejemplo, siempre que dicha matriz tiene una columna de "unos"—, se verifica:*

$$SST = SSR + SSE \tag{2.39}$$

DEMOSTRACION:

$$\begin{aligned} SST &= \| \vec{y} - \vec{\overline{y}} \|^2 && (2.40)\\ &= \| \vec{y} - X\hat{\beta} + X\hat{\beta} - \vec{\overline{y}} \|^2 && (2.41)\\ &= \| \vec{y} - X\hat{\beta} \|^2 + \| X\hat{\beta} - \vec{\overline{y}} \|^2 + 2 < \vec{y} - X\hat{\beta}, X\hat{\beta} - \vec{\overline{y}} > && (2.42) \end{aligned}$$

Pero si $\vec{\overline{y}} \in M$, $(X\hat{\beta} - \vec{\overline{y}}) \in M$, y como quiera que $\hat{\epsilon} = (\vec{y} - X\hat{\beta}) \perp M$, el último producto interno es nulo. Por consiguiente (2.42) se reduce a (2.39).

Definimos $R^2 = SSR/SST$; se denomina a $R$ *coeficiente de correlación múltiple*. Claramente, $0 \leq R^2 \leq 1$, siempre que $X$ contenga una columna constante, ya que de (2.39) se obtiene:

$$\frac{SST}{SST} = \frac{SSR}{SST} + \frac{SSE}{SST} ,$$

luego $1 = R^2 + \frac{SSE}{SST}$, y como ambos sumandos son no negativos (son cocientes de sumas de cuadrados), $R^2$ necesariamente ha de tomar valores entre $0$ y $1$.

La igualdad (2.39) es fácil de visualizar con ayuda de la ilustración esquemática en la Fig. 2.1; es una generalización $N$-dimensional del teorema de Pitágoras. Obsérvese que si $\vec{\overline{y}}$ no perteneciera a $M$, que hemos representado como el plano horizontal, ya no podría asegurarse que $\hat{\epsilon}$ y $(X\hat{\beta} - \vec{\overline{y}})$ son ortogonales.

**Observación 2.2** En la Figura 2.1 puede visualizarse $R^2$ como el coseno al cuadrado del ángulo que forman los vectores $(\vec{y} - \vec{\overline{y}})$ y $(X\hat{\beta} - \vec{\overline{y}})$. Un valor "pequeño" de $R^2$ significa que este coseno es "pequeño", y el ángulo correspondiente "grande"; es decir, que $\vec{y}$ está muy elevado sobre el plano $M$. Por el contrario, $R^2$ grande implica que el ángulo referido es pequeño, y que $\vec{y}$ está próximo a su proyección en $M$.

**Observación 2.3** Si regresamos $\vec{y}$ solamente sobre una columna de "unos", obtenemos un único coeficiente de regresión estimado, $\hat{\beta}_0$ que resulta ser igual a $\overline{y}$ (compruébese). $SST$ puede interpretarse como la suma de cuadrados de los residuos de este modelo mínimo.

Si regresamos $\vec{y}$ sobre varios regresores *incluyendo la columna de "unos"* obtenemos una suma de cuadrados de los residuos igual a $SSE$ que nunca puede ser superior a $SST$. En efecto: al añadir regresores el ajuste no puede empeorar (¿por qué?). El coeficiente $R^2$ puede verse como una medida de la mejora en el ajuste atribuible a los regresores distintos de la columna de "unos". En efecto, el numerador de $R^2$ es $SST - SSE$, diferencia de suma de cuadrados entre el modelo ampliado y el mínimo. El denominador $SST$ meramente normaliza el numerador anterior para que tome valores entre 0 y 1.

Un valor "grande" de $R^2$ podemos interpretarlo como una mejora sustancial del modelo mínimo al incluir regresores distintos de la columna de "unos". Obsérvese que para que esta interpretación sea válida, uno de los modelos (el mínimo) ha de estar anidado en el otro, es decir, sus regresores han de ser un subconjunto de los del otro.

**Observación 2.4** Si ajustamos un modelo sin columna de "unos" podemos encontrarnos con que $R^2$ definido como en el Teorema 2.3 puede ser menor que cero. Es fácil de entender: puede que los regresores ensayados no den cuenta de la variabilidad de $\vec{y}$, y $SSE$ sea por tanto grande. Si acontece que $\vec{y}$ tiene poca variabilidad en torno a su media, $SST$ será en cambio pequeño, y $SST - SSE$ puede fácilmente ser negativo.

**Observación 2.5** Cuando no hay columna de "unos" algunos programas de ordenador automáticamente sustituyen $SST$ por

$$||\vec{y}||^2$$

(suma de cuadrados de las desviaciones *respecto del origen* en lugar de respecto a la media). Ello da lugar a una definición alternativa de $R^2$ que evita que pueda ser negativa.

## 2.5. Algunos lemas sobre proyecciones.

Los siguientes resultados, de muy sencilla prueba en la mayoría de los casos, resultan útiles en demostraciones posteriores.

**Lema 2.1** *Sea $H$ un espacio de Hilbert, y $M$ un subespacio. Todo $\vec{y} \in H$ tiene expresión única en la forma: $\vec{y} = \vec{u} + \vec{v}$, con $\vec{u} \in M$ y $\vec{v} \in M^{\perp}$.*

DEMOSTRACION:

Es una consecuencia inmediata de la unicidad de la proyección (Teorema 1.1, pág. 8).

**Lema 2.2** *Prefijadas las bases en $H$ y $M \subseteq H$, la aplicación lineal que proyecta sobre $M$ tiene por asociada una única matriz $P_M$.*

DEMOSTRACION:

Es una especialización del resultado según el cual, prefijadas las bases en ambos espacios, la matriz que representa una aplicación lineal de uno en otro es única. La proyección es una aplicación lineal (véase solución al Ejercicio 1.15).

**Lema 2.3** *La matriz de proyección sobre $M$ puede ser expresada así:*

$$P_M = TT',$$

*siendo $T$ una matriz cuyas columnas forman una base ortonormal de $M \subset H$.*

DEMOSTRACION:

Sea $N$ la dimensión de $H$ y $p$ la dimensión de $M$. Sea $\vec{v}_1, \ldots, \vec{v}_p$ una base de $M$ formada por vectores ortonormales, y $T$ la matriz $N \times p$ siguiente:

$$T = \begin{pmatrix} \vec{v}_1 & | & \vec{v}_2 & | & \ldots & | & \vec{v}_p \end{pmatrix}$$

Siempre podemos completar $\{\vec{v}_1, \ldots, \vec{v}_p\}$ con $N-p$ vectores adicionales $\{\vec{v}_{p+1}, \ldots, \vec{v}_N\}$ hasta obtener una base de $H$ (véase por ej. Grafe (1985), pág. 79). Además, los $N-p$ vectores adicionales pueden tomarse ortogonales entre sí y a los de $T$, y normalizados (por ejemplo, utilizando el procedimiento de ortogonalización de Gram-Schmidt; véase Grafe (1985), pág. 93). Entonces, para cualquier $\vec{y} \in H$ tendremos:

$$\vec{y} = \underbrace{\sum_{i=1}^{p} c_i \vec{v}_i}_{\in M} + \underbrace{\sum_{j=p+1}^{N} c_j \vec{v}_j}_{\in M^{\perp}}, \tag{2.43}$$

siendo $c_i$ $(i = 1, \ldots, N)$ las coordenadas de $\vec{y}$ en la base escogida. Premultiplicando ambos lados de (2.43) por $\vec{v}_i{}'$ $(i = 1, \ldots, p)$, obtenemos:

$$\vec{v}_i{}'\vec{y} \quad = \quad \vec{v}_i{}' \sum_{j=1}^{N} c_j \vec{v}_j = \sum_{j=1}^{N} c_j (\vec{v}_i{}'\vec{v}_j) = c_i, \tag{2.44}$$

en virtud de la ortonormalidad de los vectores $\{\vec{v}_i\}$. Entonces, $\vec{u} = P_M \vec{y}$ puede escribirse así:

$$\begin{aligned}
\vec{u} &= P_M \vec{y} \\
&= \sum_{i=1}^{p} (\vec{v}_i{}'\vec{y}\ )\vec{v}_i \\
&= \begin{pmatrix} \vec{v}_1 & | & \vec{v}_2 & | & \cdots & | & \vec{v}_p \end{pmatrix} \begin{pmatrix} \vec{v}_1{}'\vec{y} \\ \vec{v}_2{}'\vec{y} \\ \vdots \\ \vec{v}_p{}'\vec{y} \end{pmatrix} \\
&= \begin{pmatrix} \vec{v}_1 & | & \vec{v}_2 & | & \cdots & | & \vec{v}_p \end{pmatrix} \begin{pmatrix} \vec{v}_1{}' \\ \vec{v}_2{}' \\ \vdots \\ \vec{v}_p{}' \end{pmatrix} \vec{y} \\
&= TT'\vec{y}
\end{aligned}$$

**Lema 2.4** *La matriz $P_M$ es simétrica idempotente.*

DEMOSTRACION:

La matriz $P_M$ es única (Lema 2.2) y puede expresarse siempre como $TT'$ (Lema 2.3). Entonces:

$$\begin{aligned}
P_M' &= (TT')' = TT' = P_M \\
P_M P_M &= TT'TT' = T(T'T)T' = TT' = P_M.
\end{aligned}$$

**Lema 2.5** *Denotamos por $R(C)$ el subespacio generado por las columnas de $C$, siendo $C$ una matriz cualquiera. $P_M$ denota la matriz de proyección sobre un cierto subespacio $M$. Entonces:*

$$R(P_M) = M.$$

DEMOSTRACION:

Claramente $R(P_M) \subseteq M$. Por otra parte, para todo $\vec{x} \in M$,

$$P_M \vec{x} = \vec{x} \Longrightarrow M \subseteq R(P_M).$$

**Lema 2.6** *Si $P_M$ es la matriz asociada al operador de proyección sobre $M$, $(I - P_M)$ es simétrica, idempotente, y está asociada al operador de proyección sobre $M^{\perp}$.*

DEMOSTRACION:

Es consecuencia inmediata de los Lemas 2.1 y 2.4.

**Lema 2.7** *Toda matriz simétrica idempotente $P$ representa una proyección ortogonal sobre el subespacio generado por las columnas de $P$.*

DEMOSTRACION:

Consideremos la identidad $\vec{y} = P\vec{y} + (I - P)\vec{y}$. Claramente, $(I - P)\vec{y} \perp P\vec{y}$ y además $(I - P)\vec{y} = \vec{y} - P\vec{y}$ es ortogonal a $P\vec{y}$. Por tanto, $P\vec{y}$ es proyección de $\vec{y}$ sobre un cierto subespacio, que, de acuerdo con el Lema 2.5, es el generado por las columnas de $P$.

**Definición 2.1** *Sea $D$ una matriz cualquiera, de orden $m \times n$. Decimos que $D^-$ es una pseudo-inversa (o inversa generalizada) de $D$ si:*

$$DD^-D = D \tag{2.45}$$

En general, $D^-$ así definida no es única. En el caso particular de que $D$ sea una matriz cuadrada de rango completo, $D^- = D^{-1}$.

**Lema 2.8** *Sea $D$ una matriz $m \times n$ cualquiera. Sea $\vec{c}$ una matriz $m \times 1$ y $\vec{z}$ un vector de variables. Si el sistema:*

$$D\vec{z} = \vec{c} \tag{2.46}$$

*es compatible, una solución viene dada por $\vec{z} = D^-\vec{c}$, siendo $D^-$ una pseudo-inversa.*

DEMOSTRACION:

De (2.45) deducimos:

$$DD^-D\vec{z} = \vec{c} \tag{2.47}$$

y sustituyendo (2.46) en (2.47):

$$DD^-\vec{c} = \vec{c} \tag{2.48}$$

$$D(D^-\vec{c}) = \vec{c} \tag{2.49}$$

lo que muestra que $D^-\vec{c}$ es solución de (2.46).

En realidad, es posible probar un resultado algo más fuerte[1]; *toda* solución de (2.46) puede expresarse como $D^-\vec{c}$ para alguna elección de $D^-$.

**Lema 2.9** *Si $M = R(X)$, entonces $P_M = X(X'X)^-X'$.*

DEMOSTRACION:

Sea $\vec{y}$ un vector cualquiera. Su proyección sobre $R(X)$ ha de ser de la forma $X\hat{\beta}$, y verificar las ecuaciones normales (2.1) en la pág. 15:

$$X'X\hat{\beta} = X'\vec{y} \tag{2.50}$$

[1] Cf. Searle (1971), Teorema 8, pág. 26.

Identificando $D = X'X$, $\vec{z} = \hat{\beta}$, y $\vec{c} = X'\vec{y}$, el lema anterior garantiza que $(X'X)^- X'\vec{y}$ será una posible solución para $\hat{\beta}$ (no necesariamente única, ya que hay múltiples $(X'X)^-$ en general); no obstante, $X(X'X)^- X'\vec{y}$ es la *única* proyección de $\vec{y}$ sobre $M$, y $X(X'X)^- X'$ es la *única* matriz de proyección. La unicidad de la proyección se demostró en el Teorema 1.1, pág. 8. La unicidad de la matriz de proyección, fue objeto del Lema 2.2.

Como se ha indicado, hay en general múltiples inversas generalizadas $D^-$, cada una de las cuales da lugar a una diferente solución[2], del sistema (2.48)–(2.49).

## CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

**2.1** ¿Que efecto tienen sobre los estimadores $\hat{\beta}$ cambios en la escala de los regresores en $X$?. Demuéstrese.

**2.2** Haciendo uso del mismo argumento empleado (en (2.38), pág. 20) para mostrar que $SSE/(N-p)$ es un estimador insesgado de $\sigma^2$, compruébese que, dada una muestra aleatoria simple $Z_1, \ldots, Z_n$, el estimador de la varianza

$$\sigma_Z^2 = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})^2$$

no es insesgado.

**2.3** Extiéndase el teorema de Gauss-Markov, para probar la afirmación hecha al final de la Sección 2.3 (pág. 20): si $\vec{c}'\vec{\beta}$ es cualquier forma lineal, en el caso de rango completo el estimador insesgado de varianza mínima de $\vec{c}'\vec{\beta}$ es $\vec{c}'\hat{\beta}$.

**2.4** La Definición 2.1 no individualiza una única inversa generalizada, salvo cuando $D$ es cuadrada de rango completo. Las siguientes condiciones, la primera de las cuáles coincide con (2.45), proporcionan una única definición de inversa generalizada (la inversa de Moore-Penrose):

$$DD^-D = D; \qquad D^-DD^- = D^-; \qquad D^-D \text{ y } DD^- \text{ simétricas.}$$

A la única matriz $D^-$ así especificada se la denomina inversa de Moore-Penrose. Sobre inversas generalizadas e inversas de Moore-Penrose puede consultarse Searle (1971) y Radhakrishna Rao and Kumar Mitra (1971)

**2.5** (↑ 1.14) Resuélvase el problema 1.14, pág. 13, haciendo uso de regresión lineal. (Ayuda: basta normalizar el primer vector y regresar el segundo sobre él. El vector de residuos de esta regresión es ortogonal al primero.)

**2.6** (↑ 2.5) Escríbase una función en R que resuelva el problema 2.5 de un modo completamente general: debe admitir como único argumento una matrix de rango completo cuyas columnas contengan los vectores a ortonormalizar, y devolver una matrix de las mismas dimensiones cuyas columnas sean los vectores ortonormalizados.

**2.7** ¿Cuándo incluir y cuándo no una columna de "unos"? En general, siempre convendrá hacerlo. Las únicas situaciones en que no será conveniente son aquéllas en que la columna de unos crearía una dependencia lineal exacta entre las columnas de la matriz $X$.

[2]Una forma de calcular una posible $D^-$ puede verse en Seber (1977), pág. 76 y en el Apéndice A.1; una monografía sobre el tema es Ben-Israel and Greville (1974).

El no incluir columna de "unos" fuerza a la recta (o hiperplano) de regresión a pasar por el origen. Salvo que haya buenos motivos para ello, no querremos forzar tal cosa en nuestra regresión, especialmente si como sucede en multitud de ocasiones el origen es arbitrario.

**2.8** (↑ 2.1)(↑ 2.7) Pensemos en la siguiente situación: un investigador está interesado en dilucidar si la temperatura de un fluido ($y$, medida en unidades adecuadas) está influida por la temperatura ($X_1$, medida en grados centígrados). Cuenta con las siguientes observaciones:

$$\vec{y} = \begin{pmatrix} 5{,}8 \\ 4{,}7 \\ 4{,}9 \\ 3{,}8 \\ 2{,}1 \end{pmatrix} \qquad X_1 = \begin{pmatrix} -10 \\ -6{,}2 \\ -2{,}5 \\ 3{,}0 \\ 4{,}6 \end{pmatrix}$$

Imaginemos que ajusta una regresión a dichos datos. Los resultados pueden verse en el siguiente fragmento en R:

```
--- Obtenido mediante R BATCH demo2a.R
> y <- c(5.8, 4.7, 4.9, 3.8, 2.1)
> X <- c(-10, -6.2, -2.5, 3.0, 4.6)
> ajuste <- lsfit(X,y,intercept=FALSE)
> ajuste$coefficients
        X
-0.447984
>
```

El coeficiente que afecta a la única variable es negativo (= -0.447984), que estaríamos tentados de interpretar así: por cada grado que aumenta la temperatura, disminuye en 0.447984 la velocidad de sedimentación. (Quedaría por ver si la estimación del coeficiente de regresión es de fiar, cuestión que abordaremos más adelante.)

Supongamos ahora que otro investigador repite el mismo análisis, pero en lugar de expresar las temperaturas en grados centígrados ($C$) lo hace en grados Fahrenheit (F) cuya relación con los centígrados viene dada por $C = \frac{5}{9}(F - 32)$ ($\Rightarrow F = \frac{9}{5}C + 32$). Los cálculos, siempre haciendo una regresión pasando por el origen, serían ahora:

```
--- Obtenido mediante R BATCH demo2b.R
> y <- c(5.8, 4.7, 4.9, 3.8, 2.1)
> X <- c(-10, -6.2, -2.5, 3.0, 4.6)        # en centígrados
> X <- (9/5)*X + 32                        # en Fahrenheit
> ajuste <- lsfit(X,y,intercept=FALSE)
> ajuste$coefficients
        X
0.1226477
```

¡Ahora el coeficiente afectando a la variable temperatura es positivo, dando la impresión de una asociación *directa* entre temperatura y velocidad de sedimentación! Claramente, tenemos motivo para preocuparnos si llegamos a conclusiones diferentes dependiendo de nuestra elección de los sistemas de medida —enteramente convencionales ambos—. El problema desaparece si incluimos una columna de unos en ambos análisis, para dar cuenta de los diferentes orígenes.

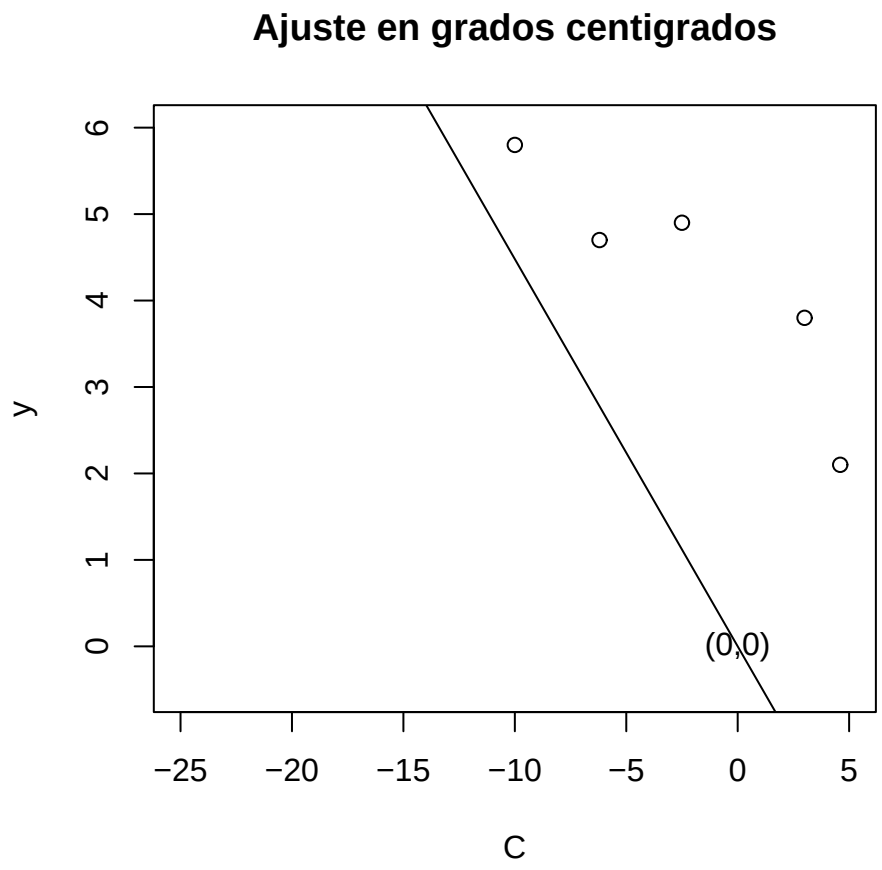
```
--- Obtenido mediante R BATCH demo2c.R
> y <- c(5.8, 4.7, 4.9, 3.8, 2.1)
> X <- c(-10, -6.2, -2.5, 3.0, 4.6)   # en grados centígrados
> ajuste <- lsfit(X,y)                # ajuste con columna de "unos".
> ajuste$coefficients
 Intercept          X
 3.8011850 -0.2066734
> X <- (9/5)*X + 32                   # en Fahrenheit
> ajuste <- lsfit(X,y)
> ajuste$coefficients
 Intercept          X
 7.4753790 -0.1148186
> ajuste$coefficients[2]*(9/5)        # el coeficiente de X coincide
         X
-0.2066734
>                                     # tras corregir el efecto de la escala
```

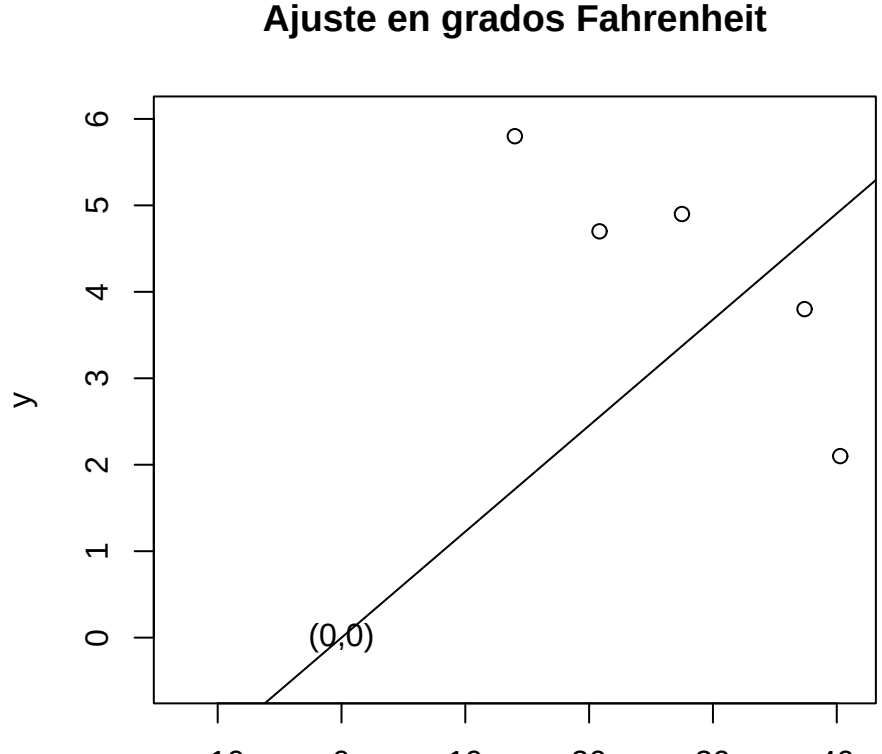
Los coeficientes de X no son ahora iguales (porque los grados Fahrenheit son más “pequeños”), pero si relacionados por un factor de escala y darían lugar a la misma conclusión de asociación inversa entre ambas magnitudes. La inversión del signo del coeficiente se explica comparando en la Figura 2.2 los puntos muestrales (en escalas comparables) y las respectivas rectas de regresión. Dichas rectas de regresión y las gráficas se han generado mediante

```
--- Obtenido mediante R BATCH demo2d.R
> #
> #  Parámetros controlando la generación de gráficos
> #
> postscript(file="demo2d.eps",horizontal=FALSE,
+            width=5,height=10)
> par(mfcol=c(2,1))
> #
> y <- c(5.8, 4.7, 4.9, 3.8, 2.1)
> C <- c(-10, -6.2, -2.5, 3.0, 4.6)        # en grados centígrados
> ajuste <- lsfit(C,y,intercept=FALSE)     # sin columna de "unos".
> plot(C,y,ylim=c(-0.5,6),xlim=c(-25,5),
+      main="Ajuste en grados centigrados")
> abline(a=0,b=ajuste$coefficients)
> text(x=0,y=0,labels="(0,0)")
> #
> F <- (9/5)*C + 32                        # en Fahrenheit
> ajuste <- lsfit(F,y,intercept=FALSE)     # sin columna de "unos".
> plot(F,y,ylim=c(-0.5,6),xlim=c(-13,41),
+      main="Ajuste en grados Fahrenheit")
> text(x=0,y=0,labels="(0,0)")
> abline(a=0,b=ajuste$coefficients)
```

Puede verse que el forzar a ambas a pasar por el origen las obliga a tener pendiente de signo opuesto para aproximar la nube de puntos.

Figura 2.2: En un ajuste sin término constante, la pendiente depende de la elección arbitraria del origen



# Capítulo 3

# Identificación. Estimación condicionada.

## 3.1. Modelos con matriz de diseño de rango deficiente.

Entre los supuestos habituales (Sección 1.3, apartados 1 a 3) estaba el de que el rango de la matriz de diseño $X$ coincide con el número de sus columnas, $p$. Cuando ésto no ocurre, sigue habiendo una única proyección de $\vec{y}$ sobre $M = R(X)$, tal como ha quedado demostrado. Ocurre sin embargo (Lema 2.9) que $\hat{\beta} = (X'X)^{-}X'\vec{y}$ no es único.

La Figura 3.1 resulta iluminante a este respecto; el plano horizontal representa $M$, y en él yacen los vectores $\vec{X}_0, \ldots, \vec{X}_{p-1}$ que lo generan. La proyección $X\hat{\beta}$ es única.

Si $\vec{X}_0, \ldots, \vec{X}_{p-1}$ son linealmente independientes, forman base del espacio que generan, y los coeficientes $\hat{\beta}_0, \ldots, \hat{\beta}_{p-1}$ que permiten expresar $P_M\vec{y}$ como combinación lineal de dichos vectores son únicos.

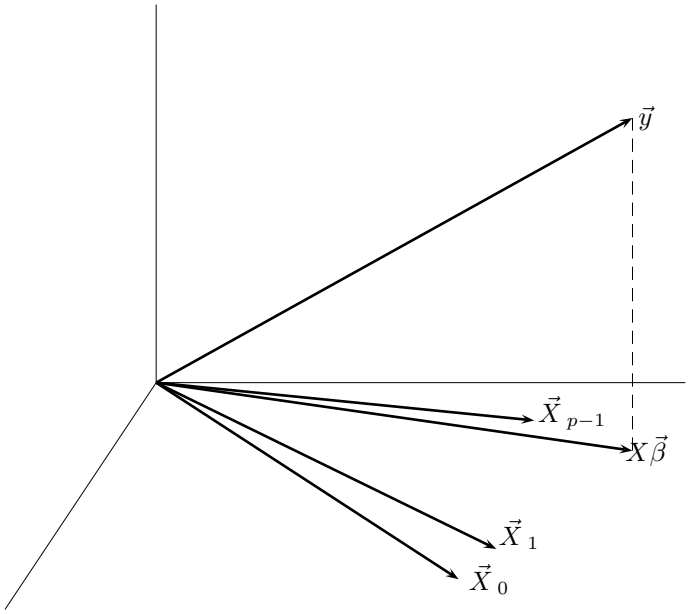
Si, como acontece en el caso de rango deficiente de la matriz $X$, los vectores $\vec{X}_0, \ldots, \vec{X}_{p-1}$ no son linealmente independientes, hay infinidad de maneras de expresar $P_M\vec{y}$ como combinación lineal de ellos. No hay por tanto una única estimación mínimo cuadrática del vector $\vec{\beta}$. Se dice que hay *multicolinealidad exacta* entre las columnas de la matriz de diseño $X$.

Incluso aunque el vector $\vec{\beta}$ no sea estimable por no estar $\hat{\beta}$ unívocamente determinado, puede haber algunos parámetros o combinaciones lineales de parámetros que sí puedan estimarse.

**Definición 3.1** *Decimos que una función lineal de los parámetros $\vec{a}\,'\vec{\beta}$ es estimable si existe un vector $\vec{c}$ de constantes tal que:*

$$E[\vec{c}\,'\vec{Y}] = \vec{a}\,'\vec{\beta}$$

Figura 3.1: Regresión en el caso de matrix $X$ de rango deficiente.




Es muy fácil caracterizar las funciones estimables, como pone de manifiesto el siguiente teorema.

**Teorema 3.1** *La función lineal $\vec{a}\,'\vec{\beta}$ es estimable si $\vec{a} \in R(X')$.*

DEMOSTRACION:

$$\vec{a}\,'\vec{\beta} \quad = \quad E[\vec{c}\,'\vec{Y}] = E[\vec{c}\,'(X\vec{\beta} + \vec{\epsilon}\,)] = \vec{c}\,'X\vec{\beta} \tag{3.1}$$

Como (3.1) ha de verificarse para cualesquiera valores de $\vec{\beta}$, ha de existir $\vec{c}$ tal que: $\vec{c}\,'X = \vec{a}\,'$, lo que demuestra que $\vec{a} \in R(X')$.

El teorema anterior incluye como caso particular el de parámetros aislados, $\beta_i$. En efecto, podemos ver $\beta_i$ como la función lineal $\vec{e}\,'_{i+1}\vec{\beta}$, en que $\vec{e}_i$ es un vector de ceros con un 1 en posición $i$–ésima. Entonces, $\beta_i$ es estimable si $\vec{e}_i \in R(X')$. La totalidad de los parámetros serán estimables si $\{\vec{e}_1, \ldots, \vec{e}_p\}$ (que son linealmente independientes) están en $R(X')$. Esto requiere que la dimensión de $R(X')$ sea $p$, es decir, que $X$ sea de rango completo.

Una última cuestión resulta de interés. Hemos visto que la inestimabilidad de los parámetros es consecuencia de la indeterminación del sistema de ecuaciones normales:

$$(X'X)\hat{\beta} = X'\vec{y}$$

Si contamos con información adicional sobre $\vec{\beta}$ que podamos imponer sobre el vector de estimadores $\hat{\beta}$, podemos añadir al anterior sistema ecuaciones adicionales que

reduzcan o resuelvan la indeterminación. Por ejemplo, si supiéramos que $A\vec{\beta} = \vec{c}$, podríamos formar el sistema:

$$(X'X)\hat{\beta} = X'\vec{y} \quad (3.2)$$

$$A\hat{\beta} = \vec{c} \quad (3.3)$$

y, dependiendo del rango de $X'X$ y $A$, obtener estimaciones únicas de $\vec{\beta}$. Se dice entonces que las relaciones $A\hat{\beta} = \vec{c}$ son *restricciones de identificación*.

Las restricciones de identificación persiguen la estimabilidad añadiendo información extramuestral a un problema en que la información muestral es insuficiente. De naturaleza completamente diferente es el problema de estimación condicionada que se estudia a continuación.

## 3.2. Estimación condicionada.

En ocasiones deseamos imponer a las estimaciones de los parámetros $\vec{\beta}$ ciertas condiciones, ya para hacer el modelo interpretable (por ej., en el caso del Análisis de Varianza, al que nos referiremos más adelante), ya porque así lo imponen criterios extra-estadísticos. Puede que el conjunto de restricciones que impongamos sea tal que, junto con las ecuaciones normales, determine un único vector de estimadores $\hat{\beta}$, en un problema que previamente admitía múltiples soluciones (se dice de tales restricciones que son identificadoras). En tal caso, todo se reduce a resolver el sistema (3.3). Las restricciones se han limitado a remover la indeterminación presente en las ecuaciones normales.

En otras ocasiones, partimos de un modelo ya identificable (con solución única para las ecuaciones normales), pero no obstante deseamos imponer una restricción que viene dictada al margen de los datos, como ilustra el ejemplo a continuación.

> **Ejemplo 3.1** Si quisiéramos estimar los parámetros de una función de producción Cobb-Douglas $Q = \alpha L^{\ell} K^{\gamma}$, podríamos desear que las estimaciones de los parámetros $\ell$ y $\gamma$ verificaran la condición $\hat{\ell} + \hat{\gamma} = 1$ (rendimientos constantes a escala). Con tres o más observaciones es perfectamente posible estimar $\alpha$, $\ell$ y $\gamma$; la restricción es innecesaria desde el punto de vista de la estimabilidad de los parámetros. No obstante, puede formar parte de la especificación que deseamos: no queremos ajustar cualquier función de producción Cobb-Douglas a nuestros datos, sino una con rendimientos constantes a la escala.

De un modo general, nos planteamos el problema siguiente:

$$\text{mín} \parallel \vec{y} - X\hat{\beta} \parallel^2 \qquad \text{condicionado a :} \quad A\hat{\beta} = \vec{c} \quad (3.4)$$

Está claro que no podemos esperar obtener la solución de este problema resolviendo un sistema como (3.3), que en general será incompatible. Resolveremos el problema por el procedimiento que conocemos, proyectando $\vec{y}$ sobre un subespacio adecuado; para ello habremos de transformar el problema en otro equivalente, que nos permita utilizar la técnica de la proyección. Previamente precisamos algunos resultados instrumentales.

**Lema 3.1** *Si $K(C)$ designa el núcleo de la aplicación lineal representada por la matriz $C$, se tiene:*

$$K(C) = [R(C')]^{\perp}$$

DEMOSTRACION:

$$\vec{x} \in K(C) \Longleftrightarrow C\vec{x} = \vec{0} \Longleftrightarrow \vec{x}\,'C' = \vec{0}\,' \Longleftrightarrow \vec{x} \perp R(C')$$

**Lema 3.2** *Si $h \subseteq M \subseteq H$, y $P_h,\ P_M$ son las matrices de proyección sobre los subespacios respectivos, se verifica: $P_M P_h = P_h P_M = P_h$*

DEMOSTRACION:

Para cualquier $\vec{v} \in H$,

$$\begin{aligned} P_h\vec{v} \in h \subseteq M \quad &\Rightarrow \quad P_M P_h \vec{v} = P_h \vec{v} \\ &\Rightarrow \quad P_M P_h = P_h \end{aligned}$$

La simetría de $P_M$ y $P_h$ (Lema 2.4) implica entonces que: $P_h = P_h' = P_h' P_M' = P_h P_M$

**Lema 3.3** *Si $h \subseteq M \subseteq H$, se tiene:*

$$P_M - P_h = P_{M \cap h^\perp}$$

DEMOSTRACION:

Partimos de la identidad,

$$P_M\vec{v} = P_h\vec{v} + (P_M\vec{v} - P_h\vec{v})$$

en la que $P_h\vec{v} \in h \subseteq M$ mientras que $(P_M\vec{v} - P_h\vec{v}) \in M$. Por otra parte,

$$\begin{aligned} < P_h\vec{v}, (P_M\vec{v} - P_h\vec{v}) > \quad &= \quad \vec{v}\,' P_h (P_M\vec{v} - P_h\vec{v}) \\ &= \quad \vec{v}\,' (P_h P_M - P_h)\vec{v} \\ &= \quad 0, \end{aligned}$$

la última igualdad en virtud del Lema 3.2. Por consiguiente, $(P_M - P_h)$, que es simétrica idempotente, proyecta sobre un subespacio ortogonal a $h$ e incluído en $M$; lo denotaremos mediante $M \cap h^\perp$.

**Lema 3.4** *Sea $B$ una matriz cualquiera, y $K(B)$ el núcleo de la aplicación lineal que representa. Sea $M$ un subespacio de $H$ y $h = M \cap K(B)$. Entonces, $M \cap h^\perp = R(P_M B')$.*

**Demostración.** En primer lugar, $M \cap h^\perp$ puede expresarse de otro modo que hará más simple la demostración. En efecto,

$$M \cap h^\perp = M \cap R(B'); \tag{3.5}$$

véase el Ejercicio 3.2 al final de este Capítulo.

Probaremos ahora que ambos subespacios considerados en el enunciado son el mismo, utilizando la expresión (3.5), y mostrando la mutua inclusión.

i) $M \cap h^\perp \subseteq R(P_M B')$. En efecto,

$$\begin{aligned} \vec{x} \in M \cap h^\perp \quad &\Longrightarrow \quad \vec{x} \in M \cap R(B') \\ &\Longrightarrow \quad \exists \vec{a}\colon \quad \vec{x} = B'\vec{a} \\ &\Longrightarrow \quad P_M\vec{x} = P_M B'\vec{a} \\ &\Longrightarrow \quad \vec{x} = P_M B'\vec{a} \\ &\Longrightarrow \quad \vec{x} \in R(P_M B') \end{aligned}$$

ii) $M \cap h^{\perp} \supseteq R(P_M B')$. Es inmediato, ya que,

$$\vec{x} \in R(P_M B') \Longrightarrow \vec{x} \in R(P_M) \Longrightarrow \vec{x} \in M$$

Sea ahora $\vec{z} \in h$. Entonces, como $h = M \cap K(B)$, $\vec{z} \in M$ y $\vec{z} \in K(B)$. Por tanto:

$$< \vec{x}, \vec{z} > \quad = \quad \vec{x}\,' \vec{z} = \vec{a}\,' B P_M \vec{z} = \vec{a}\,' B \vec{z} = 0$$

Por tanto, $\vec{x} \in M$ y además $\vec{x} \perp h$, luego $\vec{x} \in M \cap h^{\perp}$, lo que prueba ii) y finaliza la demostración del lema.

Regresemos ahora al problema (3.4) que teníamos planteado. Convendrá reparametrizar el problema mediante la transformación

$$\tilde{Y} = \vec{Y} - X\vec{\delta} \tag{3.6}$$
$$\vec{\gamma} = \vec{\beta} - \vec{\delta}, \tag{3.7}$$

siendo $\vec{\delta}$ una solución cualquiera de $A\vec{\delta} = \vec{c}$ (de no existir tal solución, no tendría sentido el problema; estaríamos imponiendo condiciones a los parámetros imposibles de satisfacer). Supondremos $X$ y $A$ de rango completo, pero es fácil generalizar el tratamiento reemplazando las inversas por inversas generalizadas. Se tiene entonces que:

$$\begin{aligned} \vec{Y} &= X\vec{\beta} + \vec{\epsilon} \Longrightarrow \vec{Y} - X\vec{\delta} = X\vec{\beta} - X\vec{\delta} + \vec{\epsilon} \Longrightarrow \tilde{Y} = X\vec{\gamma} + \vec{\epsilon} \\ A\vec{\beta} &= \vec{c} \Longrightarrow A(\vec{\gamma} + \vec{\delta}) = \vec{c} \Longrightarrow A\vec{\gamma} = \vec{c} - A\vec{\delta} \Longrightarrow A\vec{\gamma} = \vec{0} \end{aligned}$$

y el problema original (3.4) puede ahora reescribirse así:

$$\text{mín} \parallel \tilde{y} - X\hat{\gamma} \parallel^2 \qquad \text{condicionado a} \quad A\hat{\gamma} = \vec{0},$$

o, alternativamente,

$$\text{mín} \parallel \tilde{y} - X\hat{\gamma} \parallel^2 \qquad \text{condicionado a :} \quad A(X'X)^{-1}X'(X\hat{\gamma}) = \vec{0}. \tag{3.8}$$

La ecuación (3.8) muestra que el $X\hat{\gamma}$ buscado no es sino la proyección de $\tilde{y}$ sobre un cierto subespacio: $h = M \cap K(A(X'X)^{-1}X')$. Basta proyectar $\tilde{y}$ sobre $h$ para obtener $X\hat{\gamma}$ y, si $X$ es de rango completo, $\hat{\gamma}$. Si denotamos por $\hat{\gamma}_h$ las estimaciones mínimo cuadráticas condicionadas o restringidas por $A\hat{\gamma} = \vec{0}$, tenemos que:

$$X\hat{\gamma}_h = P_h \tilde{y} \tag{3.9}$$
$$= (P_M - P_{M\cap h^{\perp}})\tilde{y} \tag{3.10}$$
$$= [X(X'X)^{-1}X' - P_{M\cap h^{\perp}}]\tilde{y} \tag{3.11}$$

Pero es que, de acuerdo con el Lema 3.4,

$$M \cap h^{\perp} = R[\underbrace{X(X'X)^{-1}X'}_{P_M}\ \underbrace{X((X'X))^{-1}A'}_{B'}] = R[\underbrace{X((X'X))^{-1}A'}_{Z}]$$

Por consiguiente, $P_{M\cap h^{\perp}}$ es, de acuerdo con el Lema 2.9,

$$P_{M\cap h^{\perp}} = Z(Z'Z)^{-1}Z' \tag{3.12}$$

ecuación que, llevada a (3.11), proporciona:

$$\begin{aligned} X\hat{\gamma}_h &= X(X'X)^{-1}X'\tilde{y} - X((X'X))^{-1}A'[A((X'X))^{-1}A']^{-1}A((X'X))^{-1}X'\tilde{y} \\ &= X\hat{\gamma} - X((X'X))^{-1}A'[A((X'X))^{-1}A']^{-1}A\hat{\gamma} \end{aligned} \quad (3.13)$$

Si $X$ es de rango total, como venimos suponiendo, de (3.13) se deduce:

$$\hat{\gamma}_h = \hat{\gamma} - ((X'X))^{-1}A'[A((X'X))^{-1}A']^{-1}A\hat{\gamma} \quad (3.14)$$

Hay algunas observaciones interesantes que hacer sobre las ecuaciones (3.13) y (3.14). En primer lugar, el lado izquierdo de (3.13) es una proyección. Ello garantiza de manera automática que $\| \tilde{y} - X\hat{\gamma}_h \|^2$ es mínimo[1]. Además, el tratamiento anterior se generaliza de modo inmediato al caso de modelos de rango no completo, sin más que reemplazar en los lugares procedentes matrices inversas por las correspondientes inversas generalizadas.

En segundo lugar, dado que los estimadores mínimo cuadráticos ordinarios estiman insesgadamente los correspondientes parámetros, tomando valor medio en (3.14) vemos que:

$$E[\hat{\gamma}_h] = \vec{\gamma} - ((X'X))^{-1}A'[A((X'X))^{-1}A']^{-1}A\vec{\gamma}$$

lo que muestra que $\hat{\gamma}_h$ es un estimador insesgado de $\vec{\gamma}$ si $A\vec{\gamma} = \vec{0}$. Es decir, la insesgadez se mantiene si los parámetros *realmente* verifican las condiciones impuestas sobre los estimadores.

En tercer lugar, si definimos: $G = ((X'X))^{-1}A'[A((X'X))^{-1}A']^{-1}A$ tenemos que: $\hat{\gamma}_h = (I - G)\hat{\gamma}$. Por consiguiente,

$$\begin{aligned} \Sigma_{\hat{\gamma}_h} &= (I-G)\Sigma_{\hat{\gamma}}(I-G') \\ &= (I-G)\sigma^2((X'X))^{-1}(I-G') \\ &= \sigma^2[((X'X))^{-1} - G((X'X))^{-1} - ((X'X))^{-1}G' + G((X'X))^{-1}G'] \\ &= \sigma^2[((X'X))^{-1} - G((X'X))^{-1}G'] \end{aligned}$$

que muestra, dado que el segundo sumando tiene claramente elementos no negativos en su diagonal principal ($((X'X))^{-1}$ es definida no negativa), que $\Sigma_{\hat{\gamma}_h}$ tiene en la diagonal principal varianzas no mayores que las correspondientes en $\Sigma_{\hat{\gamma}}$. Podemos concluir, pues, que *la imposición de restricciones lineales sobre el vector de estimadores nunca incrementa su varianza*, aunque eventualmente, si las restricciones impuestas no son verificadas por los parametros a estimar, *puede introducir algún sesgo.*

Hemos razonado en las líneas anteriores sobre el modelo transformado. Podemos sustituir sin embargo (3.7) en (3.14) y obtener la expresión equivalente en términos de los parámetros originales:

$$\hat{\beta}_h = \hat{\beta} - ((X'X))^{-1}A'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - \vec{c}) \quad (3.15)$$

Retrospectivamente, podemos ver que el cambio de parámetros introducido tenía por único objeto transformar las restricciones en homogéneas, de forma que definieran un subespacio de $M$ sobre el que proyectar. Las condiciones originales no definían de modo inmediato dicho subespacio; al transformarlas en las que aparecen en (3.8) hemos visto que $X\hat{\gamma}$ debe estar en el núcleo de una cierta aplicación lineal —que automáticamente define un espacio vectorial sobre el que proyectar—.

[1] Si hubiéramos llegado al mismo resultado minimizando una suma de cuadrados por el procedimiento habitual (derivando un lagrangiano) tendríamos aún que mostrar que el punto estacionario encontrado es un mínimo y no un máximo.

**R: Ejemplo 3.1** *(estimación condicionada)*

No hay en S-PLUS ni en R una función de propósito general para realizar estimación condicionada. La extensibilidad del lenguaje hace sin embargo extraordinariamente fácil el definirla. El fragmento a continuación ilustra el modo de hacerlo y como utilizarla. No se ha buscado la eficiencia ni elegancia sino la correspondencia más directa con la teoría expuesta más arriba.

```
--- Obtenido mediante R BATCH demo3.R
> #
> #  Definimos una función para uso posterior
> #
> lscond <- function(X,y,A,d,beta0=TRUE) {
+
+   ajuste  <- lsfit(X,y,intercept=beta0)
+   betas   <- ajuste$coefficients
+   xxinv   <- solve(t(X) %*% X)
+   axxa    <- solve(A %*% xxinv %*% t(A))
+   betas.h <- betas - xxinv %*% t(A) %*% axxa %*% (A %*% betas - d)
+   betas.h <- as.vector(betas.h)
+   names(betas.h) <- names(ajuste$coefficients)
+   return(list(betas=betas,betas.h=betas.h,ajuste.inc=ajuste))
+ }
> #
> #  Generamos los datos y realizamos la estimación
> #  aplicando la teoría de modo más directo.
> #
> X <- matrix(c(1,1,1,1,1,1,1,4,12,1,4,
+               13,0,6,7,0,2,2),6,3)    # matriz de diseño
> X
     [,1] [,2] [,3]
[1,]    1    1    0
[2,]    1    4    6
[3,]    1   12    7
[4,]    1    1    0
[5,]    1    4    2
[6,]    1   13    2
> beta  <- c(2,3,4)                     # parámetros
> y <- X %*% beta + rnorm(6)            # variable respuesta
> #
> #  Especificamos la restricción beta1 = beta2 así:
> #
> A <- matrix(c(0,1,-1),1,3,byrow=TRUE)
> d <- 0
> #
> #  Estimación condicionada
> #
> resultado <- lscond(X,y,A=A,d=d,beta0=FALSE)
> #
> resultado$betas.h       # betas.h verifican la restricción
      X1       X2       X3
1.844384 3.321415 3.321415
> resultado$betas         # betas incondicionados
      X1       X2       X3
1.800731 3.060488 3.874023
```

## CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

**3.1** Sea un espacio vectorial $M$ cualquiera, de dimensión finita. Compruébese que *siempre* existe una matriz $C$ tal que $M = K(C)$. (Ayuda: considérese una matriz cuyas filas fueran una base de $M^{\perp}$).

**3.2** (↑ 3.1) Pruébese la igualdad (3.5).

**3.3** Justifíquese el paso de (3.13) a (3.14).

**3.4**  Las restricciones que hemos discutido en la Sección 3.2 son exactas. Los parámetros las verifican de modo exacto. En ocasiones se recurre a restricciones estocásticas, llevando a los parámetros a verificarlas de forma *aproximada*. Es muy fácil introducirlas. Recordemos que, al hacer estimación mínimo-cuadrática, los parámetros se fijan de modo que la suma de cuadrados de los residuos sea la mínima posible. Si tenemos restricciones $A\vec{\beta} = \vec{c}$ que queremos imponer de modo aproximado basta que añadamos las filas de $A$ a la matriz $X$ y los elementos correspondientes de $\vec{c}$ al vector $\vec{y}$ para obtener:

$$\begin{pmatrix} \vec{y} \\ \vec{c} \end{pmatrix} = \begin{pmatrix} X \\ A \end{pmatrix} \vec{\beta} + \vec{\epsilon}$$

y hagamos mínimos cuadrados ordinarios con la muestra ampliada (las filas añadidas se denominan en ocasiones *pseudo-observaciones*). La idea es que las filas añadidas funcionan como observaciones y, por tanto, el procedimiento de estimación tenderá a hacer $A\hat{\beta} \approx \vec{c}$ (para que los residuos correspondientes $\vec{c} - A\hat{\beta}$ sean "pequeños"). Aún más: podemos graduar la importancia que damos a las pseudo-observaciones (y por tanto el nivel de aproximación con que deseamos imponer las restricciones estocásticas): basta que las multipliquemos por una constante adecuada $k$ para estimar

$$\begin{pmatrix} \vec{y} \\ k\vec{c} \end{pmatrix} = \begin{pmatrix} X \\ kA \end{pmatrix} \vec{\beta} + \vec{\epsilon}.$$

Obsérvese que ahora los residuos de las pseudo-observaciones serán $k(\vec{c} - A\hat{\beta})$ y si tomamos $k$ elevado el método mínimo cuadrático tendrá que prestar atención preferente a que $A\hat{\beta} \approx \vec{c}$ se verifique con gran aproximación (porque los cuadrados de los residuos correspondientes entran en $SE$ afectados de un coeficiente $k^2$). Cuando $k \to \infty$ nos acercamos al efecto de restricciones exactas.

**3.5** (↑ 3.4)  Un caso particular de interés se presenta cuando en el problema anterior se toma $A = I$ y $\vec{c} = \vec{0}$. Se dice entonces que estamos ante el estimador *ridge* de parámetro $k$. En 8.2 abordamos su estudio y justificación con detalle.

# Capítulo 4

# Regresión con perturbaciones normales.

## 4.1. Introducción.

Si a los supuestos habituales (Sección 1.3, pág. 6) añadimos[1] el de que $\vec{\epsilon} \sim N(\vec{0}, \sigma^2 I)$, todos los resultados anteriores se mantienen; obtendremos no obstante muchos adicionales, relativos a la distribución de diferentes estadísticos. Podremos también efectuar contrastes de hipótesis diversas. Buena parte de estos resultados son consecuencia casi inmediata de alguno de los siguientes lemas.

**Lema 4.1** *Si $\vec{u} \sim N(\vec{0}, \sigma^2 I)$ y $A$ es una matriz simétrica idempotente de orden $n$ y rango $r$, entonces: $\frac{\vec{u}\,' A \vec{u}}{\sigma^2} \sim \chi^2_r$.*

DEMOSTRACION:

Sea $D$ la matriz diagonalizadora de $A$. Siendo $A$ simétrica, $D$ es una matriz ortogonal cuyas columnas son vectores propios de $A$, verificándose: $D'AD = \Lambda$, en que $\Lambda$ es una matriz en cuya diagonal principal aparecen los valores propios de $A$. Como $A$ es idempotente, $\Lambda$ es de la forma

$$\Lambda = \begin{array}{c} \begin{matrix} r & (n-r) \end{matrix} \\ \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \end{array},$$

en que $I$ es una matriz unidad de rango $r$, y los bloques de ceros que la circundan son de órdenes adecuados para completar una matriz cuadrada de orden $n \times n$.

[1] El símbolo $\sim$ denotará en lo sucesivo que el lado izquierdo es una variable aleatoria con la distribución que especifica el lado derecho.

Si hacemos el cambio de variable $\vec{v} = D'\vec{u}$ $(\Rightarrow \vec{u} = D\vec{v})$, el nuevo vector $\vec{v}$ sigue también una distribución $N(\vec{0}, \sigma^2 I)$. Entonces,

$$\frac{\vec{u}\,' A\vec{u}}{\sigma^2} = \frac{\vec{v}\,' D'AD\vec{v}}{\sigma^2} = \frac{\vec{v}\,'}{\sigma} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \frac{\vec{v}}{\sigma} = \sum_{i=1}^{r} \frac{v_i^2}{\sigma^2}. \tag{4.1}$$

Pero el lado derecho de (4.1) es una suma de cuadrados de $r$ variables aleatorias $N(0,1)$ independientes, y por tanto sigue una distribución[2] $\chi^2_r$.

**Lema 4.2** *Sea $B$ una matriz simétrica $n \times n$ y $P$ una matriz simétrica idempotente del mismo orden y rango $r$. Sea $\vec{u}$ un vector aleatorio $n$-variante, $\vec{u} \sim N(\vec{0}, \sigma^2 I)$, y supongamos que se verifica $BP = 0$. Entonces, $\vec{u}\,' B\vec{u}$ y $\vec{u}\,' P\vec{u}$ son variables aleatorias independientes.*

DEMOSTRACION:

Sea $D$ la matriz diagonalizadora de $P$. Al igual que antes, definamos $\vec{v} = D'\vec{u}$, (lo que implica $\vec{u} = D\vec{v}$). Tenemos que:

$$BP = 0 \quad \Rightarrow \quad D'BDD'PD = 0 \tag{4.2}$$

(4.3)

$$\Rightarrow \quad D'BD \begin{array}{c} \\ \\ \end{array} \begin{array}{c} \begin{array}{cc} r & (n-r) \end{array} \\ \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \end{array} = 0 \tag{4.4}$$

(4.5)

$$\Rightarrow \quad D'BD \text{ tiene sus } r \text{ primeras columnas nulas} \tag{4.6}$$

Por tanto:

$$D'BD = \begin{array}{c} \\ r \\ (n-r) \end{array} \begin{array}{c} \begin{array}{cc} r & (n-r) \end{array} \\ \begin{pmatrix} 0 & L_{12} \\ 0 & L_{22} \end{pmatrix} \end{array} = 0 \tag{4.7}$$

Como, además, $D'BD$ es simétrica, $L_{12}$ ha de ser también un bloque de ceros, y:

$$\vec{u}\,' B\vec{u} = \vec{v}\,' D'BD\vec{v} = \vec{v}\,' \begin{array}{c} \begin{array}{cc} r & (n-r) \end{array} \\ \begin{pmatrix} 0 & 0 \\ 0 & L_{22} \end{pmatrix} \end{array} \vec{v} \tag{4.8}$$

Por otra parte:

$$\vec{u}\,' P\vec{u} = \vec{v}\,' D'PD\vec{v} = \vec{v}\,' \begin{array}{c} \begin{array}{cc} r & (n-r) \end{array} \\ \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \end{array} \vec{v} \tag{4.9}$$

De (4.8) y (4.9) se deduce que ambas formas cuadráticas consideradas dependen de distintas componentes del vector $\vec{v}$, y son por tanto independientes.

**Lema 4.3** *Sea $M$ una matriz simétrica idempotente de rango $r$ y dimensiones $n \times n$. Sea $A$ una matriz que verifica $AM = 0$, y $\vec{u} \sim N(\vec{0}\,, \sigma^2 I)$. Entonces $A\vec{u}$ y $\vec{u}\,' M\vec{u}$ son variables aleatorias independientes.*

[2] El recíproco es también cierto; véase en Searle (1971), Teorema 2, pag. 57 una versión más potente de este teorema.

DEMOSTRACION:

Sea $D$ la matriz que diagonaliza $M$. Al igual que antes, definamos $\vec{v} = D'\vec{u}$ $(\Rightarrow \vec{u} = D\vec{v})$. Como $AM = 0$, y $D'MD$ es una matriz diagonal con $r$ unos y $(n-r)$ ceros en la diagonal principal, se verifica que

$$AM = ADD'MD = 0 \Rightarrow AD = \begin{array}{c} \quad r \qquad (n-r) \\ \big(0 \quad | \quad L_2 \quad \big) \end{array}, \tag{4.10}$$

es decir, $AD$ tiene sus primeras $r$ columnas de ceros. Por consiguiente,

$$A\vec{u} = AD\vec{v} = \begin{array}{c} \quad r \qquad (n-r) \\ \big(0 \quad | \quad L_2 \quad \big) \end{array}\vec{v}. \tag{4.11}$$

Como

$$\vec{u}\,'M\vec{u} = \vec{v}\,'D'MD\vec{v} = \vec{v}\,' \begin{array}{c} \quad r \quad (n-r) \\ \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \end{array}\vec{v}, \tag{4.12}$$

deducimos de (4.11) y (4.12) que ambas variables aleatorias consideradas dependen de distintas componentes de $\vec{v}$, y son consecuentemente independientes.

Podemos ahora, con ayuda de los Lemas precedentes, demostrar el siguiente resultado:

**Teorema 4.1** *Si $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$ , $\vec{\epsilon} \sim N(\vec{0}, \sigma^2 I)$, y $X$ es de orden $N \times p$ y rango $p$, se verifica:*

1. $\hat{\beta} \sim N(\vec{\beta}, \sigma^2((X\,'X))^{-1})$
2. $(\hat{\beta} - \vec{\beta})'(X\,'X)(\hat{\beta} - \vec{\beta}) \sim \sigma^2\chi^2_p$
3. $(N-p)\hat{\sigma}^2 = SSE \sim \sigma^2\chi^2_{N-p}$
4. *$\hat{\beta}$ y $\hat{\sigma}^2$ son variables aleatorias independientes.*

DEMOSTRACION:

El apartado 1) es inmediato; si se verifican los supuestos habituales, fue ya demostrado (Teorema 2.2, pág. 18) que $\hat{\beta}$ es un estimador insesgado de $\vec{\beta}$ con la matriz de covarianzas indicada. Como, además, $\hat{\beta}$ es una combinación lineal de variables aleatorias normales, es también normal.

El apartado 2) es consecuencia inmediata del Lema 4.1. Para demostrar el apartado 3) observemos que

$$\begin{aligned}
\frac{SSE}{\sigma^2} &= \frac{(\vec{Y} - X\hat{\beta})'(\vec{Y} - X\hat{\beta})}{\sigma^2} && (4.13)\\
&= \frac{(\vec{Y} - X(X'X)^{-1}X'\vec{Y})'(\vec{Y} - X(X'X)^{-1}X'\vec{Y})}{\sigma^2} && (4.14)\\
&= \frac{\vec{Y}'[I - X(X'X)^{-1}X']\vec{Y}}{\sigma^2} && (4.15)\\
&= \frac{(X\vec{\beta} + \vec{\epsilon})'[I - X(X'X)^{-1}X'](X\vec{\beta} + \vec{\epsilon})}{\sigma^2} && (4.16)\\
&= \frac{\vec{\epsilon}'[I - X(X'X)^{-1}X']\vec{\epsilon}}{\sigma^2} && (4.17)\\
&= \frac{\vec{\epsilon}'M\vec{\epsilon}}{\sigma^2} && (4.18)\\
&\sim \chi^2_{N-p}, && (4.19)
\end{aligned}$$

donde (4.19) es consecuencia inmediata del Lema 4.1, ya que $M$ es simétrica idempotente y de rango $N - p$.

