MAT-266: Colinealidad

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Colinealidad

Considere el modelo

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

donde $\mathsf{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ y $\mathsf{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$ con $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ tal que $\operatorname{rg}(\boldsymbol{X}) = p$.

Es bien conocido que cuando $\boldsymbol{X}$ es mal condicionada, el sistema de ecuaciones

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top\boldsymbol{Y},$$

puede ser muy inestable.

Observación:

Es decir, aunque $\operatorname{rg}(\boldsymbol{X}) = p$, tenemos que existe $\boldsymbol{a}$ tal que $\boldsymbol{X}\boldsymbol{a} \approx \mathbf{0}$.

Observación:

Este es un problema numérico que puede tener consecuencias inferenciales importantes, por ejemplo:

- Típicamente los coeficientes estimados $\widehat{\boldsymbol{\beta}}$ tendrán varianzas "grandes".
- Test estadísticos presentarán bajo poder y los intervalos de confianza serán muy amplios.
- Signos de algunos coeficientes son "incorrectos" (basados en conocimiento previo).
- Resultados cambian bruscamente con la eliminación de una columna de $\boldsymbol{X}$.

Colinealidad

Algunas herramientas para el diagnóstico de colinealidad, son:

(a) Examinar la matriz de correlación entre los regresores y la respuesta, esto es:

$$\begin{pmatrix} \boldsymbol{R}_{XX} & \boldsymbol{R}_{XY} \\ & 1 \end{pmatrix},$$

correlaciones altas entre dos variables pueden indicar un posible problema de colinealidad.

(b) Factores de inflación de varianza: Suponga que los datos han sido centrados y escalados, entonces

$$\boldsymbol{R}^{-1} = (\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1}, \qquad \widetilde{\boldsymbol{X}} = (x_{ij} - \overline{x}_j),$$

y los elementos diagonales de $\boldsymbol{R}^{-1}$ son llamados factores de inflación de varianza VIF_j, se puede mostrar que

$$\mathsf{VIF}_j = \frac{1}{1 - R_j^2},$$

donde R_j^2 es el coeficiente de correlación multiple de $\boldsymbol{X}_j$ "regresado" sobre el resto de variables explicativas y de ahí que un VIF_j "alto" indica R_j^2 cercano a 1 y por tanto presencia de colinealidad.

(c) Examinar los valores/vectores propios (o componentes principales) de la matriz de correlación $\boldsymbol{R}$.

(d) Número condición: Desde la SVD de $\boldsymbol{X}$ podemos escribir

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top,$$

donde $\boldsymbol{U} \in \mathbb{R}^{n \times p}$, $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_p$, $\boldsymbol{D} = \operatorname{diag}(\delta_1, \dots, \delta_p)$ y $\boldsymbol{V} \in \mathcal{O}_p$.

La detección de colinealidad puede ser llevada a cabo usando

$$\kappa(\boldsymbol{X}) = \|\boldsymbol{X}\| \|\boldsymbol{X}^+\| = \frac{\delta_1}{\delta_p},$$

y $\kappa(\boldsymbol{X})$ "grande" (> 30) es un indicador de colinealidad.

Note que, el caso de deficiencia de rango puede ser manipulado sin problemas usando SVD. En efecto,

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \boldsymbol{U}\begin{pmatrix} \boldsymbol{D}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\boldsymbol{V}^\top$$

donde $\boldsymbol{D}_1 \in \mathbb{R}^{r\times r}$, $\operatorname{rg}(\boldsymbol{X}) = r < p$. De este modo

$$\boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\begin{pmatrix} \boldsymbol{D}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \quad \Rightarrow \quad \boldsymbol{X}(\boldsymbol{V}_1, \boldsymbol{V}_2) = (\boldsymbol{U}_1, \boldsymbol{U}_2)\begin{pmatrix} \boldsymbol{D}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix},$$

desde donde sigue que

$$\boldsymbol{X}\boldsymbol{V}_1 = \boldsymbol{U}_1\boldsymbol{D}_1, \qquad \boldsymbol{X}\boldsymbol{V}_2 = \boldsymbol{0}.$$

Es decir, SVD permite "detectar" la dependencia lineal.

