

**Análisis de Regresión.
Introducción Teórica y
Práctica basada en R**

Fernando Tusell

Bilbao, Octubre 2011

Índice general

Índice general	I
Índice de figuras	IV
Índice de cuadros	v
1 El modelo de regresión lineal.	1
1.1. Planteamiento del problema.	1
1.2. Notación	3
1.3. Supuestos.	5
1.4. MCO como aproximación vectorial	7
1.5. Proyecciones.	7
1.6. Lectura recomendada.	9
2 Estimación mínimo cuadrática.	15
2.1. Obtención de los estimadores de los parámetros.	15
2.2. Una obtención alternativa	17
2.3. Propiedades del estimador mínimo cuadrático $\hat{\beta}$	18
2.4. Estimación de la varianza de la perturbación.	21
2.5. El coeficiente R^2	28
2.6. Algunos lemas sobre proyecciones.	31
2.7. Lectura recomendada	36
3 Identificación. Colinealidad exacta	43
3.1. Modelos con matriz de diseño de rango deficiente.	43
3.2. Funciones estimables.	45
3.3. Restricciones de identificación.	46
3.4. Multicolinealidad exacta y aproximada	49
3.5. Lectura recomendada.	49
4 Estimación con restricciones	50

4.1. Planteamiento del problema.	50
4.2. Lemas auxiliares.	51
4.3. Estimación condicionada.	53
5 Especificación inadecuada del modelo	60
5.1. Introducción.	60
5.2. Inclusión de regresores irrelevantes.	60
5.3. Omisión de regresores relevantes.	63
5.4. Consecuencias de orden práctico	64
6 Regresión con perturbaciones normales.	65
6.1. Introducción.	65
6.2. Contraste de hipótesis lineales.	72
6.3. Intervalos de confianza para la predicción	80
6.4. Lectura recomendada.	81
7 Regresión con R	83
7.1. Tipología de variables explicativas.	83
7.2. Factores y <i>dataframes</i>	85
7.3. Fórmulas	90
7.4. La función <code>lm</code>	97
7.5. Lectura recomendada.	105
8 Inferencia simultánea.	106
8.1. Problemas que plantea el contrastar múltiples hipótesis si- multáneas	106
8.2. Desigualdad de Bonferroni.	111
8.3. Intervalos de confianza basados en la máxima t	112
8.4. Método S de Scheffé.	114
8.5. Empleo de métodos de inferencia simultánea.	119
9 Multicolinealidad.	122
9.1. Introducción.	122
9.2. Una aproximación intuitiva	123
9.3. Detección de la multicolinealidad aproximada	125
9.4. Caracterización de formas lineales estimables.	127
9.5. Varianza en la estimación de una forma lineal.	130
9.6. Elección óptima de observaciones.	131
10 Regresión sesgada.	136
10.1. Introducción.	136
10.2. Una aproximación intuitiva.	137

10.3. Regresión ridge.	139
10.4. Regresión en componentes principales.	150
10.5. Regresión en raíces latentes	158
10.6. Lectura recomendada	162
11 Evaluación del ajuste. Diagnósticos.	165
11.1. Análisis de residuos.	165
11.2. Análisis de influencia.	170
11.3. Análisis gráfico de residuos	174
12 Selección de modelos.	180
12.1. Criterios para la comparación.	180
12.2. Selección de variables.	189
12.3. El LASSO	200
12.4. Modelos bien estructurados jerárquicamente	201
13 Transformaciones	204
13.1. Introducción	204
13.2. Transformaciones de los regresores	204
13.3. Transformaciones de la variable respuesta	207
14 Regresión con respuesta cualitativa	211
14.1. El modelo <i>logit</i>	211
A Algunos resultados en Algebra Lineal.	220
A.1. Resultados varios sobre Algebra Matricial.	220
A.2. Cálculo diferencial con notación matricial	222
A.3. Lectura recomendada	223
B Algunos prerrequisitos estadísticos.	224
B.1. Distribuciones χ^2 y \mathcal{F} descentradas	224
B.2. Estimación máximo verosímil	225
B.3. Contraste razón generalizada de verosimilitudes	226
C Regresión en S-Plus y R.	227
C.1. El sistema estadístico y gráfico S-PLUS	227
C.2. El sistema estadístico y gráfico R	227
C.3. Correspondencia de funciones para regresión y ANOVA en S-PLUS y R	234
D Procedimientos de cálculo.	235
D.1. Introducción	235

D.2. Transformaciones ortogonales.	235
D.3. Factorización QR.	238
D.4. Bibliografía	240
E Enunciados y demostraciones formales	241
E.1. Existencia y unicidad de proyecciones.	241
E.2. Proyección sobre subespacios $h = M \cap K(B)$	244
Bibliografía	246

Índice de figuras

1.1. Old Faithful Geysers: datos de 272 erupciones.	2
1.2. El vector $P_M \vec{y}$ es la proyección de \vec{y} sobre M (plano horizontal).	8
2.1. $X\hat{\beta}$ es la proyección de \vec{y} sobre M . $R^2 = \cos^2 \alpha$	29
2.2. En un ajuste sin término constante, la pendiente depende de la elección arbitraria del origen	42
3.1. Regresión en el caso de matrix X de rango deficiente.	44
3.2. Caso de un vector $\vec{\beta}$ parcialmente estimable.	45
9.1. Multicolinealidad exacta (panel superior) y aproximada (panel inferior).	124
10.1. Componentes del $ECM(\hat{\beta}^{(k)})$ en el estimador <i>ridge</i> . Las líneas de trazos y puntos representa respectivamente la varianza y (sesgo) ² de $\hat{\beta}^{(k)}$ en función de k . La curva sólida representa $ECM[\hat{\beta}^{(k)}]$. La línea horizontal es la varianza (y ECM) del estimador $\hat{\beta}$ MCO.	143
10.2. Trazas ridge y GVC para los datos <code>longley</code>	147
11.1. Una observación como a tiene residuo borrado muy grande, y gran influencia en la pendiente de la recta de regresión.	171
11.2. Gráficos para contraste de normalidad	177
12.1. Valores de C_p y \overline{R}^2 para 141 modelos ajustados a los datos <code>UScrime</code>	194

13.1. Disposición de residuos sugiriendo una transformación cuadrática del regresor X_i	205
D.1. Visualización de la transformación de Householder.	237

Índice de cuadros

C.1. Equivalencia de funciones para regresión y ANOVA en S-PLUS y R.	234
--	-----

Introducción

Lo que sigue contiene una introducción muy concisa al análisis de regresión, concebida como apoyo de las clases. 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 las secciones de COMPLEMENTOS Y EJERCICIOS al fin de cada capítulo: son parte integrante del texto a este segundo nivel y completan muchos detalles.

A lo largo del texto, tanto en demostraciones como en ejercicios o complementos se ha hecho uso abundante del símbolo de “giro peligroso” mostrado 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. La flecha vertical \uparrow remite a algún ejercicio, observación o ejemplo que son requisito previo.



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 la teoría hay fragmentos en R, que el lector puede ejecutar o tomar como modelo. Todos se han ejecutado con R versión 2.13.2.

No se ha buscado el código más terso ni la forma más rápida o elegante de hacer las cosas, sino la que ilustra mejor la teoría.

Capítulo 1

El modelo de regresión lineal.

1.1. Planteamiento del problema.

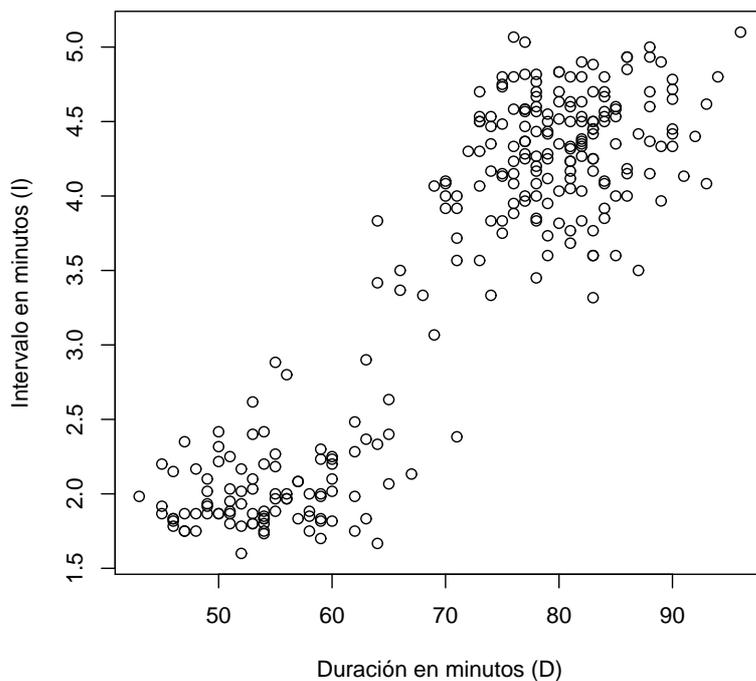
Son frecuentes en la práctica situaciones en las que se cuenta con observaciones de diversas variables, y es razonable pensar en una relación entre ellas. El poder determinar si existe esta relación —y, en su caso, una forma funcional para la misma— es de sumo interés. Por una parte, ello permitiría, 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. 2), 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 en el caso de 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).

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

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



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 ϵ . 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.

FIN DEL EJEMPLO ■

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

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

relacionando de manera aproximada los valores de Y con los que toman otras variables, X_0, \dots, X_{p-1} . Por simplicidad, limitaremos por el momento

nuestra atención a funciones $f(X_0, \dots, X_{p-1})$ lineales; el modelo resultante es el modelo de regresión lineal, que se examina en la Sección 1.2 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 ϵ 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, sino sólo que parece existir una relación entre ambas variables.

1.2. Notación

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

$$Y = \beta_0 X_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1} + \epsilon, \quad (1.1)$$

siendo:

1. $\beta_0, \dots, \beta_{p-1}$, parámetros fijos desconocidos.
2. X_0, \dots, X_{p-1} , variables explicativas no estocásticas, *regresores*, cuyos valores son fijados por el experimentador. Frecuentemente X_0 toma el valor constante “uno”.
3. ϵ 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 ϵ . 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 β_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 $\beta_0, \dots, \beta_{p-1}$. 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, ϵ 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:

$$\vec{y} = X\vec{\beta} + \vec{\epsilon}, \quad (1.2)$$

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 $(\beta_0, \dots, \beta_{p-1})'$,
- $\vec{\epsilon}$ el vector $N \times 1$ de valores de la perturbación aleatoria ϵ .

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 para la estimación el *criterio mínimo cuadrático ordinario (MCO)*. Por consiguiente, diremos que $\hat{\beta}$ es óptimo si $\|\vec{y} - X\hat{\beta}\|^2$ es mínimo, denotando $\|\cdot\|$ la norma euclídea ordinaria:

$$\|\vec{y}\|^2 \stackrel{\text{def}}{=} \sum_i y_i^2$$

(ver Definición A.2, pág. 220).

Observación 1.1 El suponer que los valores de los regresores pueden ser fijados por el analista (apartado 2, al comienzo de esta

Sección) 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 .

Observación 1.2 El criterio de seleccionar como estimadores de $\vec{\beta}$ el vector $\hat{\beta}$ minimizando $\|\vec{y} - X\hat{\beta}\|^2$ es totalmente arbitrario. En lugar de minimizar la norma euclídea ordinaria, podríamos minimizar $\|\vec{y} - X\hat{\beta}\|_{L1}$ (suma de los valores absolutos de los errores de aproximación, también llamada *norma L1*), o cualquier otra cosa. Si se emplea la norma euclídea es por conveniencia matemática y por ser un criterio “razonable” desde diversos puntos de vista.

Observación 1.3  ¿Por qué introducir la norma euclídea y no limitarnos a proponer como criterio la minimización de

$$\sum_i \left(y_i - \hat{\beta}_0 x_{i0} - \hat{\beta}_1 x_{i1} - \dots - \beta_{p-1} x_{i,p-1} \right)^2?$$

Si realizamos las demostraciones en términos de normas, servirán *sea cual fuere la norma que adoptemos*. Muchos resultados serán así “todo terreno”, trasladables de inmediato a problemas con supuestos diferentes a los realizados en la Sección 1.3 a continuación. Veremos en breve (Observación 2.1, pág. 16) ventajas adicionales de plantear y resolver el problema en términos de aproximación vectorial, minimizando una norma.

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 éste suele suceder; típicamente, la primera columna está formada por “unos”). En efecto, es claro que si:

$$\vec{Y} = \beta_0 \vec{1} + \beta_1 \vec{x}_1 + \cdots + \beta_{p-1} \vec{x}_{p-1} + \vec{\epsilon} \quad (1.3)$$

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

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

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 incorrelacionadas (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 1.4 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:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k^2 + \cdots + \beta_{p-1} x_{p-1} + \epsilon.$$

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

$$W = k z_1^{\beta_1} z_2^{\beta_2} \nu,$$

siendo ν 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 así:

$$\vec{y} = \hat{\beta}_0\vec{x}_0 + \cdots + \hat{\beta}_{p-1}\vec{x}_{p-1} + \hat{\epsilon}, \quad (1.5)$$

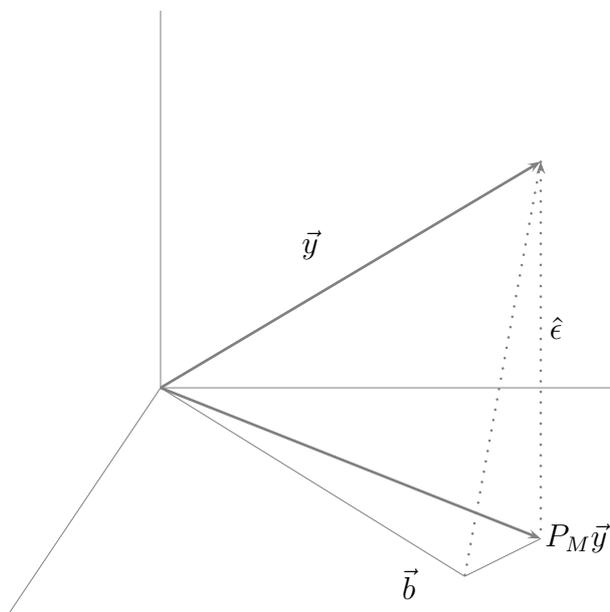
donde $\vec{x}_0, \dots, \vec{x}_{p-1}$ denotan los vectores columna de la matriz X (\vec{x}_0 será en general una columna de “unos”, como se ha indicado). Hay diferentes posibilidades en cuanto a criterio de estimación de los β . Si adoptamos el criterio MCO propuesto más arriba, consistente en minimizar $\|\hat{\epsilon}\|^2$, la ecuación (1.5) muestra que el problema puede reformularse así: ¿Cuales son los coeficientes $\hat{\beta}_0, \dots, \hat{\beta}_{p-1}$ 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) el vector \vec{y} ? Veremos inmediatamente que esta combinación lineal es lo que llamaremos *proyección* de \vec{y} sobre el subespacio generado por las columnas $\vec{x}_0, \dots, \vec{x}_{p-1}$.

1.5. Proyecciones.

Aunque en lo que sigue se hace un tratamiento generalizable, implícitamente consideramos productos internos (véase Definición A.1, pág. 220) 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 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 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 $\vec{y} \notin M$.

Figura 1.2: El vector $P_M \vec{y}$ es la proyección de \vec{y} sobre M (plano horizontal).

Siempre existe (y es única) la proyección de un vector en H sobre el subespacio M , tal como establece el teorema siguiente¹.

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

$$\|\vec{y} - \vec{u}\|^2 = \min_{\vec{z} \in M} \|\vec{y} - \vec{z}\|^2. \quad (1.6)$$

La Fig. 1.2 ilustra en tres dimensiones la noción de proyección, y hace intuitivamente evidente el Teorema 1.1. En dicha figura se ha considerado $H = \mathbb{R}^3$ y un subespacio M de dimensión dos representado como el plano horizontal. Consideremos $P_M \vec{y}$: podríamos describirlo como el obtenido al dejar caer una plomada desde el extremo de \vec{y} hasta hacer contacto con M .

Es claro que $\hat{\epsilon} = \vec{y} - P_M \vec{y}$ es ortogonal a M . Como consecuencia, para cualquier vector $\vec{b} \neq P_M \vec{y}$ en M , $\vec{y} - \vec{b}$ es la hipotenusa de un triángulo

¹Estrictamente incorrecto. El Teorema E.1, pág. 242 es una versión más elaborada del Teorema 1.1.

rectángulo, cuyos catetos son $\hat{\epsilon}$ y el segmento $\vec{b} - P_M \vec{y}$. Por tanto,

$$\|\vec{y} - \vec{b}\|^2 = \|\hat{\epsilon}\|^2 + \|\vec{b} - P_M \vec{y}\|^2 > \|\hat{\epsilon}\|^2$$

lo que demuestra la propiedad de $P_M \vec{y}$ de ser la mejor aproximación de \vec{y} en M . (Una demostración formal que va más allá de esta incompleta argumentación puede encontrarse en la Sección E.1, pág. 242.)

1.6. Lectura recomendada.

Sobre la teoría. Puede leerse como complemento a este capítulo Faraway (2005), Cap. 1 y Cap. 2, Sección 1 a 3, o los capítulos introductorios de la miríada de buenos textos que existe sobre regresión lineal: Seber (1977), Stapleton (1995), Arnold (1981), Draper and Smith (1998), Fox (2002), Peña (2002), Myers (1990), Searle (1971), Ryan (1997) o Trocóniz (1987a) son algunos de ellos.

Sobre la utilización de R. El primero de los libros citados, Faraway (2005), ilustra también el modo de emplear R para hacer regresión (pero es demasiado escueto para servir de introducción al lenguaje). R es una implementación de fuente libre del lenguaje estadístico y gráfico S (ver por ejemplo Becker et al. (1988), Chambers and Hastie (1992) o Chambers (1998)). Los textos introductorios sobre S son por ello utilizables con R. Buenos manuales incluyen Venables and Ripley (1999a) (con su complemento específico para R, Venables and Ripley (1999b)), Dalgaard (2002), o Ugarte et al. (2008). Hay documentos con extensión de libro disponibles en Internet, como Maindonald (2000) o Kuhnert and Venables (2005).

COMPLEMENTOS Y EJERCICIOS

Algunos de los ejercicios que siguen requieren hacer uso de un ordenador y un programa especializado, tal como R. En la Sección 1.6, pág. 9, se proporcionan referencias.

1.1 En 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”.

1.2 En R para crear un vector y asignarlo a la variable `x` haremos:

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

1.3 Para efectuar multitud de cálculos en R 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`.

1.4 El producto interno euclídeo de dos vectores `x` e `y` puede calcularse así:

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

o alternativamente:

```
x %*% y
```

1.5 En R rige la “regla del reciclado”, que permite operar con operandos disimilares. Por ejemplo, si:

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

entonces, tecleando

```
a + b
```

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

1.6 En 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.

1.7 La función `help` permite interrogar a R sobre el modo 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 retectarlos. Obsérvese también que el mandato `help.start()` abre una ventana de ayuda en un navegador —si es que hay alguno instalado en la máquina que empleamos—, lo que permite navegar cómodamente por la documentación.

1.8 Cuando escribimos expresiones como

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

estamos empleando funciones predefinidas (en este caso, `sum`). En 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:

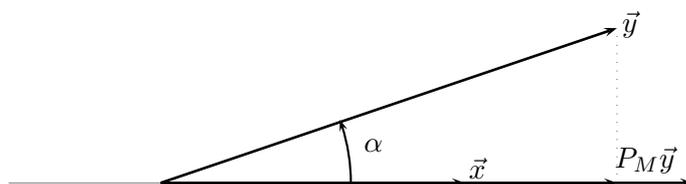
```
norma.eucl(x)
```

En realidad, la definición de una función como `eucl` es innecesaria: en R podemos emplear `x**% x` (o alternativamente `crossprod(x)`) que cumplen análogo cometido.

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

$$\langle \vec{x}, \vec{y} \rangle = \|\vec{x}\| \|\vec{y}\| \cos(\alpha)$$

siendo α el ángulo que ambos vectores forman. Esta igualdad se extiende a R^N definiendo $\cos(\alpha)$ convenientemente (véase Definición A.3, pág. 220). 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 $\langle \vec{x}, \vec{y} \rangle = \|P_M \vec{y}\|$, siendo M el subespacio generado por \vec{x} .



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

$$P_M \vec{y} = \frac{\langle \vec{x}, \vec{y} \rangle}{\langle \vec{x}, \vec{x} \rangle} \vec{x}$$

1.10 Escribese una función que, dados dos vectores arbitrarios \vec{x} e \vec{y} , obtenga el vector proyección del segundo sobre el espacio (unidimensional) generado por el primero. 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{x} , y ii) Que $(\vec{y} - \vec{z}) \perp \vec{x}$.

1.11 Demuéstrase que los siguientes cuatro vectores de R^3 son un sistema generador de dicho espacio, pero no base.

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

1.12 (\uparrow 1.11) Selecciónese, de entre los cuatro vectores indicados en el Problema 1.11, tres que formen base de R^3 .

1.13 (\uparrow 1.10) 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.

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

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

1.15  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).

1.16  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).

1.17  Como se ha indicado en la Observación 1.2, pág. 5, hay alternativas al criterio MCO. En lugar de minimizar la suma de cuadrados de los residuos, podríamos minimizar la suma de sus valores absolutos: $\sum_{i=1}^N |\hat{\epsilon}_i|$ (norma L1 del vector de residuos). Uno de sus atractivos es que los resultados resultan menos afectados por observaciones con residuo muy grande; pero es computacionalmente mucho más costosa.

Capítulo 2

Estimación mínimo cuadrática.

2.1. Obtención de los estimadores 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 $\langle \vec{v}, \vec{w} \rangle = \vec{v}'\vec{w}$, 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 ,

$$\vec{X}_0 \perp (\vec{y} - X\hat{\beta}) \quad (2.1)$$

$$\vec{X}_1 \perp (\vec{y} - X\hat{\beta}) \quad (2.2)$$

$$\vdots \quad \quad \quad \vdots \quad (2.3)$$

$$\vec{X}_{p-1} \perp (\vec{y} - X\hat{\beta}) \quad (2.4)$$

que podemos reunir en la igualdad matricial

$$X'(\vec{y} - X\hat{\beta}) = \vec{0}$$

y de aquí se deduce que:

$$X'X\hat{\beta} = X'\vec{y}. \quad (2.5)$$

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

$$\hat{\beta} = (X'X)^{-1}X'\vec{y}. \quad (2.6)$$

Obsérvese que el supuesto de rango total de la matriz X —y consiguientemente de $(X'X)$ — es requerido exclusivamente para pasar de (2.5) a (2.6). 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 (2.6) se deduce también que, en el caso de rango total, la proyección de \vec{y} sobre M viene dada por

$$P_M \vec{y} = X(X'X)^{-1}X'\vec{y}, \quad (2.7)$$

y el vector de residuos por

$$\hat{\epsilon} = \vec{y} - X\hat{\beta} \quad (2.8)$$

$$= \vec{y} - X(X'X)^{-1}X'\vec{y} \quad (2.9)$$

$$= (I - X(X'X)^{-1}X')\vec{y} \quad (2.10)$$

$$= (I - P_M)\vec{y}. \quad (2.11)$$

Observación 2.1 El ser $X\hat{\beta}$ proyección de \vec{y} sobre M garantiza sin más que $\|\hat{\epsilon}\|$ es mínimo. Si hubiéramos obtenido $\hat{\beta}$ derivando

$$\sum_i \left(y_i - \hat{\beta}_0 x_{i0} - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_{p-1} x_{i,p-1} \right)^2$$

e igualando las derivadas a cero (ver Observación 1.3, pág. 5), obtendríamos un $\hat{\beta}$ del que todo lo que podríamos afirmar es que corresponde a un punto estacionario de la expresión anterior (suma de cuadrados de los residuos). Para establecer que se trata de un mínimo, habríamos de tomar aún segundas derivadas y verificar el cumplimiento de las condiciones de segundo orden.

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 $(I - P)$, sobreenunciando 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. $\text{rango}(I - P) = N - p$.

3. Se verifica que $(I - P)X = 0$.

DEMOSTRACIÓN:

El apartado 1) es inmediato. En cuanto a 2), siendo $(I - P)$ idempotente, su rango coincide con su traza (véase Teorema A.1, pág. 220). Por tanto:

$$\text{rango}(I - P) = \text{traza}(I - P) \quad (2.12)$$

$$= \text{traza}(I) - \text{traza}(P) \quad (2.13)$$

$$= N - \text{traza}[X(X'X)^{-1}X'] \quad (2.14)$$

$$= N - \text{traza}[(X'X)^{-1}X'X] \quad (2.15)$$

$$= N - p. \quad (2.16)$$

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 $\vec{0}$. ■

2.2. Una obtención alternativa

La obtención del vector de estimadores $\hat{\beta}$ en la sección precedente tiene muchos méritos, y no es el menor el de proporcionar intuición geométrica acerca de la solución mínimo cuadrática ordinaria (MCO). Tendremos ocasiones abundantes de explotar esta intuición.

Podemos seguir una vía alternativa para llegar al mismo resultado: plantear el problema en forma de minimización respecto a $\vec{\beta}$ de la expresión:

$$\sum_{i=1}^N (y_i - \beta_0 x_{i0} - \beta_1 x_{i1} - \dots - \beta_{p-1} x_{i,p-1})^2, \quad (2.17)$$

tal como sugería la Observación 2.1. Con notación matricial, el problema puede reescribirse así:

$$\min_{\vec{\beta}} (\vec{y} - X\vec{\beta})'(\vec{y} - X\vec{\beta}). \quad (2.18)$$

La “suma de cuadrados” anterior es una forma cuadrática de matriz unidad. Haciendo uso de la fórmula (A.12), pág. 222, obtenemos las condiciones de primer orden

$$2X'(\vec{y} - X\vec{\beta}) = \vec{0}, \quad (2.19)$$

o equivalentemente

$$X' \vec{y} = (X'X) \vec{\beta}, \quad (2.20)$$

que son las ecuaciones normales (2.5).

Es fácil comprobar tomando las segundas derivadas que la solución (o soluciones, si hay más de una) del sistema de ecuaciones precedente corresponde a un mínimo y no a un máximo o punto de silla: la matriz de segundas derivadas $(X'X)$ es por construcción (semi)definida positiva.

Importa comprobar que esta aproximación al problema, a diferencia de la que hacía uso de la noción de proyección, deja en la penumbra muchas cosas que son de interés: la ortogonalidad del vector de residuos $\hat{\epsilon} = \vec{y} - X\hat{\beta}$, la idempotencia de algunas matrices, etc.

2.3. Propiedades del estimador mínimo cuadrático $\hat{\beta}$.

Notemos que $\hat{\beta}$ es *un vector aleatorio*. Aunque X se mantenga fija — cosa que podemos lograr, pues los valores de los regresores se fijan por el experimentador: recuérdese los supuestos introducidos en la Sección 1.2 —, en experimentos repetidos obtendremos cada vez un diferente vector \vec{y} de valores de la variable respuesta. En efecto, cada vez intervendrán en la formación de \vec{y} diferentes perturbaciones.

El vector $\hat{\beta} = (X'X)^{-1}X'\vec{y}$ por tanto es un vector aleatorio: “hereda” su condición de tal de \vec{y} , que a su vez la obtiene de $\vec{\epsilon}$. Tiene por ello sentido preguntarse por su vector de valores medios y por su matriz de covarianzas.

Recordemos que un estimador $\hat{\gamma}$ del parámetro γ se dice *insesgado* si

$$E[\hat{\gamma}] = \gamma.$$

En el caso de estimar un vector de parámetros, la condición análoga es

$$E[\hat{\beta}] = \vec{\beta}.$$

Recordemos también que la *matriz de covarianzas* de un vector aleatorio como $\hat{\beta}$ se define por:

$$\Sigma_{\hat{\beta}} = E[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]'$$

expresión que en el caso de ser $\hat{\beta}$ insesgado como estimador de $\vec{\beta}$ se simplifica de modo obvio a

$$\Sigma_{\hat{\beta}} = E[\hat{\beta} - \vec{\beta}][\hat{\beta} - \vec{\beta}]'$$

La matriz de covarianzas $\Sigma_{\hat{\beta}}$ tiene en su diagonal principal las varianzas de los componentes del vector $\hat{\beta}$ y fuera de la diagonal principal las covarianzas.

La insesgadez de un estimador es intuitivamente atrayente: supone que no incurrimos en derivas sistemáticas al estimar el parámetro objeto de interés. Si repitiéramos el mismo experimento muchas veces y promediáramos los valores del estimador insesgado obtenidos en cada experimento, esperaríamos que este promedio se acercará progresivamente más a su objetivo (el verdadero valor del parámetro).

Acontece que el vector de estimadores $\hat{\beta}$ disfruta de esta atractiva propiedad de insesgadez. Adicionalmente, dentro de una clase particular de estimadores es el que exhibe menores varianzas en la diagonal principal de $\Sigma_{\hat{\beta}}$ —y, en este sentido, es el que estima con mayor precisión el vector $\vec{\beta}$ —. El siguiente Teorema formaliza y demuestra estas propiedades.

Teorema 2.2 *Si se verifican los supuestos habituales (Sección 1.3, pág. 5) se cumple también 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 $\Sigma_{\hat{\beta}}$.

DEMOSTRACIÓN:

Tomando valor medio en (2.6):

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

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

$$\begin{aligned}\Sigma_{\hat{\beta}} &= E(\hat{\beta} - \vec{\beta})(\hat{\beta} - \vec{\beta})' \\ &= E[(X'X)^{-1}X'(X\vec{\beta} + \vec{\epsilon}) - \vec{\beta}][(X'X)^{-1}X'(X\vec{\beta} + \vec{\epsilon}) - \vec{\beta}]' \\ &= E[(X'X)^{-1}X'\vec{\epsilon}][(X'X)^{-1}X'\vec{\epsilon}]' \\ &= E[(X'X)^{-1}X'\vec{\epsilon}\vec{\epsilon}'X(X'X)^{-1}] \\ &= (X'X)^{-1}X'\sigma^2IX(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}.\end{aligned}$$

Para demostrar 3), consideremos cualquier estimador $\hat{\beta}_*$ alternativo a $\hat{\beta}$. 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:

$$C = (X'X)^{-1}X' + D. \quad (2.21)$$

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:

$$E[(X'X)^{-1}X'(X\vec{\beta} + \vec{\epsilon}) + D(X\vec{\beta} + \vec{\epsilon})] = \vec{\beta}, \quad (2.22)$$

$$\vec{\beta} + DX\vec{\beta} = \vec{\beta}, \quad (2.23)$$

dado que $E\vec{\epsilon} = \vec{0}$. Como (2.23) se ha de verificar sea cual fuere $\vec{\beta}$, la insesgader de $\hat{\beta}_*$ implica $DX = 0$.

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

$$\Sigma_{\hat{\beta}_*} = E[(\hat{\beta}_* - \vec{\beta})(\hat{\beta}_* - \vec{\beta})']. \quad (2.24)$$

Pero:

$$(\hat{\beta}_* - \vec{\beta}) = [(X'X)^{-1}X' + D]\vec{Y} - \vec{\beta} \quad (2.25)$$

$$= [(X'X)^{-1}X' + D](X\vec{\beta} + \vec{\epsilon}) - \vec{\beta} \quad (2.26)$$

$$= [(X'X)^{-1}X' + D]\vec{\epsilon}. \quad (2.27)$$

donde (2.27) se ha obtenido haciendo uso de $DX = 0$. Llevando (2.27) a (2.24), obtenemos:

$$\Sigma_{\hat{\beta}_*} = E\{[(X'X)^{-1}X' + D]\vec{\epsilon}\vec{\epsilon}'[(X'X)^{-1}X' + D]'\} \quad (2.28)$$

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

$$\Sigma_{\hat{\beta}_*} = (X'X)^{-1}X'\sigma^2IX(X'X)^{-1} + \sigma^2DID' \quad (2.29)$$

$$= \sigma^2(X'X)^{-1} + \sigma^2DD' \quad (2.30)$$

$$= \Sigma_{\hat{\beta}} + \sigma^2DD'. \quad (2.31)$$

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. ■

Observación 2.2 La insesgades de un estimador es una propiedad en principio atrayente, pero de ningún modo indispensable. De hecho, un estimador insesgado de un parámetro puede incluso no existir. (Para una discusión de la condición de insesgades y de sus implicaciones puede verse Lehmann (1983), Cap. 2.)

En el Capítulo 10 comprobaremos que, en ocasiones, podemos optar con ventaja por utilizar estimadores sesgados.

2.4. Estimación de la varianza de la perturbación.

El Teorema 2.2 proporciona la matriz de covarianzas del vector de estimadores $\hat{\beta}$, $\Sigma_{\hat{\beta}} = \sigma^2(X'X)^{-1}$. Pero mientras que $(X'X)$ es conocida, σ^2 es un parámetro que necesita ser estimado. Veamos como hacerlo.

Definición 2.1 Denominamos *SSE* o suma de cuadrados de los residuos al cuadrado de la norma del vector de residuos,

$$SSE \stackrel{def}{=} \|\vec{y} - X\hat{\beta}\|^2 = \|\hat{\epsilon}\|^2$$

Teorema 2.3 Una estimación insesgada de la varianza de la perturbación viene proporcionada por

$$\hat{\sigma}^2 = \frac{SSE}{N - p}$$

DEMOSTRACIÓN:

Como

$$X\hat{\beta} = P\vec{Y} = X(X'X)^{-1}X'\vec{Y}, \quad (2.32)$$

tenemos que

$$(\vec{Y} - X\hat{\beta}) = (I - P)\vec{Y} \quad (2.33)$$

$$= (I - P)(X\vec{\beta} + \vec{\epsilon}) \quad (2.34)$$

$$= (I - P)\vec{\epsilon}, \quad (2.35)$$

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 $(I - P)$,

$$SSE = \vec{\epsilon}'(I - P)\vec{\epsilon} \quad (2.36)$$

$$= \text{traza } \vec{\epsilon}'(I - P)\vec{\epsilon} \quad (2.37)$$

$$= \text{traza } (I - P)\vec{\epsilon}\vec{\epsilon}'. \quad (2.38)$$

Tomando valor medio en (2.38) tenemos:

$$E(SSE) = \text{traza } (I - P)(\sigma^2 I) = \sigma^2(N - p). \quad (2.39)$$

(El último paso ha hecho uso de la propiedad $\text{traza}(I - P) = N - p$, Teorema 2.1, pág. 16.) De (2.39) se deduce entonces que

$$E \left[\frac{SSE}{N - p} \right] = \sigma^2$$

y $\hat{\sigma}^2 \stackrel{\text{def}}{=} SSE/(N - p)$ es por tanto un estimador insesgado de σ^2 . ■

Observación 2.3 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.

Observación 2.4 El Teorema 2.3 muestra que para obtener una estimación insesgada de la varianza de la perturbación debemos dividir la suma de cuadrados de los residuos, no entre el número de residuos N , sino entre los *grados de libertad* $N - p$. Que el número de parámetros estimado debe tomarse en consideración en el denominador del estimador es intuitivamente plausible. Después de todo, si aumentáramos el número de regresores (y parámetros estimados) p hasta que $p = N$, SSE sería idénticamente cero. (Estaríamos ante un problema *sin grados de libertad*.) Sin llegar a este extremo, es claro que aumentando el número de regresores incrementamos nuestra capacidad de aproximar \vec{y} (y de reducir SSE), y esto ha de ser contrapesado reduciendo también el denominador.

Observación 2.5 El Teorema 2.3 subsume y amplía un resultado que habitualmente aparece sin demostración en los cursos elementales de Estadística: un estimador insesgado de la varianza de una población, dada una muestra i.i.d. de la misma, viene dada por

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N - 1}. \quad (2.40)$$

Este resultado puede obtenerse como caso particular del Teorema 2.3 si reparamos en lo siguiente: podemos imaginar las Y_i como generadas por

$$Y_i = \beta_0 + \epsilon_i,$$

en que β_0 es la media y ϵ_i una perturbación de media cero y misma varianza que Y_i . Si regresáramos las observaciones Y_1, \dots, Y_N sobre una columna de “unos”, $\vec{1}$, el único parámetro estimado sería:

$$\hat{\beta}_0 = (X'X)^{-1}X'\vec{Y} = (\vec{1}'\vec{1})^{-1}\vec{1}'\vec{Y} = N^{-1}\sum_{i=1}^N Y_i = \bar{Y}$$

El mejor ajuste que puede hacerse de las Y_i en términos de este único regresor es $\hat{\beta}_0\vec{1}$ y la suma de cuadrados de los residuos es por tanto $\sum_{i=1}^N (Y_i - \hat{\beta}_0\vec{1})^2 = \sum_{i=1}^N (Y_i - \bar{Y})^2$. La expresión (2.40) coincide por tanto, en este caso particular, con la dada por el Teorema 2.3.

