Capítulo 10

Regresión sesgada.

10.1. Introducción.

De acuerdo con el teorema de Gauss-Markov (Teorema 3.2, pág. 21), 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:

$$\begin{aligned} E[\hat{c}-c]^2 &= E\left[\hat{c}-E[\hat{c}]+E[\hat{c}]-c\right]^2 \\ &= E\left[\hat{c}-E[\hat{c}]\right]^2 + E\left[E[\hat{c}]-c\right]^2 + 2\underbrace{E\left[\hat{c}-E[\hat{c}]\right]\left[E[\hat{c}]-c\right]}_{=0} \\ &= \text{var}(\hat{c}) + (\text{sesgo } \hat{c})^2 \end{aligned} \tag{10.1}$$

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 (11.1).

El Capítulo 10 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 10.6, pág. 136.

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 11.1, pág. 147) 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
>
> $$\hat{\mu} = \frac{1}{2}(X_1 + X_2) \tag{10.2}$$

es un estimador insesgado de μ. Su varianza será $\mathrm{Var}(\hat{\mu}) = \sigma_1^2/4 + \sigma_2^2/4$.

¿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, $\mathrm{Var}(\hat{\mu}) = (\sigma_1^2 + \sigma_2^2)/4 = (1 + 99)/4 = 25$, mientras que $\hat{\mu}^* = X_1$, por ejemplo, sería también insesgado con $\mathrm{Var}(\hat{\mu^*}) = 1$.

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 (11.3), 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 (11.3) deducimos que

$$\begin{aligned} \mathrm{Var}(\hat{\mu}^{**}) &= \delta_1^2\sigma_1^2 + \delta_2^2\sigma_2^2 \\ &= \delta_1^2 \cdot 1 + (1-\delta_1)^2 \cdot 99 \\ &= 99 - 198\delta_1 + 100\delta_1^2 \end{aligned}$$

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 tolerasemos 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 10.5. El estimador de cualquier forma lineal $\vec{c}\,'\vec{\beta}$ puede escribirse como combinación lineal de $\vec{v}\,'_1\hat{\beta}, \vec{v}\,'_2\hat{\beta}, \ldots, \vec{v}\,'_p\hat{\beta}$, según muestra (10.29), pág. 136. Además, $\vec{v}\,'_i\hat{\beta}$ para $i = 1, \ldots, p$ son variables aleatorias incorreladas[1] con varianzas respectivas $\mathrm{Var}(\vec{v}_i\,'\hat{\beta}) = \sigma^2/\lambda_i$, (10.26), pág. 135.

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 11.1 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 11.4 prescinde de algunas $\vec{v}\,'_i\hat{\beta}$; el estimador *ridge* de la Sección 11.3 atenúa las $\vec{v}\,'_i\hat{\beta}$ más inestables. Volveremos de nuevo sobre la cuestión en la Sección 11.4.3, pág. 158.

10.3. Regresión ridge.

10.3.1. Error cuadrático medio del estimador mínimo cuadrático ordinario

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

$$\mathrm{ECM}(\hat{\beta}) = E[(\hat{\beta} - \vec{\beta}\,)'(\hat{\beta} - \vec{\beta}\,)] \tag{10.3}$$

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

[1] Independientes, si se verifica el supuesto de normalidad.

y $\Sigma_{\hat{\beta}} = \sigma^2(X'X)^{-1}$, tenemos que:

$$\begin{aligned}
\mathrm{ECM}(\hat{\beta}) &= E[\text{traza } (\hat{\beta} - \vec{\beta}\,)'(\hat{\beta} - \vec{\beta}\,)] \\
&= E[\text{traza } (\hat{\beta} - \vec{\beta}\,)(\hat{\beta} - \vec{\beta}\,)'] \\
&= \sigma^2 \text{traza } (X'X)^{-1} \\
&= \sigma^2 \text{traza } (X'X)^{-1}VV' \qquad (V = \text{diagonalizadora de } (X'X)^{-1}) \\
&= \sigma^2 \text{traza } V'(X'X)^{-1}V \\
&= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}, \qquad (10.4)
\end{aligned}$$

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.)

10.3.2. Clase de estimadores ridge

Definición 10.1 *Definiremos el estimador ridge de parámetro k así:*

$$\hat{\beta}^{(k)} = (X'X + kI)^{-1}X'\vec{Y} \qquad (10.5)$$

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:

$$\begin{aligned}
\hat{\beta}^{(k)} &= (X'X + kI)^{-1}(X'X)(X'X)^{-1}X'\vec{Y} \\
&= (X'X + kI)^{-1}(X'X)\hat{\beta} \\
&= \left[(X'X)^{-1}(X'X + kI)\right]^{-1}\hat{\beta} \\
&= \left[I + k(X'X)^{-1}\right]^{-1}\hat{\beta} \\
&= Z\hat{\beta} \qquad (10.6)
\end{aligned}$$

siendo $Z \stackrel{\text{def}}{=} [I + k(X'X)^{-1}]^{-1}$.