Considere la descomposición espectral de $\boldsymbol{X}^\top\boldsymbol{X}$, dada como

$$\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top = (\boldsymbol{U}_1, \boldsymbol{U}_2)\begin{pmatrix} \boldsymbol{\Lambda}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda}_2 \end{pmatrix}\begin{pmatrix} \boldsymbol{U}_1^\top \\ \boldsymbol{U}_2^\top \end{pmatrix},$$

donde $\boldsymbol{\Lambda}_1 = \operatorname{diag}(\lambda_1, \dots, \lambda_r)$ y $\boldsymbol{\Lambda}_2 = \operatorname{diag}(\lambda_{r+1}, \dots, \lambda_p)$, mientras que $\boldsymbol{U} = (\boldsymbol{U}_1, \boldsymbol{U}_2)$ es matriz ortogonal.

Resultado 1 (Estimador componentes principales):

Bajo los supuestos del modelo lineal en A1-A4*, el estimador componentes principales para $\boldsymbol{\beta}$ puede ser escrito como

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\mathsf{PC}} &= \boldsymbol{U}_1(\boldsymbol{U}_1^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{U}_1)^{-1}\boldsymbol{U}_1^\top\boldsymbol{X}^\top\boldsymbol{Y} \\ &= \boldsymbol{U}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{U}_1^\top\boldsymbol{X}^\top\boldsymbol{Y} \end{aligned}$$

Demostración:

Por la ortogonalidad de $\boldsymbol{U} = (\boldsymbol{U}_1, \boldsymbol{U}_2)$, sigue que

$$\boldsymbol{U}_1^\top \boldsymbol{U}_1 = \boldsymbol{I}_r, \qquad \boldsymbol{U}_2^\top \boldsymbol{U}_2 = \boldsymbol{I}_{p-r}, \qquad \boldsymbol{U}_1 \boldsymbol{U}_1^\top + \boldsymbol{U}_2 \boldsymbol{U}_2^\top = \boldsymbol{I}_p,$$

y $\boldsymbol{U}_1^\top \boldsymbol{U}_2 = \boldsymbol{0}$. Ahora,

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U} = \boldsymbol{U}_1 \boldsymbol{\Lambda}_1^{-1} \boldsymbol{U}_1^\top + \boldsymbol{U}_2 \boldsymbol{\Lambda}_2^{-1} \boldsymbol{U}_2^\top.$$

Usando que $\boldsymbol{U}_1^\top \boldsymbol{U}_2 = \boldsymbol{0}$ $(= \boldsymbol{U}_2^\top \boldsymbol{U}_1)$, sigue

$$\boldsymbol{U}_2^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{U}_2 = \boldsymbol{U}_2^\top (\boldsymbol{U}_1 \boldsymbol{\Lambda}_1^{-1} \boldsymbol{U}_1^\top + \boldsymbol{U}_2 \boldsymbol{\Lambda}_2^{-1} \boldsymbol{U}_2^\top) \boldsymbol{U}_2 = \boldsymbol{\Delta}_2^{-1}$$

De este modo, $[\boldsymbol{U}_2^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{U}_2]^{-1} = \boldsymbol{\Lambda}_2$, lo que permite escribir

$$\begin{aligned}
&(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{U}_2 [\boldsymbol{U}_2^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{U}_2]^{-1} \boldsymbol{U}_2^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&\qquad = (\boldsymbol{U}_1 \boldsymbol{\Lambda}_1^{-1} \boldsymbol{U}_1^\top + \boldsymbol{U}_2 \boldsymbol{\Lambda}_2^{-1} \boldsymbol{U}_2^\top) \boldsymbol{U}_2 \boldsymbol{\Lambda}_2 \boldsymbol{U}_2^\top (\boldsymbol{U}_1 \boldsymbol{\Lambda}_1^{-1} \boldsymbol{U}_1^\top + \boldsymbol{U}_2 \boldsymbol{\Lambda}_2^{-1} \boldsymbol{U}_2^\top) \\
&\qquad = \boldsymbol{U}_2 \boldsymbol{\Lambda}_2^{-1} \boldsymbol{U}_2^\top.
\end{aligned}$$

