IECD-325: Diagnóstico de influencia en regresión lineal

Felipe Osorio

felipe.osorio@uv.cl

Diagnóstico de influencia en regresión lineal

Suponga que se desea evaluar el efecto de eliminar una observación sobre la estimación de $\boldsymbol{\beta}$. En este caso, podemos considerar el modelo de datos eliminados

$$\boldsymbol{Y}_{(i)} = \boldsymbol{X}_{(i)}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{(i)}, \tag{1}$$

de ahí que

$$\widehat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}_{(i)}^\top \boldsymbol{Y}_{(i)}.$$

Considere

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{(i)} \\ \boldsymbol{x}_i^\top \end{pmatrix}, \qquad \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_{(i)} \\ Y_i \end{pmatrix},$$

de ahí que

$$\boldsymbol{X}^\top \boldsymbol{X} = (\boldsymbol{X}_{(i)}^\top, \boldsymbol{x}_i) \begin{pmatrix} \boldsymbol{X}_{(i)} \\ \boldsymbol{x}_i^\top \end{pmatrix} = \boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)} + \boldsymbol{x}_i \boldsymbol{x}_i^\top,$$

$$\boldsymbol{X}^\top \boldsymbol{Y} = (\boldsymbol{X}_{(i)}^\top, \boldsymbol{x}_i) \begin{pmatrix} \boldsymbol{Y}_{(i)} \\ Y_i \end{pmatrix} = \boldsymbol{X}_{(i)}^\top \boldsymbol{Y}_{(i)} + \boldsymbol{x}_i Y_i.$$

Reagrupando, podemos escribir

$$\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)} = \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{I} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i \boldsymbol{x}_i^\top),$$

cuya matriz inversa es dada por

$$\begin{aligned}(\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)})^{-1} &= (\boldsymbol{I} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i \boldsymbol{x}_i^\top)^{-1}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\ &= \Big\{\boldsymbol{I} + \frac{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i \boldsymbol{x}_i^\top}{1 - \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i}\Big\}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}.\end{aligned}$$

De este modo,

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{(i)} &= (\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}_{(i)}^\top \boldsymbol{Y}_{(i)} \\ &= \Big\{\boldsymbol{I} + \frac{1}{1 - h_{ii}}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i \boldsymbol{x}_i^\top\Big\}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}(\boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{x}_i Y_i) \\ &= \Big\{\boldsymbol{I} + \frac{1}{1 - h_{ii}}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i \boldsymbol{x}_i^\top\Big\}\big(\widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i Y_i\big) \\ &= \widehat{\boldsymbol{\beta}} - \frac{e_i}{1 - h_{ii}}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i.\end{aligned}$$

Observación:

Una característica importante de la ecuación anterior es que depende solamente de cálculos obtenidos desde el modelo con datos completos.

Sea

$$e_{j(i)} = Y_j - \widehat{Y}_{j(i)} = Y_j - \boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}_{(i)},$$

el j-ésimo residuo con la i-ésima observación eliminada, en particular

$$\begin{aligned} e_{i(i)} = Y_i - \widehat{Y}_{i(i)} = Y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{(i)} &= Y_i - \boldsymbol{x}_i^\top \left(\widehat{\boldsymbol{\beta}} - \frac{e_i}{1-h_{ii}} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i \right) \\ &= e_i + \frac{e_i h_{ii}}{1-h_{ii}} = \frac{e_i}{1-h_{ii}}, \end{aligned}$$

es conocido como residuo eliminado.

Note además que

$$\begin{aligned}\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(i)} &= \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\Big(\widehat{\boldsymbol{\beta}} - \frac{e_i}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\Big)\\ &= \frac{e_i}{1-h_{ii}}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i.\end{aligned}$$

Por otro lado,

$$\begin{aligned}(\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(i)})^\top(\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(i)}) &= \Big(\frac{e_i}{1-h_{ii}}\Big)^2\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\\ &= h_{ii}\Big(\frac{e_i}{1-h_{ii}}\Big)^2,\end{aligned}$$

es decir esta medida del efecto de remover la i-ésima observación sobre la predicción depende sólo de e_i y h_{ii}.

Podemos evaluar el efecto de la i-ésima observación sobre el estimador de σ^2 podemos considerar el estimador $s^2_{(i)}$. En efecto,

$$\mathsf{RSS}_{(i)} = \sum_{j \neq i} (Y_j - \boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}_{(i)})^2 = \mathsf{RSS} - \frac{e_i^2}{1 - h_{ii}}.$$

De ahí que

$$s^2_{(i)} = \frac{1}{n - p - 1} \left\{ (n - p) s^2 - \frac{e_i^2}{1 - h_{ii}} \right\}$$

Resultado 1:

Considere el modelo de salto en la media:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{d}_i\gamma + \boldsymbol{\epsilon}, \tag{2}$$

con $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ y $\boldsymbol{d}_i = (\boldsymbol{0}, 1, \boldsymbol{0})^\top$ un vector de ceros con un 1 en la i-ésima posición. De este modo, el estimador ML de $\boldsymbol{\beta}$ en el modelo (1) y (2) coinciden.

