

# MAT-266: Diagnóstico de influencia en regresión lineal

Felipe Osorio

[fosorios.mat.utfsm.cl](mailto:fosorios.mat.utfsm.cl)

Departamento de Matemática, UTFSM



## Diagnóstico de influencia en regresión lineal

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

$$\mathbf{Y}_{(i)} = \mathbf{X}_{(i)}\beta + \epsilon_{(i)}, \quad (1)$$

de ahí que

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

Considere

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

de ahí que

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

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



# Diagnóstico de influencia en regresión lineal

Reagrupando, podemos escribir

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

cuya matriz inversa es dada por

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

De este modo,

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



## Observación:

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

Sea

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

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

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

es conocido como [residuo eliminado](#).



# Diagnóstico de influencia en regresión lineal

Note además que

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

Por otro lado,

$$\begin{aligned}(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})^\top (\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)}) &= \left(\frac{e_i}{1 - h_{ii}}\right)^2 \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i \\ &= h_{ii} \left(\frac{e_i}{1 - h_{ii}}\right)^2,\end{aligned}$$

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



## Diagnóstico de influencia en regresión lineal

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

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

De ahí que

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



## Resultado 1:

Considere el [modelo de salto en la media](#):

$$Y = X\beta + d_i\gamma + \epsilon, \quad (2)$$

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

## Demostración:

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

$$Y = Z\theta + \epsilon, \quad Z = (X, d_i), \quad \theta = (\beta^\top, \gamma)^\top,$$

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

$$Z^\top Z = \begin{pmatrix} X^\top \\ d_i^\top \end{pmatrix} (X, d_i) = \begin{pmatrix} X^\top X & X^\top d_i \\ d_i^\top X & d_i^\top d_i \end{pmatrix} = \begin{pmatrix} X^\top X & x_i \\ x_i^\top & 1 \end{pmatrix},$$

$$Z^\top Y = \begin{pmatrix} X^\top \\ d_i^\top \end{pmatrix} Y = \begin{pmatrix} X^\top Y \\ Y_i \end{pmatrix}.$$



## Diagnóstico de influencia en regresión lineal

Sabemos que

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

luego

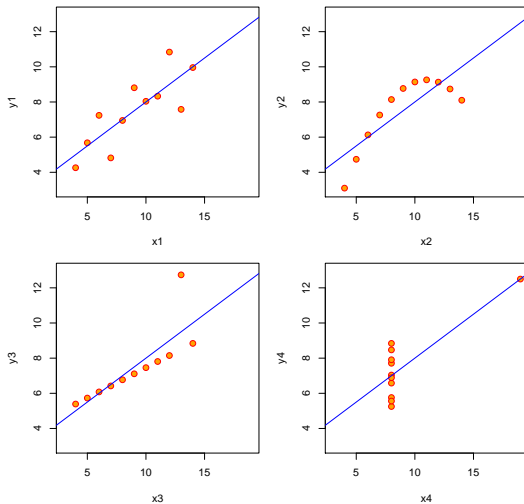
$$\begin{aligned} \hat{\beta}_* &= \left\{ (\mathbf{X}^\top \mathbf{X})^{-1} + \frac{1}{1-h_{ii}}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \right\} \mathbf{X}^\top \mathbf{Y} - \frac{Y_i}{1-h_{ii}}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i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





# Cuarteto de regresiones “idénticas” de Anscombe (1973)

Anscombe's 4 Regression data sets



## Diagnóstico de influencia en regresión lineal

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

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

Cook (1977)<sup>1</sup> propuso **determinar la influencia** de la  $i$ -ésima observación, usando

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta}_{(i)} - \hat{\beta})}{ps^2} = \frac{r_i^2}{p} \left( \frac{h_i}{1 - h_i} \right),$$

para  $i = 1, \dots, n$ , 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**.

---

<sup>1</sup>Technometrics 19, 15-18



## Diagnóstico de influencia en regresión lineal

Welsch y Kuh (1977)<sup>2</sup> propusieron medir el impacto en la  $i$ -ésima observación sobre el valor predicho como

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

y su versión estandarizada

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

Belsley, Kuh y Welsch (1980)<sup>3</sup> 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**.