R: Ejemplo 2.1 (cálculo de los estimadores MCO)

El siguiente listado crea artificialmente una matriz X y 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 `lsfit` predefinida en R, 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 factorización QR (ver Sección D.2, pág. 235, o Lawson and Hanson (1974)). Se trata de detalles que no necesitan preocuparnos por el momento. Generamos en primer lugar los datos y realizamos la estimación aplicando la teoría de modo más directo. Primero, la matriz de diseño,

```
> X <- matrix(c(1, 1, 1, 1,
+             1, 1, 1, 4, 12, 1, 4,
+             13, 0, 6, 7, 0, 2, 2),
+            6, 3)
> 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
```

A continuación, fijamos un vector $\vec{\beta}$

```
> beta <- c(2, 3, 4)
```

Finalmente, generamos los valores de la variable respuesta del modo que prescribe el modelo lineal:

```
> y <- X %*% beta + rnorm(6)
```

(La función `rnorm(n)` genera `n` variables aleatorias $N(0, 1)$.) A continuación, obtenemos los estimadores resolviendo las ecuaciones normales (2.5), pág. 15. Se muestran varias formas alternativas de hacerlo. Podemos por ejemplo escribir

```
> b <- solve(t(X) %*% X, t(X) %*%
+           y)
> b
```

```

      [,1]
[1,] 2.3517
[2,] 2.8129
[3,] 4.2329

```

(la función `solve(A,b)` proporciona una solución, si existe, del sistema de ecuaciones lineales $A\vec{x} = \vec{b}$). Una forma más rápida de calcular $(X'X)$ y $X'\vec{y}$ la proporciona la función `crossprod`. Podríamos sustituir lo anterior por

```

> b <- solve(crossprod(X),
+           crossprod(X, y))
> b

```

```

      [,1]
[1,] 2.3517
[2,] 2.8129
[3,] 4.2329

```

Podemos también escribir:

```

> XXinv <- solve(crossprod(X))
> b <- XXinv %*% crossprod(X,
+           y)
> b

```

```

      [,1]
[1,] 2.3517
[2,] 2.8129
[3,] 4.2329

```

Hemos obtenido separadamente $(X'X)^{-1}$ (que puede servirnos para estimar la matriz de covarianzas de los estimadores, $\hat{\sigma}^2(X'X)^{-1}$). La función `solve` con un único argumento matricial proporciona la matriz inversa. De cualquiera de las maneras que calculemos $\hat{\beta}$, la obtención de los residuos es inmediata:

```

> e <- y - X %*% b
> e

```

```

      [,1]
[1,]  0.42097
[2,] -0.29124
[3,]  0.15416
[4,] -0.61805
[5,]  0.53689
[6,] -0.20272

```

Podemos comprobar la ortogonalidad de los residuos a las columnas de la matriz X :

```

> t(e) %*% X
      [,1] [,2]
[1,] -2.6379e-13 -8.3933e-13
      [,3]
[1,] -5.9686e-13
> crossprod(e, X)
      [,1] [,2]
[1,] -2.6379e-13 -8.3933e-13
      [,3]
[1,] -5.9686e-13
> round(crossprod(e, X))
      [,1] [,2] [,3]
[1,]    0    0    0

```

La suma de cuadrados de los residuos y una estimación de la varianza de la perturbación pueden ahora obtenerse con facilidad:

```

> s2 <- sum(e * e)/(nrow(X) -
+      ncol(X))
> s2
[1] 0.33238

```

FIN DEL EJEMPLO ■

R: Ejemplo 2.2 Todos los cálculos anteriores pueden hacerse con mucha mayor comodidad mediante funciones de regresión especializadas. Por ejemplo,

```
> ajuste <- lsfit(X, y, intercept = FALSE)
```

hace todo lo anterior y algunas cosas más de modo mucho más eficiente. La función `lsfit` (least squares **fit**) devuelve una lista u objeto compuesto conteniendo en sus componentes los estimadores de los parámetros, los residuos y algunos resultados auxiliares asociados al método de cálculo empleado (la factorización QR aludida más arriba). Veámoslo:

```
> ajuste

$coefficients
      X1      X2      X3
2.3517 2.8129 4.2329

$residuals
[1]  0.42097 -0.29124  0.15416
[4] -0.61805  0.53689 -0.20272

$intercept
[1] FALSE

$qr
$qt
[1] -75.33003  48.78812 -23.94068
[4]  -0.66854   0.42874  -0.60529

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

```

$qlraux
[1] 1.4082 1.0362 1.9256

$rank
[1] 3

$pivot
[1] 1 2 3

$tol
[1] 1e-07

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

> resid <- ajuste$residuals
> resid

[1] 0.42097 -0.29124 0.15416
[4] -0.61805 0.53689 -0.20272

```

El argumento `intercept=FALSE` indica a la función `lsfit` que *no* debe agregarse a la matriz de diseño X una columna de “unos” (porque ya figura entre los regresores). Ordinariamente ello no sucederá, y podremos prescindir de especificar el argumento `intercept`, con lo que tomará el valor por omisión `TRUE`.

FIN DEL EJEMPLO ■

2.5. El coeficiente R^2

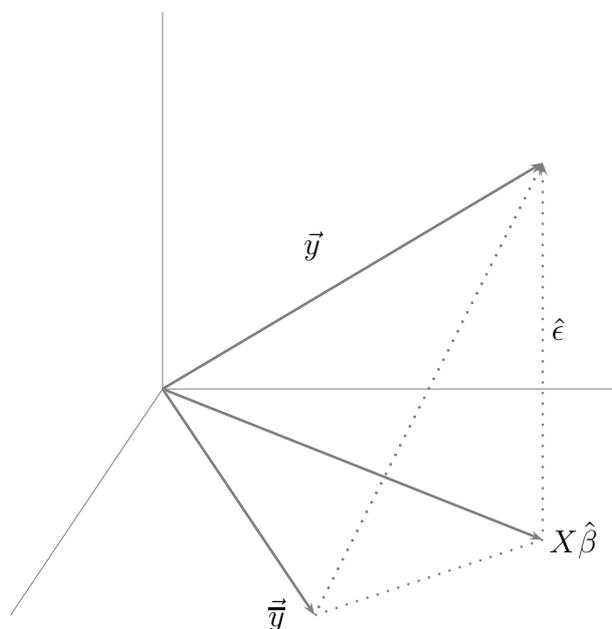
Hay una relación interesante entre SSE y otras dos sumas de cuadrados que definimos a continuación. Sea $\vec{\bar{y}}$ el vector $N \times 1$ siguiente:

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

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

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

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



Se verifica entonces el Teorema a continuación.

Teorema 2.4 Si \vec{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:

$$SST = SSR + SSE \tag{2.41}$$

DEMOSTRACIÓN:

$$SST = \|\vec{y} - \vec{\bar{y}}\|^2 \tag{2.42}$$

$$= \|\vec{y} - X\hat{\beta} + X\hat{\beta} - \vec{\bar{y}}\|^2 \tag{2.43}$$

$$= \langle (\vec{y} - X\hat{\beta}) + (X\hat{\beta} - \vec{\bar{y}}), (\vec{y} - X\hat{\beta}) + (X\hat{\beta} - \vec{\bar{y}}) \rangle \tag{2.44}$$

$$= \|\vec{y} - X\hat{\beta}\|^2 + \|X\hat{\beta} - \vec{\bar{y}}\|^2 + 2 \langle \vec{y} - X\hat{\beta}, X\hat{\beta} - \vec{\bar{y}} \rangle \tag{2.45}$$

Pero si $\vec{y} \in M$, $(X\hat{\beta} - \vec{\bar{y}}) \in M$, y como quiera que $\hat{\epsilon} = (\vec{y} - X\hat{\beta}) \perp M$, el último producto interno es nulo. Por consiguiente (2.45) se reduce a (2.41).

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 (2.41) 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 (2.41) 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{y} no perteneciera a M , que hemos representado como el plano horizontal, ya no podría asegurarse que $\hat{\epsilon}$ y $(X\hat{\beta} - \vec{y})$ son ortogonales.

Observación 2.6 En la Figura 2.1 puede visualizarse R^2 como el coseno al cuadrado del ángulo que forman los vectores $(\vec{y} - \vec{\bar{y}})$ y $(X\hat{\beta} - \vec{\bar{y}})$. 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 2.7 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 $\vec{\bar{y}}$ (se comprobó en la Observación 2.5, pág. 23). 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, su único regresor (la columna de “unos”) ha de estar entre los regresores del otro.

Observación 2.8 Si ajustamos un modelo sin columna de “unos” podemos encontrarnos con que R^2 definido como en el Teorema 2.4 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 2.9 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.

2.6. 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 vectorial, 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 $\vec{v} \in M^\perp$.*

DEMOSTRACIÓN:

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 .*

DEMOSTRACIÓN:

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 1.14). ■

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 $M \subset H$.

DEMOSTRACIÓN:

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

$$T = \left(\vec{v}_1 \mid \vec{v}_2 \mid \dots \mid \vec{v}_p \right)$$

Siempre podemos completar $\{\vec{v}_1, \dots, \vec{v}_p\}$ con $N - p$ vectores adicionales $\{\vec{v}_{p+1}, \dots, \vec{v}_N\}$ 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:

$$\vec{y} = \underbrace{\sum_{i=1}^p c_i \vec{v}_i}_{\in M} + \underbrace{\sum_{j=p+1}^N c_j \vec{v}_j}_{\in M^\perp}, \quad (2.46)$$

siendo c_i ($i = 1, \dots, N$) las coordenadas de \vec{y} en la base escogida. Premultiplicando ambos lados de (2.46) por \vec{v}_i' ($i = 1, \dots, p$), obtenemos:

$$\vec{v}_i' \vec{y} = \vec{v}_i' \sum_{j=1}^N c_j \vec{v}_j = \sum_{j=1}^N c_j (\vec{v}_i' \vec{v}_j) = c_i, \quad (2.47)$$

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

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

■

Lema 2.4 *La matriz P_M es simétrica idempotente.*

DEMOSTRACIÓN:

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:*

$$R(P_M) = M.$$

DEMOSTRACIÓN:

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

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

■

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 .

DEMOSTRACIÓN:

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 .

DEMOSTRACIÓN:

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

$$DD^-D = D \tag{2.48}$$

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:

$$D\vec{z} = \vec{c} \tag{2.49}$$

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

DEMOSTRACIÓN:

De (2.48) deducimos:

$$DD^{-}D\vec{z} = \vec{c} \quad (2.50)$$

y sustituyendo (2.49) en (2.50):

$$DD^{-}\vec{c} = \vec{c} \quad (2.51)$$

$$D(D^{-}\vec{c}) = \vec{c} \quad (2.52)$$

lo que muestra que $D^{-}\vec{c}$ es solución de (2.49). ■

En realidad, es posible probar un resultado algo más fuerte¹; *toda* solución de (2.49) puede expresarse como $D^{-}\vec{c}$ para alguna elección de D^{-} .

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

DEMOSTRACIÓN:

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 (2.5) en la pág. 15:

$$X'X\hat{\beta} = X'\vec{y} \quad (2.53)$$

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 del sistema (2.51)–(2.52).

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

2.7. Lectura recomendada

Sobre la teoría. Seber (1977), Cap. 3 cubre completamente la materia de este capítulo. Para las cuestiones de álgebra matricial, proyecciones, etc. Draper and Smith (1998) tiene un capítulo completo (el 20) mostrando el problema de la estimación MCO desde un punto de vista geométrico, similar al empleado aquí; Searle (1982), Searle (1971) y Abadir and Magnus (2005) son buenas referencias. Sobre matrices inversas generalizadas, en particular, pueden verse, además de Searle (1982), Ben-Israel and Greville (1974), Rao and Mitra (1971) y Yanai et al. (2011).

Sobre R. Son de utilidad las referencias indicadas en el Capítulo precedente. Específicamente sobre regresión con R, Cornillon and Matzner-Lober (2011) y Faraway (2005). Como se indicó, hay mucha documentación *on line* sobre R, como Venables et al. (1997) (hay traducción castellana, Venables et al. (2000), un poco desfasada), Maindonald (2000) o Kuhnert and Venables (2005); una relación actualizada puede obtenerse en <http://cran.r-project.org/>.

COMPLEMENTOS Y EJERCICIOS

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 (2.39), pág. 22) para mostrar que $SSE/(N-p)$ es un estimador insesgado de σ^2 , compruébese que, dada una muestra aleatoria simple Z_1, \dots, Z_n , el estimador de la varianza

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

no es insesgado.

2.3 Extiéndase el teorema de Gauss-Markov, para probar la afirmación hecha al final de la Sección 2.4 (pág. 21): 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 2.2, pág. 34, no individualiza una única inversa generalizada, salvo cuando D es cuadrada de rango completo. Las siguientes condiciones, la primera de las cuáles coincide con (2.48), proporcionan una única definición de inversa generalizada (la inversa de Moore-Penrose):

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

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 Rao and Mitra (1971)

2.5 (\uparrow 2.4) Cuando la función `lsfit` de R encuentra una matriz de diseño de rango incompleto, proporciona no obstante una solución de $\hat{\beta}$, haciendo un cómputo en esencia equivalente a $\hat{\beta} = (X'X)^-X'\vec{y}$. Podemos llevar a cabo el cálculo de la inversa generalizada de Moore-Penrose mediante la función `ginv` del paquete `MASS` (asociado al libro Venables and Ripley (1999a))

```
> library(MASS)
> XX <- matrix(c(2, 0, 0, 0),
+             2, 2)
> XX
```

```

      [,1] [,2]
[1,]    2    0
[2,]    0    0
> XXig <- ginv(XX)
> XXig
      [,1] [,2]
[1,]  0.5    0
[2,]  0.0    0

```

Observemos que las condiciones que definen a la inversa de Moore-Penrose se verifican.

```

> XX %*% XXig %*% XX
      [,1] [,2]
[1,]    2    0
[2,]    0    0
> XXig %*% XX %*% XXig
      [,1] [,2]
[1,]  0.5    0
[2,]  0.0    0
> XXig %*% XX
      [,1] [,2]
[1,]    1    0
[2,]    0    0
> XX %*% XXig
      [,1] [,2]
[1,]    1    0
[2,]    0    0

```

2.6 (↑ 1.13) Resuélvase el problema 1.13, 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.)

2.7 (↑ 2.6) Escribese una función en R que resuelva el problema 2.6 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.

2.8 Justifíquese la afirmación hecha en la Observación 2.7, pág. 30, de acuerdo con la cual el ajuste, medido en términos de *SSE*, no puede empeorar al añadir regresores.

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

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 queremos forzar tal cosa en nuestra regresión, especialmente si, como sucede en multitud de ocasiones, el origen es arbitrario.

2.10 (↑ 2.1)(↑ 2.9) Pensemos en la siguiente situación: un investigador está interesado en dilucidar si la velocidad de sedimentación de un fluido (y , medida en unidades adecuadas) está influida por la temperatura (X_1 , medida en grados centígrados). Cuenta con las siguientes observaciones:

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

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

```
> y <- c(5.8, 4.7, 4.9, 3.8,
+       2.1)
> X <- c(-10, -6.2, -2.5, 3,
+       4.6)
> ajuste <- lsfit(X, y, intercept = FALSE)
> ajuste$coefficients

X
-0.44798
```

El coeficiente que afecta a la única variable es negativo ($= -0,447984$), lo 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:

```
> y <- c(5.8, 4.7, 4.9, 3.8,
+       2.1)
> X <- c(-10, -6.2, -2.5, 3,
+       4.6)
> X <- (9/5) * X + 32
> ajuste <- lsfit(X, y, intercept = FALSE)
> ajuste$coefficients

      X
0.12265
```

¡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.

```
> y <- c(5.8, 4.7, 4.9, 3.8,
+       2.1)
> X <- c(-10, -6.2, -2.5, 3,
+       4.6)
> ajuste <- lsfit(X, y)
> ajuste$coefficients

Intercept      X
 3.80119 -0.20667

> X <- (9/5) * X + 32
> ajuste <- lsfit(X, y)
> ajuste$coefficients

Intercept      X
 7.47538 -0.11482

> ajuste$coefficients[2] *
+   (9/5)

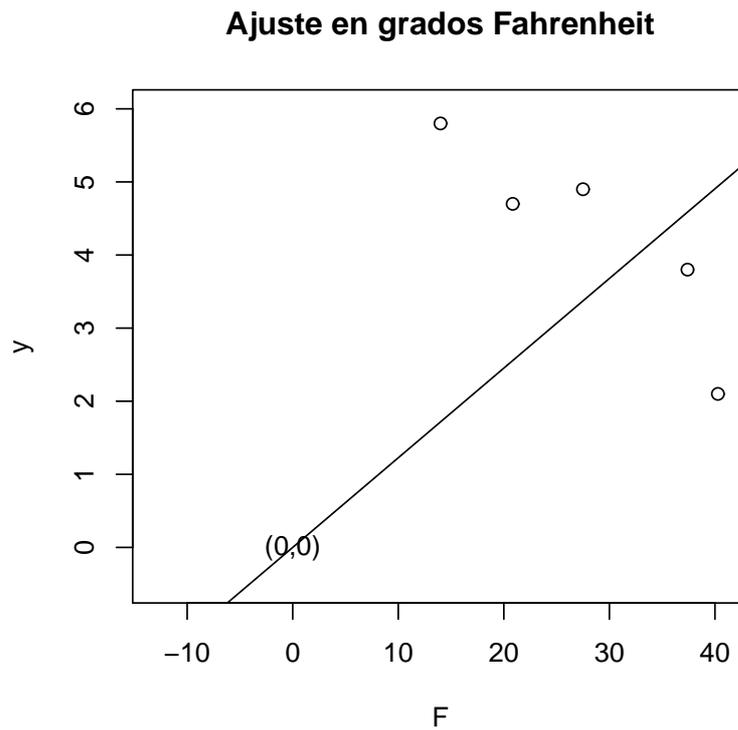
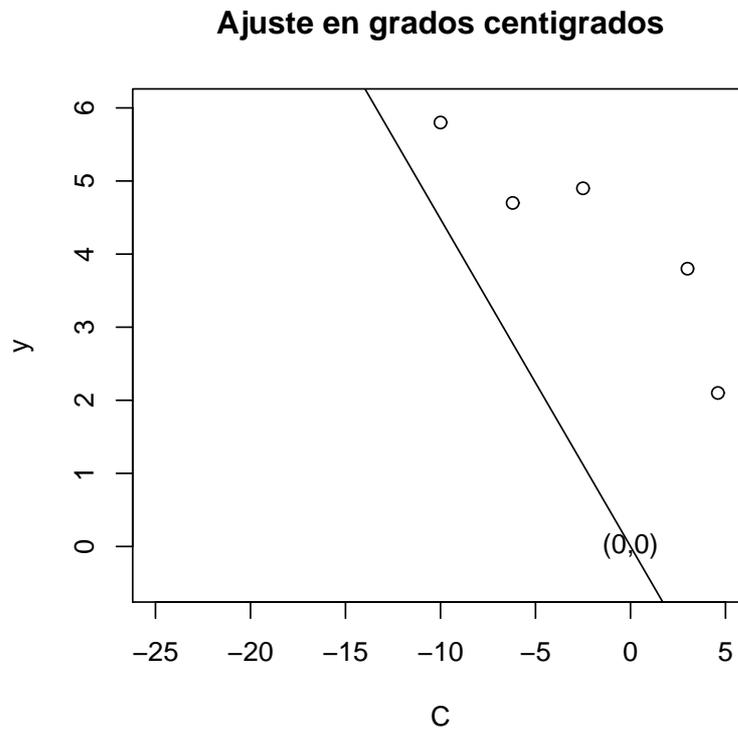
      X
-0.20667
```

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

```
> 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,
+       4.6)
> ajuste <- lsfit(C, y, intercept = FALSE)
> par(xlim = c(-25, 5))
> par(ylim = c(-0.5, 6))
> plot(C, y, ylim = c(-0.5,
+ 6), xlim = c(-25, 5))
> title(main = "Ajuste en grados centigrados")
> abline(a = 0, b = ajuste$coefficients)
> text(x = 0, y = 0, labels = "(0,0)")
> F <- (9/5) * C + 32
> ajuste <- lsfit(F, y, intercept = FALSE)
> plot(F, y, ylim = c(-0.5,
+ 6), xlim = c(-13, 41))
> title(main = "Ajuste en grados Fahrenheit")
> text(x = 0, y = 0, labels = "(0,0)")
> abline(a = 0, b = ajuste$coefficients)
> scratch <- dev.off()
```

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. Colinealidad exacta

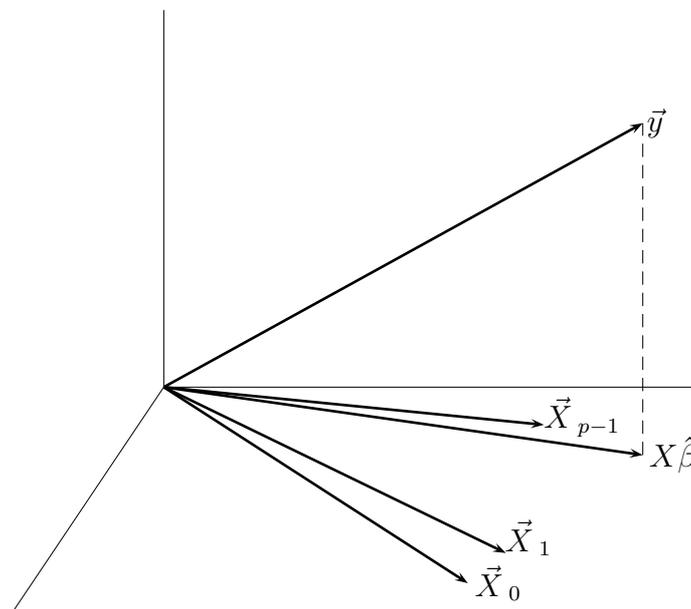
3.1. Modelos con matriz de diseño de rango deficiente.

Uno de los que hemos llamado supuestos habituales (Sección 1.3, pág. 5, apartados 1 a 3) es 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. (Recuérdese que $R(X)$ designa el subespacio generado por las columnas de X .) 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 $\vec{X}_0, \dots, \vec{X}_{p-1}$ que lo generan. La proyección $X\hat{\beta}$ es única. Si $\vec{X}_0, \dots, \vec{X}_{p-1}$ son linealmente independientes, forman base del espacio que generan, y los coeficientes $\hat{\beta}_0, \dots, \hat{\beta}_{p-1}$ 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 $\vec{X}_0, \dots, \vec{X}_{p-1}$ 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 .

Una matriz de diseño de rango deficiente es demasiado “pobre” para deslindar todos los efectos de interés: no podemos con la información disponible deslindar la relación de cada uno de los regresores con la variable respuesta, pero puede ocurrir que si lo podamos deslindar con algunos. El Ejemplo 3.1 a continuación lo ilustra.

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

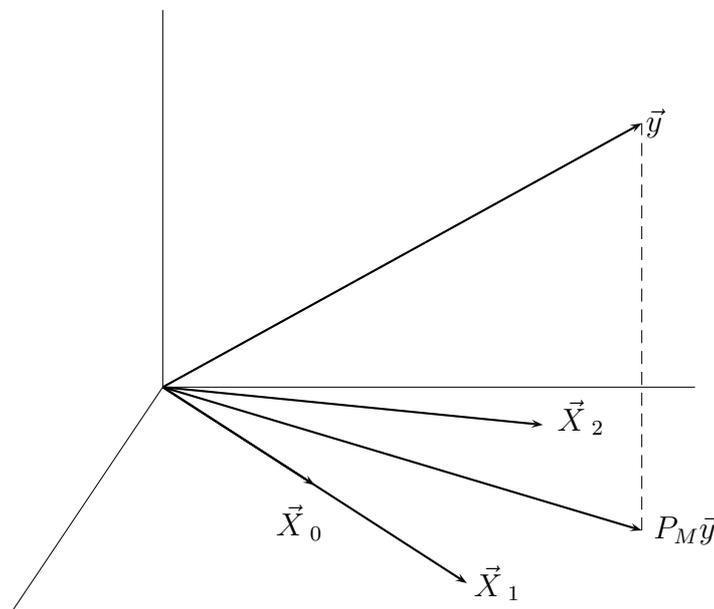
Ejemplo 3.1 Imaginemos una matriz de diseño como

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 5 \\ 2 & 4 & 1 \\ 2 & 4 & 7 \\ 1 & 2 & 8 \\ 1 & 2 & 4 \end{pmatrix}.$$

Observemos que la primera columna, \vec{X}_0 , es igual a la segunda, \vec{X}_1 , dividida entre dos. La Figura 3.2 ilustra una situación similar. Puede verse que \vec{X}_0 y \vec{X}_1 yacen uno sobre otro, difiriendo sólo en el módulo.

En un caso así, la proyección, $P_M \vec{y}$, puede expresarse de manera única como combinación lineal de \vec{X}_2 y *uno* de los vectores \vec{X}_0 ó \vec{X}_1 . Podemos estimar β_2 , pero no β_0 ó β_1 : no es posible adscribir a uno de ellos la “parte” de $P_M \vec{y}$ colineal con la dirección común de \vec{X}_0 y \vec{X}_1 .

FIN DEL EJEMPLO ■

Figura 3.2: Caso de un vector $\vec{\beta}$ parcialmente estimable.

La noción de *función estimable* a continuación permite caracterizar situaciones como la mostrada en el ejemplo anterior.

3.2. Funciones estimables.

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:

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

El Teorema a continuación permite caracterizar las funciones estimables.

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

DEMOSTRACIÓN:

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

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 $\vec{a} \in R(X')$. ■

Observación 3.1 El teorema anterior incluye como caso particular el de parámetros aislados, β_i . En efecto, podemos ver β_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, β_i es estimable si $\vec{e}_i \in R(X')$. La totalidad de los parámetros serán estimables si $\{\vec{e}_1, \dots, \vec{e}_p\}$ (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.

Observación 3.2 El enunciado del Teorema 3.1 tiene gran contenido intuitivo. Son estimables aquéllas combinaciones lineales de los parámetros cuyos coeficientes coinciden con los dados por filas de X . En efecto, si queremos estimar $\vec{a}'\vec{\beta}$ y \vec{a}' coincide con la j -ésima fila \vec{x}_j' de la matriz X , es claro que Y_j sería un estimador insesgado de $\vec{a}'\vec{\beta}$, pues:

$$E[Y_j] = E[\vec{x}_j'\vec{\beta} + \epsilon_j] = E[\vec{a}'\vec{\beta} + \epsilon_j] = \vec{a}'\vec{\beta}.$$

De manera análoga se demuestra que si \vec{a} puede expresarse como combinación lineal de filas de X , la combinación lineal análoga de observaciones en el vector \vec{Y} es un estimador insesgado de $\vec{a}'\vec{\beta}$.

3.3. Restricciones de identificación.

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*.

Ejemplo 3.2 Retomemos el Ejemplo 3.1. Vimos que $\vec{\beta}$ era parcialmente estimable, y que el problema residía en que la componente de $P_M \vec{y}$ colineal con la dirección (común) de \vec{X}_0 y \vec{X}_1 no puede ser “distribuida” entre ambos. Si, no obstante, supiéramos que $\beta_0 = 1$, el problema dejaría de existir. Por tanto, $A\vec{\beta} = 1$ con

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

es una restricción de identificación.

FIN DEL EJEMPLO ■

Una matriz de diseño de rango incompleto se puede presentar por falta de cuidado al diseñar el experimento, pero, más frecuentemente, es intencional. El Ejemplo 3.1 ilustra este punto.

R: Ejemplo 3.1 Supongamos que se investiga el efecto de tres diferentes tratamientos térmicos sobre la dureza de un acero. Podemos pensar en el modelo:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon; \quad (3.4)$$

Habremos de realizar mediciones de la dureza con varias probetas de acero elaborado con los distintos tratamientos, y estimar dicho modelo. La variable explicativa o regresor i -ésimo tomará el valor 1 cuando se emplee el tratamiento i -ésimo, y cero en caso contrario. Con esta especificación β_i , ($i = 1, 2, 3$), se interpretará como la dureza estimada derivada de utilizar el tratamiento i -ésimo. Consideremos los datos siguientes:

```
> cbind(x, y)
      [,1] [,2] [,3] [,4]
[1,]    1    0    0 4.8150
[2,]    1    0    0 4.3619
[3,]    1    0    0 4.3579
[4,]    0    1    0 4.8403
[5,]    0    1    0 5.2419
[6,]    0    1    0 6.2087
[7,]    0    0    1 3.9853
[8,]    0    0    1 4.0601
[9,]    0    0    1 3.4247
```

Podemos estimar los parámetros mediante

```

> ajuste1 <- lsfit(X, y, intercept = FALSE)
> ajuste1$coefficients

      X1      X2      X3
4.5116 5.4303 3.8234

> ajuste1$residuals

[1] 0.30342 -0.14972 -0.15371 -0.58995 -0.18841
[6] 0.77837 0.16193 0.23672 -0.39865

> SSE <- sum(ajuste1$residuals^2)
> SSE

[1] 1.3687

```

Podríamos pensar, sin embargo, en adoptar una diferente parametrización:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon; \quad (3.5)$$

En esta nueva parametrización, β_0 sería una dureza “media” y β_1 a β_3 recogerían el efecto diferencial (respecto de dicha dureza “media”) resultado de emplear cada uno de los tres tratamientos. Para introducir en el modelo β_0 multiplicando a una columna de “unos”, basta omitir el argumento `intercept=FALSE`, con lo que obtenemos:

```

> ajuste2 <- lsfit(X, y, intercept = TRUE)
> ajuste2$coefficients

Intercept      X1      X2      X3
 3.82339  0.68824  1.60690  0.00000

> ajuste2$residuals

[1] 0.30342 -0.14972 -0.15371 -0.58995 -0.18841
[6] 0.77837 0.16193 0.23672 -0.39865

> SSE <- sum(ajuste2$residuals^2)
> SSE

[1] 1.3687

```

Observemos que los dos ajustes son idénticos, como muestran los residuos, que son iguales, y $SSE = 1.3687$, igual en los dos casos; resultado lógico, dado que los subespacios que generan $\vec{X}_1, \dots, \vec{X}_3$ y estos tres vectores más la columna de “unos” son idénticos. Las proyecciones han de serlo también.

En el segundo ajuste, `lsfit` ha proporcionado *una* estimación de los parámetros, a pesar de que el rango de la matriz X ampliada con una columna de “unos” es incompleto. `lsfit` ha tomado una restricción identificadora arbitraria —ha hecho $\beta_3 = 0$ — y proporcionado *una* de las infinitas soluciones equivalentes.

La restricción adoptada hace $\beta_3 = 0$. El tratamiento 3 pasa así a convertirse en *caso de referencia* y la dureza atribuible al mismo viene medida por $\hat{\beta}_0 = 3.8234$. Los valores estimados $\hat{\beta}_1$ y $\hat{\beta}_2$ miden así las diferencias de dureza de los tratamientos 1 y 2 *respecto del caso de referencia, o tratamiento 3*.

Podríamos adoptar restricciones de identificación diferentes. Una muy habitual sería, en el caso que nos ocupa, $\beta_1 + \beta_2 + \beta_3 = 0$. Esto equivale a forzar que los efectos diferenciales de los tres tratamientos no puedan ser todos positivos o negativos. Con esta restricción, β_0 tendría la interpretación de “dureza media” y $\beta_1, \beta_2, \beta_3$ serían desviaciones respecto de esta dureza media.

FIN DEL EJEMPLO ■

3.4. Multicolinealidad exacta y aproximada

La existencia de dependencia lineal “exacta” entre las columnas de la matriz de diseño X , es, como se ha visto, fruto habitualmente de una decisión consciente. Escogemos un diseño de rango incompleto, pero lo suplementamos con restricciones de identificación que solventan el problema de la estimación y dotan a los parámetros de la interpretación que deseamos.

En la medida en que la matriz X sea de nuestra elección, siempre podemos eludir el problema. Si, por el contrario, no podemos diseñar nuestro experimento y nos vemos obligados a utilizar unos datos X, \vec{y} dados, puede ocurrir que la matriz X , aunque no precisamente de rango incompleto, proporcione una matriz $(X'X)$ “casi” singular. Esto se traduce en dificultades numéricas para resolver las ecuaciones normales, dificultades para seleccionar un modelo adecuado, grandes varianzas de los estimadores y otros inconvenientes a los que nos referiremos en el Capítulo 9.

3.5. Lectura recomendada.

Pueden verse Seber (1977), Sección 3.8, o Draper and Smith (1998), Sección 20.4, por ejemplo.

Capítulo 4

Estimación con restricciones

4.1. Planteamiento del problema.

En ocasiones deseamos imponer a las estimaciones de los parámetros $\vec{\beta}$ ciertas condiciones, ya para hacer el modelo interpretable ya porque así lo imponen criterios extra-estadísticos.

Nótese que no nos estamos refiriendo exclusivamente a restricciones de identificación. 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 (como sucedía en el Ejemplo 3.2). 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, sin embargo, 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 4.1 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 ℓ y γ verificaran la condición $\hat{\ell} + \hat{\gamma} = 1$ (rendimientos constantes a escala). Con tres o más observaciones es perfectamente posible estimar α , ℓ y γ ; 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.

FIN DEL EJEMPLO ■

De un modo general, nos planteamos el problema siguiente:

$$\text{mín } \|\vec{y} - X\hat{\beta}\|^2 \quad \text{condicionado a: } A\hat{\beta} = \vec{c} \quad (4.1)$$

Está claro que no podemos esperar obtener la solución de este problema resolviendo un sistema como (3.3), que en general será incompatible.

Hay al menos dos vías para resolver un problema como el indicado. Podemos recurrir a resolver el problema de optimización condicionada (4.1) escribiendo el lagrangiano,

$$\mathcal{L}(\beta_0, \dots, \beta_{p-1}) = \sum_{i=1}^N (y_i - \beta_0 x_{i0} - \dots - \beta_{p-1} x_{i,p-1})^2 - \vec{\lambda}' (A\hat{\beta} - \vec{c});$$

derivando respecto a $\beta_0, \dots, \beta_{p-1}$ y a los multiplicadores de Lagrange en el vector $\vec{\lambda}$, e igualando las derivadas a cero, obtendríamos una solución que mediante las condiciones de segundo orden podríamos comprobar que corresponde a un mínimo.

Resolveremos el problema por un procedimiento diferente, análogo al seguido con el problema incondicionado: 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, de algunos de los cuales nos serviremos repetidamente en lo que sigue.

4.2. Lemas auxiliares.

Lema 4.1 Si $K(C)$ designa el núcleo de la aplicación lineal representada por la matriz C , se tiene:

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

DEMOSTRACIÓN:

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

■

Lema 4.2 Si $h \subseteq M \subseteq H$, y P_h, P_M son las matrices de proyección sobre los subespacios respectivos, se verifica: $P_M P_h = P_h P_M = P_h$

DEMOSTRACIÓN:

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

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

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 4.3 Si $h \subseteq M \subseteq H$, se tiene:

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

DEMOSTRACIÓN:

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,

$$\begin{aligned} \langle P_h \vec{v}, (P_M \vec{v} - P_h \vec{v}) \rangle &= \vec{v}' P_h (P_M \vec{v} - P_h \vec{v}) \\ &= \vec{v}' (P_h P_M - P_h) \vec{v} \\ &= 0, \end{aligned}$$

la última igualdad en virtud del Lema 4.2. Por consiguiente, $(P_M - P_h)$, que es simétrica idempotente, proyecta sobre un subespacio ortogonal a h e incluido en M ; lo denotaremos mediante $M \cap h^\perp$. ■

Lema 4.4 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, $M \cap h^\perp = R(P_M B')$.

La demostración puede hallarse en el Apéndice E.2, pág. 244.

4.3. Estimación condicionada.

Los Lemas anteriores proporcionan todos los elementos para obtener de forma rápida el estimador condicionado que buscamos. (Supondremos X y A de rango completo, pero es fácil generalizar el tratamiento reemplazando las inversas por inversas generalizadas.) Aunque el desarrollo formal es algo farragoso, la idea es muy simple. Vamos a transformar el modelo de modo que las restricciones $A\vec{\beta} = \vec{c}$ se conviertan en $A\vec{\beta} = \vec{0}$.

Lo haremos mediante la transformación

$$\tilde{y} = \vec{y} - X\vec{\delta} \quad (4.2)$$

$$\vec{\gamma} = \vec{\beta} - \vec{\delta}, \quad (4.3)$$

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). Se tiene entonces que:

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

y el problema original (4.1) puede ahora reescribirse así:

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

o, alternativamente,

$$\text{mín } \|\tilde{y} - X\hat{\gamma}\|^2 \quad \text{condicionado a: } A(X'X)^{-1}X'(X\hat{\gamma}) = \vec{0}. \quad (4.4)$$

¿Qué ventajas presenta la expresión (4.4) del problema comparada con la original? Una importante: muestra que el $X\hat{\gamma}$ buscado no es sino la proyección de \tilde{y} sobre un cierto subespacio: $h = M \cap K(A(X'X)^{-1}X')$. *Hay garantía de que h es un subespacio porque M y $K(A(X'X)^{-1}X')$ lo son.* Basta proyectar \tilde{y} sobre h para obtener $X\hat{\gamma}$ y, si X es de rango completo, $\hat{\gamma}$; y esta proyección se puede obtener fácilmente con ayuda de los Lemas anteriores.

Si denotamos por $\hat{\gamma}_h$ las estimaciones mínimo cuadráticas condicionadas o restringidas por $A\hat{\gamma} = \vec{0}$, tenemos que:

$$X\hat{\gamma}_h = P_h\tilde{y} \quad (4.5)$$

$$= (P_M - P_{M \cap h^\perp})\tilde{y} \quad (4.6)$$

$$= [X(X'X)^{-1}X' - P_{M \cap h^\perp}]\tilde{y} \quad (4.7)$$

en que el paso de (4.5) a (4.6) ha hecho uso del Lema 4.3. Pero es que, de acuerdo con el Lema 4.4,

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

Por consiguiente, $P_{M \cap h^\perp}$ es, de acuerdo con el Lema 2.9, pág. 35,

$$P_{M \cap h^\perp} = Z(Z'Z)^{-1}Z', \quad (4.8)$$

ecuación que, llevada a (4.7), proporciona:

$$\begin{aligned} X\hat{\gamma}_h &= X(X'X)^{-1}X'\tilde{y} - X(X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A(X'X)^{-1}X'\tilde{y} \\ &= X\hat{\gamma} - X(X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A\hat{\gamma}, \end{aligned} \quad (4.9)$$

en que $\hat{\gamma}$ es el vector de estimadores mínimo-cuadráticos ordinarios al regresar \tilde{y} sobre X . Si X es de rango total, como venimos suponiendo, de (4.9) se deduce:

$$\hat{\gamma}_h = \hat{\gamma} - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A\hat{\gamma}. \quad (4.10)$$

(véase el Ejercicio 4.3.)

Hay algunas observaciones interesantes que hacer sobre las ecuaciones (4.9) y (4.10). En primer lugar, el lado izquierdo de (4.9) es una proyección. Ello garantiza de manera automática que $\|\tilde{y} - X\hat{\gamma}_h\|^2$ es mínimo¹. 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 (4.10) 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 insesgidez 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,

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

¹Si 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.

que muestra, dado que el segundo sumando tiene claramente elementos no negativos en su diagonal principal (la matriz $(X'X)^{-1}$ 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 parámetros a estimar, *puede introducir algún sesgo*.

Hemos razonado en las líneas anteriores sobre el modelo transformado. Podemos sustituir sin embargo (4.3) en (4.10) y obtener la expresión equivalente en términos de los parámetros originales:

$$\hat{\beta}_h = \hat{\beta} - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - \vec{c}) \quad (4.11)$$

R: Ejemplo 4.1 (*estimación condicionada*)

No hay 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.

Definimos en primer lugar 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 a continuación los datos y realizamos la estimación citándonos a la teoría del modo más directo. **X** es la matriz de diseño, **beta** contiene los parámetros e y la variable respuesta:

```
> X <- matrix(c(1, 1, 1, 1, 1, 1, 1, 4,
+   12, 1, 4, 13, 0, 6, 7, 0, 2, 2), 6,
+   3)
> 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)
> y <- X %*% beta + rnorm(6)

```

Especificamos la restricción lineal $\beta_1 = \beta_2$ tomando la matriz A y vector d siguientes:

```

> A <- matrix(c(0, 1, -1), 1, 3, byrow = TRUE)
> d <- 0

```

y a continuación realizamos la estimación condicionada:

```

> resultado <- lscond(X, y, A = A, d = d,
+   beta0 = FALSE)
> resultado$betas.h

      X1      X2      X3
2.8392 3.2647 3.2647

> resultado$betas

      X1      X2      X3
2.8037 3.0526 3.7138

```

FIN DEL EJEMPLO ■

COMPLEMENTOS Y EJERCICIOS

4.1 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).

4.2 (\uparrow 4.1) Pruébese la igualdad (E.15), pág. 244.

4.3 Justifíquese el paso de (4.9) a (4.10).

4.4  El Ejemplo 4.1 *se sale* del marco conceptual en el que nos movemos. Los regresores (K y L , ó $\log(K)$ y $\log(L)$ al linealizar la función de producción) no pueden ser fijados por el experimentador: dependen de los agentes económicos. Estamos ante *datos observados* en oposición a *datos experimentales*. Faraway (2005), Sec. 3.8, contiene una diáfana discusión de los problemas que ello conlleva. Es también interesante, aunque de más difícil lectura, Wang (1993).

4.5  Las restricciones que hemos discutido en la Sección 4.3 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

$$\begin{pmatrix} \vec{y} \\ k\vec{c} \end{pmatrix} = \begin{pmatrix} X \\ kA \end{pmatrix} \vec{\beta} + \vec{\epsilon}. \quad (4.12)$$

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 SSE afectados de un coeficiente k^2). Cuando $k \rightarrow \infty$ nos acercamos al efecto de restricciones exactas.

4.6 (↑ 4.5) $\hat{\Sigma}$ 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 10.3, pág. 139, abordamos su estudio y justificación con detalle.

4.7 (↑ 4.5) $\hat{\Sigma}$ $\hat{\Sigma}$ La estimación de (4.12) haciendo uso de las ecuaciones normales proporciona

$$\hat{\beta} = (X'X + k^2A'A)^{-1}(X'\vec{y} + k^2A'\vec{c}), \quad (4.13)$$

que admite una interpretación bayesiana. Supongamos que *a priori* $\vec{\beta} \sim N(\vec{\beta}_0, \Sigma_0)$. Dado $\vec{\beta}$, \vec{Y} se distribuye como $N(X\vec{\beta}, \sigma^2I)$. La densidad *a posteriori* de $\vec{\beta}$ es entonces

$$\begin{aligned} f(\vec{\beta}|\vec{y}, \sigma^2, \vec{\beta}_0, \Sigma_0) &\propto \exp\left\{-\frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})'(\vec{y} - X\vec{\beta})\right\} \\ &\quad \times \exp\left\{-\frac{1}{2}(\vec{\beta} - \vec{\beta}_0)'\Sigma_0^{-1}(\vec{\beta} - \vec{\beta}_0)\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left[(\vec{y} - X\vec{\beta})'(\vec{y} - X\vec{\beta})\right.\right. \\ &\quad \left.\left.+ \sigma^2(\vec{\beta} - \vec{\beta}_0)'\Sigma_0^{-1}(\vec{\beta} - \vec{\beta}_0)\right]\right\} \end{aligned}$$

Tomando el logaritmo neperiano e igualando a cero su derivada respecto a $\vec{\beta}$ tenemos entonces

$$-\frac{1}{2\sigma^2}\left[(-2X'(\vec{y} - X\vec{\beta}) + 2\sigma^2\Sigma_0^{-1}(\vec{\beta} - \vec{\beta}_0))\right] = \vec{0},$$

que proporciona

$$(X'X + \sigma^2\Sigma_0^{-1})\vec{\beta} - X'\vec{y} - \sigma^2\Sigma_0^{-1}\vec{\beta}_0 = \vec{0},$$

y por tanto la moda de la distribución *a posteriori* (que fácilmente se comprueba es normal multivariante) es:

$$\hat{\beta} = (X'X + \sigma^2\Sigma_0^{-1})^{-1}(X'\vec{y} + \sigma^2\Sigma_0^{-1}\vec{\beta}_0). \quad (4.14)$$

Comparando (4.14) con (4.13) vemos que son idénticas cuando $kA = \sigma \Sigma_0^{-\frac{1}{2}}$ y $k\vec{c} = \sigma \Sigma_0^{-\frac{1}{2}} \vec{\beta}_0$: para obtener el estimador bayesiano con información *a priori* como la indicada, basta por tanto con obtener el estimador MCO en una muestra ampliada con pseudo-observaciones.

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 Y efectivamente se genera de la siguiente manera:

$$Y = \beta_0 X_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1} + \epsilon. \quad (5.1)$$

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

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

pese a lo cual decidimos estimar el modelo

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

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

Al estimar el modelo sobreparametrizado (5.3) obtendríamos:

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

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 añadidos¹. En efecto, sustituyendo (5.2) en (5.4) tenemos:

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1} \begin{pmatrix} X' \\ Z' \end{pmatrix} \left[\begin{pmatrix} X & Z \end{pmatrix} \begin{pmatrix} \vec{\beta} \\ \vec{0} \end{pmatrix} + \vec{\epsilon} \right] \quad (5.5)$$

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

Al tomar valor medio en la ecuación anterior obtenemos:

$$E[\hat{\beta}] = \vec{\beta}, \quad (5.7)$$

$$E[\hat{\gamma}] = \vec{0}. \quad (5.8)$$

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

$$\Sigma = \sigma^2 \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1}. \quad (5.9)$$

El bloque superior izquierdo de (5.9) es la matriz de covarianzas de los $\hat{\beta}$ obtenidos en el modelo 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).

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

Haciendo uso del Teorema A.3, pág. 221, vemos que el bloque que nos interesa de (5.9) es σ^2 multiplicado por

$$(X'X)^{-1} + (X'X)^{-1}X'Z[Z'Z - Z'X(X'X)^{-1}X'Z]^{-1}Z'X(X'X)^{-1}.$$

Por simple inspección vemos que el segundo sumando es una matriz definida no negativa², y por tanto la expresión anterior tendrá en su diagonal principal elementos no menores que los de la diagonal principal de $(X'X)^{-1}$. En consecuencia, 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

$$\left(\vec{Y} - (X \ Z) \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} \right) \quad (5.10)$$

es un vector aleatorio de media cero. Denominando,

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

un desarrollo enteramente similar al que realizaremos en el Teorema 6.1, pág. 68, muestra que en el modelo sobreparametrizado

$$SSE = \vec{Y}'(I - L(L'L)^{-1}L')\vec{Y} = \vec{\epsilon}'(I - L(L'L)^{-1}L')\vec{\epsilon} \quad (5.11)$$

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

$$\hat{\sigma}^2 = \frac{SSE}{N - (p + q)} \quad (5.12)$$

es un estimador insesgado de σ^2 . (Recuérdese que el valor medio de una v.a. con distribución χ_k^2 es k , el número de grados de libertad.) El único efecto adverso de la inclusión de los q regresores irrelevantes ha sido la pérdida de otros tantos grados de libertad.

²Llamemos 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).

5.3. Omisión de regresores relevantes.

Sea $X = (X_1 : X_2)$ una matriz de diseño particionada en sendos bloques de p y r columnas. Sea $\vec{\beta}' = (\vec{\beta}'_1 : \vec{\beta}'_2)$ el correspondiente vector de $p + r$ parámetros. Consideremos el caso en que el modelo “correcto” es

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon} = X_1\vec{\beta}_1 + X_2\vec{\beta}_2 + \vec{\epsilon}, \quad (5.13)$$

pese a lo cual estimamos el modelo “escaso”

$$\vec{Y} = X_1\vec{\beta}_1 + \vec{\epsilon}. \quad (5.14)$$

Estimar (5.14) es lo mismo que estimar (5.13) junto con las restricciones $h : \vec{\beta}_2 = \vec{0}$, expresables así:

$$\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \vec{\beta}_1 \\ \vec{\beta}_2 \end{pmatrix} = \begin{pmatrix} \vec{0} \\ \vec{0} \end{pmatrix} \quad (5.15)$$

En consecuencia, podemos deducir cuanto necesitamos saber haciendo uso de los resultados en la Sección 4.3. Las siguientes conclusiones son así inmediatas:

- El estimador $\hat{\beta}_1^{(h)}$ obtenido en el modelo “escaso” (5.14) es, en general, sesgado. El sesgo puede obtenerse haciendo uso de (4.11). Tenemos así que

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

y en consecuencia

$$E[\hat{\beta}_1^{(h)} - \vec{\beta}_1] = - \left[(X'X)^{-1}A'[A(X'X)^{-1}A']^{-1} \begin{pmatrix} \vec{0} \\ \vec{\beta}_2 \end{pmatrix} \right]_{(p \times 1)} \quad (5.16)$$

en que $[M]_{(p \times q)}$ designa el bloque superior izquierdo con p filas y q columnas de la matriz M . La ecuación (5.16) muestra que el sesgo introducido depende de la magnitud de los parámetros asociados a los regresores omitidos.

- La ecuación (5.16) 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, $X'_1X_2 = 0$, la matrix $(X'X)^{-1}$ es diagonal por bloques, y

$$(X'X)^{-1}A' = \begin{pmatrix} X'_1X_1 & 0 \\ 0 & X'_2X_2 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} \quad (5.17)$$

tiene sus primeras p filas de ceros. Ello hace que el bloque considerado en (5.16) esté formado por ceros.

- El estimador de la varianza de la perturbación

$$\hat{\sigma}^2 = \frac{SSE}{N - p} = \frac{(\vec{Y} - X_1 \hat{\beta}_1^{(h)})'(\vec{Y} - X_1 \hat{\beta}_1^{(h)})}{N - p} \quad (5.18)$$

no es insesgado. En efecto, puede verse que no es de aplicación a (5.18) el Teorema 2.3, pág. 21, porque los residuos no tiene media cero.

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 menos graves, 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 sólo 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 9.

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” (5.14) 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 12. Algunas técnicas de análisis gráfico de residuos que pueden ser de ayuda en la especificación de modelos se consideran en la Sección 13.2.

Capítulo 6

Regresión con perturbaciones normales.

6.1. Introducción.

Si a los supuestos habituales (Sección 1.3, pág. 5) añadimos¹ 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 6.1 Si $\vec{u} \sim N(\vec{0}, \sigma^2 I)$ y A es una matriz simétrica idempotente de orden n y rango r , entonces: $\frac{\vec{u}' A \vec{u}}{\sigma^2} \sim \chi_r^2$.

DEMOSTRACIÓN:

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

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

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 $n \times n$.

¹El 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,

$$\frac{\vec{u}' A \vec{u}}{\sigma^2} = \frac{\vec{v}' D' A D \vec{v}}{\sigma^2} = \frac{\vec{v}'}{\sigma} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \frac{\vec{v}}{\sigma} = \sum_{i=1}^r \frac{v_i^2}{\sigma^2}. \quad (6.1)$$

Pero el lado derecho de (6.1) es una suma de cuadrados de r variables aleatorias $N(0, 1)$ independientes, y por tanto sigue una distribución² χ_r^2 . ■

Lema 6.2 *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.*

DEMOSTRACIÓN:

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:

$$BP = 0 \Rightarrow D' B D D' P D = 0 \quad (6.2)$$

$$(6.3)$$

$$\Rightarrow D' B D \begin{pmatrix} r & (n-r) \\ I & 0 \\ 0 & 0 \end{pmatrix} = 0 \quad (6.4)$$

$$(6.5)$$

$$\Rightarrow D' B D \text{ tiene sus } r \text{ primeras columnas nulas} \quad (6.6)$$

Por tanto:

$$D' B D = \begin{pmatrix} r & (n-r) \\ 0 & L_{12} \\ (n-r) & L_{22} \end{pmatrix} = 0 \quad (6.7)$$

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

$$\vec{u}' B \vec{u} = \vec{v}' D' B D \vec{v} = \vec{v}' \begin{pmatrix} r & (n-r) \\ 0 & 0 \\ 0 & L_{22} \end{pmatrix} \vec{v} \quad (6.8)$$

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

Por otra parte:

$$\vec{u}'P\vec{u} = \vec{v}'D'PD\vec{v} = \vec{v}' \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \vec{v} \quad (6.9)$$

De (6.8) y (6.9) se deduce que ambas formas cuadráticas consideradas dependen de distintas componentes del vector \vec{v} , y son por tanto independientes. ■

Lema 6.3 *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.*

DEMOSTRACIÓN:

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

$$AM = ADD'MD = 0 \Rightarrow AD = \begin{pmatrix} r & (n - r) \\ 0 & L_2 \end{pmatrix}, \quad (6.10)$$

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

$$A\vec{u} = AD\vec{v} = \begin{pmatrix} r & (n - r) \\ 0 & L_2 \end{pmatrix} \vec{v}. \quad (6.11)$$

Como

$$\vec{u}'M\vec{u} = \vec{v}'D'MD\vec{v} = \vec{v}' \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \vec{v}, \quad (6.12)$$

deducimos de (6.11) y (6.12) 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 6.1 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. $(\hat{\beta} - \vec{\beta})'(X'X)(\hat{\beta} - \vec{\beta}) \sim \sigma^2\chi_p^2$
3. $(N - p)\hat{\sigma}^2 = SSE \sim \sigma^2\chi_{N-p}^2$
4. $\hat{\beta}$ y $\hat{\sigma}^2$ son variables aleatorias independientes.

DEMOSTRACIÓN:

El apartado 1) es inmediato. Si se verifican los supuestos habituales, fue ya demostrado (Teorema 2.2, pág. 19) 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 e independientes, es también normal.

El apartado 2) es consecuencia inmediata del Lema 6.1, una vez que observamos que $(X'X)^{\frac{1}{2}}(\hat{\beta} - \vec{\beta}) \sim N(\vec{0}, \sigma^2 I)$.

Para demostrar el apartado 3) observemos que:

$$\frac{SSE}{\sigma^2} = \frac{(\vec{Y} - X\hat{\beta})'(\vec{Y} - X\hat{\beta})}{\sigma^2} \quad (6.13)$$

$$= \frac{(\vec{Y} - X(X'X)^{-1}X'\vec{Y})'(\vec{Y} - X(X'X)^{-1}X'\vec{Y})}{\sigma^2} \quad (6.14)$$

$$= \frac{\vec{Y}'[I - X(X'X)^{-1}X']\vec{Y}}{\sigma^2} \quad (6.15)$$

$$= \frac{(X\vec{\beta} + \vec{\epsilon})'[I - X(X'X)^{-1}X'](X\vec{\beta} + \vec{\epsilon})}{\sigma^2} \quad (6.16)$$

$$= \frac{\vec{\epsilon}'[I - X(X'X)^{-1}X']\vec{\epsilon}}{\sigma^2} \quad (6.17)$$

$$= \frac{\vec{\epsilon}'M\vec{\epsilon}}{\sigma^2} \quad (6.18)$$

$$\sim \chi_{N-p}^2, \quad (6.19)$$

donde (6.19) es consecuencia inmediata del Lema 6.1, ya que M es simétrica idempotente y de rango $N - p$.

Para probar 4), basta invocar el Lema 6.3, ya que

$$\hat{\beta} = (X'X)^{-1}X'\vec{Y}, \quad (6.20)$$

$$\hat{\sigma}^2 = \frac{SSE}{N - p} = \frac{\vec{Y}'[I - X(X'X)^{-1}X']\vec{Y}}{N - p}. \quad (6.21)$$

De la ecuación (6.20) deducimos (sustituyendo \vec{Y} por $X\vec{\beta} + \vec{\epsilon}$) que $\hat{\beta} = \vec{\beta} + (X'X)^{-1}X'\vec{\epsilon}$. La misma sustitución en (6.21) 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 6.3, pág. 67, demuestra la independencia de las formas lineal y cuadrática anteriores y por tanto de (6.20) y (6.21). ■

R: Ejemplo 6.1 (*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.

Generemos en primer lugar la matriz de diseño X , vector de parámetros $\vec{\beta}$ y los valores medios de la respuesta $X\vec{\beta}$:

```
> X <- matrix(c(1, 1, 1, 1, 1, 1, 9, 4,
+             12, 1, 4, 13, 0, 6, 7, 0, 2, 2), 6,
+             3)
> 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)
> Ey <- X %*% beta
```

Definiremos ahora una matriz \mathbf{b} de dimensiones 100×3 , cada una de cuyas filas guardará los parámetros estimados $\hat{\beta}$ con una muestra artificial diferente

CAPÍTULO 6. REGRESIÓN CON PERTURBACIONES NORMALES 70

```
> muestras <- 100
> b <- matrix(0, muestras, 3)
```

e iteremos, generando en cada pasada del bucle `for` un nuevo vector de perturbaciones $\hat{\epsilon}$ (mediante `rnorm`), un nuevo vector de valores de la variable respuesta \vec{y} y nuevas estimaciones $\hat{\beta}$ de los parámetros $\vec{\beta}$ (`fit$coefficients`, que se almacenan en `b[i,]`):

```
> for (i in 1:muestras) {
+   y <- Ey + rnorm(6)
+   fit <- lsfit(X, y, intercept = FALSE)
+   b[i, ] <- fit$coefficients
+ }
```

La distribución teórica de los betas es Normal, con vector de medias $(2, 3, 4)'$ y matriz de covarianzas $(X'X)^{-1}$ (la varianza de las perturbaciones generadas por `rnorm` es 1 si no se especifica otra cosa).

```
> cov.betas <- solve(t(X) %*% X)
```

Por consiguiente, un modo de verificar que los resultados empíricos son congruentes con la teoría consistiría en tipificar las estimaciones de los parámetros y comparar su distribución con una $N(0, 1)$. Podemos por ejemplo comparar la media y varianza empíricas con las teóricas,

```
> beta1.tipif <- (b[, 1] - beta[1])/sqrt(cov.betas[1,
+   1])
> mean(beta1.tipif)
```

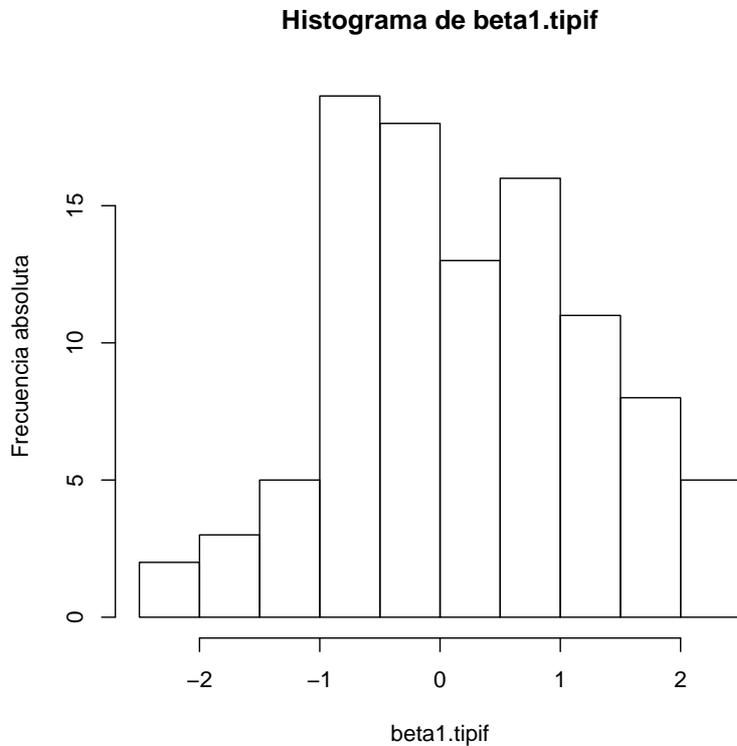
```
[1] 0.19871
```

```
> var(beta1.tipif)
```

```
[1] 1.1125
```

dibujar el histograma

```
> hist(beta1.tipif, ylab = "Frecuencia absoluta",
+   main = "Histograma de beta1.tipif")
```



o llevar a cabo algún contraste de normalidad especializado:

```
> ks.test(beta1.tipif, "pnorm")
      One-sample Kolmogorov-Smirnov test

data:  beta1.tipif
D = 0.1036, p-value = 0.2334
alternative hypothesis: two-sided
> shapiro.test(beta1.tipif)
      Shapiro-Wilk normality test
```

```
data:  beta1.tipif
W = 0.9874, p-value = 0.4679
```

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 permite ver si la aproximación asintótica es aceptable para un cierto tamaño muestral.
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. (R ofrece generadores de números aleatorios de las distribuciones más usuales, como casi cualquier otro paquete estadístico.)

FIN DEL EJEMPLO ■

6.2. 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

$$h: A\vec{\beta} = \vec{c} \quad (\text{rango de } A = q < p), \quad (6.22)$$

siendo A de dimensiones $q \times p$. Cualquier hipótesis lineal sobre los parámetros se puede expresar en la forma (6.22). 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 6.1 Llamamos hipótesis lineales a las que pueden expresarse del modo (6.22); 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 6.2 *Sea $h: A\vec{\beta} = \vec{c}$ 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 $h: A\vec{\beta} = \vec{c}$ 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 .

DEMOSTRACIÓN:

$$SSE_h - SSE = \|\vec{Y} - X\hat{\beta}_h\|^2 - \|\vec{Y} - X\hat{\beta}\|^2 \quad (6.23)$$

$$= \|\vec{Y} - X\hat{\beta} + X\hat{\beta} - X\hat{\beta}_h\|^2 - \|\vec{Y} - X\hat{\beta}\|^2 \quad (6.24)$$

$$= \|\vec{Y} - X\hat{\beta}\|^2 + \|X\hat{\beta} - X\hat{\beta}_h\|^2 - 2\langle \vec{Y} - X\hat{\beta}, X\hat{\beta} - X\hat{\beta}_h \rangle \quad (6.25)$$

$$= \|X\hat{\beta} - X\hat{\beta}_h\|^2 \quad (6.26)$$

$$= (\hat{\beta} - \hat{\beta}_h)'(X'X)(\hat{\beta} - \hat{\beta}_h). \quad (6.27)$$

Se ha hecho uso en el paso de (6.25) a (6.26) de que $\hat{\epsilon}$ es ortogonal a toda combinación lineal de las columnas de X , lo que garantiza la nulidad del producto interno en (6.25).

Haciendo uso de la ecuación (4.11), pág. 55, la expresión (6.27) se convierte en:

$$SSE_h - SSE = (A\hat{\beta} - \vec{c})'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - \vec{c}). \quad (6.28)$$

Esto finaliza la demostración del primer apartado. Por otra parte, como

$$\hat{\beta} = \vec{\beta} + (X'X)^{-1}X'\vec{\epsilon},$$

tenemos que, cuando se verifica la hipótesis h ,

$$(A\hat{\beta} - \vec{c}) = (A\hat{\beta} - A\vec{\beta}) = A(X'X)^{-1}X'\vec{\epsilon},$$

resultado que llevado a (6.28) proporciona:

$$SSE_h - SSE \stackrel{h}{=} \vec{\epsilon}' \underbrace{X(X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A(X'X)^{-1}X'}_G \vec{\epsilon} \quad (6.29)$$

Esta expresión muestra que $SSE_h - SSE$ es una forma cuadrática en variables normales (las $\vec{\epsilon}$) de matriz G que fácilmente comprobamos es idempotente. Por tanto, según el Lema 6.1, pág. 65, $SSE_h - SSE$ sigue una distribución $\sigma^2\chi_q^2$, con grados de libertad q iguales al rango de G ($= \text{rango}(A)$). Tenemos además (Teorema 6.1) que:

$$SSE = \vec{Y}'(I - P_M)\vec{Y} \sim \sigma^2\chi_{N-p}^2 \quad (6.30)$$

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

$$(I - P_M) \underbrace{X(X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}A(X'X)^{-1}X'}_G = 0.$$

El Lema 6.2 garantiza por tanto la independencia. ■

Observación 6.2 Hay cuestiones de interés sobre el Teorema 6.2. En primer lugar, es claro que, para un nivel de significación α , la región crítica estará formada por valores mayores que $\mathcal{F}_{q, N-p}^\alpha$. 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 6.2 muestra que el estadístico tiene en su numerador una forma cuadrática que crece al separarse $A\hat{\beta}$ de \vec{c} .

Observación 6.3 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:

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

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

Observación 6.4 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.

R: Ejemplo 6.2 (*contraste de una hipótesis lineal*)

Veamos el modo en que contrastaríamos una hipótesis lineal general sobre los parámetros de un modelo de regresión lineal. Nos serviremos de la función `lscond` para realizar estimación condicional presentada en el Ejemplo 4.1, pág. 55.

```
> 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))
+ }
```

Definiremos ahora una nueva función, `contraste.h`, que calcula SSE , SSE_h (utilizando `lscond`), el estadístico Q_h y su nivel de significación.

```
> contraste.h <- function(X, y, A, d, beta0 = TRUE) {
+   lscond.result <- lscond(X, y, A, d,
+     beta0 = beta0)
+   betas <- lscond.result$betas
```

CAPÍTULO 6. REGRESIÓN CON PERTURBACIONES NORMALES 76

```

+   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)
+   denom <- SSE/(nrow(X) - ncol(X))
+   Qh <- numer/denom
+   p.value <- 1 - pf(Qh, nrow(A), nrow(X) -
+     ncol(X))
+   return(list(Qh = Qh, p.value = p.value))
+ }

```

Generemos datos artificiales:

```

> X <- matrix(c(1, 1, 1, 1, 1, 1, 1, 4,
+   12, 1, 4, 13, 0, 6, 7, 0, 2, 2), 6,
+   3)
> 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)
> y <- X %*% beta + rnorm(6)

```

“Sabemos”, porque los datos han sido artificialmente generados, que $\beta_1 = 3$ y $\beta_2 = 4$. Probaremos a continuación a contrastar la hipótesis $\beta_1 = \beta_2$, que debiera ser rechazada. La matriz A y vector \vec{d} especificando dicha hipótesis pueden construirse así:

```

> A <- matrix(c(0, 1, -1), 1, 3, byrow = TRUE)
> d <- 0

```

El contraste puede entonces llevarse a cabo así:

```

> result <- contraste.h(X, y, A = A, d = d,
+   beta0 = FALSE)
> result$Qh

```

CAPÍTULO 6. REGRESIÓN CON PERTURBACIONES NORMALES 77

```
[1] 161.11
> result$p.value
[1] 0.0010548
```

Rechazaríamos por consiguiente la hipótesis contrastada para cualquier nivel de significación $\alpha > 0.0010548$.

Frecuentemente podemos obtener las sumas de cuadrados requeridas para el contraste de hipótesis de interés de manera más simple. En el caso que nos ocupa, si realmente $\beta_1 = \beta_2$,

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \quad (6.31)$$

es equivalente a

$$Y = \beta_0 X_0 + \beta_1 (X_1 + X_2) + \epsilon \quad (6.32)$$

y las sumas de cuadrados SSE y SSE_h podrían obtenerse así:

```
> SSE <- sum(lsfite(X, y)$residuals^2)
> Xmod <- cbind(X[, 1], X[, 2] + X[, 3])
> SSE.h <- sum(lsfite(Xmod, y)$residuals^2)
> Qh <- ((SSE.h - SSE)/1)/(SSE/(nrow(X) -
+      ncol(X)))
```

Puede verse que el valor de Q_h así calculado es idéntico al obtenido más arriba:

```
> Qh
[1] 161.11
```

Esta técnica de calcular las sumas de cuadrados SSE y SSE_h en dos regresiones *ad-hoc* puede ser muy frecuentemente utilizada. En el caso frecuente de hipótesis de exclusión (alguno o varios betas iguales a cero), puede obtenerse SSE_h de una regresión en que los regresores correspondientes están ausentes. Si en nuestro ejemplo quisiéramos contrastar $h : \beta_1 = \beta_2 = 0$, podríamos obtener SSE de la regresión (6.31) y SSE_h de la regresión

$$Y = \beta_0 X_0 + \epsilon,$$

para calcular el estadístico Q_h así:

```
> SSE <- sum(lsfrit(X, y)$residuals^2)
> SSE.h <- sum(lsfrit(X[, 1], y)$residuals^2)
> Qh <- ((SSE.h - SSE)/2)/(SSE/(nrow(X) -
+      ncol(X)))
> Qh

[1] 16956
```

El valor que dicho estadístico Q_h deja en a su derecha en la distribución de referencia,

```
> 1 - pf(Qh, 2, nrow(X) - ncol(X))

[1] 8.3193e-07
```

permite rechazar contundentemente la hipótesis $h : \beta_1 = \beta_2 = 0$ contrastada.

FIN DEL EJEMPLO ■

Contraste sobre coeficientes β_i aislados.

El Teorema 6.2 permite obtener como casos particulares multitud de contrastes frecuentemente utilizados. Por ejemplo, la hipótesis $h: \beta_{i-1} = 0$ 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 (recuérdese que los parámetros β se numeran a partir de β_0). En tal caso, Q_h puede escribirse así:

$$Q_h = \frac{(\hat{\beta}_{i-1} - 0)'[(X'X)_{ii}^{-1}]^{-1}(\hat{\beta}_{i-1} - 0)}{\hat{\sigma}^2} \quad (6.33)$$

donde $(X'X)_{ii}^{-1} = [A(X'X)^{-1}A']$ designa el elemento en la posición i -ésima de la diagonal principal de $(X'X)^{-1}$. Bajo la hipótesis h , (6.33) sigue una distribución $\mathcal{F}_{1, N-p}$, y como $\hat{\sigma}^2(X'X)_{ii}^{-1} = \hat{\sigma}_{\hat{\beta}_{i-1}}^2$ tenemos que:

$$\sqrt{Q_h} = \frac{\hat{\beta}_{i-1}}{\hat{\sigma}_{\hat{\beta}_{i-1}}} \sim \sqrt{\mathcal{F}_{1, N-p}} \sim t_{N-p} \quad (6.34)$$

La regla de decisión que se deduce de (6.34) es:

Rechazar $h: \beta_{i-1} = 0$ al nivel de significación α si

$$\left| \frac{\hat{\beta}_{i-1}}{\hat{\sigma}_{\hat{\beta}_{i-1}}} \right| > t_{N-p}^{\alpha/2}.$$

El estadístico $|\hat{\beta}_{i-1}/\hat{\sigma}_{\hat{\beta}_{i-1}}|$ recibe el nombre de *estadístico t* o *t-ratio*. De forma análoga se contrasta la hipótesis $h: \beta_{i-1} = c$.

Contraste de significación conjunta de la regresión.

Otra hipótesis frecuentemente de interés es: $h: \beta_1 = \dots = \beta_{p-1} = 0$ —es decir, nulidad de todos los parámetros, salvo el correspondiente a la columna de “unos”, β_0 —. En este caso,

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

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

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

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

$$\vec{c}' = (0 \quad 0 \quad \dots \quad 0)$$

Pero SSE_h en este caso particular es lo que hemos definido (Teorema 2.4, pág. 28) 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 2.4, pág. 29. El contraste de h requiere solamente conocer R^2 . Cuando h es cierta, Q_h se distribuye como una $\mathcal{F}_{p-1, N-p}$.

6.3. 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 6.3 *Se verifica lo siguiente:*

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

DEMOSTRACIÓN:

El apartado 1) se sigue inmediatamente de las ecuaciones (6.35) y (6.36) a continuación, consecuencia la primera de los supuestos habituales, y la segunda de la insesganza de $\hat{\beta}$ (Teorema 2.2, pág. 19).

$$E(Y_*) = E(\vec{x}_*' \vec{\beta} + \epsilon_*) = \vec{x}_*' \vec{\beta} \quad (6.35)$$

$$E(\hat{Y}_*) = E(\vec{x}_*' \hat{\beta}) = \vec{x}_*' \vec{\beta} \quad (6.36)$$

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

$$E(Y_* - \hat{Y}_*)^2 = E[\vec{x}_*' \vec{\beta} + \epsilon_* - \vec{x}_*' \hat{\beta}]^2 \quad (6.37)$$

$$= E[\vec{x}_*' (\vec{\beta} - \hat{\beta}) + \epsilon_*]^2 \quad (6.38)$$

$$= E[\vec{x}_*' (\vec{\beta} - \hat{\beta})]^2 + E[\epsilon_*]^2 \quad (6.39)$$

$$= E[\vec{x}_*' (\vec{\beta} - \hat{\beta}) (\vec{\beta} - \hat{\beta})' \vec{x}_*] + E[\epsilon_*]^2 \quad (6.40)$$

$$= \vec{x}_*' \Sigma_{\hat{\beta}} \vec{x}_* + \sigma^2 \quad (6.41)$$

$$= \vec{x}_*' \sigma^2 (X'X)^{-1} \vec{x}_* + \sigma^2 \quad (6.42)$$

$$= \sigma^2 [1 + \vec{x}_*' (X'X)^{-1} \vec{x}_*] \quad (6.43)$$

En el paso de (6.38) a (6.39) se ha hecho uso de la circunstancia de que $\hat{\beta}$ y ϵ_* son independientes ($\hat{\beta}$ depende solamente de $\vec{\epsilon}$, y ϵ_* es perturbación de una observación adicional, distinta de las que han servido para estimar $\hat{\beta}$ e independiente de ellas). ■

El examen de (6.43) 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: ϵ_* 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, (6.43) muestra que la varianza del error de predicción *depende de* \vec{x}_*' . Habrá determinadas Y_* cuya predicción será más precisa que la de otras. En el Capítulo 9 volveremos sobre el particular.

6.4. Lectura recomendada.

Sobre la teoría. Pueden ser consultados los manuales repetidamente citados: Seber (1977), Cap. 4, Draper and Smith (1998) Cap. 8, Stapleton (1995) Sec. 3.8, Peña (2002) Sec. 7.7 son unos cuantos.

Sobre generadores de números aleatorios, pueden consultarse Knuth (1968), Kennedy (1980), Lange (1998), Thisted (1988) y, en general, cualquier texto sobre computación estadística.

Sobre el contraste razón generalizada de verosimilitudes, puede verse Cox and Hinkley (1974) p. 313 y para su aplicación al contraste de hipótesis lineales generales, Stapleton (1995) Sec. 3.8.

Sobre la utilización de R. En el Ejemplo 4.1, pág. 55 y siguientes, se han definido las funciones `lscond` y `contraste.h` por motivos didácticos. En R hay funciones en varios paquetes que proporcionan análoga funcionalidad. Puede consultarse por ejemplo la documentación de `linear.hypothesis` (paquete `car`) y `glh.test` (paquete `gmodels`).

Por lo que hace a intervalos de confianza, que también pueden obtenerse fácilmente de acuerdo con la teoría esbozada en la Sección 6.3, puede ser de utilidad la función `confint` (paquete `stats`).

El empleo de dichas funciones, sin embargo, presupone familiaridad con la función `lm`, que es objeto de atención en el Capítulo 7 a continuación.

COMPLEMENTOS Y EJERCICIOS

6.1 Demuéstrese que si G es la matriz definida en (6.29) con A y $(X'X)$ ambas de rango completo, entonces $\text{rango}(G) = \text{rango}(A)$.

Capítulo 7

Estimación del modelo de regresión lineal con R.

En los capítulos anteriores han aparecido fragmentos de código ilustrando el modo de llevar a cabo diversos cálculos en R. Se presenta aquí la función `lm` y algunas otras, para ilustrar tanto los conceptos teóricos adquiridos como la potencia del entorno de modelización proporcionado por R.

Este capítulo es eminentemente práctico y puede ser omitido sin pérdida de continuidad por lectores que no estén interesados en utilizar R como herramienta de cálculo.

7.1. Tipología de variables explicativas.

Interesará distinguir dos tipos de variables: *cualitativas* (también llamadas categóricas) y *numéricas*. Las variables cualitativas se desglosan a su vez en *nominales* y *ordinales*.

Una variable cualitativa nominal especifica una característica o atributo que puede tomar un número entero (y habitualmente pequeño) de *niveles* o estados. Por ejemplo, una variable ZONA podría tomar los niveles o estados: “Europa”, “Africa”, “Asia”, “America” y “Oceanía”. Requeriremos que las categorías sean exhaustivas, de forma que todo caso muestral pueda recibir un valor. Si es preciso, podemos crear una categoría especial como “Otros” o “Resto”.

Una variable cualitativa ordinal se diferencia únicamente de una nominal en que hay una ordenación natural entre las categorías. Por ejemplo, en una variable como NIVEL DE ESTUDIOS podríamos tener categorías como: “Sin estudios”, “Primarios”, “Secundarios”, “Superiores”. La diferencia

esencial con las variables nominales es que hay una ordenación entre los distintos niveles: cada una de las categorías en el orden en que se hay escrito implica “más” estudios que la categoría precedente. No había, en cambio, en el ejemplo anterior una ordenación natural entre las zonas geográficas.

Las variables que hemos denominado *numéricas* pueden en principio ponerse en correspondencia con un intervalo de números reales. Sería el caso de variables como PESO ó TEMPERATURA (aunque en la práctica el número de estados que pueden tomar es finito a causa de la precisión también finita de los instrumentos de medida que empleamos).

En cierto sentido, los tres tipos de variables, en el orden en que se han descrito, reflejan una mayor finura o contenido informativo: una variable numérica puede convertirse en ordinal fijando intervalos: por ejemplo, TEMPERATURA podría convertirse en una variable ordinal con niveles “Frío”, “Templado” y “Caliente”, al precio de un cierto sacrificio de información: dos temperaturas de, por ejemplo, 80C y 93C podrían ambas convertirse en “Caliente”, perdiéndose la información de que la segunda es superior a la primera.

Análogamente, una variable ordinal puede tratarse como nominal, haciendo abstracción de su orden, también al precio de sacrificar cierta información.

Observación 7.1 En general, no interesará “degradar” una variable tratándola como un tipo inferior, aunque en algunos casos, puede convenirnos hacerlo. Por ejemplo, si examinamos la influencia de la renta sobre el consumo de un cierto bien en una muestra de familias, medir la renta en euros da al coeficiente β asociado la interpretación de “Incremento de consumo asociado a un incremento de renta de un euro”. Típicamente, tendrá un valor muy pequeño. Además, el suponer una dependencia lineal del consumo sobre la renta será en la mayoría de los casos poco realista. En tal caso, podría convenirnos redefinir la variable renta en categorías. Los coeficientes estimados serán más fácilmente interpretables, y tendremos un modelo más flexible, que no fuerza una relación lineal entre renta y consumo. (Adicionalmente, si la variable se obtiene por encuesta, los sujetos podrían ser más veraces al encuadrarse en intervalos amplios de renta que al responder directamente sobre su valor.)

7.2. Factores y *dataframes*.

R ofrece excelentes facilidades para tratar variables de diferentes tipos como regresores. En la jerga de R, una variable cualitativa se denomina *factor*.

Hay factores ordinarios, que permiten manejar variables cualitativas nominales, y factores ordenados (*ordered factors*), para variables cualitativas ordinales. El Ejemplo 7.1 a continuación ilustra la manera de operar con ellos.

R: Ejemplo 7.1 Para que una variable sea un factor, hay que especificarlo. Observemos el siguiente fragmento de código:

```
> Zona.chr <- c("Europa", "Europa", "Asia",
+              "Africa", "America", "Oceanía", "Asia")
> Zona <- as.factor(Zona.chr)
> Zona.chr

[1] "Europa" "Europa" "Asia"   "Africa"
[5] "America" "Oceanía" "Asia"

> Zona

[1] Europa Europa Asia   Africa America
[6] Oceanía Asia
Levels: Africa America Asia Europa Oceanía
```

Obsérvese que `Zona.chr` y `Zona` se imprimen de manera similar, aunque uno es una cadena de caracteres y otro un factor. La diferencia estriba en las comillas en el primer caso y la línea adicional especificando los niveles en el segundo. Podemos preguntar la clase de objeto con la función `class` o ver la estructura con la función `str` para ver la diferencia:

```
> class(Zona.chr)

[1] "character"

> class(Zona)

[1] "factor"

> str(Zona.chr)

chr [1:7] "Europa" "Europa" "Asia" ...
```

```
> str(Zona)
Factor w/ 5 levels "Africa","America",...: 4 4 3 1 2 5 3
```

Un factor tiene definidos *niveles*, en tanto una cadena de caracteres no:

```
> levels(Zona.chr)
NULL
> levels(Zona)
[1] "Africa" "America" "Asia" "Europa"
[5] "Oceanía"
```

Veamos ahora como definir un factor ordenado:

```
> Estudios <- ordered(c("Superiores", "Medios",
+ "Medios", "Primarios", "Ningunos"))
```

Si no se especifica lo contrario, el orden de los niveles se determina por el orden alfabético de sus denominaciones. Esto haría que en *Estudios* el nivel “Medios” precediera a “Ningunos”, y éste a “Primarios”, lo que es indeseable:

```
> Estudios
[1] Superiores Medios Medios Primarios
[5] Ningunos
4 Levels: Medios < Ningunos < ... < Superiores
```

Para especificar un orden, podemos crear el objeto *Estudios* así:

```
> Estudios <- ordered(c("Superiores", "Medios",
+ "Medios", "Primarios", "Ningunos",
+ "Medios", "Primarios"), levels = c("Ningunos",
+ "Primarios", "Medios", "Superiores"))
> Estudios
[1] Superiores Medios Medios Primarios
[5] Ningunos Medios Primarios
4 Levels: Ningunos < Primarios < ... < Superiores
```

Podemos de modo análogo reordenar los niveles. Si, por ejemplo, queremos revertir el orden, podemos hacerlo así:

```
> Estudios.1 <- ordered(Estudios, levels = c("Superiores",
+      "Medios", "Primarios", "Ningunos"))
```

o, mas simplemente podemos revertir el orden de los niveles mediante la función `rev`, sin necesidad de enumerarlos. Comprobemos a continuación que obtenemos en ambos casos el mismo objeto con el orden de los niveles deseado:

```
> Estudios.2 <- ordered(Estudios, levels = rev(levels(Estudios)))
> Estudios.1

[1] Superiores Medios      Medios      Primarios
[5] Ningunos   Medios      Primarios
4 Levels: Superiores < Medios < ... < Ningunos

> Estudios.2

[1] Superiores Medios      Medios      Primarios
[5] Ningunos   Medios      Primarios
4 Levels: Superiores < Medios < ... < Ningunos
```

Una manipulación que deseamos hacer de ordinario con factores no ordenados es la de poner en primer lugar uno de los niveles, el *nivel de referencia*. Podemos lograrlo cómodamente con la función `relevel`

```
> Zona

[1] Europa Europa Asia      Africa America
[6] Oceanía Asia
Levels: Africa America Asia Europa Oceanía

> Zona <- relevel(Zona, ref = "Asia")
> Zona

[1] Europa Europa Asia      Africa America
[6] Oceanía Asia
Levels: Asia Africa America Europa Oceanía
```

Veremos en el Ejemplo 7.5 la utilidad de esto. Definamos ahora dos variables numéricas:

```
> Ingresos <- c(13456, 12345, 3456, 1234,
+             6789, 4567, 2300)
> Mortalidad <- c(0.003, 0.004, 0.01, 0.02,
+               0.006, 0.005, 0.015)
```

Podemos reunir variables de diferentes tipos en una *dataframe*. A todos los efectos, es como una matriz, pero presenta la peculiaridad de que sus columnas pueden ser de diferentes tipos:

```
> Datos <- data.frame(Zona, Estudios, Ingresos,
+                    Mortalidad)
> Datos
```

	Zona	Estudios	Ingresos	Mortalidad
1	Europa	Superiores	13456	0.003
2	Europa	Medios	12345	0.004
3	Asia	Medios	3456	0.010
4	Africa	Primarios	1234	0.020
5	America	Ningunos	6789	0.006
6	Oceanía	Medios	4567	0.005
7	Asia	Primarios	2300	0.015

```
> str(Datos)
```

```
'data.frame':      7 obs. of  4 variables:
 $ Zona      : Factor w/ 5 levels "Asia","Africa",...: 4 4 1 2 3 5 1
 $ Estudios   : Ord.factor w/ 4 levels "Ningunos"<"Primarios"<...: 4 3 3 2 1 3 2
 $ Ingresos   : num  13456 12345 3456 1234 6789 ...
 $ Mortalidad: num  0.003 0.004 0.01 0.02 0.006 0.005 0.015
```

Una *dataframe* tiene la misma representación interna que una lista. Podemos referirnos a sus términos como a los elementos de una lista, o proporcionando índices de fila y columna:

```
> Datos$Ingresos
[1] 13456 12345 3456 1234 6789 4567 2300

> Datos[[3]]
[1] 13456 12345 3456 1234 6789 4567 2300

> Datos[, "Ingresos"]
[1] 13456 12345 3456 1234 6789 4567 2300
```

```
> Datos[3, 2:3]

  Estudios Ingresos
3 Medios      3456
```

FIN DEL EJEMPLO ■

Una *dataframe* provee un entorno de evaluación. Muchas funciones en R admiten un argumento *data* que permite especificar la *dataframe* en la que es preciso buscar las variables que se nombran. Adicionalmente, la instrucción `attach` hace que las columnas en una *dataframe* sean accesibles como variables definidas en el espacio de trabajo. El Ejemplo 7.2, continuación del Ejemplo 7.1, lo ilustra.

R: Ejemplo 7.2 Comencemos por eliminar del espacio de trabajo algunas variables:

```
> rm(Zona, Estudios, Ingresos, Mortalidad)
```

Si ahora tecleáramos el nombre de alguna de ellas obtendríamos un error. No obstante, tras invocar la función `attach` sus columnas son visibles como si variables en el espacio de trabajo se tratase:

```
> attach(Datos)
> Zona

[1] Europa Europa Asia Africa America
[6] Oceanía Asia
Levels: Asia Africa America Europa Oceanía
```

La función `detach` revierte el efecto de `attach`:

```
> detach(Datos)
```

Si un objeto existe en el espacio de trabajo, su valor oculta el de la columna del mismo nombre en una *dataframe* “attacheada”:

```
> Zona <- c("a", "b", "c")
> attach(Datos)
```

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

```

      Zona
> Zona
[1] "a" "b" "c"
```

FIN DEL EJEMPLO ■

7.3. Fórmulas

Bastantes funciones en R hacen uso de *fórmulas*. Permiten, entre otras cosas, especificar de modo simple modelos de regresión, simplemente nombrando a la izquierda del símbolo \sim la variable respuesta, y a la derecha las variables regresores.

Una fórmula puede proporcionarse como argumento directamente para estimar un modelo de regresión lineal ordinaria (mediante la función `lm`; un ejemplo en la Sección 7.4), regresión lineal generalizada (mediante la función `glm`) o regresión no lineal (mediante la función `nlme` en el paquete del mismo nombre). Por razones didácticas, sin embargo, exploraremos primero el modo en que los diferentes tipos de variables son tratados en una fórmula por la función `model.matrix`.

La función `model.matrix` recibe como argumentos una fórmula y, opcionalmente, una *dataframe* en la que los términos de la fórmula son evaluados. Proporciona la matriz de diseño asociada al modelo que especificamos en la fórmula.

R: Ejemplo 7.3 Supongamos que deseamos investigar la relación entre la variable `Mortalidad` y la variable `Ingresos`. Podemos construir la matriz de diseño así:

```
> X <- model.matrix(Mortalidad ~ Ingresos,
+                  data = Datos)
> X
```

	(Intercept)	Ingresos
1	1	13456
2	1	12345
3	1	3456
4	1	1234

```

5          1      6789
6          1      4567
7          1      2300
attr(,"assign")
[1] 0 1

```

Como podemos ver, se ha añadido automáticamente una columna de “unos”. Si esto fuera indeseable por algún motivo, podríamos evitarlo incluyendo como regresor “-1”.