El Teorema 11.1, que muestra la superioridad del estimador *ridge* sobre el MCO para algún valor de k, es consecuencia del Lema 11.1 a continuación.

Lema 10.1 *El error cuadrático medio del estimador ridge de parámetro k viene dado por la expresión*

$$ECM[\hat{\beta}^{(k)}] = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + \sum_{i=1}^{p} \frac{k^2\alpha_i^2}{(\lambda_i + k)^2} \qquad (10.7)$$

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

DEMOSTRACIÓN:

El ECM del estimador ridge que habremos de comparar con (11.4) es:

$$\begin{aligned} ECM[\hat{\beta}^{(k)}] &= E[(\hat{\beta}^{(k)} - \vec{\beta})'(\hat{\beta}^{(k)} - \vec{\beta})] \\ (\text{por (11.6)}) &= E[(Z\hat{\beta} - \vec{\beta})'(Z\hat{\beta} - \vec{\beta})] \\ &= E[(Z\hat{\beta} - Z\vec{\beta} + Z\vec{\beta} - \vec{\beta})'(Z\hat{\beta} - Z\vec{\beta} + Z\vec{\beta} - \vec{\beta})] \\ &= \underbrace{E[(Z\hat{\beta} - Z\vec{\beta})'(Z\hat{\beta} - Z\vec{\beta})]}_{(a)} + \underbrace{(Z\vec{\beta} - \vec{\beta})'(Z\vec{\beta} - \vec{\beta})}_{(b)} \end{aligned} \tag{10.8}$$

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:

$$\begin{aligned} (a) &= E[(\hat{\beta} - \vec{\beta})'Z'Z(\hat{\beta} - \vec{\beta})] \\ &= E[\text{traza}\{(\hat{\beta} - \vec{\beta})'Z'Z(\hat{\beta} - \vec{\beta})\}] \\ &= E[\text{traza}\{(\hat{\beta} - \vec{\beta})(\hat{\beta} - \vec{\beta})'Z'Z\}] \\ &= \text{traza}\{E(\hat{\beta} - \vec{\beta})(\hat{\beta} - \vec{\beta})'Z'Z\} \\ &= \sigma^2\text{traza}\,[(X'X)^{-1}Z'Z] \qquad (10.9) \\ &= \sigma^2\text{traza}\,\left[(X'X)^{-1}\left[I + k(X'X)^{-1}\right]^{-1}\left[I + k(X'X)^{-1}\right]^{-1}\right] \\ &= \sigma^2\text{traza}\,\left[(X'X) + kI + kI + k^2(X'X)^{-1}\right]^{-1} \\ &= \sigma^2\text{traza}\,\left\{\left[(X'X) + 2kI + k^2(X'X)^{-1}\right]^{-1}VV'\right\} \\ &= \sigma^2\text{traza}\,\left[V'[(X'X) + 2kI + k^2(X'X)^{-1}]^{-1}V\right] \qquad (10.10) \\ &= \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i + 2k + \lambda_i^{-1}k^2} \qquad (10.11) \\ &= \sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i + k)^2}. \qquad (10.12) \end{aligned}$$

La obtención de la expresión (11.9) 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 3.2, pág. 21). En el paso de (11.10) a (11.11) 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 (11.8),

$$\begin{aligned}
(b) &= (Z\vec{\beta} - \vec{\beta})'(Z\vec{\beta} - \vec{\beta}) \\
&= \vec{\beta}'(Z - I)'(Z - I)\vec{\beta} \\
&= \vec{\beta}' \left(\left[I + k(X'X)^{-1} \right]^{-1} - I \right)' \left(\left[I + k(X'X)^{-1} \right]^{-1} - I \right) \vec{\beta} \\
&= k^2 \vec{\alpha}'(\Lambda + kI)^{-2} \vec{\alpha} \qquad (10.13) \\
&= \text{traza } \left[k^2 \vec{\alpha}'(\Lambda + kI)^{-2} \vec{\alpha} \right] \\
&= \sum_{i=1}^{p} \frac{k^2 \alpha_i^2}{(\lambda_i + k)^2} \qquad (10.14)
\end{aligned}$$

El paso a (11.13) desde la expresión anterior hace uso de que $\vec{\alpha} = V'\vec{\beta}$. Sustituyendo (11.12) y (11.14) en (11.8) se obtiene (11.7)

■

El Teorema 11.1 se sigue casi inmediatamente del resultado anterior.