Es decir,

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{U}_2[\boldsymbol{U}_2^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{U}_2]^{-1}\boldsymbol{U}_2^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \boldsymbol{U}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{U}_1^\top.$$

Como $\boldsymbol{U}_1^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{U}_1 = \boldsymbol{\Lambda}_1$. Obtenemos

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\mathsf{PC}} &= [(\boldsymbol{X}^\top\boldsymbol{X})^{-1} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{U}_2[\boldsymbol{U}_2^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{U}_2]^{-1}\boldsymbol{U}_2^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}]\boldsymbol{X}^\top\boldsymbol{Y} \\ &= \boldsymbol{U}_1(\boldsymbol{U}_1^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{U}_1)^{-1}\boldsymbol{U}_1^\top\boldsymbol{X}^\top\boldsymbol{Y},\end{aligned}$$

lo que concluye la prueba.

Observación:

- Es posible notar que el estimador PC es un caso particular del estimador restringido con respecto a:
$$\boldsymbol{U}_2^\top \boldsymbol{\beta} = \mathbf{0}.$$

- $\widehat{\boldsymbol{\beta}}_{\mathsf{PC}}$ depende del 'parámetro' r. En efecto,
$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \\ &= (\boldsymbol{U}_1 \boldsymbol{\Lambda}_1^{-1} \boldsymbol{U}_1^\top + \boldsymbol{U}_2 \boldsymbol{\Lambda}_2^{-1} \boldsymbol{U}_2^\top) \boldsymbol{X}^\top \boldsymbol{Y}.\end{aligned}$$

  De este modo podemo interpretar $\widehat{\boldsymbol{\beta}}_{\mathsf{PC}}$ como una modificación del OLS que desconsidera $\boldsymbol{U}_2 \boldsymbol{\Lambda}_2^{-1} \boldsymbol{U}_2^\top$.

Una alternativa para seleccionar r, es utilizar el test F. Suponga r fijo y considere $H_0 : \boldsymbol{U}_2^\top \boldsymbol{\beta} = \boldsymbol{0}$. Tenemos el estadístico

$$F = \Big(\frac{n-p}{p-r}\Big) \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{\mathsf{PC}})^\top \boldsymbol{X}^\top \boldsymbol{X} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{\mathsf{PC}})}{\boldsymbol{Y}^\top (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{Y}}.$$

Si para un nivel α tenemos

$$F \geq F_{1-\alpha}(p-r, n-p).$$

Entonces, rechazamos H_0 y podemos seleccionar r un poco más pequeño.

Observación:

- No hay manera de verificar si las restricciones son satisfechas y en efecto este estimador es sesgado.
- Deseamos escoger r tan pequeño como posible para solucionar el problema de colinealidad y tan grande para no introducir mucho sesgo.

Hoerl y Kennard (1970)[1] propusieron usar el estimador ridge

$$\widehat{\boldsymbol{\beta}}_k = (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}, \qquad k \geq 0$$

donde k es conocido como parámetro ridge.

Note que

$$\widehat{\boldsymbol{\beta}}_k = (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

De este modo,

$$\mathsf{E}(\widehat{\boldsymbol{\beta}}_k) = (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta},$$

para $k \neq 0$, tenemos que $\widehat{\boldsymbol{\beta}}_k$ es sesgado.

[1] Technometrics **12**, 55-67.

Mientras que el error cuadrático medio de $\widehat{\boldsymbol{\beta}}_k$ es dado por:

$$\mathsf{MSE} = \mathsf{E}\{\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}\|^2\} = \operatorname{tr}\mathsf{Cov}(\widehat{\boldsymbol{\beta}}_k) + \|\,\mathsf{E}(\widehat{\boldsymbol{\beta}}_k) - \boldsymbol{\beta}\|^2.$$

En efecto,

$$\begin{aligned}\mathsf{Cov}(\widehat{\boldsymbol{\beta}}_k) &= (\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\,\mathsf{Cov}(\widehat{\boldsymbol{\beta}})\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-1}\\ &= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-1}.\end{aligned}$$