```

> X <- model.matrix(Mortalidad ~ -1 + Ingresos,
+                   data = Datos)
> X

      Ingresos
1      13456
2      12345
3       3456
4       1234
5       6789
6       4567
7       2300
attr(,"assign")
[1] 1

```

Obsérvese que la variable *Mortalidad* no juega ningún papel en la conformación de la matriz de diseño. Podríamos omitirla y dar sólo el lado derecho de la fórmula, así:

```

> X <- model.matrix(~Ingresos, data = Datos)
> X

(Intercept) Ingresos
1           1      13456
2           1      12345
3           1       3456
4           1       1234
5           1       6789
6           1       4567
7           1       2300
attr(,"assign")
[1] 0 1

```

FIN DEL EJEMPLO ■

La comodidad que proporciona la utilización de fórmulas se hace más evidente, sin embargo, cuando tenemos regresores cualitativos. El Ejemplo 7.4 lo ilustra.

R: Ejemplo 7.4 Consideremos un modelo que tiene como regresores Zona, Ingresos y Estudios. Podemos construir su matriz de diseño así:

```
> X <- model.matrix(~Zona + Estudios + Ingresos,
+ data = Datos)
```

Las variables Zona y Estudios son cualitativas. Requieren ser tratadas de manera especial, y la función `model.matrix` así lo hace. Veamos la matriz de diseño que proporciona:

```
> X
(Intercept) ZonaAfrica ZonaAmerica ZonaEuropa
1           1           0           0           1
2           1           0           0           1
3           1           0           0           0
4           1           1           0           0
5           1           0           1           0
6           1           0           0           0
7           1           0           0           0
ZonaOceanía Estudios.L Estudios.Q Estudios.C
1           0   0.67082     0.5   0.22361
2           0   0.22361    -0.5  -0.67082
3           0   0.22361    -0.5  -0.67082
4           0  -0.22361    -0.5   0.67082
5           0  -0.67082     0.5  -0.22361
6           1   0.22361    -0.5  -0.67082
7           0  -0.22361    -0.5   0.67082
Ingresos
1    13456
2    12345
3     3456
4     1234
5     6789
6     4567
7     2300
```

```

attr("assign")
[1] 0 1 1 1 1 2 2 2 3
attr("contrasts")
attr("contrasts")$Zona
[1] "contr.treatment"

attr("contrasts")$Estudios
[1] "contr.poly"

```

La variable `Ingresos` (numérica) ha sido dejada tal cual. La variable `Zona` es cualitativa nominal, y requiere ser desglosada en tantas columnas como niveles tiene (así, el β asociado a cada columna recoge el efecto del correspondiente nivel). Eso es lo que ha hecho `model.matrix`, salvo que se ha omitido uno de los niveles (el primero) para evitar la multicolinealidad exacta que se hubiera producido de otro modo. El nivel omitido (Asia) pasa así a formar parte del caso de referencia: la función `relevel` (ver Ejemplo 7.1) permitiría cambiar fácilmente el nivel que forma parte del caso de referencia.

El tratamiento de las variables ordinales como `Estudios` es algo más elaborado. En una variable ordinal hay una noción natural de proximidad entre niveles: el nivel de estudios `Medios` está más cerca del nivel `Superiores` que el nivel `Primarios`. Lo que hace `model.matrix` es conceptualmente equivalente a lo siguiente (detalles en la Observación 7.2, pág. 94):

1. Asignar a cada nivel de `Estudios` un valor entero, respetando el orden de la variable: “Ningunos”=1, “Primarios”=2, “Medios”=3 y “Superiores”=4.
2. Con la variable `Estudios` así codificada, crear tantas columnas para la variable `Estudios` como niveles tenga, de la forma: $(\text{Estudios})^0$, $(\text{Estudios})^1$, $(\text{Estudios})^2$, $(\text{Estudios})^3$.

La primera columna, que es constante, es automáticamente desechada si en la matriz de diseño existe columna de “unos”, para evitar la multicolinealidad. Las restantes son rotuladas con las letras “L” (**L**inear), “Q” (**Q**uadratic), “C” (**C**ubic), y así sucesivamente.

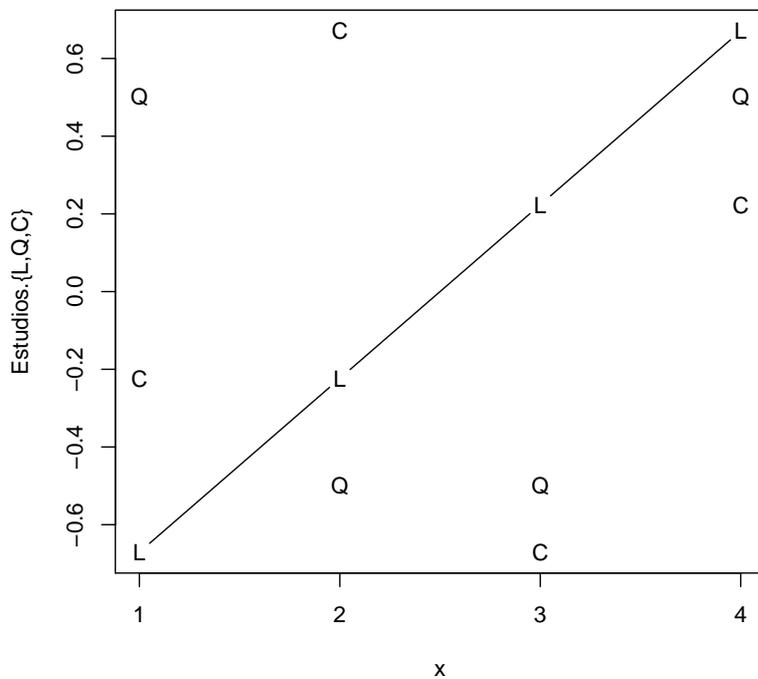
Si empleamos todas las columnas que `model.matrix` crea para una variable ordinal, obtenemos exactamente el mismo subespacio que habríamos obtenido con columnas de ceros y unos como las empleadas para una variable nominal: la ventaja de utilizar una base de dicho subespacio como la que `model.matrix` construye, es que permite en ocasiones realizar una modelización más simple: podemos, a voluntad, emplear en un modelo de regresión algunas, varias o todas

las columnas como regresores, para modelizar un efecto más o menos “suave” sobre la variable respuesta.

FIN DEL EJEMPLO ■

Observación 7.2 Se indica en el Ejemplo 7.4 que el efecto de una variable ordinal se recoge de modo *conceptualmente equivalente* a construir potencias de orden creciente de la variable ordinal codificada por valores enteros que respetan el orden. Ayudará representar gráficamente las columnas correspondientes de la matriz X frente a los enteros codificando los niveles de la variable `Estudios`. Para ello, eliminamos primero niveles duplicados y representaremos los restantes:

```
> x <- as.numeric(Datos[, "Estudios"])
> i <- !duplicated(x)
> plot(x[i], X[i, "Estudios.L"], type = "b",
+      pch = "L", xaxp = c(1, 4, 3), xlab = "x",
+      ylab = "Estudios.{L,Q,C}")
> points(x[i], X[i, "Estudios.Q"], pch = "Q")
> points(x[i], X[i, "Estudios.C"], pch = "C")
```



Hemos dibujado una línea uniendo las “L” para destacar su crecimiento lineal. Las “Q” puede verse que se sitúan sobre una parábola y las “C” sobre una función cúbica.

Un vistazo al gráfico anterior muestra, sin embargo, que el término lineal, por ejemplo, no toma los valores 1, 2, 3 4, ni el cuadrático 1, 4, 9, 16. En efecto,

```
> X[i, 6:8]
      Estudios.L Estudios.Q Estudios.C
1      0.67082      0.5      0.22361
2      0.22361     -0.5     -0.67082
4     -0.22361     -0.5      0.67082
5     -0.67082      0.5     -0.22361
```

En realidad se han rescalado las columnas y se han ortogonalizado:

```
> round(crossprod(X[i, 6:8]))
      Estudios.L Estudios.Q Estudios.C
Estudios.L      1      0      0
```

Estudios.Q	0	1	0
Estudios.C	0	0	1

Ello se hace por razones de conveniencia numérica y de interpretación.

Aunque por razones didácticas hemos construido primero la matriz de diseño y extraído luego un subconjunto de filas y columnas para ver como se codificaba la variable `Estudios`, R proporciona un modo más simple de hacerlo:

```
> contrasts(Datos[, "Estudios"])
      .L  .Q  .C
[1,] -0.67082  0.5 -0.22361
[2,] -0.22361 -0.5  0.67082
[3,]  0.22361 -0.5 -0.67082
[4,]  0.67082  0.5  0.22361
```

Observación 7.3 El anterior es el comportamiento “por omisión” de la función `model.matrix`. Podemos alterarlo especificando distintos modos de desdoblar los factores y factores ordenados. Ello se hace invocando la función `options` de modo similar al siguiente:

```
options(contrasts=c("contr.treatment", "contr.poly"))
```

La primera opción en el argumento `contrasts` se aplica a los factores, la segunda a los factores ordenados. Por ejemplo, para los factores podemos especificar que se desdoblen en tantas columnas como niveles haya, sin incluir ningún nivel en el caso de referencia. Para ello, deberemos proporcionar `contr.sum` como primer valor de `contrasts`:

```
options(contrasts=c("contr.sum", "contr.poly"))
```

Véase la documentación de `contrasts` para más detalles.

Adicionalmente, podemos invocar directamente las funciones

```
contr.sum, contr.treatment, contr.poly, contr.helmert
```

para obtener información sobre el diferente modo en que quedaría codificado un factor. Por ejemplo,

```
> NivelEstudios <- levels(Datos[, "Estudios"])
> contr.sum(NivelEstudios)
```

```

      [,1] [,2] [,3]
Ningunos    1    0    0
Primarios    0    1    0
Medios       0    0    1
Superiores  -1   -1   -1

> contr.treatment(NivelEstudios)

      Primarios Medios Superiores
Ningunos         0         0         0
Primarios         1         0         0
Medios            0         1         0
Superiores        0         0         1

> contr.poly(NivelEstudios)

      .L   .Q   .C
[1,] -0.67082  0.5 -0.22361
[2,] -0.22361 -0.5  0.67082
[3,]  0.22361 -0.5 -0.67082
[4,]  0.67082  0.5  0.22361

```

Obsérvese que mientras `contrasts` se invoca tomando como argumento un factor, las funciones `contr.sum` y similares toman como argumento *el vector de niveles* de un factor.

7.4. La función `lm`.

La función `lm` es un instrumento potente y cómodo de utilizar para el análisis de regresión lineal. Puede utilizarse con tan solo dos argumentos: una fórmula y una *dataframe* que suministra los valores para evaluar las expresiones en dicha fórmula. Por ejemplo, así:

```
ajuste <- lm(y ~ x1 + x2 + x4, data=datos)
```

La función `lm` construye entonces la matriz de diseño mediante la función `model.matrix` y estima el modelo deseado, suministrando un cúmulo de información sobre la estimación. El Ejemplo 7.5 a continuación proporciona detalles.

R: Ejemplo 7.5 Veamos en primer lugar los datos que utilizaremos. Se trata de datos correspondientes a 47 estados en EE.UU. y referidos al años 1960. Forman parte del paquete MASS (soporte

del libro Venables and Ripley (1999b)) que hemos de cargar (mediante una instrucción `library(MASS)`). Tras hacerlo, podemos obtener información detallada sobre los datos tecleando `help(UScrime)`.

```
> library(MASS)
> UScrime[1:3, 1:5]

      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)

'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 ...
```

La función `str` permite ver la estructura de cualquier objeto en R. Lo que muestra en el fragmento anterior es que `UScrime` es una *dataframe*. En este caso, todas las variables son numéricas, algunas reales (`num`) y otras enteras (`int`). Vemos también que tiene 47 filas (=observaciones) y 16 columnas (=posibles regresores).

Probemos ahora a hacer una regresión¹. La variable `y` (tasa de criminalidad) podemos relacionarla con la desigualdad (`Ineq`), probabilidad de ser encarcelado (`Prob`) y con un indicador de Estado sureño (`So`):

¹No se afirma que el modelo que ensayamos sea el mejor en ningún sentido: es sólo una ilustración. El Capítulo 12 abordará la cuestión de cómo seleccionar modelos.

```
> fit <- lm(y ~ Ineq + Prob + So, data = UScrime)
> fit
```

Call:

```
lm(formula = y ~ Ineq + Prob + So, data = UScrime)
```

Coefficients:

(Intercept)	Ineq	Prob
1538.36	-1.58	-8698.46
	So	
	242.99	

El objeto `fit`, al imprimirlo, proporciona una información muy sumaria: apenas la descripción del modelo ajustado y los coeficientes estimados. El empleo de la función `summary`, sin embargo, proporciona un estadillo con información mucho más completa.

```
> summary(fit)
```

Call:

```
lm(formula = y ~ Ineq + Prob + So, data = UScrime)
```

Residuals:

Min	1Q	Median	3Q	Max
-662.8	-163.8	-56.1	82.5	1057.4

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1538.36	345.84	4.45	6e-05
Ineq	-1.58	1.95	-0.81	0.4220
Prob	-8698.46	2725.42	-3.19	0.0026
So	242.99	169.48	1.43	0.1589

(Intercept) ***

Ineq

Prob **

So

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 353 on 43 degrees of freedom

Multiple R-squared: 0.22, Adjusted R-squared: 0.166

F-statistic: 4.05 on 3 and 43 DF, p-value: 0.0127

Desmenecemos la salida anterior. Se imprime, en primer lugar, el modelo ajustado y unos estadísticos sobre los residuos (mínimo, máximo y cuartiles, es decir, valores dejando a su izquierda el 25 %, 50 % y 75 % de los residuos; el segundo cuartil es la mediana). A continuación, tenemos un estadillo proporcionando para cada regresor mencionado al margen:

1. Su $\hat{\beta}_i$ (bajo `Estimate`).
2. Su $\hat{\sigma}_{\hat{\beta}_i}$ (bajo `Std. Error`).
3. Su estadístico t ,

$$\frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}}$$

(bajo `t value`).

4. La probabilidad bajo la hipótesis nula $H_0 : \beta_i = 0$ de obtener un valor del estadístico t tan o más alejado de cero que el obtenido (bajo `Pr(>|t|)`).

A continuación tenemos

$$\sqrt{\frac{SSE}{N-p}},$$

(`Residual standard error`), que estima σ_ϵ , los grados de libertad $N-p$, (`43 degrees of freedom`), R^2 (que toma el valor 0.22) y \bar{R}^2 (`Adjusted R-squared`; este último estadístico será introducido en el Capítulo 12). Finalmente, tenemos el estadístico Q_h para contrastar significación conjunta de la regresión, como se indica en la Sección 6.2 (`F-statistic`). Aquí toma el valor 4.05. Dicho valor deja a su derecha en una distribución $\mathcal{F}_{3,43}$ una cola de probabilidad 0.0127, que es el nivel de significación conjunto de la regresión ajustada.

El objeto compuesto `fit` contiene la información que ha permitido imprimir todos los anteriores resultados y mucha otra, cuyos nombres son autoexplicativos:

```
> attributes(fit)

$names
 [1] "coefficients" "residuals"
 [3] "effects"      "rank"
 [5] "fitted.values" "assign"
 [7] "qr"           "df.residual"
 [9] "xlevels"      "call"
[11] "terms"        "model"

$class
 [1] "lm"
```

Podemos referirnos a los componentes de `fit` y emplearlos en cálculos subsiguientes. Por ejemplo, para obtener la suma de cuadrados de los residuos, SSE, podríamos hacer:

```
> SSE <- sum(fit$residuals^2)
> SSE
[1] 5363970
```

El estadillo anterior sugería que el regresor `Prob` era muy significativo, en tanto los restantes no lo eran. Podemos contrastar la hipótesis $H_0 : \beta_{\text{Ineq}} = \beta_{\text{So}} = 0$ del modo sugerido al final del Ejemplo 6.2, pág. 77: ajustamos una segunda regresión eliminando los regresores `Ineq` y `So`,

```
> fit.h <- lm(y ~ Prob, data = UScrime)
```

calculamos la suma de cuadrados de sus residuos,

```
> SSE.h <- sum(fit.h$residuals^2)
```

y a continuación el estadístico Q_h asociado a la hipótesis y los grados de libertad del mismo:

```
> N <- nrow(UScrime)
> q <- 2
> p <- 4
> Qh <- ((SSE.h - SSE)/q)/(SSE/(N - p))
> Qh
[1] 1.0417
```

La probabilidad que el valor 1.0417 del estadístico deja en la cola a su derecha es

```
> 1 - pf(Qh, q, N - p)
[1] 0.3616
```

lo que sugiere que podemos prescindir de dichos dos regresores.

La instrucción `anova` proporciona una descomposición de la suma de cuadrados de los residuos correspondiente a cada regresor *cuando se introducen en el orden dado*. Compárese por ejemplo,

```
> anova(fit)

Analysis of Variance Table

Response: y
      Df Sum Sq Mean Sq F value Pr(>F)
Ineq   1  220530   220530    1.77 0.1907
Prob   1 1040010 1040010    8.34 0.0061 **
So     1  256417   256417    2.06 0.1589
Residuals 43 5363970 124743
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

con:

```
> fit2 <- lm(y ~ Prob + Ineq + So, data = UScrime)
> anova(fit2)

Analysis of Variance Table

Response: y
      Df Sum Sq Mean Sq F value Pr(>F)
Prob   1 1257075 1257075   10.08 0.0028 **
Ineq   1    3466    3466    0.03 0.8684
So     1  256417   256417    2.06 0.1589
Residuals 43 5363970 124743
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIN DEL EJEMPLO ■

No hay ninguna necesidad ni aparente ventaja en hacerlo así, pero a efectos puramente ilustrativos re-estimaremos la regresión anterior convirtiendo previamente la variable indicadora *So* (Estado del Sur) en una variable nominal y la variable *Ineq* en una variable ordinal (o factor ordenado). Para lo primero, basta que reemplacemos la columna *So* de la *dataframe* del siguiente modo:

```
> UScrime[, "So"] <- factor(UScrime[, "So"],
+   labels = c("Norte", "Sur"))
```

Para la segunda variable, dividiremos su recorrido en tres intervalos, y a continuación definimos un factor ordenado con tres categorías:

```
> Temp <- ordered(cut(UScrime[, "Ineq"],
+   breaks = 3), labels = c("Baja", "Media",
+   "Alta"))
> UScrime[, "Ineq"] <- Temp
```

Podemos ahora repetir la estimación anterior:

R: Ejemplo 7.6 (continuación del Ejemplo 7.5)

```
> fit3 <- lm(y ~ Prob + Ineq + So, data = UScrime)
> summary(fit3)
```

Call:

```
lm(formula = y ~ Prob + Ineq + So, data = UScrime)
```

Residuals:

Min	1Q	Median	3Q	Max
-641.9	-195.5	-55.4	124.3	1059.5

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1212.4	134.8	8.99	2.4e-11
Prob	-9013.8	2717.7	-3.32	0.0019
Ineq.L	-143.2	132.7	-1.08	0.2866
Ineq.Q	-10.6	110.4	-0.10	0.9238
SoSur	284.8	184.3	1.55	0.1298

(Intercept) ***

Prob **

Ineq.L

Ineq.Q

SoSur

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 355 on 42 degrees of freedom

Multiple R-squared: 0.232, Adjusted R-squared: 0.159

F-statistic: 3.17 on 4 and 42 DF, p-value: 0.0229

La variable ordinal `Ineq` da lugar a tres términos (constante, omitido por colineal con la columna de unos, lineal y cuadrático). La variable nominal `So` se desglosa también en dos: el nivel “Norte” se integra en el caso de referencia y el parámetro restante mide el efecto deferencial del nivel “Sur” respecto al nivel “Norte”. A título ilustrativo, podemos ajustar la anterior regresión empleando un diferente desdoblamiento del regresor cualitativo `So`:

```
> options(contrasts = c("contr.sum", "contr.poly"))
> fit4 <- lm(y ~ Prob + Ineq + So, data = UScrime)
> summary(fit4)
```

Call:

```
lm(formula = y ~ Prob + Ineq + So, data = UScrime)
```

Residuals:

Min	1Q	Median	3Q	Max
-641.9	-195.5	-55.4	124.3	1059.5

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1354.7	151.0	8.97	2.6e-11
Prob	-9013.8	2717.7	-3.32	0.0019
Ineq.L	-143.2	132.7	-1.08	0.2866
Ineq.Q	-10.6	110.4	-0.10	0.9238
So1	-142.4	92.1	-1.55	0.1298

(Intercept) ***

Prob **

Ineq.L

Ineq.Q

So1

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 355 on 42 degrees of freedom

Multiple R-squared: 0.232, Adjusted R-squared: 0.159

F-statistic: 3.17 on 4 and 42 DF, p-value: 0.0229

(Véase la Observación 7.3.) Vemos un sólo regresor asociado a `So1`, el primer nivel de `So`; el asociado al segundo nivel es su opuesto, ya que `contr.sum` fuerza los coeficientes asociados a un regresor nominal a sumar cero.

Si observamos los dos ajustes, vemos que son idénticos. Lo único que se altera es la interpretación de los parámetros. En `fit3`, el tratarse de un Estado del Sur tenía como efecto incrementar la tasa de criminalidad en 284.8, respecto de la tasa prevalente en un Estado del Norte de análogas características. La parametrización en el modelo `fit4` expresa lo mismo de otro modo: en un Estado del Norte, la criminalidad desciende en -142.4 sobre el nivel promedio de Norte y Sur, mientras que en un Estado del Sur aumenta en 142.4. La diferencia entre ambos niveles continúa siendo 284.8.

Puede encontrarse una discusión exhaustiva de las diferentes opciones de parametrización disponibles en Venables and Ripley (1999a), Sec. 6.2.

FIN DEL EJEMPLO ■

7.5. Lectura recomendada.

Sobre R. Son ya bastantes las obras que es posible consultar sobre la utilización de R como herramienta para los cálculos que requiere la regresión lineal. Una excelente referencia es Venables and Ripley (1999a). Exclusivamente orientado a modelos lineales es Faraway (2005).

Capítulo 8

Inferencia simultánea.

8.1. Problemas que plantea el contrastar múltiples hipótesis simultáneas

Evidencia contra una hipótesis

Si examinamos la teoría sobre contrastes de hipótesis presentada en la Sección 6.2 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 ordenarlos de más a menos “raros” de acuerdo con su densidad respectiva.

Ejemplo 8.1 Para una muestra X_1, \dots, X_n procedente de una distribución $N(\mu, \sigma^2)$, todos los posibles valores del estadístico \bar{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 μ , 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 μ .

FIN DEL EJEMPLO ■

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

Definición 8.1 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 8.2 En el Ejemplo 8.1, supongamos que $H_0 : \mu = 0$. Supongamos conocida $\sigma^2 = 1$. Sea una muestra con $n = 100$, e imaginemos que obtenemos un valor de \bar{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 \bar{X} igual o más alejados de μ que el que se ha presentado.

FIN DEL EJEMPLO ■

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 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.

¿Cómo de “raro” ha de ser algo para ser realmente “raro”?

El siguiente ejemplo² ilustra que un resultado aparentemente muy raro puede no serlo tanto.

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

²Paráfrasis de un célebre comentario de Bertrand Russell.

Ejemplo 8.3 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 sólo falta de ortografía *Hamlet*!

Bajo la hipótesis nula H_0 : “mono irracional”, tal resultado es absolutamente inverosímil. La probabilidad de que golpeando al azar el 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:

$$p = \left(\frac{1}{32}\right)^{635000} \approx 5,804527 \times 10^{-955771}. \quad (8.1)$$

La observación de un mono que teclaea *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 en torno a 5,8045 transcripciones exactas de *Hamlet*. Lo observado no es raro en absoluto.

FIN DEL EJEMPLO ■

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.

Análisis exploratorio e inferencia

Es importante entender lo que el Ejemplo 8.3 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 corriamos el efecto de la selección.

Ejemplo 8.4 Regresemos al Ejemplo 8.3. 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.

FIN DEL EJEMPLO ■

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

Ejemplo 8.5 De nuevo en el Ejemplo 8.3, no hay nada reprochable 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 reprochable es seleccionar el único mono que teclea *Hamlet* y reportar el hallazgo como si ese mono fuera el único observado.

FIN DEL EJEMPLO ■

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 8.6 Supongamos el modelo

$$\vec{Y} = \beta_0 \vec{X}_0 + \beta_1 \vec{X}_1 + \dots + \beta_{99} \vec{X}_{99} + \vec{\epsilon}.$$

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 $H_0 : \beta_i = 0, i = 0, \dots, 99$, podríamos (si se verifican los supuestos necesarios) emplear el contraste presentado en la Sección 6.2, pág. 79. Podríamos ser más ambiciosos e intentar al mismo tiempo ver cuál o cuales β_i son distintos de cero. Sería *incorrecto* operar así:

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

Es fácil ver por qué es incorrecto. Bajo H_0 hay probabilidad tan sólo α de que un t -ratio prefijado exceda en valor absoluto de $t_{N-p}^{\alpha/2}$. Pero la probabilidad de que *algún* t -ratio exceda de $t_{N-p}^{\alpha/2}$ es³

$$\text{Prob}(\text{Algún } \beta_i \neq 0) = 1 - (1 - \alpha)^p. \quad (8.2)$$

mayor (en ocasiones *mucho mayor*) que α . 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 $1 - 0,95^{100} =$

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

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 inadmisibles es dar un nivel de significación incorrectamente calculado.

FIN DEL EJEMPLO ■

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. 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.

8.2. Desigualdad de Bonferroni.

Consideremos k sucesos, E_i , ($i = 1, \dots, k$), cada uno de ellos con probabilidad $(1 - \alpha)$. Designamos por \bar{E}_i el complementario del suceso E_i . La probabilidad de que todos los sucesos E_i , ($i = 1, \dots, k$) acaezcan simultáneamente es:

$$\text{Prob}\{\cap_{i=1}^k E_i\} = 1 - \text{Prob}\{\overline{\cap_{i=1}^k E_i}\} = 1 - \text{Prob}\{\cup_{i=1}^k \overline{E_i}\} \geq 1 - k\alpha \quad (8.3)$$

Se conoce (8.3) 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:

$$E_1 : [(\hat{\beta}_1 \pm \hat{\sigma}_{\hat{\beta}_1} t_{N-p}^{\alpha/2}) \quad \text{cubre } \beta_1] \quad (8.4)$$

$$\vdots \quad (8.5)$$

$$E_k : [(\hat{\beta}_k \pm \hat{\sigma}_{\hat{\beta}_k} t_{N-p}^{\alpha/2}) \quad \text{cubre } \beta_k] \quad (8.6)$$

Cada E_i por separado es un suceso cuya probabilidad es $1 - \alpha$. De acuerdo con (8.3), 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 $\vec{X}_0, \dots, \vec{X}_{p-1}$ y quisiéramos obtener intervalos de confianza *simultáneos* α para los parámetros $\beta_0, \dots, \beta_{p-1}$, sería claramente incorrecto emplear los que aparecen en (8.4)–(8.6). 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 β_i fueran cubiertos por sus respectivos intervalos, sería mayor, de acuerdo con (8.3), 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 β_i quizá más de lo necesario. En lo que sigue veremos procedimientos para lograr el mismo resultado con intervalos en general más estrechos.

8.3. Intervalos de confianza basados en la máxima t .

Supongamos que tenemos k variables aleatorias independientes, t_1, \dots, t_k con distribución t -Student, y número común n de grados de libertad. La

variable aleatoria $\max\{|t_1|, \dots, |t_k|\}$ sigue una distribución que se halla tabulada⁴.

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

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

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

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

$$\text{Prob}\left\{\cap_{i=1}^k \left[\left|\frac{\vec{a}_i' \hat{\beta}}{\hat{\sigma}_{\vec{a}_i' \hat{\beta}}}\right| \leq u_{k,n}^\alpha\right]\right\} = 1 - \alpha \quad (8.7)$$

Es claro que $\vec{a}_i' \hat{\beta} / \hat{\sigma}_{\vec{a}_i' \hat{\beta}}$ ($i = 1, \dots, k$) **no** son independientes. Sin embargo, la distribución aludida del máximo valor absoluto de k variables t de Student está también tabulada cuando dichas variables tienen correlación ρ por pares. (Esto sucede en algunos casos particulares, como el de ciertos diseños de Análisis de Varianza equilibrados: la correlación ρ entre parejas de t -ratios es la misma, y fácil de calcular.)

Aún cuando la correlación ρ por pares de t -ratios no sea siempre la misma, (8.7) es de utilidad. Suministra intervalos simultáneos de confianza aproximada $1 - \alpha$. En caso de que conozcamos ρ , podemos emplear la expresión (8.7) con $u_{k,n}^\alpha$ reemplazado por $u_{k,n,\rho}^\alpha$, extraído éste último de la tabla correspondiente; en caso de que no conozcamos ρ , o ésta no sea constante, podemos utilizar $u_{k,n,\rho=0}^\alpha$, lo que hace en general los intervalos calculados con ayuda de (8.7) conservadores (es decir, la probabilidad conjunta en el lado izquierdo de (8.7) es *mayor* que $1 - \alpha$).

Es importante señalar que, si nuestro objetivo es contrastar una hipótesis del tipo $h : A\vec{\beta} = \vec{c}$ con $\text{rango}(A) > 1$, *tenemos* que emplear un contraste como el descrito en la Sección 6.2, pág. 72. El comparar cada una de las variables aleatorias $\left|(\vec{a}_i' \hat{\beta} - c_i) / \hat{\sigma}_{\vec{a}_i' \hat{\beta}}\right|$ ($i = 1, \dots, k$) con una $t_{N-p}^{\alpha/2}$ supone emplear un nivel de significación *mayor* que α . Como caso particular, es inadecuado contrastar la hipótesis $h : \beta_1 = \dots = \beta_p = 0$ comparando cada uno de los t -ratios con $t_{N-p}^{\alpha/2}$; tal contraste tendría un nivel de significación sensiblemente superior a α , en especial si p es grande.

En el caso de que el contraste conjunto rechace $h : A\vec{\beta} = \vec{c}$ y queramos saber qué filas de A son culpables del rechazo, podríamos comparar

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

$|(\vec{a}_i' \hat{\beta} - c_i) / \hat{\sigma}_{\vec{a}_i' \hat{\beta}}|$ ($i = 1, \dots, k$) con $u_{k,n}^\alpha$ ($k =$ 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 .

8.4. 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 8.1 *Sea L una matriz simétrica de orden $k \times k$ definida positiva, y \vec{c} , \vec{b} vectores k -dimensionales cualesquiera. Se verifica que:*

$$\sup_{\vec{c} \neq \vec{0}} \left[\frac{[\vec{c}' \vec{b}]^2}{\vec{c}' L \vec{c}} \right] = \vec{b}' L^{-1} \vec{b} \quad (8.8)$$

DEMOSTRACIÓN:

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

$$\vec{v} = R' \vec{c} \quad (8.9)$$

$$\vec{u} = R^{-1} \vec{b} \quad (8.10)$$

y tenemos en cuenta que por la desigualdad de Schwarz,

$$\frac{\langle \vec{u}, \vec{v} \rangle^2}{\|\vec{u}\|^2 \|\vec{v}\|^2} \leq 1 \quad (8.11)$$

entonces sustituyendo (8.9) y (8.10) en (8.11) obtenemos (8.8). ■

Podemos ahora abordar la construcción de intervalos de confianza simultáneos por el método de Scheffé. Supongamos que tenemos k hipótesis lineales $h_i: \vec{a}_i' \vec{\beta} = c_i$ ($i = 1, \dots, k$) cuyo contraste conjunto deseamos efectuar. Si denominamos:

$$A = \begin{pmatrix} \vec{a}_1' \\ \vec{a}_2' \\ \dots \\ \vec{a}_k' \end{pmatrix} \quad \vec{c} = \begin{pmatrix} c_1 \\ c_2 \\ \dots \\ c_k \end{pmatrix} \quad (8.12)$$

dichas k hipótesis se pueden escribir como $h: A\vec{\beta} = \vec{c}$. Cuando h es cierta, sabemos (Sección 6.2) que:

$$\frac{(A\hat{\beta} - \vec{c})'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - \vec{c})}{q\hat{\sigma}^2} \sim \mathcal{F}_{q, N-p} \quad (8.13)$$

siendo $q = \min(d, p)$, en que $d = \text{rango } A$ y $p = \text{rango } (X'X)$. 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:

$$1 - \alpha = \text{Prob} \left\{ (\hat{c} - \vec{c})'[A(X'X)^{-1}A']^{-1}(\hat{c} - \vec{c}) \leq q\hat{\sigma}^2 \mathcal{F}_{q, N-p}^\alpha \right\} \quad (8.14)$$

$$= \text{Prob} \left\{ (\hat{c} - \vec{c})'L^{-1}(\hat{c} - \vec{c}) \leq q\hat{\sigma}^2 \mathcal{F}_{q, N-p}^\alpha \right\} \quad (8.15)$$

en que $L = [A(X'X)^{-1}A']$. Teniendo en cuenta el Lema 8.1, obtenemos:

$$1 - \alpha = \text{Prob} \left\{ \sup_{\vec{h} \neq \vec{0}} \left[\frac{[\vec{h}'(\hat{c} - \vec{c})]^2}{\vec{h}'L\vec{h}} \right] \leq q\hat{\sigma}^2 \mathcal{F}_{q, N-p}^\alpha \right\} \quad (8.16)$$

$$= \text{Prob} \left\{ \bigcap_{\vec{h} \neq \vec{0}} \left[\left| \frac{\vec{h}'(\hat{c} - \vec{c})}{(\vec{h}'L\vec{h})^{\frac{1}{2}}} \right| \leq (q\hat{\sigma}^2 \mathcal{F}_{q, N-p}^\alpha)^{\frac{1}{2}} \right] \right\} \quad (8.17)$$

La ecuación (8.17) muestra que $(q\hat{\sigma}^2 \mathcal{F}_{q, N-p}^\alpha)^{\frac{1}{2}}$ es un valor que acota con probabilidad $1 - \alpha$ un número arbitrariamente grande de cocientes como:

$$\frac{|\vec{h}'(\hat{c} - \vec{c})|}{\sqrt{\vec{h}'L\vec{h}}} \quad (8.18)$$

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

$$\vec{h}'\hat{c} \pm \sqrt{(\vec{h}'L\vec{h})(q\hat{\sigma}^2 \mathcal{F}_{q, N-p}^\alpha)} \quad (8.19)$$

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

Esto es *más* de lo que necesitamos —pues sólo queríamos intervalos de confianza simultáneos para c_1, \dots, 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 (8.19) se reducen a:

$$\vec{h}'\hat{\beta} \pm \sqrt{(\vec{h}'(X'X)^{-1}\vec{h})(p\hat{\sigma}^2\mathcal{F}_{p,N-p}^\alpha)} \quad (8.20)$$

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 8.1 (*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_{kl} fuera “uno” cuando la k -ésima parcela se siembra con la variedad l -ésima de semilla y la variable respuesta recogiera las cosechas obtenidas en las diferentes parcelas, los parámetros β_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 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 ($h : A\vec{\beta} = \vec{c}$) 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.

Comencemos por construir una matriz de diseño y generar artificialmente las observaciones:

```
> 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(25, sd = 0.1)

```

Construyamos la matriz definiendo la hipótesis conjunta $A\vec{\beta} = \vec{c}$:

```

> p <- ncol(X)
> N <- nrow(X)
> A <- cbind(1, diag(-1, p - 1))
> 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

> q <- nrow(A)

```

Aunque por motivos didácticos hemos construido A del modo que se ha visto, hay funciones standard que permiten hacerlo con mayor comodidad.