Teorema 10.1 *Hay algún valor de $k > 0$ para el $ECM[\hat{\beta}^{(k)}]$ dado por* (11.7) *es estrictamente menor que el ECM del estimador MCO dado por* (11.4).

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 (11.7) debe coincidir con (11.4), como en efecto puede comprobarse que sucede. Derivando (**??**) 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:

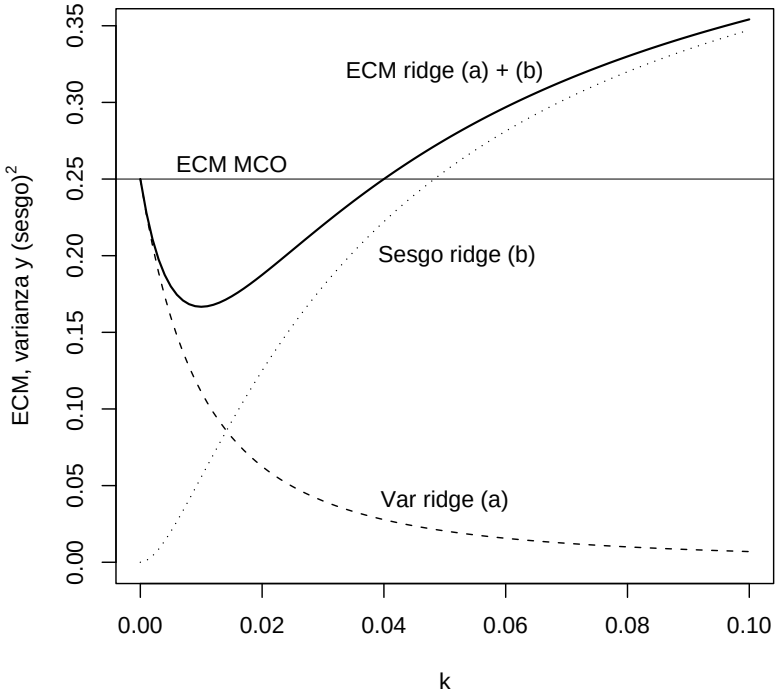
$$ECM[\hat{\beta}^{(k)}] < ECM[\hat{\beta}^{(0)}] = ECM[\hat{\beta}] \qquad (10.15)$$

lo que demuestra el teorema.

■

Una percepción intuitiva del resultado anterior la proporciona la comparación de las expresiones (11.4) y (11.8), 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 (11.4) puede hacerse arbitrariamente grande si $\lambda_i \approx 0$ para algún i. La expresión (11.12) 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 (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 11.1 muestra en un caso concreto cómo varían en función de k los componentes (a) y (b) de (11.8), 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 11.1 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.

10.3.3. Elección de k

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

Uso de trazas ridge. Se prueban diversos valores 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 (11.16) 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 (11.16) 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 (11.17) fue propuesto por Hoerl et al. (1975) y tiene una justificación bayesiana. El criterio (11.18) fue propuesto en Lawless and Wang (1976). El criterio (11.19) estima el ECM del estimador ridge insesgadamente y toma el k que minimiza dicha estimación.

> **Observación 10.2** En las ecuaciones (11.17)–(11.19), 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.

10.3.4. 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 adecuada[2] "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 5.6 y 5.7, pág. 61.

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 11.2 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 11.2; el segundo sí.