Demostración:

El resultado sigue mediante escribir el modelo en (2) como

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{d}_i), \qquad \boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \gamma)^\top,$$

y $\widehat{\boldsymbol{\theta}} = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{Y}$. Tenemos que

$$\boldsymbol{Z}^\top\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{X}^\top \\ \boldsymbol{d}_i^\top \end{pmatrix} (\boldsymbol{X}, \boldsymbol{d}_i) = \begin{pmatrix} \boldsymbol{X}^\top\boldsymbol{X} & \boldsymbol{X}^\top\boldsymbol{d}_i \\ \boldsymbol{d}_i^\top\boldsymbol{X} & \boldsymbol{d}_i^\top\boldsymbol{d}_i \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^\top\boldsymbol{X} & \boldsymbol{x}_i \\ \boldsymbol{x}_i^\top & 1 \end{pmatrix},$$

$$\boldsymbol{Z}^\top\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{X}^\top \\ \boldsymbol{d}_i^\top \end{pmatrix} \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{X}^\top\boldsymbol{Y} \\ Y_i \end{pmatrix}.$$

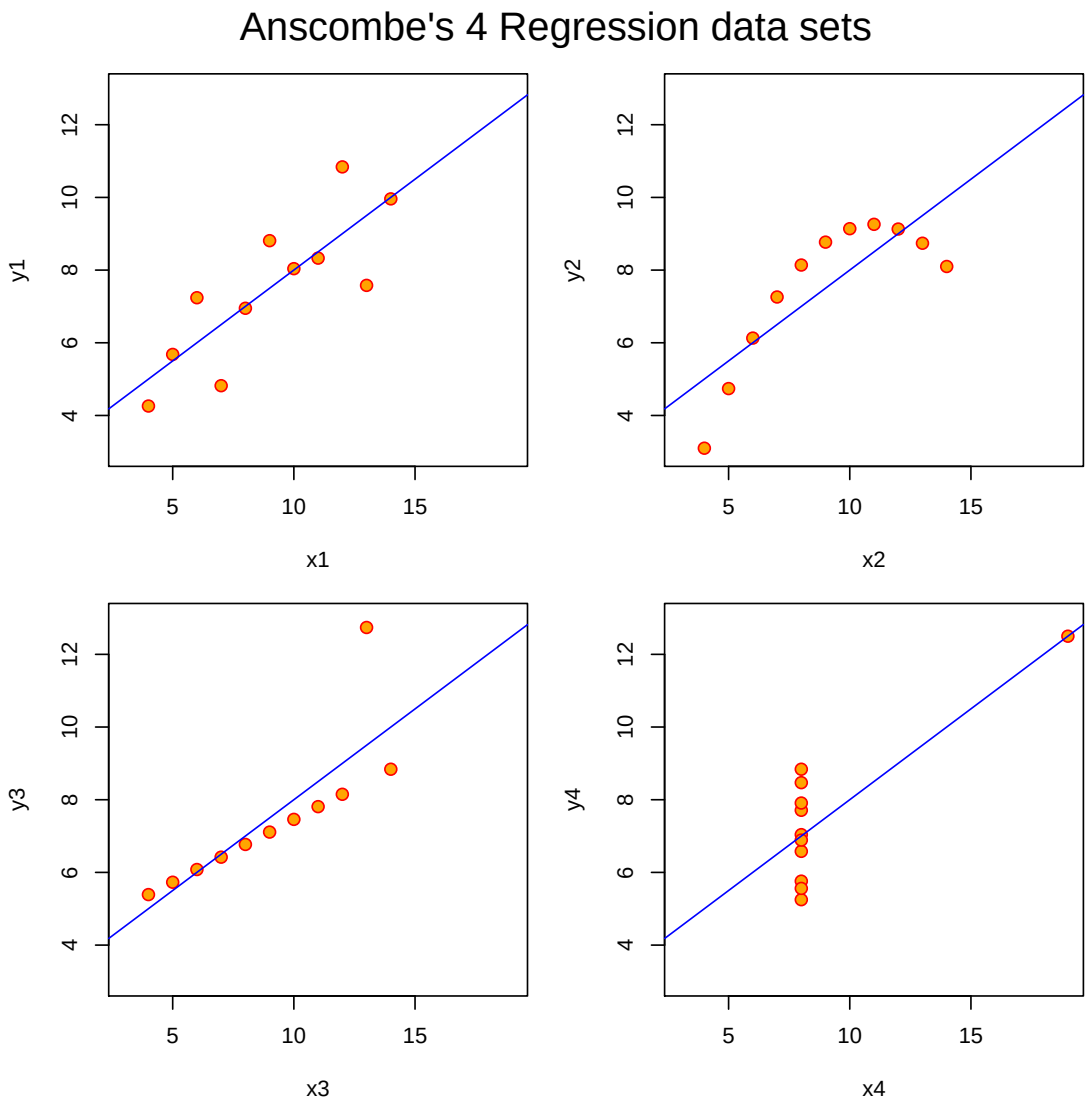
Sabemos que

$$(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} = \begin{pmatrix} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1} & -\frac{1}{1-h_{ii}}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i \\ -\frac{1}{1-h_{ii}}\boldsymbol{x}_i^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1} & \frac{1}{1-h_{ii}} \end{pmatrix},$$