```
> A <- t(contrasts(as.factor(1:5)))
> A

      1 2 3 4 5
[1,] 1 0 0 0 -1
[2,] 0 1 0 0 -1
[3,] 0 0 1 0 -1
[4,] 0 0 0 1 -1
```

que es equivalente a la A precedente.

Habiendo p betas a comparar, habrá un total de $\frac{p(p-1)}{2}$ comparaciones a efectuar. Construimos una matriz cada una de cuyas filas corresponde a una comparación:

```
> H <- matrix(0, p * (p - 1)/2, p)
> 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

      [,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
```

El siguiente fragmento de código construye ahora todos los intervalos de la forma dada por (8.20) y los imprime:

```
> 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)
> 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: ( -1.0463 , -0.94141 ) *
Intervalo comp.  1 0 -1 0 0 es: ( -1.0631 , -0.95825 ) *
Intervalo comp.  1 0 0 -1 0 es: ( -2.0886 , -1.9837 ) *
Intervalo comp.  1 0 0 0 -1 es: ( -2.067 , -1.9622 ) *
Intervalo comp.  0 1 -1 0 0 es: ( -0.069268 , 0.035591 )
Intervalo comp.  0 1 0 -1 0 es: ( -1.0947 , -0.98989 ) *
Intervalo comp.  0 1 0 0 -1 es: ( -1.0732 , -0.96834 ) *
Intervalo comp.  0 0 1 -1 0 es: ( -1.0779 , -0.97305 ) *
Intervalo comp.  0 0 1 0 -1 es: ( -1.0564 , -0.9515 ) *
Intervalo comp.  0 0 0 1 -1 es: ( -0.030881 , 0.073979 )

```

Vemos que la mayoría de intervalos de confianza simultáneos no cubren el cero. Los correspondientes a $\beta_2 - \beta_3$ y $\beta_4 - \beta_5$ si lo hacen, como esperábamos, ya que en ambas parejas los parámetros han sido fijados al mismo valor.

FIN DEL EJEMPLO ■

8.5. 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 particular⁵.

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

- 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 distribución del máximo de k ($k = 20$ en este caso) variables aleatorias con distribución t de Student ($u_{20, N-20}^\alpha$).
- 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 6.2) y el que hace uso de (8.7). 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 .

COMPLEMENTOS Y EJERCICIOS

8.1 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,

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

Tras estimar los 201 parámetros del mismo, se plantea contrastar la hipótesis como $H_0 : \beta_1 = \dots = \beta_{200}$ 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 del máximo de k variables t de Student, con grados de libertad apropiados.
- Calcular el estadístico Q_h para la hipótesis $H_0 : \hat{\beta}_1, \dots, \hat{\beta}_{200} = 0$ y comparar con $\mathcal{F}_{200, 500-201; \alpha}$.

Juzga los diferentes procedimientos, e indica con cuál (o cuáles) de ellos tendríamos garantizada una probabilidad de error de tipo I no superior al α prefijado.

8.2 Preocupado por el posible impacto de las antenas de telefonía móvil sobre la salud de los niños, un político solicita un listado 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 contraído 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 Sanidad: “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 9

Multicolinealidad.

9.1. 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 Capítulo¹ 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. ¿Cómo 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. Nótese que cuando \vec{c} es un vector de ceros con un 1 en una única posición, la primera cuestión incluye, como caso particular, la de si un parámetro concreto es estimable.

La segunda cuestión introducirá la idea de multicolinealidad aproximada. Mientras que desde un punto de vista formal la matriz de diseño es de rango deficiente o no lo es, en la práctica interesa distinguir aquellas situaciones en que la matriz de diseño es de rango “casi” deficiente. Cuando esto ocurra,

¹Basado en Silvey (1969).

en un sentido que se aclarará más abajo, todo es estimable, pero algunas formas lineales $\vec{c}'\vec{\beta}$ lo son con gran imprecisión: la varianza de su mejor estimador lineal insesgado 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. Admitido que algunas formas lineales quizá sólo pueden ser estimadas con gran varianza ¿cómo habría que escoger o ampliar X en los casos en que somos libres de ampliar la muestra?

El principal hallazgo al responder a las dos primeras cuestiones será que combinaciones lineales $\vec{c}'\vec{\beta}$ con \vec{c} aproximadamente colineal a un vector propio de $(X'X)$ de valor propio asociado “pequeño”, son las de estimación más imprecisa. La consecuencia será que haremos lo posible en nuestros diseños experimentales para que, si $\vec{c}'\vec{\beta}$ es una forma lineal de interés, no haya vectores propios de $(X'X)$ con valor propio pequeño aproximadamente en la misma dirección de \vec{c} . Recurriremos para ello a ampliar la muestra, si podemos hacerlo, o a procedimientos *ad-hoc* de manipulación de dichos valores propios pequeños para obtener estimadores diferentes del MCO. Esta cuestión se estudia en el Capítulo 10.

Realizaremos un análisis formal de la multicolinealidad en las Secciones 9.4 y siguientes. Previamente será de interés abordar la cuestión desde una perspectiva informal (en la Sección 9.2) y examinar los síntomas que evidencian problemas de multicolinealidad en una matriz de diseño (Sección 9.3).

9.2. Una aproximación intuitiva

La Figura 9.1 recoge sendas situaciones de multicolinealidad exacta (en el panel superior) y multicolinealidad aproximada (en el inferior). En el panel superior,

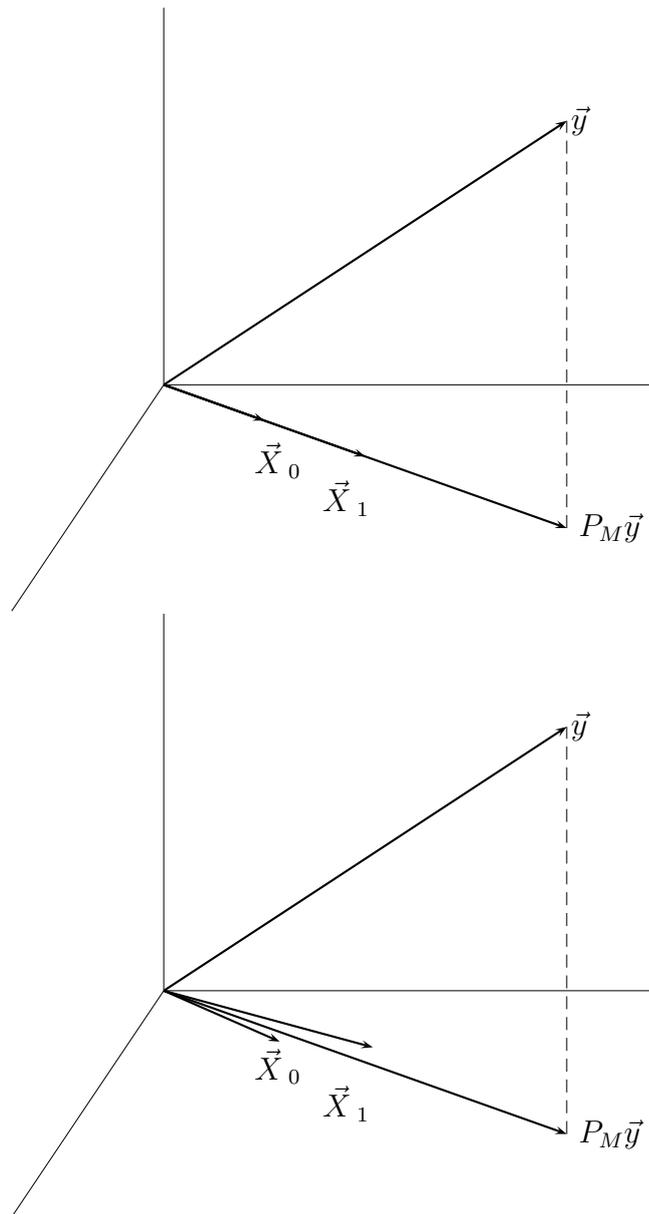
$$P_M\vec{y} = \begin{bmatrix} 5,3 \\ 1,9 \end{bmatrix} \quad \vec{X}_0 = \begin{bmatrix} 2,65 \\ 0,95 \end{bmatrix} \quad \vec{X}_1 = \begin{bmatrix} 1,325 \\ 0,475 \end{bmatrix} \quad (9.1)$$

Puede comprobarse que $\vec{X}_0 = 2 \times \vec{X}_1$, por lo que la matriz de diseño que tuviera a ambos vectores por columnas sería de rango deficiente. Consecuentemente, los estimadores MCO de los parámetros β_0 y β_1 no están unívocamente determinados. Puede comprobarse que

$$P_M\vec{y} = \hat{\beta}_0\vec{X}_0 + \hat{\beta}_1\vec{X}_1 \quad (9.2)$$

se verifica con $\hat{\beta}_0 = 2$ y $\hat{\beta}_1 = 0$ ó con $\hat{\beta}_0 = 0$ y $\hat{\beta}_1 = 4$, por ejemplo. De hecho, cualesquiera $\hat{\beta}_0, \hat{\beta}_1$ verificando $\hat{\beta}_0 + 2\hat{\beta}_1 = 2$ son una solución de (9.2).

Figura 9.1: Multicolinealidad exacta (panel superior) y aproximada (panel inferior).



En el panel inferior de la Figura 9.1,

$$P_M \vec{y} = \begin{bmatrix} 5,3 \\ 1,9 \end{bmatrix} \quad \vec{X}_0 = \begin{bmatrix} 2,75 \\ 0,75 \end{bmatrix} \quad \vec{X}_1 = \begin{bmatrix} 1,525 \\ 0,675 \end{bmatrix}; \quad (9.3)$$

puede comprobarse que ahora $P_M \vec{y} = 0,9544 \vec{X}_0 + 1,7544 \vec{X}_1$. Si, no obstante, $P_M \vec{y}$ fuera ligeramente diferente, con los mismos regresores,

$$P_M \vec{y} = \begin{bmatrix} 5,4 \\ 1,8 \end{bmatrix} \quad \vec{X}_0 = \begin{bmatrix} 2,75 \\ 0,75 \end{bmatrix} \quad \vec{X}_1 = \begin{bmatrix} 1,525 \\ 0,675 \end{bmatrix} \quad (9.4)$$

tendríamos que la solución única sería $P_M \vec{y} = 1,263 \vec{X}_0 + 1,2632 \vec{X}_1$. Una pequeña perturbación en $P_M \vec{y}$ ha originado un cambio drástico en los valores de los estimadores.

Si examinamos el panel inferior de la Figura 9.1, podemos entender fácilmente lo que sucede: los regresores son linealmente independientes y generan el plano horizontal, pero tienen una colinealidad acusada. Un leve cambio en la posición de $P_M \vec{y}$ hace que sea mucho más colineal con un regresor que con otro, y provoca una drástica modificación en los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$.

Tenemos así que si en situaciones de multicolinealidad exacta los parámetros (o algunos de entre ellos) son radicalmente inestimables, cuando el rango de la matrix X es completo, pero algunas de sus columnas son acusadamente colineales, la estimación es posible, pero imprecisa. Decimos que estamos ante una situación de *multicolinealidad aproximada*.

La multicolinealidad aproximada es, en esencia, una matriz de diseño pobre, que no permite deslindar con precisión el efecto de cada regresor sobre la variable respuesta. Es una situación muy frecuente en la práctica, a medio camino entre la multicolinealidad exacta y la ortogonalidad entre los regresores. La Sección que sigue detalla algunos síntomas que permiten percibir su existencia.

9.3. 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 aquella, aparentemente paradójica, en que todos los parámetros en $\hat{\beta}$ 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.

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 $(X'X)$. 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 λ_1/λ_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:

$$VIF(i) \stackrel{\text{def}}{=} \frac{1}{1 - R^2(i)}; \quad (9.5)$$

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 9.1 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 $\text{Var}(\hat{\beta}_i) = \sigma^2$; por tanto, $(X'X)^{ii}$ recoge el factor en que se modifica en general $\text{Var}(\hat{\beta}_i)$ 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 del VIF(i).

9.4. Caracterización de formas lineales estimables.

Teorema 9.1 *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.*

DEMOSTRACIÓN:

Observemos que el enunciado no es sino una paráfrasis del Teorema 3.1, pág. 45. La siguiente cadena de implicaciones, que puede recorrerse en ambas direcciones, establece la demostración.

$$\vec{c}'\vec{\beta} \text{ estimable} \iff \exists \vec{d}: \vec{c}'\vec{\beta} = E[\vec{d}'\vec{Y}] \quad (9.6)$$

$$\iff \vec{c}'\vec{\beta} = \vec{d}'X\vec{\beta} \quad (9.7)$$

$$\iff \vec{c}' = \vec{d}'X \quad (9.8)$$

$$\iff \vec{c} = X'\vec{d} \quad (9.9)$$

$$\iff \vec{c} \in R(X') \quad (9.10)$$

$$\iff \vec{c} \in R(X'X) \quad (9.11)$$

$$\iff \vec{c} = \alpha_1\vec{v}_1 + \cdots + \alpha_{p-j}\vec{v}_{p-j} \quad (9.12)$$

siendo $\vec{v}_1, \dots, \vec{v}_{p-j}$ los vectores propios de $(X'X)$ asociados a valores propios no nulos. El paso de (9.10) a (9.11) hace uso del hecho de que tanto las columnas de X' como las de $X'X$ generan el mismo subespacio² de R^p . La

²Es inmediato ver que $R(X'X) \subseteq R(X')$, pues si $\vec{v} \in R(X'X) \Rightarrow \exists \vec{a}: \vec{v} = X'X\vec{a} = X'\vec{d}$, 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, $X'X\vec{b} = \vec{0} \Rightarrow \vec{b}'X'X\vec{b} = \vec{d}'\vec{d} = \vec{0} \Rightarrow \vec{d} = \vec{0} \Rightarrow X\vec{b} = \vec{0}$.

equivalencia entre (9.11) y (9.12) 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:

$$Z = XV \tag{9.13}$$

$$\vec{\gamma} = V'\vec{\beta} \tag{9.14}$$

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

$$X\vec{\beta} = XVV'\vec{\beta} = Z\vec{\gamma} \tag{9.15}$$

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:

$$Z'Z = (XV)'(XV) = V'X'XV = \Lambda \tag{9.16}$$

siendo Λ 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 λ 's son no nulos, y los restantes j son cero: $\lambda_p = \lambda_{p-1} = \dots = \lambda_{p-j+1} = 0$.

Observemos que de (9.14) 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:

$$(Z'Z)\hat{\gamma} = \Lambda\hat{\gamma} = Z'\vec{y} \tag{9.17}$$

o en forma desarrollada:

$$\begin{pmatrix} \lambda_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \lambda_{p-j} & \dots & 0 \\ 0 & 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 0 \end{pmatrix} \hat{\gamma} = Z'\vec{y} \tag{9.18}$$

El sistema (9.18) es indeterminado; solo los $(p - j)$ primeros $\hat{\gamma}'s$ pueden obtenerse de él. Obsérvese además que de (9.18) se deduce que $\text{var}(\hat{\gamma}_i) \propto 1/\lambda_i$, ($i = 1, \dots, p - j$).

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

$$\vec{c}'\vec{\beta} = \vec{c}'VV'\vec{\beta} = (\vec{c}'V)\vec{\gamma} = (V'\vec{c})'\vec{\gamma} \quad (9.19)$$

y consiguientemente una estimación de $\vec{c}'\vec{\beta}$ vendrá dada por $(V'\vec{c})'\hat{\gamma}$. Por tanto, $\vec{c}'\vec{\beta}$ será estimable si $\hat{\gamma}$ es estimable, o si $\vec{c}'\vec{\beta}$ 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 (9.18), $\vec{c}'\vec{\beta}$ podrá estimarse si $(V'\vec{c})'$ tiene nulas sus últimas j coordenadas, lo que a su vez implica:

$$\vec{c} \perp \vec{v}_p \quad (9.20)$$

$$\vec{c} \perp \vec{v}_{p-1} \quad (9.21)$$

$$\vdots \quad (9.22)$$

$$\vec{c} \perp \vec{v}_{p-j+1} \quad (9.23)$$

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 (9.20)–(9.23): $\vec{c} = \alpha_1\vec{v}_1 + \dots + \alpha_{p-j}\vec{v}_{p-j}$. Toda forma estimable debe por tanto ser expresable así:

$$\vec{c}'\vec{\beta} = (\alpha_1\vec{v}_1 + \dots + \alpha_{p-j}\vec{v}_{p-j})'\vec{\beta}, \quad (9.24)$$

resultado al que habíamos llegado.

Recapitulemos: una forma lineal $\vec{c}'\vec{\beta}$ es estimable si $\vec{c} = \alpha_1\vec{v}_1 + \dots + \alpha_{p-j}\vec{v}_{p-j}$, es decir, no depende de vectores propios de $(X'X)$ asociados a valores propios nulos. Tal como sugería la Sección 9.2, podemos sin embargo esperar que formas lineales que son estrictamente estimables lo sean muy imprecisamente, en situaciones de multicolinealidad aproximada. La Sección que sigue formaliza esta intuición, mostrando que si \vec{c} depende de vectores propios de valor propio cercano a cero, la forma lineal $\vec{c}'\vec{\beta}$ será estimable sólo con gran varianza.

9.5. 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:

$$\begin{aligned}\vec{v}_i'(X'X)\hat{\beta} &= \vec{v}_i'X'\vec{Y} \\ \lambda_i\vec{v}_i'\hat{\beta} &= \vec{v}_i'X'\vec{Y}\end{aligned}$$

y tomando varianzas a ambos lados:

$$\begin{aligned}\lambda_i^2 \text{var}(\vec{v}_i'\hat{\beta}) &= \text{var}(\vec{v}_i'X'\vec{Y}) \\ &= \vec{v}_i'X'\sigma^2IX\vec{v}_i \\ &= \vec{v}_i'X'X\vec{v}_i\sigma^2 \\ &= \lambda_i\sigma^2\end{aligned}\tag{9.25}$$

De la igualdad (9.25) se deduce que:

$$\text{var}(\vec{v}_i'\hat{\beta}) = \frac{\sigma^2}{\lambda_i}\tag{9.26}$$

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

$$\begin{aligned}\text{cov}(\vec{v}_i'\hat{\beta}, \vec{v}_j'\hat{\beta}) &= \vec{v}_i'\Sigma_{\hat{\beta}}\vec{v}_j \\ &= \vec{v}_i'(X'X)^{-1}\vec{v}_j\sigma^2 \\ &= \vec{v}_i'\lambda_j^{-1}\vec{v}_j\sigma^2 \\ &= \sigma^2\lambda_j^{-1}\vec{v}_i'\vec{v}_j \\ &= 0\end{aligned}\tag{9.27}$$

La varianza de cualquier forma estimable $\vec{c}'\vec{\beta}$, teniendo en cuenta que puede escribirse como en (9.24), y haciendo uso de (9.26) y (9.27), será:

$$\begin{aligned}\text{var}(\vec{c}'\hat{\beta}) &= \text{var}[(\alpha_1\vec{v}_1 + \dots + \alpha_{p-j}\vec{v}_{p-j})'\hat{\beta}] \\ &= \alpha_1^2 \text{var}(\vec{v}_1'\hat{\beta}) + \dots + \alpha_{p-j}^2 \text{var}(\vec{v}_{p-j}'\hat{\beta}) \\ &= \alpha_1^2 \left[\frac{\sigma^2}{\lambda_1} \right] + \dots + \alpha_{p-j}^2 \left[\frac{\sigma^2}{\lambda_{p-j}} \right] \\ &= \sigma^2 \left[\frac{\alpha_1^2}{\lambda_1} + \dots + \frac{\alpha_{p-j}^2}{\lambda_{p-j}} \right]\end{aligned}\tag{9.28}$$

La expresión (9.28) es reveladora; la varianza en la estimación de $\vec{c}'\vec{\beta}$ dependerá de la varianza de la perturbación σ^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 $\vec{c} = \alpha_1\vec{v}_1 + \dots + \alpha_{p-j}\vec{v}_{p-j}$ y los α 's multiplicando a vectores propios con reducido valor propio son sustanciales, los correspondientes sumandos tenderán a dominar la expresión (9.28).

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.

Hemos razonado en esta Sección y la precedente en el caso de que j valores propios de $X'X$ son exactamente cero. Es claro que si todos los valores propios son mayores que cero, todas las formas lineales serán estimables, con varianza:

$$\text{var}(\vec{c}'\hat{\beta}) = \text{var}[(\alpha_1\vec{v}_1 + \dots + \alpha_{p-j}\vec{v}_{p-j})'\hat{\beta}] \quad (9.29)$$

$$\begin{aligned} &= \alpha_1^2 \text{var}(\vec{v}_1'\hat{\beta}) + \dots + \alpha_p^2 \text{var}(\vec{v}_p'\hat{\beta}) \\ &= \alpha_1^2 \left[\frac{\sigma^2}{\lambda_1} \right] + \dots + \alpha_p^2 \left[\frac{\sigma^2}{\lambda_p} \right] \\ &= \sigma^2 \left[\frac{\alpha_1^2}{\lambda_1} + \dots + \frac{\alpha_p^2}{\lambda_p} \right] \end{aligned} \quad (9.30)$$

9.6. Elección óptima de observaciones.



La expresión (9.28) 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 $(X'X)$, incrementándolos³. En lo que sigue, examinamos esta cuestión con más detalle.

Supongamos que tenemos un conjunto de N observaciones $(\vec{y} | X)$, 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.

Supondremos también en lo que sigue $(X'X)$ de rango completo, aunque quizá con acusada multicolinealidad⁴. Emplearemos los subíndices $N + 1$ y N para designar estimaciones respectivamente con y sin esta observación

³O suprimiéndolos. Los métodos de regresión sesgada del Capítulo 10 hacen explícita esta idea.

⁴Los resultados se pueden generalizar al caso en que $(X'X)$ es de rango deficiente, y sólo mediante la nueva fila \vec{x}_{N+1}' se hace $\vec{c}'\vec{\beta}$ estimable.

adicional. Tenemos entonces que:

$$\Sigma_{\hat{\beta}_N} = \sigma^2(X'X)^{-1} \quad (9.31)$$

$$\Sigma_{\hat{\beta}_{N+1}} = \sigma^2(X'X + \vec{x}_{N+1}\vec{x}_{N+1}')^{-1} \quad (9.32)$$

$$\sigma_{\vec{c}'\hat{\beta}_N}^2 = \sigma^2\vec{c}'(X'X)^{-1}\vec{c} \quad (9.33)$$

$$\sigma_{\vec{c}'\hat{\beta}_{N+1}}^2 = \sigma^2\vec{c}'(X'X + \vec{x}_{N+1}\vec{x}_{N+1}')^{-1}\vec{c} \quad (9.34)$$

Entonces,

$$\sigma_{\vec{c}'\hat{\beta}_N}^2 - \sigma_{\vec{c}'\hat{\beta}_{N+1}}^2 = \sigma^2\vec{c}'[(X'X)^{-1} - (X'X + \vec{x}_{N+1}\vec{x}_{N+1}')^{-1}]\vec{c} \quad (9.35)$$

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

$$\vec{a} = V'\vec{c} \quad (9.36)$$

$$\vec{z} = V'\vec{x}_{N+1} \quad (9.37)$$

$$D = V'(X'X)V \quad (9.38)$$

Entonces, (9.35) puede transformarse así:

$$\begin{aligned} \sigma_{\vec{c}'\hat{\beta}_N}^2 - \sigma_{\vec{c}'\hat{\beta}_{N+1}}^2 &= \sigma^2\vec{c}'VV'[(X'X)^{-1} - (X'X + \vec{x}_{N+1}\vec{x}_{N+1}')^{-1}]VV'\vec{c} \\ &= \sigma^2\vec{a}'[D^{-1} - V'(X'X + \vec{x}_{N+1}\vec{x}_{N+1}')^{-1}V]\vec{a} \\ &= \sigma^2\vec{a}'[D^{-1} - (V'(X'X + \vec{x}_{N+1}\vec{x}_{N+1}')V)^{-1}]\vec{a} \\ &= \sigma^2\vec{a}'[D^{-1} - (D + \vec{z}\vec{z}')^{-1}]\vec{a} \end{aligned} \quad (9.39)$$

Pero (véase Teorema A.2, pág. 221):

$$(D + \vec{z}\vec{z}')^{-1} = D^{-1} - \frac{D^{-1}\vec{z}\vec{z}'D^{-1}}{1 + \vec{z}'D^{-1}\vec{z}} \quad (9.40)$$

Sustituyendo (9.40) en (9.39):

$$\sigma_{\vec{c}'\hat{\beta}_N}^2 - \sigma_{\vec{c}'\hat{\beta}_{N+1}}^2 = \sigma^2\vec{a}' \left[\frac{D^{-1}\vec{z}\vec{z}'D^{-1}}{1 + \vec{z}'D^{-1}\vec{z}} \right] \vec{a} \quad (9.41)$$

$$= \sigma^2 \frac{\left(\sum_i \frac{a_i z_i}{\lambda_i} \right)^2}{\left(1 + \sum_i \frac{z_i^2}{\lambda_i} \right)} \quad (9.42)$$

Obsérvese que el problema de maximizar (9.35) carece de sentido si no imponemos restricciones, pues la expresión equivalente (9.42) es monótona

creciente al multiplicar \vec{z} por una constante k mayor que la unidad⁵. 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,

$$\Phi(\vec{z}) = \sigma^2 \frac{\left(\sum_i \frac{a_i z_i}{\lambda_i}\right)^2}{\left(1 + \sum_i \frac{z_i^2}{\lambda_i}\right)} - \mu \left(\sum_i z_i^2 - K^2\right) \quad (9.43)$$

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

$$\sigma^2 \frac{\left(\sum_i \frac{a_i z_i}{\lambda_i}\right) \frac{a_i}{\lambda_i} \left(1 + \sum_i \frac{z_i^2}{\lambda_i}\right) - \left(\sum_i \frac{a_i z_i}{\lambda_i}\right)^2 \frac{z_i}{\lambda_i}}{\left(1 + \sum_i \frac{z_i^2}{\lambda_i}\right)^2} - \mu z_i = 0 \quad (9.44)$$

Denominando:

$$A = \left(\sum_i \frac{a_i z_i}{\lambda_i}\right) \quad (9.45)$$

$$B = \left(1 + \sum_i \frac{z_i^2}{\lambda_i}\right) \quad (9.46)$$

las p igualdades anteriores toman la forma:

$$\frac{a_i A}{\lambda_i B} - \frac{z_i A^2}{\lambda_i B^2} - \frac{\mu z_i}{\sigma^2} = 0 \quad (9.47)$$

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

$$\mu = \frac{A^2}{K^2 B^2} \sigma^2 \quad (9.48)$$

y por consiguiente de (9.47) se obtiene:

$$\frac{a_i A}{\lambda_i B} - \frac{z_i A^2}{\lambda_i B^2} - \frac{A^2}{K^2 B^2} z_i = 0 \quad (i = 1, \dots, p) \quad (9.49)$$

$$z_i \left(\frac{1}{\lambda_i} + \frac{1}{K^2}\right) = \frac{B a_i}{A \lambda_i} \quad (i = 1, \dots, p) \quad (9.50)$$

⁵Observemos que al multiplicar \vec{z} por k el numerador queda multiplicado por k^2 , en tanto sólo una parte del denominador lo hace. Es pues claro que el numerador crece más que el denominador, y el cociente en consecuencia aumenta.

o sea:

$$z_i \propto \frac{a_i}{\lambda_i \left(\frac{1}{\lambda_i} + \frac{1}{K^2} \right)} = \frac{a_i}{1 + \frac{\lambda_i}{K^2}} \quad (9.51)$$

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

$$\vec{z} \propto (I + K^{-2}D)^{-1}\vec{a} \quad (9.52)$$

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

$$\begin{aligned} \vec{x}_{N+1} &= V\vec{z} \\ (\text{por (9.52)}) &\propto V(I + K^{-2}D)^{-1}\vec{a} \\ &= V(I + K^{-2}D)^{-1}V'V\vec{a} \\ (\text{por (9.36)}) &= V(I + K^{-2}D)^{-1}V'\vec{c} \\ &= [V(I + K^{-2}D)V']^{-1}\vec{c} \\ &= [I + K^{-2}(X'X)]^{-1}\vec{c} \end{aligned}$$

Recordemos que hemos obtenido una solución única para \vec{z} (y en consecuencia \vec{x}_{N+1}) sólo 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 (9.42) 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 σ^2 es fija, es claro que siempre preferiremos filas de módulo muy grande, pues si:

$$Y_i = m_i + \epsilon_i = \beta_0 + \dots + \beta_{p-1}x_{i,p-1} + \epsilon_i \quad (9.53)$$

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 ϵ_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 \vec{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-optimalidad⁶.

⁶Véase Silvey (1980), una monografía que trata el tema de diseño óptimo.

Capítulo 10

Regresión sesgada.

10.1. Introducción.

De acuerdo con el teorema de Gauss-Markov (Teorema 2.2, pág. 19), 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.

Si consideramos adecuado como criterio en la elección de un estimador \hat{c} su error cuadrático medio, ECM $\stackrel{\text{def}}{=} E[\hat{c} - c]^2$, y reparamos en que:

$$\begin{aligned} E[\hat{c} - c]^2 &= E[\hat{c} - E[\hat{c}] + E[\hat{c}] - c]^2 \\ &= E[\hat{c} - E[\hat{c}]]^2 + E[E[\hat{c}] - c]^2 + \underbrace{2 E[\hat{c} - E[\hat{c}]] [E[\hat{c}] - c]}_{=0} \\ &= \text{var}(\hat{c}) + (\text{sesgo } \hat{c})^2 \end{aligned} \tag{10.1}$$

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 (10.1).

El Capítulo 9 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. 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 indicó en la Sección 9.6, pág. 131.

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.

Observación 10.1 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.

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 resultado (Teorema 10.1, pág. 142) 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.

10.2. Una aproximación intuitiva.

Antes de introducir los estimadores sesgados más utilizados en la práctica, es útil ver sobre un ejemplo simple las ideas que explotan.

Ejemplo 10.1 Consideremos la siguiente situación. Tenemos dos poblaciones con media común μ y varianzas respectivas σ_1^2 , σ_2^2 . Nuestro objetivo es estimar μ , para lo que contamos con dos observaciones, una de cada población. Sean éstas X_1 , X_2 . Sabemos además que σ_2^2 es mucho mayor que σ_1^2 .

Es claro que

$$\hat{\mu} = \frac{1}{2}(X_1 + X_2) \quad (10.2)$$

es un estimador insesgado de μ . Su varianza será $\text{Var}(\hat{\mu}) = \sigma_1^2/4 + \sigma_2^2/4$.

¿Es de mínima varianza? No; y en general puede ser sumamente ineficiente. Imaginemos, por ejemplo, que $\sigma_1^2 = 1$ y $\sigma_2^2 = 99$; entonces, $\text{Var}(\hat{\mu}) = (\sigma_1^2 + \sigma_2^2)/4 = (1 + 99)/4 = 25$, mientras que $\hat{\mu}^* = X_1$, por ejemplo, sería también insesgado con $\text{Var}(\hat{\mu}^*) = 1$.

La conclusión a la que llegamos es que *es mejor prescindir de la observación X_2 —dando muy imprecisa información acerca del valor de μ — que utilizarla en pie de igualdad con X_1 .*

Si examinamos el ejemplo con más cuidado, se nos hace evidente que podemos hacerlo mejor: si nos limitamos a estimadores lineales —por simplicidad— cualquier estimador insesgado será de la forma

$$\hat{\mu}^{**} = \delta_1 X_1 + \delta_2 X_2$$

con $\delta_1 + \delta_2 = 1$ (pues de otro modo al tomar valor medio en (10.3), no obtendríamos μ , como requiere la condición de insesgadez).

Podemos a continuación plantearnos cuáles son δ_1 y $\delta_2 = 1 - \delta_1$ óptimos. De (10.3) deducimos que

$$\begin{aligned} \text{Var}(\hat{\mu}^{**}) &= \delta_1^2 \sigma_1^2 + \delta_2^2 \sigma_2^2 \\ &= \delta_1^2 \cdot 1 + (1 - \delta_1)^2 \cdot 99 \\ &= 99 - 198\delta_1 + 100\delta_1^2 \end{aligned}$$

Derivando respecto a δ_1 e igualando a cero obtenemos $\delta_1 = 99/100$ y consecuentemente $\delta_2 = 1/100$. Fácilmente se comprueba que se trata de un mínimo. El estimador insesgado de varianza mínima es por tanto:

$$\hat{\mu}^{**} = \frac{99}{100} X_1 + \frac{1}{100} X_2.$$

El resultado parece lógico; debemos ponderar las dos observaciones dando más peso a la más fiable. La segunda conclusión a que llegamos es que cuando tengamos observaciones con grado de precisión muy variable, *convendrá ponderarlas de forma inversamente proporcional a sus respectivas varianzas.*

FIN DEL EJEMPLO ■

El ejemplo anterior pretende ilustrar dos principios, que se resumen en uno: es mejor prescindir de información imprecisa que hacerle *demasiado* caso. El primer estimador construido, $\hat{\mu}^*$, prescindía directamente de X_2 ; el segundo, $\hat{\mu}^{**}$, se servía de dicha observación pero *haciéndole poco caso*.

Se ha razonado sobre estimadores a los que hemos impuesto la condición de ser insesgados, por mantener el ejemplo simple, pero esta condición es inesencial. (De hecho, como veremos a continuación, todavía sería posible mejorar $\hat{\mu}^{**}$ en términos de ECM si tolerásemos un sesgo.)

¿Qué implicaciones tiene lo anterior sobre la estimación de $\vec{\beta}$ (o, en general, de $\vec{c}'\vec{\beta}$) en un modelo lineal? Recordemos la discusión en la Sección 9.5.

El estimador de cualquier forma lineal $\vec{c}'\vec{\beta}$ puede escribirse como combinación lineal de $\vec{v}'_1\hat{\beta}, \vec{v}'_2\hat{\beta}, \dots, \vec{v}'_p\hat{\beta}$, según muestra (9.29), pág. 131. Además, $\vec{v}'_i\hat{\beta}$ para $i = 1, \dots, p$ son variables aleatorias incorreladas¹ con varianzas respectivas $\text{Var}(\vec{v}'_i\hat{\beta}) = \sigma^2/\lambda_i$, (9.26), pág. 130.

Tenemos pues $\vec{c}'\vec{\beta}$ puede escribirse como combinación lineal de “observaciones” $\vec{v}'_i\hat{\beta}$ con varianzas muy diferentes. Al igual que en el Ejemplo 10.1 al estimar μ , podemos tener interés en prescindir de algunas de estas “observaciones” $\vec{v}'_i\hat{\beta}$, ó atenuarlas, si sus varianzas son muy grandes; ello acontecerá cuando los valores propios λ_i sean muy pequeños.

Los estimadores que se presentan a continuación hacen precisamente esto. El estimador en componentes principales de la Sección 10.4 prescinde de algunas $\vec{v}'_i\hat{\beta}$; el estimador *ridge* de la Sección 10.3 atenúa las $\vec{v}'_i\hat{\beta}$ más inestables. Volveremos de nuevo sobre la cuestión en la Sección 10.4, pág. 153.

10.3. Regresión ridge.

Error cuadrático medio del estimador mínimo cuadrático ordinario

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

$$\text{ECM}(\hat{\beta}) = E[(\hat{\beta} - \vec{\beta})'(\hat{\beta} - \vec{\beta})] \quad (10.3)$$

que podemos ver también como el valor medio del cuadrado de la distancia euclídea ordinaria entre $\hat{\beta}$ y $\vec{\beta}$. Supondremos $(X'X)$ de rango total, y por tanto que $(X'X)^{-1}$ existe (este supuesto se puede relajar). Como $E[\hat{\beta}] = \vec{\beta}$ y $\Sigma_{\hat{\beta}} = \sigma^2(X'X)^{-1}$, tenemos que:

$$\begin{aligned} \text{ECM}(\hat{\beta}) &= E[\text{traza } (\hat{\beta} - \vec{\beta})'(\hat{\beta} - \vec{\beta})] \\ &= E[\text{traza } (\hat{\beta} - \vec{\beta})(\hat{\beta} - \vec{\beta})'] \\ &= \sigma^2 \text{traza } (X'X)^{-1} \\ &= \sigma^2 \text{traza } (X'X)^{-1}VV' \quad (V = \text{diagonalizadora de } (X'X)^{-1}) \\ &= \sigma^2 \text{traza } V'(X'X)^{-1}V \\ &= \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i}, \end{aligned} \quad (10.4)$$

¹Independientes, si se verifica el supuesto de normalidad.

en que los λ_i son los valores propios de la matriz $(X'X)$. (Recuérdese que los vectores propios de las matrices $(X'X)$ y $(X'X)^{-1}$ son los mismos, y los valores propios de una los inversos de los de la otra.)

Clase de estimadores ridge

Definición 10.1 Definiremos el estimador ridge de parámetro k así:

$$\hat{\beta}^{(k)} = (X'X + kI)^{-1}X'Y \quad (10.5)$$

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:

$$\begin{aligned} \hat{\beta}^{(k)} &= (X'X + kI)^{-1}(X'X)(X'X)^{-1}X'Y \\ &= (X'X + kI)^{-1}(X'X)\hat{\beta} \\ &= [(X'X)^{-1}(X'X + kI)]^{-1}\hat{\beta} \\ &= [I + k(X'X)^{-1}]^{-1}\hat{\beta} \\ &= Z\hat{\beta} \end{aligned} \quad (10.6)$$

siendo $Z \stackrel{\text{def}}{=} [I + k(X'X)^{-1}]^{-1}$.

El Teorema 10.1, que muestra la superioridad del estimador *ridge* sobre el MCO para algún valor de k , es consecuencia del Lema 10.1 a continuación.

Lema 10.1 El error cuadrático medio del estimador ridge de parámetro k viene dado por la expresión

$$ECM[\hat{\beta}^{(k)}] = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^2} + \sum_{i=1}^p \frac{k^2 \alpha_i^2}{(\lambda_i + k)^2} \quad (10.7)$$

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

DEMOSTRACIÓN:

El ECM del estimador ridge que habremos de comparar con (10.4) es:

$$\begin{aligned}
 ECM[\hat{\beta}^{(k)}] &= E[(\hat{\beta}^{(k)} - \vec{\beta})'(\hat{\beta}^{(k)} - \vec{\beta})] \\
 (\text{por (10.6)}) &= E[(Z\hat{\beta} - \vec{\beta})'(Z\hat{\beta} - \vec{\beta})] \\
 &= E[(Z\hat{\beta} - Z\vec{\beta} + Z\vec{\beta} - \vec{\beta})'(Z\hat{\beta} - Z\vec{\beta} + Z\vec{\beta} - \vec{\beta})] \\
 &= \underbrace{E[(Z\hat{\beta} - Z\vec{\beta})'(Z\hat{\beta} - Z\vec{\beta})]}_{(a)} + \underbrace{(Z\vec{\beta} - \vec{\beta})'(Z\vec{\beta} - \vec{\beta})}_{(b)}
 \end{aligned} \tag{10.8}$$

Obsérvese que el primer término (a) es la suma de varianzas de los elementos de $\hat{\beta}^{(k)}$, mientras que (b) es la suma de los sesgos al cuadrado de dichos elementos. Examinemos por separado los dos sumandos de la expresión anterior:

$$\begin{aligned}
 (a) &= E[(\hat{\beta} - \vec{\beta})'Z'Z(\hat{\beta} - \vec{\beta})] \\
 &= E[\text{traza}\{(\hat{\beta} - \vec{\beta})'Z'Z(\hat{\beta} - \vec{\beta})\}] \\
 &= E[\text{traza}\{(\hat{\beta} - \vec{\beta})(\hat{\beta} - \vec{\beta})'Z'Z\}] \\
 &= \text{traza}\{E(\hat{\beta} - \vec{\beta})(\hat{\beta} - \vec{\beta})'Z'Z\} \\
 &= \sigma^2 \text{traza} [(X'X)^{-1}Z'Z]
 \end{aligned} \tag{10.9}$$

$$\begin{aligned}
 &= \sigma^2 \text{traza} \left[(X'X)^{-1} [I + k(X'X)^{-1}]^{-1} [I + k(X'X)^{-1}]^{-1} \right] \\
 &= \sigma^2 \text{traza} \left[(X'X) + kI + kI + k^2(X'X)^{-1} \right]^{-1} \\
 &= \sigma^2 \text{traza} \left\{ [(X'X) + 2kI + k^2(X'X)^{-1}]^{-1} VV' \right\} \\
 &= \sigma^2 \text{traza} \left[V'[(X'X) + 2kI + k^2(X'X)^{-1}]^{-1} V \right]
 \end{aligned} \tag{10.10}$$

$$= \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i + 2k + \lambda_i^{-1}k^2} \tag{10.11}$$

$$= \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^2}. \tag{10.12}$$

La obtención de la expresión (10.9) hace uso de el habitual intercambio de los operadores de traza y valor medio, así como del hecho de que si $\hat{\beta}$ es el estimador MCO y $X'X$ es de rango completo, $E[(\hat{\beta} - \vec{\beta})(\hat{\beta} - \vec{\beta})'] = \sigma^2(X'X)^{-1}$ (Teorema 2.2, pág. 19). En el paso de (10.10) a (10.11) 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.

Tomando ahora el segundo término de (10.8),

$$\begin{aligned}
 (b) &= (Z\vec{\beta} - \vec{\beta})'(Z\vec{\beta} - \vec{\beta}) \\
 &= \vec{\beta}'(Z - I)'(Z - I)\vec{\beta} \\
 &= \vec{\beta}' \left([I + k(X'X)^{-1}]^{-1} - I \right)' \left([I + k(X'X)^{-1}]^{-1} - I \right) \vec{\beta} \\
 &= k^2 \vec{\alpha}' (\Lambda + kI)^{-2} \vec{\alpha} \tag{10.13}
 \end{aligned}$$

$$\begin{aligned}
 &= \text{traza} \left[k^2 \vec{\alpha}' (\Lambda + kI)^{-2} \vec{\alpha} \right] \\
 &= \sum_{i=1}^p \frac{k^2 \alpha_i^2}{(\lambda_i + k)^2} \tag{10.14}
 \end{aligned}$$

El paso a (10.13) desde la expresión anterior hace uso de que $\vec{\alpha} = V'\vec{\beta}$. Sustituyendo (10.12) y (10.14) en (10.8) se obtiene (10.7) ■

El Teorema 10.1 se sigue casi inmediatamente del resultado anterior.

Teorema 10.1 *Hay algún valor de $k > 0$ para el que $ECM[\hat{\beta}^{(k)}]$ dado por (10.7) es estrictamente menor que el ECM del estimador MCO dado por (10.4).*

DEMOSTRACIÓN:

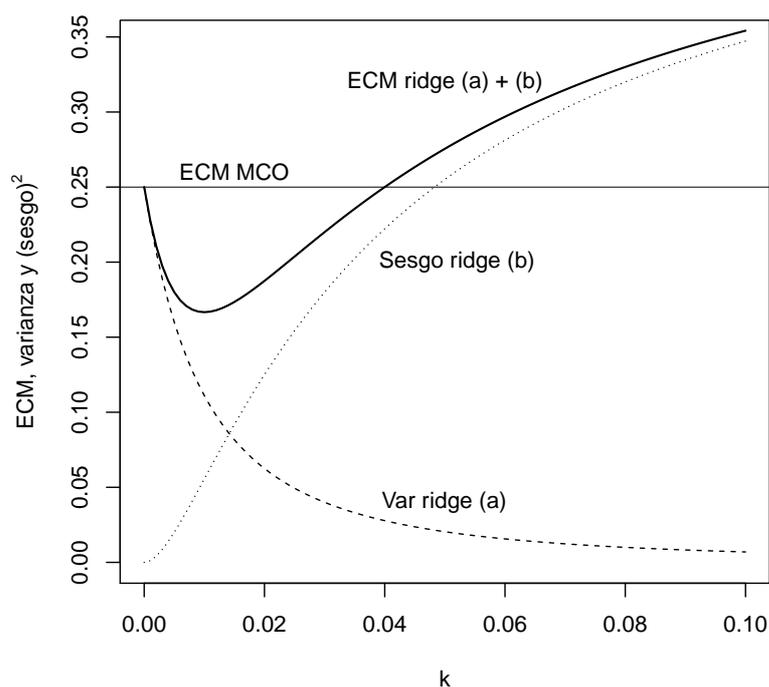
Hemos visto más arriba que cuando $k = 0$, el estimador ridge $\hat{\beta}^{(k)}$ coincide con el MCO. Por consiguiente, para $k = 0$ la expresión (10.7) debe coincidir con (10.4), como en efecto puede comprobarse que sucede. Derivando (10.7) respecto de k , es fácil comprobar que la derivada en $k = 0$ existe y es $-2\sigma^2 \sum_{i=1}^p \lambda_i^{-2}$, claramente negativa. Por consiguiente, siempre podremos (incrementando ligeramente k) lograr que:

$$ECM[\hat{\beta}^{(k)}] < ECM[\hat{\beta}^{(0)}] = ECM[\hat{\beta}] \tag{10.15}$$

lo que demuestra el teorema. ■

Una percepción intuitiva del resultado anterior la proporciona la comparación de las expresiones (10.4) y (10.8), 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 (10.4) puede hacerse arbitrariamente grande si $\lambda_i \approx 0$ para algún i . La expresión (10.12) está a

Figura 10.1: Componentes del $ECM(\hat{\beta}^{(k)})$ en el estimador *ridge*. Las líneas de trazos y puntos representa respectivamente la varianza y $(\text{sesgo})^2$ de $\hat{\beta}^{(k)}$ en función de k . La curva sólida representa $ECM[\hat{\beta}^{(k)}]$. La línea horizontal es la varianza (y ECM) del estimador $\hat{\beta}$ MCO.



cobijo de tal eventualidad, pues ninguno de los sumandos puede crecer por encima de λ_i/k^2 .

La Figura 10.1 muestra en un caso concreto cómo varían en función de k los componentes (a) y (b) de (10.8), y su suma. Como término de comparación se ha representado mediante una línea horizontal la varianza del $\hat{\beta}$ MCO (igual a su varianza, puesto que es insesgado). Puede verse que, tal como el Teorema 10.1 establece, hay valores de k en que el $ECM(\hat{\beta}^{(k)})$ desciende por debajo del $ECM(\hat{\beta})$; ocurre para valores de k menores que 0.039 aproximadamente.

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 de k 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

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

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

$$k_{GCV} = \arg \min_k \frac{\|(I - A(k))\vec{y}\|^2}{[\text{traza}(I - A(k))]^2}. \quad (10.16)$$

Sobre la justificación de dicha elección puede verse Eubank (1988) o Brown (1993), por ejemplo; no podemos entrar aquí en detalles. Baste decir que la expresión que se minimiza en (10.16) 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 (10.16) continúa siendo una suma de cuadrados de los residuos y el denominador el 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.

$$k_{HKB} = (p - 2)\hat{\sigma}^2/\hat{\beta}'\hat{\beta} \quad (10.17)$$

$$k_{LW} = (p - 2)\hat{\sigma}^2\text{traza}(X'X)/(p\hat{\beta}'(X'X)\hat{\beta}) \quad (10.18)$$

$$k_{MUR} = \arg \min_k \left[\hat{\sigma}^2 \sum_i \frac{\lambda_i - k}{\lambda_i(\lambda_i + k)} + k^2 \sum_i \frac{\hat{\alpha}_i^2}{(\lambda_i + k)^2} \right] \quad (10.19)$$

El criterio (10.17) fue propuesto por Hoerl et al. (1975) y tiene una justificación bayesiana. El criterio (10.18) fue propuesto en Lawless and Wang (1976). El criterio (10.19) estima el ECM del estimador ridge insesgadamente y toma el k que minimiza dicha estimación.

Observación 10.2 En las ecuaciones (10.17)–(10.19), p es el orden y rango de la matrix $(X'X)$. En caso de que $(X'X)$ sea de rango deficiente r , $r < p$, puede sustituirse éste por p tomando como $\vec{\beta}$ el estimador mínimo cuadrático de mínima longitud; ver detalles en Brown (1993), pág. 63.

Comentarios adicionales

Es evidente que la forma del ECM propuesto pondera por igual las discrepancias en la estimación de un β_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 adecuada² “tipificando” los $(\hat{\beta} - \vec{\beta})$.

Es habitual no sólo normalizar sino también centrar tanto las columnas de X como \vec{y} . El parámetro β_0 se sustrae así al proceso de estimación ridge, restaurándolo al final.

Finalmente, es de interés señalar que el estimador *ridge* puede verse desde distintos puntos de vista. Uno de ellos lo interpreta como un estimador bayesiano, en la línea esbozada en los Ejercicios 4.6 y 4.7, pág. 58.

R: Ejemplo 10.1 (ejemplo de regresión ridge)

El siguiente código muestra el uso de regresión ridge sobre un conjunto de datos acusadamente colineal. La Figura 10.2 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 10.2; el segundo sí.