Además,

$$\begin{aligned}\mathsf{bias}(\widehat{\boldsymbol{\beta}}_k, \widehat{\boldsymbol{\beta}}) &= \mathsf{E}(\widehat{\boldsymbol{\beta}}_k) - \boldsymbol{\beta} = (\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}\\ &= (\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-1}\big[\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} - (\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})\boldsymbol{\beta}\big]\\ &= -k(\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{\beta}.\end{aligned}$$

Considere la SVD de $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, de este modo $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top$, y podemos escribir

$$\begin{aligned}\operatorname{Cov}(\widehat{\boldsymbol{\beta}}_k) &= \sigma^2\boldsymbol{V}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-1}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top\boldsymbol{V}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-1}\boldsymbol{V}^\top \\ &= \sigma^2\boldsymbol{V}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-2}\boldsymbol{D}^2\boldsymbol{V}^\top.\end{aligned}$$

De este modo,

$$\operatorname{tr}\operatorname{Cov}(\widehat{\boldsymbol{\beta}}_k) = \sigma^2\operatorname{tr}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-2}\boldsymbol{D}^2 = \sigma^2\sum_{i=1}^p \frac{\delta_i^2}{(\delta_i^2 + k)^2},$$

donde $\delta_1, \dots, \delta_p$ son los valores singulares de $\boldsymbol{X}$. Finalmente,

$$\mathsf{MSE} = \sigma^2\sum_{i=1}^p \frac{\delta_i^2}{(\delta_i^2 + k)^2} + k^2\boldsymbol{\beta}^\top(\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-2}\boldsymbol{\beta}.$$

El estimador ridge tiene varias interpretaciones interesantes, por ejemplo:

(a) Es posible caracterizar $\widehat{\boldsymbol{\beta}}_k$ como solución del problema regularizado:

$$\min_{\beta} Q(\boldsymbol{\beta}, k), \qquad Q(\boldsymbol{\beta}, k) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + k\,\|\boldsymbol{\beta}\|^2,$$

que puede ser expresado de forma equivalente como

$$\min_{\beta} Q(\boldsymbol{\beta}), \qquad \text{sujeto a: } \|\boldsymbol{\beta}\|^2 \leq r^2,$$

y en este contexto, k corresponde a un multiplicador de Lagrange.

Observación:

Este tipo de regularización es conocida como regularización de Tikhonov.[2]

[2] Razón por la que k en ocasiones es llamado parámetro de regularización.

(b) Considere el modelo de regresión con datos aumentados:

$$\boldsymbol{Y}_a = \boldsymbol{X}_a\boldsymbol{\beta} + \boldsymbol{\epsilon}_a, \qquad \boldsymbol{\epsilon}_a \sim \mathsf{N}_{n+p}(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

donde

$$\boldsymbol{Y}_a = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{0} \end{pmatrix}, \qquad \boldsymbol{X}_a = \begin{pmatrix} \boldsymbol{X} \\ \sqrt{k}\boldsymbol{I}_p \end{pmatrix}, \qquad \boldsymbol{\epsilon}_a = \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{u} \end{pmatrix}.$$

El interés recae en escoger algún $k \geq 0$ tal que la matriz de diseño $\boldsymbol{X}_a$ tenga número condición $\kappa(\boldsymbol{X}_a)$ acotado.

Resultado 2:

Suponga que los supuestos del modelo lineal en A1-A4*, son satisfechos. Entonces,

$$\|\widehat{\boldsymbol{\beta}}_{k_2}\|^2 < \|\widehat{\boldsymbol{\beta}}_{k_1}\|^2,$$

siempre que $0 \leq k_1 < k_2$.