luego

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_* &= \left\{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} + \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right\}\boldsymbol{X}^\top \boldsymbol{Y} - \frac{Y_i}{1-h_{ii}}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i \\ &= \left\{\boldsymbol{I} + \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} - \frac{Y_i}{1-h_{ii}}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i \\ &= \widehat{\boldsymbol{\beta}} + \frac{\widehat{Y}_i}{1-h_{ii}}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i - \frac{Y_i}{1-h_{ii}}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i = \widehat{\boldsymbol{\beta}}_{(i)}. \end{aligned}$$

Cuarteto de regresiones “idénticas” de Anscombe (1973)



Basado en el elipsoide de confianza del $100(1-\alpha)\%$ para $\boldsymbol{\beta}$,

$$\frac{(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})}{ps^2} \leq F_{p,n-p}(1-\alpha).$$

Cook (1977)[1] propuso determinar la influencia de la i-ésima observación, usando

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)}-\widehat{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{(i)}-\widehat{\boldsymbol{\beta}})}{ps^2} = \frac{r_i^2}{p}\Big(\frac{h_i}{1-h_i}\Big),$$

para $i = 1, \ldots, n$, y recomendó comparar D_i con algún percentil de la distribución $F_{p,n-p}$ $(\alpha = 0.10)$. Otra alternativa más razonable puede ser usar $\alpha = 0.50$, y se ha sugerido que $D_i > 1$ es un indicador de observaciones influyentes.

[1] Technometrics 19, 15-18

Diagnóstico de influencia en regresión lineal

Welsch y Kuh (1977)[2] propusieron medir el impacto en la i-ésima observación sobre el valor predicho como

$$\text{DFFIT}_i = \widehat{Y}_i - \widehat{Y}_{i(i)} = \boldsymbol{x}_i^\top (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}) = \frac{h_{ii} e_i}{1 - h_{ii}},$$

y su versión estandarizada

$$\begin{aligned} \text{DFFITS}_i &= \frac{\widehat{Y}_i - \widehat{Y}_{i(i)}}{s_{(i)}\sqrt{h_{ii}}} = \Big(\frac{h_{ii}}{1 - h_{ii}}\Big)^{1/2} \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}} \\ &= \Big(\frac{h_{ii}}{1 - h_{ii}}\Big)^{1/2} t_i \end{aligned}$$

Belsley, Kuh y Welsch (1980)[3] sugieren poner especial atención en aquellos casos donde $\text{DFFITS}_i > 2\sqrt{p/n}$.

Observación:

En ocasiones esta medida es conocida como distancia Welsch-Kuh.

[2] Working paper No. 173, National Bureau of Economics, Cambridge

[3] Regression Diagnostics. Wiley, New York

Atkinson (1981)[4] sugirió usar una versión modificada de la distancia de Cook, como

$$\begin{aligned} AK_i &= \sqrt{\Big(\frac{n-p}{p}\Big)\Big(\frac{h_{ii}}{1-h_{ii}}\Big)}\Big|\frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}\Big| \\ &= \sqrt{\Big(\frac{n-p}{p}\Big)\Big(\frac{h_{ii}}{1-h_{ii}}\Big)}|t_i| \\ &= \sqrt{\frac{n-p}{p}}|\,\text{DFFITS}_i\,|. \end{aligned}$$

Cuando $h_{ii} = p/n, \forall i$, tenemos $AK_i = |t_i|$ debido a esto se recomienda hacer el gráfico de AK_i vs. $|t_i|$. Además podemos identificar la i-ésima observación como influyente si $AK_i > 2$.

Observación:

AK_i puede considerarse como una medida de influencia conjunta sobre $\widehat{\boldsymbol{\beta}}$ y s^2 simultáneamente.

[4] Biometrika 68, 13-20

Cook y Weisberg (1980)[5] propusieron considerar medidas generales de influencia[6] considerando,

$$D_i(\boldsymbol{M}, c) = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^\top \boldsymbol{M} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{c},$$

donde $\boldsymbol{M}$ es matriz definida positiva $p \times p$ y $c > 0$ es un factor de escala.