```
> options(digits = 4)
> options(columns = 40)
> library(MASS)
> data(longley)
> names(longley)[1] <- "y"
> longley[1:3, ]

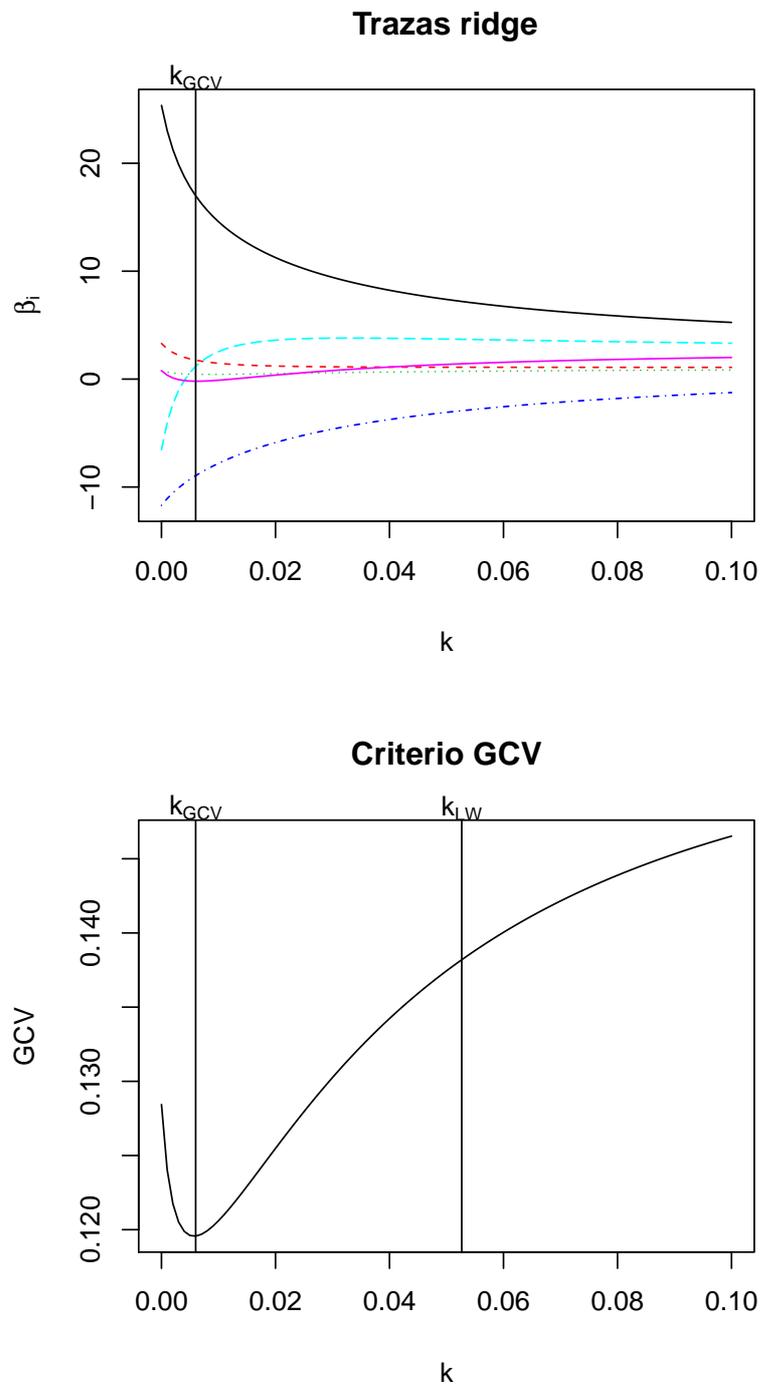
          y   GNP Unemployed Armed.Forces
1947 83.0 234.3      235.6      159.0
1948 88.5 259.4      232.5      145.6
1949 88.2 258.1      368.2      161.6
      Population Year Employed
1947      107.6 1947    60.32
1948      108.6 1948    61.12
1949      109.8 1949    60.17

> longley.mco <- lm(y ~ ., longley)
> summary(longley.mco)

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

²Es 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.

Figura 10.2: Trazas ridge y GVC para los datos longley



```

Residuals:
  Min       1Q   Median       3Q      Max
-2.009 -0.515  0.113  0.423  1.550

Coefficients:
              Estimate Std. Error t value
(Intercept) 2946.8564  5647.9766   0.52
GNP           0.2635    0.1082    2.44
Unemployed    0.0365    0.0302    1.21
Armed.Forces  0.0112    0.0155    0.72
Population   -1.7370    0.6738   -2.58
Year         -1.4188    2.9446   -0.48
Employed      0.2313    1.3039    0.18

              Pr(>|t|)
(Intercept)  0.614
GNP           0.038 *
Unemployed    0.258
Armed.Forces  0.488
Population    0.030 *
Year          0.641
Employed      0.863
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.19 on 9 degrees of freedom
Multiple R-squared: 0.993,      Adjusted R-squared: 0.988
F-statistic: 203 on 6 and 9 DF,  p-value: 4.43e-09

```

Nótese la fuerte multicolinealidad, aparente en los reducidos t -ratios y elevada R^2 . Probemos ahora regresión *ridge* con valores de k (= λ) entre 0 y 0.1 variando de milésima en milésima. Imprimiremos a continuación las estimaciones correspondientes a los tres primeros valores de k ensayados. Cuando $k = 0$, deben coincidir las estimaciones con las obtenidas por MCO.

```

> 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

> coef(longley.rr)[1:3, ]

                GNP Unemployed Armed.Forces
0.000 2947 0.2635    0.03648    0.011161
0.001 1896 0.2392    0.03101    0.009372
0.002 1166 0.2210    0.02719    0.008243
      Population   Year Employed
0.000    -1.737 -1.4188  0.23129
0.001    -1.644 -0.8766  0.10561
0.002    -1.565 -0.5011  0.03029

```

La función `select` aplicada al objeto que devuelve `lm.ridge` devuelve los valores óptimos de tres de los criterios mencionados más arriba.

```

> select(longley.rr)

modified HKB estimator is 0.006837
modified L-W estimator is 0.05267
smallest value of GCV   at 0.006

```

Podemos seleccionar el k óptimo de acuerdo, por ejemplo, al criterio GCV, y hacer regresión *ridge* con él:

```

> nGCV <- which.min(longley.rr$GCV)
> lGCV <- longley.rr$lambda[nGCV]
> lm.ridge(y ~ ., longley, lambda = lGCV)

                GNP   Unemployed
-3.144e+02  1.765e-01  1.937e-02
Armed.Forces Population   Year
 6.565e-03 -1.328e+00  2.556e-01
  Employed
-5.812e-02

```

El código a continuación genera las gráficas en la Figura 10.2.

```

> par(mfrow = c(2, 1))
> matplot(longley.rr$lambda, t(longley.rr$coef),
+       type = "l", xlab = expression(k),
+       ylab = expression(beta[i]))
> 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")
> 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)

```

FIN DEL EJEMPLO ■

10.4. Regresión en componentes principales.

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 regresión lineal que consideramos se puede escribir así:

$$\vec{y} = \vec{1}\beta_0 + W\vec{\beta}^* + \vec{\epsilon} \quad (10.20)$$

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:

$$V'(W'W)V = \Lambda \quad (\iff W'W = V\Lambda V') \quad (10.21)$$

Sean $\vec{v}_1, \dots, \vec{v}_{p-1}$ los vectores columna de V . Llamaremos *componentes principales* de W a los vectores $\vec{u}_1, \dots, \vec{u}_{p-1}$ definidos así:

$$\begin{aligned}\vec{u}_1 &= W\vec{v}_1 \\ \vec{u}_2 &= W\vec{v}_2 \\ &\vdots \\ \vec{u}_{p-1} &= W\vec{v}_{p-1}\end{aligned}\tag{10.22}$$

o abreviadamente:

$$U = WV\tag{10.23}$$

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, (10.20) puede transformarse así:

$$\vec{y} = \vec{1}\beta_0 + W\vec{\beta}^* + \vec{\epsilon}\tag{10.24}$$

$$= \vec{1}\beta_0 + WV V' \vec{\beta}^* + \vec{\epsilon}\tag{10.25}$$

$$= \vec{1}\beta_0 + U\vec{\gamma}^* + \vec{\epsilon}\tag{10.26}$$

Teniendo en cuenta (ver Problema 10.2) que $\vec{1} \perp \vec{u}_i$, ($i = 1, \dots, p-1$), el vector de estimadores puede escribirse así:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\gamma}^* \end{pmatrix} = \begin{pmatrix} \bar{y} \\ (U'U)^{-1}U'\vec{y} \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \Lambda^{-1}U'\vec{y} \end{pmatrix}\tag{10.27}$$

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 estimadores de los parámetros originales $\vec{\beta}^*$ a partir de $\hat{\gamma}^*$. 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 $\hat{\gamma}^*$:

$$\hat{\beta}_{CP}^* = V \begin{pmatrix} \hat{\gamma}_{(q)}^* \\ \vec{0} \end{pmatrix}.\tag{10.28}$$

Necesitamos por tanto criterios para escoger los estimadores $\hat{\gamma}_i$ que incluimos en $\hat{\gamma}_{(q)}^*$ y los que reemplazamos por cero en (10.28).

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 λ_i . Como quiera que la varianza de $\hat{\gamma}_i^*$ es $\sigma^2\lambda_i^{-1}$ (véase (9.26), pág. 130), una estrategia consistiría en tomar los $\hat{\gamma}_i^*$ asociados a λ_i más grande (es decir, con menos varianza), despreciando los restantes. El número de componentes principales a retener (= el número de λ_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

$$\|P_U\vec{y}\|^2 = \|U\hat{\gamma}^*\|^2 = \hat{\gamma}_1^{*2}\|\vec{u}_1\|^2 + \cdots + \hat{\gamma}_{p-1}^{*2}\|\vec{u}_{p-1}\|^2, \quad (10.29)$$

la última igualdad haciendo uso de la ortogonalidad entre las columnas de U . Entonces, $SSR = \|P_U\vec{y}\|^2$, y $SSE = \|\vec{y} - \vec{\bar{y}}\|^2 - \|U\hat{\gamma}^*\|^2$.

2. Se contrasta la hipótesis de nulidad para cada uno de los parámetros, ($H_i: \hat{\gamma}_i^* = 0, i = 1, \dots, p-1$), mediante el estadístico:

$$Q_i = \frac{N-p}{1} \times \frac{\hat{\gamma}_i^{*2}\|\vec{u}_i\|^2}{SSE} \sim \mathcal{F}_{1, N-p} \quad (10.30)$$

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; por tanto, la diferencia de suma de cuadrados explicada con y sin el regresor \vec{u}_i es precisamente $\hat{\gamma}_i^{*2}\|\vec{u}_i\|^2$.

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 $\hat{\beta}_{CP}^*$ se obtienen mediante la transformación (10.28).

Nótese que mientras que la estrategia precedente consistía en desechar componentes principales asociadas a reducido λ_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 $\hat{\gamma}_i^* \gg 0$. Puede ocurrir, sin embargo, que una componente principal asociada a un λ_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 análisis³.

Estrategia mixta. Propuesta por Jolliffe (1986), ordena los $\hat{\gamma}_i^*$ de menor a mayor λ_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 λ_i mayor, por tanto). Todos los $\hat{\gamma}_i^*$ retenidos componen el vector $\hat{\gamma}_{(q)}^*$.

Validación cruzada. Computacionalmente muy laboriosa. Puede ocurrir que al omitir distintas observaciones, dos componentes principales permuten su orden. Véanse detalles en Brown (1993).

Propiedades del estimador en componentes principales

El sesgo de $\hat{\beta}_{CP}^*$ es:

$$E[\hat{\beta}_{CP}^* - \vec{\beta}^*] = E\left[V \begin{pmatrix} \hat{\gamma}_{(q)}^* \\ 0 \end{pmatrix} - V\vec{\gamma}^*\right] = - \sum_{i=q+1}^{p-1} \hat{\gamma}_i^* \vec{v}_i \quad (10.31)$$

y su matriz de covarianzas:

$$\Sigma_{\hat{\beta}_{CP}^*} = V \left(\sigma^2 \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} \Lambda^{-1} \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} \right) V' \quad (10.32)$$

$$= \sigma^2 \sum_{i=1}^q \lambda_i^{-1} \vec{v}_i \vec{v}_i' \quad (10.33)$$

$$\leq \sigma^2 \sum_{i=1}^{p-1} \lambda_i^{-1} \vec{v}_i \vec{v}_i' \quad (10.34)$$

$$= \sigma^2 (W'W)^{-1} \quad (10.35)$$

³Pero este criterio no es unánimemente compartido. Véase Hocking (1976).

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:

$$\sigma^2 \sum_{i=q+1}^{p-1} \lambda_i^{-1} \vec{v}_i \vec{v}_i' \quad (10.36)$$

y será importante si entre las componentes principales excluidas como regresores hay alguna asociada a un λ_i muy pequeño.

Las expresiones (10.31) y (10.32)–(10.35) muestran el conflicto varianzas sesgo en el caso de la regresión en componentes principales. De (10.31) se deduce la siguiente expresión para la suma de los sesgos al cuadrado:

$$[E(\hat{\beta}_{CP}^*) - \vec{\beta}^*]' [E(\hat{\beta}_{CP}^*) - \vec{\beta}^*] = \sum_{i=q+1}^{p-1} (\hat{\gamma}_i^*)^2 \quad (10.37)$$

Es interesante comparar el estimador en componentes principales con el estimador ridge, y examinarlo a la luz del análisis efectuado en el Capítulo 9. 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 λ_i) o de reducido $Q_i \propto (\hat{\gamma}_i^*)^2 \lambda_i$; 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 expresión:

$$\hat{\beta}_{CP}^* = V \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} \hat{\gamma}^* = V \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} \Lambda^{-1} U' \vec{y} \quad (10.38)$$

con la del estimador ridge equiparable⁴:

$$\hat{\beta}^{(k)} = (W'W + kI)^{-1} W' \vec{y} \quad (10.39)$$

$$= VV'(W'W + kI)^{-1} VV'W' \vec{y} \quad (10.40)$$

$$= V(\Lambda + kI)^{-1} U' \vec{y} \quad (10.41)$$

En (10.38) sólo q columnas de $U' \vec{y}$ se utilizan; en (10.41), todas, si bien las que corresponden a componentes principales con λ_i más pequeño reciben

⁴Es decir, tras haber centrado y normado los regresores y segregado la columna de “unos”.

una ponderación menor, al ser divididas por $\lambda_i + k$ en lugar de por λ_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 10.2 (*regresión en componentes principales*)

La función `regCP` que sigue traduce directamente de la teoría expuesta el método para llevar a cabo estimación en componentes principales. 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.

```
> regCP <- function(X, y, tomar = NULL,
+   sig = 0.05) {
+   X.c <- scale(X, scale = FALSE)
+   y.c <- scale(y, scale = FALSE)
+   W <- scale(X.c, center = FALSE)/sqrt(nrow(X) -
+     1)
+   WW <- crossprod(W)
+   factores.escala <- X.c[1, ]/W[1, ]
+   N <- nrow(X)
+   p <- ncol(X)
+   res <- eigen(WW)
+   V <- res$vectors
+   landas <- res$values
+   U <- W %*% V
+   gamas <- (1/landas) * t(U) %*% y.c
+   if (is.null(tomar)) {
+     fit <- lsfit(X.c, y.c, intercept = FALSE)
+     SSE <- sum(fit$residuals^2)
+     qi <- (N - p) * (gamas * landas)^2/SSE
+     tomar <- (1:p)[sig > (1 - pf(qi,
+       1, N - p))]
+   }
+   betasCPstar <- V[, tomar] %*% gamas[tomar]
```

```

+   betasCP <- betasCPstar/factores.escala
+   m.X <- apply(X, 2, mean)
+   m.Y <- mean(y)
+   beta0 <- m.Y - sum(m.X * betasCP)
+   betasCP <- c(beta0, betasCP)
+   names(betasCP) <- c("Intercept", dimnames(X)[[2]])
+   return(list(betasCP = betasCP, landas = landas,
+             CP.usadas = tomar))
+ }

```

Veamos el modo de emplearla, con los datos `longley`, frecuentemente empleados como banco de pruebas por su muy acusada multicolinealidad:

```

> library(MASS)
> data(longley)
> y <- longley[, 1]
> X <- as.matrix(longley[, -1])
> regCP(X, y, tomar = 1:3)

$betasCP
      Intercept          GNP    Unemployed
-9.731e+02    2.459e-02    9.953e-03
Armed.Forces  Population          Year
 1.553e-02    3.391e-01    4.967e-01
      Employed
 7.239e-01

$landas
[1] 4.5478430 1.1858692 0.2517070 0.0124261
[5] 0.0018422 0.0003126

$CP.usadas
[1] 1 2 3

```

Una comprobación útil consiste en ver que el estimador en CP, cuando se utilizan todas las componente principales, coincide con el estimador MCO. Veámoslo:

```

> regCP(X, y, tomar = 1:ncol(X))

$betasCP
      Intercept          GNP    Unemployed

```

```

2946.85636      0.26353      0.03648
Armed.Forces   Population      Year
      0.01116      -1.73703     -1.41880
Employed
      0.23129

```

```

$landas
[1] 4.5478430 1.1858692 0.2517070 0.0124261
[5] 0.0018422 0.0003126

```

```

$CP.usadas
[1] 1 2 3 4 5 6

```

```

> lsfit(X, y)$coefficients

```

```

Intercept      GNP      Unemployed
2946.85636      0.26353      0.03648
Armed.Forces   Population      Year
      0.01116      -1.73703     -1.41880
Employed
      0.23129

```

Para que la función seleccione aquellas componentes principales con un nivel de significación de sus parámetros asociados prefijado, la invocamos así:

```

> regCP(X, y, sig = 0.1)

```

```

$betasCP
Intercept      GNP      Unemployed
-961.37468      0.02372      0.01373
Armed.Forces   Population      Year
      0.01991      0.33197      0.49223
Employed
      0.66205

```

```

$landas
[1] 4.5478430 1.1858692 0.2517070 0.0124261
[5] 0.0018422 0.0003126

```

```

$CP.usadas
[1] 1 2

```

FIN DEL EJEMPLO ■

10.5. Regresión en raíces latentes



Consideramos el modelo:

$$\vec{y} = \vec{1}\beta_0 + W\vec{\beta}^* + \vec{\epsilon} \quad (10.42)$$

o alternativamente:

$$\vec{y}^* = W\vec{\beta}^* + \vec{\epsilon} \quad (10.43)$$

en que tanto los regresores como la variable respuesta \vec{y}^* han sido normalizados y centrados. Es decir, $\vec{y}^* = \eta^{-1}(\vec{y} - \vec{\bar{y}})$ siendo $\eta^2 = \sum_{i=1}^N (y_i - \bar{y})^2$. Si construimos la matriz $N \times p$ siguiente:

$$A = [\vec{y}^* \mid W] \quad (10.44)$$

tenemos que la matriz $(A'A)$ es una matriz de correlación (tiene “unos” en la diagonal principal, es simétrica y semidefinida positiva). Sea $V = (\vec{v}_1 \mid \dots \mid \vec{v}_p)$ la matriz que la diagonaliza:

$$V'(A'A)V = \Lambda \iff V\Lambda V' = A'A \quad (10.45)$$

Entonces, utilizando (10.44), tenemos

$$A\vec{v}_j = v_{0j}\vec{y}^* + W\vec{v}_j^{(0)}, \quad (j = 1, \dots, p) \quad (10.46)$$

dónde $\vec{v}_j^{(0)}$ es \vec{v}_j desprovisto de su primer elemento:

$$\vec{v}_j = \begin{bmatrix} v_{0j} \\ \vec{v}_j^{(0)} \end{bmatrix}.$$

Tomando norma al cuadrado de (10.46),

$$\begin{aligned} \|A\vec{v}_j\|^2 &= \|v_{0j}\vec{y}^* + W\vec{v}_j^{(0)}\|^2 \\ &= \sum_{i=1}^N \left(\vec{y}_i^* v_{0j} + \sum_{k=1}^{p-1} W_{ik} v_{kj} \right)^2 \end{aligned} \quad (10.47)$$

en que v_{kj} es la k -ésima coordenada de $\vec{v}_j^{(0)}$. Como por otra parte

$$\begin{aligned} \|A\vec{v}_j\|^2 &= \vec{v}_j'(A'A)\vec{v}_j \\ &= \lambda_j, \end{aligned} \quad (10.48)$$

igualando (10.47) y (10.48) deducimos que si $\lambda_j \approx 0$

$$y_i^* v_{0j} \approx - \sum_{k=1}^{p-1} W_{ik} v_{kj} \quad \forall i \in [1, \dots, N] \quad (10.49)$$

Si, además, $v_{0j} \neq 0$, podemos escribir:

$$\vec{y}^* \approx -v_{0j}^{-1} W \vec{v}_j^{(0)} \stackrel{\text{def}}{=} \hat{y}_{(j)}^* \quad (10.50)$$

Como $\vec{y}^* = \eta^{-1}(\vec{y} - \vec{\bar{y}})$, $\vec{y} = \vec{\bar{y}} + \eta \vec{y}^*$ y denominando

$$\hat{y}_{(j)} = \vec{\bar{y}} + \eta \hat{y}_{(j)}^* \quad (10.51)$$

tenemos:

$$\begin{aligned} (\vec{y} - \hat{y}_{(j)})'(\vec{y} - \hat{y}_{(j)}) &= \eta^2 (\vec{y}^* - \hat{y}_{(j)}^*)'(\vec{y}^* - \hat{y}_{(j)}^*) \\ &= (v_{0j} \vec{y}^* - v_{0j} \hat{y}_{(j)}^*)'(v_{0j} \vec{y}^* - v_{0j} \hat{y}_{(j)}^*) \frac{\eta^2}{v_{0j}^2} \\ &= (A \vec{v}_j)'(A \vec{v}_j) \frac{\eta^2}{v_{0j}^2} \\ &= \frac{\lambda_j \eta^2}{v_{0j}^2} \end{aligned} \quad (10.52)$$

Nótese que la aproximación de \vec{y}^* en (10.50) y suma de cuadrados de los residuos en (10.52), hacen uso exclusivamente de una parte de la información disponible; la de que λ_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 $\hat{y}_{(i)}$ ($i = 1, \dots, p$), debidamente ponderadas por coeficientes d_i a determinar:

$$\begin{aligned} \hat{y} &= \sum_{i=1}^p d_i \hat{y}_{(i)} \\ [\text{usando (10.50) y (10.51)}] &= \sum_{i=1}^p d_i \left(\vec{\bar{y}} + W(-v_{0i}^{-1} \vec{v}_i^{(0)} \eta) \right) \\ &= \left(\sum_{i=1}^p d_i \right) \vec{\bar{y}} + W \left(- \sum_{i=1}^p d_i v_{0i}^{-1} \vec{v}_i^{(0)} \eta \right) \end{aligned}$$

Por otro lado, de (10.42) tenemos

$$\hat{\beta}_0 \vec{1} + W \hat{\beta}^*$$

que junto con la igualdad precedente proporciona:

$$\hat{\beta}_0 = \bar{y} \left(\sum_{i=1}^p d_i \right) \quad (10.53)$$

$$\hat{\beta}^* = -\eta \sum_{i=1}^p d_i v_{0i}^{-1} \vec{v}_i^{(0)} \quad (10.54)$$

Como los regresores W están centrados, es claro que $\hat{\beta}_0 = \bar{y}$, y por tanto de (10.53) se deduce $\sum_{i=1}^p d_i = 1$. Haciendo uso de (10.52), (10.53), y (10.54) obtenemos la suma de cuadrados de los residuos:

$$\begin{aligned} (\vec{y} - \hat{y})'(\vec{y} - \hat{y}) &= \eta^2 (\vec{y}^* - \hat{y}^*)' (\vec{y}^* - \hat{y}^*) \\ &= \eta^2 \left(\vec{y}^* + W \sum_{i=1}^p d_i v_{0i}^{-1} \vec{v}_i^{(0)} \right)' \left(\vec{y}^* + W \sum_{i=1}^p d_i v_{0i}^{-1} \vec{v}_i^{(0)} \right) \\ &= \eta^2 \left[\sum_{i=1}^p \left(\frac{d_i}{v_{0i}} \right) (\vec{y}^* v_{0i} + W \vec{v}_i^{(0)}) \right]' \\ &\quad \times \left[\sum_{i=1}^p \left(\frac{d_i}{v_{0i}} \right) (\vec{y}^* v_{0i} + W \vec{v}_i^{(0)}) \right] \\ &= \eta^2 \left[\sum_{i=1}^p \left(\frac{d_i}{v_{0i}} \right) A \vec{v}_i \right]' \left[\sum_{i=1}^p \left(\frac{d_i}{v_{0i}} \right) A \vec{v}_i \right] \\ &= \eta^2 \sum_{i=1}^p \left(\frac{\lambda_i d_i^2}{v_{0i}^2} \right). \end{aligned} \quad (10.55)$$

Podemos ahora minimizar la expresión (10.55) sujeta a que $\sum_{i=1}^p d_i = 1$. El lagrangiano es:

$$\Phi(\vec{d}) = \eta^2 \sum_{i=1}^p \left(\frac{\lambda_i d_i^2}{v_{0i}^2} \right) - \mu \left(\sum_{i=1}^p d_i - 1 \right) \quad (10.56)$$

cuyas derivadas

$$\frac{\partial \Phi(\vec{d})}{\partial d_i} = 2\eta^2 \left(\frac{d_i \lambda_i}{v_{0i}^2} \right) - \mu = 0 \quad (i = 1, \dots, p) \quad (10.57)$$

permiten (multiplicando cada igualdad en (10.57) por $v_{0i}^2 \lambda_i^{-1}$ y sumando) obtener:

$$\mu = 2\eta^2 \left(\sum_{i=1}^p \frac{v_{0i}^2}{\lambda_i} \right)^{-1} \quad (10.58)$$

Llevando (10.58) a (10.57) obtenemos:

$$2\eta^2 d_i \frac{\lambda_i}{v_{0i}^2} = \mu = 2\eta^2 \left(\sum_{i=1}^p \frac{v_{0i}^2}{\lambda_i} \right)^{-1} \quad (10.59)$$

y por tanto:

$$d_i = \frac{v_{0i}^2}{\lambda_i} \left(\sum_{i=1}^p \frac{v_{0i}^2}{\lambda_i} \right)^{-1} \quad (10.60)$$

Los estimadores deseados se obtienen llevando (10.60) a (10.53)–(10.54):

$$\hat{\beta}_0 = \bar{y} \quad (10.61)$$

$$\hat{\beta}^* = -\eta \frac{\sum_{i=1}^p \left(\frac{v_{0i}}{\lambda_i} \right) \vec{v}_i^{(0)}}{\sum_{i=1}^p \frac{v_{0i}^2}{\lambda_i}} \quad (10.62)$$

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 λ_i . Es lo sensato: lo primero, prima las multicolinealidades predictivas sobre las que lo son menos; lo segundo, a las multicolinealidades más fuertes (en que la igualdad aproximada (10.49) es más ajustada). Pero podemos eliminar en (10.62) términos muy inestables, cuando v_{0i} y λ_i son ambos muy pequeños, para evitar que el sumando correspondiente en (10.62) reciba gran ponderación, si parece evidente que se trata de una multicolinealidad no predictiva. La relación (10.62) se transformará entonces en:

$$\hat{\beta}^* = -\eta \frac{\sum_{i \in P} \left(\frac{v_{0i}}{\lambda_i} \right) \vec{v}_i^{(0)}}{\sum_{i \in P} \left(\frac{v_{0i}^2}{\lambda_i} \right)} \quad (10.63)$$

siendo P un subconjunto de $(1, \dots, p)$.

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 $W'W$.

10.6. Lectura recomendada

Sobre regresión *ridge*, el trabajo original es Hoerl and Kennard (1970) (ver también Hoerl et al. (1975)). Hay una enorme literatura sobre los estimadores *ridge* y en componentes principales. Pueden verse por ejemplo Brown (1993), Cap. 4, Trocóniz (1987a) Cap. 10 ó Peña (2002) Sec. 8.3.4, que relaciona el estimador *ridge* con un estimador bayesiano.

Los métodos de regresión sesgada se contemplan a veces como alternativas a los métodos de selección de variables en situaciones de acusada multicolinealidad: véase por ejemplo Miller (2002), Cap. 3. De hecho, estudiaremos en el Capítulo 12 estimadores como el LASSO y garrote no negativo que pueden también verse como métodos de regresión sesgada.

El trabajo original regresión en raíces latentes puede verse en Webster et al. (1974). Hay también descripciones completas del método en manuales como Trocóniz (1987a) (pág. 247 y ss.) o Gunst and Mason (1980), Sec. 10.2.

COMPLEMENTOS Y EJERCICIOS

10.1 Al final de la Sección 10.3 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 .

10.2 Demuéstrese que si u_i es definida como en (10.22), se verifica que $\vec{1} \perp \vec{u}_i$.

10.3 Sea una muestra formada por n observaciones, X_1, \dots, X_n , generadas por una distribución con media. Demuéstrese que, para algún c , $c\bar{X}$ es mejor estimador (en terminos de error medio cuadrático, ECM) que \bar{X} . ¿Es esto un caso particular de alguno de los procedimientos de estimación examinados en este capítulo?

10.4 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:

$$\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} \quad (10.64)$$

$$= (X'X + kI)^{-1}(X'\vec{y} + \sqrt{k}I\vec{0}) \quad (10.65)$$

$$= (X'X + kI)^{-1}X'\vec{y} \quad (10.66)$$

$$= \hat{\beta}^{(k)} \quad (10.67)$$

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 pseudo-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) 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.

10.5 Supongamos una muestra formada por pares de valores (y_i, x_i) , $i = 1, \dots, N$. 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

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & \dots & x_1^{p-1} \\ 1 & x_2 & x_2^2 & x_2^3 & \dots & x_2^{p-1} \\ 1 & x_3 & x_3^2 & x_3^3 & \dots & x_3^{p-1} \\ \vdots & & & \vdots & & \vdots \\ 1 & x_N & x_N^2 & x_N^3 & \dots & x_N^{p-1} \end{pmatrix} \quad (10.68)$$

y contrastar hipótesis tales como $H_0 : \beta_2 = \beta_3 = \dots = \beta_{p-1} = 0$ (tendencia no más que lineal), $H_0 : \beta_3 = \dots = \beta_{p-1} = 0$ (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, \dots, x_N .

Podríamos ortogonalizar los vectores columna de la matriz de diseño (por ejemplo mediante el procedimiento 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.

10.6 (\uparrow 10.5) ¿Por qué, para la finalidad perseguida en el Ejercicio 10.5, no sería de utilidad hacer regresión en componentes principales?

Capítulo 11

Evaluación del ajuste. Diagnósticos.

Ya hemos visto en lo que precede estadísticos para evaluar la bondad de ajuste de un modelo, como R^2 ; pero se trata de estadísticos que dan una idea global del ajuste. Puede ocurrir que un 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.

11.1. Análisis de residuos.

En general, como se ha indicado ya en el Capítulo 12, 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

$$\hat{\epsilon} = \vec{y} - X\hat{\beta} = (I - X(X'X)^{-1}X')\vec{y} = (I - X(X'X)^{-1}X')\vec{\epsilon} \quad (11.1)$$

Podemos contemplar los $\hat{\epsilon}_i$ como “estimaciones” de las perturbaciones ϵ_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 11.1 *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.*

DEMOSTRACIÓN:

$$\Sigma_{\hat{\epsilon}} = E[(\hat{\epsilon} - E(\hat{\epsilon}))(\hat{\epsilon} - E(\hat{\epsilon}))'] \quad (11.2)$$

Como $E(\hat{\epsilon}) = \vec{0}$, (11.2) se reduce a:

$$E\hat{\epsilon}\hat{\epsilon}' = E[(I - X(X'X)^{-1}X')\vec{y}\vec{y}'(I - X(X'X)^{-1}X)'] \quad (11.3)$$

$$= (I - X(X'X)^{-1}X')\sigma^2I \quad (11.4)$$

$$= \sigma^2(I - P), \quad (11.5)$$

que en general no tiene elementos iguales a lo largo de la diagonal principal. El apartado 2) del enunciado es inmediato a partir de (11.5), dado que $(I - P)$ es una matriz no diagonal.

Sea,

$$p_{ij} = \vec{x}_i'(X'X)^{-1}\vec{x}_j \quad (11.6)$$

un elemento genérico de la matriz P (\vec{x}_i' denota la i -ésima fila de X). De la igualdad (11.1) se deduce:

$$\hat{\epsilon}_i = (1 - p_{ii})\epsilon_i - \sum_{i \neq j} p_{ij}\epsilon_j \quad (11.7)$$

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 ϵ_i .

Los valores p_{ij} dependen sólo de la matrix de diseño y son del mayor interés, como veremos más abajo.

Residuos internamente studentizados.

Los residuos MCO definidos en (11.1) 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 (11.5) se deduce que una estimación de la varianza de $\hat{\epsilon}_i$ viene dada por $\hat{\sigma}^2(1 - p_{ii})$. Por tanto,

$$r_i = \frac{\hat{\epsilon}_i}{+\sqrt{\hat{\sigma}^2(1 - p_{ii})}} \quad (11.8)$$

para $i = 1, \dots, N$ son residuos de varianza común. Se llama *studentización* a la eliminación del efecto de un parámetro de escala (aquí σ^2) mediante división por una estimación adecuada. Se denomina *internamente studentizados* a los residuos definidos en (11.8).

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*.

Residuos externamente studentizados.

Definidos por:

$$t_i = \frac{\hat{\epsilon}_i}{+\sqrt{\hat{\sigma}^2(i)(1 - p_{ii})}} \quad (11.9)$$

son formalmente idénticos a los r_i , con la única salvedad de haberse tomado en el denominador un estimador $\hat{\sigma}^2(i)$ de σ^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 ρ (véase Sección 8.3, pág. 112) para contrastar la presencia de *outliers*. Tomaremos,

$$\hat{\sigma}^2(i) = \frac{\hat{\epsilon}'\hat{\epsilon} - \hat{\epsilon}_i(1 - p_{ii})^{-1}\hat{\epsilon}_i}{(N - p - 1)} \quad (11.10)$$

lo que permite probar el siguiente,

Teorema 11.2 Con $\hat{\sigma}^2(i)$ definido como en (11.10), bajo los supuestos habituales más el de normalidad en las perturbaciones, los residuos t_i definidos en (11.9) (externamente studentizados) siguen una distribución t de Student con $(N - p - 1)$ grados de libertad.

DEMOSTRACIÓN:

Podemos escribir $\hat{\epsilon}_i = G'_i(I - P)\bar{\epsilon}$ 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:

$$\hat{\epsilon}_i = A\bar{\epsilon} \tag{11.11}$$

Por otra parte, de (11.10) deducimos:

$$\begin{aligned} (N - p - 1)\hat{\sigma}^2(i) &= \hat{\epsilon}'[I - G_i[G'_i(I - P)G_i]^{-1}G'_i]\hat{\epsilon} \\ &= \bar{\epsilon}' \underbrace{(I - P)[I - G_i[G'_i(I - P)G_i]^{-1}G'_i](I - P)}_B \bar{\epsilon} \\ &= \bar{\epsilon}'B\bar{\epsilon} \end{aligned} \tag{11.12}$$

Es fácil comprobar que $AB = 0$, luego $\hat{\epsilon}_i$ y $\hat{\sigma}^2(i)$ son independientes (Lema 6.3, pág. 67). Por otra parte, es también fácil comprobar que B es idempotente, con rango (= traza) $(N - p - 1)$. Por consiguiente,

$$\frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2(i)(1 - p_{ii})}} = \frac{\hat{\epsilon}_i/\sqrt{\sigma^2(1 - p_{ii})}}{\sqrt{\hat{\sigma}^2(i)/\sigma^2}} \tag{11.13}$$

$$= \frac{\hat{\epsilon}_i/\sqrt{\sigma^2(1 - p_{ii})}}{\sqrt{\bar{\epsilon}'B\bar{\epsilon}/(N - p - 1)\sigma^2}} \tag{11.14}$$

Pero en el numerador y denominador de (11.14) hay respectivamente una variable aleatoria $N(0, 1)$ y una χ^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 8.3, pág. 112). 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.

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 hacerlo¹, 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.

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 así²:

$$d_i = y_i - \vec{x}_i' \hat{\beta}_{(i)} \quad (11.15)$$

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

¹Véase Theil (1971), pág. 202 y ss.

²Una denominación alternativa frecuente en la literatura es la de residuos PRESS (predictive sum of squares residuals).

$N - 1$ observaciones. En efecto, de (11.15) se deduce que:

$$\begin{aligned} d_i &= y_i - \vec{x}_i' (X'_{(i)} X_{(i)})^{-1} X'_{(i)} \vec{Y}_{(i)} \\ &= y_i - \vec{x}_i' [(X'X) - \vec{x}_i \vec{x}_i']^{-1} X'_{(i)} \vec{Y}_{(i)} \end{aligned} \quad (11.16)$$

$$= y_i - \vec{x}_i' \left[(X'X)^{-1} + \frac{(X'X)^{-1} \vec{x}_i \vec{x}_i' (X'X)^{-1}}{1 - \vec{x}_i' (X'X)^{-1} \vec{x}_i} \right] X'_{(i)} \vec{Y}_{(i)} \quad (11.17)$$

$$= y_i - \vec{x}_i' \left[\frac{(1 - p_{ii})(X'X)^{-1} + (X'X)^{-1} \vec{x}_i \vec{x}_i' (X'X)^{-1}}{1 - p_{ii}} \right] X'_{(i)} \vec{Y}_{(i)}$$

$$= y_i - \left[\frac{(1 - p_{ii}) \vec{x}_i' (X'X)^{-1} + p_{ii} \vec{x}_i' (X'X)^{-1}}{1 - p_{ii}} \right] X'_{(i)} \vec{Y}_{(i)}$$

$$\begin{aligned} &= y_i - \frac{\vec{x}_i' (X'X)^{-1} X'_{(i)} \vec{Y}_{(i)}}{1 - p_{ii}} \\ &= \frac{(1 - p_{ii}) y_i - \vec{x}_i' (X'X)^{-1} (X' \vec{Y} - \vec{x}_i y_i)}{1 - p_{ii}} \end{aligned} \quad (11.18)$$

$$\begin{aligned} &= \frac{y_i - \vec{x}_i' (X'X)^{-1} X' \vec{Y}}{1 - p_{ii}} \\ &= \frac{\hat{\epsilon}_i}{1 - p_{ii}} \end{aligned} \quad (11.19)$$

en que el paso de (11.16) a (11.17) hace uso del Teorema A.2, pág. 221. 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.

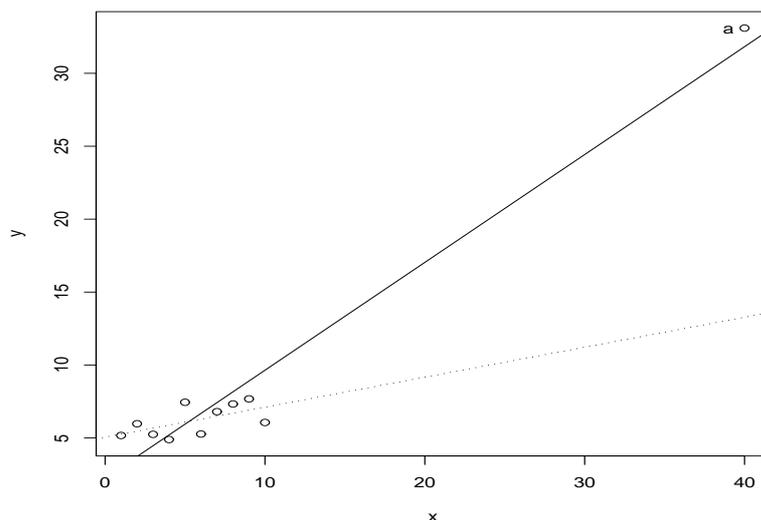
11.2. Análisis de influencia.

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 qué punto observaciones aisladas influyen 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. 11.1 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

Figura 11.1: Una observación como a tiene residuo borrado muy grande, y gran influencia en la pendiente de la recta de regresión.



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.

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:

$$\text{SIC}_i = (N - 1)(\hat{\beta} - \hat{\beta}_{(i)}). \quad (11.20)$$

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 (11.20) 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 (11.20) con el residuo borrado i -ésimo haciendo uso del siguiente lema.