Demostración:

Tenemos $\widehat{\boldsymbol{\beta}}_k = (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. De este modo,

$$\|\widehat{\boldsymbol{\beta}}_k\|^2 = \widehat{\boldsymbol{\beta}}^\top \boldsymbol{M}_k \widehat{\boldsymbol{\beta}}, \qquad \boldsymbol{M}_k = \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-2}\boldsymbol{X}^\top \boldsymbol{X}.$$

Basado en la SVD de $\boldsymbol{X}$, tenemos

$$\begin{aligned}\boldsymbol{M}_k &= \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top + k\boldsymbol{V}\boldsymbol{V}^\top)^{-2}\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top \\ &= \boldsymbol{V}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-2}\boldsymbol{D}^4\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{\Gamma}_k\boldsymbol{V}^\top,\end{aligned}$$

con

$$\boldsymbol{\Gamma}_k = \operatorname{diag}\Big(\frac{\delta_1^4}{(\delta_1^2 + k)^2}, \dots, \frac{\delta_p^4}{(\delta_p^2 + k)^2}\Big).$$

De ahí que, si $0 \leq k_1 < k_2$, entonces

$$\boldsymbol{M}_{k_1} - \boldsymbol{M}_{k_2} \geq \boldsymbol{0},$$

lo que lleva a $\widehat{\boldsymbol{\beta}}^\top \boldsymbol{M}_{k_2}\widehat{\boldsymbol{\beta}} < \widehat{\boldsymbol{\beta}}^\top \boldsymbol{M}_{k_1}\widehat{\boldsymbol{\beta}}$, siempre que $\widehat{\boldsymbol{\beta}} \neq \boldsymbol{0}$.

Observación:

Note que $\lim_{k\to\infty} \|\widehat{\boldsymbol{\beta}}_k\|^2 = 0$ y de ahí que

$$\lim_{k\to\infty} \widehat{\boldsymbol{\beta}}_k = \mathbf{0}. \qquad (1)$$

Dado que $\widehat{\boldsymbol{\beta}}_k = \boldsymbol{W}_k\widehat{\boldsymbol{\beta}}$ con $\boldsymbol{W}_k = (\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{X}$. La propiedad en (1) ha llevado a que el estimador ridge sea considerado como un estimador shrinkage, en cuyo caso

$$\boldsymbol{W}_k = (\boldsymbol{I}_p + k(\boldsymbol{X}^\top\boldsymbol{X})^{-1})^{-1}, \qquad k \geq 0,$$

es llamada matrix ridge-shrinking.

Se ha propuesto diversos estimadores de k, lo que buscan seleccionar un $\widehat{\boldsymbol{\beta}}_{\text{opt}}$ que reduzca su MSE. Algunas de estas alternativas son:

(a) Hoerl, Kennard y Baldwin (1975):[3]

$$\widehat{k}_{\mathsf{HKB}} = \frac{ps^2}{\|\widehat{\boldsymbol{\beta}}\|^2}.$$

(b) Lawless y Wang (1976):[4]

$$\widehat{k}_{\mathsf{LW}} = \frac{ps^2}{\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}}}.$$

(c) Lindley y Smith (1972):[5]

$$\widehat{k}_{\mathsf{LS}} = \frac{(n-p)(p+2)}{(n+2)}\,\frac{s^2}{\|\widehat{\boldsymbol{\beta}}\|^2}.$$

[3] Communications in Statistics: Theory and Methods **4**, 105-123.
[4] Communications in Statistics: Theory and Methods **5**, 307-323.
[5] Journal of the Royal Statistical Society, Series B **34**, 1-41.

Golub, Heath y Wahba (1979)[6] han sugerido seleccionar el parámetro ridge usando validación cruzada generalizada (GCV), la que minimiza el criterio

$$V(k) = \frac{1}{n} \frac{\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_k)^2}{\{1 - \operatorname{tr}(\boldsymbol{H}(k))/n\}^2},$$

donde

$$\boldsymbol{H}(k) = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top.$$

Es facil notar que $\widehat{Y}_k = \boldsymbol{H}(k)\boldsymbol{Y}$. En este contexto se ha definido

$$\text{edf} = \operatorname{tr} \boldsymbol{H}(k),$$

como el número de parámetros efectivos. En efecto, para $k = 0$, sigue que edf $= p$.

[6] Technometrics **21**, 215-223.