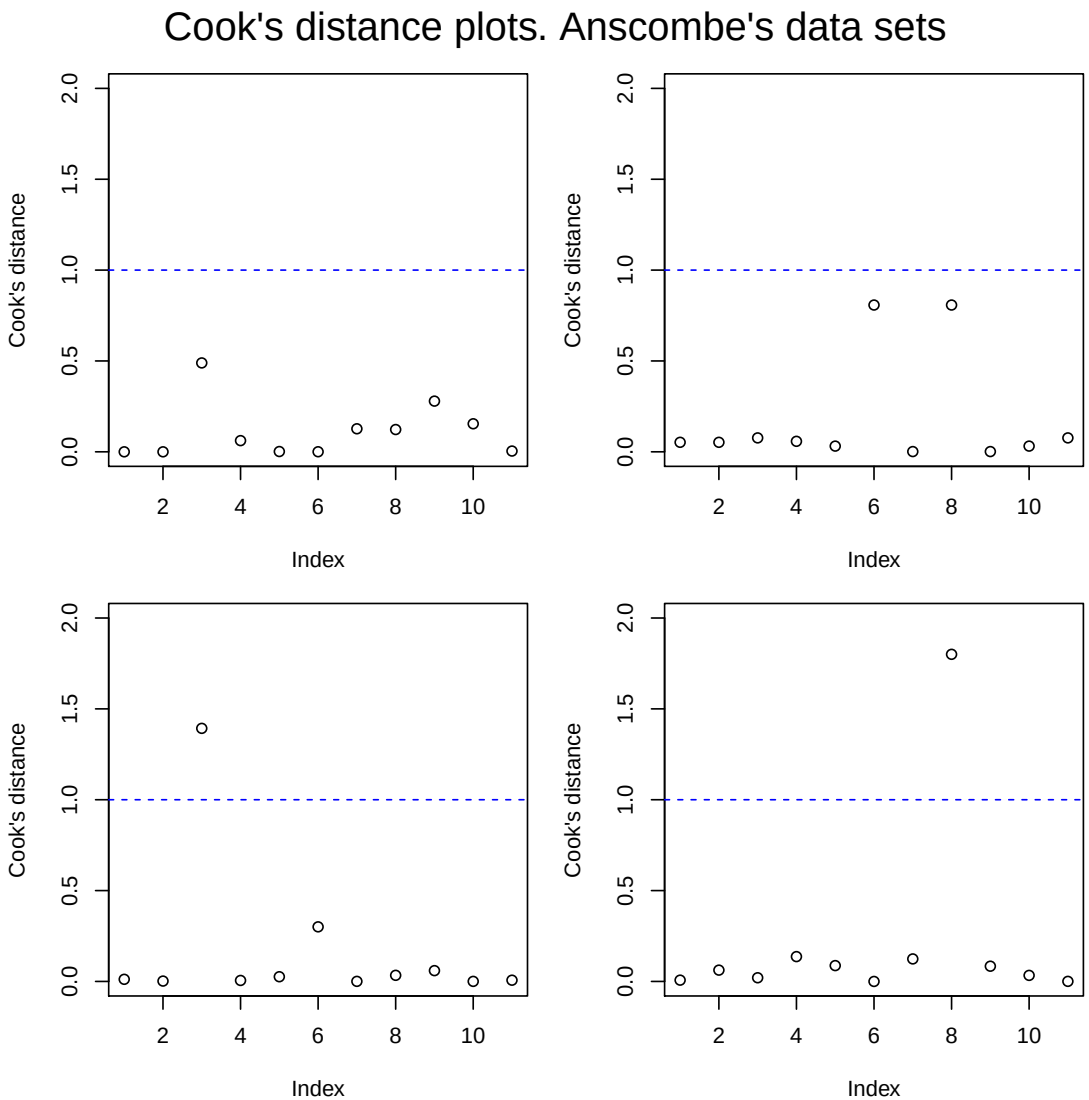
Algunas medidas de influencia:

$\boldsymbol{M}$	c	Medida	Referencia
$\boldsymbol{X}^\top \boldsymbol{X}$	ps^2	D_i	Cook (1977)
$\boldsymbol{X}^\top \boldsymbol{X}$	$ps^2_{(i)}$	$(DFFITS_i)^2$	Welsch y Kuh (1977)
$\boldsymbol{X}^\top \boldsymbol{X}$	$(n-1)^2 ps^2_{(i)}/(n-p)$	AK_i	Atkinson (1981)

[5] Technometrics 22, 495-508

[6] Basadas en la función de influencia empírica.

Distancia de Cook: Datos de Anscombe (1973)



Considere comparar $\text{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ con la matriz de covarianza que resulta de eliminar el i-ésimo caso. Esto lleva a (Belsley, Kuh y Welsch, 1980)

$$COVRATIO_i = \frac{\det\{s^2_{(i)}(\boldsymbol{X}^\top_{(i)}\boldsymbol{X}_{(i)})^{-1}\}}{\det\{s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\}} = \Big(\frac{s^2_{(i)}}{s^2}\Big)^p \frac{\det(\boldsymbol{X}^\top_{(i)}\boldsymbol{X}_{(i)})^{-1}}{\det(\boldsymbol{X}^\top\boldsymbol{X})^{-1}}$$
$$= \frac{1}{1-h_i}\Big(\frac{n-p-r_i^2}{n-p-1}\Big)^p,$$

se ha planteado como punto de corte $|COVRATIO_i - 1| > 3p/n$.

