MAT-266: Colinealidad

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Considere el modelo

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

donde $\mathsf{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ 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 $\boldsymbol{X}\boldsymbol{a} \approx \boldsymbol{0}$.

Colinealidad

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 $\boldsymbol{X}$.

EX UMBRA IN SOLEM

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

$$\boldsymbol{R}^{-1} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1},$$

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.

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.

EX UMBRA IN SOLEM

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 $\boldsymbol{\Lambda}_1$ y $\boldsymbol{\Lambda}_2$ son diagonales, 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,

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

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

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

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

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

lo que concluye la prueba.

Observación:

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

- $\widehat{\boldsymbol{\beta}}_{\text{PC}}$ 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 $\widehat{\boldsymbol{\beta}}_{\text{PC}}$ 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.

Colinealidad

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

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

De este modo,

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

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

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

Colinealidad

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

$$\mathsf{MSE} = \mathsf{E}\{\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}\|^2\} = \sigma^2 \sum_{j=1}^{n} \frac{\lambda_j^2}{(\lambda_j + k)^2} + k^2 \boldsymbol{\beta}^\top (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-2} \boldsymbol{\beta},$$

donde $\lambda_1, \dots, \lambda_p$ son los valores propios de $\boldsymbol{X}^\top \boldsymbol{X}$.

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.

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

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

donde

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

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 descomposición espectral de $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, sigue que

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

con

$$\boldsymbol{\Gamma}_k = \operatorname{diag}\Big(\frac{\lambda_1^2}{(\lambda_1 + k)^2}, \dots, \frac{\lambda_p^2}{(\lambda_p + k)^2}\Big).$$

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

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

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. En efecto, algunas de estas alternativas son:

(a) Hoerl, Kennard y Baldwin (1975):[2]

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

(b) Lawless y Wang (1976):[3]

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

(c) Lindley y Smith (1972):[4]

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

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

Golub, Heath y Wahba (1979)[5] 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 $= p$.

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