Lema 11.1 *Se verifica que*

$$(\hat{\beta} - \hat{\beta}_{(i)}) = \frac{(X'X)^{-1}\vec{x}_i\hat{\epsilon}_i}{(1 - p_{ii})} = (X'X)^{-1}\vec{x}_id_i. \quad (11.21)$$

DEMOSTRACIÓN:

$$\begin{aligned} (\hat{\beta} - \hat{\beta}_{(i)}) &= (X'X)^{-1}X'\vec{Y} - ((X'X) - \vec{x}_i\vec{x}_i')^{-1}(X'\vec{Y} - \vec{x}_iy_i) \\ &= (X'X)^{-1}X'\vec{Y} \\ &\quad - \left[(X'X)^{-1} + \frac{(X'X)^{-1}\vec{x}_i\vec{x}_i'(X'X)^{-1}}{1 - \vec{x}_i'(X'X)^{-1}\vec{x}_i} \right] (X'\vec{Y} - \vec{x}_iy_i) \\ &= (X'X)^{-1}\vec{x}_iy_i - \frac{(X'X)^{-1}\vec{x}_i\vec{x}_i'(X'X)^{-1}X'\vec{Y}}{1 - p_{ii}} \\ &\quad + \frac{(X'X)^{-1}\vec{x}_i\vec{x}_i'(X'X)^{-1}\vec{x}_iy_i}{1 - p_{ii}} \\ &= \frac{(X'X)^{-1}\vec{x}_i}{1 - p_{ii}} \left[(1 - p_{ii})y_i - \vec{x}_i'\hat{\beta} + p_{ii}y_i \right] \\ &= (X'X)^{-1}\vec{x}_i \frac{\hat{\epsilon}_i}{1 - p_{ii}} \end{aligned}$$

En consecuencia,

$$\text{SIC}_i = (N - 1)(\hat{\beta} - \hat{\beta}_{(i)}) = (N - 1)(X'X)^{-1}\vec{x}_i \frac{\hat{\epsilon}_i}{1 - p_{ii}}$$

y el cálculo de la curva de influencia muestral 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 (11.22).

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 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:

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})' S (\hat{\beta} - \hat{\beta}_{(i)})}{c} \quad (11.22)$$

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” (11.22) es: $S = (X'X)$ y $c = p\hat{\sigma}^2$. Con esta elección, la expresión (11.22) 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 matiz³.

Haciendo uso del Lema 11.1 tenemos que la distancia de Cook puede escribirse así:

$$D_i = \frac{\hat{\epsilon}_i \vec{x}_i' (X'X)^{-1} (X'X) (X'X)^{-1} \vec{x}_i \hat{\epsilon}_i}{p\hat{\sigma}^2(1 - p_{ii})^2} \quad (11.23)$$

$$= \frac{1}{p} r_i^2 \frac{p_{ii}}{1 - p_{ii}} \quad (11.24)$$

siendo r_i el i -ésimo residuo internamente studentizado.

DFFITs.

Se definen así:

$$DFFIT_i = t_i \sqrt{\frac{p_{ii}}{1 - p_{ii}}} \quad (11.25)$$

³Una relación de las mismas puede verse en Cook and Weisberg (1982), p. 124.

Se suele considerar observaciones inusuales a aquéllas con

$$| \text{DFFIT}_i | > 2\sqrt{\frac{p}{N}} \quad (11.26)$$

DFBETAS.

Se definen por:

$$\text{DFBETA}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j,(i)}}{\hat{\sigma} \sqrt{(X'X)^{-1}_{jj}}}; \quad (11.27)$$

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 (11.27) es clara: la diferencia entre la estimación de β_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 $|\text{DFBETA}_{ij}|$ con $2/\sqrt{N}$. Más detalles en Belsley et al. (1980).

11.3. 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 12) y transformaciones de las variables (capítulo 13). Referencias útiles para ampliar lo que se expone a continuación incluyen Trocóniz (1987a), Myers (1990), Ryan (1997) o Atkinson (1985).

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 después 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.

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.

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.

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.

Gráficos de normalidad de residuos

Aunque, como se ha visto (Sección 11.1 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 estadística 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 $(\Phi^{-1}(F_*(x_i)), x_i)$, en que $F_*(x_i)$ es la función de distribución empírica de la muestra, deben estar aproximadamente 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 $\Phi^{-1}(\cdot)$; o puede hacerse mediante ordenador en cuyo caso basta facilitar los datos y verificar la linealidad del gráfico resultante.

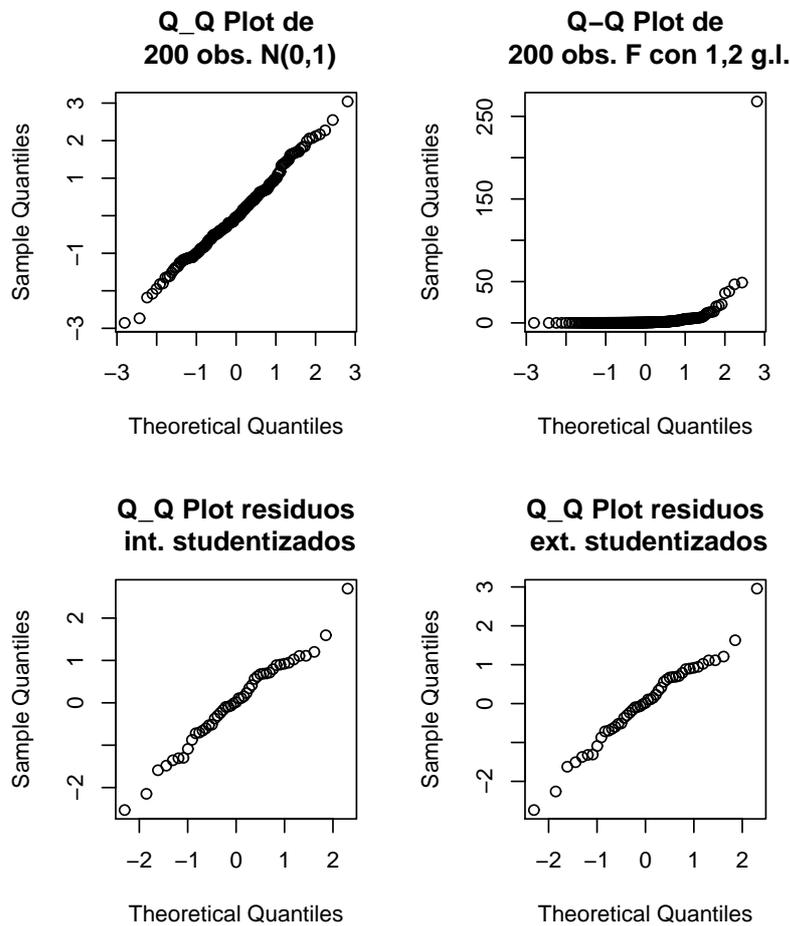
En cualquiera de los casos se cuenta con un instrumento que permite no sólo apreciar si hay desviaciones respecto de la normalidad, sino también de qué naturaleza son y a qué puntos afectan.

R: Ejemplo 11.1 (*gráficos para contraste de normalidad de residuos*)

La Figura 11.2 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.

```
> 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)
> library(MASS)
> data(UScrime)
> modelo <- lm(y ~ M + Ed +
+           Pol + M.F + U1 + U2 +
+           Prob + Ineq, data = UScrime)
```

Figura 11.2: Gráficos para contraste de normalidad



```
> qqnorm(stdres(modelo), main = "Q_Q Plot residuos\n int. studentizado")
> qqnorm(studres(modelo), main = "Q_Q Plot residuos\n ext. studentizado")
```

```
X11cairo
2
```

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.

FIN DEL EJEMPLO ■

Gráficos de residuos ordinarios frente a residuos borrados $(d_i, \hat{\epsilon}_i)$

Un residuo borrado grande 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.

COMPLEMENTOS Y EJERCICIOS

11.1 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 12

Selección de modelos.

12.1. 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 toda¹ variable que incluyamos en la regresión, tenga mucha o poca relación con la variable respuesta, reducirá SSE . Tenemos, pues, que buscar criterios más elaborados.

Maximización de \bar{R}_p^2 .

Se define el *coeficiente de determinación corregido* así:

$$\bar{R}_p^2 = 1 - [1 - R_p^2] \times \frac{N - 1}{N - p} \quad (12.1)$$

¹Las únicas excepciones son aquellas variables correspondientes a columnas de la matriz de diseño X ortogonales a \bar{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 (12.1) en la forma:

$$1 - \overline{R}_p^2 = [1 - R_p^2] \times \frac{N - 1}{N - p} \quad (12.2)$$

$$= \frac{SSE_p}{SST} \times \frac{N - 1}{N - p} \quad (12.3)$$

vemos que mientras que el primer término de la derecha de (12.3) es monótono no creciente con p , el segundo es monótono creciente. Por consiguiente, el producto de ambos² 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 12.1 *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 $Q_h > 1$.*

DEMOSTRACIÓN:³

Para contrastar la significación del $(p + 1)$ -ésimo parámetro, empleamos (Sección 6.2, pág. 72):

$$Q_h = \frac{SSE_p - SSE_{p+1}}{SSE_{p+1}} \times \frac{N - p - 1}{1} \quad (12.4)$$

$$= \frac{(R_{p+1}^2 - R_p^2)}{1 - R_{p+1}^2} \times \frac{N - p - 1}{1} \quad (12.5)$$

de donde:

$$(1 - R_{p+1}^2)Q_h = (R_{p+1}^2 - R_p^2)(N - p - 1) \quad (12.6)$$

$$Q_h - Q_h R_{p+1}^2 = (N - p - 1)R_{p+1}^2 - (N - p - 1)R_p^2 \quad (12.7)$$

$$Q_h + (N - p - 1)R_p^2 = R_{p+1}^2 [(N - p - 1) + Q_h] \quad (12.8)$$

²Expresiones 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.

³Sigue a Haitovsky (1969).

Despejando R_{p+1}^2 tenemos:

$$R_{p+1}^2 = \frac{Q_h + (N - p - 1)R_p^2}{(N - p - 1) + Q_h} \quad (12.9)$$

$$= \frac{\frac{1}{N-p-1}Q_h + R_p^2}{1 + \frac{1}{N-p-1}Q_h} \quad (12.10)$$

De (12.10) y de la definición de \bar{R}_{p+1}^2 se deduce que:

$$\bar{R}_{p+1}^2 = 1 - [1 - R_{p+1}^2] \times \frac{N - 1}{(N - p - 1)} \quad (12.11)$$

Sustituyendo en esta expresión (12.10) llegamos a:

$$\bar{R}_{p+1}^2 = 1 - \frac{[1 - R_p^2]}{\frac{N-p-1+Q_h}{N-p-1}} \times \frac{N - 1}{N - p - 1} \quad (12.12)$$

$$= 1 - [1 - R_p^2] \frac{N - 1}{N - p - 1 + Q_h} \quad (12.13)$$

$$= 1 - \underbrace{[1 - R_p^2] \frac{N - 1}{N - p}}_{\bar{R}_p^2} \underbrace{\frac{N - p}{N - p - 1 + Q_h}}_t \quad (12.14)$$

Es evidente de (12.14) que $\bar{R}_{p+1}^2 \geq \bar{R}_p^2$ si $Q_h > 1$, y viceversa⁴. Maximizar \bar{R}_p^2 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 $h: \beta_i = 0$ es cierta. Consecuentemente, el emplear este criterio en exclusiva conduciría con gran probabilidad al ajuste de modelos sobreparametrizados.

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

$$ECM = E(\hat{Y}^{(p)} - X\vec{\beta})'(\hat{Y}^{(p)} - X\vec{\beta}) \quad (12.15)$$

⁴Obsérvese que si el término t en (12.14) fuera la unidad —lo que acontece cuando $Q_h = 1$ —, el lado derecho sería precisamente \bar{R}_p^2 . Si $Q_h > 1$, t es menor que 1 y, como sólo multiplica al sustraendo en (12.14), el resultado es *mayor* que \bar{R}_p^2 .

que sumando y restando $E(\hat{Y}^{(p)})$ dentro de cada paréntesis podemos descomponer así:

$$ECM = E \left[(\hat{Y}^{(p)} - E(\hat{Y}^{(p)}))' (\hat{Y}^{(p)} - E(\hat{Y}^{(p)})) \right] + E \left[(E(\hat{Y}^{(p)}) - X\vec{\beta})' (E(\hat{Y}^{(p)}) - X\vec{\beta}) \right] \quad (12.16)$$

$$= \text{Var}(\hat{Y}^{(p)}) + (\text{Sesgo})^2. \quad (12.17)$$

El primer término no ofrece dificultad. Como

$$\hat{Y}^{(p)} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\vec{Y} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'(X\vec{\beta} + \vec{\epsilon}), \quad (12.18)$$

tenemos que

$$E[\hat{Y}^{(p)}] = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'X\vec{\beta}$$

y

$$\begin{aligned} ((\hat{Y}^{(p)} - E(\hat{Y}^{(p)}))' ((\hat{Y}^{(p)} - E(\hat{Y}^{(p)}))) &= \vec{\epsilon}\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\vec{\epsilon} \\ &= \vec{\epsilon}\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\vec{\epsilon} \\ &\sim \sigma^2\chi_p^2. \end{aligned} \quad (12.19)$$

Falta el término de sesgo. Observemos que

$$E \left[\underbrace{(\vec{Y} - \hat{Y}^{(p)})' (\vec{Y} - \hat{Y}^{(p)})}_{SSE} \right] = E \left[\underbrace{(X\vec{\beta} - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'X\vec{\beta})' (X\vec{\beta} - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'X\vec{\beta})}_{(\text{Sesgo})^2} \right] + E \left[\vec{\epsilon}'(I - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}')\vec{\epsilon} \right].$$

Por consiguiente,

$$(\text{Sesgo})^2 = E[SSE] - E[\sigma^2\chi_{N-p}^2]. \quad (12.20)$$

Sustituyendo en (12.17) tenemos entonces que

$$ECM = E \left[SSE - \sigma^2\chi_{N-p}^2 \right] + E \left[\sigma^2\chi_p^2 \right] \quad (12.21)$$

$$= E[SSE] - \sigma^2(N-p) + \sigma^2p, \quad (12.22)$$

y por consiguiente:

$$\frac{ECM}{\sigma^2} = E \left[\frac{SSE}{\sigma^2} \right] - N + 2p. \quad (12.23)$$

Minimizar esta última expresión es lo mismo que minimizar

$$E \left[\frac{SSE}{\sigma^2} \right] + 2p, \quad (12.24)$$

ya que N es constante. Como quiera que el valor medio en la expresión anterior no puede ser calculado y σ es desconocida, todo lo que podemos hacer es reemplazar (12.24) por la expresión análoga,

$$C_p = \frac{SSE}{\hat{\sigma}^2} + 2p. \quad (12.25)$$

A esta última expresión se la conoce como C_p de Mallows.

Para que se verifique la aproximación en (12.25) 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 incluidos 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 σ^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 (12.25) 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 (12.25) en 2. El efecto neto indica si el nuevo regresor es o no deseable.

Observación 12.1 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 \bar{R}_p^2 , 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 restrictivo⁵.

Observación 12.2 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

⁵La 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 σ^2 , aunque seguramente nos hace despilfarrar algunos grados de libertad). Por el contrario, al utilizar el criterio basado en \bar{R}_p^2 introducimos el nuevo regresor si $Q_h > 1$ en (12.4), 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.

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 12.1. En la Sección 12.1 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.

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 (Akaike's Information Criterion, o An Information Criterion). Consiste en seleccionar el modelo minimizando

$$AIC(p) = -2 \log_e \left[\underset{\vec{\theta}}{\text{máx}} \text{verosimilitud}(\vec{x}, \vec{\theta}) \right] + 2p$$

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 ??.

Cuando consideremos modelos de regresión lineal con normalidad, el uso de los criterios AIC y C_p daría resultados exactamente equivalentes si conociéramos σ^2 (ambos criterios difieren en tal caso en una constante; ver Venables and Ripley (1999a), pág. 185). Cuando σ^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.

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 reestimarlos.

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 12.1 Consideremos una muestra de tamaño $N = 100$. Tenemos una colección de K modelos \mathcal{M}_i , $i = 1, \dots, K$, 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 $SSE_i^{(\ell)}$, $\ell = 1, \dots, s$, (suma de cuadrados de los residuos al predecir en la muestra ℓ 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 estimación 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 11.1). Por tanto, utilizando la notación de dicha Sección,

$$\begin{aligned} SSE_i^\ell &= d_i^2 \quad (\ell = 1, \dots, N) \\ SSE_i &= N^{-1} \sum_{\ell=1}^N SSE_i^\ell. \end{aligned}$$

El modelo seleccionado es aquél al que corresponde un SSE_i más pequeño⁶.

FIN DEL EJEMPLO ■

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 dará un ajuste susceptible de mejora. Se trata de alcanzar un equilibrio entre los dos objetivos en

⁶Nótese 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 11.1.

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 probabilístico podemos describir o codificar unos datos de modo compacto asignando a los más “raros” (menos probables) los códigos más largos.

Observación 12.3 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 probabilístico. 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

$$MDL(\mathcal{M}_k; \vec{y}) = (\text{Código necesario para } \vec{y}) \quad (12.26)$$

$$+ (\text{Código necesario para } \vec{\theta}_k). \quad (12.27)$$

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 (12.27). 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.

12.2. Selección de variables.

Una aproximación ingenua al problema consistiría en estudiar la reducción en un cierto criterio (SSE , \overline{R}_p^2 , C_p , ...) originada por la introducción de cada variable, y 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.

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 incluida 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 cálculo⁷, pero aún así éste puede resultar excesivo.

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.

⁷Véase Seber (1977), pag. 349 y ss.

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 ($h: \beta_i = 0$).
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.

El procedimiento de regresión “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 mixto, 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 extremos⁸.

⁸Podría 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 distribu-

R: Ejemplo 12.1 (*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 auxiliares.

Primero generamos datos sintéticos del modo habitual. Como puede verse, hay muchos betas no significativos.

```
> set.seed(123457)
> 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)
```

Como puede verse, hay muchos betas no significativos:

```
> summary(completo)

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

Residuals:
    Min       1Q   Median       3Q      Max
-1.916 -0.550 -0.106  0.829  2.204

Coefficients:
              Estimate Std. Error
(Intercept)  -0.0706     0.2227
V1             0.0408     0.2422
V2             0.1720     0.2603
V3             1.1884     0.2397
V4            -0.0238     0.2067
```

ción diferente (véase Capítulo 8). 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.

V5	2.0035	0.2022
V6	0.2633	0.2217
V7	2.9970	0.1875
V8	-0.1074	0.2804
V9	0.0514	0.2105
V10	-0.2367	0.2148
V11	-0.2053	0.2042
V12	4.0374	0.2212
V13	0.1137	0.2161
V14	-0.2115	0.2163
V15	0.0191	0.3076
V16	0.1206	0.2328
V17	0.0318	0.1972
V18	-0.0786	0.2108
V19	0.0879	0.2569
V20	0.0162	0.1949

	t value	Pr(> t)	
(Intercept)	-0.32	0.75	
V1	0.17	0.87	
V2	0.66	0.51	
V3	4.96	2.9e-05	***
V4	-0.11	0.91	
V5	9.91	8.1e-11	***
V6	1.19	0.24	
V7	15.98	6.5e-16	***
V8	-0.38	0.70	
V9	0.24	0.81	
V10	-1.10	0.28	
V11	-1.01	0.32	
V12	18.25	< 2e-16	***
V13	0.53	0.60	
V14	-0.98	0.34	
V15	0.06	0.95	
V16	0.52	0.61	
V17	0.16	0.87	
V18	-0.37	0.71	
V19	0.34	0.73	
V20	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
```

Utilizamos ahora la función `leaps` para hacer regresión sobre todos los subconjuntos. Con 15 regresores, es un problema de talla modesta.

```
> library(leaps)
> mods <- leaps(x = X, y = y,
+             method = "Cp")
```

El objeto `mods` contiene información sobre todos los modelos estimados. Podemos ver como varía C_p y \bar{R}^2 con el número de regresores:

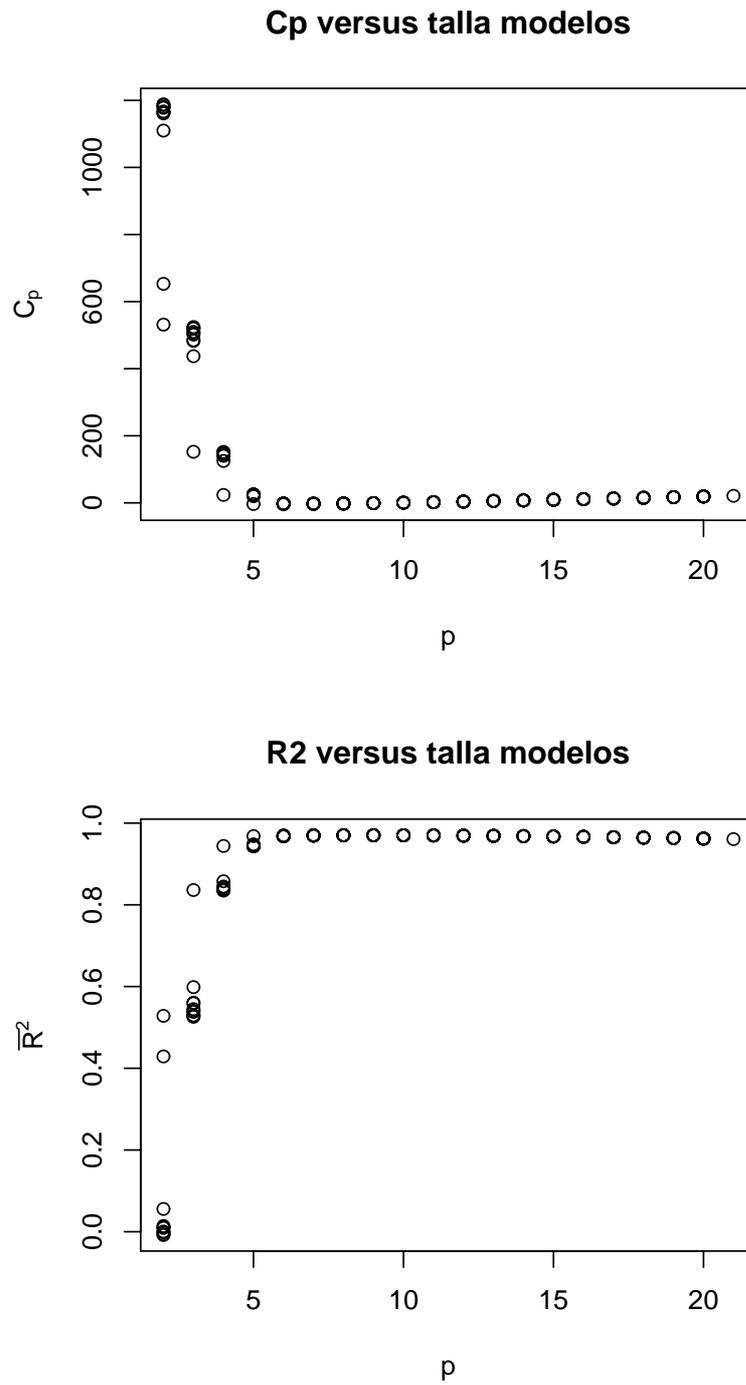
```
> postscript(file = "demo10.eps",
+           horizontal = FALSE, width = 5,
+           height = 9)
> opar <- par()
> par(mfrow = c(2, 1))
> plot(mods$size, mods$Cp,
+      main = "Cp versus talla modelos",
+      xlab = expression(p),
+      ylab = expression(C[p]))
> mods.r <- leaps(x = X, y = y,
+               method = "adjr2")
> plot(mods.r$size, mods.r$adjr2,
+      main = "R2 versus talla modelos",
+      xlab = expression(p),
+      ylab = expression(bar(R)^2))
> par(opar)
> dev.off()
```

```
X11cairo
2
```

La Figura 12.1 muestra el comportamiento típico de los criterios C_p y \bar{R}^2 . Se aprecia que, aunque de forma no muy notoria en este caso, el criterio \bar{R}^2 tiende a seleccionar modelos más parametrizados.

```
> mejores <- order(mods$Cp)[1:15]
> regres <- mods$which[mejores,
+                   ]
> dimnames(regres)[[2]] <- dimnames(datos)[[2]][1:20]
```

Figura 12.1: Valores de C_p y \bar{R}^2 para 141 modelos ajustados a los datos UScrime



```

> Cp <- mods$Cp[mejores]
> cbind(regres, Cp)

  V1 V2 V3 V4 V5 V6 V7 V8 V9 V10
5  0  0  1  0  1  1  1  0  0  0
6  0  0  1  0  1  1  1  0  0  0
6  0  0  1  0  1  1  1  0  0  1
4  0  0  1  0  1  0  1  0  0  0
6  0  0  1  0  1  1  1  0  0  0
5  0  0  1  0  1  0  1  0  0  1
6  0  0  1  0  1  1  1  0  0  0
5  0  0  1  0  1  0  1  0  0  0
7  0  0  1  0  1  1  1  0  0  1
6  0  0  1  0  1  1  1  0  0  0
6  1  0  1  0  1  1  1  0  0  0
5  1  0  1  0  1  0  1  0  0  0
6  0  0  1  0  1  1  1  0  0  0
7  0  0  1  0  1  1  1  0  0  0
6  0  0  1  0  1  1  1  0  0  0
  V11 V12 V13 V14 V15 V16 V17
5  0  1  0  0  0  0  0
6  0  1  0  1  0  0  0
6  0  1  0  0  0  0  0
4  0  1  0  0  0  0  0
6  1  1  0  0  0  0  0
5  0  1  0  0  0  0  0
6  0  1  0  0  0  0  0
5  1  1  0  0  0  0  0
7  0  1  0  1  0  0  0
6  0  1  0  0  1  0  0
6  0  1  0  0  0  0  0
5  0  1  0  0  0  0  0
6  0  1  0  0  0  0  1
7  1  1  0  1  0  0  0
6  0  1  1  0  0  0  0
  V18 V19 V20      Cp
5  0  0  0 -4.225
6  0  0  0 -3.491
6  0  0  0 -3.455
4  0  0  0 -3.453
6  0  0  0 -3.213
5  0  0  0 -3.150
6  0  1  0 -2.654
5  0  0  0 -2.550

```

```

7  0  0  0 -2.548
6  0  0  0 -2.518
6  0  0  0 -2.476
5  0  0  0 -2.405
6  0  0  0 -2.368
7  0  0  0 -2.365
6  0  0  0 -2.335

```

```

> mod1 <- lm(y ~ V3 + V4 +
+           V5 + V7 + V10 + V12 +
+           V16 + V17, data = datos)
> mod2 <- update(mod1, . ~
+               . + V1 + V2)
> 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
(Intercept) -0.03573    0.18316
V3           1.08674    0.19721
V4          -0.00741    0.16766
V5           2.03931    0.16976
V7           3.05622    0.14772
V10         -0.27977    0.19088
V12          4.10685    0.18483
V16          0.08436    0.15101
V17          0.05185    0.14567
V1           0.16370    0.18257
V2          -0.00659    0.20666

```

	t value	Pr(> t)
(Intercept)	-0.20	0.85
V3	5.51	2.5e-06 ***
V4	-0.04	0.96
V5	12.01	1.1e-14 ***
V7	20.69	< 2e-16 ***

```

V10          -1.47      0.15
V12          22.22 < 2e-16 ***
V16          0.56      0.58
V17          0.36      0.72
V1           0.90      0.38
V2          -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)
> 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
(Intercept)  0.0738     0.1596
V3           1.0693     0.1819
V4          -0.0410     0.1567
V5           1.9898     0.1603
V7           3.0484     0.1400
V12          4.1357     0.1642
              t value Pr(>|t|)
(Intercept)   0.46    0.65
V3            5.88 5.1e-07 ***
V4           -0.26    0.79
V5           12.41 5.7e-16 ***
V7           21.77 < 2e-16 ***
V12          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,
+   method = "forward")
> 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	
1	(1)	"	"	"	"	"	"	"
2	(1)	"	"	"	"	"	"	"
3	(1)	"	"	"	"	"	"	"
4	(1)	"	"	"	"	"	"	"
5	(1)	"	"	"	"	"	"	"
6	(1)	"	"	"	"	"	"	"
7	(1)	"	"	"	"	"	"	"
8	(1)	"	"	"	"	"	"	"
		V7	V8	V9	V10	V11	V12	

```

1 ( 1 ) " " " " " " " " " " "*"
2 ( 1 ) "*" " " " " " " " " " "*"
3 ( 1 ) "*" " " " " " " " " " "*"
4 ( 1 ) "*" " " " " " " " " " "*"
5 ( 1 ) "*" " " " " " " " " " "*"
6 ( 1 ) "*" " " " " " " " " " "*"
7 ( 1 ) "*" " " " " " "*" " " " "*"
8 ( 1 ) "*" " " " " " "*" " " " "*"
      V13 V14 V15 V16 V17 V18
1 ( 1 ) " " " " " " " " " " " "
2 ( 1 ) " " " " " " " " " " " "
3 ( 1 ) " " " " " " " " " " " "
4 ( 1 ) " " " " " " " " " " " "
5 ( 1 ) " " " " " " " " " " " "
6 ( 1 ) " " "*" " " " " " " " "
7 ( 1 ) " " "*" " " " " " " " "
8 ( 1 ) " " "*" " " " " " " " "
      V19 V20
1 ( 1 ) " " " "
2 ( 1 ) " " " "
3 ( 1 ) " " " "
4 ( 1 ) " " " "
5 ( 1 ) " " " "
6 ( 1 ) " " " "
7 ( 1 ) " " " "
8 ( 1 ) "*" " "

> library(MASS)
> 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

```

```

(Intercept)  0.0514    0.1518
V3           1.0256    0.1761
V5           2.0499    0.1557
V6           0.3046    0.1603
V7           3.0499    0.1346
V12          4.1077    0.1585
              t value Pr(>|t|)
(Intercept)  0.34    0.736
V3           5.82  6.1e-07 ***
V5          13.17 < 2e-16 ***
V6           1.90    0.064 .
V7          22.65 < 2e-16 ***
V12         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

```

FIN DEL EJEMPLO ■

12.3. EL LASSO

Tibshirani (1996) introdujo el método conocido como LASSO (=“least absolute shrinkage and selection operator”). Puede verse como un procedimiento a medio camino de la selección de variables y regresión ridge.

Los métodos que se han examinado en las secciones precedentes producen decisiones “todo o nada”: un regresor permanece o es excluido de la regresión, sin alternativas intermedias. En regresión ridge (cf. Sección 10.3, p. 139 y ss.), todos los regresores permanecen en el modelo, pero sus coeficientes estimados se “encogen” hacia cero; este “encogimiento”, que puede verse alternativamente como una restricción estocástica, o una distribución *a priori* sobre los parámetros, introduce un sesgo pero ayuda a reducir drásticamente la varianza.

El método LASSO participa de ambas características; aproxima los estimadores de los parámetros a cero, en ocasiones haciéndolos exactamente igual a cero (cosa que no ocurre en regresión ridge), lo que es equivalente a excluir el regresor correspondiente del modelo.

El método se describe fácilmente. Sea $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$ un modelo de regresión lineal, con $\hat{\beta} = (\beta_0, \dots, \beta_{p-1})$. El estimador LASSO se define así:

$$\hat{\beta} = \arg \min_{\hat{\beta}} (\vec{y} - X\hat{\beta})^2 \quad \text{sujeto a} \quad \sum_{i=1}^{p-1} |\beta_i| \leq t \quad (12.28)$$

en que t es un parámetro de calibrado, similar a λ en regresión ridge. Obsérvese que —al igual que en regresión ridge—, $\hat{\beta}_0$, el estimador de la ordenada en el origen, no se encoge. Obsérvese también que algunos betas pueden perfectamente ser cero.

El problema formulado en (12.28) es uno de optimización cuadrática sujeta a restricciones lineales, y es por tanto computacionalmente más complejo que MCO o regresión ridge; no obstante, existen buenos algoritmos para resolverlo.

En R, la función⁹ `lars` implementa el estimador LASSO (y otros relacionados también). La selección de t se puede hacer por validación cruzada.

12.4. 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:

$$y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon. \quad (12.29)$$

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 2.10, pág. 39); y sería inconveniente que el contraste de h dependiera del origen y de la escala empleadas. Lo menos que debemos esperar de nuestra inferencia es que sea invariante frente a cambios en las unidades de medida.

⁹En el paquete `lars`.

Si en (12.29) reemplazamos X por $Z = aX + b$, obtenemos

$$\begin{aligned} y &= \beta_0 + \beta_1(aX + b) + \beta_2(aX + b)^2 + \epsilon \\ &= (\beta_0 + \beta_1b + \beta_2b^2) + (\beta_1a + 2ab\beta_2)X + a^2\beta_2X^2 + \epsilon \\ &= \beta_0^* + \beta_1^*X + \beta_2^*X^2 + \epsilon. \end{aligned} \tag{12.30}$$

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 (12.29) o $h : \beta_2^* = 0$ en (12.30); 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 β_1 y β_2 , habrá coeficientes a, b anulando $\beta_1^* = (\beta_1a + 2ab\beta_2)$ en (12.30). 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_iX_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_iX_j .

Si regresamos al Ejercicio 2.10 en que se argüía la necesidad de utilizar un término β_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 β_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.

COMPLEMENTOS Y EJERCICIOS

12.1 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 \bar{R}_p^2 ?

12.2 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.

12.3 En la Observación 12.1 se comparan los criterios de selección de modelos consistentes en maximizar \bar{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:

$$Q_h = \frac{(SSE_A - SSE_B)}{q\hat{\sigma}^2} > \mathcal{F}_{q, N-(p+q)}^\alpha$$

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 13

Transformaciones

13.1. 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.

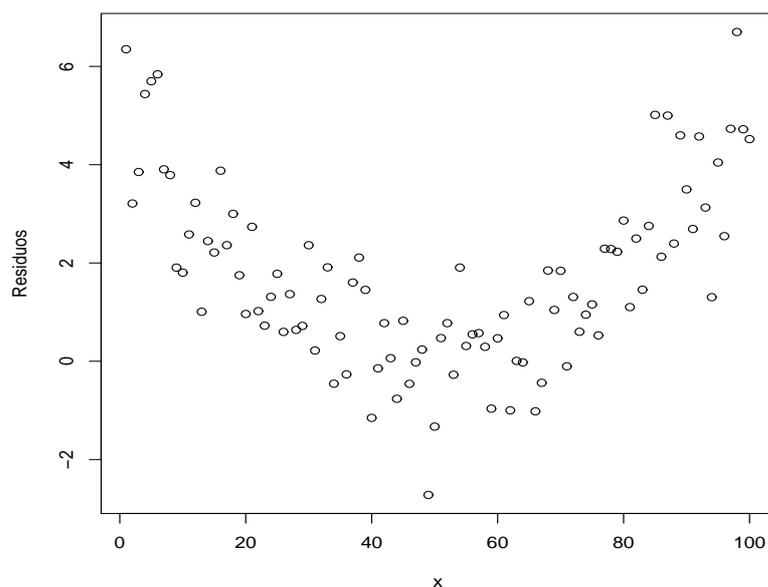
13.2. Transformaciones de los regresores

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.

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 13.1 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 13.1: Disposición de residuos sugiriendo una transformación cuadrática del regresor X_i



Transformaciones de Box-Tidwell

Consideremos los regresores X_1, \dots, X_p y transformaciones de los mismos definidas del siguiente modo:

$$W_j = \begin{cases} X_j^{\alpha_j} & \text{si } \alpha_j \neq 0, \\ \ln(X_j) & \text{si } \alpha_j = 0. \end{cases} \quad (13.1)$$

Para diferentes valores de α_j , la transformación (13.1) 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 α_j adecuado.

El modo de hacerlo propuesto por Box and Tidwell (1962) es el siguiente. Consideremos el modelo,

$$Y = \beta_0 + \beta_1 X_1^{\alpha_1} + \dots + \beta_p X_p^{\alpha_p} + \epsilon \quad (13.2)$$

$$= \beta_0 + \beta_1 W_1 + \dots + \beta_p W_p + \epsilon. \quad (13.3)$$

Si realizamos una linealización aproximada mediante un desarrollo en serie de Taylor en torno al punto $(\alpha_1, \dots, \alpha_k)' = (1, 1, \dots, 1)'$, obtenemos:

$$Y \approx \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \gamma_1 Z_1 + \dots + \gamma_p Z_p + \epsilon, \quad (13.4)$$

en donde

$$\gamma_j = \beta_j(\alpha_j - 1) \quad (13.5)$$

$$Z_j = X_j \ln(X_j). \quad (13.6)$$

Tenemos pues un modelo en el que podemos estimar los parámetros, $(\beta_0, \dots, \beta_p, \gamma_1, \dots, \gamma_p)$. De ellos podemos recuperar valores estimados de $(\alpha_1, \dots, \alpha_p)$ así:

$$\hat{\alpha}_j = \frac{\hat{\gamma}_j}{\hat{\beta}_j} + 1. \quad (13.7)$$

Podemos detenernos aquí, pero cabe pensar en un proceso iterativo de refinado de la solución obtenida. Llamemos $\hat{\alpha}_k^{(1)}$, $k = 1, \dots, p$, a los estimadores de los parámetros de transformación α_k obtenidos como primera aproximación al estimar (13.4). Podríamos ahora definir

$$W_j^{(1)} = X_j^{\hat{\alpha}_j^{(1)}} \quad (13.8)$$

$$Z_j^{(1)} = W_j^{(1)} \ln(W_j^{(1)}) \quad (13.9)$$

y estimar

$$Y = \beta_0 + \beta_1 W_1^{(1)} + \dots + \beta_p W_p^{(1)} + \gamma_1 Z_1^{(1)} + \dots + \gamma_p Z_p^{(1)} + \epsilon \quad (13.10)$$

Obtendríamos así estimaciones de $W_1^{(2)}, \dots, W_p^{(2)}$, y podríamos proseguir de modo análogo hasta convergencia, si se produce.

13.3. Transformaciones de la variable respuesta

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 13.1 Supongamos que nos planteamos escoger entre los dos modelos alternativos,

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \quad (13.11)$$

$$\log(Y) = \gamma_0 + \gamma_1 X_1 + \nu. \quad (13.12)$$

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

$$\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1})^2 \quad (13.13)$$

$$\sum (\log(Y_i) - \hat{\gamma}_0 - \hat{\gamma}_1 X_{i1})^2, \quad (13.14)$$

pero sí lo serían

$$\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1})^2 \quad (13.15)$$

$$\sum (Y_i - \exp\{\hat{\gamma}_0 + \hat{\gamma}_1 X_{i1}\})^2; \quad (13.16)$$

no obstante, véase la discusión en la Observación 13.1 que sigue.

FIN DEL EJEMPLO ■

Observación 13.1 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 12.1). 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 12.1).

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:

$$g(Y_i) = \vec{x}_i' \vec{\beta} + \epsilon_i \quad (13.17)$$

Una familia de funciones $g(\cdot)$ 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 13.2. 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 (13.17), es decir,

$$W_{(\lambda),i} = \vec{x}_i' \vec{\beta} + \epsilon_i \quad (13.18)$$

$$\vec{\epsilon} \sim N(\vec{0}, \sigma^2 I) \quad (13.19)$$

Podemos, dadas las observaciones X, \vec{y} , escribir la verosimilitud conjunta de todos los parámetros: β , σ , y λ . Dicha verosimilitud puede escribirse en función de \vec{w} así¹:

$$f_{\vec{Y}}(\vec{y}) = f_{\vec{W}}(\vec{w}) |J(\lambda)| \quad (13.20)$$

¹La variable transformada \vec{w} depende en todo caso del λ empleado en la transformación; omitimos dicha dependencia para aligerar la notación, salvo donde interese enfatizarla.

siendo $J(\lambda)$ el jacobiano de la transformación:

$$J(\lambda) = \left| \frac{\partial \vec{w}}{\partial \vec{y}} \right| = \prod_{i=1}^N y_i^{\lambda-1} \quad (13.21)$$

Por tanto:

$$\begin{aligned} \log \text{ver}(\vec{\beta}, \lambda, \sigma^2; \vec{Y}) &= \log \left(\frac{1}{\sqrt{2\pi}} \right)^N \left(\frac{1}{|\sigma^2 I|^{\frac{1}{2}}} \right) \\ &\quad \times \log \left[\exp \left\{ -\frac{1}{2} \frac{(\vec{w}_{(\lambda)} - X\vec{\beta})'(\vec{w}_{(\lambda)} - X\vec{\beta})}{\sigma^2} \right\} |J(\lambda)| \right] \\ &= -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log \sigma^2 \\ &\quad - \frac{1}{2} \frac{(\vec{w}_{(\lambda)} - X\vec{\beta})'(\vec{w}_{(\lambda)} - X\vec{\beta})}{\sigma^2} + \log \prod_{i=1}^N y_i^{\lambda-1} \\ &= -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log \sigma^2 + (\lambda - 1) \sum_{i=1}^N \log y_i \\ &\quad - \frac{1}{2} \frac{\vec{w}_{(\lambda)}'(I - X(X'X)^{-1}X')\vec{w}_{(\lambda)}}{\sigma^2} \end{aligned} \quad (13.22)$$

La expresión (13.22) 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 (2.36), pág. 22) $\vec{w}_{(\lambda)}'(I - X(X'X)^{-1}X')\vec{w}_{(\lambda)}$.