Diagnóstico de influencia en regresión lineal

Existe un repertorio bastante extenso de medidas de influencia, por ejemplo:

Medida	Punto de corte
$D_i = \frac{r_i^2}{p}\left(\frac{h_i}{1-h_i}\right)$	$F_{p,n-p}(1-\alpha)$
$\text{DFFITS}_i = \lvert t_i \rvert \sqrt{\frac{h_i}{1-h_i}}$	$2\sqrt{p/n}$
$AK_i = \text{DFFITS}_i \sqrt{\frac{n-p}{p}}$	$2\sqrt{(n-p)/n}$
$W_i = \text{DFFITS}_i \sqrt{\frac{n-1}{1-h_i}}$	$3\sqrt{p}$
$COVRATIO_i = \frac{1}{1-h_i}\left(\frac{n-p-r_i^2}{n-p-1}\right)^p$	$\lvert COVRATIO_i - 1 \rvert > 3p/n$
$h_i = \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i$	$2p/n$
$r_i = \frac{e_i}{s\sqrt{1-h_i}}$	$\approx \text{N}(0,1)$
$t_i = r_i \sqrt{\frac{n-p-1}{n-p-r_i^2}}$	$\approx t(n-p-1)$

Software Estadístico y Medidas de influencia

Belsley, Kuh y Welsch (1981) y Velleman y Welsch (1981)[7] han discutido estrategias para amenizar el cálculo de estas medidas de influencia.

Para modelos de regresión lineal algunas de estas medidas han sido implementadas en software estadístico tal como SAS, SPSS, S-PLUS/R.

En particular, R (o S-PLUS) disponen de las funciones `lm.influence` y `ls.diag` asociadas con las funciones `lm` (o `glm`) y `lsfit`, respectivamente.

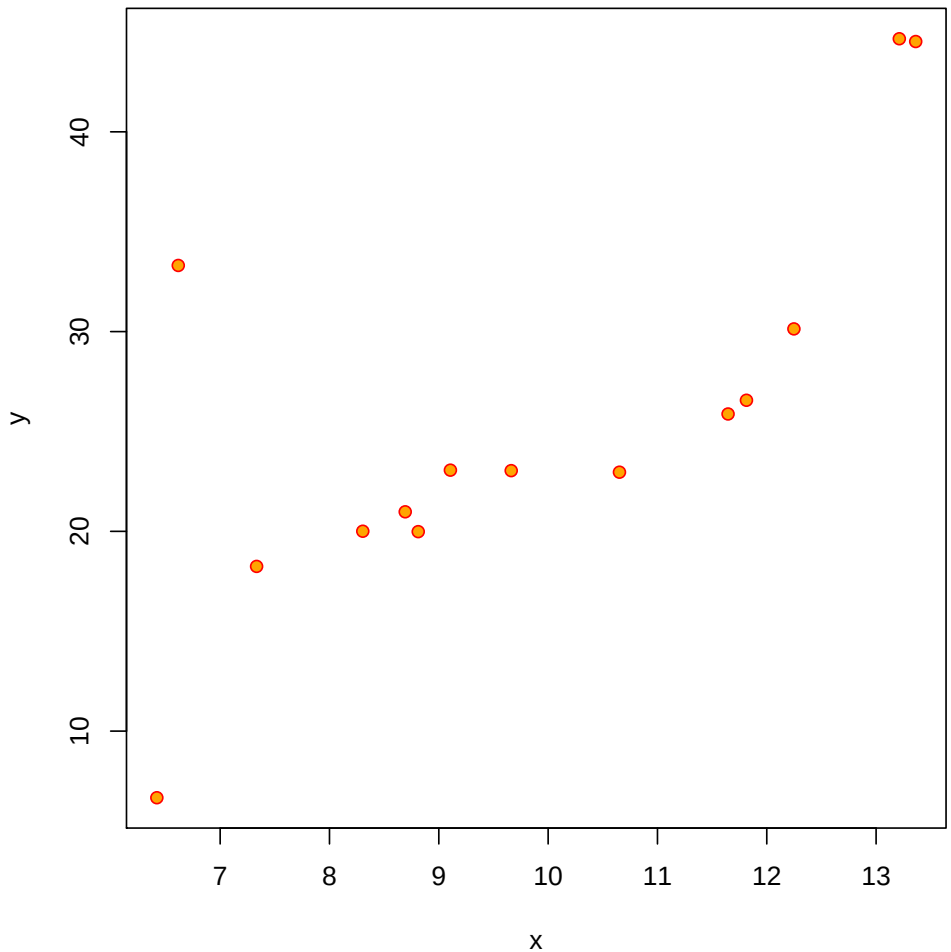
La función `lm.influence` dispone de las siguientes medidas:

```
rstandard     rstudent     dffits     dfbetas
covratio     cooks.distance     hatvalues
```

Estas cantidades pueden ser escritas de forma eficiente usando la descomposición QR o SVD.

[7] The American Statistician 35, 234-242.