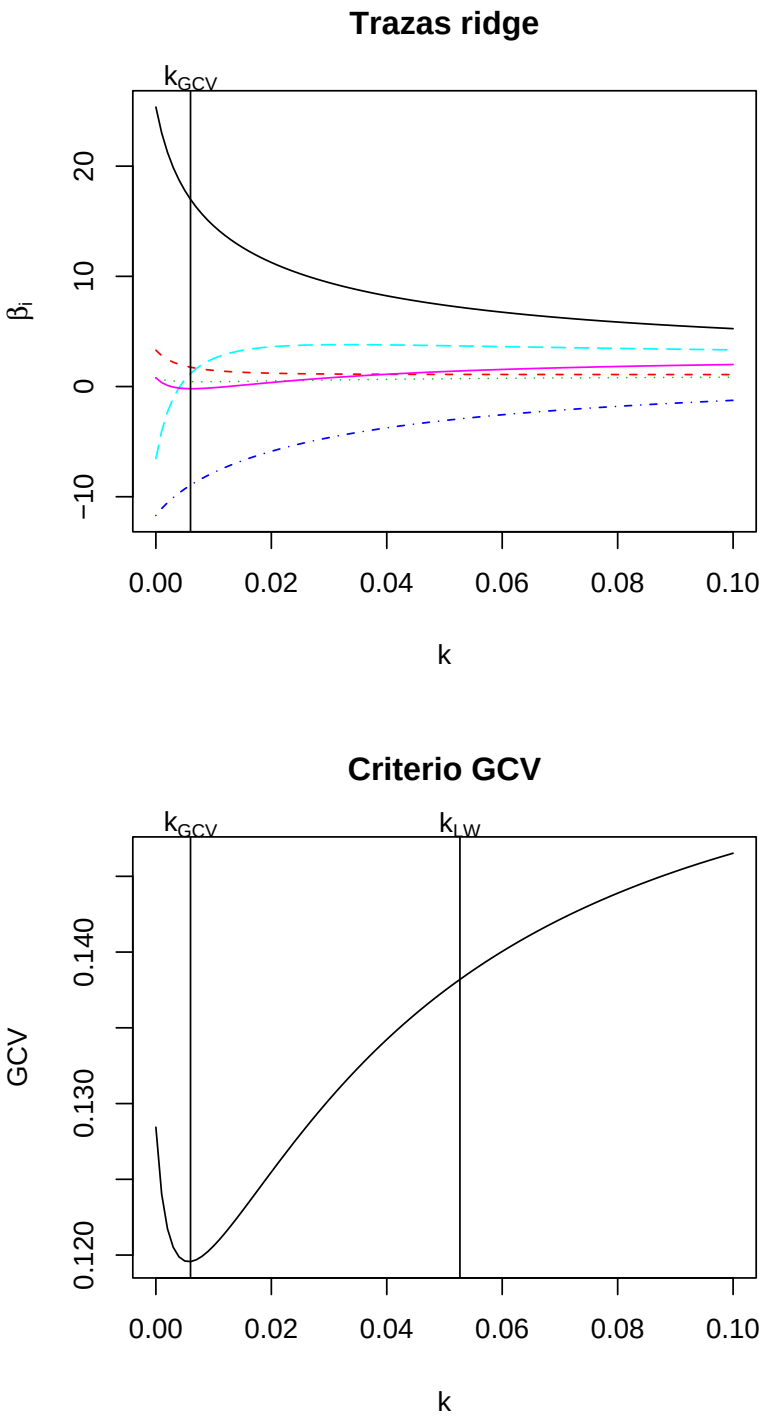
```
> options(digits = 4)
> options(columns = 40)
> library(MASS)
> data(longley)
> names(longley)[1] <- "y"
> longley[1:3, ]
        y   GNP Unemployed Armed.Forces
1947 83.0 234.3      235.6        159.0
1948 88.5 259.4      232.5        145.6
1949 88.2 258.1      368.2        161.6
     Population 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)

Residuals:
```

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

Figura 10.2: Trazas ridge y GVC para los datos `longley`



```
   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 (= `lambda`) 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 mñas 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 11.2.

```
> 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.

10.4.1. Descripción del estimador

Consideraremos, por conveniencia notacional, el modelo habitual en que la columna de "unos", si existe, ha sido segregada, y los restantes regresores han sido centrados y normalizados. Esto tiene por único efecto multiplicar los parámetros —y sus estimadores— por constantes respectivamente iguales a la norma de las columnas de X afectadas. Con este convenio, el modelo de regresion lineal que consideramos se puede escribir así:

$$\vec{y} = \vec{1}\beta_0 + W\vec{\beta}^* + \vec{\epsilon} \tag{10.20}$$

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:

$$V'(W'W)V = \Lambda \qquad (\Longleftrightarrow W'W = V\Lambda V') \tag{10.21}$$

Sean $\vec{v}_1, \ldots, \vec{v}_{p-1}$ los vectores columna de V. Llamaremos *componentes principales* de W a los vectores $\vec{u}_1, \ldots, \vec{u}_{p-1}$ definidos así:

$$\begin{aligned} \vec{u}_1 &= W\vec{v}_1 \\ \vec{u}_2 &= W\vec{v}_2 \\ &\vdots \\ \vec{u}_{p-1} &= W\vec{v}_{p-1} \end{aligned} \tag{10.22}$$

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, (11.20) puede transformarse así:

$$\vec{y} = \vec{1}\beta_0 + W\vec{\beta}^* + \vec{\epsilon} \tag{10.24}$$

$$= \vec{1}\beta_0 + WVV'\vec{\beta}^* + \vec{\epsilon} \tag{10.25}$$

$$= \vec{1}\beta_0 + U\vec{\gamma}^* + \vec{\epsilon} \tag{10.26}$$

Teniendo en cuenta (ver Problema 11.2) que $\vec{1} \perp \vec{u}_i$, $(i = 1, \ldots, p-1)$, el vector de estimadores puede escribirse así:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\gamma}^* \end{pmatrix} = \begin{pmatrix} \overline{y} \\ (U'U)^{-1}U'\vec{y} \end{pmatrix} = \begin{pmatrix} \overline{y} \\ \Lambda^{-1}U'\vec{y} \end{pmatrix} \tag{10.27}$$

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 (11.28).