Si ahora maximizamos (13.22) respecto a σ^2 , vemos que el máximo se alcanza para,

$$\hat{\sigma}_{(\lambda)}^2 = \frac{\vec{w}_{(\lambda)}'(I - X(X'X)^{-1}X')\vec{w}_{(\lambda)}}{N}$$

y el logaritmo de la verosimilitud concentrada es:

$$\log \text{ver}(\lambda; \vec{Y}) = -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log \hat{\sigma}_{(\lambda)}^2 - \frac{N}{2} + (\lambda - 1) \sum_{i=1}^N \log y_i \quad (13.23)$$

Podemos escoger como transformación aquélla cuyo λ maximice (13.23), o, de modo equivalente, tras prescindir de las constantes,

$$\log \text{ver}(\lambda; \vec{Y}) = -\frac{N}{2} \log \hat{\sigma}_{(\lambda)}^2 + (\lambda - 1) \sum_{i=1}^N \log y_i. \quad (13.24)$$

Un modo sencillo de hacerlo consiste en tomar un número adecuado de valores de λ equiespaciados en un intervalo susceptible de contener el λ óptimo, ajustar una regresión para cada λ , y calcular el correspondiente valor de (13.24). 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 λ .

Es frecuente que $\log \text{ver}(\lambda; \vec{Y})$ como función de λ sea una función relativamente plana. Ello suscita el problema de decidir si el valor de λ 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 λ y $L(\lambda)$ el valor que toma la verosimilitud, para muestras grandes se tiene que

$$2 \ln \left(\frac{L(\hat{\lambda})}{L(\lambda_0)} \right) \sim \chi_1^2; \quad (13.25)$$

por tanto, a la vista de (13.23), rechazaremos H_0 al nivel de significación α si

$$-2 \left(\frac{N}{2} \log \hat{\sigma}_{(\hat{\lambda})}^2 + (\hat{\lambda} - \lambda_0) \sum_{i=1}^N \log y_i - \frac{N}{2} \log \hat{\sigma}_{(\lambda_0)}^2 \right) > \chi_{1;\alpha}^2. \quad (13.26)$$

Utilizando la misma idea podemos construir intervalos de confianza para λ .

Capítulo 14

Regresión con respuesta cualitativa

14.1. 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,

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon} \quad (14.1)$$

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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

se deduce que ϵ 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 proporcione 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:

$$E[Y_i] = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{i,p-1} = P_i$$

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,

$$E[\epsilon_i^2] = (1 - P_i)^2 P_i + (-P_i)^2 (1 - P_i) = Q_i^2 P_i + P_i^2 Q_i = P_i Q_i. \quad (14.2)$$

La varianza de Y varía por tanto de observación a observación de acuerdo con los valores que toman los regresores. Adicionalmente, (14.2) muestra que la distribución de ϵ_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*,

$$\ell(E(Y_i)) \stackrel{\text{def}}{=} \ln \left(\frac{P_i}{1 - P_i} \right). \quad (14.3)$$

Nótese que como $E(Y_i) = P_i$, (14.3) 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:

$$\ell(E(Y_i)) = \ln \left(\frac{P_i}{1 - P_i} \right) = \vec{x}_i' \vec{\beta}, \quad (14.4)$$

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 14.1 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,

$$P_i = \frac{\exp(\vec{x}_i' \vec{\beta})}{1 + \exp(\vec{x}_i' \vec{\beta})}. \tag{14.5}$$

Interpretación de los coeficientes

Los parámetros de un modelo *logit* tienen interpretación inmediata: β_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:

	X = 1	X = 0
Y = 1	n_{11}	n_{12}
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:

$$\psi = \frac{\pi(1)/(1 - \pi(1))}{\pi(0)/(1 - \pi(0))}. \tag{14.6}$$

Entonces,

$$\begin{aligned} \ln(\psi) &= \ln\left(\frac{\pi(1)/(1 - \pi(1))}{\pi(0)/(1 - \pi(0))}\right) \\ &= \ln\left(\frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \bigg/ \frac{1}{1 + e^{\beta_0 + \beta_1}}\right) - \ln\left(\frac{e^{\beta_0}}{1 + e^{\beta_0}} \bigg/ \frac{1}{1 + e^{\beta_0}}\right) \\ &= \ln\left(\frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}}\right) \\ &= \beta_1. \end{aligned} \tag{14.7}$$

Por tanto, $\hat{\beta}_1$ estimará $\ln(\psi)$, y $\exp(\hat{\beta}_1)$ estimará ψ .

Observación 14.2 La codificación de X , al igual que la de Y , es arbitraria. La interpretación correcta de β_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 ψ 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.

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 (14.5)? 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-contróles*. 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 (*contróles*). Los coeficientes β_1, \dots, β_p son interpretables, pero β_0 no lo es. Ninguna fórmula que lo requiera —como (14.5)— puede utilizarse.

La razón es fácil de entender: $\hat{\beta}_0$ depende de la abundancia relativa de casos y contróles, y ésta es como hemos dicho arbitraria. La situación se asemeja a la que se presenta cuando construimos una tabla de contingencia 2×2 como:

	X = 1	X = 0	Total
Y = 1	n_{11}	n_{12}	$n_{1.}$
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 $n_{1\cdot}/n_{\cdot\cdot}$ como estimador de la proporción de casos con $Y = 1$ en la población (y cocientes como $n_{11}/n_{\cdot 1}$ o $n_{12}/n_{\cdot 2}$ estimarían las proporciones en las subpoblaciones caracterizadas por $X = 1$ y $X = 0$ respectivamente).

Si, por el contrario, hemos fijado los valores $n_{1\cdot}$ y $n_{2\cdot}$, es claro que dicho cociente no estima nada, sino que es resultado de una decisión arbitraria.

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

$$L(\hat{\beta}, \vec{y}, X) = \prod_{i=1}^N (P_i)^{y_i} (1 - P_i)^{1-y_i} \tag{14.8}$$

$$= \prod_{i=1}^N \left(\frac{1}{1 + \exp(\vec{x}_i' \vec{\beta})} \right)^{1-y_i} \left(\frac{\exp(\vec{x}_i' \vec{\beta})}{1 + \exp(\vec{x}_i' \vec{\beta})} \right)^{y_i} \tag{14.9}$$

$$= \prod_{i=1}^N \left(\frac{1}{1 + \tau_i} \right)^{1-y_i} \left(\frac{\tau_i}{1 + \tau_i} \right)^{y_i}, \tag{14.10}$$

con $\tau_i = \exp(\vec{x}_i' \vec{\beta})$. Tomando logaritmos en (14.10), obtenemos

$$\sum_{i=1}^N \ln \left(\frac{1}{1 + \tau_i} \right) + \sum_{i=1}^N y_i \ln(\tau_i). \tag{14.11}$$

Si derivamos (14.11) 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 (14.9) mediante un algoritmo conveniente.

Observación 14.3 La verosimilitud en (14.9) 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`).

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 $\hat{P} = (\hat{P}_1, \dots, \hat{P}_k)$.

De acuerdo con la teoría en la Sección B.3, bajo la hipótesis nula de que el modelo correcto es (14.4)

$$-2 \ln \left(\frac{L(\hat{\beta})}{L(\hat{P})} \right) \sim \chi_{k-p}, \quad (14.12)$$

en que p es el número de parámetros estimados en $\hat{\beta}$. Al cociente (14.12) 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 (14.12) positiva cuya distribución, bajo la hipótesis nula, sigue la distribución χ_{k-p}^2 indicada. Si la desviación fuera excesiva (es decir, si sobrepasa $\chi_{k-p;\alpha}^2$ para el nivel de significación α 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 β_j es significativamente diferente de cero en un cierto modelo parametrizado por $\vec{\beta}$, calcularíamos

$$-2 \ln \left(\frac{L(\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \dots, \hat{\beta}_k)}{L(\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{j-1}, \hat{\beta}_j, \hat{\beta}_{j+1}, \dots, \hat{\beta}_k)} \right), \quad (14.13)$$

que debe ser comparado con una χ_1^2 ; valores grandes de (14.13) son evidencia contra la hipótesis $h : \beta_j = 0$.

Para contrastar la hipótesis de nulidad de todos los parámetros, salvo quizá β_0 afectando a la columna de “unos”, compararíamos

$$-2 \ln \left(\frac{L(\hat{\beta}_0)}{L(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)} \right) \quad (14.14)$$

a una χ_{k-1}^2 ; la expresión (14.14) es similar a la suma de cuadrados SSR en una regresión ordinaria. El análogo a SST sería

$$-2 \ln \left(\frac{L(\hat{\beta}_0)}{L(\hat{P})} \right). \quad (14.15)$$

Esta analogía puede extenderse para obtener un estadístico similar a la C_p de Mallows así:

$$\Delta_k = -2 \ln \left(\frac{L(\hat{\beta}_0)}{L(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)} \right) - 2(k-1), \quad (14.16)$$

y una “ R^2 ” así:

$$R^2 = \frac{-2 \ln \left(\frac{L(\hat{\beta}_0)}{L(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)} \right)}{-2 \ln \left(\frac{L(\hat{\beta}_0)}{L(\hat{P})} \right)} \quad (14.17)$$

Obsérvese que en (14.16) el primer sumando de la derecha sigue asintóticamente una distribución χ_{k-1}^2 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 χ_{k-1}^2 . Mientras que el primero crece monótonamente al introducir nuevos parámetros, el segundo penaliza este crecimiento.

Observación 14.4 Escogeríamos de acuerdo con este criterio el modelo maximizando Δ_k o, alternativamente, minimizando

$$\text{AIC}_k = -2 \ln L(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k) + 2k. \quad (14.18)$$

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).

COMPLEMENTOS Y EJERCICIOS

14.1 Muéstrese que la *desviación* definida a continuación de (14.12) coincide con SSE cuando consideramos un modelo lineal ordinario con normalidad en las perturbaciones.

14.2 Compruébese derivando (14.11) 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 $\tau_i = \vec{x}_i' \vec{\beta}$.

Apéndice A

Algunos resultados en Algebra Lineal.

A.1. Resultados varios sobre Algebra Matricial.

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 $H \times H \rightarrow R$ (si es real-valorado) o en C (si es complejo valorado), tal que a cada par de vectores \vec{u}, \vec{v} corresponde $\langle \vec{u}, \vec{v} \rangle$ verificando:*

$$\langle \vec{u}, \vec{v} \rangle = \overline{\langle \vec{v}, \vec{u} \rangle} \quad (\text{A.1})$$

$$\langle \vec{u}, \vec{u} \rangle \geq 0 \quad \forall \vec{u} \in H \quad (\text{A.2})$$

$$\langle \vec{u}, \vec{u} \rangle = 0 \implies \vec{u} = 0 \quad (\text{A.3})$$

$$\langle \vec{u}, \alpha\vec{v} + \beta\vec{w} \rangle = \alpha \langle \vec{u}, \vec{v} \rangle + \beta \langle \vec{u}, \vec{w} \rangle \quad (\text{A.4})$$

Definición A.2 *Llamamos producto interno euclídeo de dos n -eplas \vec{u}, \vec{v} en R^n al definido así: $\langle \vec{u}, \vec{v} \rangle = \vec{u}'\vec{v}$. Es fácil comprobar que verifica las condiciones de la Definición A.1. La norma euclídea $\|\vec{u}\|$ del vector \vec{u} se define como $\|\vec{u}\| = +\sqrt{\langle \vec{u}, \vec{u} \rangle} = \sqrt{u_1^2 + \dots + u_n^2}$*

Definición A.3 *Dados dos vectores \vec{u}, \vec{v} en un espacio vectorial, definimos el coseno del ángulo que forman como*

$$\cos(\alpha) = \frac{\langle \vec{u}, \vec{v} \rangle}{\|\vec{u}\| \|\vec{v}\|}. \quad (\text{A.5})$$

Teorema A.2 (Sherman-Morrison-Woodbury) *Sea D una matriz simétrica $p \times p$ y \vec{a}, \vec{c} vectores $p \times 1$. Entonces,*

$$(D + \vec{a}\vec{c}')^{-1} = D^{-1} - D^{-1}\vec{a}(1 + \vec{c}'D^{-1}\vec{a})^{-1}\vec{c}'D^{-1} \quad (\text{A.6})$$

DEMOSTRACIÓN:

Multiplicando ambos lados de (A.6) 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:

$$(D + \vec{z}\vec{z}')^{-1} = D^{-1} - D^{-1}\vec{z}(1 + \vec{z}'D^{-1}\vec{z})^{-1}\vec{z}'D^{-1} \quad (\text{A.7})$$

Teorema A.3 *Si A y D son simétricas y todas las inversas existen:*

$$\begin{pmatrix} A & B \\ B' & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + FE^{-1}F' & -FE^{-1} \\ E^{-1}F' & E^{-1} \end{pmatrix} \quad (\text{A.8})$$

siendo

$$E = D - B'A^{-1}B \quad (\text{A.9})$$

$$F = A^{-1}B \quad (\text{A.10})$$

DEMOSTRACIÓN:

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.8) proporciona entonces para el bloque superior izquierdo:

$$\begin{aligned} A^{-1} + FE^{-1}F' &= (X'X)^{-1} + \\ &+ (X'X)^{-1}X'Z[Z'Z - Z'X(X'X)^{-1}X'Z]^{-1}Z'X(X'X)^{-1} \end{aligned} \quad (\text{A.11})$$

y similarmente para los demás bloques. Véase Seber (1977), pág. 390 y Myers (1990), pág. 459.

A.2. Cálculo diferencial con notación matricial

Hay aquí sólo una breve recopilación de resultados útiles. Más detalles y demostraciones en Abadir and Magnus (2005), Searle (1982) y Magnus and Neudecker (1988).

Haremos uso de las siguientes definiciones y notación.

Definición A.4 Sea \vec{x} un vector $m \times 1$ e y una función escalar de \vec{x} : $y = f(x_1, \dots, x_m) = f(\vec{x})$. Entonces:

$$\begin{pmatrix} \frac{\partial y}{\partial x} \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_m} \end{pmatrix}$$

Si $y = \vec{x}' A \vec{x}$ siendo A una matriz cuadrada cualquiera, es inmediato comprobar que:

$$\begin{pmatrix} \frac{\partial y}{\partial \vec{x}} \end{pmatrix} = (A + A') \vec{x}.$$

En el caso, frecuente, de que A sea simétrica, tenemos que:

$$\begin{pmatrix} \frac{\partial y}{\partial \vec{x}} \end{pmatrix} = 2A' \vec{x} \tag{A.12}$$

Definición A.5 Sea \vec{y} una función vectorial $(n \times 1)$ -valorada de \vec{x} , vector $m \times 1$. Entonces:

$$\begin{pmatrix} \frac{\partial \vec{y}}{\partial \vec{x}} \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_1} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_1}{\partial x_m} & \frac{\partial y_2}{\partial x_m} & \cdots & \frac{\partial y_n}{\partial x_m} \end{pmatrix}$$

Hay algunos casos particulares de interés. Si $y = \vec{a}' \vec{x} = a_1 x_1 + \dots + a_m x_m$, siendo \vec{a} un vector de constantes,

$$\frac{\partial y}{\partial \vec{x}} = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} = \vec{a};$$

si $\vec{y} = A\vec{x}$, siendo A una matriz ($n \times m$) de constantes,

$$\left(\frac{\partial \vec{y}}{\partial \vec{x}} \right) = A'.$$

Se reproducen a continuación algunos otros resultados útiles:

$$\frac{\partial \log_e |A|}{\partial A} = [A']^{-1} \quad (\text{A.13})$$

$$\frac{\partial \text{tr}(BA^{-1}C)}{\partial A} = -(A^{-1}CBA^{-1}) \quad (\text{A.14})$$

A.3. Lectura recomendada

Hay muchos manuales de álgebra lineal en que se pueden encontrar los resultados anteriores. Entre los particularmente orientados a la Estadística, pueden citarse Gentle (2007), Seber (2007), Abadir and Magnus (2005), o Searle (1982). En relación con las cuestiones numéricas específicamente relacionadas con la estimación mínimo-cuadrática es todavía de útil consulta Lawson and Hanson (1974).

Apéndice B

Algunos prerrequisitos estadísticos.

B.1. Distribuciones χ^2 y \mathcal{F} descentradas

Sean $X_i \stackrel{\text{indep}}{\sim} N(\mu_i, \sigma^2)$, ($i = 1 \dots, n$). Sea $\delta^2 = (\mu_1^2 + \dots + \mu_n^2)/\sigma^2$. Entonces, la variable aleatoria

$$Z = \frac{X_1^2 + \dots + X_n^2}{\sigma^2} \quad (\text{B.1})$$

se dice que sigue una distribución $\chi_n^2(\delta)$, o distribución χ^2 *descentrada* con *parámetro de no centralidad* δ y n grados de libertad. Algunos textos definen δ^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 χ^2 habitual o *centrada*.

Si $Z \sim \chi_m^2(\delta)$ y $V \sim \chi_n^2$ son ambas independientes, la variable aleatoria

$$W = \frac{n}{m} \frac{Z}{V} \quad (\text{B.2})$$

sigue una distribución $\mathcal{F}_{m,n}(\delta)$ o \mathcal{F} de Snedecor descentrada, con parámetro de no centralidad δ . 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

$$\hat{\beta} \underset{\sim}{\overset{asint}{\sim}} N(\vec{\beta}, \Sigma_{\hat{\beta}}) \quad (\text{B.3})$$

$$\Sigma_{\hat{\beta}} \approx [I(\hat{\beta})]^{-1} \quad (\text{B.4})$$

En la expresión anterior, $I(\hat{\beta})$ es la llamada *matriz de información* cuyo elemento genérico de lugar ij se define así:

$$[I(\hat{\beta})]_{ij} = -\frac{\partial^2 \ell(\vec{\beta}, \vec{y})}{\partial \beta_i \partial \beta_j}. \quad (\text{B.5})$$

Una consecuencia de (B.3)–(B.4) es que si $\Sigma_{\hat{\beta}}$ es de dimensión $p \times p$,

$$(\hat{\beta} - \vec{\beta})' (\Sigma_{\hat{\beta}})^{-1} (\hat{\beta} - \vec{\beta}) \sim (\hat{\beta} - \vec{\beta})' I(\hat{\beta}) (\hat{\beta} - \vec{\beta}) \sim \chi_p^2;$$

esto permite contrastar hipótesis como $H_0 : \vec{\beta} = \vec{\beta}_0$ utilizando como estadístico

$$(\hat{\beta} - \vec{\beta}_0)' I(\vec{\beta}_0) (\hat{\beta} - \vec{\beta}_0) \quad (\text{B.6})$$

o alternativamente

$$(\hat{\beta} - \vec{\beta}_0)' I(\hat{\beta}) (\hat{\beta} - \vec{\beta}_0). \quad (\text{B.7})$$

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

$$\hat{\beta}_h = \arg \max_{\vec{\beta} \in h} L(\vec{\beta}, \vec{Y}) \quad (\text{B.8})$$

$$\hat{\beta}_M = \arg \max_{\vec{\beta} \in M} L(\vec{\beta}, \vec{Y}). \quad (\text{B.9})$$

Entonces, en condiciones muy generales, que no requieren que \vec{Y} siga una distribución particular, se verifica que bajo H_0 ,

$$-2 \log_e \left(\frac{L(\hat{\beta}_h, \vec{Y})}{L(\hat{\beta}_M, \vec{Y})} \right) \sim \chi_{(p-q)}^2. \quad (\text{B.10})$$

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 $\chi_{(p-q); \alpha}^2$; valores del estadístico mayores que dicho cuantil 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 superconjunto del S 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).

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 consultarse 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 amplía 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. 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.

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 incluir 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. También 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 factorización QR de la matriz `x` de regresores. Véase la función `qr` en Becker et al. (1988). Tiene utilidad para computar algunos resultados.

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 incluir 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.

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 incluir en la matriz `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 11).

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:

<code>method</code>	Método de ajuste a emplear. Por omisión, se utiliza la factorización QR.
<code>data</code>	Una “data frame” conteniendo los datos tanto de regresores como de variable respuesta.
<code>formula</code>	Una expresión del tipo $\text{Resp} \sim \text{Regr01} + \text{Regre02} + \log(\text{Regre03})$ en que a la izquierda está el regresando y a la derecha los regresores o funciones de ellos.
<code>subset</code>	Criterio para seleccionar las filas de la tabla de datos que deseamos emplear.
<code>na.action</code>	Acción a tomar cuando algún dato en una fila de la tabla de datos es NA. Por omisión es omitir dicha fila.
<code>model,x,y</code>	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.

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).

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 11).

`stud.res` Residuos externamente studentizados (los t_i en la notación del Capítulo 11).

`cooks` Un vector conteniendo las distancias de Cook (D_i en la notación del Capítulo 11).

`dfits` Un vector conteniendo los DFITS mencionados en el Capítulo 11).

`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
multcomp	-	-	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 <i>ridge</i>

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,

$$\begin{aligned} \min_{\vec{x}} \|Q(D\vec{x} - \vec{c})\|^2 &= \min_{\vec{x}} \langle Q(D\vec{x} - \vec{c}), Q(D\vec{x} - \vec{c}) \rangle \\ &= \min_{\vec{x}} (D\vec{x} - \vec{c})' Q' Q (D\vec{x} - \vec{c}) \\ &= \min_{\vec{x}} \|D\vec{x} - \vec{c}\|^2 \end{aligned}$$

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,

$$D = HRK' \tag{D.2}$$

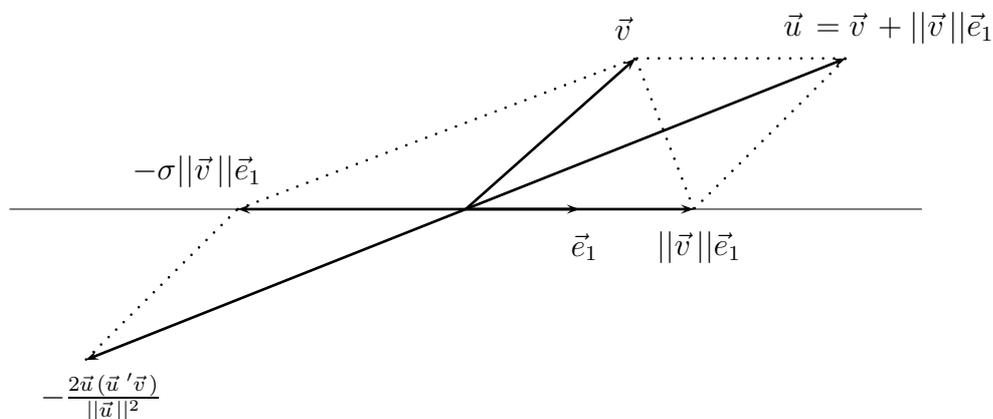
Sea el problema

$$\min_{\vec{x}} \|D\vec{x} - \vec{y}\|^2 \tag{D.3}$$

y definamos,

$$\begin{aligned} H'\vec{y} &= \vec{g} = \begin{pmatrix} \vec{g}_1 \\ \vec{g}_2 \end{pmatrix} \begin{matrix} k \\ n-k \end{matrix} \\ K'\vec{x} &= \vec{\gamma} = \begin{pmatrix} \vec{\gamma}_1 \\ \vec{\gamma}_2 \end{pmatrix} \begin{matrix} k \\ m-k \end{matrix} \end{aligned}$$

Figura D.1: Visualización de la transformación de Householder.



Sea $\tilde{\gamma}_1$ la solución (única) del sistema,

$$R_{11}\tilde{\gamma}_1 = \vec{g}_1.$$

Entonces, todas las posibles soluciones del problema (D.3) son de la forma

$$\vec{x} = K \begin{pmatrix} \tilde{\gamma}_1 \\ \tilde{\gamma}_2 \end{pmatrix},$$

con γ_2 arbitrario. Cualquiera de esas soluciones da lugar al vector de residuos

$$\vec{r} = \vec{y} - D\vec{x} = H \begin{pmatrix} \vec{0} \\ \vec{g}_2 \end{pmatrix}$$

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:

$$P\vec{v} = -\sigma\|\vec{v}\|\vec{e}_1 \tag{D.4}$$

siendo

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (\text{D.5})$$

$$\sigma = \begin{cases} +1 & \text{si } v_1 \geq 0 \\ -1 & \text{si } v_1 < 0. \end{cases} \quad (\text{D.6})$$

Esta matriz tiene por expresión,

$$P = I - 2 \frac{\vec{u} \vec{u}'}{\|\vec{u}\|^2} \quad (\text{D.7})$$

con $\vec{u} = \vec{v} + \sigma \|\vec{v}\| \vec{e}_1$.

DEMOSTRACIÓN:

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,

$$P\vec{v} = \left(I - 2 \frac{\vec{u} \vec{u}'}{\|\vec{u}\|^2} \right) \left(\frac{1}{2}\vec{u} + \frac{1}{2}\vec{z} \right) \quad (\text{D.10})$$

$$= \frac{1}{2}\vec{u} - \vec{u} + \frac{1}{2}\vec{z} \quad (\text{D.11})$$

$$= -\frac{1}{2}\vec{u} + \vec{v} - \frac{1}{2}\vec{u} \quad (\text{D.12})$$

$$= \vec{v} - \vec{u} \quad (\text{D.13})$$

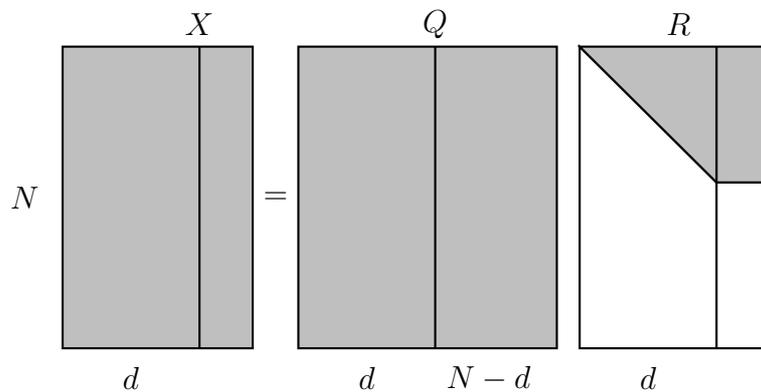
$$= -\sigma \|\vec{v}\| \vec{e}_1 \quad (\text{D.14})$$

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,



DEMOSTRACIÓN:

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,

$$P_1 X = \begin{array}{|c|c} \hline & \text{shaded area} \\ \hline \text{white column} & \text{shaded area} \\ \hline \end{array}$$

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:

$$\begin{pmatrix} 1 & \vec{0}' \\ \vec{0} & P_2^* \end{pmatrix} P_1 X = \begin{array}{|c|c|} \hline \text{[Diagrama esquemático de una matriz con una columna blanca y una columna gris]} \\ \hline \end{array}$$

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 $Q' = P_d P_{d-1} \dots P_1$ 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.

Apéndice E

Enunciados y demostraciones formales

Se incluyen aquí teoremas, desarrollos y demostraciones omitidos en el curso de la exposición, por su nivel de formalismo o por no ser esenciales.

E.1. Existencia y unicidad de proyecciones.

Definición E.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 $\langle \cdot, \cdot \rangle$ y correspondiente norma $\|\vec{v}\|^2 = \langle \vec{v}, \vec{v} \rangle$. Decimos que $\{\vec{v}_n\}$ es una sucesión de Cauchy si para cualquier $\delta > 0$ hay un $N(\delta)$ tal que $\forall m, n \geq N(\delta), \|\vec{v}_n - \vec{v}_m\| < \delta$; es decir, si prefijado un δ 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 δ .

Definición E.2 Sea H un espacio vectorial como en la Definición E.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 , infinito-dimensional y separable. Cualquier subespacio vectorial de un espacio de Hilbert, es a su vez espacio de Hilbert.

Teorema E.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:

$$\|\vec{y} - \vec{v}\|^2 = \min_{\vec{z} \in M} \|\vec{y} - \vec{z}\|^2. \quad (\text{E.1})$$



Demostración. Veamos¹ primero la existencia. Sea $d = \min_{\vec{z} \in M} \|\vec{y} - \vec{z}\|^2$. Entonces, necesariamente existirá en M algún vector \vec{v}_1 tal que: $\|\vec{y} - \vec{v}_1\|^2 \leq d + 1$; de no haberlo, mín $\|\vec{y} - \vec{z}\|^2$ tendría que ser mayor que $d + 1$, contra la hipótesis. Análogamente, para cualquier número natural n existirá \vec{v}_n verificando: $\|\vec{y} - \vec{v}_n\|^2 \leq d + 1/n$. 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:

$$D = \|(\vec{y} - \vec{v}_n) - (\vec{y} - \vec{v}_m)\|^2 + \|(\vec{y} - \vec{v}_n) + (\vec{y} - \vec{v}_m)\|^2 \quad (\text{E.2})$$

Podemos escribir:

$$\begin{aligned} D &= \|(\vec{y} - \vec{v}_n)\|^2 + \|(\vec{y} - \vec{v}_m)\|^2 - 2 \langle (\vec{y} - \vec{v}_m), (\vec{y} - \vec{v}_n) \rangle \\ &\quad + \|(\vec{y} - \vec{v}_n)\|^2 + \|(\vec{y} - \vec{v}_m)\|^2 + 2 \langle (\vec{y} - \vec{v}_m), (\vec{y} - \vec{v}_n) \rangle \\ &= 2\|(\vec{y} - \vec{v}_n)\|^2 + 2\|(\vec{y} - \vec{v}_m)\|^2. \end{aligned} \quad (\text{E.3})$$

Por otra parte, tenemos:

$$\begin{aligned} D &= \|(\vec{v}_m - \vec{v}_n)\|^2 + \|2\vec{y} - 2(\frac{1}{2})(\vec{v}_n + \vec{v}_m)\|^2 \\ &= \|(\vec{v}_m - \vec{v}_n)\|^2 + 4\|\vec{y} - (\frac{1}{2})(\vec{v}_n + \vec{v}_m)\|^2. \end{aligned} \quad (\text{E.4})$$

Igualando (E.3) y (E.4) obtenemos:

$$\begin{aligned} \|\vec{v}_m - \vec{v}_n\|^2 &= 2\|\vec{y} - \vec{v}_n\|^2 + 2\|\vec{y} - \vec{v}_m\|^2 \\ &\quad - 4\|\vec{y} - (\frac{1}{2})(\vec{v}_n + \vec{v}_m)\|^2. \end{aligned} \quad (\text{E.5})$$

Como la norma al cuadrado del último término de (E.5) es al menos d , tenemos:

$$\|\vec{v}_m - \vec{v}_n\|^2 \leq 2\|(\vec{y} - \vec{v}_n)\|^2 + 2\|(\vec{y} - \vec{v}_m)\|^2 - 4d \quad (\text{E.6})$$

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

$$\|(\vec{y} - \vec{v}_n)\|^2 \leq d + \delta/4 \quad (\text{E.7})$$

$$\|(\vec{y} - \vec{v}_m)\|^2 \leq d + \delta/4. \quad (\text{E.8})$$

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.

Sustituyendo ésto en (E.5) obtenemos:

$$\|(\vec{v}_m - \vec{v}_n)\|^2 \leq 2(d + \delta/4) + 2(d + \delta/4) - 4d = \delta, \quad (\text{E.9})$$

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 $\|\vec{y} - \vec{v}\|^2 = d$.

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

$$\|\vec{y} - \vec{v} - \alpha\vec{z}\|^2 = \|\vec{y} - \vec{v}\|^2 + \alpha^2\|\vec{z}\|^2 - 2\alpha\langle\vec{y} - \vec{v}, \vec{z}\rangle \quad (\text{E.10})$$

$$= d + \alpha^2\|\vec{z}\|^2 - 2\alpha\langle\vec{y} - \vec{v}, \vec{z}\rangle \quad (\text{E.11})$$

$$\geq d. \quad (\text{E.12})$$

Por tanto:

$$\alpha^2\|\vec{z}\|^2 - 2\alpha\langle\vec{y} - \vec{v}, \vec{z}\rangle \geq 0, \quad (\text{E.13})$$

$$\alpha^2\|\vec{z}\|^2 \geq 2\alpha\langle\vec{y} - \vec{v}, \vec{z}\rangle. \quad (\text{E.14})$$

Como (E.14) se ha de cumplir para cualquier posible valor de α , ha de suceder que $\langle\vec{y} - \vec{v}, \vec{z}\rangle = 0$, 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 1.1). 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,

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

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

Observación E.1 ¿Qué trascendencia tiene en el enunciado del Teorema E.1 que H (y, en consecuencia, su subespacio M) tengan estructura de espacio de Hilbert? Examinando la demostración del Teorema E.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 .

Observación E.2   ¿Debemos preocuparnos de verificar que estamos ante un espacio de Hilbert? ¿Cómo hacerlo? Cuando los regresores generan un espacio de dimension finita, nada de ello es preciso. Cuando se hace análisis de series temporales, la mejor predicción lineal en el momento t del valor de la misma en $t + 1$ (predicción una etapa hacia adelante) se hace proyectando y_{t+1} sobre el subespacio que generan $y_t, y_{t-1}, y_{t-2}, \dots$ (todo el “pasado” de la serie). Este “pasado”, al menos en principio, puede ser infinito dimensional y aquí sí tiene objeto suponer que genera un espacio de Hilbert para garantizar la existencia de la proyección.

Nótese, incidentalmente, que en este problema emplearíamos una norma que no sería la euclídea ordinaria, sino la inducida por el producto interno $\langle y_t, y_s \rangle = E[y_t y_s]$ (supuesta estacionariedad y media cero). Pueden verse más detalles en la obra ya citada Anderson (1971), Sección 7.6. Ejemplos del uso del espacio de Hilbert en series temporales pueden verse en Davis (1977), Cap. 2, o Shumway and Stoffer (2006), Apéndice B.1.

E.2. Proyección sobre subespacios $h = M \cap K(B)$.

El **Lema 4.4** decía:

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,

$$M \cap h^\perp = M \cap R(B'); \tag{E.15}$$

véase el Ejercicio 4.2, pág. 57.

Probaremos ahora que ambos subespacios considerados en el enunciado son el mismo, utilizando la expresión (E.15), y mostrando la mutua inclusión.

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

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

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

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

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:

$$\langle \vec{x}, \vec{z} \rangle = \vec{x}' \vec{z} = \vec{a}' B P_M \vec{z} = \vec{a}' B \vec{z} = 0$$

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. ■

Bibliografía

- Abadir, K. and Magnus, J. (2005). *Matrix Algebra*. Cambridge Univ. Press.
- Akaike, H. (1972). Use of an Information Theoretic Quantity for Statistical Model Identification. In *Proc. 5th. Hawaii Int. Conf. on System Sciences*, pp. 249–250.
- Akaike, H. (1974). Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. 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. W. (1971). *The Statistical Analysis of Time Series*. New York: Wiley.
- Ansley, C. F. (1985). Quick Proofs of Some Regression Theorems Via the QR Algorithm. *As*, 39, 55–59.
- Arnold, S. F. (1981). *The Theory of Linear Models and Multivariate Analysis*. New York: Wiley.
- Atkinson, A. C. (1985). *Plots, Transformations and Regression*. Oxford Univ. Press.
- Barnett, V. and Lewis, T. (1978). *Outliers in Statistical Data*. New York: Wiley.
- Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). *The New S Language. A Programming Environment for Data Analysis and Graphics*. Pacific Grove, California: Wadsworth & Brooks/Cole.

- Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.
- Ben-Israel, A. and Greville, T. N. E. (1974). *Generalized Inverses: Theory and Applications*. New York: Wiley.
- Bishop, C. M. (1996). *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.
- Box, G. E. P. and Tidwell, P. W. (1962). Transformations of the Independent Variables. *Technometrics*, 4, 531–550.
- Brown, P. J. (1993). *Measurement, Regression and Calibration*. Clarendon Press/Oxford, Signatura: 519.235.5 BRO.
- Chambers, J. M. (1998). *Programming with Data*. Mathsoft.
- Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*. Pacific Grove, Ca.: Wadsworth & Brooks/Cole.
- Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.
- Cornillon, P.-A. and Matzner-Lober, E. (2011). *Régression avec R*. Springer Verlag.
- 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.
- Dalgaard, P. (2002). *Introductory Statistics with R*. Statistics and Computing, Springer-Verlag, Signatura: 519.682 DAL.
- Davis, M. H. A. (1977). *Linear Estimation and Stochastic Control*. Chapman and Hall.
- de Leeuw, J. (2000). Information Theory and an Extension of the Maximum Likelihood Principle by Hirotugu Akaike. Disponible en <http://www.stat.ucla.edu/~deleeuw/work/research.phtml>.

- Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. Wiley, third edition, Signatura: 519.233.5 DRA.
- Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker.
- Faraway, J. J. (2005). *Linear Models with R*. Chapman & Hall/CRC, Signatura: 519.233 FAR.
- Fox, J. (2002). *An R and S-Plus Companion to Applied Regression*. Sage Pub.
- Garthwaite, P. H., Jolliffe, I. T., and Jones, B. (1995). *Statistical Inference*. London: Prentice Hall.
- Gentle, J. (2007). *Matrix Algebra: Theory, Computations, and Applications in Statistics*. Springer.
- Goodhall, C. R. (1993). Computation Using the QR Decomposition. In C. R. Rao, editor, *Handbook of Statistics*, chapter 13, pp. 467–508, Amsterdam: North-Holland.
- Grafe, J. H. (1985). *Matemáticas Universitarias*. Madrid: MacGraw-Hill.
- Gunst, R. F. and Mason, R. L. (1980). *Regression Analysis and Its Applications. A Data Oriented Approach*. New York: Marcel Dekker, Inc.
- Haitovsky, Y. (1969). A Note on Maximization of \overline{R}^2 . *As*, 23, 20–21.
- Harrell, F. E. (2001). *Regression Modelling Strategies*. Springer-Verlag, Signatura: 519.233.5 HAR.
- 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. M. (1980). *Identification of Outliers*. London: Chapman & Hall.
- Haykin, S. (1998). *Neural Networks. A Comprehensive Foundation*. Prentice Hall, second edition.
- Hocking, R. R. (1976). The Analysis and Selection of Variables in Linear Regression. *Biometrics*, 32, 1–49.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Non-Orthogonal Problems. *Technometrics*, 12, 55–67.

- Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975). Ridge Regression: Some Simulations. *Cstat*, 4, 105–123.
- Hosmer, D. W. 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. T. (1986). *Principal Components Analysis*. New York: Springer-Verlag.
- Kennedy, W. J. (1980). *Statistical Computing*. New York: Marcel Dekker.
- Kleinbaum, D. G. (1994). *Logistic Regression. A Self-Learning Test*. Springer Verlag.
- Knuth, D. (1986). *The T_EX Book*. Reading, Mass.: Addison Wesley.
- Knuth, D. K. (1968). Fundamental Algorithms. In *The Art of Computer Programming*, volume 1, Reading, Mass.: Addison-Wesley.
- Kuhnert, P. and Venables, W. (2005). *An Introduction to R: Software for Statistical Modelling and Computing*. CSIRO Mathematical and Information Sciences, Cleveland, Australia.
- Lange, K. (1998). *Numerical Analysis for Statisticians*. Springer, Signatura: 519.6 LAN.
- Lawless, J. F. 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. L. Dowe, K. B. Korb, and J. 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. E. (1975). Tables for the Approximate Test for Outliers in Linear Regression. *Technometrics*, 17, 473–476.

- Magnus, J. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley.
- Maindonald, J. H. (2000). *Data Analysis and Graphics Using R - An Introduction*.
- Miller, A. (2002). *Subset Selection In Regression, Second Editon*. Chapman & Hall/CRC.
- Myers, R. H. (1990). *Classical and Modern Regression with Applications*. Boston: PWS-KENT Pub. Co.
- Peña, D. (2002). *Regresión y Diseño de Experimentos*. Alianza Editorial.
- Rao, C. R. and Mitra, S. K. (1971). *Generalized Inverse of Matrices and Its Applications*. John Wiley & Sons, New York [etc.].
- Ripley, B. D. (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. P. (1997). *Modern Regression Methods*. Wiley, Signatura: 519.233.4 RYA.
- Searle, S. R. (1971). *Linear Models*. New York: Wiley.
- Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. Wiley.
- Seber, G. (2007). *A Matrix Handbook for Statisticians*. Wiley.
- Seber, G. A. F. (1977). *Linear Regression Analysis*. New York: Wiley.
- Shapiro, S. S. and Francia, R. S. (1972). An Approximate Analysis of Variance Test for Normality. *Jasa*, 67, 215–216.
- Shapiro, S. S. and Wilk, M. B. (1965). An Analysis of Variance Test for Normality (complete Samples). *Biometrika*, 52, 591–611.
- Shumway, R. H. and Stoffer, D. S. (2006). *Time Series Analysis and Its Applications. With R Examples*. Springer Verlag.
- Silvey, S. D. (1969). Multicollinearity and Imprecise Estimation. *Jrssb*, 31, 539–552.
- Silvey, S. D. (1980). *Optimal Design*. London: Chapman & Hall.

- Stapleton, J. H. (1995). *Linear Statistical Models*. New York: Wiley.
- Theil, H. (1971). *Principles of Econometrics*. New York: Wiley.
- Thisted, R. A. (1988). *Elements of Statistical Computing*. New York: Chapman & Hall.
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288.
- 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.
- Ugarte, M., Militino, A., and Arnholt, A. (2008). *Probability and Statistics with R*. CRC Press.
- 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, libremente disponible en Internet.
- 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. D. (1999b). R Complements to *Modern Applied Statistics with S-Plus*. En <http://www.stats.ox.ac.uk/pub/MASS3>.
- Wang, C. (1993). *Sense and Nonsense of Statistical Inference*. New York: Marcel Dekker.
- Webster, J. T., Gunst, R. F., and Mason, R. L. (1974). Latent Root Regression Analysis. *Technometrics*, 16, 513–522.
- Yanai, H., Takeuchi, K., and Takane, Y. (2011). *Projection Matrices, Generalized Inverse Matrices and Singular Value Decomposition*, volume 34. Springer Verlag.