Influencia de múltiples casos



Influencia de múltiples casos

Sea $I = (i_1, \ldots, i_m)^\top$ vector de índices m-dimensional, tal que $1 \leq i_j \leq m$. La distancia de Cook adopta la forma:

$$D_I(\boldsymbol{M}, c) = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)})^\top \boldsymbol{M} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)})}{c}.$$

Usando que

$$\widehat{\boldsymbol{\beta}}_{(I)} = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_I^\top (\boldsymbol{I} - \boldsymbol{H}_I)^{-1} \boldsymbol{e}_I,$$

lleva a una fórmula más conveniente para la distancia de Cook, D_I

$$D_I = \frac{\boldsymbol{e}_I^\top (\boldsymbol{I} - \boldsymbol{H}_I)^{-1} \boldsymbol{H}_I (\boldsymbol{I} - \boldsymbol{H}_I)^{-1} \boldsymbol{e}_I}{ps^2}.$$

Observación:

D_I puede ser evaluado en $\binom{n}{m}$ posibles subconjuntos de casos.

Influencia de múltiples casos

Múltiples datos atípicos pueden ser detectados mediante el modelo

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

donde $\boldsymbol{D}_m \in \mathbb{R}^{n\times m}$ cuya i-ésima columna es $\boldsymbol{d}_{i_k}$ y $\boldsymbol{\phi}$ es vector m-dimensional. La hipótesis $H : \boldsymbol{\phi} = \mathbf{0}$ lleva al estadístico

$$t_I^2 = \Big(\frac{n-p-m}{m}\Big)\frac{\boldsymbol{e}_I^T(\boldsymbol{I}-\boldsymbol{H}_I)^{-1}\boldsymbol{e}_I}{(n-p)s^2 - \boldsymbol{e}_I^T(\boldsymbol{I}-\boldsymbol{H}_I)^{-1}\boldsymbol{e}_I},$$

con distribución nula (bajo normalidad) dada por $F(m, n-p-m)$.

Influencia de múltiples casos

- Lamentablemente este problema tiene un costo computacional muy alto.
- Típicamente las técnicas de eliminación de un único caso no son suficiente.
- Técnicas para multiples casos pueden sufrir del efecto de enmascaramiento.

Outliers y colinealidad en regresión lineal

Outliers y su relación con el problema de colinealidad:

- Propuestas para robustificar el estimador ridge
  (Holland, 1973; Askin y Montgomery, 1980; Lawrence y Marsh, 1984 y Silvapulle, 1991).
- Diagnóstico de influencia en regresión ridge
  (Steece, 1986; Walker y Birch, 1988; Billor y Loynes, 1999; Shi y Wang, 1999 y Labra, Aoki y Rojas, 2007; Ogueda y Osorio, 2025).
- Colinealidad inducida por outliers
  (Mason y Gunst, 1985; Hadi, 1988 y Walker, 1989).

Medidas de diagnóstico en regresión ridge

Walker y Birch (1988)[8], propusieron las siguientes medidas de diagnóstico:

- Leverage:

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

- Distancia de Cook:

$$D_i^* = \frac{(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k(i))^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k(i))}{ps^2},$$

$$D_i^{**} = \frac{(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k(i))^\top (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k(i))}{ps^2},$$

para $i = 1, \ldots, n$.

[8] Technometrics 30, 221-227.

Outliers que inducen colinealidad

Hadi (1988)[9] y Walker (1989)[10], proponen detectar observaciones que inducen o ocultan colinealidad, mediante:

$$\gamma_i = \frac{\kappa_{(i)} - \kappa}{\kappa}, \qquad i = 1, \ldots, n,$$

donde $\kappa_{(i)} = \kappa(\boldsymbol{X}_{(i)})$ y $\kappa = \kappa(\boldsymbol{X})$.

Concluyeron que:

Aquellas observaciones que afectan el condicionamiento de $\boldsymbol{X}$ frecuentemente tienen alto leverage.

[9] Computational Statistics & Data Analysis 7, 143-159

[10] Communications in Statistics: Theory and Methods 18, 1675-1690.

Cemento Portland (Woods, Steinour y Starke, 1932)

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).

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 $\kappa(\widetilde{\boldsymbol{X}}) = 37.106$).