Para probar 4) basta invocar el Lema 4.3, ya que

$$\begin{aligned}
\hat{\beta} &= ((X'X))^{-1}X'\vec{Y}\,, && (4.20)\\
\hat{\sigma}^2 &= \frac{SSE}{N-p} = \frac{\vec{Y}'[I - X(X'X)^{-1}X']\vec{Y}}{N-p}. && (4.21)
\end{aligned}$$

De la ecuación (4.20) deducimos (sustituyendo $\vec{Y}$ por $X\vec{\beta}+\vec{\epsilon}$) que $\hat{\beta} = \vec{\beta}+X(X'X)^{-1}X'\vec{\epsilon}$. La misma sustitución en (4.21) muestra que

$$\hat{\sigma}^2 = \frac{\vec{\epsilon}'[I - X(X'X)^{-1}X']\vec{\epsilon}}{N-p}.$$

Como

$$((X'X))^{-1}X'[I - X(X'X)^{-1}X'] = 0.$$

el Lema 4.3 demuestra la independencia de las formas lineal y cuadrática anteriores y por tanto de (4.20) y (4.21).

**R: Ejemplo 4.1** *(ejemplo de simulación)*

El código que sigue tiene por objeto ilustrar cómo examinaríamos empíricamente la concordancia entre lo que la teoría predice y lo que podemos obtener en la práctica. Lo que se hace es generar múltiples muestras artificiales, obtener de ellas múltiples observaciones del estadístico de interés (aquí, $\hat{\beta}$) y examinar el ajuste de la distribución empírica de los mismos a la teórica.

```
--- Obtenido mediante R BATCH demo4.R
> #
> #  La idea es generar múltiples instancias del mismo problema
> #  de regresión (con la misma X y los mismos betas) muestreando
> #  en cada ocasión unas perturbaciones diferentes. Obtenemos
> #  así múltiples estimaciones de los betas, cuya distribución
```

```
> #  debería adecuarse a la que predice la teoría.
> #
> X <- matrix(c(1,1,1,1,1,1,9,4,12,1,4,
+               13,0,6,7,0,2,2),6,3)    # matriz de diseño
> X
     [,1] [,2] [,3]
[1,]    1    9    0
[2,]    1    4    6
[3,]    1   12    7
[4,]    1    1    0
[5,]    1    4    2
[6,]    1   13    2
> beta  <- c(2,3,4)                     # parámetros
> Ey <- X %*% beta                      # E(variable respuesta)
> #
> #  Hasta la línea anterior hay cálculos que solo se requiere
> #  realizar una vez. Vamos ahora a generar 100 muestras artificiales
> #  del vector Y y a estimar los betas para cada una de ellas. Nos
> #  servimos de for() {  } para especificar un conjunto de
> #  instrucciones que ha de ser repetido muchas veces.
> #
> muestras <- 100
> b <- matrix(0,muestras,3)             # matriz para guardar resultados
> for (i in 1:muestras) {
+   y     <- Ey + rnorm(6)              # y = X %*% beta + epsilon
+   fit   <- lsfit(X,y,intercept=FALSE)
+   b[i,] <- fit$coefficients           # guardamos los betas de la
+                                       # i-esima iteración en la
+                                       # i-esima fila de b
+ }
> #
> #  La distribución teórica de los betas es Normal, con vector de
> #  medias (2,3,4) y matriz de covarianzas inversa(X'X) (la
> #  varianza de las perturbaciones generadas por rnorm() es 1).
> #
> cov.betas <- solve(t(X) %*% X)
> #
> #  Tomemos, por ejemplo, el primer beta. Los valores estimados
> #  en las 100 replicaciones del experimento están en la primera
> #  columna de la matriz b. Tipificándolas,
> #
> beta1.tipif <- (b[,1] - beta[1]) / sqrt(cov.betas[1,1])
> #
> #  obtendremos 100 observaciones procedentes de una N(0,1).
> #  Para comprobar la adecuación de lo obtenido a la teoría,
> #  podemos calcular los momentos...
> #
> mean(beta1.tipif)                  # razonablemente cerca de 0
[1] -0.001101161
> var(beta1.tipif)                   # razonablemente cerca de 1
[1] 0.8326672
> #
> #  dibujar el histograma...
> #
```

```
> hist(beta1.tipif)
> #
> #  o llevar a cabo algún contraste especializado:
> #
> library(ctest)
Warning message:
package 'ctest' has been merged into 'stats'
> ks.test(beta1.tipif,"pnorm")       # Kolmogorov-Smirnov,

One-sample Kolmogorov-Smirnov test

data:  beta1.tipif
D = 0.0727, p-value = 0.6654
alternative hypothesis: two.sided

>                                    # 1 población.
> shapiro.test(beta1.tipif)          # Shapiro-Wilk

Shapiro-Wilk normality test

data:  beta1.tipif
W = 0.9928, p-value = 0.8774

> #
> #  Vemos que el ajuste a la distribución teórica es bueno (sería
> #  aún mejor si el número de muestras tomadas fuera >> 100).
> #
> rm(beta1.tipif,cov.betas,b,fit,X,beta,Ey,muestras)
```

Lo que antecede ilustra, reducido a sus rasgos esenciales, el llamado método de Monte-Carlo. Puede parecer un ejercicio ocioso en el caso que nos ocupa (ya "sabíamos" cómo se distribuye $\hat{\beta}$ ¿a que viene comprobarlo mediante una simulación?). Sin embargo, tiene una enorme aplicación práctica por varias razones:

1. En ocasiones no conocemos la distribución teórica de los estadísticos de interés para muestras finitas. Todo lo que podemos obtener teóricamente es la distribución asintótica (la distribución cuando el tamaño muestral tiende a infinito). En este caso, la simulación proporciona un método para ver si para un cierto tamaño muestral la aproximación asintótica es aceptable.

2. En otras ocasiones, ni siquiera la distribución asintótica es obtenible analíticamente. Este es el caso más frecuente en la práctica. De nuevo el método de Monte-Carlo proporciona un método para obtener aproximaciones a la distribución de cualquier estadístico.

El uso del método de Monte-Carlo reposa en la posibilidad de generar mediante un ordenador números aleatorios con la distribución que deseemos. En este ejemplo, se ha empleado `rnorm` para generar variables aleatorias normales. (Sobre generadores de números aleatorios puede consultarse Knuth (1968), Kennedy (1980) y, en general, cualquier texto sobre computación estadística; R y S-PLUS ofrecen generadores de números aleatorios de las distribuciones más usuales, como casi cualquier otro paquete estadístico.)

## 4.2. Contraste de hipótesis lineales.

El problema que nos planteamos es el siguiente: dado el modelo lineal $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$ con los supuestos habituales más normalidad, queremos, con ayuda de una muestra, contrastar la siguiente hipótesis lineal:

$$h\colon A\vec{\beta} = \vec{c} \qquad (\text{rango de } A = q < p) \tag{4.22}$$

En la forma (4.22) se puede escribir cualquier hipótesis lineal sobre los parámetros. En particular, mediante adecuada elección de $A$ se pueden hacer contrastes de nulidad de uno o varios parámetros, de igualdad de dos o más de ellos, etc.

> **Observación 4.1** Llamamos hipótesis lineales a las que pueden expresarse del modo (4.22); multitud de hipótesis de interés admiten tal expresión, como se verá en lo que sigue. Hay hipótesis, sin embargo, que no pueden escribirse de tal forma. Por ejemplo, restricciones de no negatividad sobre los parámetros ($\beta_i > 0$) o sobre el módulo de $\vec{\beta}$ (cosas como $\beta_1^2 + \beta_2^2 = 1$).

La forma de efectuar el contraste es la habitual. Se busca un estadístico que bajo la hipótesis nula $h$ siga una distribución conocida; si el valor obtenido en el muestreo de dicho estadístico es "raro" de acuerdo con lo esperable cuando $h$ es cierta, rechazaremos la hipótesis nula. El estadístico de contraste y su distribución se deducen del siguiente teorema:

**Teorema 4.2** *Sea $h\colon A\vec{\beta} = \vec{c}$ una hipótesis lineal, $\hat{\beta}_h$ el vector de estimadores mínimo cuadráticos condicionados por $h$, y $SSE_h = \| \vec{Y} - X\hat{\beta}_h \|^2$ Bajo los supuestos habituales más el de normalidad en las perturbaciones, se verifica:*

1. $SSE_h - SSE = (A\hat{\beta} - \vec{c}\,)'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - \vec{c}\,)$
2. *Si $h\colon A\vec{\beta} = \vec{c}$ es cierta,*

$$Q_h = \frac{(SSE_h - SSE)/q}{SSE/(N-p)} \sim \mathcal{F}_{q,N-p}$$

*en que $q \leq p$ es el rango de $A$.*

DEMOSTRACION:

$$\begin{aligned}
SSE_h - SSE &= \| \vec{Y} - X\hat{\beta}_h \|^2 - \| \vec{Y} - X\hat{\beta} \|^2 && (4.23)\\
&= \| \vec{Y} - X\hat{\beta} + X\hat{\beta} - X\hat{\beta}_h \|^2 - \| \vec{Y} - X\hat{\beta} \|^2 && (4.24)\\
&= \| \vec{Y} - X\hat{\beta} \|^2 + \| X\hat{\beta} - X\hat{\beta}_h \|^2 - \| \vec{Y} - X\hat{\beta} \|^2 \\
&\quad +2 < (\vec{Y} - X\hat{\beta}\,), (X\hat{\beta} - X\hat{\beta}_h) > && (4.25)\\
&= \| X\hat{\beta} - X\hat{\beta}_h \|^2 && (4.26)\\
&= (\hat{\beta} - \hat{\beta}_h)'(X'X)(\hat{\beta} - \hat{\beta}_h) && (4.27)
\end{aligned}$$

Se ha hecho uso en el paso de (4.25) a (4.26) de que $\hat{\epsilon}$ es ortogonal a toda combinación lineal de las columnas de $X$, lo que garantiza la nulidad del producto interno en (4.25).

Haciendo uso de la ecuación (3.15), pág. 36, la expresión (4.27) se convierte en:

$$\begin{aligned} SSE_h &- SSE = \\ &= (A\hat{\beta} - \vec{c}\,)'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - \vec{c}\,) \qquad (4.28) \\ &\stackrel{h}{=} \vec{\epsilon}\,' \underbrace{X((X'X))^{-1}A'[A(X'X)^{-1}A']^{-1}A((X'X))^{-1}X'}_{G} \vec{\epsilon} \qquad (4.29) \end{aligned}$$

La igualdad del primer miembro con la expresión (4.28) finaliza la demostración de 1). La expresión (4.29) en la línea siguiente (válida sólo bajo el supuesto de que $h$ se cumple) es de utilidad, porque muestra que $SSE_h - SSE$ es una forma cuadrática en variables normales (las $\epsilon$) de matriz $G$ que fácilmente comprobamos es idempotente. Tenemos por otra parte (Teorema 4.1) que:

$$SSE = \vec{Y}\,'(I - P_M)\vec{Y} \sim \sigma^2\chi^2_{N-p} \qquad (4.30)$$

Por otra parte, $SSE_h - SSE$ ya se ha visto que sigue distribución $\chi^2$ con grados de libertad que coincidirán con el rango de $G$ (= rango($A$)).

Para demostrar que $Q_h$ en el enunciado es una variable aleatoria con distribución $\mathcal{F}$ de Snedecor, basta comprobar que numerador y denominador son independientes: pero ésto es inmediato, ya que

$$(I - P_M)X((X'X))^{-1}A'[A(X'X)^{-1}A']^{-1}A((X'X))^{-1}X' = 0;$$

el Lema 4.2 garantiza por tanto la independencia.

**Observación 4.2** Hay cuestiones de interés sobre el Teorema 4.2. En primer lugar, es claro que, para un nivel de significación $\alpha$, la región crítica estará formada por valores mayores que $\mathcal{F}^{\alpha}_{q,N-p}$. En efecto, son grandes discrepancias entre $SSE_h$ y $SSE$ las que cabe considerar evidencia contra $h$. Desde otro punto de vista, el apartado 1) del Teorema 4.2 muestra que el estadístico tiene en su numerador una forma cuadrática que crece al separarse $A\hat{\beta}$ de $\vec{c}$.

**Observación 4.3** La presentación es puramente heurística; se ha propuesto el estadístico $Q_h$ y encontrado su distribución, indicándose, sin otro apoyo que el sentido común, qué valores debemos considerar en la región crítica. Podríamos llegar a un resultado análogo si construyéramos un estadístico de contraste basado en la razón generalizada de verosimilitudes:

$$\Lambda = \frac{\text{máx}_{\hat{\beta}}\, g(\hat{\beta}; \vec{y}, X)}{\text{máx}_{\hat{\beta}_h}\, g(\hat{\beta}_h; \vec{y}, X)}$$

siendo $\hat{\beta}_h$ aquellos $\hat{\beta}$ verificando $h\colon A\hat{\beta} = \vec{c}$. Ello proporciona una justificación al estadístico anterior[3].

**Observación 4.4** Del enunciado del teorema anterior se sigue con facilidad que cuando $h$ no es cierta (y en consecuencia $A\vec{\beta} - \vec{c} = \vec{d} \neq \vec{0}$, $Q_h$ sigue una distribución $\mathcal{F}$ de Snedecor no central, con parámetro de no centralidad $\delta^2 = \vec{t}\,'\vec{t}$ (véase Apéndice B.1), siendo

$$\vec{t} = [A((X'X))^{-1}A']^{-\frac{1}{2}}(A\vec{\beta} - \vec{c}).$$

Ello permite calcular fácilmente la potencia de cualquier contraste frente a alternativas prefijadas, si se dispone de tablas o ábacos de la $\mathcal{F}$ de Snedecor no central. En R se dispone de la función `pf` que admite un parámetro de no centralidad. Alternativamente, puede estimarse la potencia por simulación.

[3] Cf. Cox and Hinkley (1974) p. 313 y ss.

**R: Ejemplo 4.2** *(contraste de una hipótesis lineal)*

```
--- Obtenido mediante R BATCH demo5.R
> #
> #  Esta función ya fue definida; se reproduce para
> #  comodidad de referencia
> #
> lscond <- function(X,y,A,d,beta0=TRUE) {
+
+   ajuste  <- lsfit(X,y,intercept=beta0)
+   betas   <- ajuste$coefficients
+   xxinv   <- solve(t(X) %*% X)
+   axxa    <- solve(A %*% xxinv %*% t(A))
+   betas.h <- betas -
+       xxinv %*% t(A) %*% axxa %*% (A %*% betas - d)
+   betas.h <- as.vector(betas.h)
+   names(betas.h) <- names(ajuste$coefficients)
+   return(list(betas=betas,betas.h=betas.h,
+               ajuste.inc=ajuste))
+ }
> #
> #  Definimos ahora otra cuyo objeto es contrastar la hipótesis
> #  la hipótesis lineal h: A beta = c.
> #
> contraste.h <- function(X,y,A,d,beta0=TRUE) {
+   lscond.result <- lscond(X,y,A,d,beta0=beta0)
+   betas   <- lscond.result$betas
+   betas.h <- lscond.result$betas.h
+   SSE     <- sum((y - X %*% betas)^2)
+   SSE.h   <- sum((y - X %*% betas.h)^2)
+   numer   <- (SSE.h - SSE)/nrow(A)    # supone A rango completo
+   denom   <- SSE/(nrow(X) - ncol(X))
+   Qh  <- numer / denom
+   p.value <- 1 - pf(Qh,nrow(A),       # p-value, valor en la cola.
+                 nrow(X)-ncol(X))
+   return(list(Qh=Qh,p.value=p.value))
+ }
> #
> #  Generamos los datos
> #
> X <- matrix(c(1,1,1,1,1,1,1,4,12,1,4,
+               13,0,6,7,0,2,2),6,3)    # matriz de diseño
> X
     [,1] [,2] [,3]
[1,]    1    1    0
[2,]    1    4    6
[3,]    1   12    7
[4,]    1    1    0
[5,]    1    4    2
[6,]    1   13    2
> beta  <- c(2,3,4)                     # parámetros
> y <- X %*% beta + rnorm(6)            # variable respuesta
> #
> #  Especificamos la restricción beta1 = beta2 así:
> #
```

```
> A <- matrix(c(0,1,-1),1,3,byrow=TRUE)
> d <- 0
> #
> #  Estimación condicionada
> #
> result <- contraste.h(X,y,A=A,d=d,beta0=FALSE)
> #
> result$p.value            # Si menor que "alfa", rechazamos
[1] 0.1035325
>                           # al nivel de significación "alfa".
> rm(result,X,y,beta,A,d)
```

### 4.2.1. Contraste sobre coeficientes $\beta_i$ aislados.

El Teorema 4.2 permite obtener como casos particulares multitud de contrastes frecuentemente utilizados. Por ejemplo, la hipótesis $h\colon \beta_i = 0$ puede contrastarse tomando $\vec{c} = \vec{0}$ y $A = \begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \end{pmatrix}$, ocupando el único "uno" la posición $i$-ésima. En tal caso, $Q_h$ puede escribirse así:

$$Q_h = \frac{(\hat{\beta}_i - 0)'[(X'X)^{-1}_{ii}]^{-1}(\hat{\beta}_i - 0)}{\hat{\sigma}^2} \tag{4.31}$$

donde $(X'X)^{-1}_{ii} = [A((X'X))^{-1}A']$ designa el elemento en la posición $i$-ésima de la diagonal principal de $((X'X))^{-1}$. Bajo la hipótesis $h$, (4.31) sigue una distribución $\mathcal{F}_{1,N-p}$, y como $\hat{\sigma}^2(X'X)^{-1}_{ii} = \hat{\sigma}^2_{\hat{\beta}_i}$ tenemos que:

$$\sqrt{Q_h} = \frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}} \sim \sqrt{\mathcal{F}_{1,N-p}} \sim t_{N-p} \tag{4.32}$$

La regla de decisión que se deduce de (4.32) es:

> "Rechazar $h\colon \beta_i = 0$ al nivel de significación $\alpha$ si $|\hat{\beta}_i/\hat{\sigma}_{\hat{\beta}_i}| > t^{\alpha/2}(N-p)$."

El estadístico $|\hat{\beta}_i/\hat{\sigma}_{\hat{\beta}_i}|$ recibe el nombre de *estadístico t* o *t-ratio*. De forma análoga se contrasta la hipótesis $h\colon \beta_i = c$.

### 4.2.2. Contraste de significación conjunta de la regresión.

Otra hipótesis frecuentemente de interés es: $h\colon \beta_1 = \cdots = \beta_{p-1} = 0$ —es decir, nulidad de todos los parámetros, salvo el correspondiente a la columna de "unos", $\beta_0$—. En este caso,

$$SSE_h = \sum_{i=1}^{N} (Y_i - \overline{Y})^2$$

y la hipótesis $h$ puede expresarse en la forma $A\vec{\beta} = \vec{c}$ siendo:

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix} = \left(\vec{0} \mid I\right)$$

una matriz con $(p-1)$ filas y $p$ columnas, y:

$$\vec{c}' = \begin{pmatrix} 0 & 0 & \cdots & 0 \end{pmatrix}$$

Pero $SSE_h$ en este caso particular es lo que hemos definido (Teorema 2.3, pág. 21) como $SST$. Por tanto,

$$\begin{aligned} Q_h &= \frac{(SST - SSE)/(p-1)}{SSE/(N-p)} \\ &= \frac{N-p}{p-1} \times \frac{(SST - SSE)}{SSE} \\ &= \frac{N-p}{p-1} \times \frac{R^2}{(1-R^2)} \end{aligned}$$

siendo $R$ el coeficiente de correlación múltiple definido en el Teorema 2.3, pág. 21. El contraste de $h$ requiere solamente conocer $R^2$. Cuando $h$ es cierta, $Q_h$ se distribuye como una $\mathcal{F}_{p-1,N-p}$.

**R: Ejemplo 4.3** *(use de la función `lm` y contraste de hipótesis)*

La función `lsfit` proporciona esencialmente todo lo necesario, pero no en la forma más cómoda ni directamente utilizable. En general, preferiremos utilizar la función `lm` que permite especificar los modelos de regresión de forma simbólica, con variables nombradas, y proporciona una salida más completa que `lsfit`.

El fragmento que sigue ilustra su funcionamiento sobre un conjunto de datos que utilizaremos repetidamente en lo que sigue. Entre otras cosas, es de interés notar que para contrastar hipótesis de nulidad de grupos de variables, todo lo que necesitamos son las respectivas sumas de cuadrados de los residuos para calcular $Q_h$, que podemos obtener de sendas regresiones con y sin las variables objeto del contraste.

```
--- Obtenido mediante R BATCH demo6.R
> UScrime[1:3,1:5]                          #  Veamos los datos.
    M So  Ed Po1 Po2
1 151  1  91  58  56
2 143  0 113 103  95
3 142  1  89  45  44
> str(UScrime)                              #  Es una dataframe.
'data.frame': 47 obs. of  16 variables:
 $ M   : int  151 143 142 136 141 121 127 131 157 140 ...
 $ So  : int  1 0 1 0 0 0 1 1 1 0 ...
 $ Ed  : int  91 113 89 121 121 110 111 109 90 118 ...
 $ Po1 : int  58 103 45 149 109 118 82 115 65 71 ...
 $ Po2 : int  56 95 44 141 101 115 79 109 62 68 ...
 $ LF  : int  510 583 533 577 591 547 519 542 553 632 ...
 $ M.F : int  950 1012 969 994 985 964 982 969 955 1029 ...
 $ Pop : int  33 13 18 157 18 25 4 50 39 7 ...
 $ NW  : int  301 102 219 80 30 44 139 179 286 15 ...
 $ U1  : int  108 96 94 102 91 84 97 79 81 100 ...
 $ U2  : int  41 36 33 39 20 29 38 35 28 24 ...
 $ GDP : int  394 557 318 673 578 689 620 472 421 526 ...
 $ Ineq: int  261 194 250 167 174 126 168 206 239 174 ...
 $ Prob: num  0.0846 0.0296 0.0834 0.0158 0.0414 ...
```

```
 $ Time: num  26.2 25.3 24.3 29.9 21.3 ...
 $ y   : int  791 1635 578 1969 1234 682 963 1555 856 705 ...
> fit <- lm(y ~ Ineq + Prob + Time,        #  Regresión lineal
+               data=UScrime)
> fit                         #  objeto compuesto

Call:
lm(formula = y ~ Ineq + Prob + Time, data = UScrime)

Coefficients:
(Intercept)         Ineq         Prob         Time
  1287.1147       0.4877   -8144.7085      -3.5002

> attributes(fit)             #  con mucha información;
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

$class
[1] "lm"

> summary(fit)                #  summary() proporciona resumen

Call:
lm(formula = y ~ Ineq + Prob + Time, data = UScrime)

Residuals:
    Min      1Q  Median      3Q     Max
-619.13 -213.34  -29.74  156.89 1048.73

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1287.1147   332.0345   3.876 0.000358 ***
Ineq            0.4877     1.6310   0.299 0.766371
Prob        -8144.7085  3163.9883  -2.574 0.013575 *
Time           -3.5002     9.0321  -0.388 0.700278
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 360.9 on 43 degrees of freedom
Multiple R-Squared: 0.186,Adjusted R-squared: 0.1292
F-statistic: 3.276 on 3 and 43 DF,  p-value: 0.02995

> #
> #  Obtenemos directamente los t-ratios y R2 y los
> #  niveles de significación, lo que permite el contraste
> #  directo de hipótesis sobre parámetros aislados y sobre
> #  significación conjunta de la regresión.
> #
> #  Si quisiéramos efectuar contrastes de exclusión de variables,
> #  podemos hacerlo comparando sumas de cuadrados de dos regresiones.
> #  Por ejemplo, para contrastar nulidad de coeficientes de Ineq y
> #  Time en la regresión precedente, podríamos hacer lo siguiente:
```

```
> #
> fit.h <- lm(y ~ Prob, data=UScrime)
> SSE   <- sum(fit$residuals^2)
> SSE.h <- sum(fit.h$residuals^2)
> N <- nrow(UScrime)
> q <- 2              # Rango hipótesis contrastada
> p <- 4              # Número regresores modelo no restringido
> Qh <- ((SSE.h - SSE)/ q) / (SSE / (N-p)) # Estadístico Qh
> 1 - pf(Qh,q,N-p)                         # Nivel significación
[1] 0.9155674
```

## 4.3. Construcción de intervalos de confianza para la predicción.

Supongamos de nuevo que trabajamos sobre el modelo $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$ con los supuestos habituales más el de normalidad en las perturbaciones. Frecuentemente es de interés, además de la estimación de los parámetros, la utilización del modelo con finalidad predictiva.

Sea $\vec{x}_*$ un vector $p \times 1$ de valores a tomar por los regresores. La correspondiente $Y_*$ será: $Y_* = \vec{x}_*{}'\vec{\beta} + \epsilon_*$. Una predicción $\hat{Y}_*$ del valor a tomar por la $Y_*$ es: $\hat{Y}_* = \vec{x}_*{}'\hat{\beta}$.

**Teorema 4.3** *Se verifica lo siguiente:*

1. $E(Y_* - \hat{Y}_*) = 0$
2. $E(Y_* - \hat{Y}_*)^2 = \sigma^2(1 + \vec{x}_*{}'((X'X))^{-1}\vec{x}_*)$

DEMOSTRACION:

El apartado 1) se sigue inmediatamente de las ecuaciones (4.33) y (4.34) a continuación, consecuencia la primera de los supuestos habituales, y la segunda de la insesgadez de $\hat{\beta}$ (Teorema 2.2, pág. 18).

$$E(Y_*) = E(\vec{x}_*{}'\vec{\beta} + \epsilon_*) = \vec{x}_*{}'\vec{\beta} \quad (4.33)$$
$$E(\hat{Y}_*) = E(\vec{x}_*{}'\hat{\beta}) = \vec{x}_*{}'\vec{\beta} \quad (4.34)$$

Se dice que $\hat{Y}_*$ es una predicción *insesgada* de $Y_*$. Observemos que:

$$\begin{aligned}
E(Y_* - \hat{Y}_*)^2 &= E[\vec{x}_*{}'\vec{\beta} + \vec{\epsilon}_* - \vec{x}_*{}'\hat{\beta}\,]^2 && (4.35)\\
&= E[\vec{x}_*{}'(\vec{\beta} - \hat{\beta}) + \epsilon_*\,]^2 && (4.36)\\
&= E[\vec{x}_*{}'(\vec{\beta} - \hat{\beta})]^2 + E[\epsilon_*\,]^2 && (4.37)\\
&= E[\vec{x}_*{}'(\vec{\beta} - \hat{\beta})(\vec{\beta} - \hat{\beta})'\vec{x}_*] + E[\epsilon_*\,]^2 && (4.38)\\
&= \vec{x}_*{}'\Sigma_{\hat{\beta}}\vec{x}_* + \sigma^2 && (4.39)\\
&= \vec{x}_*{}'\sigma^2((X'X))^{-1}\vec{x}_* + \sigma^2 && (4.40)\\
&= \sigma^2[1 + \vec{x}_*{}'((X'X))^{-1}\vec{x}_*] && (4.41)
\end{aligned}$$

En el paso de (4.36) a (4.37) se ha hecho uso de la circunstancia de que $\hat{\beta}$ y $\epsilon_*$ son independientes ($\hat{\beta}$ depende solamente de $\vec{\epsilon}$, y $\epsilon_*$ es perturbación de una observación adicional, distinta de las que han servido para estimar $\hat{\beta}$ e independiente de ellas).

El examen de (4.41) muestra dos cosas. Una, que la varianza del error de predicción es *mayor o igual* que la varianza de la perturbación (ya que $\vec{x}_*\,'((X\,'X))^{-1}\vec{x}_*$ es una forma cuadrática semidefinida positiva). Esto es lógico: $\epsilon_*$ es del todo impredecible, y, *además*, la predicción $\hat{Y}_*$ incorpora una fuente adicional de error, al emplear $\hat{\beta}$ en lugar de $\vec{\beta}$.

Por otra parte, (4.41) muestra que la varianza del error de predicción *depende de* $\vec{x}_*\,'$. Habrá determinadas $Y_*$ cuya predicción será más precisa que la de otras. Más abajo volveremos sobre el particular.

## CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

**4.1** Demuéstrese que si $G$ es la matriz definida en (4.29) y $(X\,'X)$ es de rango completo, entonces rango($G$) = rango($A$).

# Capítulo 5

# Especificación inadecuada del modelo

## 5.1. Introducción.

En lo que antecede hemos dado por supuesto que el modelo lineal que se estima es el "correcto", es decir, que la variable aleatoria $\vec{Y}$ efectivamente se genera de la siguiente manera:

$$\vec{Y} = \beta_0 \vec{X}_0 + \beta_1 \vec{X}_1 + \ldots + \beta_{p-1} \vec{X}_{p-1} + \vec{\epsilon} \tag{5.1}$$

En la práctica, sin embargo, no tenemos un conocimiento preciso del mecanismo que genera las $Y$'s. Tenemos, todo lo más, una lista de variables susceptibles de formar parte de la ecuación (5.1) en condición de regresores.

De ordinario, por ello, incurriremos en errores en la especificación, que pueden ser de dos naturalezas:

1. Incluir en (5.1) regresores irrelevantes.
2. Omitir en (5.1) regresores que hubieran debido ser incluidos.

Estudiamos en lo que sigue el efecto de estos dos tipos de mala especificación..

## 5.2. Inclusión de regresores irrelevantes.

Supongamos que

$$\vec{Y} \quad = \quad X\vec{\beta} + \vec{\epsilon} \tag{5.2}$$

pese a lo cual decidimos estimar el modelo

$$\vec{Y} \quad = \quad X\vec{\beta} + Z\vec{\gamma} + \vec{\epsilon} \tag{5.3}$$

¿Qué ocurre con los estimadores de los parámetros $\vec{\beta}$?

Al estimar el modelo sobreparametrizado (5.3) obtendríamos:

$$\begin{pmatrix}\hat{\beta}\\ \hat{\gamma}\end{pmatrix} = \begin{pmatrix}X'X & X'Z\\ Z'X & Z'Z\end{pmatrix}^{-1}\begin{pmatrix}X'\\ Z'\end{pmatrix}\vec{Y} \tag{5.4}$$

De esta ecuación inmediatamente se deduce que, en el caso particular de columnas $Z$ ortogonales a las columnas en $X$, los estimadores de $\vec{\beta}$ proporcionados por (5.3) son idénticos a los que se obtendrían de (5.2). En efecto, si existe tal ortogonalidad, la matriz inversa en (5.4) es una matriz diagonal por bloques y $\hat{\beta} = (X'X)^{-1}X'\vec{Y}$.

Fuera de este caso particular, los estimadores de $\vec{\beta}$ procedentes de (5.4) son diferentes a los que se obtendría de estimar (5.2).

Sin embargo, (5.4) *proporciona estimadores insesgados*, sean cuales fueren los regresores irrelevantes añadidos[1]. En efecto, sustituyendo (5.2) en (5.4) tenemos:

$$\begin{pmatrix}\hat{\beta}\\ \hat{\gamma}\end{pmatrix} = \begin{pmatrix}X'X & X'Z\\ Z'X & Z'Z\end{pmatrix}^{-1}\begin{pmatrix}X'\\ Z'\end{pmatrix}\begin{bmatrix}X & Z\end{bmatrix}\begin{pmatrix}\vec{\beta}\\ \vec{0}\end{pmatrix} + \vec{\epsilon} \tag{5.5}$$

$$= \begin{pmatrix}\vec{\beta}\\ \vec{0}\end{pmatrix} + \begin{pmatrix}X'X & X'Z\\ Z'X & Z'Z\end{pmatrix}^{-1}\begin{pmatrix}X'\vec{\epsilon}\\ Z'\vec{\epsilon}\end{pmatrix} \tag{5.6}$$

y al tomar valor medio en la ecuación anterior obtenemos:

$$E[\hat{\beta}] = \vec{\beta} \tag{5.7}$$

$$E[\hat{\gamma}] = \vec{0} \tag{5.8}$$

De la misma ecuación (5.6) obtenemos que la matriz de covarianzas del vector $(\hat{\beta}'\ \hat{\gamma}')'$ es:

$$\Sigma = \sigma^2\begin{pmatrix}X'X & X'Z\\ Z'X & Z'Z\end{pmatrix}^{-1} \tag{5.9}$$

El bloque superior izquierdo de (5.9) es la matriz de covarianzas de los $\hat{\beta}$ obtenidos en el modelos sobreparametrizado. Debemos comparar dicho bloque con $\sigma^2(X'X)^{-1}$, matriz de covarianzas de los $\hat{\beta}$ obtenidos al estimar el modelo (5.2).

Haciendo uso del Teorema A.3, pág. 180, vemos que el bloque que nos interesa de (5.9) es:

$$((X'X))^{-1} + ((X'X))^{-1}X'Z[Z'Z - Z'X((X'X))^{-1}X'Z]^{-1}Z'X((X'X))^{-1} \tag{5.10}$$

Por simple inspección vemos que el segundo sumando de la expresión anterior es una matriz definida no negativa[2], y por tanto (5.10) tendrá en su diagonal principal elementos no menores que los de la diagonal principal de $(X'X)^{-1}$. En consecuencia,

[1] De los que lo único que supondremos es que no introducen combinaciones lineales exactas que hagan inestimables los parámetros.

[2] Llamemos $G$ a dicho segundo sumando. Para mostrar que es definida no negativa, basta ver que para cualquier $\vec{a}$ se verifica $\vec{a}'G\vec{a} \geq 0$. Pero $\vec{a}'G\vec{a} = \vec{b}'(Z'Z - Z'X(X'X)^{-1}XZ)^{-1}\vec{b}$ con $\vec{b} = Z'X(X'X)^{-1}\vec{a}$; ya sólo tenemos que comprobar que $(Z'Z - Z'X(X'X)^{-1}XZ)^{-1}$ es definida no negativa, o equivalentemente que $(Z'Z - Z'X(X'X)^{-1}XZ)$ lo es. Esto último es inmediato: $(Z'Z - Z'X(X'X)^{-1}XZ) = Z'(I - X(X'X)^{-1}X)Z$, y $\vec{d}'Z'(I - X(X'X)^{-1}X)Z\vec{d}$ puede escribirse como $\vec{e}\,'(I - X(X'X)^{-1}X)\vec{e}$ con $\vec{e} = Z\vec{d}$. La matriz de la forma cuadrática en $\vec{e}$ es la conocida matriz de coproyección, definida no negativa por ser idempotente (con valores propios cero o uno).

la inclusión de regresores irrelevantes no disminuye, y en general incrementa, las varianzas de los estimadores de los parámetros relevantes. No afecta sin embargo a su insesgadez.

De cuanto antecede se deduce que

$$\left(\vec{Y} - \begin{pmatrix} X & Z \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix}\right) \tag{5.11}$$

es un vector aleatorio de media cero. Denominando,

$$\begin{aligned} L &= \begin{pmatrix} X & Z \end{pmatrix} \\ \hat{\delta} &= \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} \end{aligned}$$

un desarrollo enteramente similar al realizado en el Teorema 4.1, pág. 41, muestra que en el modelo sobreparametrizado

$$SSE = \vec{Y}\,'(I - L(L'L)^{-1}L')\vec{Y} = \vec{\epsilon}\,'(I - L(L'L)^{-1}L')\vec{\epsilon} \tag{5.12}$$

es, bajo los supuestos habituales más normalidad, una forma cuadrática con distribución $\sigma^2\chi^2_{N-(p+q)}$, en que $p$ y $q$ son respectivamente los rangos de $X$ y $Z$. En consecuencia,

$$\hat{\sigma}^2 = \frac{SSE}{N-(p+q)} \tag{5.13}$$

es un estimador insesgado de $\sigma^2$. El único efecto adverso de la inclusión de los $q$ regresores irrelevantes ha sido la pérdida de otros tantos grados de libertad.

## 5.3. Omisión de regresores relevantes.

Sea $X = (X_1 \vdots X_2)$ y $\vec{\beta}\,' = (\vec{\beta}\,'_1 \vdots \vec{\beta}\,'_2)$. Consideremos el caso en que el modelo "correcto" es

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon} = X_1\vec{\beta}_1 + X_2\vec{\beta}_2 + \vec{\epsilon} \tag{5.14}$$

pese a lo cual estimamos el modelo "escaso"

$$\vec{Y} = X_1\vec{\beta}_1 + \vec{\epsilon} \tag{5.15}$$

Estimar (5.15) es lo mismo que estimar (5.14) junto con las restricciones $h : \vec{\beta}_2 = \vec{0}$, expresables así:

$$\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \vec{\beta}_1 \\ \vec{\beta}_2 \end{pmatrix} = \begin{pmatrix} \vec{0} \\ \vec{0} \end{pmatrix} \tag{5.16}$$

En consecuencia, podemos deducir cuanto necesitamos saber haciendo uso de los resultados en la Sección 3.2. Las siguientes conclusiones son así inmediatas.

- El estimador $\hat{\beta}_1^{(h)}$ obtenido en el modelo "escaso" (5.15) es, en general, sesgado. El sesgo puede obtenerse haciendo uso de (3.15). Tenemos así que,

$$\begin{pmatrix} \hat{\beta}_1^{(h)} \\ \vec{0} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - \vec{0}) \quad (5.17)$$

  y en consecuencia:

$$E[\hat{\beta}_1^{(h)} - \vec{\beta}_1] = -\left[(X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}\begin{pmatrix} \vec{0} \\ \vec{\beta}_2 \end{pmatrix}\right]_{(p\times 1)} \quad (5.18)$$

  en que $[M]_{(p\times q)}$ designa el bloque superior izquierdo con $p$ filas y $q$ columnas de la matriz $M$. La ecuación (5.18) muestra que el sesgo introducido depende de la magnitud de los parámetros asociados a los regresores omitidos.

- La ecuación (5.18) muestra también que hay un caso particular en que $\hat{\beta}_1^{(h)}$ es insesgado para $\vec{\beta}_1$; cuando las columnas de $X_1$ y las de $X_2$ son ortogonales, $X_1'X_2 = 0$, la matrix $(X'X)^{-1}$ es diagonal por bloques, y

$$(X'X)^{-1}A' = \begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix}^{-1}\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} \quad (5.19)$$

  tiene sus primeras $p$ filas de ceros. Ello hace que el bloque considerado en (5.18) esté formado por ceros.

- El estimador de la varianza de la perturbación

$$\hat{\sigma}^2 = \frac{SSE}{N-p} = \frac{(\vec{Y} - X\hat{\beta}_1^{(h)})'(\vec{Y} - X\hat{\beta}_1^{(h)})}{N-p} \quad (5.20)$$

  **no** es insesgado. En efecto, $(\vec{Y} - X\hat{\beta}_1^{(h)})$ es un vector aleatorio de media no nula, y sin dificultad se prueba que $SSE$ en (5.20), convenientemente reescalado, sigue una distribución $\chi^2$ no central, cuyo valor medio supera $\sigma^2(N-p)$; véase el Ejercicio 5.2.

## 5.4. Consecuencias de orden práctico

Los resultados de las dos Secciones anteriores pueden ayudarnos a tomar decisiones a la hora de especificar un modelo. Hemos visto que sobreparametrizar no introduce sesgos: tan sólo incrementa la varianza de los estimadores y resta grados de libertad. Errar "por exceso" tendrá por ello en general consecuencias leves, y tanto menos importantes cuanto mayor sea el tamaño muestral. La pérdida de un grado de libertad adicional originada por la inclusión de un parámetro es menos importante cuando los grados de libertad restantes $(N-p)$ siguen siendo muchos.

La sóla circunstancia en que la inclusión de un regresor innecesario puede perjudicar gravemente la estimación se presenta cuando la muestra es muy pequeña o el parámetro adicional es aproximadamente combinación lineal de los ya presentes. A esta última cuestión volveremos en el Capítulo 7.

Omitir regresores relevantes tiene consecuencias en general más graves y que no se atenúan al crecer el tamaño muestral: el sesgo de $\hat{\beta}_1^{(h)}$ en el modelo "escaso" no decrece hacia cero al crecer $N$.

En este capítulo hemos rastreado las consecuencias de dos posibles errores de especificación "puros": falta o sobra de regresores. En la práctica los dos tipos de errores se pueden presentar conjuntamente y sus efectos se combinan.

Conocidos los problemas de una mala especificación se plantea el problema de cómo lograr una buena. Esta cuestión se trata en el Capítulo 10. Algunas técnicas de análisis gráfico de residuos que pueden ser de ayuda en la especificación de modelos se consideran en la Sección 11.2.1.

## CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

**5.1** La distribución $\chi^2(\delta)$ no central con parámetro de descentramiento $\delta$ se introduce en Trocóniz (1987a), Sección 22.2.VIII. Obténgase razonadamente su media.

**5.2** Demuéstrese que el numerador de (5.20) dividido entre $\sigma^2$ sigue una distribución $\chi^2(\delta)$ no central. ¿Cuál es el parámetro de descentramiento?

# Capítulo 6

# Inferencia simultánea.

## 6.1. Problemas que plantea el contrastar múltiples hipótesis simultáneas

### 6.1.1. Evidencia contra una hipótesis

Si examinamos la teoría sobre contrastes de hipótesis presentada en la Sección 4.2 veremos que el método ha sido el habitual en Estadística no bayesiana. Los pasos se pueden esquematizar así:

1. Fijar una hipótesis $H_0$ sobre los parámetros de un modelo.
2. Seleccionar un estadístico cuya distribución sea conocida cuando $H_0$ es cierta y que se desvía de modo predecible de dicha distribución cuando $H_0$ no es cierta.
3. Calcular el valor del estadístico en una determinada muestra.
4. **Si el valor de dicho estadístico es anómalo** respecto de lo que esperaríamos bajo $H_0$, **rechazar** $H_0$**.**

La lógica subyacente es: "Como cuando $H_0$ es cierta es difícil que se de un valor del estadístico como el observado, lo más plausible es que $H_0$ no sea cierta."

Cuando el estadístico que empleamos en el contraste tiene una distribución continua, todos los valores posibles tienen probabilidad cero. No obstante, podemos escalafonarlos en más o menos "raros" de acuerdo con su densidad respectiva.

> **Ejemplo 6.1** Para una muestra $X_1, \ldots, X_n$ procedente de una distribución $N(\mu, \sigma^2)$, todos los posibles valores del estadístico $\overline{X}$ tienen probabilidad cero. No obstante, la distribución de dicho estadístico —una $N(\mu, \sigma^2/n)$— genera de modo frecuente observaciones en las cercanías de $\mu$, y sólo raramente valores en las colas. Consideraremos a estos últimos "raros" y favoreciendo el rechazo de $H_0$. Tienen densidad menor que los cercanos a $\mu$.

Tendrá interés en lo que sigue la noción de *nivel de significación empírico*[1].

**Definición 6.1** *Llamamos* nivel de significación empírico *asociado al valor observado de un estadístico a la probabilidad de obtener en el muestreo (bajo $H_0$) valores tan o más raros que el obtenido.*

> **Ejemplo 6.2** En el Ejemplo 6.1, supongamos que $H_0 : \mu = 0$. Supongamos conocida $\sigma^2 = 1$. Sea una muestra con $n = 100$, e imaginemos que obtenemos un valor de $\overline{X}$ de 0.196 ($= 1{,}96 \times \sqrt{100^{-1}}$). El nivel de significación empírico (u *observado*) sería 0.05, porque bajo $H_0$ hay probabilidad 0.05 de observar valores de $\overline{X}$ igual o más alejados de $\mu$ que el que se ha presentado.

Si en ocasiones al abordar un contraste de hipótesis prefijamos de antemano el nivel de significación que deseamos utilizar (y la región crítica), es muy frecuente el realizar el contraste sin una región crítica preespecificada y tomar el nivel de significación empírico como una medida del acuerdo (o desacuerdo) de la evidencia con la hipótesis de interés. Niveles de significación empíricos muy pequeños habrían así de entenderse como evidencia contra la hipótesis nula objeto de contraste.

### 6.1.2. ¿Cómo de "raro" ha de ser algo para ser realmente "raro"?

El siguiente ejemplo[2] ilustra que un resultado aparentemente muy raro puede no serlo tanto.

> **Ejemplo 6.3** Consideremos un mono frente a una máquina de escribir. Imaginemos que tras un periodo de tiempo observamos el conjunto de folios tecleados por el mono y constatamos que ¡ha escrito sin una sóla falta de ortografía *Hamlet*!
>
> Bajo la hipótesis nula $H_0$: "mono irracional", tal resultado es absolutamente inverosímil. La probabilidad de que golpeando al azar un teclado un mono logre tal cosa es ridículamente baja. Supongamos que una obra como *Hamlet* requiriera, entre blancos y caracteres, de 635000 digitaciones. Supongamos que hay 26 letras más caracteres de puntuación, etc. totalizando 32 posibilidades de digitación. Componer *Hamlet* totalmente al azar consistiría en apretar la tecla correcta sucesivamente 635.000 veces, algo que, suponiendo las 32 posibilidades de digitación equiprobables, tendría probabilidad:
>
> $$p = \left(\frac{1}{32}\right)^{635000} \approx 5{,}804527 \times 10^{-955771}. \tag{6.1}$$
>
> La observación de un mono que teclea *Hamlet* sería prácticamente imposible bajo $H_0$: habríamos de rechazar $H_0$ y pensar en alguna alternativa (¿quizá Shakespeare reencarnado en un mono?)
>
> Imaginemos ahora una multitud de monos a los que situamos frente a máquinas de escribir, haciéndoles teclear a su entero arbitrio 635.000 digitaciones. Específicamente, imaginemos $10^{955771}$ monos. Supongamos que examinando el trabajo de cada uno de ellos, nos topamos con que el mono $n$-ésimo ¡ha compuesto *Hamlet*! ¿Lo separaríamos de sus congéneres para homenajearlo como reencarnación de Shakespeare? Claramente no; porque, entre tantos, no es extraño que uno, por puro azar, haya tecleado *Hamlet*. De hecho, si todos los conjuntos de 635.000 digitaciones son equiprobables, del trabajo de $10^{955771}$ monos esperaríamos obtener un promedio en torno a $5{,}8045$ transcripciones exactas de *Hamlet*. Lo observado no es raro en absoluto.

[1] O *p-value*, en la literatura inglesa.

[2] Paráfrasis de un célebre comentario de Bertrand Russell.

El ejemplo anterior, deliberadamente extremo e inverosímil, ilustra un punto importante. Algo, aparentemente lo mismo, puede ser raro o no dependiendo del contexto. Observar un mono tecleando *Hamlet* es rarísimo, pero si *seleccionamos* el mono entre una miríada de ellos *precisamente porque ha tecleado Hamlet*, ya no podemos juzgar el suceso observado del mismo modo. ¡Hemos seleccionado la observación por su rareza, no podemos extrañarnos de que sea rara!

Cuando seleccionamos la evidencia, hemos de tenerlo en cuenta al hacer inferencia. De otro modo, estaremos prejuzgando el resultado.

### 6.1.3. Análisis exploratorio e inferencia

Es importante entender lo que el Ejemplo 6.3 intenta transmitir. El error, frecuente en el trabajo aplicado, es *seleccionar la evidencia* e ignorar este hecho *al producir afirmaciones o resultados de tipo inferencial* como rechazar tal o cual hipótesis con nivel de significación $p$, construir tal o cual intervalo con confianza $(1-p)$. Es el valor de $p$ que reportamos el que resulta completamente irreal a menos que corrijamos el efecto de la selección.

**Ejemplo 6.4** Regresemos al Ejemplo 6.3. Imaginemos la segunda situación descrita en que uno entre los $10^{955771}$ monos examinados compone *Hamlet*. Sería incorrecto rechazar la hipótesis $H_0$: "Los monos son irracionales." atribuyendo a esta decisión un nivel de significación de $5{,}804525 \times 10^{-955771}$. Por el contrario, la probabilidad de que ninguno de los monos hubiera tecleado *Hamlet* sería:

$$\begin{aligned} p_0 &= (1-p)^{10^{955771}} \\ &= \left[1 - \left(\frac{1}{32}\right)^{635000}\right]^{10^{955770}} \\ &\approx 0{,}0030138, \end{aligned}$$

el último valor calculado haciendo uso de una aproximación de Poisson (con media $\lambda = 5{,}804527$). Por tanto, la probabilidad de observar una o más transcripciones de *Hamlet* (un suceso tan raro o más raro que el observado, bajo $H_0$) ¡es tan grande como $1 - 0{,}0030138 = 0{,}9969862$! Difícilmente consideraríamos evidencia contra la hipótesis nula algo que, bajo $H_0$, acontece con probabilidad mayor que 0.99.

Nada nos impide, sin embargo, hacer análisis exploratorio: examinar nuestros datos, y seleccionar como interesante la evidencia que nos lo parezca.

**Ejemplo 6.5** De nuevo en el Ejemplo 6.3, no hay nada reprobable en examinar el trabajo de cada uno de los monos y detenernos con toda atención a examinar al animal que produce *Hamlet*. Seguramente le invitaríamos a seguir escribiendo. Sería del mayor interés que *ese mono* produjera a continuación *Macbeth*.

Lo que es reprobable es seleccionar el único mono que teclea *Hamlet* y reportar el hallazgo como si ese mono fuera el único observado.

### 6.1.4. Inferencia simultánea y modelo de regresión lineal ordinario

Pero ¿qué tiene ésto que ver con el modelo de regresión lineal, objeto de nuestro estudio?

Bastante. En ocasiones, hemos de hacer uso de modelos con un número grande de parámetros. Cuando ello ocurre, hay muchas hipótesis que podemos plantearnos

contrastar. Si lo hacemos, hemos de ser conscientes de que algunas hipótesis serán objeto de rechazo con una probabilidad mucho mayor que el nivel de significación nominal empleado para contrastar cada una de ellas. El siguiente ejemplo lo aclara.

**Ejemplo 6.6** Supongamos el modelo

$$\vec{Y} \quad = \quad \beta_0 \vec{X}_{\,0} + \beta_1 \vec{X}_{\,1} + \ldots + \beta_{99} \vec{X}_{\,99} + \vec{\epsilon}.$$

Supongamos, por simplicidad, normalidad de las perturbaciones y ortogonalidad de las columnas de la matriz de diseño. Dicho modelo tiene su origen en nuestra completa ignorancia acerca de cuál de las cien variables regresoras consideradas, si es que alguna, influye sobre la respuesta.

Si quisiéramos contrastar la hipótesis $H_0 : \beta_i = 0,\ i = 0, \ldots, 99$, podríamos (si se verifican los supuestos necesarios) emplear el contraste presentado en la Sección 4.2.2. Podríamos ser más ambiciosos e intentar al mismo tiempo ver cuál o cuales $\beta_i$ son distintos de cero. Sería *incorrecto* operar así:

1. Contrastar las hipótesis $H_{0i} :\ \beta_i = 0$ al nivel de significación $\alpha$ comparando cada $t$-ratio en valor absoluto con $t^{\alpha/2}_{N-p}$.
2. Si algún $t$-ratio excede $t^{\alpha/2}_{N-p}$, rechazar la hipótesis $H_{0i}$, y por consiguiente $H_0$, reportando un nivel de significación $\alpha$.

Es fácil ver por qué es incorrecto. Bajo $H_0$ hay probabilidad tan sólo $\alpha$ de que un $t$-ratio prefijado exceda en valor absoluto de $t^{\alpha/2}_{N-p}$. Pero la probabilidad de que *algún* $t$-ratio exceda de $t^{\alpha/2}_{N-p}$ es[3]

$$\text{Prob}(\text{Algún } \beta_i \neq 0) = 1 - (1 - \alpha)^p. \tag{6.2}$$

mayor (en ocasiones *mucho mayor*) que $\alpha$. Tomemos por ejemplo el caso examinado en que $p = 100$ y supongamos $\alpha = 0{,}05$. La probabilidad de obtener algún $t$-ratio fuera de límites es $1 - 0{,}95^{100} = 0{,}9940$. Lejos de tener un nivel de significación de $\alpha = 0{,}05$, el que tenemos es de $0{,}9940$. Contrastar la hipótesis $H_0$ de este modo tiene una probabilidad de falsa alarma de 0.9940.

Si nuestro propósito fuera puramente exploratorio, nada debe disuadirnos de estimar el modelo con los cien regresores y examinar luego las variables asociadas a $t$-ratios mayores, quizá estimando un modelo restringido con muestra adicional. Lo que es inadmisible es dar un nivel de significación incorrectamente calculado.

El problema de inferencias distorsionadas es grave y muchas veces indetectable. Pensemos en el investigador que hace multitud de regresiones, quizá miles, a cuál más descabellada. Por puro azar, encuentra una pocas con $R^2$ muy alto, escribe un artículo y lo publica. Si el experimento es reproducible, cabe esperar que otros investigadores tratarán de replicarlo y, al no lograrlo —el $R^2$ alto era casualidad—, la superchería quedará al descubierto. Pero si la investigación versa sobre, por ejemplo, Ciencias Sociales, en que con frecuencia una y sólo una muestra está disponible, todo lo que sus colegas podrán hacer es reproducir sus resultados con la única muestra a mano. A menos que el primer investigador tenga la decencia de señalar que el alto $R^2$ obtenido era el más alto entre miles de regresiones efectuadas (lo que permitiría calcular correctamente el nivel de significación y apreciar de un modo realista su valor como evidencia), es fácil que su trabajo pase por ciencia.

De nuevo es preciso insistir: no hay nada objetable en la realización de miles de regresiones, quizá con carácter exploratorio. Tampoco es objetable el concentrar la atención en la única (o las pocas) que parecen prometedoras. Al revés, ello es muy sensato.

[3] Bajo la hipótesis de independencia entre los respectivos $t$-ratios, hipótesis que se verifica en por la normalidad y la ortogonalidad entre las columnas de la matriz de diseño.

Lo que es objetable es reportar dichas regresiones como si fueran las únicas realizadas, el resultado de estimar un modelo prefijado de antemano, dando la impresión de que la evidencia muestral sustenta una hipótesis o modelo pre-establecidos, cuando lo cierto es que la hipótesis o modelo han sido escogidos a la vista de los resultados.

## 6.2. Desigualdad de Bonferroni.

Consideremos $k$ sucesos, $E_i,\ (i = 1, \ldots, k)$, cada uno de ellos con probabilidad $(1 - \alpha)$. La probabilidad de que todos acaezcan simultáneamente es:

$$\text{Prob}\{\cap_{i=1}^{k} E_i\} = 1 - \text{Prob}\{\overline{\cap_{i=1}^{k} E_i}\} = 1 - \text{Prob}\{\cup_{i=1}^{k} \overline{E_i}\} \geq 1 - k\alpha \qquad (6.3)$$

Se conoce (6.3) como *desigualdad de Bonferroni de primer orden*. Es una igualdad si los $\overline{E_i}$ son disjuntos. Muestra que la probabilidad conjunta de varios sucesos puede, en general, ser muy inferior a la de uno cualquiera de ellos. Por ejemplo, si $k = 10$ y $\text{Prob}\{E_i\} = 0{,}95 = 1 - 0{,}05$, la desigualdad anterior solo permite garantizar que $\text{Prob}\{\cap_{i=1}^{k} E_i\} \geq 1 - 10 \times 0{,}05 = 0{,}50$.

Consideremos ahora el modelo $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$ y los siguientes sucesos:

$$E_1 \quad : [(\hat{\beta}_1 \pm \hat{\sigma}_{\hat{\beta}_1} t_{N-p}^{\alpha/2}) \quad \text{cubre } \beta_1] \qquad (6.4)$$

$$\vdots \qquad (6.5)$$

$$E_k \quad : [(\hat{\beta}_k \pm \hat{\sigma}_{\hat{\beta}_k} t_{N-p}^{\alpha/2}) \quad \text{cubre } \beta_k] \qquad (6.6)$$

Cada $E_i$ por separado es un suceso cuya probabilidad es $1 - \alpha$. De acuerdo con (6.3), sin embargo, todo cuanto podemos asegurar acerca de $\text{Prob}\{\cap_{i=1}^{k} E_i\}$ es que su probabilidad es superior a $1 - k\alpha$.

Las implicaciones son importantes. Si regresáramos $\vec{Y}$ sobre $\vec{X}_0, \ldots, \vec{X}_{p-1}$ y quisiéramos obtener intervalos de confianza *simultáneos* $\alpha$ para los parámetros $\beta_0, \cdots, \beta_{p-1}$, sería claramente incorrecto emplear los que aparecen en (6.4)–(6.6). Si actuásemos de este modo, el nivel de confianza conjunto no sería el deseado de $1 - \alpha$, sino que tan sólo podríamos afirmar que es mayor que $1 - k\alpha$.

Si queremos intervalos de confianza *simultáneos* al nivel $1-\alpha$, podríamos construir intervalos para cada uno de los parámetros con un nivel de confianza $\psi = \frac{\alpha}{k}$. Haciendo ésto, tendríamos que la probabilidad de que *todos* los $\beta_i$ fueran cubiertos por sus respectivos intervalos, sería mayor, de acuerdo con (6.3), que $1-k\psi = 1-k(\frac{\alpha}{k}) = 1-\alpha$. Ello se logra, sin embargo, al coste de ensanchar el intervalo de confianza correspondiente a cada $\beta_i$ quizá mas de lo necesario. En lo que sigue veremos procedimientos para lograr el mismo resultado con intervalos en general más estrechos.

## 6.3. Intervalos de confianza basados en la máxima $t$.

Supongamos que tenemos $k$ variables aleatorias independientes, $t_1, \ldots, t_k$ con distribución $t$-Student, y número común $n$ de grados de libertad. La variable aleatoria $\text{máx}\{|t_1|, \ldots, |t_k|\}$ sigue una distribución que se halla tabulada[4].

[4] Véase, por ej., Seber (1977), Apéndice E.

Sea $u^{\alpha}_{k,n}$ el cuantil $1-\alpha$ de dicha distribución, es decir, un valor que resulta superado con probabilidad $\alpha$ por $\max\{|t_1|,\ldots,|t_k|\}$. Entonces,

$$\text{Prob}\{\cap_{i=1}^{k}[|t_i| \leq u^{\alpha}_{k,n}]\} = 1-\alpha,$$

dado que si $u^{\alpha}_{k,n}$ acota con probabilidad $1-\alpha$ al máximo, acota simultáneamente con la misma probabilidad la totalidad de las variables aleatorias .

