Estimaciones inestables y ausencia de normalidad en el modelo de regresión lineal

Román Salmerón Gómez
romansg@ugr.es
Víctor Blanco Izquierdo
vblanco@ugr.es

Universidad de Granada
Departamento de Métodos Cuantitativos para la Economía y la Empresa

I Jornadas de Análisis Cuantitativo en Economía
9 y 10 de Mayo 2018

Índice

Introducción

Econometría

Econometría es una rama de la Economía que proporciona una base para refinar o refutar conocimiento teórico y conseguir signos y magnitudes de las relaciones de variables que se desean analizar.

Con tal objetivo se han de estimar los coeficientes del modelo lineal general:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \cdots + \beta_p X_p + u = X\beta + u \qquad (1)$$

Aplicando MCO se llega al sistema de ecuaciones normales:

$$(X'X)\,\beta = X'Y$$

Para que tenga solución única debe existir $(X'X)^{-1}$:

$$\widehat{\beta} = (X'X)^{-1} X'Y$$

Si las variables que forman X son linealmente dependientes, entonces no existe $(X'X)^{-1}$ (multicolinealidad perfecta).

Introducción

¿Qué ocurre si son casi linealmente independientes? (multicolinealidad aproximada): resultados inestables y contradictorios.

- Pequeños cambios en los datos pueden suponer cambios sustanciales en las estimaciones de los coeficientes de los regresores.
- Tendencia a no rechazar que los coeficientes de los regresores son cero.
- Coeficiente de determinación alto y, en consecuencia, tendencia a rechazar que todos los coeficientes son cero de forma simultánea.

Posibles soluciones:

- Mejora del diseño muestral, aumento del tamaño de la muestra o usar información a priori.
- Eliminar variables que se consideran problemáticas.
- Usar métodos de estimación alternativos a MCO.

Índice

Estimaciones Inestables: simulación de datos

Inicialmente genero 50 observaciones para $X \sim N(1,100)$ y $Z \sim N(1,100)$ ($corr(X,Z) = -0{,}1186327$).
A continuación, genero Z^* perturbando ligeramente Z: $Z^*(i) = Z(i) + 0{,}5 \cdot (-1)^i$.
Finalmente, genero la variable dependiente como $Y = 5 + 2 \cdot X - 4 \cdot Z + u$ donde $u \sim N(0,4)$.

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z + u \quad \rightarrow \quad \widehat{\beta}_0 = 4{,}9451, \widehat{\beta}_1 = 1{,}9997, \widehat{\beta}_2 = -3{,}9761$$
$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z^* + u \quad \rightarrow \quad \widehat{\beta}_0 = 4{,}8872, \widehat{\beta}_1 = 1{,}9485, \widehat{\beta}_2 = -3{,}9491$$

Estimaciones Inestables: simulación de datos

Inicialmente genero 50 observaciones para $X \sim N(1,100)$ y $Z \sim N(1,100)$ ($corr(X,Z) = -0{,}1186327$).
A continuación, genero Z^* perturbando ligeramente Z: $Z^*(i) = Z(i) + 0{,}5 \cdot (-1)^i$.
Finalmente, genero la variable dependiente como $Y = 5 + 2 \cdot X - 4 \cdot Z + u$ donde $u \sim N(0,4)$.

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z + u \quad \rightarrow \quad \widehat{\beta}_0 = 4{,}9451, \widehat{\beta}_1 = 1{,}9997, \widehat{\beta}_2 = -3{,}9761$$
$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z^* + u \quad \rightarrow \quad \widehat{\beta}_0 = 4{,}8872, \widehat{\beta}_1 = 1{,}9485, \widehat{\beta}_2 = -3{,}9491$$

Si en cambio genero Z como $Z(i) = 1 + 5 \cdot X(i) + (-1)^i$ ($corr(X,Z) = 0{,}9998075$) y lo demás igual:

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z + u \quad \rightarrow \quad \widehat{\beta}_0 = 5{,}2036, \widehat{\beta}_1 = 3{,}5849, \widehat{\beta}_2 = -4{,}3172$$
$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z^* + u \quad \rightarrow \quad \widehat{\beta}_0 = 3{,}7645, \widehat{\beta}_1 = -3{,}6104, \widehat{\beta}_2 = -2{,}9782$$

Coeficientes significativos, modelo globalmente válido y $R^2 = 0{,}9999$.