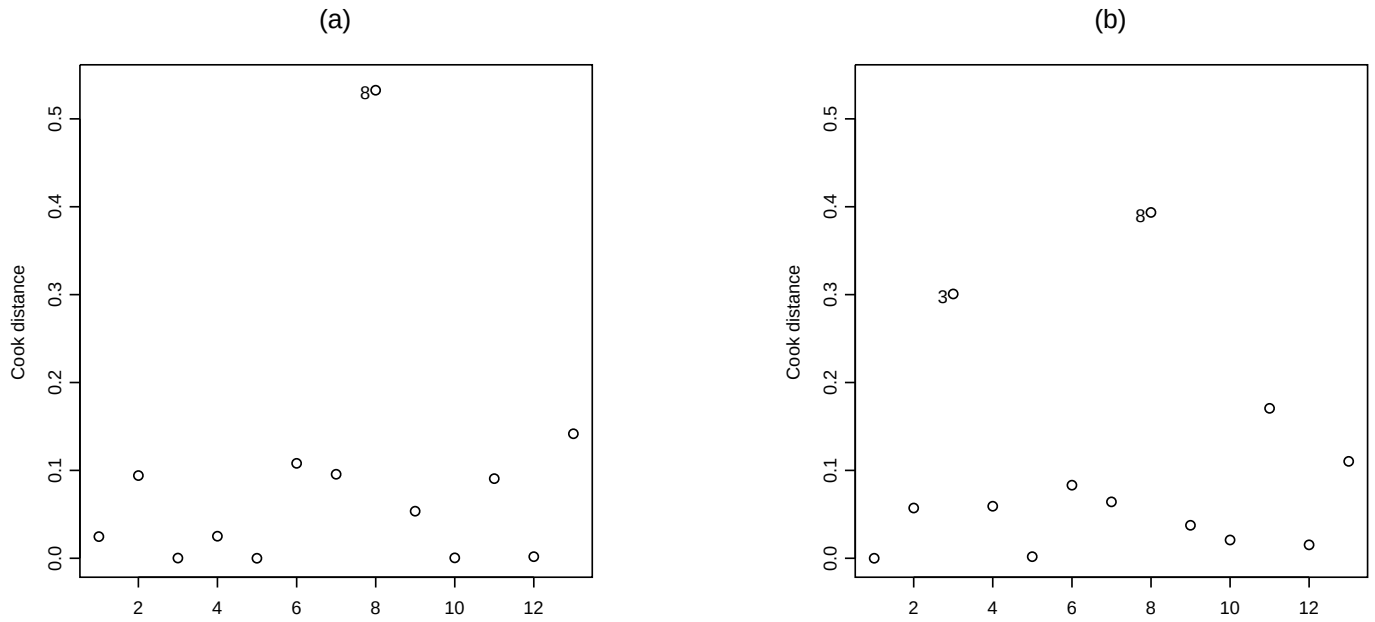
Por otro lado, Hald (1952), Gorman y Toman (1966) y Daniel y Wood (1980) adoptan un modelo con intercepto (modelo no homogéneo). En cuyo caso $\kappa(\boldsymbol{X}) = 249.578$, sugiriendo la presencia de colinealidad. El aumento en el número condición se debe a que existe una relación lineal aproximada,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

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

de modo que incluir el intercepto causa una colinealidad severa.

Cemento Portland

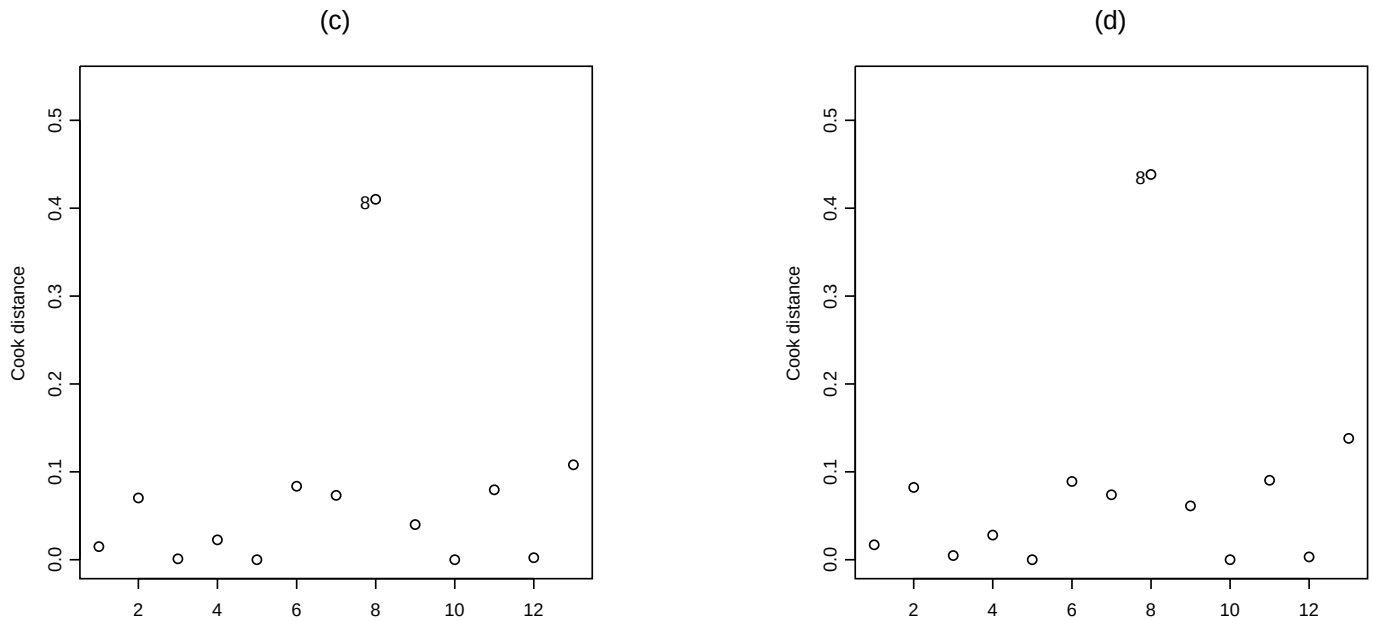
Distancia de Cook:



Estimación mínimos cuadrados: Modelo homogéneo (a) y no homogéneo (b).