Si $\vec{a}_i\,'\hat{\beta}/\hat{\sigma}_{\vec{a}_i\,'\hat{\beta}}$ $(i=1,\ldots,k)$ fueran independientes, y la hipótesis nula $h:\ \vec{a}_i\,'\vec{\beta} = 0$ $(i=1,\ldots,k)$ fuera cierta, tendríamos que:

$$\text{Prob}\left\{\bigcap_{i=1}^{k}\left[\left|\frac{\vec{a}_i\,'\hat{\beta}}{\hat{\sigma}_{\vec{a}_i\,'\hat{\beta}}}\right| \leq u^{\alpha}_{k,n}\right]\right\} = 1-\alpha \tag{6.7}$$

Es claro que $\vec{a}_i\,'\hat{\beta}/\hat{\sigma}_{\vec{a}_i\,'\hat{\beta}}$ $(i=1,\ldots,k)$ **no** son independientes en general. Sin embargo, la distribución aludida del máximo valor absoluto de $k$ variables $t$ de Student está también tabulada cuando dichas variables tienen correlación $\rho$ por pares. Ni siquiera ésto es cierto en general[5]. Aún así, (6.7) es de utilidad. Suministra intervalos simultáneos de confianza aproximada $1-\alpha$. En caso de que conociéramos $\rho$, emplearíamos la expresión (6.7) con $u^{\alpha}_{k,n}$ reemplazado por $u^{\alpha}_{k,n,\rho}$, extraído éste último de la tabla correspondiente.

Es importante señalar que, si nuestro objetivo es contrastar una hipótesis del tipo $h\colon A\vec{\beta} = \vec{c}$ con rango$(A) > 1$, *tenemos* que emplear un contraste como el descrito en la Sección 4.2. El comparar cada una de las variables aleatorias $\left|(\vec{a}_i\,'\hat{\beta} - c_i)/\hat{\sigma}_{\vec{a}_i\,'\hat{\beta}}\right|$ $(i=1,\ldots,k)$ con una $t^{\alpha/2}_{N-p}$ supone emplear un nivel de significación *mayor* que $\alpha$. Como caso particular, es inadecuado contrastar la hipótesis $h\colon \beta_0 = \cdots = \beta_p = 0$ comparando cada uno de los $t$-ratios con $t^{\alpha/2}_{N-p}$; tal contraste tendría un nivel de significación sensiblemente superior a $\alpha$, en especial si $p$ es grande.

En el caso de que el contraste conjunto rechace $h\colon A\vec{\beta} = \vec{c}$ y queramos saber qué filas de $A$ son culpables del rechazo, podríamos comparar $\left|(\vec{a}_i\,'\hat{\beta} - c_i)/\hat{\sigma}_{\vec{a}_i\,'\hat{\beta}}\right|$ $(i=1,\ldots,k)$ con $u^{\alpha}_{k,n}$ ($k$ = número de filas de $A$). Nótese que es perfectamente posible rechazar la hipótesis conjunta y no poder rechazar ninguna de las hipótesis parciales correspondientes a las filas de $A$.

## 6.4. Método S de Scheffé.

Este método permite la construcción de un número arbitrario de intervalos de confianza simultáneos, de manera muy simple. Necesitaremos el siguiente lema:

**Lema 6.1** *Sea $L$ una matriz simétrica de orden $k\times k$ definida positiva, y $\vec{c}$, $\vec{b}$ vectores $k\times 1$ cualesquiera en $R$. Se verifica que:*

$$\sup_{\vec{c}\neq\vec{0}}\left(\frac{[\vec{c}\,'\vec{b}]^2}{\vec{c}\,'L\vec{c}}\right) \quad = \quad \vec{b}\,'L^{-1}\vec{b} \tag{6.8}$$

[5] Salvo en casos particulares, como el de ciertos diseños de Análisis de Varianza equilibrados, en que $\rho$ es además muy fácil de calcular.

DEMOSTRACION:

Siendo $L$ definida positiva, existe una matriz $R$ cuadrada no singular tal que: $L = RR'$. Si definimos:

$$\vec{v} = R'\vec{c} \tag{6.9}$$
$$\vec{u} = R^{-1}\vec{b} \tag{6.10}$$

y tenemos en cuenta que por la desigualdad de Schwarz,

$$\frac{<\vec{u},\vec{v}>^2}{\parallel \vec{u} \parallel^2 \parallel \vec{v} \parallel^2} \leq 1 \tag{6.11}$$

entonces sustituyendo (6.9) y (6.10) en (6.11) obtenemos (6.8).

Podemos ahora abordar la demostración del método de Scheffé. Supongamos que tenemos $k$ hipótesis lineales $h_i\colon \vec{a}_i'\vec{\beta} = c_i \ (i = 1, \ldots, k)$ cuyo contraste conjunto deseamos efectuar. Si denominamos:

$$A = \begin{pmatrix} \vec{a}_1\,' \\ \vec{a}_2\,' \\ \cdots \\ \vec{a}_k\,' \end{pmatrix} \quad \vec{c} = \begin{pmatrix} c_1 \\ c_2 \\ \cdots \\ c_k \end{pmatrix} \tag{6.12}$$

dichas $k$ hipótesis se pueden escribir como $h\colon A\vec{\beta} = \vec{c}$. Cuando $h$ es cierta, sabemos (Sección 4.2) que:

$$\frac{(A\hat{\beta} - \vec{c})'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - \vec{c})}{q\hat{\sigma}^2} \sim \mathcal{F}_{q,N-p} \tag{6.13}$$

siendo $q = \min(d, p)$, en que $d = \operatorname{rango} A$ y $p = \operatorname{rango}(X'X)$. Las inversas pueden ser inversas generalizadas, si los rangos de las matrices así lo exigen.

Llamemos $\hat{c}$ a $A\hat{\beta}$. Bajo $h$, sabemos que:

$$1 - \alpha = \operatorname{Prob}\left\{(\hat{c} - \vec{c})'[A(X'X)^{-1}A']^{-1}(\hat{c} - \vec{c}) \leq q\hat{\sigma}^2\mathcal{F}^{\alpha}_{q,N-p}\right\} \tag{6.14}$$
$$= \operatorname{Prob}\left\{(\hat{c} - \vec{c})'L^{-1}(\hat{c} - \vec{c}) \leq q\hat{\sigma}^2\mathcal{F}^{\alpha}_{q,N-p}\right\} \tag{6.15}$$

en que $L = [A((X'X))^{-1}A']$. Teniendo en cuenta el Lema 6.1, obtenemos:

$$1 - \alpha = \operatorname{Prob}\left\{\sup_{\vec{h}\neq\vec{0}}\left(\frac{[\vec{h}\,'(\hat{c} - \vec{c})]^2}{\vec{h}\,'L\vec{h}}\right) \leq q\hat{\sigma}^2\mathcal{F}^{\alpha}_{q,N-p}\right\} \tag{6.16}$$
$$= \operatorname{Prob}\left\{\bigcap_{\vec{h}\neq\vec{0}}\left[\left|\frac{\vec{h}\,'(\hat{c} - \vec{c})}{\left(\vec{h}\,'L\vec{h}\right)^{\frac{1}{2}}}\right| \leq \left(q\hat{\sigma}^2\mathcal{F}^{\alpha}_{q,N-p}\right)^{\frac{1}{2}}\right]\right\} \tag{6.17}$$

La ecuación (6.17) muestra que $\left(q\hat{\sigma}^2\mathcal{F}^{\alpha}_{q,N-p}\right)^{\frac{1}{2}}$ es un valor que acota con probabilidad $1 - \alpha$ un número arbitrariamente grande de cocientes como:

$$\frac{\left|\vec{h}\,'(\hat{c} - \vec{c})\right|}{\sqrt{\vec{h}\,'L\vec{h}}} \tag{6.18}$$

Por consiguiente, cuantos intervalos para $\vec{h}\,'\vec{c}$ construyamos de la forma:

$$\vec{h}\,'\hat{c} \pm \sqrt{(\vec{h}\,'L\vec{h})(q\hat{\sigma}^2\mathcal{F}^{\alpha}_{q,N-p})} \tag{6.19}$$

tendrán confianza *simultánea* $1-\alpha$.

Esto es *más* de lo que necesitamos —pues sólo queríamos intervalos de confianza simultáneos para $c_1,\ldots,c_k$—. El método de Scheffé proporciona intervalos de confianza conservadores (más amplios, en general, de lo estrictamente necesario).

Obsérvese que, en el caso particular en que $A = I_{p\times p}$, los intervalos de confianza en (6.19) se reducen a:

$$\vec{h}\,'\hat{\beta} \pm \sqrt{(\vec{h}\,'((X'X))^{-1}\vec{h})(p\hat{\sigma}^2\mathcal{F}^{\alpha}_{p,N-p})} \tag{6.20}$$

expresión que será frecuente en la práctica. Cuando el conjunto de hipótesis simultáneas que se contrastan configure una matriz $A$ de rango $q < p$, será sin embargo conveniente tener en cuenta este hecho, ya que obtendremos intervalos menos amplios.

**R: Ejemplo 6.1** *(uso del método de Scheffé)*

El siguiente código implementa el método de Scheffé para contrastar la igualdad entre todas las parejas de parámetros intervinientes en un modelo. La matriz de diseño es una matriz de ceros y unos. Si, por ejemplo, $X_{ij}$ fuera "uno" cuando la $j$-ésima parcela se siembra con la variedad $i$-ésima de semilla y la variable respuesta recogiera las cosechas obtenidas en las diferentes parcelas, los parámetros $\beta_i$ serían interpretables como la productividad de las diferentes variedades de semilla (suponemos que no hay otros factores en juego; las parcelas son todas homogéneas. En la Parte II se considera este problema con mayor generalidad).

En una situación como la descrita tendría interés contrastar todas las hipótesis del tipo: $h_{ij} : \beta_i - \beta_j = 0$. Aquellas parejas para las que no se rechazase corresponderían a variedades de semilla no significativamente diferentes.

Fácilmente se ve que el contraste de todas las hipótesis de interés agrupadas ($h : A\equiv\vec{c}$) no es de gran interés: no nos interesa saber si *hay* algunas variedades de semilla diferentes, sino *cuáles son*. Fácilmente se ve también que, incluso para un número moderado de variedades de semilla, hay bastantes parejas que podemos formar y el realizar múltiples contrastes como $h_{ij} : \beta_i - \beta_j = 0$ requerirá el uso de métodos de inferencia simultánea.

```
--- Obtenido mediante R BATCH demo7.R
> options(digits=5)
> #                                          #
> #  Generamos artificalmente los datos
> #
> options(warn=-1)          # la instrucción siguiente genera un
>                           # warning benigno.
>
> X <- matrix(c(rep(1,5),rep(0,25)),25,5)
> X
      [,1] [,2] [,3] [,4] [,5]
 [1,]    1    0    0    0    0
 [2,]    1    0    0    0    0
 [3,]    1    0    0    0    0
 [4,]    1    0    0    0    0
 [5,]    1    0    0    0    0
```

```
 [6,]    0    1    0    0    0
 [7,]    0    1    0    0    0
 [8,]    0    1    0    0    0
 [9,]    0    1    0    0    0
[10,]    0    1    0    0    0
[11,]    0    0    1    0    0
[12,]    0    0    1    0    0
[13,]    0    0    1    0    0
[14,]    0    0    1    0    0
[15,]    0    0    1    0    0
[16,]    0    0    0    1    0
[17,]    0    0    0    1    0
[18,]    0    0    0    1    0
[19,]    0    0    0    1    0
[20,]    0    0    0    1    0
[21,]    0    0    0    0    1
[22,]    0    0    0    0    1
[23,]    0    0    0    0    1
[24,]    0    0    0    0    1
[25,]    0    0    0    0    1
> b <- c(3,4,4,5,5)
> y <- X %*% b + rnorm(20,sd=0.1)
> #
> p <- ncol(X)                 # número de parámetros
> N <- nrow(X)                 # número de observaciones
> A <- cbind(1,diag(-1,p-1))   # las comparaciones pueden tomarse
>                              # como combinaciones lineales de las
>                              # filas de A
> q <- p-1                     # Este es el rango de A.
> A
     [,1] [,2] [,3] [,4] [,5]
[1,]    1   -1    0    0    0
[2,]    1    0   -1    0    0
[3,]    1    0    0   -1    0
[4,]    1    0    0    0   -1
> H <- matrix(0,p*(p-1)/2,p)   # matriz de comparaciones.
> j    <- 0
> for (i in ((p-1):1)) {
+   H[(j+1):(j+i),(p-i):p] <- cbind(1,diag(-1,i))
+   j <- j + i
+ }
> H                            # esta es la matriz de comparaciones
      [,1] [,2] [,3] [,4] [,5]
 [1,]    1   -1    0    0    0
 [2,]    1    0   -1    0    0
 [3,]    1    0    0   -1    0
 [4,]    1    0    0    0   -1
 [5,]    0    1   -1    0    0
 [6,]    0    1    0   -1    0
 [7,]    0    1    0    0   -1
 [8,]    0    0    1   -1    0
 [9,]    0    0    1    0   -1
[10,]    0    0    0    1   -1
> #
```

```
> fit   <- lsfit(X,y,intercept=FALSE)
> betas <- fit$coefficients
> s2    <- sum(fit$residuals^2) / (N - p)
> qsf   <- q*s2*qf(0.05,q,N-p)
> xxi   <- solve(t(X) %*% X)
> L     <- A %*% xxi %*% t(A)
> #
> #  El siguiente bucle construye todos los intervalos de confianza
> #  simultáneos. Nótese que ejecuciones sucesivas darán normalmente
> #  valores diferentes, dado que cada vez se genera una muestra
> #  artificial diferente
> #
> for (i in 1:nrow(H)) {
+   cat("Intervalo comp. ",H[i,])
+   z <- sqrt(t(H[i,]) %*% xxi %*% H[i,] * qsf)
+   d <- t(H[i,]) %*% betas
+   cat(" es: (",d - z," , ",d+z,")")
+   if((d-z < 0) && (d+z > 0))
+     cat("\n")
+   else
+     cat(" * \n")
+ }
Intervalo comp.  1 -1 0 0 0 es: ( -0.9869  ,  -0.89046 ) *
Intervalo comp.  1 0 -1 0 0 es: ( -1.0243  ,  -0.92784 ) *
Intervalo comp.  1 0 0 -1 0 es: ( -1.9448  ,  -1.8484 ) *
Intervalo comp.  1 0 0 0 -1 es: ( -2.0482  ,  -1.9518 ) *
Intervalo comp.  0 1 -1 0 0 es: ( -0.085598  ,  0.010846 )
Intervalo comp.  0 1 0 -1 0 es: ( -1.0061  ,  -0.90968 ) *
Intervalo comp.  0 1 0 0 -1 es: ( -1.1095  ,  -1.0131 ) *
Intervalo comp.  0 0 1 -1 0 es: ( -0.96875  ,  -0.8723 ) *
Intervalo comp.  0 0 1 0 -1 es: ( -1.0722  ,  -0.97572 ) *
Intervalo comp.  0 0 0 1 -1 es: ( -0.15163  ,  -0.055188 ) *
```

## 6.5. Empleo de métodos de inferencia simultánea.

Si el desarrollo anterior es formalmente simple, puede no ser obvio, en cambio, en que situaciones es de aplicación. Las notas siguientes esbozan algunas ideas sobre el particular[6].

- Emplearemos inferencia simultánea cuando *a priori*, y por cualquier motivo, estemos interesados en múltiples contrastes (o intervalos de confianza) y queramos que el nivel de significación conjunto sea $1 - \alpha$. Esta situación se presenta con relativa rareza en la práctica estadística.

- Más importante, emplearemos los métodos anteriores cuando la elección de hipótesis o parámetros objeto de contraste o estimación *se haga a la vista de los resultados*. Esta situación es muy frecuente en el análisis exploratorio. Sería incorrecto, por ejemplo, estimar una ecuación con veinte regresores, seleccionar aquel $\hat{\beta}_i$ con el máximo t-ratio, y comparar dicho t-ratio con una $t$ de Student con grados de libertad adecuados. Dado que hemos seleccionado el $\hat{\beta}_i$ de interés como el de mayor t-ratio, hemos de comparar éste con los cuantiles de la

[6] Puede consultarse también Trocóniz (1987a) Cap. 5 y Cox and Hinkley (1974), Sec. 7.4.

distribución del máximo de $k$ ($k = 20$ en este caso) variables aleatorias con distribución $t$ de Student ($u^{\alpha}_{20,N-20}$).

- Por último, conviene resaltar la diferencia entre el contraste de varias hipótesis simultáneas $\vec{a}_i{}'\vec{\beta} = c_i$ agrupadas en $A\vec{\beta} = \vec{c}$ mediante $Q_h$ (Sección 4.2) y el que hace uso de (6.7). El primero es perfectamente utilizable; el segundo será, en general, conservador —menos rechazos de los que sugiere el nivel de significación nominal—, pero tiene la ventaja de arrojar luz sobre cuales de las "subhipótesis" $\vec{a}_i{}'\vec{\beta} = c_i$ son responsables del rechazo, caso de que se produzca. Esta información queda sumergida al emplear $Q_h$.

Hay otros métodos de inferencia simultánea, si bien de carácter menos general. En la Sección 13.2.2 nos ocuparemos de uno de ellos (método de Tukey o del recorrido estudentizado).

## CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

**6.1** Un investigador sospecha que la concentración de una toxina en la sangre puede estar relacionada con la ingesta de algún tipo de alimento. Realiza un completo estudio en que para $N = 500$ sujetos mide la concentración de dicha toxina y las cantidades consumidas de 200 diferentes tipos de alimento. Cree razonable proponer como modelo explicativo,

$$Y \quad = \quad \beta_0 + \beta_1 X_1 + \ldots + \beta_{200} X_{200} + \epsilon.$$

Tras estimar los 201 parámetros del mismo, se plantea hacer contrastes de hipótesis como $H_0 : \beta_i = 0$, y considera las siguientes posibilidades:

- Comparar cada uno de los t-ratios $\hat{\beta}_i / \hat{\sigma}_{\hat{\beta}_i}$ con el cuantil $t_{N-p;\alpha/2}$.
- Idem con el cuantil correspondiente de una distribución $u_k$, del máximo de $k$ variables $t$ de Student.
- Calcular el estadístico $Q_h$ para la hipótesis $H_0 : \hat{\beta}_1, \ldots, \hat{\beta}_{200} = 0$ y comparar con $\mathcal{F}_{200,5000-200;\alpha}$.
- Hacer iii) y luego i).
- Hacer iii) y luego iv).

Juzga los diferente procedimientos, e indica con cuál (o cuáles) de ellos tendríamos al menos garantizada una probabilidad de error de tipo I no superior al $\alpha$ prefijado.

**6.2** Preocupado por el posible impacto de las antenas de telefonía móvil sobre la salud de los niños, un político solicita un lista completo de las 15320 escuelas del país a menos de 500 metros de una antena. Investiga la probabilidad de contraer leucemia y la probabilidad de que por puro azar se presenten los casos de leucemia que se han registrado en dichas escuelas.

Aparece un caso llamativo: en la escuela X con 650 niños hay tres que han contraido la enfermedad, lo que, de acuerdo con los cálculos realizados por nuestro político, asistido por un epidemiólogo, acontecería por azar con probabilidad 0,0003. Al día siguiente acude al Parlamento y pide la dimisión del Ministro de Sanindad: "Hay —dice– evidencia concluyente de que las antenas de telefonía móvil influyen en la prevalencia de la leucemia entre la población infantil. Un evento como el registrado en la escuela X sólo se presentaría por azar con probabilidad 0,0003". Comenta.

# Capítulo 7

# Multicolinealidad.

## 7.1. Introducción.

Hemos visto (Capítulo 3) que, en presencia de multicolinealidad exacta entre las columnas de la matriz de diseño $X$, la proyección de $\vec{y}$ sobre $M = R(X)$ sigue siendo única, pero no hay una única estimación de $\vec{\beta}$. Decíamos entonces que el vector de parámetros no estaba identificado. Este Capítulo[1] analiza esta cuestión con mayor detalle. En particular, aborda las siguientes cuestiones:

1. ¿Es estimable una cierta combinación lineal $\vec{c}\,'\vec{\beta}$ de los parámetros?

2. Si $\vec{c}\,'\vec{\beta}$ es estimable, ¿cuál es la varianza de la estimación?. ¿De qué depende la precisión con que pueden estimarse distintas combinaciones lineales de los parámetros?

3. ¿Como escoger la matriz de diseño $X$ —u observaciones adicionales a la misma— si el objetivo es estimar determinadas combinaciones lineales $\vec{c}\,'\vec{\beta}$ con varianza mínima?

Responder a la primera requiere que caractericemos las formas lineales estimables. La segunda pondrá de manifiesto que la varianza en la estimación de $\vec{c}\,'\vec{\beta}$ depende de la dirección del vector $\vec{c}$ en $R(X'X)$. La tercera cuestión hace referencia a un tema de gran interés; el de diseño óptimo, en los casos en que somos libres de escoger o ampliar $X$.

Las dos cuestiones anteriores desvelan en buena medida la naturaleza de la multicolinealidad aproximada. Quizá el problema de más interés práctico sea el previo de detectar su presencia; la Sección 7.5 trata brevemente esta cuestión.

[1] Basado en Silvey (1969).

## 7.2. Caracterización de formas lineales estimables.

**Teorema 7.1** *La forma lineal $\vec{c}\,'\vec{\beta}$ es estimable si, y solo si, $\vec{c}$ es una combinación lineal de los vectores propios de $X'X$ asociados a valores propios no nulos.*

DEMOSTRACION:

Observemos que el enunciado no es sino una paráfrasis del Teorema 3.1 (Capítulo 3). La siguiente cadena de implicaciones, que puede recorrerse en ambas direcciones, establece la demostración.

$$\begin{aligned}
\vec{c}\,'\vec{\beta} \text{ estimable} &\iff \exists \vec{d}\colon \vec{c}\,'\vec{\beta} = E[\vec{d}\,'\vec{Y}\,] && (7.1)\\
&\iff \vec{c}\,'\vec{\beta} = \vec{d}\,'X\vec{\beta} && (7.2)\\
&\iff \vec{c}\,' = \vec{d}\,'X && (7.3)\\
&\iff \vec{c} = X'\vec{d} && (7.4)\\
&\iff \vec{c} \in R(X') && (7.5)\\
&\iff \vec{c} \in R(X'X) && (7.6)\\
&\iff \vec{c} = \alpha_1\vec{v}_1 + \cdots + \alpha_{p-j}\vec{v}_{p-j} && (7.7)
\end{aligned}$$

siendo $\vec{v}_1, \ldots, \vec{v}_{p-j}$ los vectores propios de $(X'X)$ asociados a valores propios no nulos. El paso de (7.5) a (7.6) hace uso del hecho de que tanto las columnas de $X'$ como las de $X'X$ generan el mismo subespacio[2] de $R^p$. La equivalencia entre (7.6) y (7.7) hace uso del hecho de que los vectores propios de $R(X'X)$ asociados a valores propios no nulos generan $R(X'X)$.

Hay una forma alternativa de llegar al resultado anterior, que resulta interesante en sí misma y útil para lo que sigue. Sea $V$ la matriz diagonalizadora de $X'X$, y definamos:

$$\begin{aligned}
Z &= XV && (7.8)\\
\vec{\gamma} &= V'\vec{\beta} && (7.9)
\end{aligned}$$

Entonces, como $VV' = I$ tenemos que:

$$X\vec{\beta} = XVV'\vec{\beta} = Z\vec{\gamma} \tag{7.10}$$

y por consiguiente el modelo $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$ se transforma en: $\vec{Y} = Z\vec{\gamma} + \vec{\epsilon}$.

El cambio de variables y parámetros ha convertido la matriz de diseño en una matriz de columnas ortogonales:

$$Z'Z = (XV)'(XV) = V'X'XV = \Lambda \tag{7.11}$$

siendo $\Lambda$ una matriz cuya diagonal principal contiene los valores propios de $X'X$. Sin pérdida de generalidad los supondremos ordenados de forma que los $p - j$ primeros $\lambda's$ son no nulos, y los restantes $j$ son cero: $\lambda_p = \lambda_{p-1} = \cdots = \lambda_{p-j+1} = 0$.

[2]Es inmediato ver que $R(X'X) \subseteq R(X')$, pues si $\vec{v} \in R(X'X) \Rightarrow \exists \vec{a}\colon\ \vec{v} = X'X\vec{a} = X'\vec{d}$, siendo $\vec{d} = X\vec{a}$. Por otra parte, $R(X'X)$ no es subespacio propio de $R(X')$, pues ambos tienen la misma dimensión. Para verlo, basta comprobar que toda dependencia lineal entre las columnas de $X'X$ es una dependencia lineal entre las columnas de $X$. En efecto, $X'X\vec{b} = \vec{0} \Rightarrow \vec{b}'X'X\vec{b} = \vec{d}'\vec{d} = \vec{0} \Rightarrow \vec{d} = \vec{0} \Rightarrow X\vec{b} = \vec{0}$.

Observemos que de (7.9) se deduce, dado que $V$ es ortogonal, que $\vec{\beta} = V\vec{\gamma}$. Por consiguiente, es equivalente el problema de estimar $\vec{\beta}$ al de estimar $\vec{\gamma}$, pues el conocimiento de un vector permite con facilidad recuperar el otro. Las ecuaciones normales al estimar $\vec{\gamma}$ son:

$$(Z'Z)\hat{\gamma} = \Lambda\hat{\gamma} = Z'\vec{y} \tag{7.12}$$

o en forma desarrollada:

$$\begin{pmatrix} \lambda_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \lambda_{p-j} & \dots & 0 \\ 0 & 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 0 \end{pmatrix} \hat{\gamma} = Z'\vec{y} \tag{7.13}$$

El sistema (7.13) es indeterminado; solo los $(p-j)$ primeros $\hat{\gamma}'s$ pueden obtenerse de él. Obsérvese además que de (7.13 ) se deduce que var($\gamma_i$) $\propto 1/\lambda_i$, $(i = 1, \dots, p-j)$.

Consideremos una forma lineal cualquiera $\vec{c}\,'\vec{\beta}$. Tenemos que:

$$\vec{c}\,'\vec{\beta} = \vec{c}\,'VV'\vec{\beta} = (\vec{c}\,'V)\vec{\gamma} = (V'\vec{c})'\vec{\gamma} \tag{7.14}$$

y consiguientemente una estimación de $\vec{c}\,'\hat{\beta}$ vendrá dada por $(V'\vec{c})'\hat{\gamma}$. Por tanto, $\vec{c}\,'\vec{\beta}$ será estimable si $\hat{\gamma}$ es estimable, *o si $\vec{c}\,'\hat{\beta}$ depende sólo de aquellos $\hat{\gamma}'s$ que pueden ser estimados*. Es decir, en el caso de rango $(p-j)$ correspondiente a las ecuaciones normales (7.13), $\vec{c}\,'\vec{\beta}$ podrá estimarse si $(V'\vec{c})'$ tiene nulas sus últimas $j$ coordenadas, lo que a su vez implica:

$$\vec{c} \perp \vec{v}_p \tag{7.15}$$
$$\vec{c} \perp \vec{v}_{p-1} \tag{7.16}$$
$$\vec{c} \perp \vec{v}_{p-2} \tag{7.17}$$
$$\vdots \tag{7.18}$$
$$\vec{c} \perp \vec{v}_{p-j+1} \tag{7.19}$$

Por consiguiente, para que $\vec{c}\,'\vec{\beta}$ sea estimable, $\vec{c}$ debe poder escribirse como combinación lineal de los vectores propios de $(X'X)$ que no figuran en (7.15)–(7.19): $\vec{c} = \alpha_1\vec{v}_1 + \dots + \alpha_{p-j}\vec{v}_{p-j}$, y toda forma estimable $\vec{c}\,'\vec{\beta}$ debe poder expresarse así:

$$\vec{c}\,'\vec{\beta} = (\alpha_1\vec{v}_1 + \dots + \alpha_{p-j}\vec{v}_{p-j})'\vec{\beta} \tag{7.20}$$

## 7.3. Varianza en la estimación de una forma lineal.

Si premultiplicamos ambos lados de las ecuaciones normales $(X'X)\hat{\beta} = X'\vec{Y}$ por $\vec{v}_i$, $(i = 1, \dots, p-j)$, tenemos:

$$\vec{v}_i'(X'X)\hat{\beta} = \vec{v}_i'X'\vec{Y} \tag{7.21}$$
$$\lambda_i\vec{v}_i'\hat{\beta} = \vec{v}_i'X'\vec{Y} \tag{7.22}$$

y tomando varianzas a ambos lados:

$$\begin{aligned}
\lambda_i^2 \text{var}(\vec{v}_i{}'\hat{\beta}) &= \text{var}(\vec{v}_i{}'X'\vec{Y}\,) && (7.23)\\
&= \vec{v}_i{}'X'\sigma^2 I X\vec{v}_i && (7.24)\\
&= \vec{v}_i{}'X'X\vec{v}_i\sigma^2 && (7.25)\\
&= \lambda_i\sigma^2 && (7.26)
\end{aligned}$$

De la igualdad (7.26) se deduce que:

$$\text{var}(\vec{v}_i{}'\hat{\beta}) = \frac{\sigma^2}{\lambda_i} \tag{7.27}$$

Además, para cualquier $i \neq j$ se tiene:

$$\begin{aligned}
\text{cov}(\vec{v}_i{}'\hat{\beta}, \vec{v}_j{}'\hat{\beta}) &= \vec{v}_i{}'\Sigma_{\hat{\beta}}\vec{v}_j && (7.28)\\
&= \vec{v}_i{}'((X'X))^{-1}\vec{v}_j\sigma^2 && (7.29)\\
&= \vec{v}_i{}'\lambda_j{}^{-1}\vec{v}_j\sigma^2 && (7.30)\\
&= \sigma^2\lambda_j{}^{-1}\vec{v}_i{}'\vec{v}_j && (7.31)\\
&= 0 && (7.32)
\end{aligned}$$

La varianza de cualquier forma estimable $\vec{c}\,'\vec{\beta}$, teniendo en cuenta que puede escribirse como en (7.20), y haciendo uso de (7.27) y (7.32), será:

$$\begin{aligned}
\text{var}(\vec{c}\,'\hat{\beta}) &= \text{var}[(\alpha_1\vec{v}_1 + \cdots + \alpha_{p-j}\vec{v}_{p-j})'\hat{\beta}] && (7.33)\\
&= \alpha_1^2\text{var}(\vec{v}_1{}'\hat{\beta}) + \cdots + \alpha_{p-j}^2\text{var}(\vec{v}_{p-j}{}'\hat{\beta}) && (7.34)\\
&= \alpha_1^2\left[\frac{\sigma^2}{\lambda_1}\right] + \cdots + \alpha_{p-j}^2\left[\frac{\sigma^2}{\lambda_{p-j}}\right] && (7.35)\\
&= \sigma^2\left[\frac{\alpha_1^2}{\lambda_1} + \cdots + \frac{\alpha_{p-j}^2}{\lambda_{p-j}}\right] && (7.36)
\end{aligned}$$

La expresión (7.36) es reveladora; la varianza en la estimación de $\vec{c}\,'\vec{\beta}$ dependerá de la varianza de la perturbación $\sigma^2$ y de la dirección de $\vec{c}$. Si $\vec{c}$ no puede expresarse como combinación lineal de los vectores propios con valor propio no nulo, $\vec{c}\,'\vec{\beta}$ no es estimable. Si $\vec{c} = \alpha_1\vec{v}_1 + \cdots + \alpha_{p-j}\vec{v}_{p-j}$ y los $\alpha's$ multiplicando a vectores propios con reducido valor propio son sustanciales, los correspondientes sumandos tenderán a dominar la expresión (7.36).

En definitiva, la varianza en la estimación de una forma lineal $\vec{c}\,'\vec{\beta}$ depende, fundamentalmente, de cuán colineal es $\vec{c}$ con vectores propios de reducido valor propio.

## 7.4. Elección óptima de observaciones adicionales*.

La expresión (7.36) y comentario posterior muestran que, para guarecernos de varianzas muy grandes en la estimación de algunas formas lineales, debemos actuar sobre los valores propios más pequeños de $(X'X)$, incrementándolos[3]. En lo que sigue, examinamos esta cuestión con más detalle.

[3]O suprimiéndolos. Los métodos de regresión sesgada de la Sección siguiente hacen explícita esta idea.

Supongamos que tenemos un conjunto de $N$ observaciones $(\vec{y} \mid X)$, y nos planteamos ampliar $X$ con una fila adicional $\vec{x}_{N+1}{}'$ (e $\vec{y}$ con el correspondiente valor observado de $Y$) de modo que se reduzca al máximo la varianza en la estimación de una determinada forma lineal $\vec{c}\,'\vec{\beta}$ en que estamos interesados. Emplearemos los subíndices $N+1$ y $N$ para designar estimaciones con y sin esta observación adicional. Tenemos entonces que:

$$\Sigma_{\hat{\beta}_N} = \sigma^2((X'X))^{-1} \tag{7.37}$$
$$\Sigma_{\hat{\beta}_{N+1}} = \sigma^2(X'X + \vec{x}_{N+1}\vec{x}_{N+1}{}')^{-1} \tag{7.38}$$
$$\sigma^2_{\vec{c}\,'\hat{\beta}_N} = \sigma^2\vec{c}\,'((X'X))^{-1}\vec{c} \tag{7.39}$$
$$\sigma^2_{\vec{c}\,'\hat{\beta}_{N+1}} = \sigma^2\vec{c}\,'(X'X + \vec{x}_{N+1}\vec{x}_{N+1}{}')^{-1}\vec{c} \tag{7.40}$$

Entonces,

$$\sigma^2_{\vec{c}\,'\hat{\beta}_N} - \sigma^2_{\vec{c}\,'\hat{\beta}_{N+1}} = \sigma^2\vec{c}\,'[((X'X))^{-1} - (X'X + \vec{x}_{N+1}\vec{x}_{N+1}{}')^{-1}]\vec{c} \tag{7.41}$$

y el problema es encontrar $\vec{x}_{N+1}$ maximizando esta expresión. Sea $V$ la matriz que diagonaliza a $(X'X)$. Denominemos:

$$\vec{a} = V'\vec{c} \tag{7.42}$$
$$\vec{z} = V'\vec{x}_{N+1} \tag{7.43}$$
$$D = V'(X'X)V \tag{7.44}$$

Entonces, (7.41) puede transformarse así:

$$\sigma^2_{\vec{c}\,'\hat{\beta}_N} - \sigma^2_{\vec{c}\,'\hat{\beta}_{N+1}} = \sigma^2\vec{c}\,'VV'[((X'X))^{-1} - (X'X + \vec{x}_{N+1}\vec{x}_{N+1}{}')^{-1}]VV'\vec{c} \tag{7.45}$$
$$= \sigma^2\vec{a}\,'[D^{-1} - V'(X'X + \vec{x}_{N+1}\vec{x}_{N+1}{}')^{-1}V]\vec{a} \tag{7.46}$$
$$= \sigma^2\vec{a}\,'[D^{-1} - (V'(X'X + \vec{x}_{N+1}\vec{x}_{N+1}{}')V)^{-1}]\vec{a} \tag{7.47}$$
$$= \sigma^2\vec{a}\,'[D^{-1} - (D + \vec{z}\,\vec{z}\,')^{-1}]\vec{a} \tag{7.48}$$

Pero (véase Teorema A.2, pág. 179):

$$(D + \vec{z}\,\vec{z}\,')^{-1} = D^{-1} - \frac{D^{-1}\vec{z}\,\vec{z}\,'D^{-1}}{1 + \vec{z}\,'D^{-1}\vec{z}} \tag{7.49}$$

Sustituyendo (7.49) en (7.48):

$$\sigma^2_{\vec{c}\,'\hat{\beta}_N} - \sigma^2_{\vec{c}\,'\hat{\beta}_{N+1}} = \sigma^2\vec{a}\,'\left[\frac{D^{-1}\vec{z}\,\vec{z}\,'D^{-1}}{1 + \vec{z}\,'D^{-1}\vec{z}}\right]\vec{a} \tag{7.50}$$
$$= \sigma^2\frac{\left(\sum_i \frac{a_i z_i}{\lambda_i}\right)^2}{\left(1 + \sum_i \frac{z_i^2}{\lambda_i}\right)} \tag{7.51}$$

Obsérvese que el problema de maximizar (7.41) carece de sentido si no imponemos restricciones, pues la expresión equivalente (7.51) es monótona creciente al multiplicar

$\vec{z}$ por una constante $k > 1$. Necesitamos una restricción del tipo $\vec{z}\,'\vec{z} = \sum_i z_i^2 = K^2$ para obtener una solución única. Formando entonces el lagrangiano,

$$\Phi(\vec{z}) = \sigma^2 \frac{\left(\sum_i \frac{a_i z_i}{\lambda_i}\right)^2}{\left(1 + \sum_i \frac{z_i^2}{\lambda_i}\right)} - \mu\left(\sum_i z_i^2 - K^2\right) \tag{7.52}$$

y derivando respecto a $z_i$, $(i = 1, \ldots, p)$, obtenemos $p$ igualdades de la forma:

$$\sigma^2 \frac{\left(\sum_i \frac{a_i z_i}{\lambda_i}\right)\frac{a_i}{\lambda_i}\left(1 + \sum_i \frac{z_i^2}{\lambda_i}\right) - \left(\sum_i \frac{a_i z_i}{\lambda_i}\right)^2 \frac{z_i}{\lambda_i}}{\left(1 + \sum_i \frac{z_i^2}{\lambda_i}\right)^2} - \mu z_i = 0 \tag{7.53}$$

Denominando:

$$A = \left(\sum_i \frac{a_i z_i}{\lambda_i}\right) \tag{7.54}$$

$$B = \left(1 + \sum_i \frac{z_i^2}{\lambda_i}\right) \tag{7.55}$$

las $p$ igualdades anteriores toman la forma:

$$\frac{a_i}{\lambda_i}\frac{A}{B} - \frac{z_i}{\lambda_i}\frac{A^2}{B^2} - \frac{\mu z_i}{\sigma^2} = 0 \tag{7.56}$$

Multiplicando por $z_i$ cada una de las anteriores igualdades y sumándolas, puede despejarse:

$$\mu = \frac{A^2}{K^2 B^2}\sigma^2 \tag{7.57}$$

y por consiguiente de (7.56) se obtiene:

$$\frac{a_i}{\lambda_i}\frac{A}{B} - \frac{z_i}{\lambda_i}\frac{A^2}{B^2} - \frac{A^2}{K^2 B^2} z_i = 0 \qquad (i = 1, \ldots, p) \tag{7.58}$$

$$z_i\left(\frac{1}{\lambda_i} + \frac{1}{K^2}\right) = \frac{B}{A}\frac{a_i}{\lambda_i} \qquad (i = 1, \ldots, p) \tag{7.59}$$

o sea:

$$z_i \propto \frac{a_i}{\lambda_i\left(\frac{1}{\lambda_i} + \frac{1}{K^2}\right)} = \frac{a_i}{1 + \frac{\lambda_i}{K^2}} \tag{7.60}$$

para $i = 1, \ldots, p$. Las anteriores $p$ igualdades pueden expresarse en notación matricial así:

$$\vec{z} \propto (I + K^{-2}D)^{-1}\vec{a} \tag{7.61}$$

Por tanto, la fila a añadir a $X$ para mejorar al máximo la estimación de $\vec{c}\,'\vec{\beta}$ será:

$$\vec{x}_{N+1} = V\vec{z} = (I + K^{-2}X'X)^{-1}\vec{c} \tag{7.62}$$

Este resultado se mantiene[4] si $\vec{c}\,'\vec{\beta}$ era inicialmente inestimable y se hace estimable mediante la adición a la matriz de diseño de $\vec{x}_{N+1}$.

Recordemos que hemos obtenido una solución única para $\vec{z}$ solo mediante la imposición de una restricción de escala $\sum_i z_i^2 = K^2$. Es decir, podemos determinar la dirección de $\vec{z}$, pero no su norma. El examen de (7.51) hace evidente que una norma tan grande como sea posible es lo deseable.

Cabe hacer dos comentarios sobre esta última afirmación. El primero, que es lógico que así sea. Si $\sigma^2$ es fija, es claro que siempre preferiremos filas de módulo muy grande, pues si:

$$Y_i = m_i + \epsilon_i = \beta_0 + \cdots + \beta_{p-1}x_{i,p-1} + \epsilon_i \tag{7.63}$$

incrementar el módulo de $\vec{x}_{N+1}$ equivale a incrementar $|m_i|$; y haciendo $|m_i| \gg \epsilon_i$ podemos reducir en términos relativos el peso de $\epsilon_i$ en $y_i$.

En la práctica, sin embargo, hay un límite al valor de $|m_i|$, cuyo crecimiento desaforado podría llevarnos a regiones en las que las $Y_i$ dejan de ser una función aproximadamente lineal de los regresores. Por ejemplo, si el modelo intenta ajustar una constante biológica como función lineal de ciertos tipos de nutrientes, hay un límite práctico a los valores que pueden tomar los regresores: el impuesto por las cantidades que los sujetos bajo estudio pueden ingerir.

En definitiva, el desarrollo anterior suministra la *dirección* en que debe tomarse una observación adicional para mejorar al máximo la varianza en la estimación de $\vec{c}\,'\vec{\beta}$. Tomaremos $x_{N+1}$ tan grande como sea posible en dicha dirección. Si no tuviéramos una forma estimable única como objetivo, una estrategia sensata consistiría en tomar observaciones de forma que se incrementasen los menores valores propios de la matriz $(X'X)$. Podríamos también aceptar como criterio el de maximizar el determinante de $X'X$. Este criterio se conoce como de D-optimalidad[5].

## 7.5. Detección de la multicolinealidad aproximada

Hay algunos indicios y estadísticos que pueden ayudar en el diagnóstico de multicolinealidad.

**Elevado $R^2$ y todos los parámetros no significativos.** La multicolinealidad aproximada se pone de manifiesto en elevadas varianzas de los parámetros estimados que, como consecuencia, son de ordinario no significativos y frecuentemente toman signos contrarios a los previstos.

Una situación típica es aquélla, aparentemente paradójica, en que todos los parámetros en $\vec{\beta}$ son no significativos y sin embargo $R^2$ es muy elevado. ¡Parece que ningún regresor ayuda a ajustar el regresando, y sin embargo todos en conjunto lo hacen muy bien! Ello se debe a que la multicolinealidad no permite deslindar la contribución de cada regresor.

[4] Véase Silvey (1969) para más detalles.

[5] Véase Silvey (1980), una monografía que trata el tema de diseño óptimo.

**Valores propios y número de condición de** $(X'X)$**.** La existencia de relaciones lineales aproximadas entre las columnas de $X$ se traduce en relaciones lineales aproximadas entre las columnas de $(X'X)$ (ver nota al pie de la página 71). Los métodos usuales para examinar el condicionamiento de una matriz en análisis numérico son por tanto de aplicación. En particular, puede recurrirse a calcular los valores propios de la matriz $(X'X)$; uno o mas valores propios muy pequeños (cero, en caso de multicolinealidad perfecta) son indicativos de multicolinealidad aproximada.

A menudo se calcula el "número de condición" de la matriz $(X'X)$, definido como $\lambda_1/\lambda_p$; números de condición "grandes" evidencian gran disparidad entre el mayor y menor valor propio, y consiguientemente multicolinealidad aproximada. Hay que notar, sin embargo, que se trata de un indicador relativo, que, en particular, depende de la escala en que se miden las respectivas columnas de la matriz $X$ —algo perfectamente arbitrario—.

**Factores de incremento de varianza (VIF).** Otra práctica muy usual consiste en regresar cada columna de $X$ sobre las restantes; un $R^2$ muy elevado en una o más de dichas regresiones evidencia una relación lineal aproximada entre la variable tomada como regresando y las tomadas como regresores.

Llamemos $R^2(i)$ al $R^2$ resultante de regresar $\vec{X}_i$ sobre las restantes columnas de $X$. Se define el *factor de incremento de varianza* (variance inflation factor) VIF$(i)$ así:

$$\text{VIF}(i) \stackrel{\text{def}}{=} \frac{1}{1 - R^2(i)}; \tag{7.64}$$

valores de VIF$(i)$ mayores que 10 (equivalentes a $R^2(i) > 0{,}90$) se consideran indicativos de multicolinealidad afectando a $\vec{X}_i$ junto a alguna de las restantes columnas de $X$.

> **Observación 7.1** El nombre de "factores de incremento de varianza" tiene la siguiente motivación. Supongamos que $X$ tiene sus columnas normalizadas de modo que $(X'X)$ es una matriz de correlación (elementos diagonales unitarios). La varianza de $\hat{\beta}_i$ es $\sigma^2(X'X)^{ii}$, en que $(X'X)^{ii}$ denota el elemento en la fila y columna $i$ de la matriz $((X'X))^{-1}$.
>
> Si $X$ tuviera sus columnas ortogonales, $(X'X)$ (y por tanto $((X'X))^{-1}$) serían matrices unidad y Var$(\hat{\beta}_i) = \sigma^2$; por tanto, $(X'X)^{ii}$ recoge el factor en que se modifica en general Var$(\hat{\beta}_i)$ respecto de la situación de mínima multicolinealidad (= regresores ortogonales). Se puede demostrar que $(X'X)^{ii} = (1 - R^2(i))^{-1}$, lo que muestra que se trata precisamente de VIF$(i)$.

# Capítulo 8

# Regresión sesgada.

## 8.1. Introducción.

De acuerdo con el teorema de Gauss-Markov (Teorema 2.2, pág. 18), los estimadores mínimo cuadráticos ordinarios (MCO) son los de varianza mínima en la clase de los estimadores lineales insesgados. Cualesquiera otros que consideremos, si son lineales y de varianza menor, habrán de ser sesgados.

> **Observación 8.1** De ahí la denominación colectiva de métodos de regresión sesgada. Denominaciones alternativas son *regresión regularizada* o métodos de estimación *por encogimiento* ("shrinkage estimators"), está última abarcando un conjunto de estimadores mucho más amplio que el considerado aquí.

Pese a ello, si consideramos adecuado como criterio en la elección de un estimador su ECM (error cuadrático medio) y reparamos en que:

$$
\begin{aligned}
E[\hat{c}-c]^2 &= E\left[\hat{c}-E[\hat{c}]+E[\hat{c}]-c\right]^2 && (8.1)\\
&&& (8.2)\\
&= E\left[\hat{c}-E[\hat{c}]\right]^2 + E\left[E[\hat{c}]-c\right]^2 + 2\underbrace{E\left[\hat{c}-E[\hat{c}]\right]\left[E[\hat{c}]-c\right]}_{=0} && (8.3)\\
&= \mathrm{var}(\hat{c}) + (\text{sesgo } \hat{c})^2 && (8.4)
\end{aligned}
$$

podemos plantearnos la siguiente pregunta: ¿Es posible reducir el ECM en la estimación tolerando un sesgo? Si la respuesta fuera afirmativa, podríamos preferir el estimador resultante que, aunque sesgado, tendría un ECM menor, producido por una disminución en la varianza capaz de compensar el segundo sumando en (8.4).

El Capítulo anterior ponía de manifiesto que vectores propios de $(X'X)$ con valor propio asociado nulo o muy pequeño eran responsables de la inestimabilidad (en el caso extremo de valores propios exactamente cero) o estimación muy imprecisa de formas lineales $\vec{c}\,'\vec{\beta}$ en los parámetros. Si es posible ampliar la matriz de diseño, se decía,

conviene hacerlo de modo crezcan lo más posible los valores propios más pequeños. Analizaremos ahora las implicaciones del análisis realizado.

Si los valores propios pequeños son causantes de elevada varianza en las estimaciones, caben varias soluciones:

1. Incrementarlos mediante observaciones adicionales, según se indica en el Capítulo anterior.
2. Incrementarlos mediante procedimientos "ad-hoc", que no requieren la toma de observaciones adicionales *(ridge regression)*.
3. Prescindir, simplemente, de ellos *(regresión en componentes principales y regresión en raíces latentes)*.

Nos ocuparemos de procedimientos tomando las alternativas 2) y 3) para reducir la varianza de los estimadores. De acuerdo con los comentarios anteriores, los procedimientos que diseñemos habrán perdido la condición de insesgados. Si se utilizan, es con la fundada creencia de que, en presencia de multicolinealidad acusada, la reducción de varianza que se obtiene compensa la introducción de sesgo. Existe incluso un teorema que demuestra la existencia de un estimador sesgado que domina (en términos de ECM) al MCO; su aplicación práctica está limitada por el hecho de que no es inmediato saber *cuál* precisamente es este estimador.

## 8.2. Regresión ridge.

### 8.2.1. Error cuadrático medio del estimador mínimo cuadrático ordinario

Dado que hay varios parámetros a estimar, definiremos como ECM del estimador MCO:

$$\text{ECM}(\hat{\beta}) = E[(\hat{\beta} - \vec{\beta})'(\hat{\beta} - \vec{\beta})] \tag{8.5}$$

que podemos ver también como el valor medio del cuadrado de la distancia euclídea ordinaria entre $\hat{\beta}$ y $\vec{\beta}$. Como $E[\hat{\beta}] = \vec{\beta}$ y $\Sigma_{\hat{\beta}} = \sigma^2((X'X))^{-1}$, tenemos que:

$$\begin{aligned}
\text{ECM}(\hat{\beta}) &= E[\text{traza } (\hat{\beta} - \vec{\beta})'(\hat{\beta} - \vec{\beta})] \\
&= E[\text{traza } (\hat{\beta} - \vec{\beta})(\hat{\beta} - \vec{\beta})'] \\
&= \sigma^2 \text{traza } ((X'X))^{-1} \\
&= \sigma^2 \text{traza } ((X'X))^{-1}VV' \qquad (V = \text{diagonalizadora de } ((X'X))^{-1}) \\
&= \sigma^2 \text{traza } V'((X'X))^{-1}V \\
&= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}
\end{aligned} \tag{8.6}$$

### 8.2.2. Clase de estimadores ridge

**Definición 8.1** *Definiremos el estimador ridge*[1] *de parámetro $k$ así:*

$$\hat{\beta}^{(k)} = ((X'X) + kI)^{-1}X'\vec{Y} \tag{8.7}$$

[1] Véase Hoerl and Kennard (1970), o cualquiera de los manuales citados en la Bibliografía.

*siendo $k$ una constante positiva a determinar.*

El estimador ridge es idéntico al MCO en el caso particular en que $k = 0$. La relación entre ambos para un valor arbitrario de $k$ queda de manifiesto en la siguiente cadena de igualdades:

$$\begin{aligned}
\hat{\beta}^{(k)} &= ((X'X)+kI)^{-1}(X'X)((X'X))^{-1}X'\vec{Y} && (8.8)\\
&= ((X'X)+kI)^{-1}(X'X)\hat{\beta} && (8.9)\\
&= \left[((X'X))^{-1}((X'X)+kI)\right]^{-1}\hat{\beta} && (8.10)\\
&= \left[I+k(X'X)^{-1}\right]^{-1}\hat{\beta} && (8.11)\\
&= Z\hat{\beta} && (8.12)
\end{aligned}$$

**Lema 8.1** *El error cuadrático medio del estimador ridge de parámetro $k$ viene dado por la expresión*

$$ECM[\hat{\beta}^{(k)}] = \sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^2}+\sum_{i=1}^{p}\frac{k^2\alpha_i^2}{(\lambda_i+k)^2} \quad (8.13)$$

*en que los $\lambda_i$ son los valores propios de la matrix $(X'X)$ y $\vec{\alpha} = V'\vec{\beta}$, siendo $V$ una matriz cuyas columnas son vectores propios de $(X'X)$.*

DEMOSTRACION:

De acuerdo con (8.12), el ECM del estimador ridge que habremos de comparar con (8.6) es:

$$\begin{aligned}
E[(\hat{\beta}^{(k)}-\vec{\beta})'(\hat{\beta}^{(k)}-\vec{\beta})] &= E[(Z\hat{\beta}-\vec{\beta})'(Z\hat{\beta}-\vec{\beta})]\\
&= E[(Z\hat{\beta}-Z\vec{\beta}+Z\vec{\beta}-\vec{\beta})'(Z\hat{\beta}-Z\vec{\beta}+Z\vec{\beta}-\vec{\beta})]\\
&= \underbrace{\sigma^2\text{traza}\,[((X'X))^{-1}Z'Z]}_{(a)}\\
&\quad+\underbrace{\vec{\beta}'(Z-I)'(Z-I)\vec{\beta}}_{(b)} \qquad (8.14)
\end{aligned}$$

Examinemos por separado los dos sumandos de (8.14).

$$\begin{aligned}
(a) &= \sigma^2\text{traza}\left[((X'X))^{-1}\left[I+k(X'X)^{-1}\right]^{-1}\left[I+k(X'X)^{-1}\right]^{-1}\right]\\
&= \sigma^2\text{traza}\left[(X'X)+kI+kI+k^2((X'X))^{-1}\right]^{-1}\\
&= \sigma^2\text{traza}\left\{\left[(X'X)+2kI+k^2((X'X))^{-1}\right]^{-1}VV'\right\}\\
&= \sigma^2\text{traza}\left[V'[(X'X)+2kI+k^2((X'X))^{-1}]^{-1}V\right] && (8.15)\\
&= \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i+2k+\lambda_i^{-1}k^2} && (8.16)\\
&= \sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^2} && (8.17)
\end{aligned}$$

En el paso de (8.15) a (8.16) se ha empleado el hecho de que si $V$ diagonaliza a $(X'X)$ diagonaliza también a cada una de las matrices en el corchete, y por consiguiente a la matriz inversa de la contenida en el corchete. Por otra parte:

$$\begin{aligned}(b) &= \vec{\beta}'(Z-I)'(Z-I)\vec{\beta} \\ &= \vec{\beta}'\left(\left[I+k(X'X)^{-1}\right]^{-1}-I\right)'\left(\left[I+k(X'X)^{-1}\right]^{-1}-I\right)\vec{\beta} \\ &= k^2\vec{\alpha}'(\Lambda+kI)^{-2}\vec{\alpha} \qquad (\text{siendo } \vec{\alpha}=V'\vec{\beta}) \\ &= \text{traza}\ \left[k^2\vec{\alpha}'(\Lambda+kI)^{-2}\vec{\alpha}\right] \\ &= \sum_{i=1}^{p}\frac{k^2\alpha_i^2}{(\lambda_i+k)^2} \end{aligned} \tag{8.18}$$

Por consiguiente, el ECM del estimador $\hat{\beta}^{(k)}$ es, de acuerdo con (8.14), (8.17), y (8.18):

$$ECM[\hat{\beta}^{(k)}] = \sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^2}+\sum_{i=1}^{p}\frac{k^2\alpha_i^2}{(\lambda_i+k)^2} \tag{8.19}$$

■

**Teorema 8.1** *Hay algún valor de $k>0$ para el que (8.19) es estrictamente menor que el ECM del estimador MCO dado por (8.6).*

DEMOSTRACION:

Hemos visto más arriba que cuando $k=0$, el estimador ridge $\hat{\beta}^{(k)}$ coincide con el MCO. Por consiguiente, para $k=0$ la expresión (8.19) debe coincidir con (8.6), como en efecto puede comprobarse que sucede. Derivando (8.19) respecto de $k$, es fácil comprobar que la derivada en $k=0$ existe y es $-2\sigma^2\sum_{i=1}^{p}\lambda_i^{-2}$, claramente negativa. Por consiguiente, siempre podremos (incrementando ligeramente $k$) lograr que:

$$ECM[\hat{\beta}^{(k)}] < ECM[\hat{\beta}^{(0)}] = ECM[\hat{\beta}] \tag{8.20}$$

lo que demuestra el teorema.

■

Una percepción intuitiva del resultado anterior puede lograrse comparando las expresiones (8.6) y (8.14), valores medios respectivamente de $(\hat{\beta}-\vec{\beta})'(\hat{\beta}-\vec{\beta})$ y $(\hat{\beta}^{(k)}-\vec{\beta})'(\hat{\beta}^{(k)}-\vec{\beta})$. Se observa que (8.6) puede hacerse arbitrariamente grande si $\lambda_i\approx 0$ para algún $i$. La expresión (8.17) está a cobijo de tal eventualidad, pues ninguno de los sumandos puede crecer por encima de $\lambda_i/k^2$. La definición (8.7) puede verse como una manipulación de la diagonal principal de $(X'X)$ tendente a incrementar cada valor propio en $k$. No deja de ser sorprendente que una manipulación tan trivial tenga la potencialidad de reducir el ECM.

### 8.2.3. Elección de $k$

Sabemos que existe un $k$ (de hecho, un intervalo de valores de $k$ mejorando el ECM del estimador MCO; pero nada en la discusión anterior nos permite decidir cuál es su valor. En la práctica, se recurre a alguna o varias de las siguientes soluciones:

**Uso de trazas ridge.** Se prueban diversos valores representándose las diferentes estimaciones del vector $\vec{\beta}$ *(trazas ridge)*; se retiene entonces aquel valor de $k$ a partir del cual se estabilizan las estimaciones.

La idea es intuitivamente atrayente: pequeños incrementos de $k$ partiendo de cero tienen habitualmente un efecto drástico sobre $\vec{\beta}$, al coste de introducir algún sesgo. Incrementaremos $k$ por tanto hasta que parezca que su influencia sobre $\vec{\beta}$ se atenúa —hasta que las trazas ridge sean casi horizontales. El decidir dónde ocurre esto es no obstante bastante subjetivo.

**Elección de $k$ por validación cruzada.** La idea es también muy simple, aunque computacionalmente algo laboriosa. Sea $\hat{y}_{(i),k}$ la predicción que hacemos de la observación $y_i$ cuando empleamos el estimador ridge de parámetro $k$ *obtenido con una muestra de la que excluimos la observación $i$-ésima*. Definamos

$$CV(k) \quad = \quad \sum_{i=1}^{N} (y_i - \hat{y}_{(i),k})^2;$$

es decir, $CV(k)$ es la suma de cuadrados de los residuos obtenidos al ajustar cada observación con una regresión que la ha dejado fuera al estimar los parámetros. Entonces,

$$k_{CV} = \arg \min_k CV(k),$$

y la idea es emplear este valor $k_{CV}$. En principio, calcular $CV(k)$ para un valor de $k$ requeriría llevar a cabo $N$ regresiones, excluyendo cada vez una observación distinta. En la práctica, el cálculo puede agilizarse de modo considerable

**Elección de $k$ por validación cruzada generalizada (GCV).** Es un criterio estrechamente emparentado con el anterior. Sean

$$\begin{aligned} A(k) &= X((X'X) + kI)^{-1}X' \\ \hat{y} &= X\hat{\beta}^{(k)} = A(k)\vec{y}; \end{aligned}$$

entonces, elegimos

$$k_{GCV} \quad = \quad \arg \min_k \frac{||(I - A(k))\vec{y}||^2}{[\text{traza}(I - A(k))]^2}. \tag{8.21}$$

Sobre la justificación de dicha elección puede verse Eubank (1988) o Brown (1993), por ejemplo; no podemos entrar aquí en detalles. Baste decir que (8.21) se reduce a $SSE/(N-p)^2$ cuando $k = 0$ (mínimos cuadrados ordinarios), como resulta inmediato de la definición de $A(k)$; una expresión cuya minimización parece razonable. Para otros valores de $k$ el numerador de (8.21) continúa siendo una suma de cuadrados de los residuos y el denominador al cuadrado del número de *grados de libertad equivalentes*.