Considere la SVD de $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ y escriba el modelo en su forma canónica:

$$\boldsymbol{Z} = \boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{u}, \qquad \boldsymbol{u} = \boldsymbol{U}^\top\boldsymbol{\epsilon} \sim \mathsf{N}_p(\boldsymbol{0}, \sigma^2\boldsymbol{I}),$$

donde $\boldsymbol{Z} = \boldsymbol{U}^\top\boldsymbol{Y}$, $\boldsymbol{\alpha} = \boldsymbol{V}^\top\boldsymbol{\beta}$. De este modo, es fácil notar que

$$\widehat{\boldsymbol{\alpha}}_k = (\boldsymbol{D}^2 + k\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{Z},$$

es decir,

$$\widehat{\alpha}_{k,j} = \frac{\delta_j z_j}{\delta_j^2 + k}, \qquad j = 1, \dots, p.$$

Por otro lado,

$$\begin{aligned}
\text{edf} = \operatorname{tr}\boldsymbol{H}(k) &= \operatorname{tr}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top + k\boldsymbol{I})^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top \\
&= \operatorname{tr}\boldsymbol{U}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-1}\boldsymbol{D}^2\boldsymbol{U}^\top = \operatorname{tr}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-1}\boldsymbol{D}^2 \\
&= \sum_{j=1}^p \frac{\delta_j^2}{\delta_j^2 + k}.
\end{aligned}$$

Además,

$$\widehat{\boldsymbol{Y}}_k = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_k = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top\widehat{\boldsymbol{\beta}}_k = \boldsymbol{U}\boldsymbol{D}\widehat{\boldsymbol{\alpha}}_k.$$

Lo que permite escribir

$$V(k) = \frac{1}{n}\frac{\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_k\|^2}{\{\operatorname{tr}(\boldsymbol{I} - \boldsymbol{H}(k))/n\}^2} = \frac{\|\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{D}\widehat{\boldsymbol{\alpha}}_k\|^2}{(1 - \mathsf{edf}\,/n)^2},$$

Observación:

Las consideraciones anteriores ofrecen un procedimiento sencillo para evaluar $V(k)$.

Además, en términos del modelo canónico, tenemos:

$$\widehat{k}_{\mathsf{HKB}} = \frac{ps^2}{\|\widehat{\boldsymbol{\alpha}}\|^2}, \qquad \widehat{k}_{\mathsf{LW}} = \frac{ps^2}{\|\boldsymbol{Z}\|^2},$$

con $\widehat{\boldsymbol{\alpha}} = \boldsymbol{D}^{-1}\boldsymbol{Z}$ $(= \widehat{\boldsymbol{\alpha}}_0)$, $\boldsymbol{Z} = \boldsymbol{U}^\top\boldsymbol{Y}$ y $s^2 = \|\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{D}\widehat{\boldsymbol{\alpha}}\|^2/(n-p)$.

Cemento Portland (Woods, Steinour y Starke, 1932)[7]

Ejemplo (Datos de cemento Portland):

Estudio experimental relacionando la emisión de calor durante la producción y endurecimiento de 13 muestras de cementos Portland. Woods, Steinour y Starke (1932) consideraron cuatro compuestos para los clinkers desde los que se produce el cemento.

La respuesta (Y) es la emisión de calor después de 180 días de curado, medido en calorías por gramo de cemento. Los regresores son los porcentajes de los cuatro compuestos principales: aluminato tricálcico (X_1), silicato tricálcico (X_2), ferrito aluminato tetracálcico (X_3) y silicato dicálcico (X_4).

[7] Industrial and Engineering Chemistry **24**, 1207-1214.

Cemento Portland (Woods, Steinour y Starke, 1932)

Siguiendo a Woods, Steinour y Starke (1932) consideramos un modelo lineal sin intercepto (modelo homogéneo), cuyo número condición escalado es $\kappa(\boldsymbol{X}) = 9.432$, esto es, $\boldsymbol{X}$ es bien condicionada (variables centradas $\kappa(\widetilde{\boldsymbol{X}}) = 37.106$).

