

MAT-266: Colinealidad

Felipe Osorio

`fosorios.mat.utfsm.cl`

Departamento de Matemática, UTFSM



Considere el modelo

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

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

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

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

puede ser muy inestable.

Observación:

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



Observación:

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

- ▶ Tipicamente los coeficientes estimados $\hat{\beta}$ 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 X .



Algunas herramientas para el diagnóstico de colinealidad, son:

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

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

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

$$\mathbf{R}^{-1} = (\widetilde{\mathbf{X}}^{\top} \widetilde{\mathbf{X}})^{-1}, \quad \widetilde{\mathbf{X}} = (x_{ij} - \bar{x}_j),$$

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

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

donde R_j^2 es el coeficiente de correlación múltiple de \mathbf{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 \mathbf{R} .
- (d) **Número condición**: Desde la SVD de \mathbf{X} podemos escribir

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

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

La detección de colinealidad puede ser llevada a cabo usando

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

y $\kappa(\mathbf{X})$ “grande” (> 30) es un indicador de colinealidad.



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

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

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

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

desde donde sigue que

$$\mathbf{X} \mathbf{V}_1 = \mathbf{U}_1 \mathbf{D}_1, \quad \mathbf{X} \mathbf{V}_2 = \mathbf{0}.$$

Es decir, SVD permite **“detectar”** la dependencia lineal.



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

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

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

Resultado 1 (Estimador componentes principales):

Bajo los supuestos del modelo lineal en [A1-A4*](#), el estimador componentes principales para β puede ser escrito como

$$\begin{aligned} \hat{\beta}_{\text{PC}} &= \mathbf{U}_1 (\mathbf{U}_1^\top \mathbf{X}^\top \mathbf{X} \mathbf{U}_1)^{-1} \mathbf{U}_1^\top \mathbf{X}^\top \mathbf{Y} \\ &= \mathbf{U}_1 \mathbf{\Lambda}_1^{-1} \mathbf{U}_1^\top \mathbf{X}^\top \mathbf{Y} \end{aligned}$$



Demostración:

Por la ortogonalidad de $U = (U_1, U_2)$, sigue que

$$U_1^\top U_1 = I_r, \quad U_2^\top U_2 = I_{p-r}, \quad U_1 U_1^\top + U_2 U_2^\top = I_p,$$

y $U_1^\top U_2 = 0$. Ahora,

$$(X^\top X)^{-1} = U \Lambda^{-1} U^\top = U_1 \Lambda_1^{-1} U_1^\top + U_2 \Lambda_2^{-1} U_2^\top.$$

Usando que $U_1^\top U_2 = 0$ ($= U_2^\top U_1$), sigue

$$U_2^\top (X^\top X)^{-1} U_2 = U_2^\top (U_1 \Lambda_1^{-1} U_1^\top + U_2 \Lambda_2^{-1} U_2^\top) U_2 = \Lambda_2^{-1}$$

De este modo, $[U_2^\top (X^\top X)^{-1} U_2]^{-1} = \Lambda_2$, lo que permite escribir

$$\begin{aligned} (X^\top X)^{-1} U_2 [U_2^\top (X^\top X)^{-1} U_2]^{-1} U_2^\top (X^\top X)^{-1} \\ = (U_1 \Lambda_1^{-1} U_1^\top + U_2 \Lambda_2^{-1} U_2^\top) U_2 \Lambda_2 U_2^\top (U_1 \Lambda_1^{-1} U_1^\top + U_2 \Lambda_2^{-1} U_2^\top) \\ = U_2 \Lambda_2^{-1} U_2^\top. \end{aligned}$$



Es decir,

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

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

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

lo que concluye la prueba.



Observación:

- Es posible notar que el estimador PC es un caso particular del **estimador restringido** con respecto a:

$$U_2^\top \beta = 0.$$

- $\hat{\beta}_{PC}$ depende del 'parámetro' r . En efecto,

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \\ &= (\mathbf{U}_1 \mathbf{\Lambda}_1^{-1} \mathbf{U}_1^\top + \mathbf{U}_2 \mathbf{\Lambda}_2^{-1} \mathbf{U}_2^\top) \mathbf{X}^\top \mathbf{Y}.\end{aligned}$$

De este modo podemos interpretar $\hat{\beta}_{PC}$ como una modificación del OLS que **desconsidera** $\mathbf{U}_2 \mathbf{\Lambda}_2^{-1} \mathbf{U}_2^\top$.



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

$$F = \left(\frac{n-p}{p-r} \right) \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{PC}})^\top \mathbf{X}^\top \mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{PC}})}{\mathbf{Y}^\top (\mathbf{I} - \mathbf{H}) \mathbf{Y}}.$$

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**

$$\hat{\beta}_k = (\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{Y}, \quad k \geq 0$$

donde k es conocido como **parámetro ridge**.

Note que

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

De este modo,

$$E(\hat{\beta}_k) = (\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} \beta,$$

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

¹Technometrics 12, 55-67.