10.4.2. Estrategias de selección de componentes principales

Hay varias estrategias. Una discusión más pormenorizada que el resumen a continuación puede encontrarse en Brown (1993) o en Jolliffe (1986).

Elección basada en λ_i. Como quiera que la varianza de $\hat{\gamma}_i^*$ es $\sigma^2\lambda_i^{-1}$ (véase (10.26), pág. 135), 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

$$\|P_U\vec{y}\,\|^2 = \|U\hat{\gamma}^*\|^2 = \hat{\gamma}_1^{*2}\|\vec{u}_1\,\|^2 + \cdots + \hat{\gamma}_{p-1}^{*2}\|\vec{u}_{p-1}\,\|^2, \tag{10.29}$$

   la última igualdad haciendo uso de la ortogonalidad entre las columnas de U. Entonces, $SSR = \|P_U\vec{y}\,\|^2$, y $SSE = \|\vec{y} - \vec{\overline{y}}\|^2 - \|U\hat{\gamma}^*\|^2$.

2. Se contrasta la hipótesis de nulidad para cada uno de los parámetros, $(H_i\colon \hat{\gamma}_i^* = 0,\ \ i = 1, \ldots, p-1)$, mediante el estadístico:

$$Q_i = \frac{N-p}{1} \times \frac{\hat{\gamma}_i^{*2}\|\vec{u}_i\,\|^2}{SSE} \sim \mathcal{F}_{1,N-p} \tag{10.30}$$

   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 (11.28).

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 análisis[3].

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).

10.4.3. Propiedades del estimador en componentes principales

El sesgo de $\hat{\beta}^*_{CP}$ es:

$$E[\hat{\beta}^*_{CP} - \vec{\beta}^*] = E\left[V\begin{pmatrix}\hat{\gamma}^*_{(q)}\\ \vec{0}\end{pmatrix} - V\vec{\gamma}^*\right] = -\sum_{i=q+1}^{p-1}\hat{\gamma}^*_i\vec{v}_i \tag{10.31}$$

y su matriz de covarianzas:

$$\begin{aligned}
\Sigma_{\hat{\beta}^*_{CP}} &= V\left(\sigma^2\begin{pmatrix}I_q & 0\\ 0 & 0\end{pmatrix}\Lambda^{-1}\begin{pmatrix}I_q & 0\\ 0 & 0\end{pmatrix}\right)V' && (10.32)\\
&= \sigma^2\sum_{i=1}^{q}\lambda_i^{-1}\vec{v}_i\,\vec{v}_i\,' && (10.33)\\
&\leq \sigma^2\sum_{i=1}^{p-1}\lambda_i^{-1}\vec{v}_i\,\vec{v}_i\,' && (10.34)\\
&= \sigma^2(W'W)^{-1} && (10.35)
\end{aligned}$$

[3]Pero 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:

$$\sigma^2 \sum_{i=q+1}^{p-1} \lambda_i^{-1} \vec{v}_i \, \vec{v}_i{}' \tag{10.36}$$

y será importante si entre las componentes principales excluídas como regresores hay alguna asociada a un λ_i muy pequeño.

Las expresiones (11.31) y (11.32)–(11.35) muestran el conflicto varianza-sesgo en el caso de la regresión en componentes principales. De (11.31) se deduce la siguiente expresión para la suma de los sesgos al cuadrado:

$$[E(\hat{\beta}^*_{CP}) - \vec{\beta}^*]' [E(\hat{\beta}^*_{CP}) - \vec{\beta}^*] = \sum_{i=q+1}^{p-1} (\hat{\gamma}^*_i)^2 \tag{10.37}$$

Es interesante comparar el estimador en componentes principales con el proporcionado por el estimador ridge, y examinarlo a la luz del análisis efectuado en el Capítulo 10. 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:

$$\hat{\beta}^*_{CP} = V \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} \hat{\gamma}^* = V \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} \Lambda^{-1} U' \vec{y} \tag{10.38}$$

con la del estimador ridge equiparable[4]:

$$\begin{aligned} \hat{\beta}^{(k)} &= (W'W + kI)^{-1} W' \vec{y} && (10.39) \\ &= VV'(W'W + kI)^{-1} VV' W' \vec{y} && (10.40) \\ &= V(\Lambda + kI)^{-1} U' \vec{y} && (10.41) \end{aligned}$$