**Otros criterios.** Nos limitamos a mencionarlos. Detalles adicionales pueden encontrarse en Brown (1993) o en los trabajos originales de sus respectivos proponentes.

$$k_{HKB} = (r-2)\hat{\sigma}^2/\hat{\beta}'\hat{\beta} \quad (8.22)$$

$$k_{LW} = (r-2)\hat{\sigma}^2 \text{traza}(X'X)/(r\hat{\beta}'(X'X)\hat{\beta}) \quad (8.23)$$

$$k_{MUR} = \arg\min_k \left[\hat{\sigma}^2 \sum_i \frac{\lambda_i - k}{\lambda_i(\lambda_i + k)} + k^2 \sum_i \frac{\hat{\alpha}_i^2}{(\lambda_i + k)^2}\right] \quad (8.24)$$

El criterio (8.22) fue propuesto por Hoerl et al. (1975) y tiene una justificación bayesiana. El criterio (8.23) fue propuesto en Lawless and Wang (1976). El criterio (8.24) estima el ECM del estimador ridge insesgadamente y toma el $k$ que minimiza dicha estimación.

### 8.2.4. Comentarios adicionales

Es evidente que la forma del ECM propuesto pondera por igual las discrepancias en la estimación de un $\beta_i$ cuyo valor real es muy grande que aquéllas en la estimación de uno cuyo valor real es muy pequeño. Por ello, es aconsejable antes de emplear el procedimiento normalizar los regresores. Alternativamente podría reproducirse el desarrollo anterior empleando como ECM una expresión del tipo: $(\hat{\beta} - \vec{\beta})'M(\hat{\beta} - \vec{\beta})$, siendo $M$ una matriz definida positiva adecuada[2].

Finalmente, es habitual no sólo normalizar sino también centrar tanto las columnas de $X$ como $\vec{y}$. El parámetro $\beta_0$ se sustrae así al proceso de estimación ridge, restaurándolo al final.

**R: Ejemplo 8.1** *(ejemplo de regresión ridge)*

El siguiente código muestra el uso de regresión ridge sobre un conjunto de datos acusadamente colineal. La Figura 8.1 muestra las trazas ridge de los seis parámetros estimados y el valor del criterio GCV para distintos valores de $k$. En ambas gráficas, que comparten la escala de abscisas, se ha trazado una recta vertical al nivel de $k_{GCV}$. Los valores de $k_{HKB}$ y $k_{LW}$ son también output de la función lm.ridge y podrían haberse utilizado. El primero es prácticamente idéntico a $k_{GCV}$ y no se ha representado en la Figura 8.1; el segundo sí.

```
--- Obtenido mediante R BATCH demo8.R
> #
> #  La biblioteca MASS contiene una función para hacer regresión
> #  ridge de manera fácil y cómoda.
> #
> options(digits=4)
> options(columns=40)
> library(MASS)

Attaching package: 'MASS'


The following object(s) are masked _by_ .GlobalEnv :

 UScrime
```

[2]Es decir, empleando una métrica distinta de la euclídea ordinaria para medir la discrepancia entre $\hat{\beta}$ y $\vec{\beta}$; $M = (X'X)$ sería una elección natural.

```
> data(longley)                  #  datos con acusada
> names(longley)[1] <- "y"
>                                #  multicolinealidad
> longley[1:3,]
        y   GNP Unemployed Armed.Forces Population Year Employed
1947 83.0 234.3      235.6        159.0      107.6 1947    60.32
1948 88.5 259.4      232.5        145.6      108.6 1948    61.12
1949 88.2 258.1      368.2        161.6      109.8 1949    60.17
>
> lm(y ~ ., longley)             #  MCO

Call:
lm(formula = y ~ ., data = longley)

Coefficients:
 (Intercept)           GNP    Unemployed  Armed.Forces    Population
   2946.8564        0.2635        0.0365        0.0112       -1.7370
        Year      Employed
     -1.4188        0.2313

> lm.ridge(y ~ ., longley)       #  Por omisión de lambda, MCO
                     GNP   Unemployed Armed.Forces   Population         Year
  2946.85636     0.26353      0.03648     0.01116     -1.73703     -1.41880
    Employed
     0.23129
> #
> #  Todas las regresiones ridge para lambda desde 0 a 0.1 en
> #  incrementos de 0.0001
> #
> longley.rr <- lm.ridge(y ~ ., longley,
+                lambda = seq(0,0.1,0.001))
> summary(longley.rr)
       Length Class  Mode
coef   606    -none- numeric
scales   6    -none- numeric
Inter    1    -none- numeric
lambda 101    -none- numeric
ym       1    -none- numeric
xm       6    -none- numeric
GCV    101    -none- numeric
kHKB     1    -none- numeric
kLW      1    -none- numeric
> #
> #  Proporciona lambda óptimo según tres diferentes criterios.
> #
> select(longley.rr)
modified HKB estimator is 0.006837
modified L-W estimator is 0.05267
smallest value of GCV  at 0.006
> #
> #  Lugar que ocupa el lambda que minimiza GCV
> #
> nGCV <- order(longley.rr$GCV)[1]
```

```
> lGCV <- longley.rr$lambda[nGCV]   #  Lambda minimizador.
> #
> #  Hacemos ahora regresión ridge con el lambda seleccionado.
> #
> lm.ridge(y ~ ., longley,lambda=lGCV)
                      GNP   Unemployed Armed.Forces   Population         Year
  -3.144e+02    1.765e-01    1.937e-02    6.565e-03   -1.328e+00    2.556e-01
    Employed
  -5.812e-02
>
> postscript(file="demo8.eps",horizontal=FALSE,
+           width=5,height=9)
> par(mfrow=c(2,1))
> matplot(longley.rr$lambda,
+         t(longley.rr$coef),type="l",
+         xlab=expression(k),
+         ylab=expression(beta[i]))  #  Trazas ridge; podríamos
>                                    #  usar plot(longley.rr)
> abline(v=lGCV)
> mtext(expression(k[GCV]),side=3,
+       at=lGCV)
> title(main="Trazas ridge")
> plot(longley.rr$lambda,
+      longley.rr$GCV,type="l",
+      xlab=expression(k),ylab="GCV",
+      main="Criterio GCV")          #  GCV; forma típica
> abline(v=lGCV)
> mtext(expression(k[GCV]),side=3,
+       at=lGCV)
> abline(v=longley.rr$kLW)
> mtext(expression(k[LW]),side=3,
+       at=longley.rr$kLW)
```

## 8.3. Regresión en componentes principales.

### 8.3.1. Descripción del estimador

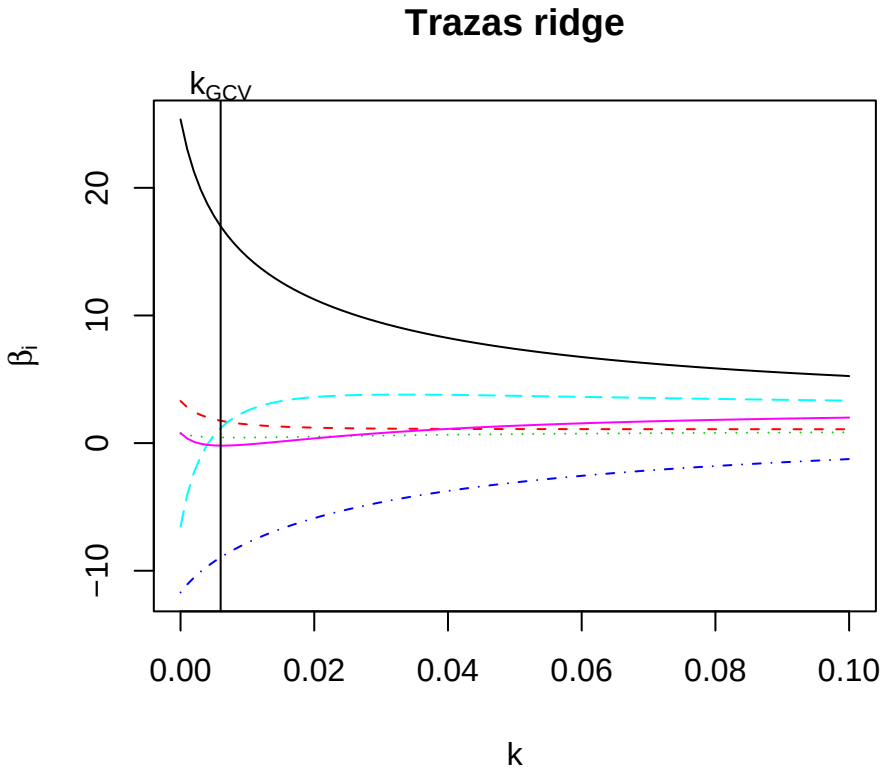
Consideraremos, por conveniencia notacional, el modelo habitual en que la columna de "unos", si existe, ha sido segregada, y los restantes regresores han sido centrados y normalizados. Esto tiene por único efecto multiplicar los parámetros —y sus estimadores— por constantes respectivamente iguales a la norma de las columnas de $X$ afectadas. Con este convenio, el modelo de regresion lineal que consideramos se puede escribir así:

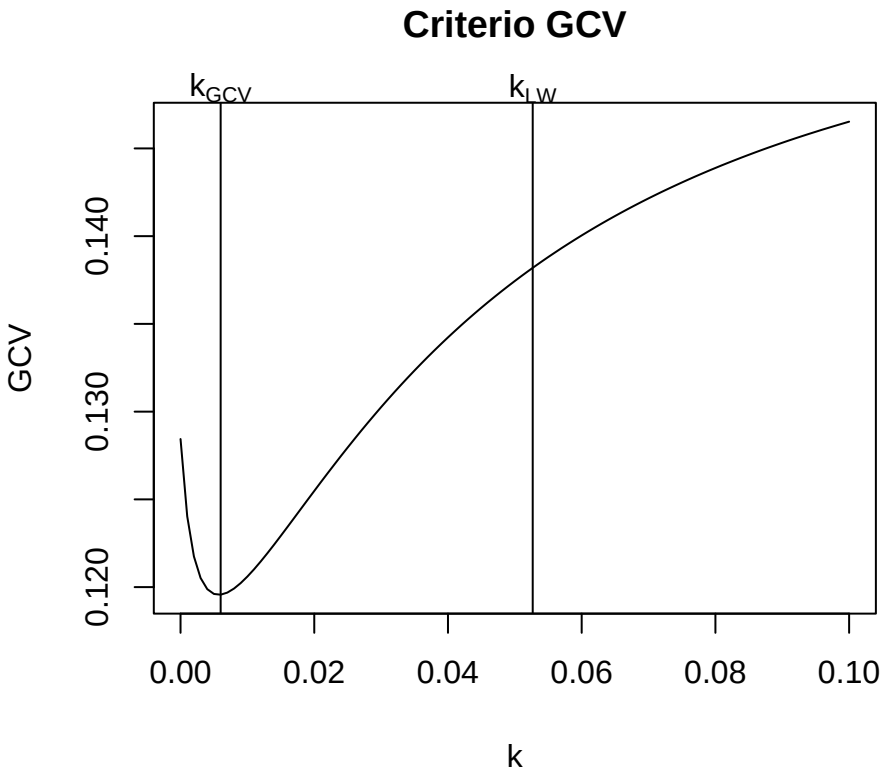
$$\vec{y} = \vec{1}\beta_0 + W\vec{\beta}^* + \vec{\epsilon} \tag{8.25}$$

Supondremos, consistentemente con la notación anterior, que $\vec{\beta}^*$ es un vector $(p-1)\times 1$, y $W$ una matriz $N \times (p-1)$. La matriz $W'W$ es una matriz con "unos" en la diagonal principal, simétrica, y definida no negativa. Existe siempre una diagonalizadora ortogonal $V$ tal que:

$$V'(W'W)V = \Lambda \qquad (\Longleftrightarrow W'W = V\Lambda V') \tag{8.26}$$

Figura 8.1: Trazas ridge y GVC para los datos `longley`



Sean $\vec{v}_1, \ldots, \vec{v}_{p-1}$ los vectores columna de $V$. Llamaremos *componentes principales* de $W$ a los vectores $\vec{u}_1, \ldots, \vec{u}_{p-1}$ definidos así:

$$\begin{aligned} \vec{u}_1 &= W\vec{v}_1 \\ \vec{u}_2 &= W\vec{v}_2 \\ &\vdots \\ \vec{u}_{p-1} &= W\vec{v}_{p-1} \end{aligned} \tag{8.27}$$

o abreviadamente:

$$U = WV \tag{8.28}$$

La matriz $U$ es $N \times (p-1)$, con columnas combinación lineal de las de $W$. Es además aparente que las columnas de $U$ son ortogonales: $U'U = V'(W'W)V = \Lambda$, y que generan el mismo subespacio de $R^N$ que las de $W$.

Siendo $V$ ortogonal, (8.25) puede transformarse así:

$$\vec{y} = \vec{1}\beta_0 + W\vec{\beta}^* + \vec{\epsilon} \tag{8.29}$$

$$= \vec{1}\beta_0 + WVV'\vec{\beta}^* + \vec{\epsilon} \tag{8.30}$$

$$= \vec{1}\beta_0 + U\vec{\gamma}^* + \vec{\epsilon} \tag{8.31}$$

Teniendo en cuenta (ver Problema 8.2) que $\vec{1} \perp \vec{u}_i$, $(i = 1, \ldots, p-1)$, el vector de estimadores puede escribirse así:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\gamma}^* \end{pmatrix} = \begin{pmatrix} \overline{\vec{y}} \\ (U'U)^{-1}U'\vec{y} \end{pmatrix} = \begin{pmatrix} \overline{\vec{y}} \\ \Lambda^{-1}U'\vec{y} \end{pmatrix} \tag{8.32}$$

Todo lo que hemos hecho hasta el momento es tomar una diferente base del espacio de proyección —la formada por las columnas de $U$ en lugar de la formada por las columnas de $W$—. Llegados a este punto, tenemos que recuperar los parámetros originales $\vec{\beta}^*$ a partir de los $\hat{\gamma}^*$. Si lo hacemos mediante

$$\hat{\beta}^* = V\hat{\gamma}^*$$

estaremos obteniendo exactamente los estimadores MCO. La idea del estimador en componentes principales $\hat{\beta}^*_{CP}$ es emplear sólo algunos de los términos en $\hat{\gamma}^*$:

$$\hat{\beta}^*_{CP} = V \begin{pmatrix} \hat{\gamma}^*_{(q)} \\ \vec{0} \end{pmatrix}. \tag{8.33}$$

Necesitamos por tanto criterios para escoger los parámetros $\hat{\gamma}_i$ que incluirmos en $\hat{\gamma}^*_{(q)}$ y los que reemplazamos por cero en (8.33).

### 8.3.2. Estrategias de selección de componentes principales

Hay varias estrategias. Una discusión más pormenorizada que el resumen a continuación puede encontrarse en Brown (1993) o en Jolliffe (1986).

**Elección basada en $\lambda_i$.** Como quiera que la varianza de $\hat{\gamma}_i$ es $\sigma^2\lambda_i^{-1}$ (véase (7.27), pág. 74), una estrategia consistiría en tomar los $\hat{\gamma}^*_i$ asociados a $\lambda_i$ más grande (es decir, con menos varianza), despreciando los restantes. El número de componentes principales a retener (= el número de $\lambda_i$'s "grandes") es en buena medida subjetivo.

Nótese que puede ocurrir que componentes asociadas a parámetros $\hat{\gamma}^*_i$ con mucha varianza —y por tanto desechados— tengan no obstante gran poder predictivo de $\vec{y}$. En este caso, podría ser preferible emplear la estrategia a continuación.

**Elección basada en el contraste de nulidad de los $\hat{\gamma}_i^*$.** Se procede así:

1. Se calcula $||P_U\vec{y}||^2 = ||U\hat{\gamma}^*||^2 = \hat{\gamma}_1^{*2}||\vec{u}_1\,||^2 + \cdots + \hat{\gamma}_{p-1}^{*2}||\vec{u}_{p-1}\,||^2$, esta última igualdad haciendo uso de la ortogonalidad entre las columnas de $U$. Entonces, $SSR = ||P_U\vec{y}||^2$, y $SSE = ||\vec{y} - \vec{\overline{y}}||^2 - ||U\hat{\gamma}^*||^2$.

2. Se contrasta la hipótesis de nulidad para cada uno de los parámetros, $(H_i\colon \gamma_i = 0,\ \ i = 1, \ldots, p-1)$, mediante el estadístico:

$$Q_i = \frac{N-p}{1} \times \frac{\hat{\gamma}_i^{*2}||\vec{u}_i\,||^2}{SSE} \sim \mathcal{F}_{1,N-p} \tag{8.34}$$

   que sigue la distribución indicada bajo los supuestos habituales más normalidad cuando $H_i$ es cierta. Obsérvese que, gracias a ser ortogonales las columnas de $U$, la fracción de $SSR$ atribuible a cada regresor es independiente de los que pueda haber ya incluidos en la ecuación de regresión.

3. Se introducen todos los regresores cuyo estadístico $Q_i$ supere un nivel prefijado. Sin pérdida de generalidad, supondremos que éstos son los $q$ primeros, formando el vector $\hat{\gamma}_{(q)}^*$.

4. Los $\hat{\beta}$ se obtienen mediante la transformación (8.33).

Nótese que mientras que la estrategia precedente consistía en desechar componentes principales asociadas a reducido $\lambda_i$, la presente propone desechar las asociadas a reducido $Q_i$; frecuentemente, no suele haber conflicto entre ambos objetivos: $||\vec{u}_i||^2 = \lambda_i \approx 0 \Rightarrow Q_i \approx 0$ a menos que simultáneamente $\gamma_i \gg 0$. Puede ocurrir, sin embargo, que una componente principal asociada a un $\lambda_i$ muy pequeño tenga apreciable valor predictivo (si $\hat{\gamma}_i$ es grande). Procedería incluir dicha componente principal como predictor si el valor de $Q_i$ lo justifica y la predicción es el objetivo del análisis[3].

**Estrategia mixta.** Propuesta por Jolliffe (1986), ordena los $\hat{\gamma}_i^*$ de menor a mayor $\lambda_i$ y realiza *en este orden* un contraste como el del apartado anterior sobre cada uno de ellos. Cuando se encuentra el primer $\hat{\gamma}_i^*$ significativo, se retiene junto a todos los que le siguen (con $\lambda_i$ mayor, por tanto). Todos los $\hat{\gamma}_i^*$ retenidos componen el vector $\hat{\gamma}^*$.

**Validación cruzada.** Computacionalmente muy laboriosa. Puede ocurrir que al omitir distintas observaciones, dos componentes principales permuten su orden. Véanse detalles en Brown (1993).

### 8.3.3. Propiedades del estimador en componentes principales

El sesgo de $\hat{\beta}_{CP}^*$ es:

$$E[\hat{\beta}_{CP}^* - \vec{\beta}^*] \;\; = \;\; E\left[V\begin{pmatrix}\hat{\gamma}_{(q)}^* \\ \vec{0}\end{pmatrix} - V\vec{\gamma}^{\,*}\right] = -\sum_{i=q+1}^{p-1}\gamma_i\vec{v}_i \tag{8.35}$$

[3]Pero este criterio no es unánimemente compartido. Véase Hocking (1976).

y su matriz de covarianzas:

$$\Sigma_{\hat{\beta}^*_{CP}} = V \left( \sigma^2 \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} \Lambda^{-1} \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} \right) V' \tag{8.36}$$

$$= \sigma^2 \sum_{i=1}^{q} \lambda_i^{-1} \vec{v}_i \, \vec{v}_i{}' \tag{8.37}$$

$$\leq \sigma^2 \sum_{i=1}^{p-1} \lambda_i^{-1} \vec{v}_i \, \vec{v}_i{}' \tag{8.38}$$

$$= \sigma^2 (W'W)^{-1} \tag{8.39}$$

en que el símbolo $\leq$ indica elementos no mayores en la diagonal principal. La diferencia entre la matriz de covarianzas de los estimadores MCO y la de los estimadores en componentes principales es:

$$\sigma^2 \sum_{i=q+1}^{p-1} \lambda_i^{-1} \vec{v}_i \, \vec{v}_i{}' \tag{8.40}$$

y será importante si entre las componentes principales excluídas como regresores hay alguna asociada a un $\lambda_i$ muy pequeño.

Las expresiones (8.35) y (8.36)–(8.39) muestran el conflicto varianza-sesgo en el caso de la regresión en componentes principales. De (8.35) se deduce la siguiente expresión para la suma de los sesgos al cuadrado:

$$[E(\hat{\beta}^*_{CP}) - \vec{\beta}]' [E(\hat{\beta}^*_{CP}) - \vec{\beta}] = \sum_{i=q+1}^{p-1} \gamma_i^2 \tag{8.41}$$

Es interesante comparar el resultado anterior con el proporcionado por el estimador ridge, y examinarlo a la luz del análisis efectuado en el Capítulo 7. En realidad, todo cuanto hace el estimador en componentes principales es reparametrizar el modelo, estimarlo por MCO, y obtener los estimadores de los parámetros originales despreciando información (algunos $\hat{\gamma}_i$) de gran varianza (si se sigue el criterio de despreciar sin más componentes principales con pequeño $\lambda_i$) o de reducido $Q_i \propto \gamma_i^{*2} \lambda_i$; este último estadístico puede contemplarse como relación señal/ruido.

El estimador ridge no hace una elección tan drástica sino que, mediante la introducción del parámetro $k$, atenúa las componentes principales responsables en mayor medida de la varianza de $\hat{\beta}$. Esto se hace evidente si comparamos la siguiente expresion:

$$\hat{\beta}^*_{CP} = V \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} \Lambda^{-1} U' \vec{y} = V \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} \hat{\gamma}^* \tag{8.42}$$

con la del estimador ridge equiparable[4]:

$$\hat{\beta}^{(k)} = (W'W + kI)^{-1} W' \vec{y} \tag{8.43}$$

$$= VV'(W'W + kI)^{-1} VV'W'\vec{y} \tag{8.44}$$

$$= V(\Lambda + kI)^{-1} U' \vec{y} \tag{8.45}$$

[4]Es decir, tras haber centrado los regresores y segregado la columna de "unos".

En (8.42) solo $q$ columnas de $U'\vec{y}$ se utilizan; en (8.45), todas, si bien las que corresponden a componentes principales con $\lambda_i$ más pequeño reciben una ponderación menor, al ser divididas por $\lambda_i + k$ en lugar de por $\lambda_i$. Por ejemplo, si $\lambda_1 = 5$, $\lambda_4 = ,002$ y $k = 0{,}01$, la primera columna de $U'\vec{y}$ sería dividida por $5{,}01 \approx 5$, mientras que la cuarta resultaría dividida por $0{,}012 \gg 0{,}002$, es decir, su ponderación se reduciría a la sexta parte. de la original.

**R: Ejemplo 8.2** `--- Obtenido mediante R BATCH demo9.R`

```
> #
> #  La función regCP admite como argumentos la matriz de regresores,
> #  el vector respuesta, y uno de dos argumentos:
> #
> #     tomar :  Vector de índices de las componentes principales a
> #              retener. Por ejemplo, tomar=1:3 tomaría las tres primeras.
> #
> #     sig   :  nivel de significación de las componentes principales a
> #              retener. Se toman todas aquéllas --sea cual fuere su valor
> #              propio asociado-- significativas al nivel sig.
> #
> #  La función es ineficiente, no hace comprobación de errores y tiene
> #  sólo interés didáctico.
> #
> options(digits=4)
> options(columns=30)
> regCP <- function(X,y,tomar=NULL,sig=0.05) {
+
+ X.c <- scale(X,scale=FALSE)               #  datos centrados
+ W   <- scale(X.c,center=FALSE) /
+                  sqrt(nrow(X)-1)          #  datos centrados y normalizados
+ WW  <- crossprod(W)                       #  matriz de momentos
+ factores.escala <- X.c[1,] / W[1,]        #  para restaurar los betas a las
+                                           #  unidades originales
+ N <- nrow(X) ; p <- ncol(X)               #  Núm. observaciones y parámetros.
+ res    <- eigen(WW)
+ V      <- res$vectors                     #  Vectores propios de W'W
+ landas <- res$values                      #  Valores  propios de W'W
+ U      <- W %*% V                         #  Componentes principales
+ gamas  <- (1 / landas) * t(U) %*% y       #  Falla si algún landa == 0.
+
+ if (is.null(tomar)) {                     #  Si no se ha indicado que
+   fit   <- lsfit(X,y)                     #  CP tomar, se contrastan
+   SSE   <- sum(fit$residuals^2)           #  todas al nivel de significación
+   qi    <- (N-p) * (gamas*landas)^2 / SSE #  sig
+   tomar <- (1:p)[qi > (1 - pf(qi,1,N-p))]
+ }
+ betas   <- V[,tomar] %*% gamas[tomar]     #  Los betas obtenidos se corrigen
+ betasCP <- betas / factores.escala        #  con los factores de escala
+
+ m.X <- apply(X,2,mean)                    #  Se calculan las medias de las
+ m.Y <- mean(y)                            #  X y de la y...
+ beta0   <- m.Y - sum(m.X*betasCP)         #  ... y con ellas, beta0.
+ #
+ betasCP <- c(beta0,betasCP)
+ names(betasCP) <- c("Intercept",          #  Rotulado coeficientes, para
```

```
+                   dimnames(X)[[2]])       #  mayor legibilidad.
+ return(list(betasCP=betasCP,landas=landas,
+        CP.usadas=tomar))
+ }
>
> library(MASS)

Attaching package: 'MASS'


The following object(s) are masked _by_ .GlobalEnv :

 UScrime

> data(longley)                             #  datos multicolineales
> y <- longley [,1]                         #  Primera columna es respuesta
> X <- as.matrix(longley[,-1])              #  Resto columnas regresores
> #
> #  Veamos ahora como funciona regCP. Si quisiéramos tomar, por ej.,
> #  tres componentes principales, la invocaríamos así:
> #
> regCP(X,y,tomar=1:3)
$betasCP
   Intercept          GNP   Unemployed Armed.Forces   Population         Year
  -9.731e+02    2.459e-02    9.953e-03    1.553e-02    3.391e-01    4.967e-01
    Employed
   7.239e-01

$landas
[1] 4.5478430 1.1858692 0.2517070 0.0124261 0.0018422 0.0003126

$CP.usadas
[1] 1 2 3

> #
> #  Si tomamos tantas componentes principales como regresores hay, hemos
> #  de obtener precisamente la misma solución que con MCO
> #
> regCP(X,y,tomar=1:ncol(X))
$betasCP
   Intercept          GNP   Unemployed Armed.Forces   Population         Year
  2946.85636      0.26353      0.03648      0.01116     -1.73703     -1.41880
    Employed
     0.23129

$landas
[1] 4.5478430 1.1858692 0.2517070 0.0124261 0.0018422 0.0003126

$CP.usadas
[1] 1 2 3 4 5 6

> lsfit(X,y)$coefficients                   #  Comprobación
   Intercept          GNP   Unemployed Armed.Forces   Population         Year
  2946.85636      0.26353      0.03648      0.01116     -1.73703     -1.41880
```

```
    Employed
     0.23129
> #
> #  Para dejar que la función seleccione el número de componentes
> #  tomando aquéllas significativas al nivel, por ejemplo, 0.10,
> #
> regCP(X,y,sig=0.10)
$betasCP
   Intercept          GNP   Unemployed Armed.Forces   Population         Year
  -961.37468      0.02372      0.01373      0.01991      0.33197      0.49223
    Employed
     0.66205

$landas
[1] 4.5478430 1.1858692 0.2517070 0.0124261 0.0018422 0.0003126

$CP.usadas
[1] 1 2
```

## 8.4. Regresión en raíces latentes*.

Consideramos el modelo[5]

$$\vec{y} = \vec{1}\beta_0 + W\vec{\beta}^* + \vec{\epsilon} \tag{8.46}$$

o alternativamente:

$$\vec{y}^* = W\vec{\beta}^* + \vec{\epsilon} \tag{8.47}$$

en que tanto los regresores como la variable respuesta $\vec{y}^*$ han sido normalizados y centrados. Es decir, $\vec{y}^* = \eta^{-1}(\vec{y} - \overline{\vec{y}})$ siendo $\eta^2 = \sum_{i=1}^{N}(y_i - \overline{y})^2$. Si construímos la matriz $N \times p$ siguiente:

$$A = [\vec{y}^* \mid W] \tag{8.48}$$

tenemos que la matriz $(A'A)$ es una matriz de correlación (tiene "unos" en la diagonal principal, es simétrica y semidefinida positiva). Sea $V = (\vec{v}_1 \mid \cdots \mid \vec{v}_p)$ la matriz que la diagonaliza:

$$V'(A'A)V = \Lambda \Longleftrightarrow V\Lambda V' = A'A \tag{8.49}$$

Entonces:

$$A\vec{v}_j = v_{0j}\vec{y}^* + W\vec{v}_j^{(0)} \qquad (j = 1, \ldots, p) \tag{8.50}$$

y:

$$||v_{0j}\vec{y}_i^* + W\vec{v}_j^{(0)}||^2 = \sum_{i=1}^{N}\left(\vec{y}_i^* v_{0j} + \sum_{k=1}^{p-1} W_{ik}v_{kj}\right)^2 \tag{8.51}$$

$$= \vec{v}_j\,'(A'A)\vec{v}_j \tag{8.52}$$

$$= \lambda_j \tag{8.53}$$

[5]El trabajo original sobre este procedimiento puede verse en Webster et al. (1974). Puede consultarse también Trocóniz (1987a), pág. 247 y ss.

donde $v_{kj}$ es la $k$-ésima coordenada de $\vec{v}_j^{(0)}$, $W_{ik}$ el elemento en la fila $i$-ésima y columna $k$-ésima de la matriz $W$, y denominamos $\vec{v}_j^{(0)}$ a $\vec{v}_j$ desprovisto de su primer elemento; $\vec{v}_j\,' = (v_{0j} \mid \vec{v}_j^{(0)}\,')$.

Por consiguiente, cuando $\lambda_j \approx 0$ el módulo del lado derecho de (8.50) es cercano a cero, o lo que es lo mismo:

$$y_i^* v_{0j} \approx -\sum_{k=1}^{p-1} W_{ik} v_{kj} \qquad \forall i \in [1, \ldots, N] \tag{8.54}$$

Si $v_{0j} \neq 0$, podemos escribir:

$$\vec{y}^* \approx \hat{y}_{(j)}^* \stackrel{\text{def}}{=} -v_{0j}^{-1} W \vec{v}_j^{(0)} \tag{8.55}$$

Como $\vec{y}^* = \eta^{-1}(\vec{y} - \vec{\overline{y}})$, $\vec{y} = \vec{\overline{y}} + \eta \vec{y}^*$ y denominando $\hat{y}_{(j)} = \vec{\overline{y}} + \eta \hat{y}_{(j)}^*$ tenemos:

$$\begin{aligned}
(\vec{y} - \hat{y}_{(j)})\,'(\vec{y} - \hat{y}_{(j)}) &= \eta^2 (\vec{y}^* - \hat{y}_{(j)}^*)\,'(\vec{y}^* - \hat{y}_{(j)}^*) && (8.56)\\
&= (v_{0j}\vec{y}^* - v_{0j}\hat{y}_{(j)}^*)\,'(v_{0j}\vec{y}^* - v_{0j}\hat{y}_{(j)}^*)\frac{\eta^2}{v_{0j}^2} && (8.57)\\
&= (A\vec{v}_j)\,'(A\vec{v}_j)\frac{\eta^2}{v_{0j}^2} && (8.58)\\
&= \frac{\lambda_j \eta^2}{v_{0j}^2} && (8.59)
\end{aligned}$$

Nótese que la aproximación de $\vec{y}^*$ en (8.55), y correspondiente suma de cuadrados de los residuos en (8.59, hacen uso exclusivamente de una parte de la información disponible; la de que $\lambda_j$ es aproximadamente cero para un determinado $j$. Podemos pensar en hacer uso de toda la información disponible aproximando $\vec{y}$ mediante una combinación lineal de $\hat{y}_{(i)} \quad (i = 1, \ldots, p)$, debidamente ponderadas por coeficientes $d_i$ a determinar:

$$\begin{aligned}
\hat{y} &= \sum_{i=1}^{p} d_i \hat{y}_{(i)} && (8.60)\\
&= \sum_{i=1}^{p} d_i [\vec{\overline{y}} + W(-{v_{0i}}^{-1}\vec{v}_i^{(0)}\eta)] && (8.61)\\
&= \left(\sum_{i=1}^{p} d_i\right)\vec{\overline{y}} + W\left(-\sum_{i=1}^{p} d_i {v_{0i}}^{-1}\vec{v}_i^{(0)}\eta\right) && (8.62)\\
&= \hat{\beta}_0\vec{1} + W\hat{\beta}^* && (8.63)
\end{aligned}$$

lo que proporciona:

$$\begin{aligned}
\hat{\beta}_0 &= \overline{y}\vec{1}\left(\sum_{i=1}^{p} d_i\right) && (8.64)\\
\hat{\beta}^* &= -\eta\sum_{i=1}^{p} d_i {v_{0i}}^{-1}\vec{v}_i^{(0)} && (8.65)
\end{aligned}$$

Como los regresores $W$ están centrados, es claro que $\hat{\beta}_0 = \overline{y}$, y por tanto de (8.64) se deduce $\sum_{i=1}^{p} d_i = 1$. Haciendo uso de (8.59), (8.64), y (8.65) obtenemos:

$$\begin{aligned}
(\vec{y}-\hat{y})'(\vec{y}-\hat{y}) &= \eta^2(\vec{y}^*-\hat{y}^*)'(\vec{y}^*-\hat{y}^*) \\
&= \eta^2\left(\vec{y}^* + W\sum_{i=1}^{p} d_i {v_{0i}}^{-1}\vec{v}_i^{(0)}\right)'\left(\vec{y}^* + W\sum_{i=1}^{p} d_i {v_{0i}}^{-1}\vec{v}_i^{(0)}\right) \\
&= \eta^2\left[\sum_{i=1}^{p}\left(\frac{d_i}{v_{0i}}\right)(\vec{y}^* v_{0i} + W\vec{v}_i^{(0)})\right]' \\
&\quad \times\left[\sum_{i=1}^{p}\left(\frac{d_i}{v_{0i}}\right)(\vec{y}^* v_{0i} + W\vec{v}_i^{(0)})\right] \\
&= \eta^2\left[\sum_{i=1}^{p}\left(\frac{d_i}{v_{0i}}\right)A\vec{v}_i\right]'\left[\sum_{i=1}^{p}\left(\frac{d_i}{v_{0i}}\right)A\vec{v}_i\right] \\
&= \eta^2\sum_{i=1}^{p}\left(\frac{\lambda_i d_i^2}{{v_{0i}}^2}\right) \qquad (8.66)
\end{aligned}$$

Podemos ahora minimizar la expresión (8.66) sujeta a que $\sum_{i=1}^{p} d_i = 1$. El lagrangiano es:

$$\Phi(\vec{d}) = \eta^2\sum_{i=1}^{p}\left(\frac{\lambda_i d_i^2}{{v_{0i}}^2}\right) - \mu\left(\sum_{i=1}^{p} d_i - 1\right) \qquad (8.67)$$

cuyas derivadas

$$\frac{\partial\Phi(\vec{d})}{\partial d_i} = 2\eta^2\left(\frac{d_i\lambda_i}{{v_{0i}}^2}\right) - \mu = 0 \qquad (i = 1,\ldots,p) \qquad (8.68)$$

permiten (multiplicando cada igualdad en (8.68) por ${v_{0i}}^2\lambda_i^{-1}$ y sumando) obtener:

$$\mu = 2\eta^2\left(\sum_{i=1}^{p}\frac{v_{0i}^2}{\lambda_i}\right)^{-1} \qquad (8.69)$$

Llevando (8.69) a (8.68) obtenemos:

$$2\eta^2 d_i\frac{\lambda_i}{v_{0i}^2} = \mu = 2\eta^2\left(\sum_{i=1}^{p}\frac{v_{0i}^2}{\lambda_i}\right)^{-1} \qquad (8.70)$$

y por tanto:

$$d_i = \frac{v_{0i}^2}{\lambda_i}\left(\sum_{i=1}^{p}\frac{v_{0i}^2}{\lambda_i}\right)^{-1} \qquad (8.71)$$

Los estimadores deseados se obtienen llevando (8.71) a (8.64)–(8.65):

$$\hat{\beta}_0 = \overline{y} \qquad (8.72)$$

$$\hat{\beta}^* = -\eta\frac{\sum_{i=1}^{p}\left(\frac{v_{0i}}{\lambda_i}\right)\vec{v}_i^{(0)}}{\sum_{i=1}^{p}\frac{v_{0i}^2}{\lambda_i}} \qquad (8.73)$$

Podríamos detenernos aquí, pero hay más. Cabe distinguir dos tipos de multicolinealidades entre las columnas de la matriz $[\vec{y}^* \mid W]$; aquéllas en que $v_{0i} \gg 0$ (que llamaremos *(multicolinealidades predictivas)*, y aquéllas en que $v_{0i} \approx 0$ *(multicolinealidades no predictivas)*; las primeras permiten despejar $\vec{y}^*$, y son aprovechables para la predicción, en tanto las segundas son multicolinealidades fundamentalmente entre los regresores.

El estimador anterior pondera cada $\vec{v}_i^{(0)}$ en proporción directa a $v_{0i}$ e inversa a $\lambda_i$; es lo sensato. Pero podemos eliminar en (8.73) términos muy inestables, cuando $v_{0i}$ y $\lambda_i$ son ambos muy pequeños, para evitar que el sumando correspondiente en (8.73) reciba gran ponderación, si parece evidente que se trata de una multicolinealidad no predictiva. La relación (8.73) se transformará entonces en:

$$\hat{\beta}^* = -\eta \frac{\sum_{i \in P} \left( \frac{v_{0i}}{\lambda_i} \right) \vec{v}_i^{(0)}}{\sum_{i \in P} \left( \frac{v_{0i}^2}{\lambda_i} \right)} \tag{8.74}$$

siendo $P$ un subconjunto de $(1, \ldots, p)$.

La determinación de $P$ es una tarea eminentemente subjetiva; se suele desechar una multicolinealidad cuando $\lambda_i < 0{,}10$ y $v_{0i} < 0{,}10$, si además $\vec{v}_i^{(0)}$ "se aproxima" a un vector propio de $W'W$.

## CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

**8.1** Al final de la Sección 8.2 se proponía emplear un criterio del tipo

$$(\hat{\beta} - \vec{\beta})' M (\hat{\beta} - \vec{\beta})$$

con $M = (X'X)$. Dése una justificación para esta elección de $M$.

**8.2** Demuéstrese que si $u_i$ es definida como en (8.27), se verifica que $\vec{1} \perp \vec{u}_i$.

**8.3** Sea una muestra formada por $n$ observaciones, $X_1, \ldots, X_n$, generadas por una distribución con media. Demuéstrese que, para algún $c$, $c\overline{X}$ es mejor estimador (en terminos de error medio cuadrático, ECM) que $\overline{X}$. ¿Es esto un caso particular de alguno de los procedimientos de estimación examinados en este capítulo?

**8.4** Es fácil realizar regresión *ridge* incluso con programas pensados sólo para hacer regresión mínimo cuadrática ordinaria. Basta prolongar el vector $\vec{y}$ con $p$ ceros, y la matriz $X$ con $p$ filas adicionales: las de la matriz $\sqrt{k} I_{p \times p}$. Llamamos $\tilde{X}$ e $\tilde{y}$ a la matriz de regresores y vector respuesta así ampliados. Al hacer regresión ordinaria de $\tilde{y}$ sobre $\tilde{X}$ obtenemos:

$$\begin{aligned}
\hat{\beta} &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} && (8.75)\\
&= (X'X + kI)^{-1}(X'\vec{y} + \sqrt{k} I \vec{0}) && (8.76)\\
&= (X'X + kI)^{-1} X'\vec{y} && (8.77)\\
&= \hat{\beta}^{(k)} && (8.78)
\end{aligned}$$

Alternativamente, se puede formar $\tilde{X}$ añadiendo a $X$ las filas de una matriz unidad, y realizar regresión ponderada (dando a cada observación "normal" peso unitario

y a las $p$ seudo-observaciones añadidas peso $\sqrt{k}$). La alteración de los pesos es habitualmente más cómoda que la creación de una nueva matriz de regresores. Este será de ordinario el método a utilizar cuando hayamos de probar muchos valores diferentes de $k$ y dispongamos de un programa para hacer regresión mínimo cuadrática ponderada. Las funciones `lsfit` y `lm` (disponibles en R y S-PLUS) admiten ambas el uso de pesos y por tanto se prestan al uso descrito. La librería MASS contiene no obstante la función `lm.ridge`, que hace estimación ridge de modo más cómodo para el usuario.

**8.5** Supongamos una muestra formada por pares de valores $(y_i, x_i)$, $i = 1, \ldots, N$. La variable $Y$ es peso, la variable $X$ es edad, y las observaciones corresponden a $N$ diferentes sujetos. Estamos interesados en especificar la evolución del peso con la edad. Podríamos construir la matrix de diseño

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & \ldots & x_1^{p-1} \\ 1 & x_2 & x_2^2 & x_2^3 & \ldots & x_2^{p-1} \\ 1 & x_3 & x_3^2 & x_3^3 & \ldots & x_3^{p-1} \\ \vdots & & & \vdots & & \vdots \\ 1 & x_N & x_N^2 & x_N^3 & \ldots & x_N^{p-1} \end{pmatrix} \tag{8.79}$$

y contrastar hipótesis tales como $H_0 : \beta_2 = \beta_3 = \ldots = \beta_{p-1} = 0$ (tendencia no más que lineal), $H_0 : \beta_3 = \ldots = \beta_{p-1} = 0$ (tendencia no más que cuadrática), etc. Sucede sin embargo, como es fácil comprobar, que una matriz como la anterior adolece de una acusada multicolinealidad, sean cuales fueren los valores $x_1, \ldots, x_N$.

Podríamos ortogonalizar los vectores columna de la matriz de diseño (por ejemplo mediante el procedimiendo de Gram-Schmidt: véase Grafe (1985) o cualquier libro de Algebra Lineal), para obtener una nueva matriz de diseño. Los nuevos vectores columna generan el mismo espacio y el contraste puede hacerse del mismo modo que con los originales, pero sin problemas de multicolinealidad.

Otra posibilidad es sustituir las potencias creciente de $x_i$ en las columnas de $X$ por polinomios ortogonales evaluados para los mismos valores $x_i$ (ver por ejemplo Seber (1977), Dahlquist and Björck (1974), o cualquier texto de Análisis Numérico).

Ambos procedimientos tienen por finalidad encontrar una base ortogonal o aproximadamente ortogonal generando el mismo espacio que los vectores columna originales de la matriz de diseño.

**8.6** (↑ 8.5) ¿Por qué, para la finalidad perseguida en el Ejercicio 8.5, no sería de utilidad hacer regresión en componentes principales?

# Capítulo 9

# Evaluación del ajuste. Diagnósticos.

Ya hemos visto en lo que precede estadísticos para evaluar la bondad de ajuste de un modelo, como $\overline{R}^2$; pero se trata de estadísticos que dan una idea global del ajuste. Puede ocurrir que un $\overline{R}^2$ encubra el hecho de que localmente —para unas ciertas observaciones— el ajuste es muy deficiente.

En lo que sigue abordaremos esta cuestión, considerando instrumentos para examinar el ajuste localmente (para observaciones individuales). Examinaremos también la cuestión íntimamente relacionada de cuándo una observación (o varias) son muy influyentes, en el sentido de condicionar de modo importante la estimación del modelo.

## 9.1. Análisis de residuos.

En general, como se ha indicado ya en el Capítulo 10, no conocemos la forma en que se generan los valores de la variable respuesta $\vec{Y}$ . Todos los modelos que ajustemos son en alguna medida provisionales, y su adecuación a los datos debe ser objeto de análisis. El desarrollo que se hace a continuación sigue principalmente a Cook and Weisberg (1982). Otras referencias de utilidad son Hawkins (1980), Barnett and Lewis (1978), Belsley et al. (1980), Myers (1990) y Trocóniz (1987a).

La forma más natural de examinar el ajuste consiste en considerar los residuos

$$\hat{\epsilon} = \vec{y} - X\hat{\beta} = (I - X(X'X)^{-1}X')\vec{y} = (I - X(X'X)^{-1}X')\vec{\epsilon} \qquad (9.1)$$

Podemos contemplar los $\hat{\epsilon}_i$ como "estimaciones" de las perturbaciones $\epsilon_i$ (inobservables) que han intervenido en la generación de las $Y_i$. Veremos sin embargo que, en general, sólo vagamente reproduce $\hat{\epsilon}$ el comportamiento de $\vec{\epsilon}$. En particular,

**Teorema 9.1** *Bajo los supuestos habituales se verifica que:*

1. *Los residuos no son, en general, homoscedásticos, incluso cuando las perturbaciones lo son.*

2. *Los residuos no son, en general, incorrelados, incluso cuando las perturbaciones lo son.*

DEMOSTRACION:

$$\Sigma_{\hat{\epsilon}} = E[(\hat{\epsilon} - E(\hat{\epsilon}))(\hat{\epsilon} - E(\hat{\epsilon}))']$$ (9.2)

Como $E(\hat{\epsilon}) = \vec{0}$, (9.2) se reduce a:

$$\begin{aligned} E\hat{\epsilon}\hat{\epsilon}' &= E[(I - X(X'X)^{-1}X')\vec{y}\vec{y}'(I - X(X'X)^{-1}X')'] & (9.3)\\ &= (I - X(X'X)^{-1}X')\sigma^2 I & (9.4)\\ &= \sigma^2(I - P), & (9.5)\end{aligned}$$

que en general no tiene elementos iguales a lo largo de la diagonal principal. El apartado 2) del enunciado es inmediato a partir de (9.5), dado que $(I - P)$ es una matriz no diagonal.

Sea,

$$p_{ij} = \vec{x}_i\,'((X'X))^{-1}\vec{x}_j$$ (9.6)

un elemento genérico de la matriz $P$ ($\vec{x}_i\,'$ denota la $i$-ésima fila de $X$). De la igualdad (9.1) se deduce:

$$\hat{\epsilon}_i = (1 - p_{ii})\epsilon_i - \sum_{i \neq j} p_{ij}\epsilon_j$$ (9.7)

Por tanto, el residuo $i$-ésimo es un promedio ponderado de la perturbación correspondiente a dicha observación y las de todas las demás observaciones, con ponderaciones $(1-p_{ii})$ y $(-p_{ij})$. Dependiendo de los valores que tomen estos coeficientes, $\hat{\epsilon}_i$ recogerá con desigual fidelidad el valor de $\epsilon_i$.

Los valores $p_{ij}$ dependen sólo de la matrix de diseño y son del mayor interés, como veremos más abajo.

### 9.1.1. Residuos internamente studentizados.

Los residuos MCO definidos en (9.1) son, por causa de su heterocedasticidad, desaconsejables para la detección de observaciones anormales o diagnóstico de modelos de regresión. Es sin embargo fácil corregir dicha heterocedasticidad. De (9.5) se deduce que una estimación de la varianza de $\hat{\epsilon}_i$ viene dada por $\hat{\sigma}^2(1 - p_{ii})$. Por tanto,

$$r_i = \frac{\hat{\epsilon}_i}{+\sqrt{\hat{\sigma}^2(1 - p_{ii})}}$$ (9.8)

para $i = 1, \ldots, N$ son residuos de varianza común. Se llama *studentización* a la eliminación del efecto de un parámetro de escala (aquí $\sigma^2$) mediante división por una estimación adecuada. Se denomina *internamente studentizados* a los residuos definidos en (9.8).

Es de notar que, a pesar de su denominación, los $r_i$ no siguen una distribución $t$ de Student, pues numerador y denominador no son independientes ($\hat{\epsilon}_i$ ha intervenido en el cómputo de $\hat{\sigma}^2$). Es fácil demostrar, sin embargo, que bajo los supuestos habituales más el de normalidad en las perturbaciones, $r_i^2/(N-p)$ sigue una distribución beta $B(\frac{1}{2}, \frac{1}{2}(N-p-1))$.

Al tener los $r_i$ la misma varianza, se prestan mejor a ser examinados gráficamente para identificar posibles observaciones anómalas o *outliers*.

### 9.1.2. Residuos externamente studentizados.

Definidos por:

$$t_i = \frac{\hat{\epsilon}_i}{+\sqrt{\hat{\sigma}^2(i)(1-p_{ii})}} \tag{9.9}$$

son formalmente idénticos a los $r_i$, con la única salvedad de haberse tomado en el denominador un estimador $\hat{\sigma}^2(i)$ de $\sigma^2$ que no hace uso de $\hat{\epsilon}_i$. Mediante una elección adecuada de $\hat{\sigma}^2(i)$ puede lograrse que $t_i$ siga una distribución $t$ de Student con $(N-p-1)$ grados de libertad. Esto permite, entre otras cosas, hacer uso de la distribución del máximo de $k$ variables $t$ de Student con correlación por pares $\rho$ (véase Sección 6.3, pág. 63) para contrastar la presencia de *outliers*. Tomaremos,

$$\hat{\sigma}^2(i) = \frac{\hat{\epsilon}'\hat{\epsilon} - \hat{\epsilon}_i(1-p_{ii})^{-1}\hat{\epsilon}_i}{(N-p-1)} \tag{9.10}$$

lo que permite probar el siguiente,

**Teorema 9.2** *Con $\hat{\sigma}^2(i)$ definido como en (9.10), bajo los supuestos habituales más el de normalidad en las perturbaciones, los residuos $t_i$ definidos en (9.9) (externamente studentizados) siguen una distribución $t$ de Student con $(N-p-1)$ grados de libertad.*

DEMOSTRACION:

Podemos escribir $\hat{\epsilon}_i = G_i'(I-P)\vec{\epsilon}$ siendo $G_i'$ de dimensión $1 \times N$, con un único "uno" en posición $i$-ésima y ceros en los demás lugares. Llamando $A = G_i'(I-P)$ tenemos que:

$$\hat{\epsilon}_i = A\vec{\epsilon} \tag{9.11}$$

Por otra parte, de (9.10) deducimos:

$$\begin{aligned}(N-p-1)\hat{\sigma}^2(i) &= \hat{\epsilon}\,'[I - G_i[G_i'(I-P)G_i]^{-1}G_i']\hat{\epsilon} \\ &= \vec{\epsilon}\,' \underbrace{(I-P)[I - G_i[G_i'(I-P)G_i]^{-1}G_i'](I-P)}_{B}\vec{\epsilon} \\ &= \vec{\epsilon}\,' B\vec{\epsilon} \end{aligned} \tag{9.12}$$

Es fácil comprobar que $AB = 0$, luego $\hat{\epsilon}_i$ y $\hat{\sigma}^2(i)$ son independientes (Lema 4.3, pág. 40). Por otra parte, es también fácil comprobar que $B$ es idempotente, con rango (= traza) $(N-p-1)$. Por consiguiente,

$$\frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2(i)(1-p_{ii})}} = \frac{\hat{\epsilon}_i/\sqrt{\sigma^2(1-p_{ii})}}{\sqrt{\hat{\sigma}^2(i)/\sigma^2}} \tag{9.13}$$

$$= \frac{\hat{\epsilon}_i/\sqrt{\sigma^2(1-p_{ii})}}{\sqrt{\vec{\epsilon}\,' B\vec{\epsilon}/(N-p-1)\sigma^2}} \tag{9.14}$$

Pero en el numerador y denominador de (9.14) hay respectivamente una variable aleatoria $N(0,1)$ y una $\chi^2$ dividida entre sus grados de libertad, ambas independientes, lo que demuestra el Teorema.

Para contrastar la hipótesis de presencia de *outliers*, podemos comparar el mayor de los residuos externamente studentizados con el cuantil apropiado de la distribución del máximo valor absoluto de $k$ variables aleatorias $t$ de Student (Sección 6.3, pág. 63). Supondremos que son incorrelados, salvo que podamos calcular fácilmente su correlación por pares, como sucede a menudo en Análisis de Varianza. El texto Seber (1977) reproduce en su Apéndice E tablas adecuadas. Alternativamente, podemos comparar el mayor residuo internamente studentizado con los valores críticos en las tablas de Lund (1975), o emplear la desigualdad de Bonferroni.

### 9.1.3. Residuos BLUS.

La studentización, tanto interna como externa, elimina la heterocedasticidad de los residuos, pero no la mutua correlación. No es posible obtener un vector de $N$ residuos incorrelados y ortogonales a las columnas de $X$. La razón se ve fácilmente: $\hat{\epsilon} \perp R(X)$ es un vector aleatorio de $N$ coordenadas, pero constreñido a yacer en un subespacio $(N-p)$ dimensional. Su distribución en $R^N$ es degenerada, y su matriz de covarianzas de rango $(N-p)$ (supuesta $X$ de rango completo). ***Ninguna*** transformación ortogonal puede convertir tal matriz en diagonal de rango $N$.

Si es posible, sin embargo, obtener $(N-p)$ residuos incorrelados, homoscedásticos, y de media 0; de hecho, hay multitud de maneras de hacerlo[1], dependiendo del subconjunto de $(N-p)$ residuos que escojamos.

Tales residuos, denominados BLUS (o ELIO), son de utilidad para contrastar homoscedasticidad (suministrando una alternativa al conocido método de Goldfeld-Quandt), normalidad, etc. Un tratamiento detallado puede encontrarse en Theil (1971), Cap. 5.

### 9.1.4. Residuos borrados.

Sean $X_{(i)}$ e $\vec{Y}_{(i)}$ la matriz de diseño y vector respuesta desprovistos de la observación $i$-ésima. Sea $\hat{\beta}_{(i)}$ el vector de estimadores de los parámetros obtenido sin dicha observación, es decir, $\hat{\beta}_{(i)} = (X'_{(i)}X_{(i)})^{-1}X'_{(i)}\vec{Y}_{(i)}$. Se llama *residuos borrados (deleted residuals)* a los $d_i$ definidos así[2]:

$$d_i = y_i - \vec{x}_i{}'\hat{\beta}_{(i)} \tag{9.15}$$

Un $d_i$ muy pequeño o nulo indicaría que la observación $i$-ésima no se separa en su comportamiento del recogido por la regresión sobre las restantes $N-1$ observaciones. Lo contrario es cierto si $d_i$ es muy grande.

Hay una relación muy simple que permite calcular los $d_i$ sin necesidad de realizar $N$ regresiones diferentes sobre todos los conjuntos posibles de $N-1$ observaciones.

[1] Véase Theil (1971), pág. 202 y ss.

[2] Una denominación alternativa frecuente en la literatura es la de residuos PRESS (predictive sum of squares residuals).

En efecto, de (9.15) se deduce que:

$$
\begin{aligned}
d_i &= y_i - \vec{x}_i\,'(X'_{(i)}X_{(i)})^{-1}X'_{(i)}\vec{Y}_{(i)} \\
&= y_i - \vec{x}_i\,'\left[(X\,'X) - \vec{x}_i\vec{x}_i\,'\right]^{-1}X'_{(i)}\vec{Y}_{(i)} && (9.16)\\
&= y_i - \vec{x}_i\,'\left[((X\,'X))^{-1} + \frac{((X\,'X))^{-1}\vec{x}_i\vec{x}_i\,'((X\,'X))^{-1}}{1-\vec{x}_i\,'((X\,'X))^{-1}\vec{x}_i}\right]X'_{(i)}\vec{Y}_{(i)} && (9.17)\\
&= y_i - \vec{x}_i\,'\left[\frac{(1-p_{ii})((X\,'X))^{-1} + ((X\,'X))^{-1}\vec{x}_i\vec{x}_i\,'((X\,'X))^{-1}}{1-p_{ii}}\right]X'_{(i)}\vec{Y}_{(i)} \\
&= y_i - \left[\frac{(1-p_{ii})\vec{x}_i\,'((X\,'X))^{-1} + p_{ii}\vec{x}_i\,'((X\,'X))^{-1}}{1-p_{ii}}\right]X'_{(i)}\vec{Y}_{(i)} \\
&= y_i - \frac{\vec{x}_i\,'((X\,'X))^{-1}X'_{(i)}\vec{Y}_{(i)}}{1-p_{ii}} \\
&= \frac{(1-p_{ii})y_i - \vec{x}_i\,'((X\,'X))^{-1}(X'\vec{Y} - \vec{x}_i y_i)}{1-p_{ii}} && (9.18)\\
&= \frac{y_i - \vec{x}_i\,'((X\,'X))^{-1}X'\vec{Y}}{1-p_{ii}} \\
&= \frac{\hat{\epsilon}_i}{1-p_{ii}} && (9.19)
\end{aligned}
$$

en que el paso de (9.16) a (9.17) hace uso del Teorema A.2, pág. 179. Veremos en lo que sigue que $d_i$ está relacionado con la influencia que la observación $i$-ésima tiene sobre la estimación de los parámetros.

## 9.2. Análisis de influencia.

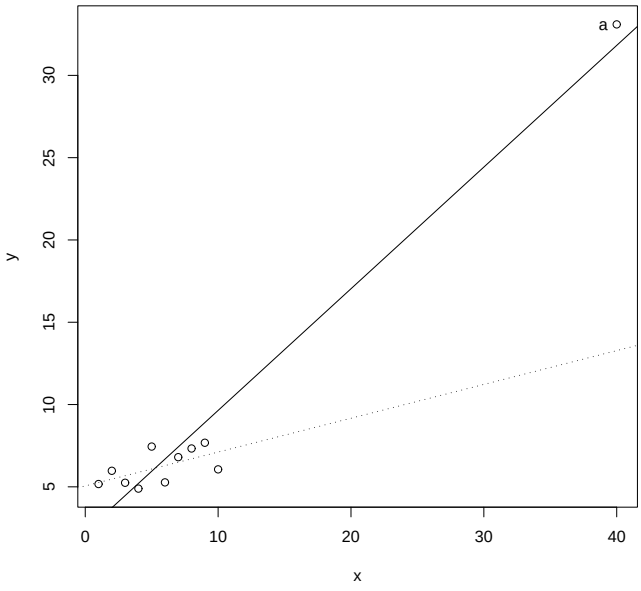
Es en general indeseable que la estimación de un parámetro dependa de modo casi exclusivo de una sola observación o de unas pocas, de manera que su eliminación conduzca a resultados completamente diferentes. En general, cuando esto ocurre, es necesario particionar la muestra o replantear el modelo. En todo caso, es necesario saber hasta que punto observaciones aisladas influencian las estimaciones de los parámetros para obrar en consecuencia.

Puede parecer que para determinar qué observaciones influyen más en el resultado de la estimación basta mirar los residuos, brutos o studentizados. Ello es verdad, pero sólo en parte: puede haber observaciones extraordinariamente influyentes que resulten muy bien ajustadas por la regresión, como el ejemplo de la Fig. 9.1 pone de manifiesto.