---

<sup>2</sup>Working paper No. 173, National Bureau of Economics, Cambridge

<sup>3</sup>Regression Diagnostics. Wiley, New York



## Diagnóstico de influencia en regresión lineal

Atkinson (1981)<sup>4</sup> sugirió usar una versión modificada de la distancia de Cook, como

$$\begin{aligned} AK_i &= \sqrt{\left(\frac{n-p}{p}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right)} \left| \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}} \right| \\ &= \sqrt{\left(\frac{n-p}{p}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right)} |t_i| \\ &= \sqrt{\frac{n-p}{p}} |DFFITS_i|. \end{aligned}$$

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  $\hat{\beta}$  y  $s^2$  simultáneamente.

---

<sup>4</sup>Biometrika 68, 13-20



# Diagnóstico de influencia en regresión lineal

Cook y Weisberg (1980)<sup>5</sup> propusieron considerar medidas generales de influencia<sup>6</sup> considerando,

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

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

Algunas medidas de influencia:

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

---

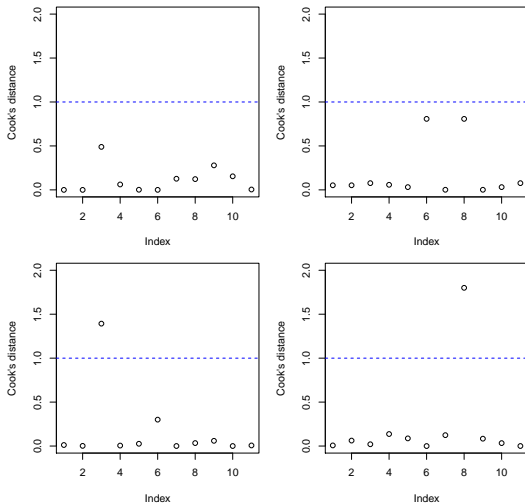
<sup>5</sup>Technometrics 22, 495-508

<sup>6</sup>Basadas en la función de influencia empírica.



# Distancia de Cook: Datos de Anscombe (1973)

Cook's distance plots. Anscombe's data sets



## Diagnóstico de influencia en regresión lineal

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

$$\begin{aligned} COVRATIO_i &= \frac{\det\{s_{(i)}^2(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1}\}}{\det\{s^2(\mathbf{X}^\top \mathbf{X})^{-1}\}} = \left(\frac{s_{(i)}^2}{s^2}\right)^p \frac{\det(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1}}{\det(\mathbf{X}^\top \mathbf{X})^{-1}} \\ &= \frac{1}{1 - h_i} \left(\frac{n - p - r_i^2}{n - p - 1}\right)^p, \end{aligned}$$

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
$D_i = \frac{r_i^2}{p} \left( \frac{h_i}{1-h_i} \right)$	$F_{p, n-p}(1-\alpha)$
$DFFITS_i =  t_i  \sqrt{\frac{h_i}{1-h_i}}$	$2\sqrt{p/n}$
$AK_i = DFFITS_i \sqrt{\frac{n-p}{p}}$	$2\sqrt{(n-p)/n}$
$W_i = DFFITS_i \sqrt{\frac{n-1}{1-h_i}}$	$3\sqrt{p}$
$COVRATIO_i = \frac{1}{1-h_i} \left( \frac{n-p-r_i^2}{n-p-1} \right)^p$	$ COVRATIO_i - 1  > 3p/n$
$h_i = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i$	$2p/n$
$r_i = \frac{e_i}{s\sqrt{1-h_i}}$	$\approx N(0, 1)$
$t_i = r_i \sqrt{\frac{n-p-1}{n-p-r_i^2}}$	$\approx t(n-p-1)$





Belsley, Kuh y Welsch (1981) y Velleman y Welsch (1981)<sup>7</sup> 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:

<b>rstandard</b>	<b>rstudent</b>	<b>dffits</b>	<b>dfbetas</b>
<b>covratio</b>	<b>cooks.distance</b>	<b>hatvalues</b>	

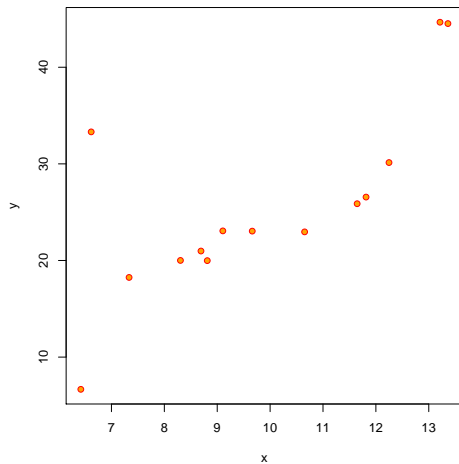
Estas cantidades **pueden ser escritas de forma eficiente** usando la descomposición QR o SVD.