Estimaciones Inestables: regresión simple

Las estimaciones inestables se deben a la relación lineal entre X y Z.
Solución: la regresión simple, ya que mide la relación entre variables en ausencia de multicolinealidad (esencial).

Estimaciones Inestables: regresión simple

Las estimaciones inestables se deben a la relación lineal entre X y Z.
Solución: la regresión simple, ya que mide la relación entre variables en ausencia de multicolinealidad (esencial).
Un poquito de teoría, si estandarizamos las variables:

$$y = \beta_1 x + \beta_2 z + u, \quad \widehat{\beta}_1 = \frac{\gamma_1 - \rho\gamma_2}{1-\rho^2}, \quad \widehat{\beta}_2 = \frac{\gamma_2 - \rho\gamma_1}{1-\rho^2},$$

donde $\rho = corr(x, z)$, $\gamma_1 = corr(x, y)$, $\gamma_2 = corr(z, y)$. Además:

$$y = \alpha_1 \cdot x + v \to \widehat{\alpha}_1 = \gamma_1, \quad y = \delta_1 \cdot z + w \to \widehat{\delta}_1 = \gamma_2$$

Estimaciones Inestables: regresión simple

Las estimaciones inestables se deben a la relación lineal entre X y Z.
Solución: la regresión simple, ya que mide la relación entre variables en ausencia de multicolinealidad (esencial).
Un poquito de teoría, si estandarizamos las variables:

$$y = \beta_1 x + \beta_2 z + u, \quad \widehat{\beta}_1 = \frac{\gamma_1 - \rho\gamma_2}{1-\rho^2}, \quad \widehat{\beta}_2 = \frac{\gamma_2 - \rho\gamma_1}{1-\rho^2},$$

donde $\rho = corr(x,z)$, $\gamma_1 = corr(x,y)$, $\gamma_2 = corr(z,y)$. Además:

$$y = \alpha_1 \cdot x + v \rightarrow \widehat{\alpha}_1 = \gamma_1, \quad y = \delta_1 \cdot z + w \rightarrow \widehat{\delta}_1 = \gamma_2$$

En nuestros ejemplos ($Y = 5 + 2 \cdot X - 4 \cdot Z + u$):

$Y = \alpha_0 + \alpha_1 \cdot X + v$	$\rightarrow$	$\widehat{\alpha}_1 = 2{,}415$	$Y = \alpha_0 + \alpha_1 \cdot X + v$	$\rightarrow$	$\widehat{\alpha}_1 = -17{,}9728$
$Y = \delta_0 + \delta_1 \cdot Z + w$	$\rightarrow$	$\widehat{\delta}_1 = -4{,}2456$	$Y = \delta_0 + \delta_1 \cdot Z + w$	$\rightarrow$	$\widehat{\delta}_1 = -3{,}5995$
$Y = \delta_0 + \delta_1 \cdot Z^* + w$	$\rightarrow$	$\widehat{\delta}_1 = -4{,}242$	$Y = \delta_0 + \delta_1 \cdot Z^* + w$	$\rightarrow$	$\widehat{\delta}_1 = -3{,}601$

-17.9728 marca la relación lineal existente entre X e Y libre de la relación lineal entre X y Z o Z^*

Estimaciones Inestables: alternativa

Si me empeño en la regresión múltiple, una alternativa a los MCO es hacer que las estimaciones de $Y = X\beta + u$ se parezcan a las obtenidas a partir de las regresiones auxiliares $Y = \alpha_0 + \alpha_1 X_i + v$:

$$min\ c_1 \cdot SCR + c_2 \cdot ||\widehat{\beta}_{-0} - \widehat{\alpha}||^2, \quad c_1, c_2 \geq 0$$

donde:

- $SCR = Y^t Y - 2 \cdot \widehat{\beta}^t X^t Y + \widehat{\beta}^t X^t X \widehat{\beta}$ donde $\beta = (\beta_0\ \beta_1\ \ldots\ \beta_p)^t = (\beta_0\ \beta_{-0})^t$,
- $||\widehat{\beta}_{-0} - \widehat{\alpha}||^2 = \widehat{\beta}^t_{-0}\widehat{\beta}_{-0} - 2 \cdot \widehat{\beta}^t_{-0}\widehat{\alpha} + \widehat{\alpha}^t\widehat{\alpha}$ siendo $\widehat{\alpha} = (\widehat{\alpha}_1\ \ldots\ \widehat{\alpha}_p)$ con $\widehat{\alpha}_i = \frac{cov(X_i, Y)}{var(X_i)}$.

Estimaciones Inestables: alternativa

Si me empeño en la regresión múltiple, una alternativa a los MCO es hacer que las estimaciones de $Y = X\beta + u$ se parezcan a las obtenidas a partir de las regresiones auxiliares $Y = \alpha_0 + \alpha_1 X_i + v$:

$$min\ c_1 \cdot SCR + c_2 \cdot ||\widehat{\beta}_{-0} - \widehat{\alpha}||^2, \quad c_1, c_2 \geq 0$$

donde:

- $SCR = Y^t Y - 2 \cdot \widehat{\beta}^t X^t Y + \widehat{\beta}^t X^t X \widehat{\beta}$ donde $\beta = (\beta_0\ \beta_1\ \ldots\ \beta_p)^t = (\beta_0\ \beta_{-0})^t$,
- $||\widehat{\beta}_{-0} - \widehat{\alpha}||^2 = \widehat{\beta}^t_{-0}\widehat{\beta}_{-0} - 2 \cdot \widehat{\beta}^t_{-0}\widehat{\alpha} + \widehat{\alpha}^t\widehat{\alpha}$ siendo $\widehat{\alpha} = (\widehat{\alpha}_1\ \ldots\ \widehat{\alpha}_p)$ con $\widehat{\alpha}_i = \frac{cov(X_i, Y)}{var(X_i)}$.

Problema: derivamos respecto a $\widehat{\beta}$ y tenemos $\widehat{\beta}_{-0}$.

Estimaciones Inestables: alternativa

Si me empeño en la regresión múltiple, una alternativa a los MCO es hacer que las estimaciones de $Y = X\beta + u$ se parezcan a las obtenidas a partir de las regresiones auxiliares $Y = \alpha_0 + \alpha_1 X_i + v$:

$$min\ c_1 \cdot SCR + c_2 \cdot ||\widehat{\beta}_{-0} - \widehat{\alpha}||^2, \quad c_1, c_2 \geq 0$$

donde:

- $SCR = Y^t Y - 2 \cdot \widehat{\beta}^t X^t Y + \widehat{\beta}^t X^t X \widehat{\beta}$ donde $\beta = (\beta_0\ \beta_1\ \dots\ \beta_p)^t = (\beta_0\ \beta_{-0})^t$,
- $||\widehat{\beta}_{-0} - \widehat{\alpha}||^2 = \widehat{\beta}^t_{-0}\widehat{\beta}_{-0} - 2 \cdot \widehat{\beta}^t_{-0}\widehat{\alpha} + \widehat{\alpha}^t\widehat{\alpha}$ siendo $\widehat{\alpha} = (\widehat{\alpha}_1\ \dots\ \widehat{\alpha}_p)$ con $\widehat{\alpha}_i = \frac{cov(X_i, Y)}{var(X_i)}$.