Cemento Portland

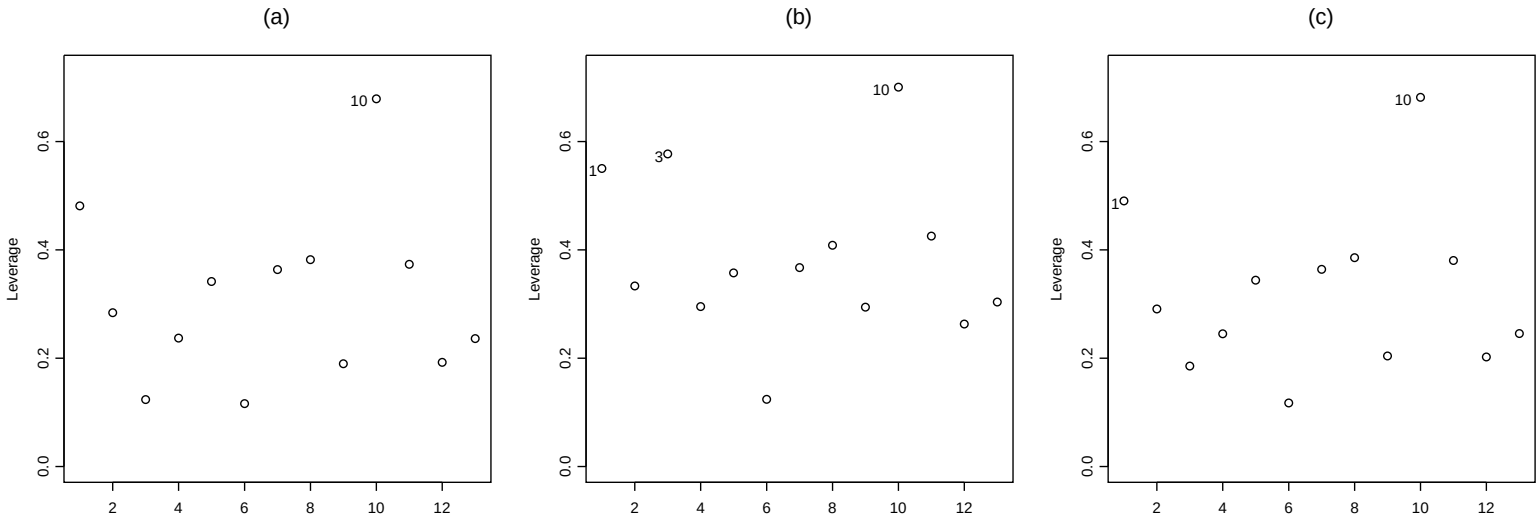
Distancia de Cook:



Estimación ridge: D_i^* (c) y D_i^{**} (d).

Cemento Portland

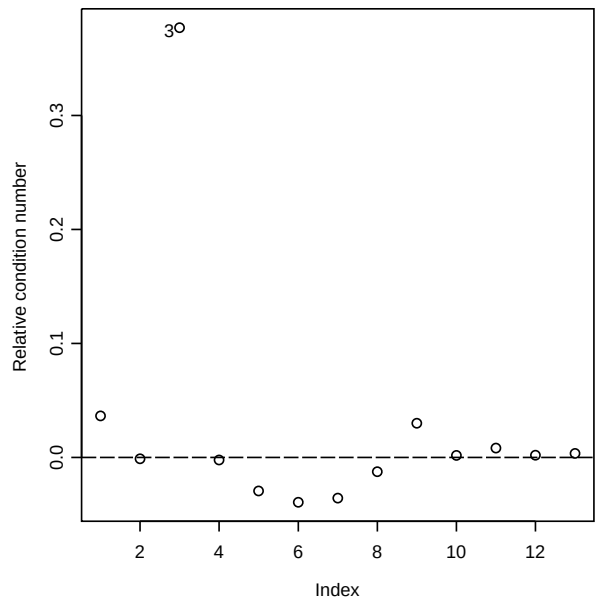
Leverage:



Modelo homogéneo (a), modelo no homogéneo (b) y ridge (c).

Cemento Portland

Número condición relativo: $\gamma_i = (\kappa_{(i)} - \kappa)/\kappa$



Es decir, obs. 3 afecta el condicionamiento de $\boldsymbol{X}$. En efecto, $\kappa(\boldsymbol{X}_{(3)}) = 343.658$ mientras que, $\kappa(\boldsymbol{X}) = 249.578$ (oculta una colinealidad).