En (11.38) sólo q columnas de $U'\vec{y}$ se utilizan; en (11.41), todas, si bien las que corresponden a componentes principales con λ_i más pequeño reciben una

[4]Es decir, tras haber centrado y normado los regresores y segregado la columna de "unos".

ponderación menor, al ser divididas por λ_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:

$$\vec{y} = \vec{1}\beta_0 + W\vec{\beta}^* + \vec{\epsilon} \tag{10.42}$$

o alternativamente:

$$\vec{y}^* = W\vec{\beta}^* + \vec{\epsilon} \tag{10.43}$$

en que tanto los regresores como la variable respuesta $\vec{y}^*$ han sido normalizados y centrados. Es decir, $\vec{y}^* = \eta^{-1}(\vec{y} - \vec{\overline{y}})$ siendo $\eta^2 = \sum_{i=1}^{N}(y_i - \overline{y})^2$. Si construímos la matriz $N \times p$ siguiente:

$$A = [\vec{y}^* \mid W] \tag{10.44}$$

tenemos que la matriz $(A'A)$ es una matriz de correlación (tiene "unos" en la diagonal principal, es simétrica y semidefinida positiva). Sea $V = (\vec{v}_1 \mid \cdots \mid \vec{v}_p)$ la matriz que la diagonaliza:

$$V'(A'A)V = \Lambda \Longleftrightarrow V\Lambda V' = A'A \tag{10.45}$$

Entonces, utilizando (11.44), tenemos

$$A\vec{v}_j = v_{0j}\vec{y}^* + W\vec{v}_j^{(0)}, \qquad (j = 1, \ldots, p) \tag{10.46}$$

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 (11.46),

$$\begin{aligned} \|A\vec{v}_j\|^2 &= \|v_{0j}\vec{y}_i^* + W\vec{v}_j^{(0)}\|^2 \\ &= \sum_{i=1}^{N}\left(\vec{y}_i^* v_{0j} + \sum_{k=1}^{p-1} W_{ik}v_{kj}\right)^2 \end{aligned} \tag{10.47}$$

en que v_{kj} es la k-ésima coordenada de $\vec{v}_j^{(0)}$. Como por otra parte

$$\begin{aligned} \|A\vec{v}_j\|^2 &= \vec{v}_j\,'(A'A)\vec{v}_j \\ &= \lambda_j, \end{aligned} \tag{10.48}$$

igualando (11.47) y (11.48) deducimos que si $\lambda_j \approx 0$

$$y_i^* v_{0j} \approx -\sum_{k=1}^{p-1} W_{ik} v_{kj} \qquad \forall i \in [1, \ldots, N] \tag{10.49}$$

Si, además, $v_{0j} \neq 0$, podemos escribir:

$$\vec{y}^* \approx -v_{0j}^{-1} W \vec{v}_j^{(0)} \quad \stackrel{\text{def}}{=} \quad \hat{y}^*_{(j)} \tag{10.50}$$

Como $\vec{y}^* = \eta^{-1}(\vec{y} - \vec{\overline{y}})$, $\vec{y} = \vec{\overline{y}} + \eta \vec{y}^*$ y denominando

$$\hat{y}_{(j)} = \vec{\overline{y}} + \eta \hat{y}^*_{(j)} \tag{10.51}$$

tenemos:

$$\begin{aligned}
(\vec{y} - \hat{y}_{(j)})'(\vec{y} - \hat{y}_{(j)}) &= \eta^2 (\vec{y}^* - \hat{y}^*_{(j)})'(\vec{y}^* - \hat{y}^*_{(j)}) \\
&= (v_{0j}\vec{y}^* - v_{0j}\hat{y}^*_{(j)})'(v_{0j}\vec{y}^* - v_{0j}\hat{y}^*_{(j)}) \frac{\eta^2}{v_{0j}^2} \\
&= (A\vec{v}_j)'(A\vec{v}_j)\frac{\eta^2}{v_{0j}^2} \\
&= \frac{\lambda_j \eta^2}{v_{0j}^2}
\end{aligned} \tag{10.52}$$