Claramente, el punto $a$ tiene una notable influencia en la estimación de la pendiente de la recta, hasta el punto de que su omisión daría lugar a un resultado completamente diferente (la recta dibujada con trazo discontinuo). Sin embargo, su residuo MCO es muy pequeño; un exámen de los residuos MCO —o incluso de los residuos *studentizados*— difícilmente delataría ninguna anormalidad.

El examen de los residuos borrados detectaría una situación como la mencionada: $a$ tendría un residuo borrado grande. Pero todavía es posible un análisis más sofisticado, que tenga en cuenta, en particular, los parámetros sobre los que una observación es muy influyente. Abordamos este análisis a continuación.

Figura 9.1: Una observación como $a$ tiene residuo borrado muy grande, y gran influencia en la pendiente de la recta de regresión.



### 9.2.1. La curva de influencia muestral.

La forma obvia de examinar la influencia de la observación $i$-ésima consiste en comparar los vectores de estimadores obtenidos con y sin dicha observación: $\hat{\beta}$ y $\hat{\beta}_{(i)}$ respectivamente. En consecuencia, definimos la *curva de influencia muestral* (SIC) así:

$$\text{SIC}_i \quad = \quad (N-1)(\hat{\beta} - \hat{\beta}_{(i)}). \tag{9.20}$$

El factor $(N-1)$ tiene por misión corregir el efecto del tamaño muestral: en igualdad de todo lo demás, una observación altera la estimación tanto menos cuanto más grande sea la muestra.

La expresión (9.20) es vector-valorada: recoge, debidamente amplificadas por $(N-1)$, por la razón apuntada, las diferencias que introduce la inclusión de la observación $i$-ésima sobre cada uno de los $p$ parámetros estimados. Podemos relacionar (9.20) con el residuo borrado $i$-ésimo haciendo uso del siguiente lema.

**Lema 9.1** *Se verifica que*

$$(\hat{\beta} - \hat{\beta}_{(i)}) \quad = \quad \frac{((X'X))^{-1}\vec{x}_i\hat{\epsilon}_i}{(1-p_{ii})} = ((X'X))^{-1}\vec{x}_i d_i. \tag{9.21}$$

DEMOSTRACION:

$$\begin{aligned}
(\hat{\beta} - \hat{\beta}_{(i)}) &= ((X'X))^{-1}X'\vec{Y} - ((X'X) - \vec{x}_i\vec{x}_i')^{-1}(X'\vec{Y} - \vec{x}_i y_i) \\
&= ((X'X))^{-1}X'\vec{Y} \\
&\quad - \left[((X'X))^{-1} + \frac{((X'X))^{-1}\vec{x}_i\vec{x}_i'((X'X))^{-1}}{1-\vec{x}_i'((X'X))^{-1}\vec{x}_i}\right](X'\vec{Y} - \vec{x}_i y_i) \\
&= ((X'X))^{-1}\vec{x}_i y_i - \frac{((X'X))^{-1}\vec{x}_i\vec{x}_i'((X'X))^{-1}X'\vec{Y}}{1-p_{ii}} \\
&\quad + \frac{((X'X))^{-1}\vec{x}_i\vec{x}_i'((X'X))^{-1}\vec{x}_i y_i}{1-p_{ii}} \\
&= \frac{((X'X))^{-1}\vec{x}_i}{1-p_{ii}}\left[(1-p_{ii})y_i - \vec{x}_i'\hat{\beta} + p_{ii}y_i\right] \\
&= ((X'X))^{-1}\vec{x}_i\frac{\hat{\epsilon}_i}{1-p_{ii}}
\end{aligned}$$

En consecuencia,

$$\text{SIC}_i \quad = \quad (N-1)(\hat{\beta} - \hat{\beta}_{(i)}) = (N-1)((X'X))^{-1}\vec{x}_i\frac{\hat{\epsilon}_i}{1-p_{ii}}$$

y el cálculo de la curva de influencia muestral $\text{SIC}_i$ correspondiente a la observación $i$ no requiere realizar una regresión para cada $i$; todos los cálculos se se pueden hacer con ayuda de los residuos ordinarios y diagonal de la matriz de proyección correspondientes a la matriz de proyección $X(X'X)^{-1}X'$.

Diferentes versiones de la curva de influencia disponibles en regresión lineal puede encontrarse en Cook and Weisberg (1982) y Belsley et al. (1980). Alternativas como la *curva de influencia empírica EIC* y otras, difieren de la curva de influencia muestral presentada en el grado en que se corrige $\hat{\epsilon}_i$ (en la EIC se divide entre $(1-p_{ii})^2$, en lugar de entre $(1-p_{ii})$ como en (9.22).

### 9.2.2. Distancia de Cook.

Tal y como se indica más arriba, la curva de influencia en cualquiera de sus versiones es, en nuestro caso, un vector $p \times 1$ ($p$ = número de parámetros). La coordenada $k$-ésima de $\mathrm{SIC}_i$ proporciona información sobre la influencia de la observación $i$-ésima en la estimación de $\hat{\beta}_k$. Aunque esta información pormenorizada sea útil, en ocasiones queremos una única medida resumen de la influencia de una observación.

Sea $\hat{\beta}_{(i)}$ el vector de estimadores obtenido sin hacer uso de la observación $i$-ésima, y $\hat{\beta}$ el computado con la muestra completa. Una posibilidad es ponderar las discrepancias en una única expresión como:

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})' S (\hat{\beta} - \hat{\beta}_{(i)})}{c} \tag{9.22}$$

siendo $S$ una matriz definida no negativa y $c$ una constante positiva. Puesto que $\hat{\beta} \sim (\vec{\beta}\,, \sigma^2((X\,'X))^{-1})$, una elección posible que aproximadamente "normaliza" (9.22) es: $S = (X\,'X)$ y $c = p\hat{\sigma}^2$. Con esta elección, la expresión (9.22) se denomina *distancia de Cook* y es una medida global de la influencia de la observación $(\vec{x}_i, y_i)$. Hay otras posibles elecciones de $S$ y $c$ con diferencias, en general, sólo de matiz[3].

Haciendo uso del Lema 9.1 tenemos que la distancia de Cook puede escribirse así:

$$D_i = \frac{\hat{\epsilon}_i \vec{x}_i\,'((X\,'X))^{-1}(X\,'X)((X\,'X))^{-1}\vec{x}_i\hat{\epsilon}_i}{p\hat{\sigma}^2(1-p_{ii})^2} \tag{9.23}$$

$$= \frac{1}{p} r_i^2 \frac{p_{ii}}{1-p_{ii}} \tag{9.24}$$

siendo $r_i$ el $i$-ésimo residuo internamente studentizado.

### 9.2.3. DFFITS.

Se definen así:

$$\mathtt{DFFIT}_i = t_i \sqrt{\frac{p_{ii}}{1-p_{ii}}} \tag{9.25}$$

Se suele considerar observaciones inusuales a aquéllas con

$$|\mathtt{DFFIT}_i| > 2\sqrt{\frac{p}{N}} \tag{9.26}$$

### 9.2.4. DFBETAS.

Se definen por:

$$\mathsf{DFBETA}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j,(i)}}{\hat{\sigma}\sqrt{((X\,'X))^{-1}_{jj}}}; \tag{9.27}$$

Los estadísticos DFBETA permiten evaluar la influencia de la observación $i$-ésima sobre el parámetro $j$-ésimo. En cierto modo desglosan la información que la distancia de Cook resume en un único estadístico por observación. La motivación de la expresión

[3] Una relación de las mismas puede verse en Cook and Weisberg (1982), p. 124.

(9.27) es clara: la diferencia entre la estimación de $\beta_j$-ésimo con y sin la observación $i$-ésima se divide por una estimación de la desviación típica de $\hat{\beta}_j$.

El criterio que se sigue es el de comparar $|\mathsf{DFBETA}_{ij}|$ con $2/\sqrt{N}$. Más detalles en Belsley et al. (1980).

## 9.3. Análisis gráfico de residuos

Al margen del uso que pueda hacerse de los residuos en cualquiera de sus variedades para, por ejemplo, contrastar hipótesis de presencia de *outliers*, etc., con frecuencia será conveniente construir algunos gráficos. Es mucha, en efecto, la información que cabe obtener de ellos. Presentamos a continuación algunos de estos gráficos; otros aparecerán en contexto en los capítulos dedicados a selección de modelos (Capítulo 10) y transformaciones de las variables (capítulo 11). Referencias útiles para ampliar lo que se expone a continuación incluyen Trocóniz (1987a), Myers (1990), Ryan (1997) o Atkinson (1985).

### 9.3.1. Gráficos de residuos frente a índice de observación $(i, \hat{\epsilon}_i)$

Frecuentemente, el índice de cada observación es el tiempo, es decir, las observaciones han sido tomadas secuencialmente una despues de otra. El representar $\vec{\epsilon}_i$ frente a $i$ nos podría poner de manifiesto rupturas temporales —por ejemplo, una brusca disminución del tamaño de los residuos a partir de un cierto $i$—. En ocasiones podemos ver también en un gráfico de esta naturaleza pautas como agrupamiento de residuos, que puede convenir investigar.

Pueden emplearse residuos ordinarios o *studentizados* en cualquiera de sus variedades.

### 9.3.2. Gráficos de residuos frente a variables incluidas $(x_{ij}, \hat{\epsilon}_i)$

Los residuos ordinarios son por construcción ortogonales a cualquiera de los regresores. No obstante, un gráfico de esta naturaleza puede aportar información acerca del modo en que un regresor interviene en la generación de la respuesta: por ejemplo, podríamos ver una pauta de relación no lineal entre $\hat{\epsilon}_i$ y $x_{ij}$, sugiriendo que $x_{ij}$ debe suplementarse con un término cuadrático, entrar como función exponencial, etc.

### 9.3.3. Gráficos de residuos frente a variables excluidas $(x^*_{ij}, \hat{\epsilon}_i)$

La idea es similar a la del apartado precedente, pero $x^*_{ij}$ son ahora los valores de una variable no incluida (y candidato a serlo) en la regresión. Un gráfico de esta naturaleza permitiría ver si la parte no explicada de la respuesta (los residuos) tiene alguna relación evidente con la nueva variable. En su caso, dependiendo de la pauta que dibujaran los residuos, tendríamos pistas acerca de si dicha variable $\vec{x}^*_j$ ha de incluirse tal cual o tras alguna transformación funcional.

### 9.3.4. Gráficos de variable añadida $(\hat{\epsilon}_{Y|X_{-j}}, \hat{\epsilon}_{X_j|X_{-j}})$

La idea es similar a la del apartado anterior. Se dibujan los residuos de la regresión de $Y$ sobre todas las variables *menos* $X_j$ sobre los residuos de regresar dicha variable

sobre todas las demás. Los residuos de ambas regresiones recogen, respectivamente, las partes de $Y$ y $X_j$ ortogonales al subespacio generado por las restantes variables.

Si hubiera alguna pauta en dicha gráfica, podríamos interpretarla como relación entre $Y$ y $X_j$ eliminado en ambas el efecto de las restantes variables.

### 9.3.5. Gráficos de normalidad de residuos

Aunque, como se ha visto (Sección 9.1.1 y siguiente), los residuos *studentizados* no siguen una distribución normal, a efectos prácticos y para tamaños muestrales moderados (Trocóniz (1987a), pág. 174, indica que suele bastar $N > 20$) la aproximación a la normalidad es muy buena, si las perturbaciones son a su vez normales.

Hay multitud de pruebas utilizables para contrastar ajuste a una distribución. La de Kolmogorov-Smirnov (véase Trocóniz (1987b), pág. 255) es de uso general con muestras grandes y distribuciones continuas —lo que incluye a la normal—. Hay contrastes como el de Shapiro-Wilk descrito en Shapiro and Wilk (1965) y Shapiro and Francia (1972), especializados en el contraste de la hipótesis de normalidad.

Tan útil como pueda ser una prueba estadistica convencional de normalidad, en ocasiones es útil un instrumento que permita visualizar la naturaleza y alcance de la desviación respecto a la normalidad, si existe. Los gráficos en papel normal cumplen esta finalidad.

El principio es muy simple: dada una muestra $\{x_i\}_{i=1}^N$, si procede de una distribución normal los puntos $(x_i, \Phi^{-1}(F_*(x_i)))$, en que $F_*(x_i)$ es la función de distribución empírica de la muestra, deben estar aproximdamente alineados. Véase por ejemplo Trocóniz (1987b), pág. 270.

El gráfico puede hacerse manualmente sobre papel especial ("papel normal") en que la escala vertical absorbe la transformación $\Phi^{-1}(.)$; o puede hacerse mediante ordenador en cuyo caso basta facilitar los datos y verificar la linealidad del gráfico resultante.

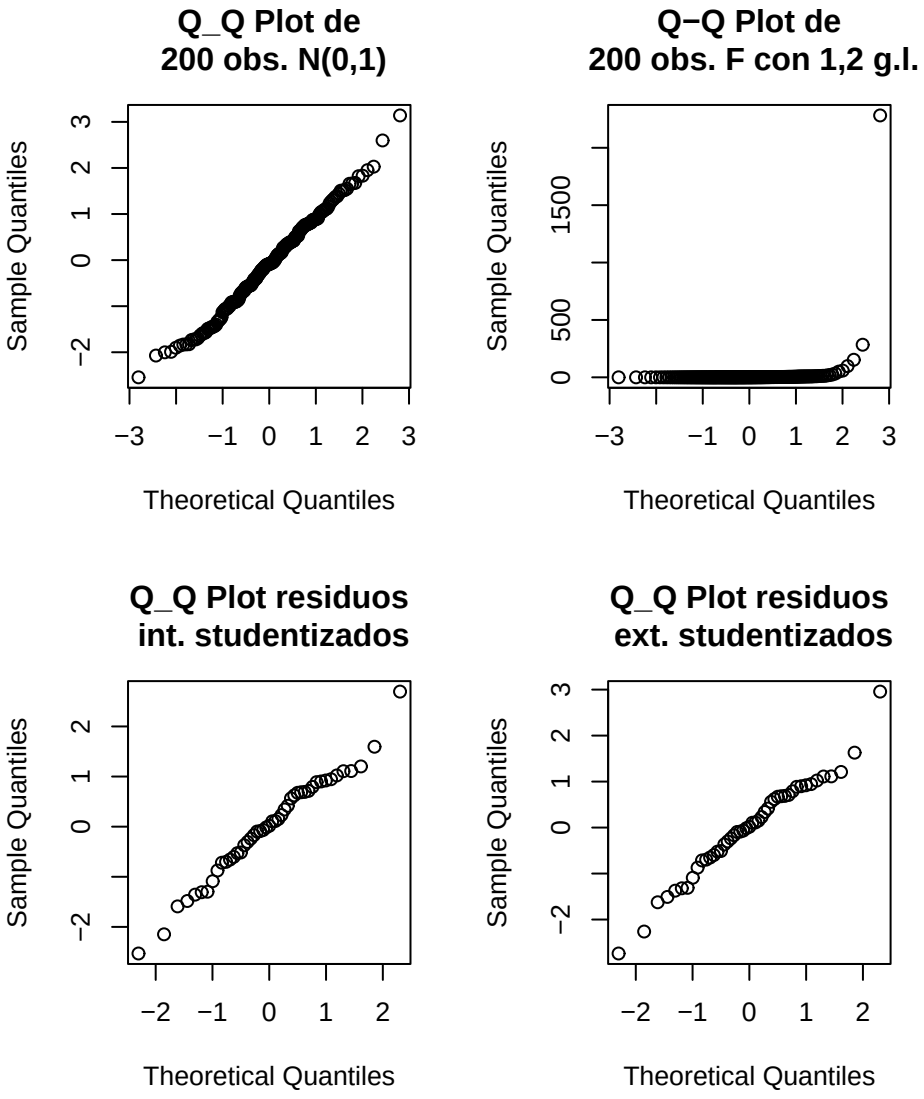
En cualquiera de los casos se cuenta con un instrumento que permite no sólo apreciar si hay desviaciones respecto de la normalidad, sino también de qué naturaleza son y a qué puntos afectan.

**R: Ejemplo 9.1** *(gráficos para contraste de normalidad de residuos)*

La Figura 9.2 se genera mediante el fragmento de código reproducido a continuación. Los dos primeros paneles recogen sendos gráficos de normalidad para una muestra normal y una muestra procedente de una $\mathcal{F}_{1,2}$; puede verse la llamativa desviación de la normalidad en este último caso.

```
--- Obtenido mediante R BATCH demo9a.R
> #
> #  Ejemplo de uso de gráficas de normalidad
> #
> postscript(file="demo9a.eps",horizontal=FALSE,
+            paper="a4",width=5,height=6)
> par(mfrow=c(2,2))
> muestra <- rnorm(200)
> qqnorm(muestra,main="Q_Q Plot de\n 200 obs. N(0,1)")
> muestra <- rf(200,1,2)
> qqnorm(muestra,main="Q-Q Plot de\n 200 obs. F con 1,2 g.l.")
> rm(muestra)
> #
> #  Probemos ahora con los residuos interna y externamente
```

Figura 9.2: Gráficos para contraste de normalidad



```
> #  estudentizados de una regresión
> #
> library(MASS)

Attaching package: 'MASS'


The following object(s) are masked _by_ .GlobalEnv :

 UScrime

> data(UScrime)
> #
> #  Ajustamos un modelo a la variable y
> #
> modelo <- lm(y ~ M + Ed + Po1 + M.F + U1 + U2 +
+                  Prob + Ineq, data =UScrime)
> #
> #  Extraemos y dibujamos los residuos. Obsérvese que
> #  NO emplearíamos para estos gráficos residuos
> #  ordinarios, por sus diferentes varianzas.
> #
> qqnorm(stdres(modelo),
+        main="Q_Q Plot residuos\n int. studentizados")
> qqnorm(studres(modelo),
+        main="Q_Q Plot residuos\n ext. studentizados")
> q()
```

Los siguientes dos paneles muestran los gráficos de normalidad correspondientes a los residuos interna y externamente *studentizados* de un mismo modelo. Puede constatarse que son casi idénticos y que sugieren un buen ajuste de la muestra a la hipótesis de normalidad.

### 9.3.6. Gráficos de residuos ordinarios frente a residuos borrados $(d_i, \hat{\epsilon}_i)$

Un residuo borrado no necesariamente es indicativo de que una observación sea muy influyente. Lo realmente sintomático es una gran divergencia entre el residuo ordinario y el residuo borrado, pues ello indica que al omitir la observación correspondiente los resultados varían mucho, al menos en el ajuste de la observación $i$-ésima.

Por ello se propone como gráfico útil en el diagnóstico de un modelo el de $\hat{\epsilon}_i$ frente a $d_i$. En general, deberíamos observar puntos aproximadamente sobre la bisectriz: $d_i \approx \hat{\epsilon}_i$. Puntos muy separados de la bisectriz corresponderían a observaciones que alteran sustancialmente la regresión.

#### CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

**9.1** Demuéstrese que $r_i^2/(N-p)$, bajo los supuestos habituales más normalidad, sigue una distribución beta, $B(\frac{1}{2}, \frac{1}{2}(N-p-1))$.

# Capítulo 10

# Selección de modelos.

## 10.1. Criterios para la comparación.

En ocasiones, ajustamos un modelo de regresión teniendo una idea clara de las variables que debemos incluir como regresores. Es más frecuente, sin embargo, el caso en que sólo tenemos una idea aproximada de la forma adecuada para nuestro modelo, y debemos decidir con criterio estadístico qué regresores deben ser incluidos.

Para enfrentar este tipo de situaciones necesitamos, por una parte, criterios de bondad de ajuste, capaces de permitirnos comparar distintos modelos ajustados a una misma muestra. Por otra, necesitamos estrategias de selección de variables que construyan de manera automática o semi-automática subconjuntos de todos los modelos posibles susceptibles de incluir el "mejor". Examinaremos en esta Sección el primer punto.

Es claro que no podemos preferir un modelo a otro simplemente porque su $SSE$ es menor, dado que toda[1] variable que incluyamos en la regresión, tenga mucha o poca relación con la variable respuesta, reducirá $SSE$. Tenemos, pues, que buscar criterios más elaborados.

### 10.1.1. Maximización de $\overline{R}^2_p$.

Se define el *coeficiente de determinación corregido* así:

$$\overline{R}^2_p = 1 - [1 - R^2_p] \times \frac{N-1}{N-p} \tag{10.1}$$

[1] Las únicas excepciones son aquellas variables correspondientes a columnas de la matriz de diseño $X$ ortogonales a $\vec{y}$, o que son combinación lineal exacta de columnas correspondientes a variables ya presentes entre los regresores.

haciendo referencia el subíndice $p$ al número de regresores presentes en el modelo. Si reescribimos la ecuación (10.1) en la forma:

$$1 - \overline{R}_p^2 = [1 - R_p^2] \times \frac{N-1}{N-p} \quad (10.2)$$

$$= \frac{SSE_p}{SST} \times \frac{N-1}{N-p} \quad (10.3)$$

vemos que mientras que el primer término de la derecha de (10.3) es monótono no creciente con $p$, el segundo es monótono creciente. Por consiguiente, el producto de ambos[2] puede crecer o decrecer al crecer $p$.

Es frecuente por ello utilizar $\overline{R}_p^2$ como criterio de ajuste. Aunque útil, veremos sin embargo que debe complementarse con otros criterios. Su exclusiva aplicación da lugar con gran probabilidad a modelos sobreparametrizados, como pone de manifiesto el siguiente teorema.

**Teorema 10.1** *El estadístico $\overline{R}_p^2$ crece con la introducción de un parámetro en la ecuación de regresión si el estadístico $Q_h$ asociado al contraste de significación de dicho parámetro verifica $Q_h > 1$.*

DEMOSTRACION:[3]

Para contrastar la significación del $(p+1)$-ésimo parámetro, empleamos (Sección 4.2, pág. 45):

$$Q_h = \frac{SSE_p - SSE_{p+1}}{SSE_{p+1}} \times \frac{N-p-1}{1} \quad (10.4)$$

$$= \frac{(R_{p+1}^2 - R_p^2)}{1 - R_{p+1}^2} \times \frac{N-p-1}{1} \quad (10.5)$$

de donde:

$$(1 - R_{p+1}^2)Q_h = (R_{p+1}^2 - R_p^2)(N-p-1) \quad (10.6)$$

$$Q_h - Q_h R_{p+1}^2 = (N-p-1)R_{p+1}^2 - (N-p-1)R_p^2 \quad (10.7)$$

$$Q_h + (N-p-1)R_p^2 = R_{p+1}^2 \left[(N-p-1) + Q_h\right] \quad (10.8)$$

Despejando $R_{p+1}^2$ tenemos:

$$R_{p+1}^2 = \frac{Q_h + (N-p-1)R_p^2}{(N-p-1) + Q_h} \quad (10.9)$$

$$= \frac{\frac{1}{N-p-1}Q_h + R_p^2}{1 + \frac{1}{N-p-1}Q_h} \quad (10.10)$$

[2] Expresiones como la anterior con un término función de la suma de cuadrados de los residuos y otro interpretable como "penalización" por la introducción de parámetros adicionales, son ubicuas en la literatura estadística. La $C_p$ de Mallows que se examina más abajo tiene la misma forma, como muchos criterios de ajuste utilizados sobre todo en el análisis de series temporales: Criterio de Información de Akaike (AIC), FPE, BIC, etc.

[3] Sigue a Haitovsky (1969).

De (10.10) y de la definición de $\overline{R}^2_{p+1}$ se deduce que:

$$\overline{R}^2_{p+1} = 1 - [1 - R^2_{p+1}] \times \frac{N-1}{(N-p-1)} \tag{10.11}$$

Sustituyendo en esta expresión (10.10) llegamos a:

$$\overline{R}^2_{p+1} = 1 - \frac{[1-R^2_p]}{\frac{N-p-1+Q_h}{N-p-1}} \times \frac{N-1}{N-p-1} \tag{10.12}$$

$$= 1 - [1 - R^2_p]\frac{N-1}{N-p-1+Q_h} \tag{10.13}$$

$$= \underbrace{1 - [1 - R^2_p]\frac{N-1}{N-p}}_{\overline{R}^2_p} \underbrace{\frac{N-p}{N-p-1+Q_h}}_{t} \tag{10.14}$$

Es evidente de (10.14) que $\overline{R}^2_{p+1} \geq \overline{R}^2_p$ si $Q_h > 1$, y viceversa[4]. Maximizar $\overline{R}^2_p$ implica introducir en la ecuación de regresión todos aquellos regresores cuyo estadístico $Q_h$ sea superior a la unidad; pero esto ocurre con probabilidad $\approx 0{,}50$ incluso cuando $h\colon \beta_i = 0$ es cierta. Consecuentemente, el emplear este criterio en exclusiva conduciría con gran probabilidad al ajuste de modelos sobreparametrizados.

### 10.1.2. Criterio $C_p$ de Mallows.

Supongamos que la variable aleatoria $Y$ se genera realmente como prescribe el modelo $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$ , no obstante lo cual ajustamos el modelo equivocado $Y = \tilde{X}\tilde{\beta} + \vec{\epsilon}$ con $p$ parámetros. Una vez estimado, dicho modelo suministra las predicciones $\hat{Y}^{(p)}$. Un criterio para evaluar la adecuación del modelo estimado al real, sería el error cuadrático medio

$$ECM = E(\hat{Y}^{(p)} - X\vec{\beta})'(\hat{Y}^{(p)} - X\vec{\beta}) \tag{10.15}$$

que sumando y restando $E(\hat{Y}^{(p)})$ dentro de cada paréntesis podemos descomponer así:

$$ECM = E\left[(\hat{Y}^{(p)} - E(\hat{Y}^{(p)}))'(\hat{Y}^{(p)} - E(\hat{Y}^{(p)}))\right] + E\left[(E(\hat{Y}^{(p)}) - X\vec{\beta})'(E(\hat{Y}^{(p)}) - X\vec{\beta})\right] \tag{10.16}$$

$$= \text{Var}(\hat{Y}^{(p)}) + (\text{Sesgo})^2. \tag{10.17}$$

El primer término no ofrece dificultad. Como

$$\hat{Y}^{(p)} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\vec{Y} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'(X\vec{\beta} + \vec{\epsilon}), \tag{10.18}$$

tenemos que

$$E[\hat{Y}^{(p)}] = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'X\vec{\beta}$$

y

$$\begin{aligned}((\hat{Y}^{(p)} - E(\hat{Y}^{(p)}))'((\hat{Y}^{(p)} - E(\hat{Y}^{(p)})) &= \vec{\epsilon}\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\vec{\epsilon} \\ &= \vec{\epsilon}\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\vec{\epsilon} \\ &\sim \sigma^2\chi^2_p. \end{aligned} \tag{10.19}$$

[4]Obsérvese que si el término $t$ en (10.14) fuera la unidad —lo que acontece cuando $Q_h = 1$—, el lado derecho sería precisamente $\overline{R}^2_p$. Si $Q_h > 1$, $t$ es menor que 1 y, como sólo multiplica al sustraendo en (10.14), el resultado es *mayor* que $\overline{R}^2_p$.

Falta el término de sesgo. Observemos que

$$E[\underbrace{(\vec{Y} - \hat{Y}^{(p)})'(\vec{Y} - \hat{Y}^{(p)})}_{SSE}] = \underbrace{E\left[(X\vec{\beta} - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'X\vec{\beta})'(X\vec{\beta} - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'X\vec{\beta})\right]}_{\text{(Sesgo)}^2} + E\left[\vec{\epsilon}\,'(I - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}')\vec{\epsilon}\right].$$

Por consiguiente,

$$(\text{Sesgo})^2 = E[SSE] - E[\sigma^2\chi^2_{N-p}]. \tag{10.20}$$

Sustituyendo en (10.17) tenemos entonces que

$$\begin{aligned} ECM &= E\left[SSE - \sigma^2\chi^2_{N-p}\right] + E\left[\sigma^2\chi^2_p\right] && (10.21)\\ &= E[SSE] - \sigma^2(N-p) + \sigma^2 p, && (10.22)\end{aligned}$$

y por consiguiente:

$$\frac{ECM}{\sigma^2} = E\left[\frac{SSE}{\sigma^2}\right] - N + 2p. \tag{10.23}$$

Minimizar esta última expresión es lo mismo que minimizar

$$E\left[\frac{SSE}{\sigma^2}\right] + 2p, \tag{10.24}$$

ya que $N$ es constante. Como quiera que el valor medio en la expresión anterior no puede ser calculado y $\sigma$ es desconocida, todo lo que podemos hacer es reemplazar (10.24) por la expresión análoga,

$$C_p = \frac{SSE}{\hat{\sigma}^2} + 2p. \tag{10.25}$$

A esta última expresión se la conoce como $C_p$ de Mallows.

Para que se verifique la aproximación en (10.25) es preciso que $\hat{\sigma}^2 \approx \sigma^2$, lo que se consigue si la muestra es lo suficientemente grande y $\hat{\sigma}^2 = SSE^{(N-p-k)}/(N-p-k)$, estando entre los $(p+k)$ regresores incluídos los $p$ necesarios. Incluso aunque entre dichos $(p+k)$ regresores haya algunos innecesarios, $\hat{\sigma}^2$ es insesgado; el precio que se paga por emplear más parámetros de los debidos en la estimación de $\sigma^2$ es una reducción en el número de grados de libertad (véase Sección 5.2).

De acuerdo con el criterio de Mallows, seleccionaremos el modelo que minimice $C_p$. La expresión (10.25) es otro ejemplo de criterio de ajuste con penalización. Cada nuevo parámetro que introducimos, reduce quizá $SSE$, pero esta reducción tiene un precio: el incremento del segundo sumando de (10.25) en 2. El efecto neto indica si el nuevo regresor es o no deseable.

> **Observación 10.1** De acuerdo con el criterio $C_p$ de Mallows, dada una ecuación de regresión con unos ciertos regresores presentes, introduciremos un nuevo regresor si éste puede "pagar" su inclusión reduciendo $SSE$ en, al menos, dos veces $\hat{\sigma}^2$. La maximización de $\overline{R}^2_p$, en cambio, requeriría en análoga situación introducir el mismo regresor si disminuye $SSE$ en al menos una vez $\hat{\sigma}^2$. El criterio $C_p$ de Mallows es más restrictivo[5].

[5]La comparación es aproximada tan sólo. El valor de $\hat{\sigma}^2$ que se emplea en el criterio $C_p$ se obtiene, típicamente, ajustando el modelo más parametrizado (esto minimiza el riesgo de introducir sesgos en la estimación de $\sigma^2$, aunque seguramente nos hace despilfarrar algunos grados de libertad). Por el contrario, al utilizar el criterio basado en $\overline{R}^2_p$ introducimos el nuevo regresor si $Q_h > 1$ en (10.4), es decir, si la disminución $SSE_p - SSE_{p+1}$ en la suma de cuadrados de los residuos es mayor que $\hat{\sigma}^2 = SSE_{p+1}/(N-p-1)$, varianza estimada *en el modelo con* $p+1$ *regresores*.

**Observación 10.2** Un estadístico se enfrenta con frecuencia a este dilema en su trabajo. ¿Hasta dónde procede llevar la complejidad del modelo a emplear? ¿Qué mejora en el ajuste de un modelo a la muestra justifica la adición de un nuevo parámetro?. O, si se prefiere, ¿Cuán afilada debe ser la navaja de Ockham? En el caso del modelo de regresión lineal, el criterio $C_p$ suministra seguramente una navaja con el filo adecuado; argumentos alternativos llevan a criterios equivalentes o similares al $C_p$. Es un hecho notable y llamativo que por diversas vías se llegue siempre a análogos resultados, que tienen en común el medir la complejidad del modelo empleado como una función lineal o aproximadamente lineal del número de sus parámetros; más sobre esto en la Sección 10.1.5. En la Sección 10.1.4 se introduce la idea de la *validación cruzada*, que proporciona una forma alternativa de evaluar la bondad de ajuste de un modelo soslayando el empleo de una penalización basada en el número de parámetros.

### 10.1.3. Criterio AIC

Relacionado con el criterio $C_p$ de Mallows, aunque válido de modo mucho más general y motivado de modo muy diferente, está el criterio AIC (**A**kaike's **I**nformation **C**riterion, o **A**n **I**nformation **C**riterion). Consiste en seleccionar el modelo minimizando

$$AIC(p) = -2\log_e\left[\max_{\vec{\theta}} \text{ verosimilitud}(\vec{x}\,,\vec{\theta}\,)\right] + 2p$$

El primer término en la expresión anterior es, como en la $C_p$ de Mallows, una medida de bondad de ajuste (disminuye al crecer el máximo de la verosimilitud); el segundo penaliza el número de parámetros en $\vec{\theta}$. Puede verse una justificación en Akaike (1972) (y en Akaike (1974), Akaike (1991)). Una explicación simplificada que sigue esencialmente a de Leeuw (2000) puede encontrarse en Tusell (2003), Sección 9.2.2.

Cuando consideremos modelos de regresión lineal con normalidad, el uso de los criterios AIC y $C_p$ daría resultados exactamente equivalentes si conociéramos $\sigma^2$ (ambos criterios difieren en tal caso en una constante; ver Venables and Ripley (1999a), pág. 185). Cuando $\sigma^2$ es desconocida y ha de ser estimada a partir de los datos, ambos criterios pueden diferir, pero son a efectos prácticos intercambiables. El criterio AIC no obstante es de ámbito mucho más general, y puede ser utilizado dondequiera que tengamos una verosimilitud, sea o no normal la distribución generadora de la muestra.

### 10.1.4. Residuos borrados y validación cruzada

Hemos visto que el problema de emplear como criterio para la selección de modelos alguno de los estadísticos de ajuste obvios (suma de cuadrados residual, $R^2$, o similar) estriba en que hay que tomar en consideración el diferente número de parámetros en cada modelo.

El problema consiste en que, al incrementar el número de parámetros, el modelo puede "seguir" más a la muestra, ajustando no sólo el comportamiento predecible sino incluso el puramente aleatorio Se adapta muy bien a *una* muestra —la que hemos empleado para estimarlo—, pero quizá no a otras.

Una solución consistiría en estimar los modelos con una muestra (muestra de entrenamiento o aprendizaje) y evaluarlos examinando su comportamiento en la predicción de *otra* diferente (muestra de validación). Actuando así, estaríamos a salvo de impresiones excesivamente optimistas: la suma de cuadrados de los residuos o $R^2$ que calculáramos para cada modelo reflejaría su capacidad de generalización: su comportamiento con otras observaciones distintas de las que han servido para estimarlo.

Lamentablemente, esto requiere dividir nuestra disponibilidad de observaciones en dos grupos: uno para estimar y otro para validar. El obtener un diagnóstico realista por este procedimiento requiere sacrificar en aras de la validación una preciosa fracción de muestra que habría permitido, quizá, estimar mejor.

¿Realmente es esto así? No; una vez que hemos decidido por el procedimiento anterior de fraccionar la muestra en dos para seleccionar el modelo mejor, podemos emplear *todas* las observaciones en reestimarlo.

La idea de la *validación cruzada* incorpora una mejora adicional al planteamiento anterior. No tenemos necesariamente que usar sólo una fracción de la muestra para validar. Podemos dividir la muestra en dos (o más) partes y emplear todas ellas en la validación. El ejemplo que sigue detalla los pasos a seguir haciendo validación cruzada por mitades.

**Ejemplo 10.1** Consideremos una muestra de tamaño $N = 100$. Tenemos una colección de $K$ modelos $\mathcal{M}_i$, $i = 1, \ldots, K$, posiblemente con diferente número de parámetros, de entre los que queremos seleccionar uno. Podemos dividir la muestra en dos trozos, $A$ y $B$, de tamaños respectivos $N_A = N_B = 50$, y proceder así:

1. Con la muestra $A$ estimaremos cada uno de los modelos $\mathcal{M}_i$.
2. Examinaremos el ajuste de los modelos así estimados a la muestra $B$, computando sumas de cuadrados residuales para cada uno de los modelos, $SSE_i^{(A)}$.
3. Con la muestra $B$ estimaremos cada uno de los modelos $\mathcal{M}_i$.
4. Examinaremos el ajuste de los modelos así estimados a la muestra $A$, computando sumas de cuadrados residuales para cada uno de los modelos, $SSE_i^{(B)}$
5. Tanto $SSE_i^{(A)}$ como $SSE_i^{(B)}$ son estimaciones de las sumas de cuadrados de los residuos del modelo $\mathcal{M}_i$, cuando se utiliza en predicción sobre una muestra diferente de la que se ha empleado en su estimación. Podemos promediar ambas para obtener un único estadístico, $SSE_i = \frac{1}{2}(SSE_i^{(A)} + SSE_i^{(B)})$.
6. Seleccionaremos el modelo $\mathcal{M}_i$ tal que $SSE_i$ es mínimo.

Observemos que nada nos constriñe a dividir la muestra en dos partes; podríamos dividirla en $s$ partes, y proceder exactamente del mismo modo: utilizaríamos sucesivamente $s-1$ partes para estimar y la restante para evaluar $SSE_i^{(\ell)}$, $\ell = 1, \ldots, s$, (suma de cuadrados de los residuos al predecir en la muestra $\ell$ mediante el modelo $\mathcal{M}_i$ estimado con las restantes observaciones). Promediando los $s$ valores $SSE_i^{(\ell)}$ obtendríamos el $SSE_i$ del modelo $\mathcal{M}_i$.

El caso extremo consistiría en tomar $s = N$, y realizar el proceso dejando cada vez fuera una única observación (validación cruzada de tipo *leave one out*).

En muchas situaciones esta estrategia puede requerir un esfuerzo de cálculo formidable: ¡cada modelo ha de ser reestimado $(N-1)$ veces, dejando cada vez fuera de la muestra de estimacion una observación diferente! En regresión lineal, sin embargo, la diferencia entre la predicción de la observación $i$-ésima haciendo uso de todas las restantes y el valor observado de la misma es, simplemente, el residuo borrado, de cómoda y rápida obtención (véase Sección 9.1.4). Por tanto, utilizando la notación de dicha Sección,

$$\begin{aligned} SSE_i^{\ell} &= d_\ell^2 \qquad (\ell = 1, \ldots, N) \\ SSE_i &= N^{-1} \sum_{\ell=1}^{N} SSE_i^{\ell}. \end{aligned}$$

El modelo seleccionado es aquél al que corresponde un $SSE_i$ más pequeño[6].

### 10.1.5. Complejidad estocástica y longitud de descripción mínima*

En esencia, seleccionar un modelo entraña adoptar un compromiso entre la bondad de ajuste y la complejidad, medida por el número de sus parámetros. Sabemos que un modelo lineal suficientemente parametrizado podría ajustar perfectamente la muestra, pero que ello no significa que sea idóneo: puede tener muy poca capacidad de generalización. Por el contrario, un modelo que no incluya los parámetros suficientes dara un ajuste susceptible de mejora. Se trata de alcanzar un equilibrio entre los dos objetivos en contradicción: un modelo dando buen ajuste y con los mínimos parámetros precisos.

Una aproximación intuitivamente atrayente al problema es la siguiente: tratemos de dar una descripción tan corta como sea posible de la evidencia (la muestra). Esto puede de nuevo verse como una apelación al principio de Ockham: construir "explicaciones" de la realidad que hacen uso del mínimo número de entidades.

La aproximación propuesta exige medir la longitud de la descripción que hagamos, y podemos para ello hacer uso de la Teoría de la Información. No podemos elaborar esta cuestión con detalle aquí (véase una buena introducción en Rissanen (1989), y detalles en Legg (1996)). En esencia, dado un modelo probabilistico podemos describir o codificar unos datos de modo compacto asignando a los más "raros" (menos probables) los códigos más largos.

> **Observación 10.3** Esta estrategia, de sentido común, es la que hace que al codificar en el alfabeto telegráfico de Morse la letra "e" (muy frecuente en inglés) se adoptara el código `.`, reservando los códigos más largos para caracteres menos frecuentes (ej: `- . . -` para la "x").

*Además* de codificar los datos tenemos que codificar los parámetros del modelo probabilistico. La longitud total de descripción de la muestra $\vec{y}$ cuando hacemos uso del modelo probabilístico $\mathcal{M}_k$ haciendo uso del vector de parámetros $\vec{\theta}_k$ es entonces

$$MDL(\mathcal{M}_k; \vec{y}) = (\text{Código necesario para } \vec{y}) \quad (10.26)$$

$$+ \quad (\text{Código necesario para } \vec{\theta}_k). \quad (10.27)$$

Un mal ajuste hará que el primer sumando sea grande; los datos muestrales se desvían mucho de lo que el modelo predice. Un modelo con un perfecto ajuste tendría un primer sumando nulo (porque las $\vec{y}$ se deducirían exactamente del modelo, y no requerirían ser codificadas), pero requeriría quizá muchos parámetros incrementando el segundo sumando.

El criterio MDL propone seleccionar el modelo $\mathcal{M}_k$ que minimiza (10.27). En el caso de modelos de regresión, el criterio MDL da resultados íntimamente emparentados asintóticamente con los precedentes (suma de cuadrados PRESS y $C_p$); véanse detalles en Rissanen (1989), Cap. 5.

## 10.2. Selección de variables.

Una aproximación ingenua al problema consistiría en estudiar la reducción en un cierto criterio ($SSE$, $\overline{R}^2_p$, $C_p$, ...) originada por la introducción de cada variable, y

[6]Nótese que $SSE_i$ es lo que se conoce también como suma de cuadrados de los residuos predictiva o PRESS; véase nota a pie de página de la Sección 9.1.4.

retener como regresores todas aquellas variables que dieran lugar a una reducción significativa. Desgraciadamente, esta estrategia no tiene en cuenta el hecho de que, a menos que las columnas de la matriz de diseño $X$ sean ortogonales, la reducción en $SSE$ originada por la inclusión de una variable depende de qué otras variables estén ya presentes en la ecuación ajustada.

Se impone, pues, emplear procedimientos más sofisticados. Relacionamos algunos de los más utilizados.

### 10.2.1. Regresión sobre todos los subconjuntos de variables.

De acuerdo con el párrafo anterior, la adopción de una estrategia ingenua podría dificultar el hallazgo de un modelo adecuado. Por ejemplo, puede bien suceder que una variable $X_i$, que debiera ser incluída en el modelo, no origine una reducción significativa de $SSE$ cuando la introducimos después de $X_j$. Si esto ocurre, es claro que $X_i$ no mostrará sus buenas condiciones como regresor mas que si es introducida con $X_j$ ausente.

Una posible solución sería, dados $p$ regresores, formar todos los posibles subconjuntos de regresores y efectuar todas las posibles regresiones, reteniendo aquélla que, de acuerdo con el criterio de bondad de ajuste que hayamos adoptado, parezca mejor.

El inconveniente es el gran volumen de cálculo que es preciso realizar. Piénsese que con $p$ regresores pueden estimarse $2^p - 1$ diferentes regresiones. Si $p = 5$, $2^p - 1 = 31$; pero si $p = 10$, $2^p - 1 = 1023$, y para $p > 20$ habría que realizar por encima de un millón de regresiones. Hay procedimientos para reducir y agilizar el cálculo[7], pero aún así éste puede resultar excesivo.

### 10.2.2. Regresión escalonada *(stepwise regression)*.

Se trata de un procedimiento muy utilizado que, aunque no garantiza obtener la mejor ecuación de regresión, suministra modelos que habitualmente son óptimos o muy próximos al óptimo, con muy poco trabajo por parte del analista. Describiremos el procedimiento de regresión escalonada "hacia adelante" *(forward selection procedure)*; la regresión escalonada "hacia atrás" *(backward elimination)* o mixta son variantes fáciles de entender.

En cada momento, tendremos una ecuación de regresión provisional, que incluye algunas variables (regresores incluidos) y no otras (regresores ausentes). Al comienzo del procedimiento, la ecuación de regresión no incluye ningún regresor. El modo de operar es entonces el siguiente:

1. Calcular los estadísticos $Q_h$ para todos los regresores ausentes $(h\colon \beta_i = 0)$.
2. Sea $Q_h^*$ el máximo estadístico de los calculados en 1). Si $Q_h^* < \mathcal{F}$, siendo $\mathcal{F}$ un umbral prefijado, finalizar; la ecuación provisional es la definitiva. Si, por el contrario, $Q_h^* \geq \mathcal{F}$, se introduce la variable correspondiente en la ecuación de regresión.
3. Si no quedan regresores ausentes, finalizar el procedimiento. En caso contrario, reiniciar los cálculos en 1).

En suma, se trata de introducir las variables de una en una, por orden de mayor contribución a disminuir $SSE$, y mientras la disminución sea apreciable.

[7] Véase Seber (1977), pag. 349 y ss.

El procedimiento de regresion "hacia atrás" procede de manera análoga, pero se comienza con una ecuación que incluye todos los regresores, y se van excluyendo de uno en uno, mientras el incremento en $SSE$ que dicha exclusión origine no sea excesivo. En el procedimiento míxto, por fin, se alterna la inclusión y exclusión de variables en la recta de regresión; ello permite que una variable incluida sea posteriormente desechada cuando la presencia de otra u otras hacen su contribución a la reducción de $SSE$ insignificante.

Los criterios de entrada y salida de variables se fijan especificando sendos valores $\mathcal{F}_{\text{entrada}}$ y $\mathcal{F}_{\text{salida}}$ que deben ser superados (no alcanzados) por el $Q_h^*$ correspondiente para que una variable pueda ser incluida (excluida) en la regresión. Ambos umbrales pueden ser el mismo. Mediante su selección adecuada, puede lograrse un algoritmo "hacia adelante" puro (fijando $\mathcal{F}_{\text{salida}} = 0$, con lo que se impide el abandono de cualquier variable introducida), "hacia atrás" puro (fijando $\mathcal{F}_{\text{entrada}}$ muy grande, y comenzando con una ecuación de regresión que incluye todas las variables), o un procedimiento mixto arbitrariamente próximo a cualquiera de los dos extremos[8].

> **R: Ejemplo 10.1** *(selección automática de modelos)* El ejemplo siguiente muestra el uso de las funciones `leaps` (en el paquete del mismo nombre) para hacer regresión sobre todos los subconjuntos con criterios $R^2$, $\overline{R}^2$ ó $C_p$, `stepAIC` (en el paquete `MASS`) para hacer regresión escalonada con criterio AIC y algunas otras funciones ancilares.
>
> La Figura 10.1 muestra el comportamiento típico de los criterios $C_p$ y $\overline{R}^2$. Se aprecia que, aunque de forma no muy notoria en este caso, el criterio $\overline{R}^2$ tiende a seleccionar modelos más parametrizados.

```
--- Obtenido mediante R BATCH demo10.R
```

## 10.3. Modelos bien estructurados jerárquicamente

La facilidad con que los algoritmos presentados en este Capítulo producen modelos candidatos no debe hacer que el analista delegue demasiado en ellos. Un modelo ha de ser consistente con los conocimientos fiables que se tengan acerca del fenómeno bajo estudio. Debe ser también interpretable. Prestemos algo de atención a este último requerimiento.

Imaginemos un modelo como el siguiente:

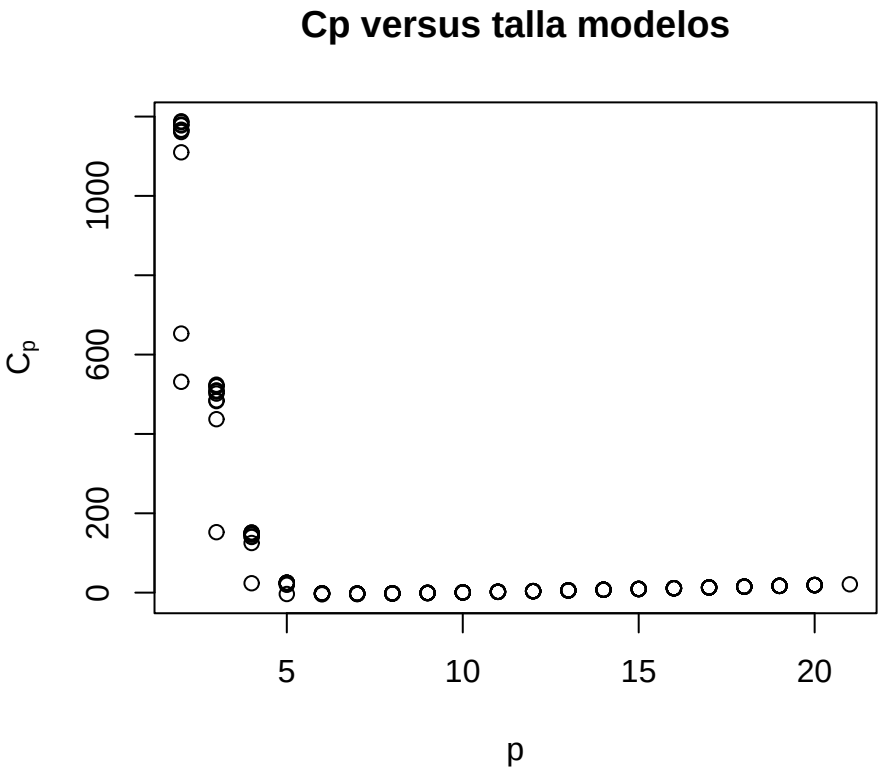
$$y \quad = \quad \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon. \tag{10.28}$$

En un caso así, frecuentemente el interés se centrará en dilucidar si la relación de $X$ con $Y$ es lineal o cuadrática —es decir, en contrastar la hipótesis $h : \beta_2 = 0$—.

Es frecuentemente el caso que $X$ se mide en unidades en que tanto la escala como el origen son arbitrarios (como ocurría, por ejemplo, en el Ejercicio 2.8, pág. 27); y sería inconveniente que el contraste de $h$ dependiera del origen y de la escala empleadas.

[8] Podría pensarse en fijar niveles de significación para la entrada y salida de variables. Esto no se hace porque serían considerablemente arduos de computar; obsérvese que en un procedimiento *stepwise* se selecciona para entrar o salir de la ecuación de regresión la variable con un $Q_h$ mayor (menor). Bajo la hipótesis de nulidad del correspondiente parámetro, un $Q_h$ cualquiera se distribuye como una $\mathcal{F}$ de Snedecor con grados de libertad apropiados. *El mayor* (o menor) de los estadísticos $Q_h$ en cada etapa, sigue una distribución diferente (véase Capítulo 6). El nivel de significación asociado al contraste implícito en la inclusión o exclusión de un regresor *no es* la probabilidad a la derecha (o izquierda) de $\mathcal{F}_{\text{entrada}}$ (o $\mathcal{F}_{\text{salida}}$) en una distribución $\mathcal{F}$ con grados de libertad apropiados.

Figura 10.1: Valores de $C_p$ y $\overline{R}^2$ para 141 modelos ajustados a los datos `UScrime`



Lo menos que debemos esperar de nuestra inferencia es que sea invariante frente a cambios en las unidades de medida.

Si en (10.28) reemplazamos $X$ por $Z = aX + b$, obtenemos

$$\begin{aligned} y &= \beta_0 + \beta_1(aX + b) + \beta_2(aX + b)^2 + \epsilon \\ &= (\beta_0 + \beta_1 b + \beta_2 b^2) + (\beta_1 a + 2ab\beta_2)X + a^2\beta_2 X^2 + \epsilon \\ &= \beta_0^* + \beta_1^* X + \beta_2^* X^2 + \epsilon. \end{aligned} \tag{10.29}$$

En este nuevo modelo, $\beta_2^* = a^2\beta_2$ absorbiendo el cambio de escala en la $X$. Es fácil ver que es equivalente contrastar $h : \beta_2 = 0$ en (10.28) o $h : \beta_2^* = 0$ en (10.29); el contraste de la hipótesis "efecto cuadrático de $X$ sobre $Y$", al menos, no se altera por el cambio de unidades. Sin embargo, sean cuales fueren $\beta_1$ y $\beta_2$, habrá coeficientes $a$, $b$ anulando $\beta_1^* = (\beta_1 a + 2ab\beta_2)$ en (10.29). Ello hace ver que:

- No tiene sentido contrastar efecto lineal en un modelo que incluye término cuadrático, porque el contraste tendría un resultado diferente dependiendo de las unidades de medida.
- La inclusión de un término en $X^2$ *debe* ir acompañada de un término lineal y constante, si queremos que el modelo sea invariante frente a cambios en el origen y la escala.

La conclusión que extraemos es que los términos de orden superior deben estar acompañados de todos los términos de orden inferior —es decir, si incluimos un término cúbico, deben también existir términos cuadráticos y lineales, etc.—. Un modelo que cumpla con dicho requisito se dice que está jerárquicamente estructurado y en él podemos contrastar no nulidad del coeficiente del término jerárquico de orden superior, pero no de los inferiores. La misma conclusión es de aplicación a términos recogiendo interacciones: si introducimos una variable compuesta como $X_i X_j$ en el modelo, $X_i$ y $X_j$ deben también ser incluidas. Se suele decir que un modelo jerárquicamente bien estructurado verifica *restricciones de marginalidad* y que, por ejemplo, $X_i$ y $X_j$ son ambas marginales a $X_i X_j$.

Si regresamos al Ejercicio 2.8 en que se argüía la necesidad de utilizar un término $\beta_0$ veremos que se trata del mismo problema: necesitamos el término jerárquico inferior (la constante) cuando incluimos $X$ dado que las unidades y el origen son arbitrarios. No es imposible que un modelo sin $\beta_0$ sea adecuado, pero lo normal es lo contrario.

Dependiendo de los programas que se utilicen, un algoritmo puede eliminar del modelo de regresión un término jerárquico inferior manteniendo otro de orden superior. Es responsabilidad del analista garantizar que ello no ocurra, manteniendo la interpretabilidad de los parámetros en toda circunstancia.

## CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

**10.1** Supongamos que hacemos regresión escalonada "hacia adelante". ¿Qué valor de $\mathcal{F}_{\text{entrada}}$ equivaldría a introducir regresores en el modelo en tanto en cuanto incrementen $\overline{R}_p^2$?

**10.2** Las estrategias de regresión escalonada descritas (hacia adelante, hacia atrás, o mixta) exploran un subconjunto de los modelos posibles, añadiendo (omitiendo) en cada momento el regresor que parece con mayor (menor) capacidad explicativa de la variable respuesta. Puede perfectamente alcanzarse un óptimo

local, al llegarse a un modelo en el que no es posible mejorar el criterio elegido ($C_p$, o cualquier otro) añadiendo u omitiendo regresores, pese a existir otro modelo mejor en términos de dicho criterio. ¿Mejoran nuestras expectativas de encontrar el óptimo global mediante regresión escalonada cuando las columnas de la matriz $X$ de regresores son ortogonales? Justifíquese la respuesta.

**10.3** En la Observación 10.1 se comparan los criterios de selección de modelos consistentes en maximizar $\overline{R}_p^2$ y $C_p$, viendo que el segundo es en general más restrictivo.

Consideremos ahora dos posibles modelos $A$ y $B$ de regresión con sumas de cuadrados de los residuos respectivamente $SSE_A$ y $SSE_B$. El primer modelo utiliza sólo un subconjunto de los regresores presentes en el segundo (por tanto, $SSE_A \geq SSE_B$).

Para escoger entre los modelos $A$ y $B$ podríamos adoptar uno de los siguientes criterios:

1. Seleccionar el modelo $B$ si la disminución en la suma de cuadrados respecto al modelo $A$ es estadísticamente significativa, es decir, si:

$$Q_h = \frac{(SSE_A - SSE_B)}{q\hat{\sigma}^2} > \mathcal{F}^{\alpha}_{q,N-(p+q)}$$

   siendo $p$ el número de parámetros presentes en $A$ y $q$ el de los adicionales presentes en $B$.

2. Seleccionar el modelo $B$ si su estadístico $C_p$ es menor.

Supongamos además que el modelo $B$ es el más parametrizado de los posibles (incluye todas las variables de que disponemos). ¿Qué relación existe entre ambos criterios?

# Capítulo 11

# Transformaciones

## 11.1. Introducción

Nada nos obliga a utilizar los regresores o la variable respuesta tal cual; es posible que la relación que buscamos entre una y otros requiera para ser expresada realizar alguna transformación. Por ejemplo, si regresáramos el volumen de sólidos aproximadamente esféricos sobre sus mayores dimensiones, obtendríamos probablemente un ajuste muy pobre; sería mucho mejor, en cambio, regresando el volumen sobre *el cubo* de la mayor dimensión —dado que la fórmula del volumen de una esfera es $\frac{4}{3}\pi r^3$, y cabría esperar una relación similar en los sólidos aproximadamente esféricos que manejamos—.

En el ejemplo anterior, bastaba tomar un regresor —la mayor dimensión— y elevarla al cubo para obtener un ajuste mejor. Además, la naturaleza del problema y unos mínimos conocimientos de Geometría sugieren el tipo de transformación que procede realizar. En otros casos, la transformación puede distar de ser obvia. En ocasiones, es la variable respuesta la que conviene transformar. En las secciones que siguen se muestran algunos procedimientos para seleccionar un modelo, acaso transformando regresores, variable respuesta, o ambas cosas.

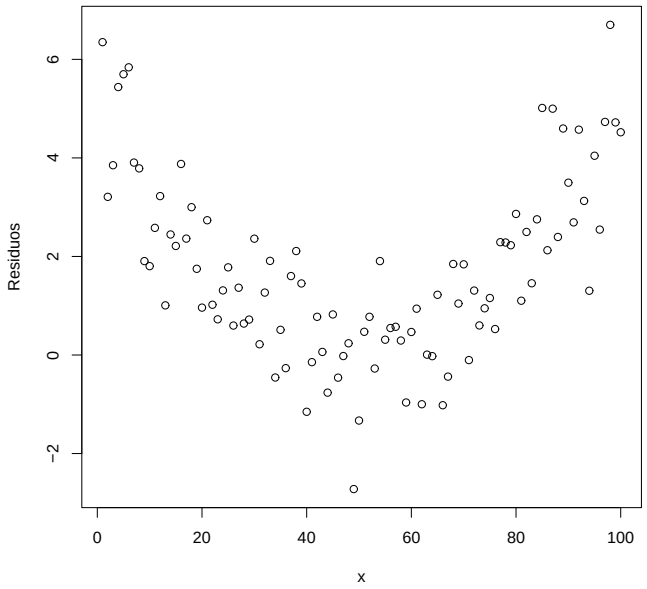
## 11.2. Transformaciones de los regresores

En ocasiones, teoría o conocimientos previos acerca del funcionamiento del fenómeno bajo análisis puede sugerir transformaciones en los regresores. Alternativamente podemos recurrir a métodos exploratorios, gráficos o no. En lo que sigue se mencionan algunas posibilidades.

### 11.2.1. Gráficos de residuos frente a regresores

Se trata de representar gráficamente los residuos en ordenadas frente a cada uno de los regresores en abscisas. La motivación es muy simple: los residuos recogen la fracción de la respuesta que el modelo no ha podido recoger. Si observamos alguna pauta al representar dichos residuos frente a un regresor, podemos intuir la transformación precisa en dicho regresor. Por ejemplo, en la Figura 11.1 se muestran residuos que frente a los valores de $X_i$ toman forma de parábola; ello sugiere introducir el regresor $X_i^2$. En efecto, esto permitiría recoger una parte de $Y$ de la que el modelo actual no da cuenta, y que por este motivo aflora en los residuos.