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

$$\text{MSE} = \text{E}\{\|\hat{\beta}_k - \beta\|^2\} = \text{tr Cov}(\hat{\beta}_k) + \|\text{E}(\hat{\beta}_k) - \beta\|^2.$$

En efecto,

$$\begin{aligned}\text{Cov}(\hat{\beta}_k) &= (\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} \text{Cov}(\hat{\beta}) \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1} \\ &= \sigma^2 (\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1}.\end{aligned}$$

Además,

$$\begin{aligned}\text{bias}(\hat{\beta}_k, \hat{\beta}) &= \text{E}(\hat{\beta}_k) - \beta = (\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} \beta - \beta \\ &= (\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1} [\mathbf{X}^\top \mathbf{X} \beta - (\mathbf{X}^\top \mathbf{X} + k\mathbf{I}) \beta] \\ &= -k (\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1} \beta.\end{aligned}$$

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

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

De este modo,

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

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

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



El estimador ridge tiene varias interpretaciones interesantes, por ejemplo:

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

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

que puede ser expresado de forma equivalente como

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

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

Observación:

Este tipo de regularización es conocida como **regularización de Tikhonov**.²

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

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

$$\mathbf{Y}_a = \mathbf{X}_a \boldsymbol{\beta} + \boldsymbol{\epsilon}_a, \quad \boldsymbol{\epsilon}_a \sim N_{n+p}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

donde

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

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

Resultado 2:

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

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

siempre que $0 \leq k_1 < k_2$.



Demostración:

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

$$\|\hat{\beta}_k\|^2 = \hat{\beta}^\top \mathbf{M}_k \hat{\beta}, \quad \mathbf{M}_k = \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-2} \mathbf{X}^\top \mathbf{X}.$$

Basado en la SVD de \mathbf{X} , tenemos

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

con

$$\mathbf{\Gamma}_k = \text{diag} \left(\frac{\delta_1^4}{(\delta_1^2 + k)^2}, \dots, \frac{\delta_p^4}{(\delta_p^2 + k)^2} \right).$$

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

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

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



Observación:

Note que $\lim_{k \rightarrow \infty} \|\hat{\beta}_k\|^2 = 0$ y de ahí que

$$\lim_{k \rightarrow \infty} \hat{\beta}_k = \mathbf{0}. \quad (1)$$

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

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

es llamada matrix ridge-shrinking.



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

(a) Hoerl, Kennard y Baldwin (1975):³

$$\hat{k}_{\text{HKB}} = \frac{ps^2}{\|\hat{\beta}\|^2}.$$

(b) Lawless y Wang (1976):⁴

$$\hat{k}_{\text{LW}} = \frac{ps^2}{\hat{\beta}^\top \mathbf{X}^\top \mathbf{X} \hat{\beta}}.$$

(c) Lindley y Smith (1972):⁵

$$\hat{k}_{\text{LS}} = \frac{(n-p)(p+2)}{(n+2)} \frac{s^2}{\|\hat{\beta}\|^2}.$$

³Communications in Statistics: Theory and Methods **4**, 105-123.

⁴Communications in Statistics: Theory and Methods **5**, 307-323.

⁵Journal 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

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

donde

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

Es fácil notar que $\hat{\mathbf{Y}}_k = \mathbf{H}(k)\mathbf{Y}$. En este contexto se ha definido

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

como el **número de parámetros efectivos**. En efecto, para $k = 0$, sigue que $\text{edf} = p$.

⁶Technometrics **21**, 215-223.



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

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

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

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

es decir,

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

Por otro lado,

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



Además,

$$\hat{\mathbf{Y}}_k = \mathbf{X}\hat{\boldsymbol{\beta}}_k = \mathbf{UDV}^\top \hat{\boldsymbol{\beta}}_k = \mathbf{UD}\hat{\boldsymbol{\alpha}}_k.$$

Lo que permite escribir

$$V(k) = \frac{1}{n} \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_k\|^2}{\{\text{tr}(\mathbf{I} - \mathbf{H}(k))/n\}^2} = \frac{\|\mathbf{Y} - \mathbf{UD}\hat{\boldsymbol{\alpha}}_k\|^2}{(1 - \text{edf}/n)^2},$$

Observación:

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

Además, en términos del modelo canónico, tenemos:

$$\hat{k}_{\text{HKB}} = \frac{ps^2}{\|\hat{\boldsymbol{\alpha}}\|^2}, \quad \hat{k}_{\text{LW}} = \frac{ps^2}{\|\mathbf{Z}\|^2},$$

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



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).

⁷Industrial 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 $\kappa(\mathbf{X}) = 9.432$, esto es, \mathbf{X} es bien condicionada (variables centradas $\kappa(\widetilde{\mathbf{X}}) = 37.106$).

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 $\kappa(\mathbf{X}) = 249.578$, 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.

⁸Statistical Theory with Engineering Application, Wiley.

⁹Technometrics **8**, 27-51.

¹⁰Fitting 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")
```

¹¹Proximamente, 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
```



Cemento Portland (Woods, Steinour y Starke, 1932)

```
# 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
```



Cemento Portland (Woods, Steinour y Starke, 1932)

```
# 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
```



Cemento Portland (Woods, Steinour y Starke, 1932)

```
# 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



Cemento Portland (Woods, Steinour y Starke, 1932)

```
# 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

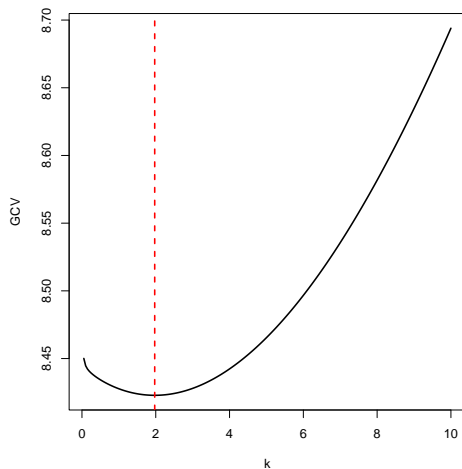
Estimated ridge parameter: 1.9716

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
κ	9.4325	249.5783	92.4131	130.8854	10.7852