Nótese que la aproximación de $\vec{y}^*$ en (11.50) y suma de cuadrados de los residuos en (11.52), 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 $\hat{y}_{(i)} \quad (i = 1, \ldots, p)$, debidamente ponderadas por coeficientes d_i a determinar:

$$\begin{aligned}
\hat{y} &= \sum_{i=1}^{p} d_i \hat{y}_{(i)} \\
\text{[usando (11.50) y (11.51)]} &= \sum_{i=1}^{p} d_i \left(\vec{\overline{y}} + W(-{v_{0i}}^{-1} \vec{v}_i^{(0)} \eta) \right) \\
&= \left(\sum_{i=1}^{p} d_i \right) \vec{\overline{y}} + W \left(-\sum_{i=1}^{p} d_i {v_{0i}}^{-1} \vec{v}_i^{(0)} \eta \right)
\end{aligned}$$

Por otro lado, de (11.42) tenemos

$$\hat{\beta}_0 \vec{1} + W \hat{\beta}^*$$

que junto con la igualdad precedente proporciona:

$$\hat{\beta}_0 = \overline{y}\left(\sum_{i=1}^{p} d_i\right) \tag{10.53}$$

$$\hat{\beta}^* = -\eta \sum_{i=1}^{p} d_i {v_{0i}}^{-1} \vec{v}_i^{(0)} \tag{10.54}$$

Como los regresores W están centrados, es claro que $\hat{\beta}_0 = \overline{y}$, y por tanto de (11.53) se deduce $\sum_{i=1}^{p} d_i = 1$. Haciendo uso de (11.52), (11.53), y (11.54) obtenemos la suma de cuadrados de los residuos:

$$\begin{aligned}
(\vec{y}-\hat{y})'(\vec{y}-\hat{y}) &= \eta^2(\vec{y}^*-\hat{y}^*)'(\vec{y}^*-\hat{y}^*) \\
&= \eta^2\left(\vec{y}^* + W\sum_{i=1}^{p} d_i {v_{0i}}^{-1}\vec{v}_i^{(0)}\right)'\left(\vec{y}^* + W\sum_{i=1}^{p} d_i {v_{0i}}^{-1}\vec{v}_i^{(0)}\right) \\
&= \eta^2\left[\sum_{i=1}^{p}\left(\frac{d_i}{v_{0i}}\right)(\vec{y}^* v_{0i} + W\vec{v}_i^{(0)})\right]' \\
&\quad \times\left[\sum_{i=1}^{p}\left(\frac{d_i}{v_{0i}}\right)(\vec{y}^* v_{0i} + W\vec{v}_i^{(0)})\right] \\
&= \eta^2\left[\sum_{i=1}^{p}\left(\frac{d_i}{v_{0i}}\right)A\vec{v}_i\right]'\left[\sum_{i=1}^{p}\left(\frac{d_i}{v_{0i}}\right)A\vec{v}_i\right] \\
&= \eta^2\sum_{i=1}^{p}\left(\frac{\lambda_i d_i^2}{{v_{0i}}^2}\right). \qquad (10.55)
\end{aligned}$$

Podemos ahora minimizar la expresión (11.55) sujeta a que $\sum_{i=1}^{p} d_i = 1$. El lagrangiano es:

$$\Phi(\vec{d}) = \eta^2\sum_{i=1}^{p}\left(\frac{\lambda_i d_i^2}{{v_{0i}}^2}\right) - \mu\left(\sum_{i=1}^{p} d_i - 1\right) \tag{10.56}$$

cuyas derivadas

$$\frac{\partial\Phi(\vec{d})}{\partial d_i} = 2\eta^2\left(\frac{d_i\lambda_i}{{v_{0i}}^2}\right) - \mu = 0 \qquad (i = 1,\ldots,p) \tag{10.57}$$

permiten (multiplicando cada igualdad en (11.57) por ${v_{0i}}^2\lambda_i^{-1}$ y sumando) obtener:

$$\mu = 2\eta^2\left(\sum_{i=1}^{p}\frac{v_{0i}^2}{\lambda_i}\right)^{-1} \tag{10.58}$$

Llevando (11.58) a (11.57) obtenemos:

$$2\eta^2 d_i \frac{\lambda_i}{v_{0i}^2} = \mu = 2\eta^2 \left(\sum_{i=1}^{p} \frac{v_{0i}^2}{\lambda_i} \right)^{-1} \tag{10.59}$$

y por tanto:

$$d_i = \frac{v_{0i}^2}{\lambda_i} \left(\sum_{i=1}^{p} \frac{v_{0i}^2}{\lambda_i} \right)^{-1} \tag{10.60}$$

Los estimadores deseados se obtienen llevando (11.60) a (11.53)–(11.54):

$$\hat{\beta}_0 = \overline{y} \tag{10.61}$$

$$\hat{\beta}^* = -\eta \frac{\sum_{i=1}^{p} \left(\frac{v_{0i}}{\lambda_i} \right) \vec{v}_i^{(0)}}{\sum_{i=1}^{p} \frac{v_{0i}^2}{\lambda_i}} \tag{10.62}$$