Figura 11.1: Disposición de residuos sugiriendo una transformación cuadrática del regresor $X_i$




### 11.2.2. Transformaciones de Box-Tidwell

Consideremos los regresores $X_1, \ldots, X_p$ y transformaciones de los mismos definidas del siguiente modo:

$$W_j = \begin{cases} X_j^{\alpha_j} & \text{si } \alpha_j \neq 0, \\ \ln(X_j) & \text{si } \alpha_j = 0. \end{cases} \tag{11.1}$$

Para diferentes valores de $\alpha_j$, la transformación (11.1) incluye muchos casos particulares de interés: transformación cuadrado, raíz cuadrada, logaritmo, etc. Un $\alpha_j = 1$ significaría que el regresor aparece sin ninguna transformación. El problema está en seleccionar para cada regresor el $\alpha_j$ adecuado.

El modo de hacerlo propuesto por Box and Tidwell (1962) es el siguiente. Consideremos el modelo,

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1^{\alpha_1} + \ldots + \beta_p X_p^{\alpha_p} + \epsilon \quad (11.2) \\ &= \beta_0 + \beta_1 W_1 + \ldots + \beta_p W_p + \epsilon. \quad (11.3) \end{aligned}$$

Si realizamos una linealización aproximada mediante un desarrollo en serie de Taylor en torno al punto $(\alpha_1, \ldots, \alpha_k)' = (1, 1, \ldots, 1)'$, obtenemos:

$$Y \approx \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \gamma_1 Z_1 + \ldots + \gamma_p Z_p + \epsilon, \quad (11.4)$$

en donde

$$\begin{aligned} \gamma_j &= \beta_j(\alpha_j - 1) \quad (11.5) \\ Z_j &= X_j \ln(X_j). \quad (11.6) \end{aligned}$$

Tenemos pues un modelo en el que podemos estimar los parámetros, $(\beta_0, \ldots, \beta_p, \gamma_1, \ldots, \gamma_p)$. De ellos podemos recuperar valores estimados de $(\alpha_1, \ldots, \alpha_p)$ así:

$$\hat{\alpha}_j = \frac{\hat{\gamma}_j}{\hat{\beta}_j} + 1. \quad (11.7)$$

Podemos detenernos aquí, pero cabe pensar en un proceso iterativo de refinado de la solución obtenida. Llamemos $\hat{\alpha}_k^{(1)}$, $k = 1, \ldots, p$, a los estimadores de los parámetros de transformación $\alpha_k$ obtenidos como primera aproximación al estimar (11.4). Podríamos ahora definir

$$\begin{aligned} W_j^{(1)} &= X_j^{\alpha_j^{(1)}} \quad (11.8) \\ Z_j^{(1)} &= W_j^{(1)} \ln(W_j^{(1)}) \quad (11.9) \end{aligned}$$

y estimar

$$Y = \beta_0 + \beta_1 W_1^{(1)} + \ldots + \beta_p W_p^{(1)} + \gamma_1 Z_1^{(1)} + \ldots + \gamma_p Z_p^{(1)} + \epsilon, \quad (11.10)$$

Obtendríamos así estimaciones de $W_1^{(2)}, \ldots, W_p^{(2)}$, y podríamos proseguir de modo análogo hasta convergencia, si se produce.

## 11.3. Transformaciones de la variable respuesta

### 11.3.1. Generalidades

Además de transformar los regresores, o en lugar de hacerlo, podemos transformar la variable respuesta $Y$. Es importante tener en cuenta que si realizamos transformaciones no lineales de la $Y$ los modelos ya no serán directamente comparables en términos de, por ejemplo, $R^2$ o suma de cuadrados residual. Comparaciones de esta naturaleza requerirían reformular el modelo en las variables originales.

**Ejemplo 11.1** Supongamos que nos planteamos escoger entre los dos modelos alternativos,

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \epsilon \quad (11.11) \\ \log(Y) &= \gamma_0 + \gamma_1 X_1 + \nu. \quad (11.12) \end{aligned}$$

La transformación $\log$ deforma la escala de la $Y$; si el logaritmo es decimal, por ejemplo, valores de $Y$ entre 1 y 1000 quedan convertidos en valores entre 0 y 3 (si hubiera valores de $Y$ cercanos a cero, por el contrario, al tomar logaritmos se separarían hacia $-\infty$). Esta deformación puede ser bastante drástica, y afectar mucho a la suma de cuadrados de los residuos, independientemente del poder predictivo del único regresor $X_1$.

Para efectuar la comparación podemos convertir todo a unidades comunes. Así, no serían comparables las sumas de cuadrados

$$\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1})^2 \tag{11.13}$$

$$\sum (\log(Y_i) - \hat{\gamma}_0 - \hat{\gamma}_1 X_{i1})^2, \tag{11.14}$$

pero sí lo serían

$$\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1})^2 \tag{11.15}$$

$$\sum (Y_i - \exp\{\hat{\gamma}_0 + \hat{\gamma}_1 X_{i1}\})^2; \tag{11.16}$$

no obstante, véase la discusión en la Observación 11.1 que sigue.

**Observación 11.1** Las sumas de cuadrados de los residuos de dos modelos son comparables cuando ambos poseen el mismo número de parámetros estimados. Si no es el caso, y los modelos son lineales, podemos corregir el efecto del diferente número de parámetros penalizando la suma de cuadrados (por ejemplo, adoptando criterios como la $C_p$ de Mallows; véase la Sección 10.1.2). En el caso en que se hace alguna transformación, ¿hay que "contarla" como parámetro? En cierto modo, la transformación efectuada es una manipulación tendente a mejorar el ajuste a los datos, y *habría que tener esto en cuenta, especialmente si la transformación se escoge a la vista de los datos.*

No está claro, sin embargo, cómo "contar" una transformación. Una posibilidad que elude el problema es renunciar a penalizar la correspondiente suma de cuadrados y hacer validación cruzada (ver la Sección 10.1.4).

### 11.3.2. La transformación de Box-Cox.

En ocasiones puede resultar inadecuado suponer que la variable respuesta $Y$ está relacionada linealmente con las $X$, y, sin embargo, ser plausible un modelo como el siguiente:

$$g(Y_i) \quad = \quad \vec{x_i}\,'\vec{\beta} + \epsilon_i \tag{11.17}$$

Una familia de funciones $g(.)$ de particular interés y flexibilidad es la proporcionada por la llamada *transformación de Box-Cox,* sustancialmente idéntica a la adoptada para los regresores en la Sección 11.2.2. Definamos,

$$W_{(\lambda)} = g(Y;\lambda) = \begin{cases} (Y^\lambda - 1)/\lambda & \text{cuando } \lambda \neq 0, \\ \ln Y & \text{cuando } \lambda = 0. \end{cases}$$

y supongamos que $W_{(\lambda)}$ se genera de acuerdo con (11.17), es decir,

$$W_{(\lambda),i} \quad = \quad \vec{x_i}\,'\vec{\beta} + \epsilon_i \tag{11.18}$$

$$\vec{\epsilon} \quad \sim \quad N(\vec{0}, \sigma^2 I) \tag{11.19}$$

Podemos, dadas las observaciones $X, \vec{y}$, escribir la verosimilitud conjunta de todos los parámetros: $\beta$, $\sigma$, y $\lambda$. Dicha verosimilitud puede escribirse en función de $\vec{w}$ así[1]:

$$f_{\vec{Y}}(\vec{y}) \quad = \quad f_{\vec{W}}(\vec{w})\,|J(\lambda)| \tag{11.20}$$

siendo $J(\lambda)$ el jacobiano de la transformación:

$$J(\lambda) \quad = \quad \left|\frac{\partial \vec{w}}{\partial \vec{y}}\right| = \prod_{i=1}^{N} y_i^{\lambda-1} \tag{11.21}$$

Por tanto:

$$\begin{aligned}
\log \text{ver}(\vec{\beta}, \lambda, \sigma^2; \vec{Y}) \quad &= \quad \log\left(\frac{1}{\sqrt{2\pi}}\right)^N \left(\frac{1}{|\sigma^2 I|^{\frac{1}{2}}}\right) \\
&\quad \times \log\left[\exp\left\{-\frac{1}{2}\frac{(\vec{w}_{(\lambda)} - X\vec{\beta})'(\vec{w}_{(\lambda)} - X\vec{\beta})}{\sigma^2}\right\}|J(\lambda)|\right] \\
&= \quad -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 \\
&\quad -\frac{1}{2}\frac{(\vec{w}_{(\lambda)} - X\vec{\beta})'(\vec{w}_{(\lambda)} - X\vec{\beta})}{\sigma^2} + \log\prod_{i=1}^{N} y_i^{\lambda-1} \\
&= \quad -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 + (\lambda-1)\sum_{i=1}^{N}\log y_i \\
&\quad -\frac{1}{2}\frac{\vec{w}_{(\lambda)}{}'(I - X(X'X)^{-1}X')\vec{w}_{(\lambda)}}{\sigma^2}
\end{aligned} \tag{11.22}$$

La expresión (11.22) se ha obtenido maximizando la precedente respecto de $\vec{\beta}$. El máximo, en efecto, se alcanza para aquél valor de $\vec{\beta}$ que minimiza $(\vec{w}_{(\lambda)} - X\vec{\beta})'(\vec{w}_{(\lambda)} - X\vec{\beta})$, y éste es precisamente el $\hat{\beta}$ mínimo cuadrático. La suma de cuadrados de los residuos es entonces (véase (2.35), pág. 20) $\vec{w}'_{(\lambda)}(I - X(X'X)^{-1}X')\vec{w}_{(\lambda)}$.

Si ahora maximizamos (11.22) respecto a $\sigma^2$, vemos que el máximo se alcanza para,

$$\hat{\sigma}^2_{(\lambda)} \quad = \quad \frac{\vec{w}'_{(\lambda)}(I - X(X'X)^{-1}X')\vec{w}_{(\lambda)}}{N}$$

y el logaritmo de la verosimilitud concentrada es:

$$\log \text{ver}(\lambda; \vec{Y}) \quad = \quad -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\hat{\sigma}^2_{(\lambda)} - \frac{N}{2} + (\lambda-1)\sum_{i=1}^{N}\log y_i \tag{11.23}$$

Podemos escoger como transformación aquélla cuyo $\lambda$ maximice (11.23), o, de modo equivalente, tras prescindir de las constantes,

$$\log \text{ver}(\lambda; \vec{Y}) \quad = \quad -\frac{N}{2}\log\hat{\sigma}^2_{(\lambda)} + (\lambda-1)\sum_{i=1}^{N}\log y_i. \tag{11.24}$$

[1] La variable transformada $\vec{w}$ depende en todo caso del $\lambda$ empleado en la transformación; omitimos dicha dependencia para aligerar la notación, salvo donde interese enfatizarla.

Un modo sencillo de hacerlo consiste en tomar un número adecuado de valores de $\lambda$ equiespaciados en un intervalo susceptible de contener el $\lambda$ óptimo, ajustar una regresión para cada $\lambda$, y calcular el correspondiente valor de (11.24). Frecuentemente se suele tomar el intervalo $-2 \leq \lambda \leq 2$ (que incluye como casos particulares la transformación raíz cuadrada ($\lambda = \frac{1}{2}$), cuadrado ($\lambda = 2$), logaritmo ($\lambda = 0$), raíz cuadrada negativa, etc.), y dentro de él unas cuantas decenas de valores de $\lambda$.

Es frecuente que $\log \text{ver}(\lambda; \vec{Y})$ como función de $\lambda$ sea una función relativamente plana. Ello suscita el problema de decidir si el valor de $\lambda$ que la maximiza es significativamente distinto de 1 (lo que supondría que no es preciso hacer ninguna transformación). Podemos recurrir a un contraste razón de verosimilitudes (véase B.3). Bajo la hipótesis $H_0 : \lambda = \lambda_0$, si $\hat{\lambda}$ denota el estimador máximo verosímil de $\lambda$ y $L(\lambda)$ el valor que toma la verosimilitud, para muestras grandes se tiene que

$$2 \ln \left( \frac{L(\hat{\lambda})}{L(\lambda_0)} \right) \sim \chi^2_1; \tag{11.25}$$

por tanto, a la vista de (11.23), rechazaremos $H_0$ al nivel de significación $\alpha$ si

$$-2 \left( \frac{N}{2} \log \hat{\sigma}^2_{(\hat{\lambda})} + (\hat{\lambda} - \lambda_0) \sum_{i=1}^{N} \log y_i - \frac{N}{2} \log \hat{\sigma}^2_{(\lambda_0)} \right) > \chi^2_{1;\alpha}. \tag{11.26}$$

Utilizando la misma idea podemos construir intervalos de confianza para $\lambda$.

# Capítulo 12

# Regresión con respuesta cualitativa

## 12.1. El modelo *logit*.

Con frecuencia se presentan situaciones en que la variable respuesta a explicar toma sólo uno de dos estados, a los que convencionalmente asignamos valor 0 ó 1. Por ejemplo, variables de renta, habitat, educación y similares pueden influenciar la decisión de compra de un cierto artículo. Podríamos así plantearnos el estimar,

$$\vec{Y} \quad = X\vec{\beta} + \vec{\epsilon} \tag{12.1}$$

en que $Y$ es una variable tomando dos valores: 1 (= "Compra") ó 0 (= "No compra").

Nada parecería, en principio, impedir el empleo del modelo lineal estudiado en una situación como ésta. Pero hay varias circunstancias que debemos considerar.

1. No tiene ya sentido suponer una distribución normal en las perturbaciones. En efecto, para cualesquiera valores que tomen los regresores, de

   $$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

   se deduce que $\epsilon$ sólo puede tomar uno de dos valores: la diferencia que separa a la $Y_i$ (0 ó 1) de la combinación lineal de regresores que constituye su "parte explicada".

2. Tratándose de una respuesta que puede tomar valor 0 ó 1, interpretaríamos $\hat{Y}_i$ como su valor medio dados los valores de los regresores. Al poder tomar $Y_i$ sólo los valores 0 y 1, su valor medio es $P_i$, la probabilidad del valor 1. Por tanto, valores de $\hat{Y}_i$ entre 0 y 1 son interpretables. Pero nada impide que el modelo proporciones predicciones mayores que 1 (o menores que 0), circunstancia molesta.

3. Tampoco podemos ya suponer que hay homoscedasticidad. En efecto, si tomamos valor medio en la expresión anterior tenemos:

$$E[Y_i] = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1} = P_i$$

En consecuencia, $Y_i$ toma valor 1 con probabilidad $P_i$ y valor 0 con probabilidad $Q_i = 1 - P_i$ y,

$$\epsilon_i = \begin{cases} 1 - P_i & \text{con probabilidad } P_i \\ -P_i & \text{con probabilidad } Q_i = 1 - P_i. \end{cases}$$

Entonces,

$$E[\epsilon_i^2] = (1 - P_i)^2 P_i + (-P_i)^2 (1 - P_i) = Q_i^2 P_i + Q_i P_i^2 = P_i Q_i. \quad (12.2)$$

La varianza de $Y$ varía por tanto de observación a observación de acuerdo con los valores que toman los regresores. Adicionalmente, (12.2) muestra que la distribución de $\epsilon_i$ sería binaria de parámetro $P_i$.

El tercer inconveniente podría resolverse haciendo uso de regresión ponderada, para corregir el efecto de la heterocedasticidad. No obstante, suele emplearse una aproximación alternativa que da cuenta también de los dos primeros. El modelo lineal ordinario hace depender linealmente de las variables $X$ la *media* de la variable respuesta, $E(Y_i)$. Podemos en lugar de ello hacer depender de los regresores *una función* de la media $E(Y_i)$; por ejemplo, la conocida como *logit*,

$$\ell(E(Y_i)) \quad \stackrel{\text{def}}{=} \quad \ln\left(\frac{P_i}{1 - P_i}\right). \quad (12.3)$$

Nótese que como $E(Y_i) = P_i$, (12.3) es efectivamente una función de la media. Obsérvese también que $\ell(E(Y_i))$ toma valores de modo continuo entre $-\infty$ y $+\infty$. Podemos pensar en hacer que $\ell(E(Y_i))$, y no $E(Y_i)$, dependa linealmente de los regresores:

$$\ell(E(Y_i)) \quad = \quad \ln\left(\frac{P_i}{1 - P_i}\right) = \vec{x}_i{}' \vec{\beta}, \quad (12.4)$$

y a continuación especificar la distribución de $Y_i$ en torno a su media $E(Y_i)$. Ya hemos visto que una distribución binaria es una elección natural si $Y_i$ es una variable 0/1.

**Observación 12.1** Transformar *la media* $E(Y_i)$ es un enfoque alternativo al de transformar $Y_i$, y en muchos aspectos un refinamiento. Una transformación de la respuesta como, por ejemplo, las de la familia de Box-Cox, tiene que cumplir varios objetivos, generalmente contradictorios. Por un lado, deseamos que la variable respuesta se acerque a la normalidad. Por otro, que la varianza sea homogénea, y la dependencia de los regresores lineal.

El enfoque de hacer depender linealmente de los regresores una función de la media de la variable respuesta es mucho más flexible. Podemos escoger la función de la media que sea más aproximadamente función lineal de los regresores, y especificar separadamente la distribución de la variable respuesta en torno a su media. El enfoque goza así de una enorme flexibilidad.

Despejando $P_i$ de la expresión anterior,

$$P_i \quad = \quad \frac{\exp(\vec{x}_i{}' \vec{\beta})}{1 + \exp(\vec{x}_i{}' \vec{\beta})}. \quad (12.5)$$

### 12.1.1. Interpretación de los coeficientes

Los parámetros de un modelo *logit* tienen interpretación inmediata: $\beta_i$ es el efecto de un cambio unitario en $X_i$ sobre el *logit* o logaritmo de la razón de posibilidades *(log odds)*. Pero pueden en ocasiones ser interpretados de manera más directamente relacionada con magnitudes de interés. Consideremos primero el caso más simple, en que tenemos un único regresor dicotómico, $X$, codificado con valores 0/1. El resultado de clasificar una muestra de $N$ sujetos con arreglo a los valores observados de $Y$ (respuesta) y $X$ (regresor) puede imaginarse en una tabla de doble entrada como la siguiente:

| | $\mathbf{X=1}$ | $\mathbf{X=0}$ |
|---|---|---|
| $\mathbf{Y=1}$ | $n_{11}$ | $n_{12}$ |
| $\mathbf{Y=0}$ | $n_{21}$ | $n_{22}$ |

Si el modelo *logit* es de aplicación, las probabilidades de cada celda en la tabla anterior vendrían dadas por las expresiones que aparecen en la tabla siguiente:

| | $\mathbf{X=1}$ | $\mathbf{X=0}$ |
|---|---|---|
| $\mathbf{Y=1}$ | $\pi(1)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$ | $\pi(0)=\frac{e^{\beta_0}}{1+e^{\beta_0}}$ |
| $\mathbf{Y=0}$ | $1-\pi(1)=\frac{1}{1+e^{\beta_0+\beta_1}}$ | $1-\pi(0)=\frac{1}{1+e^{\beta_0}}$ |

Definamos la *razón de posibilidades relativa (relative odds ratio)* así:

$$\psi = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}. \tag{12.6}$$

Entonces,

$$\begin{aligned}
\ln(\psi) &= \ln\left(\frac{\pi(1)\,/\,(1-\pi(1))}{\pi(0)/(1-\pi(0))}\right) \\
&= \ln\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\Big/\frac{1}{1+e^{\beta_0+\beta_1}}\right) - \ln\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\Big/\frac{1}{1+e^{\beta_0}}\right) \\
&= \ln\left(\frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}\right) \\
&= \beta_1. \qquad (12.7)
\end{aligned}$$

Por tanto, $\hat{\beta}_1$ estimará $\ln(\psi)$, y $\exp(\hat{\beta}_1)$ estimará $\psi$.

**Observación 12.2** La codificación de $X$, al igual que la de $Y$, es arbitraria. La interpretación correcta de $\beta_1$ es "incremento de $\ln(\psi)$ cuando $X$ se incrementa en una unidad". Por tanto, como se ha indicado, si la presencia de una característica se codifica mediante $X=1$ y su ausencia mediante $X=0$, $\ln(\hat{\psi})=\hat{\beta}_1$ y $\hat{\psi}=\exp(\hat{\beta}_1)$. Pero si la presencia de la misma característica se codifica mediante $X=a$ y su ausencia mediante $X=b$, cálculos similares a los realizados muestran que $\ln(\psi)=\beta_1(a-b)$. A la hora de interpretar los coeficientes de un modelo logit es necesario por tanto tener en cuenta la codificación utilizada.

Interpretamos $\psi$ como indicando *aproximadamente* cuánto más probable es que $Y$ tome el valor 1 cuando $X = 1$ que cuando $X = 0$. *Aproximadamente*, porque

$$\frac{\pi(1)}{\pi(0)} \approx \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}$$

si y sólo si

$$\frac{1-\pi(0)}{1-\pi(1)} \approx 1.$$

Ello acontece, por ejemplo, cuando $Y = 1$ se presenta muy raramente en la población —como cuando estudiamos la incidencia de una enfermedad muy rara, tanto para sujetos tratados ($X = 1$) como no tratados ($X = 0$)—. En este último caso, $\exp(\hat{\beta}_1)$ se interpretaría como una estimación de la relación de riesgos. Un $\hat{\beta}_1 > 0$ significará, por tanto, que $X = 1$ incrementa el riesgo de que $Y = 1$, y viceversa.

### 12.1.2. La importancia del diseño muestral

¿Sólo podemos estimar, y aún aproximadamente, la razón de riesgos $\pi(1)/\pi(0)$? ¿Qué impediría estimar el riesgo $P_i$ correspondiente a unos determinados valores de los regresores, $\vec{x}_i$, haciendo uso de el análogo muestral de (12.5)? Es importante observar (véase Kleinbaum (1994) para una discusión completa de esto) que en ocasiones ello no será posible.

Se hace preciso distinguir dos situaciones que pueden dar lugar a los mismos datos pero reflejan modos de obtenerlos radicalmente diferentes. En el primer caso tenemos un *diseño de exposición*, típico en trabajos epidemiológicos, en que una muestra *fijada de antemano sin conocer el valor de la variable respuesta $Y$ y representativa del total de la población en riesgo* se sigue a lo largo de un periodo de tiempo al cabo del cual se conoce el valor de $Y$. En este caso, podríamos estimar el riesgo $P_i$ como se ha dicho.

Completamente diferente es el diseño muestral de *casos-controles*. En este caso seleccionamos la muestra *a la vista de los valores de $Y_i$*. Típicamente, si examinamos un evento que se presenta raramente, como una enfermedad poco frecuente, tomaremos todos los individuos enfermos de que dispongamos *(casos)*, completando la muestra con un número arbitrario de sanos *(controles)*. Los coeficientes $\beta_1, \ldots, \beta_p$ son interpretables, pero $\beta_0$ no lo es. Ninguna fórmula que lo requiera —como (12.5)— puede utilizarse.

La razón es fácil de entender: $\hat{\beta}_0$ depende de la abundancia relativa de casos y controles, y ésta es como hemos dicho arbitraria. La situación se asemeja a la que se presenta cuando construimos una tabla de contingencia $2 \times 2$ como:

| | $\mathbf{X = 1}$ | $\mathbf{X = 0}$ | **Total** |
|---|---|---|---|
| $\mathbf{Y = 1}$ | $n_{11}$ | $n_{12}$ | $n_{1.}$ |
| $\mathbf{Y = 0}$ | $n_{21}$ | $n_{22}$ | $n_{2.}$ |
| **Total** | $n_{,1}$ | $n_{,2}$ | $n_{..}$ |

Si hemos escogido los sujetos completamente al azar, es razonable tomar el cociente $n_{1.}/n_{..}$ como estimador de la proporción de casos con $Y = 1$ en la población (y cocientes como $n_{11}/n_{,1}$ o $n_{12}/n_{,2}$ estimarían las proporciones en las subpoblaciones caracterizadas por $X = 1$ y $X = 0$ respectivamente).

Si, por el contrario, hemos fijado los valores $n_{1.}$ y $n_{2.}$, es claro que dicho cociente no estima nada, sino que es resultado de una decisión arbitraria.

### 12.1.3. Estimación

Consideremos una muestra de tamaño $N$, formada por observaciones $(y_i, \vec{x}_i)$. Para cada observación, $y_i$ es 0 ó 1. El modelo *logit*, sin embargo, le atribuye una probabilidad $P_i$ (si se trata de un "1") ó $1 - P_i$ (si se trata de un "0"). Por consiguiente, la verosimilitud de la muestra es

$$
\begin{aligned}
L(\hat{\beta}, \vec{y}, X) &= \prod_{i=1}^{N} (P_i)^{y_i} (1 - P_i)^{1-y_i} && (12.8)\\
&= \prod_{i=1}^{N} \left( \frac{1}{1 + \exp(\vec{x}_i{}' \vec{\beta})} \right)^{1-y_i} \left( \frac{\exp(\vec{x}_i{}' \vec{\beta})}{1 + \exp(\vec{x}_i{}' \vec{\beta})} \right)^{y_i} && (12.9)\\
&= \prod_{i=1}^{N} \left( \frac{1}{1 + \tau_i} \right)^{1-y_i} \left( \frac{\tau_i}{1 + \tau_i} \right)^{y_i}, && (12.10)
\end{aligned}
$$

con $\tau_i = \exp(\vec{x}_i{}' \vec{\beta})$. Tomando logaritmos en (12.10), obtenemos

$$
\sum_{i=1}^{N} \ln \left( \frac{1}{1 + \tau_i} \right) + \sum_{i=1}^{N} y_i \ln(\tau_i). \tag{12.11}
$$

Si derivamos (12.11) respecto de $\vec{\beta}$ e igualamos el vector de derivadas a cero, obtenemos un sistema no lineal; no obstante, puede resolverse numéricamente para obtener el vector de estimadores $\hat{\beta}$. Alternativamente, podría procederse a la maximización directa de (12.9) mediante un algoritmo conveniente.

> **Observación 12.3** La verosimilitud en (12.9) es la ordinaria o incondicional. En determinadas circunstancias —notablemente en estudios con casos y controles emparejados respecto de variables de estratificación cuyos coeficientes carecen de interés— podríamos desear realizar estimación máximo verosímil condicional. Sobre el fundamento de esto puede verse Cox and Hinkley (1978), pág. 298 y siguientes, Kleinbaum (1994) o Hosmer and Lemeshow (1989), Cap. 7. En R puede estimarse un modelo logit mediante máxima verosimilitud condicional utilizando la función `clogit` (en el paquete `survival`).

### 12.1.4. Contrastes y selección de modelos

Necesitamos criterios para decidir sobre la inclusión o no de parámetros, y para comparar modelos. La teoría para ello deriva del contraste razón generalizada de verosimilitudes (ver B.3).

Consideremos un modelo saturado, proporcionando el mejor ajuste posible. Llamaremos a éste modelo *modelo base* o *modelo de referencia*: se tratará en general de un modelo claramente sobreparametrizado, pero que proporciona un término de comparación útil. Requerirá, en principio, un parámetro por cada combinación de valores de los regresores, y proporcionará valores ajustados $\hat{P} = (\hat{P}_1, \ldots, \hat{P}_k)$.

De acuerdo con la teoría en la Sección B.3, bajo la hipótesis nula de que el modelo correcto es (12.4)

$$
-2 \ln \left( \frac{L(\hat{\beta})}{L(\hat{P})} \right) \sim \chi_{k-p}, \tag{12.12}
$$

en que $p$ es el número de parámetros estimados en $\hat{\beta}$. Al cociente (12.12) se le denomina *desviación* respecto del modelo de referencia parametrizado por $\hat{P}$.

El adoptar un modelo menos parametrizado que el de referencia, implica una disminución de la verosimilitud y una desviación (12.12) positiva cuya distribución, bajo la hipótesis nula, sigue la distribución $\chi^2_{k-p}$ indicada. Si la desviación fuera excesiva (es decir, si sobrepasa $\chi^2_{k-p;\alpha}$ para el nivel de significación $\alpha$ que hayamos escogido), rechazaríamos la hipótesis nula.

Análogo criterio podemos seguir para hacer contrastes sobre un único parámetro o sobre grupos de parámetros. Por ejemplo, para contrastar si el parámetro $\beta_j$ es significativamente diferente de cero en un cierto modelo parametrizado por $\vec{\beta}$, calcularíamos

$$-2\ln\left(\frac{L(\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_{j-1},\hat{\beta}_{j+1},\ldots,\hat{\beta}_k)}{L(\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_{j-1},\hat{\beta}_j,\hat{\beta}_{j+1},\ldots,\hat{\beta}_k)}\right), \tag{12.13}$$

que debe ser comparado con una $\chi^2_1$; valores grandes de (12.13) son evidencia contra la hipótesis $h:\beta_j=0$.

Para contrastar la hipótesis de nulidad de todos los parámetros, salvo quizá $\beta_0$ afectando a la columna de "unos", compararíamos

$$-2\ln\left(\frac{L(\hat{\beta}_0)}{L(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_k)}\right) \tag{12.14}$$

a una $\chi^2_{k-1}$; la expresión (12.14) es similar a la suma de cuadrados SSR en una regresión ordinaria. El análogo a SST sería

$$-2\ln\left(\frac{L(\hat{\beta}_0)}{L(\hat{P})}\right). \tag{12.15}$$

Esta analogía puede extenderse para obtener un estadístico similar a la $C_p$ de Mallows así:

$$\Delta_k \quad = \quad -2\ln\left(\frac{L(\hat{\beta}_0)}{L(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_k)}\right) - 2(k-1), \tag{12.16}$$

y una "$R^2$" así:

$$R^2 \quad = \quad \frac{-2\ln\left(\frac{L(\hat{\beta}_0)}{L(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_k)}\right)}{-2\ln\left(\frac{L(\hat{\beta}_0)}{L(\hat{P})}\right)} \tag{12.17}$$

Obsérvese que en (12.16) el primer sumando de la derecha sigue asintóticamente una distribución $\chi^2_{k-1}$ con grados de libertad bajo el supuesto de que el modelo más parametrizado no añade realmente nada. Los grados de libertad —y por tanto el valor esperado de dicho sumando— crecen con el número de parámetros ajustados. El segundo término que se sustrae a continuación es, precisamente, el valor medio de una $\chi^2_{k-1}$. Mientras que el primero crece monótonamente al introducir nuevos parámetros, el segundo penaliza este crecimiento.

**Observación 12.4** Escogeríamos de acuerdo con este criterio el modelo maximizando $\Delta_k$ o, alternativamente, minimizando

$$\text{AIC}_k \quad = \quad -2\ln L(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_k) + 2k. \tag{12.18}$$

La expresión anterior se conoce como criterio AIC (="An Information Criterion" o "Akaike Information Criterion", por su proponente). Puede ser obtenido de diversos modos, incluido un argumento haciendo uso de Teoría de la Información: véase Akaike (1972).

**R: Ejemplo 12.1** *(estimación de modelos logit mediante* `glm`*)*

```
--- Obtenido mediante R BATCH demo11.R
>  invisible(options(echo = TRUE))
>  options(digits=5)
>  options(columns=40)
>  set.seed(123457)
> #
> #   Creamos datos sintéticos con parámetros conocidos.
> #
>  X <- matrix(rnorm(1000),ncol=20)
>  betas <- rep(0,20)
>  betas[c(3,5,7,12)] <- 1:4
>  y <- X %*% betas + rnorm(50)
>  datos <- as.data.frame(cbind(X,y))
>  dimnames(datos)[[2]][21] <- "y"
>  completo <- lm(y ~ .,datos)
>  summary(completo)

Call:
lm(formula = y ~ ., data = datos)

Residuals:
   Min     1Q Median     3Q    Max
-1.916 -0.550 -0.107  0.829  2.204

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.0706     0.2227   -0.32     0.75
V1            0.0408     0.2422    0.17     0.87
V2            0.1720     0.2603    0.66     0.51
V3            1.1884     0.2397    4.96  2.9e-05 ***
V4           -0.0238     0.2067   -0.11     0.91
V5            2.0035     0.2022    9.91  8.1e-11 ***
V6            0.2633     0.2217    1.19     0.24
V7            2.9970     0.1875   15.98  6.5e-16 ***
V8           -0.1074     0.2804   -0.38     0.70
V9            0.0514     0.2105    0.24     0.81
V10          -0.2367     0.2148   -1.10     0.28
V11          -0.2053     0.2042   -1.01     0.32
V12           4.0374     0.2212   18.25  < 2e-16 ***
V13           0.1137     0.2161    0.53     0.60
V14          -0.2115     0.2163   -0.98     0.34
V15           0.0191     0.3076    0.06     0.95
V16           0.1206     0.2328    0.52     0.61
V17           0.0318     0.1972    0.16     0.87
V18          -0.0786     0.2108   -0.37     0.71
V19           0.0879     0.2569    0.34     0.73
V20           0.0162     0.1949    0.08     0.93
---
```

```
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 1.2 on 29 degrees of freedom
Multiple R-Squared: 0.977,Adjusted R-squared: 0.961
F-statistic:   61 on 20 and 29 DF,  p-value: <2e-16

>  #
>  #  Utilicemos fuerza bruta (con 15 regresores, no hay problema. Con más
>  #  puede tardar bastante en una máquina lenta). Necesitamos la función
>  #  "leaps" y dar regresores y respuesta como una matriz y un vector
>  #
>  library(leaps)
>  mods <- leaps(x=X,y=y,method="Cp")  # mods contiene información sobre
>                                      # todos los modelos estimados.
>  postscript(file="demo10.eps",
+             horizontal=FALSE,
+             width=5,height=9)
>  par(mfrow=c(2,1))
>  plot(mods$size,mods$Cp,
+       main="Cp versus talla modelos",
+       xlab=expression(p),
+       ylab=expression(C[p]))
>
>  mods.r <- leaps(x=X,y=y,method="adjr2") # R2 como criterio,
>                                          # selecciona modelos "mayores".
>  plot(mods.r$size,mods.r$adjr2,
+       main="R2 versus talla modelos",
+       xlab=expression(p),
+       ylab=expression(bar(R)^2))
>
>  mejores <- order(mods$Cp)[1:15]         # Los 15 mejores según Cp.
>  regres  <- mods$which[mejores,]
>  dimnames(regres)[[2]] <-                # Para fácil legibilidad.
+             dimnames(datos)[[2]][1:20]
>  Cp  <- mods$Cp[mejores]                 # Las Cp’s correspondientes.
>  cbind(regres,Cp)                        # Los mejores modelos
  V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20
5  0  0  1  0  1  1  1  0  0   0   0   1   0   0   0   0   0   0   0   0
6  0  0  1  0  1  1  1  0  0   0   0   1   0   1   0   0   0   0   0   0
6  0  0  1  0  1  1  1  0  0   1   0   1   0   0   0   0   0   0   0   0
4  0  0  1  0  1  0  1  0  0   0   0   1   0   0   0   0   0   0   0   0
6  0  0  1  0  1  1  1  0  0   0   1   1   0   0   0   0   0   0   0   0
5  0  0  1  0  1  0  1  0  0   1   0   1   0   0   0   0   0   0   0   0
6  0  0  1  0  1  1  1  0  0   0   0   1   0   0   0   0   0   0   1   0
5  0  0  1  0  1  0  1  0  0   0   1   1   0   0   0   0   0   0   0   0
7  0  0  1  0  1  1  1  0  0   1   0   1   0   1   0   0   0   0   0   0
6  0  0  1  0  1  1  1  0  0   0   0   1   0   0   1   0   0   0   0   0
6  1  0  1  0  1  1  1  0  0   0   0   1   0   0   0   0   0   0   0   0
5  1  0  1  0  1  0  1  0  0   0   0   1   0   0   0   0   0   0   0   0
6  0  0  1  0  1  1  1  0  0   0   0   1   0   0   0   0   1   0   0   0
7  0  0  1  0  1  1  1  0  0   0   1   1   0   1   0   0   0   0   0   0
6  0  0  1  0  1  1  1  0  0   0   0   1   1   0   0   0   0   0   0   0
       Cp
5 -4.2251
```

```
6 -3.4911
6 -3.4553
4 -3.4531
6 -3.2133
5 -3.1498
6 -2.6538
5 -2.5504
7 -2.5475
6 -2.5181
6 -2.4759
5 -2.4051
6 -2.3677
7 -2.3654
6 -2.3347
>  #
>  #  Estimemos el mejor de acuerdo con el criterio Cp.
>  #
>  mod1 <- lm(y ~ V3 + V4 + V5 + V7 + V10 + V12 + V16 + V17,data=datos)
>  #
>  #
>  #  Vemos que el "mejor" modelo de acuerdo con Cp reproduce bastante
>  #  bien el mecanismo que genera los datos; ha incluido tres variables
>  #  extra innecesarias.
>  #
>  #  Podemos probar modelos competidores, añadiendo o quitando variables
>  #  sin reestimar todo.
>  #
>  mod2 <- update(mod1, . ~ . + V1 + V2)    # añadimos dos variables
>  summary(mod2)

Call:
lm(formula = y ~ V3 + V4 + V5 + V7 + V10 + V12 + V16 + V17 +
    V1 + V2, data = datos)

Residuals:
   Min     1Q Median     3Q    Max
-1.611 -0.762  0.122  0.627  2.237

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03573    0.18316   -0.20     0.85
V3           1.08674    0.19721    5.51  2.5e-06 ***
V4          -0.00741    0.16766   -0.04     0.96
V5           2.03931    0.16976   12.01  1.1e-14 ***
V7           3.05622    0.14772   20.69  < 2e-16 ***
V10         -0.27977    0.19088   -1.47     0.15
V12          4.10685    0.18483   22.22  < 2e-16 ***
V16          0.08436    0.15101    0.56     0.58
V17          0.05185    0.14567    0.36     0.72
V1           0.16370    0.18257    0.90     0.38
V2          -0.00659    0.20666   -0.03     0.97
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
Residual standard error: 1.11 on 39 degrees of freedom
Multiple R-Squared: 0.973,Adjusted R-squared: 0.966
F-statistic:  141 on 10 and 39 DF,  p-value: <2e-16

>  mod3 <- update(mod1, . ~ .-V10-V16-V17)  # eliminamos tres variables
>  summary(mod3)

Call:
lm(formula = y ~ V3 + V4 + V5 + V7 + V12, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0289 -0.6955  0.0539  0.7177  2.5956

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0738     0.1596    0.46     0.65
V3            1.0693     0.1819    5.88  5.1e-07 ***
V4           -0.0410     0.1567   -0.26     0.79
V5            1.9898     0.1603   12.41  5.7e-16 ***
V7            3.0484     0.1400   21.77  < 2e-16 ***
V12           4.1357     0.1642   25.19  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 1.09 on 44 degrees of freedom
Multiple R-Squared: 0.971,Adjusted R-squared: 0.967
F-statistic:  293 on 5 and 44 DF,  p-value: <2e-16

>  #
>  #
>  m <- regsubsets(y ~ .,datos,       # Como alternativa tenemos regsubsets;
+                 method="forward")  # hace también regresión escalonada.
>  summary(m)
Subset selection object
Call: regsubsets.formula(y ~ ., datos, method = "forward")
20 Variables  (and intercept)
    Forced in Forced out
V1      FALSE      FALSE
V2      FALSE      FALSE
V3      FALSE      FALSE
V4      FALSE      FALSE
V5      FALSE      FALSE
V6      FALSE      FALSE
V7      FALSE      FALSE
V8      FALSE      FALSE
V9      FALSE      FALSE
V10     FALSE      FALSE
V11     FALSE      FALSE
V12     FALSE      FALSE
V13     FALSE      FALSE
V14     FALSE      FALSE
V15     FALSE      FALSE
V16     FALSE      FALSE
```

```
V17     FALSE      FALSE
V18     FALSE      FALSE
V19     FALSE      FALSE
V20     FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: forward
         V1  V2  V3  V4  V5  V6  V7  V8  V9  V10 V11 V12 V13 V14 V15 V16 V17
1  ( 1 ) " " " " " " " " " " " " " " " " " " " " " " "*" " " " " " " " " " "
2  ( 1 ) " " " " " " " " " " " " "*" " " " " " " " " "*" " " " " " " " " " "
3  ( 1 ) " " " " " " " " "*" " " "*" " " " " " " " " "*" " " " " " " " " " "
4  ( 1 ) " " " " "*" " " "*" " " "*" " " " " " " " " "*" " " " " " " " " " "
5  ( 1 ) " " " " "*" " " "*" "*" "*" " " " " " " " " "*" " " " " " " " " " "
6  ( 1 ) " " " " "*" " " "*" "*" "*" " " " " " " " " "*" " " "*" " " " " " "
7  ( 1 ) " " " " "*" " " "*" "*" "*" " " " " " "*" " " "*" " " "*" " " " " "
8  ( 1 ) " " " " "*" " " "*" "*" "*" " " " " " "*" " " "*" " " "*" " " " " "
         V18 V19 V20
1  ( 1 ) " " " " " "
2  ( 1 ) " " " " " "
3  ( 1 ) " " " " " "
4  ( 1 ) " " " " " "
5  ( 1 ) " " " " " "
6  ( 1 ) " " " " " "
7  ( 1 ) " " " " " "
8  ( 1 ) " " "*" " "
>  #
>  #  En la librería MASS tenemos también la función stepAIC, que emplea
>  #  el criterio AIC, aproximadamente equivalente a Cp
>  #
>  library(MASS)

Attaching package: 'MASS'


The following object(s) are masked _by_ .GlobalEnv :

 UScrime

>  step <- stepAIC(completo,scope= y ~ . ,
+                       direction="both",
+                       trace=FALSE)
>  summary(step)

Call:
lm(formula = y ~ V3 + V5 + V6 + V7 + V12, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9495 -0.6503 -0.0349  0.5244  2.6196

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0514     0.1518    0.34    0.736
V3            1.0256     0.1761    5.82  6.1e-07 ***
V5            2.0499     0.1557   13.17  < 2e-16 ***
```

```
V6            0.3046     0.1603    1.90    0.064 .
V7            3.0499     0.1346   22.65  < 2e-16 ***
V12           4.1077     0.1585   25.91  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.05 on 44 degrees of freedom
Multiple R-Squared: 0.973,Adjusted R-squared: 0.97
F-statistic:  317 on 5 and 44 DF,  p-value: <2e-16

>
```

## CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

**12.1** Muéstrese que la *desviación* definida a continuación de (12.12) coincide con SSE cuando consideramos un modelo lineal ordinario con normalidad en las perturbaciones.

**12.2** Compruébese derivando (12.11) que los estimadores máximo verosímiles de los parámetros $\vec{\beta}$ son soluciones del sistema de ecuaciones:

$$\sum_{i=1}^{N} \vec{x}_i \left( y_i - \frac{\tau_i}{1+\tau_i} \right) = \vec{0},$$

en que $\tau_i = \vec{x}_i{}' \vec{\beta}$.

# Parte II

# Análisis de Varianza

# Capítulo 13

# Análisis de varianza con efectos fijos.

## 13.1. Introducción.

Se presenta con frecuencia el problema de comparar el efecto que diversas circunstancias de tipo cualitativo ejercen sobre un cierto fenómeno. El problema inicial, que ha estimulado sobremanera en sus orígenes la investigación sobre Análisis de Varianza (ANOVA en lo sucesivo) se describe, en su forma más simple, así: tenemos varios tipos *(niveles)* de simientes, y estamos interesados en comparar su rendimiento. La variable endógena o respuesta es la cosecha obtenida por unidad de superficie. Para efectuar la comparación citada contamos con la posibilidad de hacer diversos ensayos, sembrando parcelas (que suponemos homogéneas) con semillas iguales o diferentes. Todos los ensayos efectuados con un mismo nivel del tratamiento (parcelas sembradas con el mismo tipo de semilla en nuestro ejemplo) reciben el nombre de *replicaciones* de dicho nivel.

En multitud de circunstancias, podemos pensar en un modelo del tipo:

$$y_{ij} = \mu_i + \epsilon_{ij} \tag{13.1}$$

$(i = 1, \ldots, I;\ j = 1, \ldots, J)$ que muestra la cosecha obtenida en la $j$-ésima parcela de las sembradas con la $i$-ésima semilla (o $j$-ésima replicacion del nivel $i$) como dependiendo de un parámetro $\mu_i$ que describe el rendimiento de la semilla $i$-ésima, mas una perturbación $\epsilon_{ij}$ recogiendo factores aleatorios (ambientales, etc.). Si se verifican los supuestos habituales (incorrelación y homoscedasticidad de las $\epsilon_{ij}$, etc.), (13.1) puede verse como un caso particular del modelo de regresión lineal, cuyo rasgo distintivo consiste en que los regresores toman valores 0 ó 1.

Por ejemplo, si hubiera tres tipos de semillas y sembráramos 2 parcelas con cada tipo, tendríamos que el problema de estimar los rendimientos $\mu_i$ (y contrastar hipótesis

sobre los mismos) requeriría estimar el siguiente modelo:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}. \tag{13.2}$$

*Toda* la teoría desarrollada para el caso general puede trasladarse sin más al caso particular del modelo ANOVA [1]. La razón por la que en ocasiones es objeto de estudio separado radica en que los cálculos son más simples en muchos casos de interés (frecuentemente se pueden realizar a mano). Por otra parte, se adoptan a menudo parametrizaciones que, aunque exactamente equivalentes a (13.1), dan lugar a una interpretación más intuitiva, al precio de hacer las matrices de diseño de rango deficiente; ello no tiene otra repercusión que la de obligar al empleo de inversas generalizadas en lugar de inversas ordinarias.

## 13.2. Análisis de varianza equilibrado con un tratamiento.

Decimos que un diseño es *equilibrado* cuando se realiza el mismo número de replicaciones con cada nivel de tratamiento. El ejemplo (13.2) de la Sección anterior es equilibrado. Como veremos pronto, esto simplifica los cálculos en extremo.

Frecuentemente se reescribe (13.2) así:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \alpha_1^A \\ \alpha_2^A \\ \alpha_3^A \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}, \tag{13.3}$$

y, en general, (13.1) de la Sección anterior adopta la forma

$$y_{ij} \quad = \quad \alpha + \alpha_i^A + \epsilon_{ij} \tag{13.4}$$

con $(i = 1, \ldots, I;\ j = 1, \ldots, J)$. Se dice que el parámetro $\alpha$ recoge la media global, y $\alpha_i^A$ los efectos diferenciales de cada nivel del tratamiento A (en el caso presente, en que hay un único tratamiento, podríamos sin problema prescindir del superíndice A).

Se deduce inmediatamente de (13.3) que la matriz de diseño $X$ es de rango deficiente: la primera columna es suma de las restantes. Por tanto, los parámetros no son estimables. Se recurre por ello a una restricción de identificación (véase Sección 3.1,

[1] La mayoría de los manuales citados en lo que precede contienen capítulos sobre ANOVA. Las peculiaridades de los modelos ANOVA han dado lugar sin embargo a bibliografía especializada. Un clásico sobre el tema es Scheffé (1959), y Fisher and McDonald (1978) es una monografía útil. Los libros Peña (1987) y, particularmente, Trocóniz (1987a) contienen ambos un tratamiento detallado del Análisis de Varianza y Diseño Experimental. Bajo el nombre de Análisis de Varianza se incluyen también modelos de efectos aleatorios que difieren de los estudiados aquí. Hay generalizaciones multivariantes (modelos MANOVA) de las que tampoco nos ocupamos.

final) como la siguiente:

$$\sum_{i=1}^{I} \alpha_i^A = 0. \tag{13.5}$$

**Observación 13.1** La restricción (13.5) no tiene sólo el fin utilitario de solucionar el problema de multicolinealidad; hace además interpretables los parámetros $\alpha_i^A$, que son el efecto atribuíble al $i$-ésimo nivel de tratamiento *adicional a la media global*. En el ejemplo descrito más arriba, si el rendimiento de la semilla $i$-ésima fuera $\mu_i$ y la media de los rendimientos de todos los tipos de semillas analizador fuera $\mu$, entonces $\alpha_i^A = \mu_i - \mu$, es decir, la diferencia entre el rendimiento de la semilla $i$-ésima y el rendimiento medio global.

La restricción (13.5) no implica ninguna pérdida de generalidad ni fuerza el modelo; si $\sum_{i=1}^{I} \alpha_i^A = \delta$, siempre podríamos sustraer $\delta/I$ de cada $\alpha_i^A$ e incluirlo en $\alpha$, para lograr un modelo exactamente equivalente que verifica (13.5). El razonamiento es el mismo que hacíamos en la Sección 1.3 al indicar que suponer $E[\vec{\epsilon}\,] = \vec{0}$ en un modelo de regresión lineal con término constante no implica ninguna pérdida de generalidad.

Con el modelo en la forma (13.4), la hipótesis de más frecuente interés será la de nulidad de todos los $\alpha_i^A$ (= igualdad de rendimiento de todas las semillas). En caso de rechazarse, tendríamos interés en ver qué parámetros $\alpha_i^A$ son significativamente distintos (mayores o menores) que el resto, pues ello apuntaría a semillas significativamente mejores o peores.

Hay diversas maneras de abordar la estimación de (13.4). Designemos por $\vec{v}_0, \vec{v}_1, \ldots, \vec{v}_I$ los vectores columna de la matriz de diseño. El problema de estimación consiste en encontrar coeficientes $\hat{\alpha}, \hat{\alpha_i^A} \quad (i = 1, \ldots, I)$ minimizando

$$\sum_{i=1}^{I}\sum_{j=1}^{J} \hat{\epsilon}_{ij}^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} (y_{ij} - \hat{\alpha} - \hat{\alpha_i^A})^2 = ||\vec{y} - \hat{\alpha}\vec{v}_0 - \sum_{i=1}^{I} \hat{\alpha_i^A}\vec{v}_i||^2 \tag{13.6}$$

que verifiquen además (13.5). Si escribimos (13.4) en forma expandida tenemos:

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1J} \\ y_{21} \\ \vdots \\ y_{2J} \\ \vdots \\ y_{I1} \\ \vdots \\ y_{IJ} \end{pmatrix} = \begin{array}{c} \begin{matrix} \vec{v}_0 & \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_I \end{matrix} \\ \begin{pmatrix} 1 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \ldots & 1 \end{pmatrix} \end{array} \begin{pmatrix} \alpha \\ \alpha_1^A \\ \alpha_2^A \\ \vdots \\ \alpha_I^A \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1J} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2J} \\ \vdots \\ \epsilon_{I1} \\ \vdots \\ \epsilon_{IJ} \end{pmatrix}. \tag{13.7}$$

Sea $h$ el subespacio generado por $\vec{v}_0$ y $M_A$ el generado por $\vec{v}_1, \ldots, \vec{v}_I$. La restricción $\sum_{i=1}^{I} \hat{\alpha_i^A} = 0$ impide que $\sum_{i=1}^{I} \hat{\alpha_i^A}\vec{v}_i$ sea cualquier vector de $M_A$; es fácil ver que las combinaciones lineales de la forma indicada son precisamente aquellos vectores de $M_A$ ortogonales a $h$. En efecto, si $\sum_{i=1}^{I} \hat{\alpha_i^A} = 0$,

$$\left(\sum_{i=1}^{I} \hat{\alpha_i^A}\vec{v}_i\right)' \vec{v}_0 = \sum_{i=1}^{I} \hat{\alpha_i^A}(\vec{v}_i\,'\vec{v}_0) = J\sum_{i=1}^{I} \hat{\alpha_i^A} = 0 \tag{13.8}$$

La restricción de identificación restringe pues las combinaciones lineales factibles $\sum_{i=1}^{I} \hat{\alpha_i^A}\vec{v}_i$ a estar en $M_A \cap h^\perp$. Por consiguiente,

1. $\hat{\alpha}\vec{v}_0$ es la proyección de $\vec{y}$ sobre $h$,
2. $\sum_{i=1}^{I} \hat{\alpha_i^A}\vec{v}_i$ es la proyección de $\vec{y}$ sobre $M_A \cap h^\perp$.

y los coeficientes pueden ser determinados con facilidad por separado gracias a la mutua ortogonalidad de estos subespacios:

$$\begin{aligned}
\hat{\alpha} &= (\vec{v}_0\,'\vec{v}_0)^{-1}\vec{v}_0\,'\vec{y} && (13.9)\\
&= \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} && (13.10)\\
&\stackrel{\text{def}}{=} y_{..} && (13.11)
\end{aligned}$$

Haciendo uso del Lema 3.3, pág. 34, y dado que $\vec{v}_0 = \vec{v}_1 + \ldots + \vec{v}_I$ (y por tanto $h \subset M_A$) tenemos que $P_{M_A \cap h^\perp} = P_{M_A} - P_h$ y por consiguiente:

$$\sum_{i=1}^{I} \hat{\alpha_i^A}\vec{v}_i = P_{M_A\cap h^\perp}\vec{y} = P_{M_A}\vec{y} - P_h\vec{y} = \sum_{i=1}^{I}\delta_i\vec{v}_i - y_{..}\vec{v}_0. \qquad (13.12)$$

La estructura de la matriz cuyas columnas son $\vec{v}_1, \ldots, \vec{v}_I$ hace muy fácil calcular los $\delta_i$:

$$\begin{aligned}
\begin{pmatrix}\delta_1\\ \vdots \\ \delta_I\end{pmatrix} &= \left[\begin{pmatrix}\vec{v}_1\,'\\ \vdots \\ \vec{v}_I\,'\end{pmatrix}\begin{pmatrix}\vec{v}_1 & \ldots & \vec{v}_I\end{pmatrix}\right]^{-1}\begin{pmatrix}\vec{v}_1\,'\\ \vdots \\ \vec{v}_I\,'\end{pmatrix}\vec{y} && (13.13)\\
&= \begin{pmatrix}J & 0 & \ldots & 0\\ 0 & J & \ldots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \ldots & J\end{pmatrix}^{-1}\begin{pmatrix}\sum_{j=1}^{J} y_{1j}\\ \vdots \\ \sum_{j=1}^{J} y_{Ij}\end{pmatrix} && (13.14)\\
&\stackrel{\text{def}}{=} \begin{pmatrix}y_{1.}\\ \vdots \\ y_{I.}\end{pmatrix}. && (13.15)
\end{aligned}$$

Llevando este resultado a (13.12):

$$\begin{aligned}
\sum_{i=1}^{I}\hat{\alpha_i^A}\vec{v}_i &= \sum_{i=1}^{I} y_{i.}\vec{v}_i - y_{..}\vec{v}_0 && (13.16)\\
&= \sum_{i=1}^{I} y_{i.}\vec{v}_i - y_{..}\sum_{i=1}^{I}\vec{v}_i && (13.17)\\
&= \sum_{i=1}^{I}(y_{i.} - y_{..})\vec{v}_i. && (13.18)
\end{aligned}$$

y como los $\vec{v}_i$ son linealmente independientes, de (13.18) se deduce que $\hat{\alpha_i^A} = (y_{i.} - y_{..})$. La estimación, pues, es muy sencilla en un modelo equilibrado: todos los cálculos

se reducen al cómputo de medias aritméticas de observaciones en cada nivel (las $y_{i.}$), y de la media global $y_{..}$. Además, como $h$ y $M_A \cap h^\perp$ son mutuamente ortogonales, tenemos que

$$\vec{y} = P_h\vec{y} + P_{M_A\cap h^\perp}\vec{y} + P_{M_A^\perp}\vec{y} = \hat{\alpha}\vec{v}_0 + \sum_{i=1}^{I}\hat{\alpha_i^A}\vec{v}_i + \hat{\epsilon} \tag{13.19}$$

es una descomposición de $\vec{y}$ en partes ortogonales, lo que permite escribir

$$||\vec{y}\,||^2 = ||P_h\vec{y}\,||^2 + ||P_{M_A\cap h^\perp}\vec{y}\,||^2 + ||P_{M_A^\perp}\vec{y}\,||^2, \tag{13.20}$$

o equivalentemente,

$$\begin{aligned} ||\vec{y}\,||^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}^2 &= \hat{\alpha}^2||\vec{v}_0||^2 + \sum_{i=1}^{I}(\hat{\alpha_i^A})^2||\vec{v}_i||^2 + ||\hat{\epsilon}||^2 && (13.21)\\ &= IJy_{..}^2 + J\sum_{i=1}^{I}(y_{i.}-y_{..})^2 \\ &\quad + \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-(y_{i.}-y_{..})-y_{..})^2. && (13.22)\end{aligned}$$

La relación (13.22) es la igualdad fundamental en el análisis de varianza, en el caso de modelos con un único tratamiento. Hay dos circunstancias de interés a señalar:

1. Cada sumando es una forma cuadrática de matriz idempotente. Por ejemplo, el primer sumando en (13.20) es:

   $$||P_h\vec{y}\,||^2 = <P_h\vec{y}, P_h\vec{y}> = \vec{y}\,'P_h'P_h\vec{y} = \vec{y}\,'P_h\vec{y} \tag{13.23}$$

   y análogamente para los restantes términos.

2. Dado que $h$ y $M_A \cap h^\perp$ son ortogonales, la eliminación de $\vec{v}_1, \ldots, \vec{v}_I$ en el modelo para nada modifica $||P_h\vec{y}\,||^2$.

   Por tanto, si en lugar de ajustar el modelo $\vec{y} = \alpha\vec{v}_0 + \sum_{i=1}^{I}\alpha_i^A\vec{v}_i + \vec{\epsilon}$ ajustásemos el más simple $\vec{y} = \alpha\vec{v}_0 + \vec{\eta}$, obtendríamos una descomposición (13.20)-(13.22) de la suma de cuadrados de la siguiente forma:

   $$\begin{aligned} ||\vec{y}\,||^2 &= ||P_h\vec{y}\,||^2 + ||P_h^\perp\vec{y}\,||^2 && (13.24)\\ &= ||\hat{\alpha}\vec{v}_0||^2 + ||\hat{\eta}||^2 && (13.25)\\ &= IJy_{..}^2 + ||\hat{\eta}||^2 && (13.26)\\ &= IJy_{..}^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-y_{..})^2. && (13.27)\end{aligned}$$

La comparación de (13.27) y (13.22) muestra que la suma de cuadrados de los residuos se ha incrementado precisamente en $J\sum_{i=1}^{I}(y_{i.}-y_{..})^2$; ésta es la fracción de la suma de cuadrados de $\vec{y}$ "explicada" por los distintos niveles del tratamiento A. Análogamente se razonaría sobre los demás sumandos de (13.22) que son respectivamente las fracciones de suma de cuadrados de $\vec{y}$ atribuibles a la media global, al mencionado tratamiento A, y al residuo.