---

<sup>7</sup>The American Statistician 35, 234-242.



# Influencia de múltiples casos



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

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

Usando que

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

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

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

*Observación:*

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



Múltiples datos atípicos pueden ser detectados mediante el modelo

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

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

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

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



- ▶ 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 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).
- ▶ Colinealidad **inducida** por outliers  
(Mason y Gunst, 1985; Hadi, 1988 y Walker, 1989).



Walker y Birch (1988)<sup>8</sup>, propusieron las siguientes medidas de diagnóstico:

- Leverage:

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

- Distancia de Cook:

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

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

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

---

<sup>8</sup>Technometrics 30, 221-227.

Hadi (1988)<sup>9</sup> y Walker (1989)<sup>10</sup>, proponen detectar observaciones que inducen o ocultan colinealidad, mediante:

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

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

### Concluyeron que:

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

---

<sup>9</sup>Computational Statistics & Data Analysis 7, 143-159

<sup>10</sup>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  $\kappa(\mathbf{X}) = 9.432$ , esto es,  $\mathbf{X}$  es bien condicionada (variables centradas  $\kappa(\widetilde{\mathbf{X}}) = 37.106$ ).

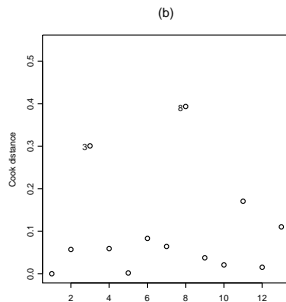
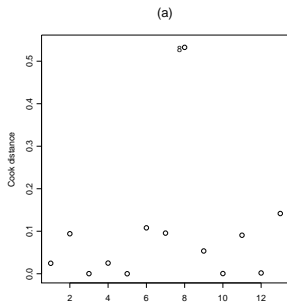
Por otro lado, Hald (1952), Gorman y Toman (1966) y Daniel y Wood (1980) adoptan un modelo **con intercepto (modelo no homogéneo)**. En cuyo caso  $\kappa(\mathbf{X}) = 249.578$ , sugiriendo la presencia de colinealidad. El aumento en el número condición se debe a que existe una relación lineal aproximada,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

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

de modo que incluir el intercepto causa una colinealidad severa.

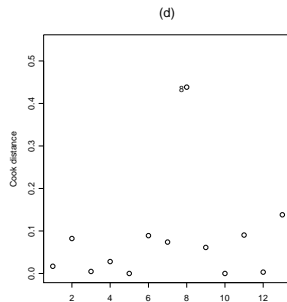
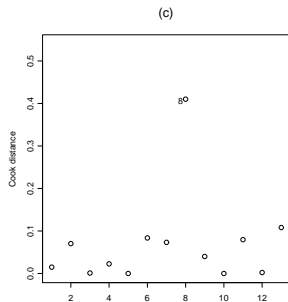


Distancia de Cook:



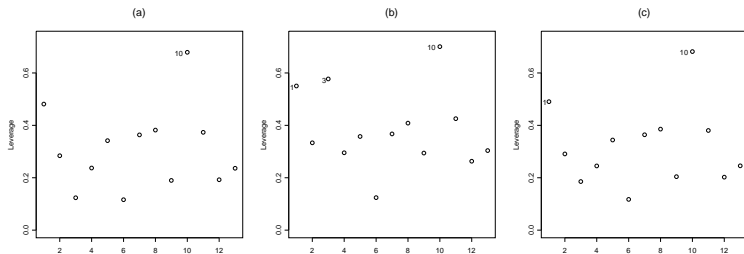
Estimación mínimos cuadrados: Modelo homogéneo (a) y no homogéneo (b).

Distancia de Cook:



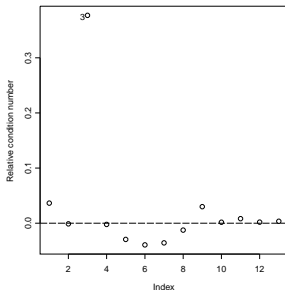
Estimación ridge:  $D_i^*$  (c) y  $D_i^{**}$  (d).

Leverage:



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

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



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