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 (11.49) es más ajustada). Pero podemos eliminar en (11.62) términos muy inestables, cuando v_{0i} y λ_i son ambos muy pequeños, para evitar que el sumando correspondiente en (11.62) reciba gran ponderación, si parece evidente que se trata de una multicolinealidad no predictiva. La relación (11.62) se transformará entonces en:

$$\hat{\beta}^* = -\eta \frac{\sum_{i \in P} \left(\frac{v_{0i}}{\lambda_i} \right) \vec{v}_i^{(0)}}{\sum_{i \in P} \left(\frac{v_{0i}^2}{\lambda_i} \right)} \tag{10.63}$$

siendo P un subconjunto de $(1, \ldots, p)$.

La determinación de P es una tarea eminentemente subjetiva; se suele desechar una multicolinealidad cuando $\lambda_i < 0{,}10$ y $v_{0i} < 0{,}10$, si además $\vec{v}_i^{(0)}$ "se aproxima" a un vector propio de $W'W$.

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 13 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 11.3 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 (11.22), se verifica que $\vec{1} \perp \vec{u}_i$.

10.3 Sea una muestra formada por n observaciones, $X_1, \ldots, X_n$, generadas por una distribución con media. Demuéstrese que, para algún c, $c\overline{X}$ es mejor estimador (en terminos de error medio cuadrático, ECM) que $\overline{X}$. ¿Es esto un caso particular de alguno de los procedimientos de estimación examinados en este capítulo?

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:

$$\begin{aligned}
\hat{\beta} &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} && (10.64)\\
&= (X'X + kI)^{-1}(X'\vec{y} + \sqrt{k} I \vec{0}) && (10.65)\\
&= (X'X + kI)^{-1} X'\vec{y} && (10.66)\\
&= \hat{\beta}^{(k)} && (10.67)
\end{aligned}$$

Alternativamente, se puede formar $\tilde{X}$ añadiendo a X las filas de una matriz unidad, y realizar regresión ponderada (dando a cada observación "normal" peso unitario y a las p seudo-observaciones añadidas peso $\sqrt{k}$). La alteración de los pesos es habitualmente más cómoda que la creación de una nueva matriz de regresores. Este será de ordinario el método a utilizar cuando hayamos de probar muchos valores diferentes de k y dispongamos de un programa para hacer regresión mínimo cuadrática ponderada. Las funciones `lsfit` y `lm` (disponibles en R) 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), $i = 1, \ldots, N$. La variable Y es peso, la variable X es edad,

y las observaciones corresponden a N diferentes sujetos. Estamos interesados en especificar la evolución del peso con la edad. Podríamos construir la matrix de diseño

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & \dots & x_1^{p-1} \\ 1 & x_2 & x_2^2 & x_2^3 & \dots & x_2^{p-1} \\ 1 & x_3 & x_3^2 & x_3^3 & \dots & x_3^{p-1} \\ \vdots & & & \vdots & & \vdots \\ 1 & x_N & x_N^2 & x_N^3 & \dots & x_N^{p-1} \end{pmatrix} \tag{10.68}$$

y contrastar hipótesis tales como $H_0 : \beta_2 = \beta_3 = \ldots = \beta_{p-1} = 0$ (tendencia no más que lineal), $H_0 : \beta_3 = \ldots = \beta_{p-1} = 0$ (tendencia no más que cuadrática), etc. Sucede sin embargo, como es fácil comprobar, que una matriz como la anterior adolece de una acusada multicolinealidad, sean cuales fueren los valores $x_1, \ldots, x_N$.

Podríamos ortogonalizar los vectores columna de la matriz de diseño (por ejemplo mediante el procedimiendo de Gram-Schmidt: véase Grafe (1985) o cualquier libro de Algebra Lineal), para obtener una nueva matriz de diseño. Los nuevos vectores columna generan el mismo espacio y el contraste puede hacerse del mismo modo que con los originales, pero sin problemas de multicolinealidad.

Otra posibilidad es sustituir las potencias creciente de x_i en las columnas de X por polinomios ortogonales evaluados para los mismos valores x_i (ver por ejemplo Seber (1977), Dahlquist and Björck (1974), o cualquier texto de Análisis Numérico).

Ambos procedimientos tienen por finalidad encontrar una base ortogonal o aproximadamente ortogonal generando el mismo espacio que los vectores columna originales de la matriz de diseño.

10.6 (↑ 11.5) ¿Por qué, para la finalidad perseguida en el Ejercicio 11.5, no sería de utilidad hacer regresión en componentes principales?