### 13.2.1. Contraste de hipótesis.

Consideremos la hipótesis $h\colon \alpha_i^A = 0 \quad (i = 1, \ldots, I)$. De acuerdo con el Teorema 4.2, pág. 45, emplearemos para su contraste el estadístico:

$$Q_h = \frac{(\mathrm{SSE}_h - \mathrm{SSE})/q}{\mathrm{SSE}/(N-p)} \tag{13.28}$$

que bajo la hipótesis nula $h$ se distribuye como $\mathcal{F}_{q,N-p}$. Tal como acabamos de ver, sin embargo,

$$\begin{aligned} \mathrm{SSE}_h - \mathrm{SSE} &= ||P_{M_A \cap h^\perp}\vec{y}\,||^2 = J\sum_{i=1}^{I}(y_{i\cdot} - y_{\cdot\cdot})^2 && (13.29)\\ \mathrm{SSE} &= ||P_{M_A^\perp}\vec{y}\,||^2. && (13.30) \end{aligned}$$

La hipótesis $h$ puede expresarse en la notación empleada en el Capítulo 4.2 así:

$$\underset{\text{filas}}{(I-1)} \quad \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha_i^A \\ \vdots \\ \alpha_I^A \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{13.31}$$

Bastan $I-1$ filas en la matriz en (13.31) porque, debido a la restricción de identificación, si $I-1$ parámetros son cero el restante ha de ser también cero. Por tanto $q = I-1$. Esta misma circunstancia hace que $p = 1 + (I-1) = I$, aunque aparentemente haya un total de $I+1$ parámetros; solo $I$ de entre ellos son libres.

Aunque se pueden aplicar directamente los resultados obtenidos en el Capítulo 4.2, es ilustrativo derivar directamente la distribución de (13.28) bajo $h$ para mostrar que es una $\mathcal{F}_{I-1,IJ-I}$. Bajo $h$, $\vec{y} = \alpha\vec{v}_0 + \vec{\epsilon}$, y por tanto

$$\begin{aligned} P_{M_A\cap h^\perp}\vec{y} &= P_{M_A}\vec{y} - P_h\vec{y} && (13.32)\\ &= (\alpha\vec{v}_0 + P_{M_A}\vec{\epsilon}) - (\alpha\vec{v}_0 + P_h\vec{\epsilon}) \qquad \text{ya que } \vec{v}_0 \in h \subset M_A && (13.33)\\ &= (P_{M_A} - P_h)\vec{\epsilon}. && (13.34) \end{aligned}$$

Por tanto, el numerador de $Q_h$ en (13.28) es

$$||(P_{M_A} - P_h)\vec{\epsilon}||^2 = \vec{\epsilon}\,'(P_{M_A} - P_h)\vec{\epsilon} \sim \sigma^2\chi^2_{I-1}, \tag{13.35}$$

en que nos hemos servido del Lema 4.1 (pág. 39) y del hecho de que $(P_{M_A} - P_h)$ es simétrica idempotente de rango $I-1$. Análogamente, el denominador de $Q_h$ es

$$\begin{aligned} ||P_{M_A^\perp}\vec{y}\,||^2 &= ||(I - P_{M_A})\vec{y}\,||^2 && (13.36)\\ &= ||(I - P_{M_A})\vec{\epsilon}||^2 && (13.37)\\ &= \vec{\epsilon}\,'(I - P_{M_A})\vec{\epsilon} && (13.38)\\ &\sim \sigma^2\chi^2_{IJ-I}. && (13.39) \end{aligned}$$

Como $(P_{M_A} - P_h)P_{M_A^\perp} = 0$, el Lema 4.2 garantiza la independencia de (13.35) y (13.39). En definitiva, bajo la hipótesis nula $h$,

$$\begin{aligned} Q_h &= \frac{(\mathrm{SSE}_h - \mathrm{SSE})/q}{\mathrm{SSE}/(N-p)} && (13.40)\\ &= \frac{J\sum_{i=1}^{I}(y_{i.} - y_{..})^2/(I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - y_{i.})^2/(IJ-I)} && (13.41)\\ &= \frac{(IJ-I)}{(I-1)}\frac{J\sum_{i=1}^{I}(y_{i.} - y_{..})^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - y_{i.})^2} && (13.42)\\ &\sim \mathcal{F}_{I-1,IJ-I}. && (13.43) \end{aligned}$$

Habitualmente se resumen los resultados de un análisis de varianza en un estadillo como el de la Tabla 13.1.

Cuando la hipótesis de nulidad de los respectivos efectos no es cierta, el estadístico correspondiente sigue una distribución $\mathcal{F}$ de Snedecor descentrada (véase B.1). La región crítica está formada por valores grandes del estadístico, evidenciadores de grandes discrepancias entre $\mathrm{SSE}_h$ y SSE.

### 13.2.2. Distribución del recorrido studentizado.

Supongamos que $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2) \quad (i = 1, \ldots, n)$. Sea,

$$w = \max_i X_i - \min_i X_i \tag{13.44}$$

y $s^2$ un estimador de $\sigma^2$ con $\nu$ grados de libertad (es decir, tal que $\nu s^2/\sigma^2 \sim \chi^2_\nu$). Supongamos, además, que $s^2$ es independiente de $X_i \quad (i = 1, \ldots, n)$. Entonces,

$$W = \frac{w}{s} \tag{13.45}$$

sigue una distribución conocida y tabulada (la *distribución del recorrido studentizado*)[2].

### 13.2.3. Búsqueda de diferencias significativas.

El contraste utilizando el estadístico (13.28) no indica qué niveles de tratamiento difieren significativamente, en el caso de producirse el rechazo de la hipótesis nula de igualdad. Tenemos a nuestra disposición toda la teoría en el Capítulo 6 para abordar multiples contrastes simultáneos de hipótesis como: $H_{kl} : \alpha_k^A - \alpha_l^A = 0$ con nivel de significación conjunto $\alpha$.

Observemos sin embargo que bajo la hipótesis nula de igualdad de los parámetros $\alpha_k^A, \alpha_l^A$ las variables aleatorias $\hat{\alpha}_k^A, \hat{\alpha}_l^A$ tienen igual media, igual varianza (esto último debido a ser el modelo equilibrado), y son además independientes de $\hat{\sigma}^2$ (Teorema 4.2, pág. 45). Por tanto, si hay $n$ hipótesis del tipo $H_{kl}$ (a veces denominadas *comparaciones*) y $W^\alpha_{n,\nu}$ es el cuantil $(1-\alpha)$ de la distribución del recorrido studentizado con parámetros $n, \nu$, tenemos que

[2] Tablas en Beyer (1968), reproducidas en el Apéndice **??**.

Cuadro 13.1: Análisis de varianza con un tratamiento.

| **Efecto** | **Suma de cuadrados** | **Grados de libertad** | **Estadístico de contraste** | **Distribución bajo $\mathbf{H_0}$: efecto nulo** |
|---|---|---|---|---|
| Efecto A | $J\sum_{i=1}^{I}(y_{i.}-y_{..})^2$ | $I-1$ | $\dfrac{(IJ-I)J\sum_{i=1}^{I}(y_{i.}-y_{..})^2}{(I-1)\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-y_{i.})^2}$ | $\mathcal{F}_{I-1,IJ-I}$ |
| Media | $IJy_{..}^2$ | 1 | $\dfrac{(IJ-I)IJy_{..}^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-y_{i.})^2}$ | $\mathcal{F}_{1,IJ-I}$ |
| Residuo | $\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-y_{i.})^2$ | $IJ-I$ | | |
| Total | $\sum_{i=1}^{I}\sum_{j=1}^{J}y_{ij}^2$ | $IJ$ | | |

$$\text{Prob}\left\{\bigcap_{k,l}\left[\left|\frac{\hat{\alpha}_k^A - \hat{\alpha}_l^A}{\hat{\sigma}}\right| \leq W_{n,\nu}^{\alpha}\right]\right\} = 1-\alpha, \tag{13.46}$$

lo que sugiere una forma de hacer el contraste simultáneo.

Obsérvese que (13.46) sólo sirve para contrastes de tipo comparación, en tanto que los métodos de Scheffé o máxima $t$ (Capítulo 6) eran mucho más generales. Como a menudo sucede, el método del recorrido studentizado es preferible a los métodos más generales (= da intervalos de confianza menos amplios) en el tipo de situación para el que ha sido diseñado[3].

## 13.3. Aleatorización. Factores de bloque

La aleatorización de la experimentación es una idea relativamente nueva, primero introducida por el estadístico británico Sir R.A. Fisher. En el pasado, la asignación de las unidades experimentales, las parcelas en el caso del ejemplo en la Sección 13.1, a los respectivos tratamientos, los tipos de semilla en el mismo ejemplo, se ha hecho de una manera sistemática o en algunos casos incluso subjetiva. El uso impropio de ciertas reglas o procedimientos para realizar esta asignación puede originar muchos problemas.

> **Ejemplo 13.1** En un estudio médico diseñado para comparar un tratamiento standard para una enfermedad con un tratamiento nuevo, pero potencialmente arriesgado, el grupo de pacientes recibiendo el nuevo tratamiento podría estar formado completamente por personas voluntarias. Estas personas normalmente estarían en condiciones de salud más precaria que las personas que fuesen a recibir el tratamiento standard, siendo este estado de salud relativamente peor el que las impulsa a aceptar el riesgo de un nuevo tratamiento.
>
> Aún en el caso en que ambos tratamientos fuesen igualmente efectivos, el análisis de los resultados al final del estudio seguramente mostraría un estado de salud peor en los pacientes asignados al nuevo tratamiento que en los pacientes asignados al tratamiento standard. Obviamente la conclusión del estudio sería, en este caso particular, que el tratamiento nuevo es menos efectivo que el tratamiento standard. La fuente de *sesgo* en este caso se suele llamar *sesgo de selección*, debido a que las unidades experimentales para los dos niveles de tratamiento no eran similares[4].

Como veremos más adelante, el *sesgo de selección* puede ser minimizado mediante el uso de la *aleatorización*. El uso de la *aleatorización* tiende a establecer un equilibrio entre las unidades experimentales asignadas a cada tratamiento, respecto a factores distintos a los tratamientos y que puedan influir de una u otra manera en el resultado del experimento.

Hemos visto cómo sesgos significativos pueden ocurrir cuando las unidades experimentales se autoasignan a los diversos tratamientos. Lo mismo puede ocurrir cuando la asignación se hace de manera sistemática o de manera subjetiva.

> **Ejemplo 13.2** Supongamos que tenemos una lista con las calificaciones, ordenadas de mayor a menor, de 50 estudiantes en la asignatura A1. Queremos comparar dos métodos distintos de enseñanza en la asignatura A2, muy relacionada a

[3] Ver detalles en Scheffé (1959), p. 76 y ss.

[4] En este caso la similitud se refiere a que dentro de lo posible las unidades experimentales sólo deben diferir en el tratamiento aplicado y no en su condición de salud inicial.

la anterior, en la que se han matriculado dichos estudiantes, asignando al primer método (método nuevo) los primeros 25 estudiantes de la asignatura A1 (es decir, los que tienen calificación más alta) y al segundo método los 25 restantes. Una comparación de los dos métodos de enseñanza para la asignatura A2 no sólo reflejará la diferencia entre los tratamientos, sino también la diferencia entre los dos grupos, posiblemente dada por las calificaciones obtenidas en la asignatura A1.

**Ejemplo 13.3** La asignación subjetiva puede darse, por ejemplo, en un experimento en que se desea comparar dos tipos de semilla y el agricultor siembra (sin utilizar ninguna razón específica para hacerlo) un tipo de semilla en las parcelas que se encuentran en la parte alta de su finca, y el otro tipo de semilla en las que se encuentran en la parte baja de la finca, en este caso tierras más fértiles. Una comparación de los dos tipos de semillas no sólo reflejará la diferencia entre las semillas, sino también la posible diferencia de fertilidad entre las tierras altas y bajas de su finca.

La aleatorización distribuye las unidades experimentales al azar con objeto de eliminar cualquier efecto, aparente o escondido, que pudiese estar presente entre los tratamientos. Se trata de que las comparaciones entre tratamientos midan sólamente la diferencia entre ellos. En resumen, la aleatorización tiende a eliminar la influencia de variables externas que no están controladas directamente en el experimento. Estas variables externas pueden dar lugar a la aparición de un sesgo de selección, y por tanto oscurecer el efecto de los tratamientos sobre las unidades experimentales.

Si, por otro lado, nos consta que una variable externa afecta de manera directa a las unidades experimentales, es una buena idea el efectuar la aleatorización de una manera que tome en cuenta esta circunstancia. En el caso mencionado en el Ejemplo 13.3, sembrando ambos tipos de semillas tanto en la parte alta como en la parte baja de la finca. En cada parte de la finca, y por separado, se aleatorizará el tipo de semilla a sembrar en las respectivas parcelas. Llamaremos *factores de bloque* a las variables externas que afecten directamente a las unidades experimentales, y que por tanto sea necesario *controlar* (por ejemplo, parte de la finca). Además llamaremos *bloque* a toda combinación de *niveles* de los *factores de bloque*. Supongamos que hubiese otro *factor de bloque* en el experimento de las semillas, que bien podría ser cualquier factor presente en ambas partes de la finca. Por ejemplo, alguna variedad de insecto, digamos que los hay de tipo A y de tipo B, que afecte el proceso de germinación de la semilla, y que se sabe precisamente en qué parcelas de las partes alta y baja de la finca se encuentra (es decir, ambos tipos están en ambas partes de la finca y nosotros sabemos dónde). En este caso tendríamos que un *bloque* sería terreno alto e insecto A, y así sucesivamente.

**Ejemplo 13.4** Para ilustrar la idea de aleatorización tomemos nuevamente el Ejemplo 13.3 en que tenemos 3 tipos de semilla y 6 parcelas. Queremos asignar aleatoriamente 2 parcelas a cada tipo de semilla. Numeramos las parcelas arbitrariamente del 1 al 6 y buscamos 6 números aleatorios en cualquier tabla de números aleatorios (o en un generador que suele venir en las calculadoras), colocándolos junto a las parcelas numeradas. En nuestro caso hemos generado los números aleatorios correspondientes a una variable aleatoria uniforme entre 0 y 1. Esquemáticamente lo que tendríamos sería:

| Parcela | Número Aleatorio |
|---|---|
| 1 | 0.3615 |
| 2 | 0.6711 |
| 3 | 0.2095 |
| 4 | 0.2767 |
| 5 | 0.7546 |
| 6 | 0.4547 |

El siguiente paso consiste en ordenar las parcelas de acuerdo al número aleatorio correspondiente y una vez hecho ésto se asigna las dos primeras parcelas al tipo de semilla 1, las dos siguientes al tipo de semilla 2 y las dos últimas al tipo de semilla 3. Es decir:

| Parcela | Número Aleatorio | Tipo de Semilla |
|---|---|---|
| 3 | 0.2095 | 1 |
| 4 | 0.2767 | 1 |
| 1 | 0.3615 | 2 |
| 6 | 0.4547 | 2 |
| 2 | 0.6711 | 3 |
| 5 | 0.7546 | 3 |

Por tanto en nuestro experimento sembraremos las parcelas 3 y 4 con la semilla tipo 1, las parcelas 1 y 6 con la semilla tipo 2, y las parcelas 2 y 5 con la semilla tipo 3.

## CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

**13.1** En los ejemplos más arriba en que se precisaba aleatorizar un diseño se ha descrito conceptualmente el proceso para hacerlo generando números aleatorios. No es necesario en la práctica generar explícitamente números aleatorios y ordenarlos: la función `sample` (existente tanto en S-PLUS como en R) permite obtener con comodidad el mismo resultado.

**13.2** En todo modelo, los supuestos no son de ordinario sino aproximaciones, más o menos razonables. Una cuestión de interés es qué ocurre en los modelos de Análisis de Varianza cuando se incumple: i) El supuesto de normalidad de las perturbaciones, o ii) El supuesto de homoscedasticidad de las perturbaciones. El análisis teórico y la experiencia muestran que el segundo incumplimiento es mucho más grave que el primero.

Un modo rápido de adquirir una cierta apreciación de la influencia de los respectivos incumplimientos consiste en estudiar el comportamiento de los diferentes estadísticos mediante simulación. Obténganse aleatoriamente 200 muestras generadas por un modelo ANOVA (por ejemplo, con un tratamiento, cuatro niveles y cinco replicaciones en cada nivel), en el caso en que el tratamiento no tiene ningún efecto y la distribución de las perturbaciones es, por ejemplo, uniforme o de Cauchy. Compárese el histograma de valores del estadístico para el contraste de $H_0$ : "Tratamiento sin ningún efecto" con su distribución teórica.

**13.3** La Compañía Cereales Novedosos S.A. desea comparar 3 tipos diferentes de envases para un nuevo cereal que intentan lanzar al mercado. Quince

supermercados, con volúmenes de venta aproximadamente iguales, fueron seleccionados como unidades experimentales. Se asignó aleatoriamente cada tipo de envase a 5 supermercados. Supondremos además que todas las condiciones relevantes y distintas del tipo de envase, tales como el precio, espacio físico en los escaparates, y esfuerzos promocionales especiales, se mantuvieron iguales para todos los supermercados en el experimento. Las ventas obtenidas durante una semana determinada para cada tipo de envase fueron:

| **Tipo de envase** | **Ventas semanales** | | | | |
|---|---|---|---|---|---|
| Tipo 1 | 4 | 15 | 21 | 10 | 16 |
| Tipo 2 | 16 | 11 | 9 | 22 | 19 |
| Tipo 3 | 13 | 23 | 22 | 18 | 19 |

1. Construye la tabla ANOVA para este experimento.
2. ¿Existe diferencia en las ventas atribuible a los métodos de envase de cereales? ($\alpha = 0{,}05$). Si la hubiese, ¿cuál de los métodos de envase es estadísticamente diferente a los otros?
3. ¿Hay alguna observación anómala (= con residuo studentizado muy grande)?
4. ¿Hay alguna observación notoriamente influyente? ¿Sobre algún parámetro en particular?
5. ¿Cuáles de entre los supuestos necesarios podrían ser inadecuados en una situación como la analizada? (Ayuda: dada la dispersión de las observaciones presentadas, ¿es razonable suponer que las ventas —siempre no negativas— siguen una distribución normal? ¿Qué efecto tendría el incumplimiento de este supuesto sobre el análisis efectuado?)

**13.4** Un economista obtiene datos sobre las mejoras en la productividad el año pasado para una muestra de empresas que producen accesorios para equipos electrónicos. Las empresas fueron clasificadas de acuerdo al promedio de los gastos en investigación y a desarrollo en los últimos tres años (niveles bajo, moderado y alto). Los resultados del estudio se detallan a continuación (la mejora en productividad se mide en una escala de 0 a 10). Supongamos que el modelo visto en este capítulo es adecuado.

| **Gasto en I+D** | **Mejora productividad** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Bajo | 8.9 | 8.5 | 8.6 | 5.1 | 6.1 | 8.5 | 5.3 | 6.4 | 5.4 | 7.4 |
| Moderado | 7.8 | 6.8 | 8.4 | 7.7 | 6.3 | 7.7 | 9.3 | 7.1 | 7.2 | 6.1 |
| Alto | 8.9 | 8.7 | 7.7 | 9.7 | 8.6 | 9.0 | 8.9 | 8.8 | 8.7 | 8.5 |

1. Construir la tabla ANOVA para estos datos.
2. Basándonos en la mejora de productividad de la empresa, ¿existe diferencia atribuible a los distintos niveles de gasto en I+D ($\alpha = 0{,}05$)? Si la hubiese, ¿las empresas que pertenecen a cuál de los niveles de gasto difieren estadísticamente de las demás?
3. ¿Hay alguna observación anómala (= con residuo studentizado muy grande)?
4. ¿Hay alguna observación notoriamente influyente? ¿Sobre algún parámetro en particular?

**13.5** Supongamos que una empresa está considerando utilizar una de tres políticas para seleccionar a los supervisores en distintas áreas de trabajo. La política A dicta que se promocionará a los trabajadores que estén dentro de la empresa haciendo que participen en cursos de perfeccionamiento impartidos por la misma empresa; la política B también promocionará a los trabajadores de dentro de la empresa haciendo que participen en cursos de perfeccionamiento, pero que en este caso serán impartidos por una Universidad determinada. La política C consiste en seleccionar supervisores experimentados de entre personas que no pertenezcan a la empresa. Supongamos en este caso (aunque de manera irreal) que los supuestos de ANOVA son válidos y que los posibles sesgos han sido eliminados. Los datos son porcentaje de efectividad de cada uno de los 10 trabajadores en cada grupo, de acuerdo a la escala de la empresa (de 0 a 100).

| **Política** | **Productividad** | | | | | | |
|---|---|---|---|---|---|---|---|
| A | 87 | 50 | 68 | 76 | 62 | 59 | 63 |
| B | 56 | 63 | 44 | 58 | 57 | 42 | 53 |
| C | 35 | 26 | 36 | 37 | 39 | 56 | 42 |

1. Construir la tabla ANOVA para estos datos.
2. Basándonos en el porcentaje de efectividad de los supervisores, ¿existe diferencia en la productividad atribuible a las diversas políticas utilizadas por la empresa para seleccionar sus supervisores $(\alpha = 0{,}05)$? Si la hubiese, ¿qué política o políticas difieren estadísticamente de las demás?
3. ¿Hay alguna observación anómala (= con residuo studentizado muy grande)?
4. ¿Hay alguna observación notoriamente influyente? ¿Sobre algún parámetro en particular?

# Capítulo 14

# Análisis de varianza con dos y tres tratamientos.

## 14.1. Introducción.

Entendidas las ideas básicas, el desarrollo de todos los modelos cruzados[1] de ANOVA es similar y sumamente sencillo. Consideremos ahora dos tratamientos (por ej., semillas y fertilizantes), que denominaremos A y B, con $I$ y $J$ niveles respectivamente. Combinamos los $I$ niveles del tratamiento A con los $J$ del tratamiento B replicando cada combinación $K$ veces. Hay dos versiones de modelo que podemos plantear:

$$y_{ijk} = \alpha + \alpha_i^A + \alpha_j^B + \epsilon_{ijk} \qquad \text{(modelo aditivo)} \tag{14.1}$$

$$y_{ijk} = \alpha + \alpha_i^A + \alpha_j^B + \alpha_{ij}^{AB} + \epsilon_{ijk} \qquad \text{(modelo no aditivo).} \tag{14.2}$$

## 14.2. Modelo aditivo.

El modelo (14.1), que estudiamos a continuación, supone que semilla y fertilizante inciden sobre la variable respuesta aditivamente: el efecto de una determinada combinación semilla-fertilizante es la suma de efectos de semilla y fertilizante; $\alpha_i^A$ y $\alpha_j^B$. El modelo (14.2), con más parámetros, sería el indicado cuando la naturaleza del problema analizado sugiriera la posibilidad de sinergias entre parejas de niveles de tratamiento (alguna semilla reaccionando de manera particularmente favorable o desfavorable a la presencia de algún fertilizante); los parámetros $\alpha_{ij}^{AB}$ (llamados interacciones) pretenden recoger ésto. Volveremos sobre el particular.

[1] En oposición a anidados, de los que nos ocuparemos someramente más adelante.

Restringiéndonos de momento al modelo (14.1), podemos expresarlo vectorialmente así:

$$\vec{y} = \alpha\vec{v}_0 + \sum_{i=1}^{I} \alpha_i^A \vec{v}_i + \sum_{j=1}^{J} \alpha_j^B \vec{w}_j + \vec{\epsilon}, \tag{14.3}$$

o, de modo expandido, tal como aparece a continuación:

$$\begin{pmatrix} y_{111} \\ \vdots \\ y_{11K} \\ \vdots \\ y_{211} \\ \vdots \\ y_{21K} \\ \vdots \\ \vdots \\ y_{IJ1} \\ \vdots \\ y_{IJK} \end{pmatrix} = \begin{array}{c} \begin{matrix} \vec{v}_0 & \vec{v}_1 & \vec{v}_2 & \dots & \vec{v}_I & \vec{w}_1 & \vec{w}_2 & \dots & \vec{w}_J \end{matrix} \\ \begin{pmatrix} 1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 \end{pmatrix} \end{array} \begin{pmatrix} \alpha \\ \alpha_1^A \\ \vdots \\ \alpha_I^A \\ \alpha_1^B \\ \vdots \\ \alpha_J^B \end{pmatrix} + \begin{pmatrix} \epsilon_{111} \\ \vdots \\ \epsilon_{11K} \\ \vdots \\ \epsilon_{211} \\ \vdots \\ \epsilon_{21K} \\ \vdots \\ \vdots \\ \epsilon_{IJ1} \\ \vdots \\ \epsilon_{IJK} \end{pmatrix}$$

Llamaremos $h$, $M_A$ y $M_B$ a los espacios generados por la columna $\vec{v}_0$, las columnas $\vec{v}_1, \dots, \vec{v}_I$, y las columnas $\vec{w}_1, \dots, \vec{w}_J$ respectivamente. Hay claramente dos relaciones lineales exactas entre las columnas de la matriz $X$: $\sum_{i=1}^{I} \vec{v}_i = \vec{v}_0$ y $\sum_{j=1}^{J} \vec{w}_j = \vec{v}_0$. La matriz $X$ es de rango $1 + (I-1) + (J-1)$. Para hacer estimables —e interpretables— los parámetros, imponemos las dos restricciones de identificación:

$$\sum_{i=1}^{I} \alpha_i^A = 0 \qquad\qquad \sum_{j=1}^{J} \alpha_j^B = 0. \tag{14.4}$$

Es fácil ver, al igual que en el modelo con un sólo tratamiento, que las restricciones (14.4) tienen el efecto de hacer las combinaciones lineales factibles de $M_A$ y de $M_B$ ortogonales al subespacio $h$. Del mismo modo puede verse que si los $\alpha_i^A, \alpha_j^B$ ($i = 1, \dots, I$ y $j = 1, \dots, J$) verifican las restricciones (14.4), $\sum_{j=1}^{J} \hat{\alpha_j^B} \vec{w}_j \perp \vec{v}_i$, cualquiera que sea $i$. Por consiguiente, $M_A \cap h^\perp$ y $M_B \cap h^\perp$ son subespacios ortogonales a $h$ y entre sí, y

$$\begin{aligned} \vec{y} &= P_h\vec{y} + P_{M_A \cap h^\perp}\vec{y} + P_{M_B \cap h^\perp}\vec{y} + (I - P_{M_A \cap h^\perp} - P_{M_B \cap h^\perp} - P_h)\vec{y} \\ &= \hat{\alpha}\vec{v}_0 + \sum_{j=1}^{J} \hat{\alpha_j^B} \vec{w}_j + \sum_{i=1}^{I} \hat{\alpha_i^A} \vec{v}_i + \hat{\epsilon} \end{aligned} \tag{14.5}$$

es una descomposición de $\vec{y}$ en partes mutuamente ortogonales, lo que permite estimar los parámetros separadamente:

$$\begin{aligned}
\hat{\alpha} &= (\vec{v}_0\,'\vec{v}_0)^{-1}\vec{v}_0\,'\vec{y} && (14.6)\\
&= \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk} && (14.7)\\
&\stackrel{\text{def}}{=} y_{...} && (14.8)
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^{I}\hat{\alpha_i^A}\vec{v}_i &= P_{M_A\cap h^\perp}\vec{y} && (14.9)\\
&= P_{M_A}\vec{y} - P_h\vec{y} && (14.10)\\
&= \sum_{i=1}^{I} y_{i..}\vec{v}_i - y_{...}\vec{v}_0 && (14.11)\\
&= \sum_{i=1}^{I} y_{i..}\vec{v}_i - y_{...}\sum_{i=1}^{I}\vec{v}_i && (14.12)\\
&= \sum_{i=1}^{I}(y_{i..} - y_{...})\vec{v}_i && (14.13)\\
\Rightarrow \hat{\alpha_i^A} &= (y_{i..} - y_{...}) \qquad (i = 1,\ldots,I) && (14.14)
\end{aligned}$$

y del mismo modo:

$$\hat{\alpha_j^B} = (y_{.j.} - y_{...}) \qquad (j = 1,\ldots,J). \tag{14.15}$$

Los residuos se obtienen por diferencia:

$$\begin{aligned}
\hat{\epsilon}_{ijk} &= y_{ijk} - (y_{i..} - y_{...}) - (y_{.j.} - y_{...}) - y_{...} && (14.16)\\
&= y_{ijk} - y_{i..} - y_{.j.} + y_{...} && (14.17)
\end{aligned}$$

Los resultados se resumen habitualmente en un estado similar al de la Tabla 14.1, con:

$$\begin{aligned}
\hat{\epsilon_{ijk}} &= y_{ijk} - y_{i..} - y_{.j.} + y_{...}\\
\hat{\sigma}^2 &= \frac{1}{IJK - I - J + 1}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\hat{\epsilon_{ijk}}^2 && (14.18)
\end{aligned}$$

## 14.3. Modelo con interacción.

Consideramos ahora un modelo como

$$y_{ijk} = \alpha + \alpha_i^A + \alpha_j^B + \alpha_{ij}^{AB} + \epsilon_{ijk} \tag{14.19}$$

o, vectorialmente,

$$\vec{y} = \alpha\vec{v}_0 + \sum_{i=1}^{I}\alpha_i^A\vec{v}_i + \sum_{j=1}^{J}\alpha_j^B\vec{w}_j + \sum_{i=1}^{I}\sum_{j=1}^{J}\alpha_{ij}^{AB}\vec{z}_{ij} + \vec{\epsilon}, \tag{14.20}$$

Cuadro 14.1: Análisis de Varianza con dos tratamientos replicados (modelo aditivo).

| **Efecto** | **Suma de cuadrados** | **Grados de libertad** | **Estadístico de contraste** | **Distribución bajo $\mathbf{H_0}$: efecto nulo** |
|---|---|---|---|---|
| Efecto A | $JK\sum_{i=1}^{I}(y_{i..}-y_{...})^2$ | $I-1$ | $\dfrac{JK\sum_{i=1}^{I}(y_{i..}-y_{...})^2}{(I-1)\hat{\sigma}^2}$ | $\mathcal{F}_{I-1,IJK-I-J+1}$ |
| Efecto B | $IK\sum_{i=1}^{I}(y_{.j.}-y_{...})^2$ | $J-1$ | $\dfrac{IK\sum_{i=1}^{I}(y_{.j.}-y_{...})^2}{(J-1)\hat{\sigma}^2}$ | $\mathcal{F}_{J-1,IJK-I-J+1}$ |
| Media | $IJKy_{...}^2$ | $1$ | $IJKy_{...}^2/\hat{\sigma}^2$ | $\mathcal{F}_{1,IJK-I-J+1}$ |
| Residuo | $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\hat{\epsilon_{ijk}}^2$ | $IJK-I-J+1$ | | |
| Total | $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}y_{ijk}^2$ | $IJK$ | | |

que de forma expandida puede expresarse así:

$$\begin{pmatrix} y_{111} \\ \vdots \\ y_{11K} \\ y_{121} \\ \vdots \\ y_{12K} \\ \vdots \\ \vdots \\ y_{IJ1} \\ \vdots \\ y_{IJK} \end{pmatrix} = \begin{array}{c} \begin{matrix} \vec{v}_0 & \vec{v}_1 & \vec{v}_2 & \dots & \vec{v}_I & \vec{w}_1 & \vec{w}_2 & \dots & \vec{w}_J & \vec{z}_{11} & \dots & \vec{z}_{IJ} \end{matrix} \\ \begin{pmatrix} 1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & 1 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \dots & 0 & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 & 0 & \dots & 1 \end{pmatrix} \end{array} \begin{pmatrix} \alpha \\ \alpha_1^A \\ \vdots \\ \alpha_I^A \\ \alpha_1^B \\ \vdots \\ \alpha_J^B \\ \alpha_{11}^{AB} \\ \vdots \\ \alpha_{IJ}^{AB} \end{pmatrix} + \vec{\epsilon}$$

El vector $\vec{z}_{ij}$ tiene "unos" en los lugares enfrentados a términos $y_{ijk} \quad (k = 1, \dots, K)$, y ceros en los restantes. Las restricciones identificando a los parámetros son ahora:

$$\sum_{i=1}^{I} \alpha_i^A = 0 \qquad \sum_{j=1}^{J} \alpha_j^B = 0 \tag{14.21}$$

$$\sum_{i=1}^{I} \alpha_{ij}^{AB} = 0, \qquad \forall j \tag{14.22}$$

$$\sum_{j=1}^{J} \alpha_{ij}^{AB} = 0, \qquad \forall i \tag{14.23}$$

y es de nuevo fácil comprobar que dichas restricciones ortogonalizan las combinaciones lineales factibles en cada uno de los sumandos de (14.20). Por consiguiente,

$$\vec{y} = \hat{\alpha}\vec{v}_0 + \sum_{i=1}^{I} \hat{\alpha_i^A}\vec{v}_i + \sum_{j=1}^{J} \hat{\alpha_j^B}\vec{w}_j + \sum_{i=1}^{I}\sum_{j=1}^{J} \hat{\alpha_{ij}^{AB}}\vec{z}_{ij} + \hat{\epsilon} \tag{14.24}$$

es una descomposición de $\vec{y}$ cuyos tres primeros sumandos son respectivamente proyecciones sobre $h$, $M_A \cap h^\perp$, $M_B \cap h^\perp$, y $M_I = M_{AB} - M_A \cap h^\perp - M_B \cap h^\perp - h$, siendo $M_{AB}$ el subespacio generado por los vectores $\{\vec{z}_{ij}\}_{i=1,I}^{j=1,J}$. De (14.20) se deduce entonces:

$$\begin{aligned} ||\vec{y}\ ||^2 &= ||P_h\vec{y}\ ||^2 + ||P_{M_A \cap h^\perp}\vec{y}\ ||^2 + ||P_{M_B \cap h^\perp}\vec{y}\ ||^2 + ||P_{M_I}\vec{y}\ ||^2 + ||\hat{\epsilon}||^2 \qquad (14.25) \\ &= \hat{\alpha}^2||\vec{v}_0||^2 + \sum_{i=1}^{I}(\hat{\alpha_i^A})^2||\vec{v}_i||^2 + \sum_{j=1}^{J}(\hat{\alpha_j^B})^2||\vec{w}_j||^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(\hat{\alpha_{ij}^{AB}})^2||\vec{z}_{ij}||^2 \\ &\quad + \sum_{i,j,k=1}^{I,J,K} \hat{\epsilon_{ijk}}^2 \qquad (14.26) \end{aligned}$$

Explotando la ortogonalidad por bloques y siguiendo un procedimiento enteramente análogo al empleado con el modelo aditivo, llegaríamos a los siguientes resultados:

$$\hat{\alpha} = y_{...} \tag{14.27}$$
$$\hat{\alpha_i^A} = y_{i..} - y_{...} \tag{14.28}$$
$$\hat{\alpha_j^B} = y_{.j.} - y_{...} \tag{14.29}$$
$$\hat{\alpha_{ij}^{AB}} = y_{ij.} - y_{i..} - y_{.j.} + y_{...} \tag{14.30}$$
$$\hat{\epsilon_{ijk}} = y_{ijk} - y_{ij.} \tag{14.31}$$

Llevados a (14.26), proporcionan:

$$\begin{aligned}\sum_{i=1}^{I} y_{ijk}^2 &= IJKy_{...}^2 + JK\sum_{i=1}^{I}(y_{i..} - y_{...})^2 + IK\sum_{j=1}^{J}(y_{.j.} - y_{...})^2 \\ &\quad + K\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - y_{ij.})^2\end{aligned} \tag{14.32}$$

Los cálculos y contrastes suelen agruparse en un cuadro como el presentado en la Tabla 14.2. Los grados de libertad se obtienen sin más que considerar el rango de las matrices de proyección correspondientes. Por ejemplo,

$$\begin{aligned}\text{rango } P_{M_I} &= \text{traza } P_{M_I} \\ &= \text{traza } P_{M_{AB}} - \text{traza } P_{M_A \cap h^\perp} - \text{traza } P_{M_B \cap h^\perp} - \text{traza } P_h \\ &= \text{traza } P_{M_{AB}} - \text{traza } P_{M_A} + \text{traza } P_h - \text{traza } P_{M_B} + \text{traza } P_h - \text{traza } P_h \\ &= IJ - I - J + 1\end{aligned}$$

y análogamente con las demás.

Cuadro 14.2: Análisis de Varianza equilibrado con dos tratamientos replicados (modelo con interacción)

| **Efecto** | **Suma de cuadrados** | **Grados de libertad** | **Estadístico de contraste** | **Distribución bajo $\mathbf{H_0}$: efecto nulo** |
|---|---|---|---|---|
| Efecto A | $JK\sum_{i=1}^{I}(y_{i..}-y_{...})^2$ | $I-1$ | $\dfrac{JK\sum_{i=1}^{I}(y_{i..}-y_{...})^2}{(I-1)\hat{\sigma}^2}$ | $\mathcal{F}_{I-1,IJ(K-1)}$ |
| Efecto B | $IK\sum_{j=1}^{J}(y_{.j.}-y_{...})^2$ | $J-1$ | $\dfrac{IK\sum_{j=1}^{J}(y_{.j.}-y_{...})^2}{(J-1)\hat{\sigma}^2}$ | $\mathcal{F}_{J-1,IJ(K-1)}$ |
| Interacción | $K\sum_{i=1}^{I}\sum_{j=1}^{J}(\alpha_{ij}^{\hat{AB}})^2$ | $(IJ-I-J+1)$ | $\dfrac{K\sum_{i=1}^{I}\sum_{j=1}^{J}(\alpha_{ij}^{\hat{AB}})^2}{(IJ-I-J+1)\hat{\sigma}^2}$ | $\mathcal{F}_{IJ-I-J+1,IJ(K-1)}$ |
| Media | $IJKy_{...}^2$ | 1 | $IJKy_{...}^2/\hat{\sigma}^2$ | $\mathcal{F}_{1,IJ(K-1)}$ |
| Residuo | $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\hat{\epsilon_{ijk}}^2$ | $IJ(K-1)$ | | |
| Total | $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}y_{ijk}^2$ | $IJK$ | | |

## 14.4. Aleatorización de la experimentación

Siguiendo las ideas mencionadas en la Sección 13.3 sobre la necesidad de asignar adecuadamente las unidades experimentales a los diversos tratamientos en un experimento, ilustraremos con un ejemplo simple la manera de aleatorizar cuando se trata de un experimento equilibrado con dos tratamientos.

**Ejemplo 14.1** Supondremos que se tienen dos factores, que por ejemplo podrían ser tipo de semilla y tipo de fertilizante. Tenemos $I = 2$ tipos de semillas y $J = 3$ tipos de fertilizantes. Contamos con 12 parcelas que debemos asignar *aleatoriamente* a las distintas combinaciones de tratamientos. El experimento consiste en observar el efecto que el tipo de semilla y el tipo de fertilizante tienen en la cosecha obtenida por unidad de superficie.

Numeramos las parcelas arbitrariamente del 1 al 12 y siguiendo un proceso similar al utilizado en el caso de un experimento equilibrado con un tratamiento, asignamos 6 parcelas al tipo de semilla 1 y 6 parcelas al tipo de semilla 2, a la vez que asignamos 4 parcelas a cada tipo de fertilizante. Es decir la asignación de parcelas a tipo de semilla y fertilizante quedaría así:

| **Parcela** | **Número Aleatorio** | **Tipo de Semilla** | **Tipo de Fertilizante** |
|---|---|---|---|
| 12 | 0.0129 | 1 | 1 |
| 2 | 0.0261 | 1 | 1 |
| 8 | 0.0391 | 1 | 2 |
| 3 | 0.0707 | 1 | 2 |
| 4 | 0.1284 | 1 | 3 |
| 1 | 0.1855 | 1 | 3 |
| 5 | 0.2612 | 2 | 1 |
| 6 | 0.4036 | 2 | 1 |
| 10 | 0.4869 | 2 | 2 |
| 9 | 0.7985 | 2 | 2 |
| 7 | 0.8358 | 2 | 3 |
| 11 | 0.9861 | 2 | 3 |

Lo cual normalmente se mostraría de manera esquemática como:

| | **Fertilizante** | | |
|---|---|---|---|
| **Semilla** | **1** | **2** | **3** |
| **1** | Parcelas 2 y 12 | Parcelas 3 y 8 | Parcelas 1 y 4 |
| **2** | Parcelas 5 y 6 | Parcelas 9 y 10 | Parcelas 7 y 11 |

## 14.5. Análisis de varianza equilibrado con tres tratamientos.

La extensión y posterior generalización de los resultados anteriores a modelos con tres tratamientos es inmediata, con ciertos matices que debemos clarificar. Consideremos ahora tres tratamientos (por ejemplo: semillas, fertilizantes y tipo de riego utilizado), que denominaremos A, B y C, con $I$, $J$ y $K$ niveles respectivamente. Combinamos los $I$ niveles del tratamiento A con los $J$ del tratamiento B y con los $K$ del tratamiento C replicando cada combinación $L$ veces. Al igual que cuando describíamos el modelo equilibrado con dos tratamientos, en este caso tendremos interacciones dobles (es

decir, AB, AC y BC) y una interacción triple (es decir, ABC) entre los tratamientos presentes en el experimento. Llamaremos modelo aditivo al modelo que no contiene ningún tipo de interacción, modelo no aditivo de primer orden al que contiene tan sólo las interacciones dobles, y modelo no aditivo de segundo orden al que contiene todas las interacciones dobles y la triple.

**Modelo aditivo:**

$$y_{ijkl} = \alpha + \alpha_i^A + \alpha_j^B + \alpha_k^C + \epsilon_{ijkl} \tag{14.33}$$

**Modelo no aditivo de primer orden:**

$$y_{ijkl} = \alpha + \alpha_i^A + \alpha_j^B + \alpha_k^C + \alpha_{ij}^{AB} + \alpha_{ik}^{AC} + \alpha_{jk}^{BC} + \epsilon_{ijkl} \tag{14.34}$$

**Modelo no aditivo de segundo orden:**

$$y_{ijkl} = \alpha + \alpha_i^A + \alpha_j^B + \alpha_k^C + \alpha_{ij}^{AB} + \alpha_{ik}^{AC} + \alpha_{jk}^{BC} + \alpha_{ijk}^{ABC} + \epsilon_{ijkl} \tag{14.35}$$

El significado de las interacciones dobles es similar al de análisis de varianza con dos tratamientos, manteniendo el tercer tratamiento fijo en uno de sus niveles. La inclusión en el modelo de la interacción triple, o en este caso el uso del modelo no aditivo de segundo orden, se justificaría si la naturaleza del problema analizado sugiriese la posibilidad de sinergias entre parejas de niveles de dos tratamientos cualesquiera (de entre los tres tratamientos) para los distintos niveles del tercer tratamiento (por ejemplo, alguna semilla reaccionando de manera particularmente favorable o desfavorable a la presencia de algún fertilizante, siendo dicha reacción diferente dependiendo del tipo de riego que se esté utilizando).

Cualquiera de los tres modelos mencionados puede ser expresado ya sea vectorialmente o en forma expandida de manera inmediata siguiendo las ideas del análisis de varianza equilibrado con dos tratamientos, e imponiendo de la misma forma las restricciones que identifiquen a los parámetros en este nuevo modelo. Los cálculos y contrastes suelen agruparse en un cuadro como el presentado en las Tablas 14.3 y 14.4 (para el caso del modelo no aditivo de segundo orden). El estadístico para contrastar la nulidad de cada efecto se obtiene como cociente de la respectiva suma de cuadrados y la asociada al residuo (dividida cada una por sus grados de libertad). Bajo la hipótesis de nulidad, los estadísticos siguen las distribuciones que aparecen en la tabla.

Cuadro 14.3: Análisis de Varianza equilibrado con tres tratamientos replicados (modelo no aditivo de segundo orden)

| **Efecto** | **Suma de cuadrados** | **Grados de libertad** | **Distribución bajo $\mathbf{H_0}$: efecto nulo** |
|---|---|---|---|
| Efecto A | $JKL\sum_{i=1}^{I}(y_{i...}-y_{....})^2$ | $I-1$ | $\mathcal{F}_{I-1,IJKL-IJK}$ |
| Efecto B | $IKL\sum_{j=1}^{J}(y_{.j..}-y_{....})^2$ | $J-1$ | $\mathcal{F}_{J-1,IJKL-IJK}$ |
| Efecto C | $IJL\sum_{k=1}^{K}(y_{..k.}-y_{....})^2$ | $K-1$ | $\mathcal{F}_{K-1,IJKL-IJK}$ |
| Interacción AB | $KL\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij..}-y_{i...}-y_{.j..}+y_{....})^2$ | $IJ-I-J+1$ | $\mathcal{F}_{IJ-I-J+1,IJKL-IJK}$ |
| Interacción AC | $JL\sum_{i=1}^{I}\sum_{k=1}^{K}(y_{i.k.}-y_{i...}-y_{..k.}+y_{....})^2$ | $IK-I-K+1$ | $\mathcal{F}_{IK-I-K+1,IJKL-IJK}$ |

Cuadro 14.4: Análisis de Varianza equilibrado con tres tratamientos replicados (modelo no aditivo de segundo orden). Continuación.

| **Efecto** | **Suma de cuadrados** | **Grados de libertad** | **Distribución bajo $\mathbf{H_0}$: efecto nulo** |
|---|---|---|---|
| Interacción BC | $IL\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{.jk.} - y_{.j..} - y_{..k.} + y_{....})^2$ | $JK - J - K + 1$ | $\mathcal{F}_{JK-J-K+1,IJKL-IJK}$ |
| Interacción ABC | $L\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk.} + y_{i...} + y_{.j..} + y_{..k.} - y_{ij..} - y_{i.k.} - y_{.jk.} - y_{....})^2$ | $IJK - IJ - IK - JK + I + J + K - 1$ | $\mathcal{F}_{IJK-IJ-IK-JK+I+J+K-1,IJKL-IJK}$ |
| Media | $IJKLy_{....}^2$ | 1 | $\mathcal{F}_{1,IJKL-IJK}$ |
| Residuo | $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}(y_{ijkl} - y_{ijk.})^2$ | $IJKL - IJK$ | |
| Total | $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}y_{ijkl}^2$ | $IJKL$ | |

## CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

**14.1** Consideremos un experimento para medir la pérdida de calor que se tiene cuando se usan 4 tipos de cristal térmico para ventanas (Tratamiento 1), y 5 temperaturas exteriores prefijadas (Tratamiento 2). Se prueban tres cristales de cada tipo para cada temperatura exterior y se mide la correspondiente pérdida de calor expresada porcentualmente. Los datos obtenidos son:

| | **Tipo de cristal** | | | |
|---|---|---|---|---|
| **Temperatura exterior** | **A** | **B** | **C** | **D** |
| 15 | 2.9 | 7.2 | 10.0 | 12.1 |
| | 3.6 | 8.0 | 11.2 | 13.3 |
| | 4.0 | 7.7 | 10.6 | 12.6 |
| 10 | 5.4 | 10.3 | 12.6 | 14.7 |
| | 6.3 | 9.8 | 13.1 | 14.5 |
| | 5.8 | 9.4 | 12.7 | 15.2 |
| 5 | 7.2 | 10.7 | 13.1 | 15.2 |
| | 6.7 | 11.3 | 14.3 | 15.8 |
| | 6.2 | 10.4 | 13.7 | 15.9 |
| 0 | 7.7 | 11.7 | 14.3 | 17.2 |
| | 7.4 | 12.1 | 14.7 | 16.8 |
| | 7.9 | 11.4 | 15.0 | 16.1 |
| –5 | 9.8 | 13.3 | 16.8 | 18.5 |
| | 9.6 | 13.7 | 16.7 | 18.9 |
| | 9.7 | 14.2 | 16.5 | 18.5 |

1. Construye la tabla ANOVA para este experimento equilibrado con dos tratamientos, utilizando el modelo adecuado (es decir, aditivo o no aditivo). Si decides usar el modelo aditivo, justifica esta decisión (es decir, ¿hay algún tipo de cristal que tenga una pérdida de calor significativamente diferente a la que cabría esperar de la consideración separada de los efectos “cristal” y “temperatura”?).

2. Contrasta la hipótesis de que los cuatro tipos de cristal tienen igual pérdida de calor.

3. Contrasta la hipótesis de que la pérdida de calor es la misma para las distintos temperaturas.

4. ¿Hay alguna observación anómala (= con residuo studentizado muy grande)?

5. ¿Hay alguna observación notoriamente influyente? ¿Sobre algún parámetro en particular?

**14.2** Un experimento consiste en analizar el efecto que sobre la producción de plátanos (en cajas de 2 kilogramos por racimo) tienen 4 tipos de terreno y 2 tipos de fertilizantes. Se observa la producción promedio en una parcela para cada combinación terreno-fertilizante. Los datos se muestran a continuación:

| | **Fertilizante** | |
|---|---|---|
| **Terreno** | **A** | **B** |
| T1 | 0.9 | 1.4 |
| | 1.1 | 1.5 |
| T2 | 0.7 | 1.2 |
| | 0.8 | 1.0 |
| T3 | 1.8 | 0.8 |
| | 2.2 | 1.0 |
| T4 | 1.4 | 1.3 |
| | 1.2 | 1.5 |

1. Construye la tabla ANOVA para este experimento equilibrado con dos tratamientos, utilizando el modelo adecuado (es decir, aditivo o no aditivo). Si decides usar el modelo aditivo, justifica esta decisión (es decir, hay algún tipo de terreno que tenga una producción de plátanos significativamente diferente a la que cabría esperar de la consideración separada de los efectos "terreno" y "fertilizante").
2. Contrasta la hipótesis de que los cuatro tipos de terreno tienen igual producción de plátanos.
3. Contrasta la hipótesis de que la producción de plátanos es la misma para los distintos fertilizantes.
4. ¿Hay alguna observación anómala (= con residuo studentizado muy grande)?
5. ¿Hay alguna observación notoriamente influyente? ¿Sobre algún parámetro en particular?

**14.3** En un experimento realizado en la Universidad de Iowa[2], se quería probar la rapidez de un programa de ordenador para calcular probabilidades en una distribución t de Student no centrada. Se observaron los tiempos (en microsegundos) que el programa usó para calcular $P(T \leq t|\nu, \delta)$, para distintos valores de $t$ y grados de libertad $\nu$ cuando el parámetro de nocentralidad era $\delta = 2$. Supón que las asunciones del modelo ANOVA con dos tratamientos son válidas. Los datos se muestran a continuación:

| | **Grados de libertad** | |
|---|---|---|
| | $\nu = 5$ | $\nu = 51$ |
| $t = 0$ | 363 | 272 |
| | 360 | 321 |
| | 214 | 312 |
| $t = 2$ | 1934 | 4719 |
| | 2205 | 5066 |
| | 1994 | 4769 |

1. Construye la tabla ANOVA para este experimento equilibrado con dos tratamientos, utilizando el modelo adecuado (es decir, aditivo o no aditivo). Si decides usar el modelo aditivo, justifica esta decisión (es decir, hay algún valor de t que utilice un tiempo significativamente diferente a la que cabría esperar de la consideración separada de los efectos "valor de t" y "grados de libertad").

[2]Estos datos aparecen en un folleto escrito por Rusell Lenth en 1989 y titulado "Experimental Design. 22S:158 #1. Spring 89" que era utilizado para la enseñanza del curso de Diseño de Experimentos en dicha Universidad.

2. Contrasta la hipótesis de que el programa utiliza un tiempo igual para los dos valores de t.
3. Contrasta la hipótesis de que el tiempo utilizado por el programa es el mismo para los distintos grados de libertad.
4. ¿Hay alguna observación anómala (= con residuo studentizado muy grande)?
5. ¿Hay alguna observación notoriamente influyente? ¿Sobre algún parámetro en particular?

**14.4** Un experimento consiste en analizar el efecto que sobre la producción de plátanos (en cajas de 2 kilogramos por racimo) tienen 4 tipos de terreno, 2 tipos de fertilizantes y dos tipos de riego (R1 y R2). Se observa la producción promedio en una parcela para cada combinación terreno-fertilizante-riego. Los datos se muestran a continuación:

| | **Riego R1** | | **Riego R2** | |
|---|---|---|---|---|
| **Fertilizante** / **Terreno** | **A** | **B** | **A** | **B** |
| T1 | 0.8 | 1.3 | 1.2 | 1.4 |
| | 1.0 | 1.4 | 1.3 | 1.6 |
| T2 | 0.5 | 1.1 | 0.9 | 1.4 |
| | 0.6 | 0.8 | 0.8 | 1.2 |
| T3 | 1.4 | 0.5 | 1.7 | 1.0 |
| | 2.0 | 1.0 | 2.5 | 1.1 |
| T4 | 1.4 | 1.3 | 1.7 | 1.3 |
| | 1.2 | 1.5 | 1.5 | 1.5 |

1. Construye la tabla ANOVA para este experimento equilibrado con tres tratamientos, utilizando el modelo adecuado (es decir, aditivo, no aditivo de primer orden o no aditivo de segundo orden). Justifica tu decisión efectuando los contrastes adecuados.
2. Contrasta la hipótesis de que los cuatro tipos de terreno tienen igual producción de plátanos.
3. Contrasta la hipótesis de que la producción de plátanos es la misma para los distintos fertilizantes.
4. Contrasta la hipótesis de que la producción de plátanos es la misma para los distintos tipos de riego.
5. ¿Hay alguna observación anómala (= con residuo studentizado muy grande)?
6. ¿Hay alguna observación notoriamente influyente? ¿Sobre algún parámetro en particular?

# Capítulo 15

# Otros diseños.

## 15.1. Introducción.

Como hemos visto en el modelo con tres tratamientos, la generalización de este tipo de resultados a modelos con más tratamientos es inmediata, y no requiere mayor aclaración.

Acontece, sin embargo, que el volumen de experimentacion requerido para estimar un modelo con muchos tratamientos puede fácilmente exceder de lo permisible. Por ejemplo, con dos tratamientos de tres niveles cada uno, son precisos nueve experimentos. Si deseamos efectuar $k$ replicaciones, el número de experimentos será $9k$.

Si el número de tratamientos simultáneos que investigamos es de cinco, y suponemos de nuevo que cada uno posee tres niveles, el número de experimentos a realizar replicando $k$ veces es de $243k$ $(= 3^5 \times k)$. Es fácil entender que un modelo factorial completo puede ser imposible o extremadamente caro de estimar.

En casos particulares es posible, sin embargo, salir del paso con un número mucho menor de experimentos, con el único requisito de que podamos prescindir de ciertos parámetros, por ser razonable suponerlos *a priori* iguales a cero. Algunos de los casos más habituales son los presentados a continuación.

## 15.2. Modelos no completos. Cuadrados latinos.

En ocasiones no es posible realizar todas las experimentaciones que requiere un modelo *completo* —es decir, que ensaya todas y cada una de las combinaciones posibles con los niveles de cada tratamiento—. Con algunas limitaciones, es posible sin embargo obtener bastante información realizando un número menor de ensayos. La descripción sumaria que hacemos del modelo de cuadrado latino ilustra la manera en que esto puede llevarse a cabo. Supongamos tres tratamientos, con las siguientes dos importantes restricciones:

1. Nos consta que no hay interacciones.
2. Cada tratamiento tiene el mismo número de niveles.

Imaginemos, para trabajar sobre un caso concreto, que el número común $n$ de niveles es tres. Supondremos además que se verifican los supuestos habituales, que hay normalidad en las perturbaciones, y que los valores observados de la variable respuesta se generan así:

$$y_{ijk} = \alpha + \alpha_i^A + \alpha_j^B + \alpha_k^C + \epsilon_{ijk} \tag{15.1}$$

Es costumbre el nombrar mediante letras latinas los niveles del último tratamiento. Como en nuestro ejemplo son tres, los denotaremos por $a$, $b$, y $c$. El diseño cuadrado latino consiste en realizar los experimentos de forma que cada nivel de cada tratamiento resulte combinado el mismo número de veces (una, si no hay replicación) con cada uno de los demás. Puede representarse este diseño como un cuadro en que cada una de las letras $a$, $b$, y $c$ apareciera una sola vez en cada fila y columna. Por ejemplo,

| I/J | 1 | 2 | 3 |
|---|---|---|---|
| 1 | a | b | c |
| 2 | b | c | a |
| 3 | c | a | b |

es un cuadrado latino representando un diseño en que se hubieran ensayado las siguientes combinaciones de niveles:

| I | J | K |
|---|---|---|
| 1 | 1 | $a$ |
| 1 | 2 | $b$ |
| 1 | 3 | $c$ |
| 2 | 1 | $b$ |
| 2 | 2 | $c$ |
| 2 | 3 | $a$ |
| 3 | 1 | $c$ |
| 3 | 2 | $a$ |
| 3 | 3 | $b$ |

En este caso concreto, (15.1) podría escribirse así:

$$\begin{pmatrix} y_{11a} \\ y_{12b} \\ y_{13c} \\ y_{21b} \\ y_{22c} \\ y_{23a} \\ y_{31c} \\ y_{32a} \\ y_{33b} \end{pmatrix} = \begin{array}{c} \begin{matrix} \vec{v}_0 & \vec{v}_1 & \vec{v}_2 & \vec{v}_3 & \vec{w}_1 & \vec{w}_2 & \vec{w}_3 & \vec{z}_1 & \vec{z}_2 & \vec{z}_3 \end{matrix} \\ \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} \end{array} \begin{pmatrix} \alpha \\ \alpha_1^A \\ \alpha_2^A \\ \alpha_3^A \\ \alpha_1^B \\ \alpha_2^B \\ \alpha_3^B \\ \alpha_1^C \\ \alpha_2^C \\ \alpha_3^C \end{pmatrix} + \begin{pmatrix} \epsilon_{11a} \\ \epsilon_{12b} \\ \epsilon_{13c} \\ \epsilon_{21b} \\ \epsilon_{22c} \\ \epsilon_{23a} \\ \epsilon_{31c} \\ \epsilon_{32a} \\ \epsilon_{33b} \end{pmatrix} \tag{15.2}$$

Es fácil ver que si imponemos las restricciones de identificación

$$\sum_{i=1}^{I} \alpha_i^A = 0 \qquad \sum_{j=1}^{J} \alpha_j^B = 0 \qquad \sum_{k=1}^{K} \alpha_k^C = 0, \tag{15.3}$$

Cuadro 15.1: Análisis de Varianza. Cuadrado Latino.

| **Efecto** | **Suma de cuadrados** | **Grados de libertad** | **Estadístico de contraste** | **Distribución bajo $\mathbf{H_0}$: efecto nulo** |
|---|---|---|---|---|
| Efecto A | $n\sum_{i=1}^{I}(y_{i..}-y_{...})^2$ | $n-1$ | $\frac{n\sum_{i=1}^{I}(y_{i..}-y_{...})^2}{(n-1)\hat{\sigma}^2}$ | $\mathcal{F}_{n-1,n^2-3n+2}$ |
| Efecto B | $n\sum_{j=1}^{J}(y_{.j.}-y_{...})^2$ | $n-1$ | $\frac{n\sum_{j=1}^{J}(y_{.j.}-y_{...})^2}{(n-1)\hat{\sigma}^2}$ | $\mathcal{F}_{n-1,n^2-3n+2}$ |
| Efecto C | $n\sum_{k=1}^{K}(y_{..k}-y_{...})^2$ | $n-1$ | $\frac{n\sum_{k=1}^{K}(y_{..k}-y_{...})^2}{(n-1)\hat{\sigma}^2}$ | $\mathcal{F}_{n-1,n^2-3n+2}$ |
| Media | $n^2y_{...}^2$ | 1 | $n^2y_{...}^2/\hat{\sigma}^2$ | $\mathcal{F}_{1,n^2-3n+2}$ |
| Residuo | $\sum_{i,j,k}\hat{\epsilon_{ijk}}^2$ | $n^2-3n+2$ | | |
| Total | $\sum_{i,j,k}y_{ijk}^2$ | $n^2$ | | |

las combinaciones lineales factibles generan subespacios $h$, $M_A \cap h^{\perp}$, $M_B \cap h^{\perp}$, y $M_C \cap h^{\perp}$ mutuamente ortogonales. El desarrollo es entonces paralelo al efectuado en las Secciones anteriores, y por ello omitiremos detalles. Los cálculos necesarios se resumen en un estadillo como el que se presenta (Tabla 15.1).