Por otro lado, Hald (1952),[8] Gorman y Toman (1966)[9] y Daniel y Wood (1980)[10] adoptan un modelo con intercepto (modelo no homogéneo). En cuyo caso $\kappa(\boldsymbol{X}) = 249.578$, sugiriendo la presencia de colinealidad. El aumento en el número condición se debe a que existe una relación lineal aproximada, pues

$$x_1 + x_2 + x_3 + x_4 \approx 100,$$

de modo que incluir el intercepto causa una colinealidad severa.

[8] Statistical Theory with Engineering Application, Wiley.

[9] Technometrics **8**, 27-51.

[10] Fitting Equations to Data: Computer analysis of multifactor data, Wiley.

R script para el cálculo del número condición (escalado)[11]

```
scaled.condition <- function(x)
{ # scaled condition number
  colScales <- apply(x, 2, function(x) sum(x^2))
  z <- scale(x, center = FALSE, scale = sqrt(colScales))
  d <- svd(z)$d
  p <- length(d)
  cn <- d[1] / d[p]
  obj <- list(condition = cn, values = d, x.scaled = z)
  obj
}
```

En R debemos hacer:

```
# interpreta el código en el script
> source("scaled.condition.R")
```

[11] Proximamente, versión C disponible en 'fastmatrix'

Cemento Portland (Woods, Steinour y Starke, 1932)

```
> scaled.condition(portland[,-1])
$condition
[1] 9.432457

$values
[1] 1.7672193 0.7439735 0.5369706 0.1873551

$x.scaled
           x1        x2         x3         x4
1  0.20741310 0.1430170 0.12529947 0.48888862
2  0.02963044 0.1595189 0.31324867 0.42370347
3  0.32593487 0.3080366 0.16706596 0.16296287
4  0.32593487 0.1705203 0.16706596 0.38296275
5  0.20741310 0.2860340 0.12529947 0.26888874
6  0.32593487 0.3025359 0.18794920 0.17925916
7  0.08889133 0.3905464 0.35501516 0.04888886
8  0.02963044 0.1705203 0.45943138 0.35851832
9  0.05926089 0.2970353 0.37589840 0.17925916
10 0.62223929 0.2585307 0.08353298 0.21185174
11 0.02963044 0.2200261 0.48031462 0.27703689
12 0.32593487 0.3630431 0.18794920 0.09777772
13 0.29630443 0.3740444 0.16706596 0.09777772
attr(,"scaled:scale")
       x1        x2        x3        x4
 33.74907 181.79659  47.88528 122.72734
```

Cemento Portland (Woods, Steinour y Starke, 1932)

```
# carga biblioteca 'fastmatrix'
# disponible en: https://faosorios.github.io/fastmatrix/
> library(fastmatrix)

# carga base de datos en directorio de trabajo
> load("portland.rda")

# ajuste de modelo homogéneo
> f0 <- ols(y ~ -1 + x1 + x2 + x3 + x4, data = portland)
> f0

Call:
ols(formula = y ~ -1 + x1 + x2 + x3 + x4, data = portland)

Coefficients:
    x1      x2      x3      x4
2.1930  1.1533  0.7585  0.4863

Degrees of freedom: 13 total; 9 residual
Residual standard error: 2.417739
```

Cemento Portland (Woods, Steinour y Starke, 1932)

```
# ajuste de modelo no homogéneo
> f1 <- ols(y ~ x1 + x2 + x3 + x4, data = portland)
> f1

Call:
ols(formula = y ~ x1 + x2 + x3 + x4, data = portland)

Coefficients:
(Intercept)       x1       x2       x3        x4
    62.4054   1.5511   0.5102   0.1019   -0.1441

Degrees of freedom: 13 total; 8 residual
Residual standard error: 2.446008
```

```
# ajuste usando regresión ridge
> z0 <- ridge(y ~ x1 + x2 + x3 + x4, data = portland, lambda = 10,
+              method = "grid")
> z0

Call:
ridge(formula = y ~ x1 + x2 + x3 + x4, data = portland, lambda = 10,
    method = "grid")

Coefficients:
(Intercept)          x1          x2          x3          x4
    0.08568     2.16549     1.15860     0.73845     0.48948

Optimal ridge parameter: 1.9598

Number of observations: 13
Effective number of parameters: 3.9796
Scale parameter estimate: 4.0553
```

```
# explorando 'elementos' del objeto 'ridge'
> attributes(z0)
$names
 [1] "dims"          "coefficients"  "scale"          "fitted.values"
 [5] "residuals"     "RSS"           "edf"            "pen"
 [9] "GCV"           "HKB"           "LW"             "lambda"
[13] "optimal"       "call"          "method"         "xlevels"
[17] "terms"

$class
[1] "ridge"

# extrayendo estimadores HKB, LW, y 'optimal'
> opt <- z0$optimal
> HKB <- z0$HKB
> LW <- z0$LW

> opt
[1] 1.959799
> HKB
[1] 0.007676109
> LW
[1] 0.003212916
```