El desarrollo es fácilmente generalizable, tanto a cuadrados latinos replicados (podríamos por ejemplo adoptar un diseño que realizara $2n^2$ experimentos, escogidos de modo que conformaran dos cuadrados latinos superponibles) como a modelos de orden superior.

## 15.3. Modelos de orden superior.

Cualquier modelo cruzado puede ser desarrollado sobre las líneas de lo expuesto anteriormente. Hay así modelos completos de tres tratamientos (con posibilidad de interacciones entre pares y entre tríadas de niveles), y modelos con cuatro y mas tra-

tamientos, completos o no (entre estos últimos está el llamado cuadrado grecolatino, construído como el latino combinando en cada celda del cuadro una letra latina y una griega; como el cuadrado latino requiere la ausencia de interacciones).

En general, rara vez se usan modelos completos con mas de tres o cuatro tratamientos, en vista del enorme número de experimentos que sería necesario realizar. Se prefiere en el caso de tener que estudiar muchos tratamientos el empleo de modelos incompletos que, si es posible prescindir a priori de algunas interacciones, permiten una drástica disminución de la experimentacion y del costo. Trocóniz (1987a) contiene una cobertura amplia y referencias. Otras obras de interés son Raktoe (1981), Dey (1985), y Petersen (1985).

## 15.4. Modelos anidados.

Hay modelos ANOVA muy diferentes a los estudiados sucintamente aquí. Un ejemplo servirá de introducción a los *modelos anidados*. Supongamos que estamos interesados en estudiar la eficiencia de varias granjas en la producción de animales de engorde. Tenemos tres granjas, dentro de cada una de las cuales escogemos cinco animales cuyo progreso en el engorde registramos. Supongamos tambien que la velocidad de engorde depende acaso de la granja (dieta, factores ambientales, ...) y del animal escogido, habiendo algunos genéticamente mejor dispuestos que otros. Supongamos también que de cada animal tenemos $K$ observaciones (obtenidas por ejemplo en semanas consecutivas que podamos considerar homogéneas a efectos de engorde).

Claramente, no tendría sentido etiquetar los cinco animales y formular un modelo cruzado $3 \times 5$ como los estudiados anteriormente. Cada una de las cinco columnas no recogería nada común; no se trata de cinco animales engordados sucesivamente en las tres granjas, sino de *quince* animales. El animal numero 1 de la granja 1 es distinto del animal número 1 de la granja 2, y carece de sentido pensar en un único $\alpha_{,1}$ recogiendo el "efecto animal 1".

Mas sentido tendría plantearse un modelo como:

$$y_{ijk} = \alpha + \alpha_i^A + \alpha_{ij}^{A \leftarrow B} + \epsilon_{ijk} \tag{15.4}$$

De esta manera, $\alpha_i^A$ recogería el efecto granja —si lo hubiera—, cada uno de los $\alpha_{ij}^{A \leftarrow B}$ el efecto *del animal j-ésimo en la granja i-ésima*, y $\epsilon_{ijk}$ la perturbación afectando a la $k$-ésima observación realizada sobre el animal $ij$-ésimo. Un modelo como (15.4) se dice que es anidado. Es un ejercicio útil escribir su matriz de diseño para un número pequeño de niveles y replicaciones, comparándola con la de un modelo cruzado. En este caso concreto, y suponiendo que tenemos $I = 2$ granjas, $J = 2$ animales en cada granja y $K = 2$ observaciones para cada combinación granja-animal, (15.4) podría escribirse así:

$$\begin{pmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \end{pmatrix} = \begin{array}{c} \begin{matrix} \vec{v}_0 & \vec{v}_1 & \vec{v}_2 & \vec{w}_1 & \vec{w}_2 & \vec{w}_3 & \vec{w}_4 \end{matrix} \\ \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{pmatrix} \alpha \\ \alpha_1^A \\ \alpha_2^A \\ \alpha_{11}^{A \leftarrow B} \\ \alpha_{12}^{A \leftarrow B} \\ \alpha_{21}^{A \leftarrow B} \\ \alpha_{22}^{A \leftarrow B} \end{pmatrix} + \begin{pmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{221} \\ \epsilon_{222} \end{pmatrix} \tag{15.5}$$

Es fácil desarrollar un cuadro de análisis de varianza similar al presentado para los modelos cruzados, por tanto lo proponemos como un ejercicio al final del capítulo.

## 15.5. Modelos de bloques aleatorizados.

En la Sección 13.3 mencionamos la necesidad de controlar variables externas *(factores de bloque)* que afectasen directamente a las unidades experimentales, y que por tanto fuese necesario controlar. Recordaremos que el efecto que los *factores de bloque* tengan sobre las unidades experimentales no se encuentra dentro de los objetivos del experimento que se lleva a cabo. En el Ejemplo 13.3, pág. 152, queríamos comparar dos tipos de semillas y habíamos citado como posible *factor de bloque* la parte de la finca en la que nos encontremos (alta o baja). El interés del agricultor estaba en la comparación del tipo de semilla y no en la comparación del tipo de terreno o parte de la finca en la que siembra las semillas, lo cual ilustra la idea que habíamos planteado.

La descripción sumaria que hacemos del modelo de bloques aleatorizados ilustra la manera en que ésto puede llevarse a cabo. Supongamos un tratamiento y un *factor de bloque* con las siguientes restricciones:

1. Hay razones fundadas para pensar en la existencia de un efecto claro del *factor de bloque* sobre las unidades experimentales.

2. Nos consta que no hay interacción alguna entre el *factor de bloque* y el tratamiento de interés en el experimento.

3. El tratamiento tiene $I$ posibles niveles, y el *factor de bloque* tiene $J$ posibles niveles.

4. Hay una observación por tratamiento y por bloque, y aleatoriamente se asignará el nivel del tratamiento a las unidades experimentales dentro de cada bloque.

El modelo planteado se puede expresar con la siguiente ecuación:

$$y_{ij} = \alpha + \alpha_i^A + \beta_j^B + \epsilon_{ij}, \tag{15.6}$$

donde $\beta_j^B,\ j = 1, ..., J$ representa el efecto del *factor de bloque*. El desarrollo es entonces paralelo al efectuado en las secciones anteriores, y por ello omitiremos detalles. Los cálculos necesarios se resumen en un estadillo como el que se presenta en la Tabla 15.2.

Trocóniz (1987a) contiene una cobertura amplia y referencias. Otras obras de interés son Montgomery (1984), y Box et al. (1978). En algunos casos no es posible utilizar todas las combinaciones de los tratamientos en cada uno de los *bloques*. Esto suele ocurrir debido a que no se tiene la infraestructura para realizar el experimento en estas condiciones o simplemente porque la misma naturaleza del experimento no permite realizarlo. Este tipo de experimentos pueden expresarse matemáticamente utilizando el llamado modelo de bloques aleatorizados incompletos, del cual tenemos una cobertura amplia tanto en Trocóniz (1987a) como en Montgomery (1984).

Cuadro 15.2: Análisis de Varianza. Bloques Aleatorizados.

| **Efecto** | **Suma de cuadrados** | **Grados de libertad** | **Estadístico de contraste** | **Dist. bajo $\mathbf{H_0}$: efecto nulo** |
|---|---|---|---|---|
| Efecto A (Tratamiento) | $J\sum_{i=1}^{I}(y_{i.}-y_{..})^2$ | $I-1$ | $\frac{J\sum_{i=1}^{I}(y_{i.}-y_{..})^2}{(I-1)\hat{\sigma}^2}$ | $\mathcal{F}_{I-1,IJ-I-J+1}$ |
| Efecto B (Factor de Bloque) | $I\sum_{j=1}^{J}(y_{.j}-y_{..})^2$ | $J-1$ | $\frac{I\sum_{j=1}^{J}(y_{.j}-y_{..})^2}{(J-1)\hat{\sigma}^2}$ | $\mathcal{F}_{J-1,IJ-I-J+1}$ |
| Media | $IJy_{..}^2$ | 1 | $IJy_{..}^2/\hat{\sigma}^2$ | $\mathcal{F}_{1,IJ-I-J+1}$ |
| Residuo | $\sum_{i,j}(y_{ij}-y_{i.}-y_{.j}+y_{..})^2$ | $(IJ-I-J+1)$ | $\hat{\sigma}^2$ | |
| Total | $\sum_{i,j}y_{ij}^2$ | $IJ$ | | |

## 15.6. Otros modelos.

Dentro de los capítulos correspondientes a Análisis de Varianza hemos querido introducir algunos de los modelos más frecuentemente utilizados, motivando su uso ya sea teóricamente como con ejemplos que ilustran ciertas situaciones prácticas que requieran de este tipo de modelos. Existen otros modelos que entran claramente en la categoría de ampliamente utilizados, pero cuya motivación y explicación va más allá de los objetivos iniciales de estos apuntes. Sin embargo, mencionaremos algunos de estos modelos y alguna de las referencias relevantes a las cuales podría recurrir un lector interesado en estos tópicos.

Una buena compilación de lo que son los modelos factoriales, modelos factoriales fraccionados y lo que significa el que los efectos estén *mezclados o confundidos* en este tipo de experimentos se puede encontrar en Trocóniz (1987a), Box et al. (1978) o en Montgomery (1984).

Todos los modelos vistos asumen una estructura de análisis de varianza con efectos fijos. Existen modelos de efectos aleatorios (también llamados de *componentes de la varianza*), en que los niveles de un tratamiento determinado pueden variar aleatoriamente dentro de una población de niveles. Referencias para este tipo de modelos podrían ser Trocóniz (1987a) y Scheffé (1959).

Si además de los tratamientos o efectos presentes en el experimento se tuviese variables explicativas de tipo cuantitativo, tales como las que teníamos en regresión, nuestro modelo entraría a formar parte de lo que llamamos *Análisis de la Covarianza*. Nuevamente tanto Trocóniz (1987a) como Scheffé (1959) contienen una cobertura amplia y referencias sobre este tema.

### CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

**15.1** En los modelos anidados de la Sección 15.4 construye el cuadro de análisis de varianza

**15.2** Un estudio intenta determinar si el volumen del sonido en las propagandas televisivas tiene algún efecto sobre el hecho de que la persona recuerde la propaganda y si además éste varía con el producto que se esté promocionando. Se escogen 16 personas, 1 para cada uno de los 16 grupos definidos de acuerdo a la edad (clase 1: muy jóvenes; 2; 3; 4: edad madura) y nivel de educación alcanzado (clase 1: EGB; 2; 3; 4: Post-graduado). Cada persona observó uno de los cuatro tipos de propagandas (A: volumen alto, producto X; B: volumen bajo, producto X; C: volumen alto, producto Y; D: volumen bajo, producto Y) de acuerdo con el diseño de cuadrados latinos que se muestra a continuación. Durante la siguiente semana se pidió a las personas que mencionasen todo lo que pudiesen recordar sobre la propaganda. Las calificaciones que se muestran en la tabla de datos están basadas en el número de detalles que recordaron sobre las propagandas, estandarizadas adecuadamente (véase Neter et al. (1985)).

| | **Nivel educativo** | | | |
|---|---|---|---|---|
| **Grupo de Edad** | **1** | **2** | **3** | **4** |
| **1** | 83(D) | 64(A) | 78(C) | 76(B) |
| **2** | 70(B) | 81(C) | 64(A) | 87(D) |
| **3** | 67(C) | 67(B) | 76(D) | 64(A) |
| **4** | 56(A) | 72(D) | 63(B) | 64(C) |

1. Construye la tabla ANOVA para este modelo.

2. Contrasta la hipótesis de que los cuatro grupos de edad tienen igual efecto sobre el hecho que la persona recuerde la propaganda.
3. Contrasta la hipótesis de que los cuatro niveles de educación tienen igual efecto sobre el hecho que la persona recuerde la propaganda.
4. Contrasta la hipótesis de que los cuatro tipos de propaganda tienen igual efecto sobre el hecho que la persona recuerde la propaganda.
5. ¿Hay alguna observación anómala (= con residuo studentizado muy grande)?
6. ¿Hay alguna observación notoriamente influyente? ¿Sobre algún parámetro en particular?

**15.3** Un investigador desea estudiar el efecto de tres dietas experimentales que se diferencian en su contenido total de grasas. El contenido total de grasas en la sangre suele ser utilizado para predecir enfermedades coronarias. Quince individuos, cuyo peso no sobrepasaba en más de un $20\,\%$ su peso ideal, fueron separados en 5 *bloques* de acuerdo a su edad (es decir, se está utilizando la edad de los individuos como *factor de bloque*), teniendo 3 individuos a cada grupo de edad. Dentro de cada *bloque*, las tres dietas experimentales fueron aleatoriamente asignadas a los tres sujetos. Después de un período determinado de tiempo, se obtuvieron mediciones de la reducción en el nivel de grasas (en gramos por litro) en la sangre. Los datos se muestran a continuación (véase Neter et al. (1985)).

| | **Contenido en grasas** | | |
|---|---|---|---|
| **Bloque** | **Muy Bajo** | **Bastante Bajo** | **Bajo** |
| Edad: 15 - 24 | 0.73 | 0.67 | 0.15 |
| Edad: 25 - 34 | 0.86 | 0.75 | 0.21 |
| Edad: 35 - 44 | 0.94 | 0.81 | 0.26 |
| Edad: 45 - 54 | 1.40 | 1.32 | 0.75 |
| Edad: 55 - 64 | 1.62 | 1.41 | 0.78 |

1. ¿Por qué piensas que la edad de los individuos ha sido utilizada como *factor de bloque*?
2. Construye la tabla ANOVA para este modelo.
3. Contrasta la hipótesis de que los tres tipos de dieta causan una reducción similar del contenido total de grasas.
4. Contrasta la hipótesis de que la edad de los individuos es realmente un *factor de bloque*.
5. ¿Hay alguna observación anómala (= con residuo studentizado muy grande)?
6. ¿Hay alguna observación notoriamente influyente? ¿Sobre algún parámetro en particular?

# Apéndice A

# Algunos resultados en Algebra Lineal.

**Teorema A.1** *El rango y la traza de una matriz idempotente coinciden.*

**Definición A.1** *En un espacio vectorial $V$ llamamos* producto interno *a una aplicación de $H \times H \longrightarrow R$ (si es real-valorado) o en $C$ (si es completo valorado), tal que a cada par de vectores $\vec{u}\,,\vec{v}$ corresponde $<\vec{u}\,,\vec{v}\,>$ verificando:*

$$<\vec{u}\,,\vec{v}\,>=\overline{<\vec{v}\,,\vec{u}\,>} \tag{A.1}$$

$$<\vec{u}\,,\vec{u}\,>\geq 0 \qquad \forall \vec{u}\in H \tag{A.2}$$

$$<\vec{u}\,,\vec{u}\,>=0 \Longrightarrow \vec{u}\,=0 \tag{A.3}$$

**Definición A.2** *Llamamos producto interno euclídeo de dos $n$-eplas $\vec{u}\,,\vec{v}$ en $R^n$ al definido así: $<\vec{u}\,,\vec{u}\,>=\vec{u}\,'\vec{v}$. Es fácil comprobar que verifica las condiciones en la Definición A.1. La* norma euclídea $||\vec{u}\,||$ *del vector $\vec{u}$ se define como $||\vec{u}\,|| = +\sqrt{<\vec{u}\,,\vec{u}\,>}$*

**Definición A.3** *Dados dos vectores $\vec{u}\,,\vec{v}$ en un espacio vectorial, definimos el coseno del ángulo que forman como*

$$\cos(\alpha)=\frac{<\vec{u}\,,\vec{v}\,>}{||\vec{u}\,||||\vec{v}\,||}. \tag{A.4}$$

**Teorema A.2** (Sherman-Morrison-Woodbury) *Sea $D$ una matriz simétrica $p\times p$ y $\vec{a}\,$,$\vec{c}$ vectores $p\times 1$. Entonces,*

$$(D+\vec{a}\,\vec{c}\,')^{-1}=D^{-1}-D^{-1}\vec{a}\,(1+\vec{c}\,'D^{-1}\vec{a}\,)^{-1}\vec{c}\,'D^{-1} \tag{A.5}$$

DEMOSTRACION:

Multiplicando ambos lados de (A.5) por $(D + \vec{a}\,\vec{c}\,')$ se llega a la igualdad $I = I$. En particular, si $\vec{a} = \vec{c} = \vec{z}$, la relación anterior produce:

$$(D + \vec{z}\vec{z}')^{-1} = D^{-1} - D^{-1}\vec{z}(1 + \vec{z}'D^{-1}\vec{z})^{-1}\vec{z}'D^{-1} \tag{A.6}$$

**Teorema A.3** *Si $A$ y $D$ son simétricas y todas las inversas existen:*

$$\begin{pmatrix} A & B \\ B' & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + FE^{-1}F' & -FE^{-1} \\ E^{-1}F' & E^{-1} \end{pmatrix} \tag{A.7}$$

*siendo*

$$E = D - B'A^{-1}B \tag{A.8}$$
$$F = A^{-1}B \tag{A.9}$$

DEMOSTRACION:

Basta efectuar la multiplicación matricial correspondiente.

Un caso particular de interés se presenta cuando la matriz particionada cuya inversa deseamos es del tipo:

$$\begin{pmatrix} (X'X) & X'Z \\ Z'X & Z'Z \end{pmatrix}$$

La aplicación de (A.7) proporciona entonces para el bloque superior izquierdo:

$$\begin{aligned} A^{-1} + FE^{-1}F' &= (X'X)^{-1} \\ &\quad + (X'X)^{-1}X'Z[Z'Z - Z'X(X'X)^{-1}X'Z]^{-1}Z'X(X'X)^{-1} \end{aligned} \tag{A.10}$$

y similarmente para los demás bloques. Véase Seber (1977), pág. 390 y Myers (1990), pág. 459.

# Apéndice B

# Algunos prerrequisitos estadísticos.

## B.1. Distribuciones $\chi^2$ y $\mathcal{F}$ descentradas

Sean $X_i \overset{\text{indep}}{\sim} N(\mu_i, \sigma^2)$, $(i = 1 \ldots, n)$. Sea $\delta^2 = (\mu_1^2 + \ldots + \mu_n^2)/\sigma^2$. Entonces, la variable aleatoria

$$Z = \frac{X_1^2 + \ldots + X_n^2}{\sigma^2} \tag{B.1}$$

se dice que sigue una distribución $\chi_n^2(\delta)$, o distribución $\chi^2$ *descentrada* con *parámetro de no centralidad* $\delta$ y $n$ grados de libertad. Algunos textos definen $\delta^2$ o $\frac{1}{2}\delta^2$ como parámetro de no centralidad; la notación que empleamos es congruente con las Tablas en **??** . Claramente, si $\delta = 0$ se tiene la $\chi^2$ habitual o *centrada*.

Si $Z \sim \chi_m^2(\delta)$ y $V \sim \chi_n^2$ son ambas independientes, la variable aleatoria

$$W = \frac{n}{m}\frac{Z}{V} \tag{B.2}$$

sigue una distribución $\mathcal{F}_{m,n}(\delta)$ o $\mathcal{F}$ de Snedecor descentrada, con parámetro de no centralidad $\delta$. Si $V$ siguiera una distribución $\chi_n^2(\gamma)$, tendríamos que $W$ sería una $\mathcal{F}$ de Snedecor doblemente descentrada, habitualmente denotada como $\mathcal{F}_{m,n}(\delta, \gamma)$. Siempre nos referiremos al primer tipo, en que solo el numerador es descentrado.

La $\mathcal{F}$ de Snedecor descentrada es una distribución definida en el semieje real positivo, cuya forma es similar a la de su homóloga centrada. Su moda está tanto mas desplazada a la derecha cuanto mayor sea el parámetro de no centralidad. El examen del estadístico de contraste $Q_h$ introducido en la Sección 12 hace evidente que cuando

la hipótesis contrastada no es cierta, la distribución de $Q_h$ es descentrada. Ello permite, como ya se indicó, calcular con facilidad la potencia de cualquier contraste, si se dispone de tablas de la $\mathcal{F}_{m,n}(\delta)$. El apéndice A.4 proporciona tablas que permiten calcular la potencia de los contrastes en análisis de varianza directamente, prefijada una alternativa.

## B.2. Estimación máximo verosímil

Se realiza maximizando la función de verosimilitud $L(\vec{\beta}\,,\vec{y}\,)$ o, equivalentemente, su logaritmo, $\ell(\vec{\beta}\,,\vec{y}\,)$. Sea $\hat{\beta}$ el vector que maximiza $\ell(\vec{\beta}\,,\vec{y}\,)$. En condiciones muy generales, se tiene que para muestras grandes

$$\hat{\beta} \quad \overset{asint}{\sim} \quad N(\vec{\beta}\,,\Sigma_{\hat{\beta}}) \tag{B.3}$$

$$\Sigma_{\hat{\beta}} \quad \approx \quad \Big[I(\hat{\beta})\Big]^{-1} \tag{B.4}$$

En la expresión anterior, $I(\hat{\beta})$ es la llamada *matriz de información* cuyo elemento genérico de lugar $ij$ se define así:

$$\Big[I(\hat{\beta})\Big]_{ij} \quad = \quad -\frac{\partial^2 \ell(\vec{\beta}\,,\vec{y}\,)}{\partial\beta_i \partial\beta_j}. \tag{B.5}$$

Una consecuencia de (B.3)–(B.4) es que si $\Sigma_{\hat{\beta}}$ es de dimensión $p \times p$,

$$(\hat{\beta}-\vec{\beta}\,)'(\Sigma_{\hat{\beta}})^{-1}(\hat{\beta}-\vec{\beta}\,) \sim (\hat{\beta}-\vec{\beta}\,)' I(\hat{\beta})(\hat{\beta}-\vec{\beta}\,) \sim \chi^2_p;$$

esto permite contrastar hipótesis como $H_0 : \vec{\beta} = \vec{\beta}_{\,0}$ utilizando como estadístico

$$(\hat{\beta}-\vec{\beta}_{\,0})' I(\vec{\beta}_{\,0})(\hat{\beta}-\vec{\beta}_{\,0}) \tag{B.6}$$

o alternativamente

$$(\hat{\beta}-\vec{\beta}_{\,0})' I(\hat{\beta})(\hat{\beta}-\vec{\beta}_{\,0}). \tag{B.7}$$

Asintóticamente ambos contrastes son equivalentes, y ambos se conocen como *contrastes de Wald*; pueden consultarse más detalles en Lehmann (1983), Cap. 6 o Garthwaite et al. (1995), Cap. 3 y 4.

## B.3. Contraste razón generalizada de verosimilitudes

Supongamos una hipótesis nula $H_0$ que prescribe para el vector de parámetros un subespacio $h$. Supongamos $h$ es un subespacio de $M$, y $\dim(h) = q < p = \dim(H)$. Supongamos, finalmente, que $L(\vec{\beta}\,,\vec{Y}\,)$ es la función de verosimilitud y

$$\hat{\beta}_h \quad = \quad \arg\max_{\vec{\beta}\,\in h} L(\vec{\beta}\,,\vec{Y}\,) \tag{B.8}$$

$$\hat{\beta}_M \quad = \quad \arg\max_{\vec{\beta}\,\in M} L(\vec{\beta}\,,\vec{Y}\,). \tag{B.9}$$

Entonces, en condiciones muy generales, que no requieren que $\vec{Y}$ siga una distribución particular, se verifica que bajo $H_0$,

$$-2\log_e\left(\frac{L(\hat{\beta}_h,\vec{Y})}{L(\hat{\beta}_M,\vec{Y})}\right)\sim\chi^2_{(p-q)}. \tag{B.10}$$

Por lo tanto, un contraste de la hipótesis $H_0$ puede obtenerse comparando el estadístico en el lado izquierdo de (B.10) con el cuantil $\chi^2_{(p-q);\alpha}$; valores del estadístico mayores que dicho cualtil conducirán al rechazo de la hipótesis nula.

# Apéndice C

# Regresión en S-PLUS y R.

## C.1. El sistema estadístico y gráfico S-PLUS

El lenguaje y sistema estadístico S fue desarrollado en ATT a principios de los ochenta. Es una síntesis afortunada de simplicidad, sintaxis consistente, flexibilidad, e integración con el sistema operativo UNIX, sobre el que se desarrolló y para el que fue principalmente desarrollado.

Incorpora conceptos y ventajas de muchos lenguajes. El manejo de vectores y matrices, y la facilidad para definirlos, empalmarlos, y operar con ellos recuerda al lenguaje `APL`. El uso de listas es reminiscente de `LISP`. La sintaxis, el convenio de paso de argumentos por valor, y la forma de definir funciones son similares a los que existen en `C`. Sobre todo ello, S añade un conjunto bastante rico de funciones primitivas que hace fácil programar casi cualquier procedimiento. Las facilidades gráficas son también excelentes.

La referencia fundamental para utilizar S es Becker et al. (1988). Hay una versión comercial de S (S-PLUS, de Insightful, Inc.) que es un super-conjunto del S-PLUS descrito en Becker et al. (1988); para ella existen manuales específicos. Las funciones más modernas —entre ellas, algunas de interés para análisis de regresión— están descritas en Chambers and Hastie (1992).

Además de los manuales, S-PLUS permite acceder "on line" a la descripción de cualquier función. Basta teclear en una sesión `help(xxxxx)`, en que `xxxxx` es el nombre de la función.

## C.2. El sistema estadístico y gráfico R

R comenzó siendo un paquete estadístico "no muy diferente" de S, cuya funcionalidad pretendía replicar manteniendo una filosofía de código fuente disponible. Puede verse una descripción en Ihaka and Gentleman (1996). Adicionalmente puede consul-

tarse Venables et al. (1997) (traducción castellana Venables et al. (2000)), o el manual Venables and Ripley (1999a) y sus complementos Venables and Ripley (1999b).

En la actualidad continúa manteniendo una buena compatibilidad aunque con diferencias sustanciales en su arquitectura (que por lo general sólo precisa conocer el usuario avanzado). No replica toda la funcionalidad de S-PLUS en algunos aspectos, pero la amplia en otros. Esta siendo muy activamente desarrollado por la comunidad universitaria e investigadora internacional. Su fácil extensibilidad y disponibilidad gratuita hace que sea el paquete en que primero se implementan métodos que tardan en encontrar hueco en los paquetes comerciales.

En `http://cran.r-project.org/` o sus espejos en los cinco continentes pueden encontrarse las versiones más recientes para multitud de sistemas operativos, las fuentes y los añadidos que la comunidad de usuarios ha ido contribuyendo.

Las secciones siguientes describen algunas funciones específicas para análisis de regresión. Dado que pueden producirse modificaciones de una versión a otra, la información autorizada y definitiva debe buscarse en los manuales de S-PLUS. Las mismas funciones están disponibles en R, con funcionalidad equivalente pero posibles ligeras diferencias en los argumentos y resultados. De nuevo la consulta de los manuales o ayuda "on line" es obligada para contrastar lo que sigue.

Finalmente, en la Sección C.3 se presenta una tabla recogiendo la correspondencia entre algunas funciones similares de S-PLUS y R.

### C.2.1. La función `lsfit`.

Es el principal bloque constructivo de cualquier procedimiento de regresión. Ajusta una regresión (opcionalmente ponderada) y devuelve una lista con los coeficientes estimados, los residuos, y otra variada información de interés. La sintaxis es la siguiente:

```
lsfit(x, y, wt=<<ver texto>>, intercept=T, tolerance=1.e-07,
      yname=NULL)
```

**Argumentos.** Los argumentos obligatorios son los siguientes:

`x` Vector o matriz de regresores. **No** es preciso incluír una columna de "unos": se incluye automáticamente a menos que especifiquemos `intercept=F`. Ha de tener tantas filas como el argumento `y`. Puede tener valores perdidos. `x` puede ser un vector cuando estamos regresando solo sobre una variable.

`y` Variable respuesta. Es un vector, o una matriz. Si se trata de una matriz, se regresa *cada una de sus columnas* sobre los regresores en `x`. De esta manera, una sola invocación de `lsfit` puede realizar un gran número de regresiones, cuando los regresores son comunes a todas ellas. Tambien se permiten valores perdidos.

Los restantes argumentos son optativos. Si no se especifican, se supone que sus valores son los que aparecen en el ejemplo de sintaxis más arriba. Sus significados son los siguientes:

| | |
|---|---|
| `wt` | Vector de ponderaciones, si se quiere realizar regresión ponderada. Ha de tener la misma longitud que `y`. Salvo que se especifique, la regresión pondera igualmente todas las observaciones. |
| `intercept` | Si es `T`, se incluye una columna de "unos". Si no deseamos columna de "unos", es preciso especificar `intercept=F`. |
| `tolerance` | Valor numérico para especificar cuando consideramos una matriz singular. |
| `yname` | Nombre de la variable `y` en la regresión. |

**Resultados.** La función `lsfit` devuelve una lista con los siguientes componentes:

| | |
|---|---|
| `coef` | Vector $\hat{\beta}$ de estimadores, en forma de matriz con una columna para cada regresión, si se han hecho varias a la vez. |
| `residuals` | Vector (o matriz, si `y` era una matriz) conteniendo los residuos ordinarios $\hat{\epsilon}$. |
| `wt` | Si especificamos ponderaciones, nos son devueltas inalteradas. Esto es útil si guardamos la lista de resultados, pues permite con posterioridad saber a qué tipo de regresión corresponden. |
| `intercept` | Valor lógico, `T` ó `F`. |
| `qr` | Objeto representando la descomposición QR de la matriz `x` de regresores. Véase la función `qr` en Becker et al. (1988). Tiene utilidad para computar algunos resultados. |

### C.2.2. La función `leaps`.

La función `leaps` realiza *all-subsets* regresión. No debe invocarse con un número excesivo de regresores, al crecer el esfuerzo de cálculo exponencialmente con éste.

La sintaxis es:

```
leaps(x, y, wt, int=TRUE, method=''Cp'', nbest=10, names, df=nrow(x))
```

**Argumentos.** Los argumentos `x`, `y`, `wt` tienen el mismo significado que en la función `lsfit`. El argumento `int` se utiliza para indicar si se desea incluír columna de "unos" (por omisión, sí). Los demás argumentos tienen los siguientes significados:

| | |
|---|---|
| `method` | Argumento alfanumérico (entre dobles comillas, por tanto) especificando el criterio que se desea emplear en la selección de las mejores regresiones. Puede ser "Cp" ($C_p$ de Mallows, el valor por omisión), "r2" (el $R^2$), y "adjr2" (valor $\overline{R}^2$). |
| `nbest` | Número de regresiones que deseamos para cada tamaño de modelo. |
| `names` | Vector de nombres de los regresores. |
| `df` | Grados de libertad de `y` (puede no coincidir con el número de filas si ha sido previamente objeto de alguna manipulación. Un caso frecuente en Economía es la desestacionalización, que consume grados de libertad. |

**Resultados.** Retorna una lista con cuatro elementos:

| | |
|---|---|
| `Cp` | Criterio de ajuste especificado como argumento. |
| `size` | Número de regresores (incluyendo, en su caso, la columna de "unos". |
| `label` | Vector de nombres de los regresores. |
| `which` | Matriz lógica. Tiene tantas filas como subconjuntos de regresores devueltos, y la fila $i$-ésima tiene valores `T` ó `F` según el regresor correspondiente haya sido o no seleccionado en el $i$-ésimo subconjunto. |

### C.2.3. La función `hat`.

Se invoca así:

```
hat(x, int=TRUE)
```

en que `x` es argumento obligatorio y es la matriz de regresores. El argumento `int` toma el valor `T` por omisión y señala si se desea incluír en la matrix `x` columna de "unos".

La función devuelve un vector con los elementos diagonales de la matriz de proyección $X(X'X)^{-1}X'$ (los $p_{ii}$ del Capítulo 9).

### C.2.4. Data frames.

S-PLUS tiene funciones que hacen aún más simple y cómodo el análisis de regresión. Permiten además, con una sintaxis común, ajustar modelos muy diversos, lineales y no lineales, paramétricos y no paramétricos o semiparamétricos.

Tales funciones admiten como entrada "data frames" (a las que en lo sucesivo nos referiremos como "tablas de datos". Podemos pensar en ellas como en matrices cuyos elementos no tienen porqué ser todos del mismo tipo. Por ejemplo, cabe que una columna sea numérica y otra tenga como datos literales. Esto simplifica mucho las cosas. Podríamos utilizar como tabla de datos,

$$\begin{array}{ccc} 2.3 & 4.3 & \text{Viejo} \\ 3.2 & 1.2 & \text{Viejo} \\ \vdots & \vdots & \vdots \\ 4.1 & 2.3 & \text{Joven} \end{array}$$

y la variable en la tercera columna (cualitativa) sería convertida de manera automática en columnas de "unos" y "ceros". No habríamos de preocuparnos de hacerlo nosotros y, por añadidura, si la columna tercera tuviera $n$ categorías y existiera ya una columna de "unos" en el modelo, se generarían automáticamente $(n-1)$ columnas en evitación de una multicolinealidad exacta.

El output, finalmente, quedaría rotulado adecuadamente.

Una tabla de datos puede, como una matriz, tener valores perdidos `NA` y dimensiones nombradas. En general, no habremos de preocuparnos de hacerlo. Cuando se tiene un fichero `ASCII`, la función `read.table` permite leerla y dar nombres a filas y columnas de modo muy cómodo: véase Chambers and Hastie (1992), pág. 567.

### C.2.5. La función `lm`.

La función `lm` ajusta un modelo lineal. La sintaxis es:

```
lm(formula,data,weights,subset,na.action,method="qr",model=F,x=F,y=F,...)
```

**Argumentos.** El argumento `weights` se utiliza para hacer regresión ponderada, de modo similar a como se hace con `lsfit`. Los demás argumentos tienen los siguientes significados:

| | |
|---|---|
| `method` | Método de ajuste a emplear. Por omisión, se utiliza la factorización QR. |
| `data` | Una "data frame" conteniendo los datos tanto de regresores como de variable respuesta. |
| `formula` | Una expresión del tipo `Resp` $\sim$ `Regr01 + Regre02 + log(Regre03)` en que a la izquierda está el regresando y a la derecha los regresores o funciones de ellos. |
| `subset` | Criterio para seleccionar las filas de la tabla de datos que deseamos emplear. |
| `na.action` | Acción a tomar cuando algún dato en una fila de la tabla de datos es `NA`. Por omisión es omitir dicha fila. |
| `model,x,y` | Seleccionando estos argumentos como `T` se obtienen como resultado. |

**Resultados.** Retorna un objeto de tipo `lm.object`, una estructura de datos compuesta que contiene los resultados del ajuste. Hay funciones especializadas en extraer los resultados y presentarlos de modo ordenado. Por ejemplo, `summary()`, `residuals()`,

`coefficients()` o `effects()`. Por otra parte, el carácter objeto-orientado de S-PLUS (una descripción de esto referida a XLISP-STAT en la Sección **??**) hace que funciones como `print()` aplicadas a un objeto de tipo `lm.object` "sepan" como imprimirlo.

Debe invocarse tras `lm` y `ls` y sobre los objetos que éstas devuelven.

### C.2.6. La función `lm.influence`.

La sintaxis es:

```
lm.influence(ajuste)
```

**Argumentos.** `ajuste` es un objeto de tipo `lm.object` devuelto por `lm`.

**Resultados.** La función `lm.influence` devuelve (salvo una constante) los coeficientes de la curva de influencia muestral (SIC).

### C.2.7. La función `ls.diag`.

La sintaxis es:

```
ls.diag(ls)
```

**Argumentos.** La función `ls.diag` se invoca con un objeto de tipo `ls` (devuelto por `lsfit`) por argumento.

**Resultados.** Produce como resultado una lista con los componentes siguientes:

| | |
|---|---|
| `std.dev` | $= \sigma = \sqrt{\frac{SSE}{N-p}}$. |
| `hat` | Los $p_{ii}$, elementos diagonales de la matriz de proyección $P = X((X'X))^{-1}X'$. |
| `std.res` | Residuos internamente studentizados (los $r_i$ en la notación del Capítulo 9). |
| `stud.res` | Residuos externamente studentizados (los $t_i$ en la notación del Capítulo 9). |
| `cooks` | Un vector conteniendo las distancias de Cook ($D_i$ en la notación del Capítulo 9). |
| `dfits` | Un vector conteniendo los DFITS mencionados en el Capítulo 9). |
| `correlation` | Matriz de correlación de los parámetros estimados (es decir, la matriz de correlación obtenida de la de covarianzas $\hat{\sigma}^2((X'X))^{-1}$). |
| `std.err` | Desviaciones típicas estimadas de los parámetros estimados, $\hat{\sigma}_{\hat{\beta}_i}$. |
| `cov.unscaled` | Matriz de momentos $((X'X))^{-1}$. |

## C.3. Correspondencia de funciones para regresión y ANOVA en S-PLUS y R

Cuadro C.1: Equivalencia de funciones para regresión y ANOVA en S-PLUS y R.

| En S-PLUS | En R | Paquete: | Funcionalidad: |
|---|---|---|---|
| `add1` | `add1` | `base` | Añadir un regresor |
| `drop1` | `drop1` | `base` | Eliminar un regresor |
| `leaps` | `leaps` | `leaps` | Regresión sobre todos los subconjuntos |
| `ls.diag` | `ls.diag` | `base` | Diagnósticos |
| `lsfit` | `lsfit` | `base` | Ajuste recta regresión |
| `lm` | `lm` | `base` | Ajuste recta de regresión |
| `lm.influence` | `lm.influence` | `base` | Análisis de influencia |
| `multicomp` | - | - | Inferencia simultánea |
| - | `regsubsets` | `leaps` | Regresión sobre todos los subconjuntos |
| `step` | `step` | `base` | Regresión escalonada |
| `stepwise` | - | - | Regresión escalonada |
| - | `stepAIC` | `MASS` | Regresión escalonada |
| - | `p.adjust` | `base` | Ajuste $p$ por simultaneidad |
| - | `pairwise.t.test` | `ctest` | Contrastes más usuales |
| - | `lm.ridge` | `MASS` | Regresión *ridge* |

Además de las indicadas en la Tabla C.1, en R se dispone del paquete `multcomp` con varias funciones específicas para inferencia simultánea.

# Apéndice D

# Procedimientos de cálculo.

## D.1. Introducción

La resolución de las ecuaciones normales,

$$(X'X)\vec{\beta} = X'\vec{Y}$$

requiere, en su aproximación más directa, la obtención de la inversa (ordinaria o generalizada) de $(X'X)$. Hay procedimientos mucho menos costosos desde el punto de vista del cálculo que, además, permiten en algunos casos intuiciones interesantes y demostraciones de gran simplicidad.

En lo que sigue se presenta uno de los métodos de cálculo más utilizados, y la construcción en que se basa (la *factorización QR*). Se detalla también la correspondencia entre la notación empleada y los resultados de algunas funciones de S que hacen uso de dicha factorización.

## D.2. Transformaciones ortogonales.

Sea el problema,

$$\min_{\vec{x}} ||D\vec{x} - \vec{c}||^2 \tag{D.1}$$

Podemos ver el problema como el de encontrar la combinación lineal de las columnas de $D$ que mejor aproxima $\vec{c}$, en términos de norma de la discrepancia. Dicho problema queda inalterado cuando realizamos una misma transformación ortogonal de las columnas de $D$ y del vector $\vec{c}$. En efecto,

$$\begin{aligned}
\min_{\vec{x}} ||Q(D\vec{x} - \vec{c})||^2 &= \min_{\vec{x}} < Q(D\vec{x} - \vec{c}), Q(D\vec{x} - \vec{c}) > \\
&= \min_{\vec{x}} (D\vec{x} - \vec{c})'Q'Q(D\vec{x} - \vec{c}) \\
&= \min_{\vec{x}} ||D\vec{x} - \vec{c}||^2
\end{aligned}$$

al ser $Q$ ortogonal.

**Definición D.1** *Sea $D$ una matriz de orden $n \times m$. Supongamos que puede expresarse del siguiente modo:*

$$D = HRK'$$

*en que:*

*(i) $H$ es $n \times n$ y ortogonal.*

*(ii) $R$ es $n \times m$ de la forma,*

$$\begin{pmatrix} R_{11} & 0 \\ 0 & 0 \end{pmatrix}$$

*con $R_{11}$ cuadrada de rango completo $k \leq \min(m, n)$.*

*(iii) $K$ es $m \times m$ ortogonal.*

*Se dice que $HRK'$ es una descomposición ortogonal de $D$.*

En general, hay más de una descomposición ortogonal, dependiendo de la estructura que quiera imponerse a $R$. Si requerimos que $R$ sea diagonal, tenemos la *descomposición en valores singulares*. Podemos también requerir que $R$ sea triangular superior, o triangular inferior, obteniendo diferentes descomposiciones de $D$.

La elección de una descomposición ortogonal adecuada simplifica enormemente la solución de (D.1). Los resultados fundamentales vienen recogidos en el siguiente teorema.

**Teorema D.1** *Sea $D$ una matriz de orden $n \times m$ y rango $k$, admitiendo la descomposición ortogonal,*

$$D \quad = \quad HRK'. \tag{D.2}$$

*Sea el problema*

$$\min_{\vec{x}} \; ||D\vec{x} - \vec{y}||^2 \tag{D.3}$$

*y definamos,*

$$\begin{aligned} H'\vec{y} &= \vec{g} = \begin{pmatrix} \vec{g}_1 \\ \vec{g}_2 \end{pmatrix} \begin{matrix} k \\ n-k \end{matrix} \\ K'\vec{x} &= \vec{\gamma} = \begin{pmatrix} \vec{\gamma}_1 \\ \vec{\gamma}_2 \end{pmatrix} \begin{matrix} k \\ m-k \end{matrix}. \end{aligned}$$

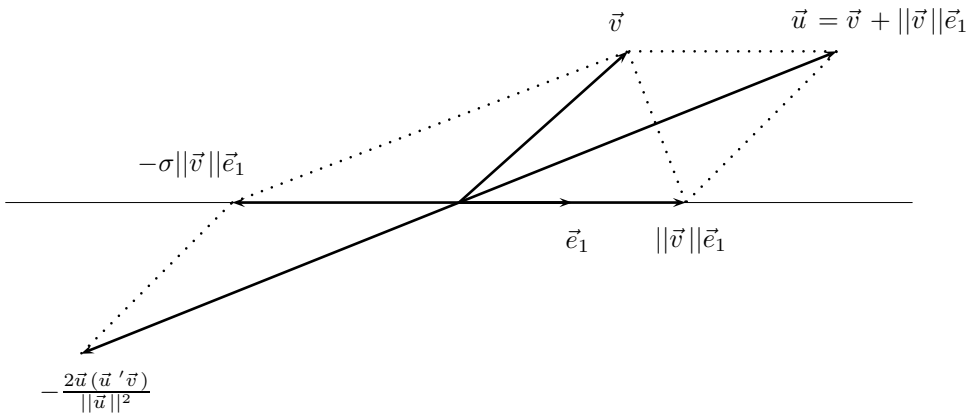
*Sea $\tilde{\gamma}_1$ la solución (única) del sistema,*

$$R_{11}\tilde{\gamma}_1 \quad = \quad \vec{g}_1.$$

*Entonces, todas las posibles soluciones del problema (D.3) son de la forma*

$$\vec{x} \quad = \quad K \begin{pmatrix} \tilde{\gamma}_1 \\ \vec{\gamma}_2 \end{pmatrix},$$

Figura D.1: Visualización de la transformación de Householder.




*con $\gamma_2$ arbitrario. Cualquiera de esas soluciones da lugar al vector de residuos*

$$\vec{r} \quad = \quad \vec{y} - D\vec{x} = H \begin{pmatrix} \vec{0} \\ \vec{g}_2 \end{pmatrix}$$

*y en consecuencia, $||\vec{r}|| = ||\vec{g}_2||$.*

Existe un resultado interesante que muestra cómo es posible encontrar una transformación ortogonal que rota (y quizá refleja) un vector $\vec{v}$ hasta abatirlo sobre el subespacio generado por otro, $\vec{e}_1$ . Se denomina *transformación de Householder*, y se obtiene de manera muy cómoda y simple como muestra el teorema siguiente.

**Teorema D.2** *Sea $\vec{v}$ cualquier vector $m \times 1$ distinto de $\vec{0}$. Existe una matriz ortogonal $P$ $m \times m$ tal que:*

$$P\vec{v} \quad = \quad -\sigma||\vec{v}||\vec{e}_1 \tag{D.4}$$

*siendo*

$$\vec{e}_1 \quad = \quad \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{D.5}$$

$$\sigma \quad = \quad \begin{cases} +1 & \textit{si } v_1 \geq 0 \\ -1 & \textit{si } v_1 < 0. \end{cases} \tag{D.6}$$

*Esta matriz tiene por expresión,*

$$P \quad = \quad I - 2\frac{\vec{u}\,\vec{u}\,'}{||\vec{u}\,||^2} \tag{D.7}$$

*con $\vec{u} = \vec{v} + \sigma||\vec{v}||\vec{e}_1$ .*

DEMOSTRACION:

Entonces (ver Figura D.1),

$$\vec{u} = \vec{v} + \sigma||\vec{v}||\vec{e}_1 \quad \text{(D.8)}$$
$$\vec{z} = \vec{v} - \sigma||\vec{v}||\vec{e}_1 \quad \text{(D.9)}$$

son ortogonales y $\vec{v} = \frac{1}{2}\vec{u} + \frac{1}{2}\vec{z}$. Tenemos en consecuencia,

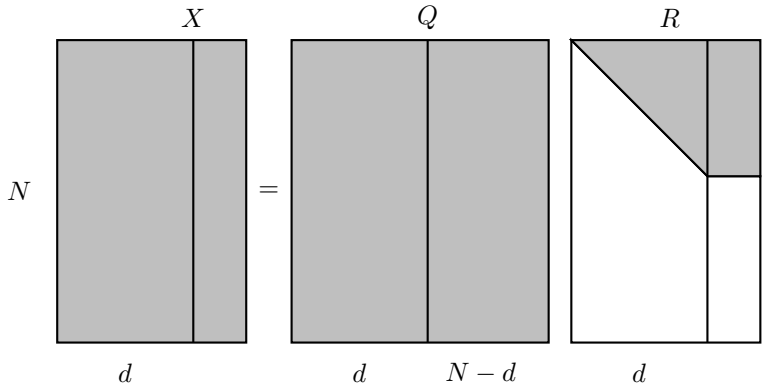
$$P\vec{v} = \left(I - 2\frac{\vec{u}\,\vec{u}\,'}{||\vec{u}\,||^2}\right)\left(\frac{1}{2}\vec{u} + \frac{1}{2}\vec{z}\right) \quad \text{(D.10)}$$
$$= \frac{1}{2}\vec{u} - \vec{u} + \frac{1}{2}\vec{z} \quad \text{(D.11)}$$
$$= -\frac{1}{2}\vec{u} + \vec{v} - \frac{1}{2}\vec{u} \quad \text{(D.12)}$$
$$= \vec{v} - \vec{u} \quad \text{(D.13)}$$
$$= -\sigma||\vec{v}||\vec{e}_1 \quad \text{(D.14)}$$

## D.3. Factorización QR.

**Teorema D.3** *Sea una matriz $X$ de orden $(N \times p)$ y rango $d \leq \min(N, p)$. Existe siempre una matriz ortogonal $Q$ de orden $(N \times N)$ y una matriz $R$ trapezoidal superior verificando:*

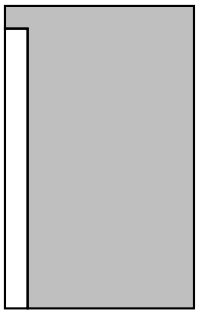
$$X = QR \quad \text{(D.15)}$$

*Esquemáticamente,*




DEMOSTRACION:

La prueba es constructiva, y reposa en la aplicación reiterada de la transformación de Householder a las columna de la matriz $X$. Sea $\vec{x}_1$ la primera de dichas columnas. Existe una transformación de Householder, de matriz ortogonal $P_1$ que abate dicha primera columna sobre el $\vec{e}_1$ de la base canónica de $R^n$. Es decir,

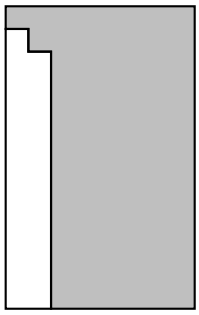
$P_1 X =$ 

Llamemos $X_1$ a la matriz así obtenida, y consideremos su segunda columna eliminado su primer elemento. Los restantes, pueden verse como un vector en $R^{N-1}$, que puede tambien abatirse sobre el primer vector $\vec{e}_1$ de la base canónica de dicho subespacio multiplicando por una matriz de Householder $P_2^*$. Entonces,

$$\begin{pmatrix} 1 & \vec{0}\,' \\ \vec{0} & P_2^* \end{pmatrix} P_1 \tag{D.16}$$

reduce la matriz $X$ de la forma que esquemáticamente se muestra a continuación:

$$\begin{pmatrix} 1 & \vec{0}\,' \\ \vec{0} & P_2^* \end{pmatrix} P_1 X = \quad$$



Por consiguiente, si llamamos

$$P_2 = \begin{pmatrix} 1 & \vec{0}\,' \\ \vec{0} & P_2^* \end{pmatrix}$$

el producto $P_2 P_1$ reduce las dos primeras columnas de $X$ a forma escalonada. Como tanto $P_1$ como $P_2$ son ortogonales, su producto también lo es. Fácilmente se comprueba que el proceso puede continuarse hasta obtener un producto de matrices ortogonales $Q' = P_d P_{d-1} \ldots P_1$ que deja $X$ con sus $d$ primeras columnas "escalonadas". Además, como el rango de $X$ era $d$, necesariamente las últimas $N - d$ filas de $R$ son de ceros.

En definitiva, $Q'X = R$ y por tanto $X = QR$, lo que prueba el teorema.

## D.4. Bibliografía

Hay abundante literatura sobre la factorización QR y procedimientos similares de aplicación al problema (D.1). Casi cualquier texto de Cálculo Numérico contiene una discusión de la factorización QR. Una referencia fundamental que continúa vigente es Lawson and Hanson (1974). Una exposición breve, clara, y con abundantes referencias a la literatura más reciente puede encontrarse en Goodhall (1993). Ansley (1985) muestra como, al margen y además de su utilidad como procedimiento numérico, la factorización QR arroja luz sobre, y simplifica la demostración de, bastantes resultados en regresión lineal.

# Bibliografía


Akaike, H. (1972). Use of an Information Theoretic Quantity for Statistical Model Identification. In *Proc. 5th. Hawai Int. Conf. on System Sciences*, pp. 249–250.

Akaike, H. (1974). Information Theroy and an Extension of the Maximum Likelihood Principle. In B. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pp. 267–281, Budapest: Akademia Kiado.

Akaike, H. (1991). Information Theory and an Extension of the Maximum Likelihood Principle. In Johnson and Kotz, editors, *Breakthroughs in Statistics*, volume 1, p. 610 y ss., Springer Verlag.

Anderson, T. (1971). *The Statistical Analysis of Time Series*. New York: Wiley.

Ansley, C. (1985). Quick Proofs of Some Regression Theorems via the QR Algorithm. *Appl. Statist*, 39, 55–59.

Arnold, S. (1981). *The Theory of Linear Models and Multivariate Analysis*. New York: Wiley.

Atkinson, A. (1985). *Plots, Transformations and Regression*. Oxford Univ. Press.

Barnett, V. and Lewis, T. (1978). *Outliers in Statistical Data*. New York: Wiley.

Becker, R., Chambers, J., and Wilks, A. (1988). *The New S Language. A Programming Environment for Data Analysis and Graphics*. Pacific Grove, California: Wadsworth & Brooks/Cole.

Belsley, D., Kuh, E., and Welsch., R. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Ben-Israel, A. and Greville, T. (1974). *Generalized Inverses: Theory and Aplications*. New York: Wiley.

Beyer, W. (1968). *Handbook of Tables for Probability and Statistics*. Cleveland, Ohio: The American Rubber Co, second edition.

Bishop, C. (1996). *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Box, G., Hunter, W., and Hunter, J. (1978). *Statistics for Experimenters*. New York: Wiley.

Box, G. and Tidwell, P. (1962). Transformations of the independent variables. *Technometrics*, 4, 531–550.

Brown, P. (1993). *Measurement, Regression and Calibration*. Clarendon Press/Oxford, Signatura: 519.235.5 BRO.

Chambers, J. and Hastie, T. (1992). *Statistical Models in S*. Pacific Grove, Ca.: Wadsworth & Brooks/Cole.

Cook, R. and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall, 1979th edition.

Cox, D. R. and Hinkley, D. V. (1978). *Problems and Solutions in Theoretical Statistics*. London: Chapman & Hall.

Dahlquist, G. and Björck, Å. (1974). *Numerical Methods*. Englewood Cliffs, N.J.: Prentice Hall.

de Leeuw, J. (2000). Information Theroy and an Extension of the Maximum Likelihood Principle by Hirotugu Akaike. Disponible en `http://www.stat.ucla.edu/~deleeuw/work/research.phtml`.

Dey, A. (1985). *Orthogonal Fractional Factorial Designs*. Nueva Delhi: Wiley Eastern Ltd.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker.

Fisher, L. and McDonald, J. (1978). *Fixed Effects Analysis of Variance*. New York: Academic Press.

Garthwaite, P., Jolliffe, I., and Jones, B. (1995). *Statistical Inference*. London: Prentice Hall.

Goodhall, C. (1993). Computation Using the QR Decomposition. In C. Rao, editor, *Handbook of Statistics*, chapter 13, pp. 467–508, Amsterdam: North-Holland.

Grafe, J. (1985). *Matemáticas Universitarias*. Madrid: MacGraw-Hill.

Haitovsky, Y. (1969). A note on maximization of $\overline{R}^2$. *Appl. Statist*, 23, 20–21.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer-Verlag, Signatura: 519.237.8 HAS.

Hawkins, D. (1980). *Identification of Outliers*. London: Chapman & Hall.

Haykin, S. (1998). *Neural Networks. A comprehensive Foundation*. Prentice Hall, second edition.

Hocking, R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, 32, 1–49.

Hoerl, A. and Kennard, R. (1970). Ridge Regression: Biased Estimation for Non-orthogonal Problems. *Technometrics*, 12, 55–67.

Hoerl, A., Kennard, R., and Baldwin, K. (1975). Ridge regression: some simulations. *Communications in Statistics*, 4, 105–123.

Hosmer, D. and Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *J. of Comp. and Graphical Stats.*, 5, 299–314.

Jolliffe, I. (1986). *Principal Components Analysis*. New York: Springer-Verlag.

Kennedy, W. (1980). *Statistical Computing*. New York: Marcel Dekker.

Kleinbaum, D. (1994). *Logistic Regression. A Self-Learning Test*. Springer Verlag.

Knuth, D. (1968). Fundamental Algorithms. In *The Art of Computer Programming*, volume 1, Reading, Mass.: Addison-Wesley.

Knuth, D. (1986). *The TEXbook*. Reading, Mass.: Addison Wesley.

Lawless, J. and Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics*, 5, 307–323.

Lawson, C. L. and Hanson, R. J. (1974). *Solving Least Squares Problems*. Englewood Cliffs, N.J.: Prentice-Hall.

Legg, S. (1996). Minimum Information Estimation of Linear Regression Models. In D. Dowe, K. Korb, and J. Oliver, editors, *ISIS: Information, Statistics and Induction in Science*, pp. 103–111, Singapore: World Scientific.

Lehmann, E. L. (1983). *Theory of Point Estimation*. New York: Wiley.

Lund, R. (1975). Tables for the approximate test for outliers in linear regression. *Technometrics*, 17, 473–476.

Montgomery, D. (1984). *Design and Analysis of Experiments*. New York: Wiley, second edition.

Myers, R. (1990). *Classical and Modern Regression with Applications*. Boston: PWS-KENT Pub. Co.

Neter, J., Wasserman, W., and Kutner, M. (1985). *Applied Linear Statistical Models*. Homewood, Illinois: Irwin, second edition.

Peña, D. (1987). *Estadística Modelos y Métodos. 2. Modelos Lineales y Series Temporales*. Madrid: Alianza Editorial.

Petersen, R. (1985). *Design and Analysis of Experiments*. New York: Marcel Dekker.

Radhakrishna Rao, C. and Kumar Mitra, S. (1971). *Generalized Inverse of Matrices and its Applications*. John Wiley & Sons, New York [etc.].

Raktoe, B. (1981). *Factorial Designs*. New York: Wiley.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, 519.237.8 RIP.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific.

Ryan, T. (1997). *Modern Regression Methods*. Wiley, Signatura: 519.233.4 RYA.

Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.

Searle, S. (1971). *Linear Models*. New York: Wiley.

Seber, G. (1977). *Linear Regression Analysis*. New York: Wiley.

Shapiro, S. and Francia, R. (1972). An Approximate Analysis of Variance Test for Normality. *Journal of the American Statistical Association*, 67, 215–216.

Shapiro, S. and Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.

Silvey, S. (1969). Multicollinearity and Imprecise Estimation. *Journal of the Royal Statistical Society, Ser. B*, 31, 539–552.

Silvey, S. (1980). *Optimal Design*. London: Chapman & Hall.

Theil, H. (1971). *Principles of Econometrics*. New York: Wiley.

Trocóniz, A. F. (1987a). *Modelos Lineales*. Bilbao: Serv. Editorial UPV/EHU.

Trocóniz, A. F. (1987b). *Probabilidades. Estadística. Muestreo*. Madrid: Tebar-Flores.

Tusell, F. (2003). Estadística Matemática. 154 p., notas de clase.

Venables, B., Smith, D., Gentleman, R., and Ihaka, R. (1997). *Notes on R: A Programming Environment for Data Analysis and Graphics*. Dept. of Statistics, University of Adelaide and University of Auckland, Available at `http://cran.at.r-project.org/doc/R-intro.pdf`.

Venables, B., Smith, D., Gentleman, R., Ihaka, R., and Mächler, M. (2000). *Notas sobre R: Un entorno de programación para análisis de datos y gráficos*. Traducción española de A. González y S. González.

Venables, W. and Ripley, B. (1999a). *Modern Applied Statistics with* S-PLUS. New York: Springer-Verlag, third edition.

Venables, W. and Ripley, B. (1999b). 'R' Complements to Modern Applied Statistics with S-PLUS. En `http://www.stats.ox.ac.uk/pub/MASS3`.

Webster, J., Gunst, R., and Mason, R. (1974). Latent Root Regression Analysis. *Technometrics*, 16, 513–522.

# Índice alfabético