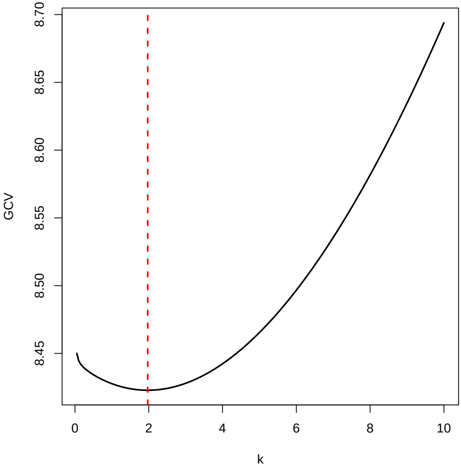
Cemento Portland (Woods, Steinour y Starke, 1932)

```
# estimador ridge usando HKB
> z1 <- ridge(y ~ x1 + x2 + x3 + x4, data = portland, lambda = HKB,
+             method = "none")
> z1

Call:
ridge(formula = y ~ x1 + x2 + x3 + x4, data = portland, lambda = HKB,
    method = "none")

Coefficients:
(Intercept)          x1          x2          x3          x4
     8.5870      2.1046      1.0648      0.6681      0.3996

Ridge parameter: 0.0077

Number of observations: 13
Effective number of parameters: 4.1369
Scale parameter estimate: 4.0005
```

```
# estimador ridge usando LW
> z2 <- ridge(y ~ x1 + x2 + x3 + x4, data = portland, lambda = LW,
+             method = "none")
> z2

Call:
ridge(formula = y ~ x1 + x2 + x3 + x4, data = portland, lambda = LW,
    method = "none")

Coefficients:
(Intercept)          x1          x2          x3          x4
    17.1889      2.0162      0.9762      0.5776      0.3127

Ridge parameter: 0.0032

Number of observations: 13
Effective number of parameters: 4.2749
Scale parameter estimate: 3.9478
```

```
# estimador ridge usando GCV
> z3 <- ridge(y ~ x1 + x2 + x3 + x4, data = portland, method = "GCV")
> z3

Call:
ridge(formula = y ~ x1 + x2 + x3 + x4, data = portland, method = "GCV")

Coefficients:
(Intercept)          x1          x2          x3          x4
    0.08545     2.16534     1.15864     0.73834     0.48950

Estimated ridge parameter: 1.9716

Number of observations: 13
Effective number of parameters: 3.9795
Scale parameter estimate: 5.0902

# 'k' óptimo usando GCV
> opt <- z3$lambda
> opt
[1] 1.971571
```

Cemento Portland (Woods, Steinour y Starke, 1932)



Cemento Portland (Woods, Steinour y Starke, 1932)

Resumen de estimación para los datos de cemento:

Parámetro	OLS		Ridge		
	homogéneo	no homogéneo	HKB	LW	GCV
β_0	—	62.4054	8.5870	17.1889	0.0855
β_1	2.1930	1.5511	2.1046	2.0162	2.1653
β_2	1.1533	0.5102	1.0648	0.9762	1.1586
β_3	0.7585	0.1019	0.6681	0.5776	0.7383
β_4	0.4863	-0.1441	0.3996	0.3127	0.4895
σ^2	4.0469	3.6818	4.0005	3.9478	5.0902
k	—	—	0.0077	0.0032	1.9716
edf	4.0000	5.0000	4.1369	4.2749	3.9795