Problema: derivamos respecto a $\widehat{\beta}$ y tenemos $\widehat{\beta}_{-0}$.
Solución: consideramos el modelo en desviaciones:

$$y = \beta_1 x_1 + \cdots + \beta_p x_p + u = x \cdot \beta_{-0} + u, \quad y = Y - \overline{Y},\ x_i = X_i - \overline{X}$$

Objetivo:

$$min\ c_1 \cdot SCR_{-0} + c_2 \cdot ||\widehat{\beta}_{-0} - \widehat{\alpha}||^2, \quad c_1, c_2 \geq 0$$

donde $SCR_{-0} = Y^t Y - 2 \cdot \widehat{\beta}^t_{-0} X^t Y + \widehat{\beta}^t_{-0} X^t X \widehat{\beta}_{-0}$

Estimaciones Inestables: alternativa

Derivando respecto a $\widehat{\beta}_{-0}$:

$$\widehat{\beta}_{-0}(c_1, c_2) = \left(c_1 \cdot x^t x + c_2 I_p\right)^{-1} \cdot (c_1 \cdot I_p + c_2 \cdot V) \cdot x^t y$$

donde $V = diag(\sum x_{1t}^2 \cdots \sum x_{pt}^2)^{-1}$.

Estimamos la constante: $\widehat{\beta}_0 = \overline{Y} - \sum_{i=1}^{p} \widehat{\beta}_{-0}(c_1, c_2)(i) \cdot \overline{X}_i$

Estimaciones Inestables: alternativa

Derivando respecto a $\widehat{\beta}_{-0}$:

$$\widehat{\beta}_{-0}(c_1, c_2) = \left(c_1 \cdot x^t x + c_2 I_p\right)^{-1} \cdot (c_1 \cdot I_p + c_2 \cdot V) \cdot x^t y$$

donde $V = diag(\sum x_{1t}^2 \cdots \sum x_{pt}^2)^{-1}$.

Estimamos la constante: $\widehat{\beta_0} = \overline{Y} - \sum_{i=1}^{p} \widehat{\beta}_{-0}(c_1, c_2)(i) \cdot \overline{X}_i$

$c_1 = 1, c_2 = 0$ tenemos MCO $\qquad$ $c_1 = 0, c_2 = 1$ tenemos que $\widehat{\beta}_{-0}(0, 1) = \widehat{\alpha}$

Estimaciones Inestables: alternativa

Derivando respecto a $\widehat{\beta}_{-0}$:

$$\widehat{\beta}_{-0}(c_1, c_2) = \left(c_1 \cdot x^t x + c_2 I_p\right)^{-1} \cdot \left(c_1 \cdot I_p + c_2 \cdot V\right) \cdot x^t y$$

donde $V = diag(\sum x_{1t}^2 \cdots \sum x_{pt}^2)^{-1}$.

Estimamos la constante: $\widehat{\beta}_0 = \overline{Y} - \sum_{i=1}^{p} \widehat{\beta}_{-0}(c_1, c_2)(i) \cdot \overline{X}_i$

$c_1 = 1, c_2 = 0$ tenemos MCO $\qquad$ $c_1 = 0, c_2 = 1$ tenemos que $\widehat{\beta}_{-0}(0,1) = \widehat{\alpha}$

Similitud con el estimador cresta:

$$\begin{aligned} \widehat{\beta}_{-0}(c) &= \left(x^t x + c \cdot I_p\right)^{-1} \cdot \left(I_p + c \cdot V\right) \cdot x^t y, \quad c = \frac{c_2}{c_1},\ c_1 > 0 \\ \widehat{\beta}(k) &= \left(x^t x + k \cdot I_p\right)^{-1} x^t y, \quad k \geq 0 \end{aligned}$$

Estimaciones Inestables: alternativa

Derivando respecto a $\widehat{\beta}_{-0}$:

$$\widehat{\beta}_{-0}(c_1, c_2) = \left(c_1 \cdot x^t x + c_2 I_p\right)^{-1} \cdot \left(c_1 \cdot I_p + c_2 \cdot V\right) \cdot x^t y$$

donde $V = diag(\sum x_{1t}^2 \cdots \sum x_{pt}^2)^{-1}$.

Estimamos la constante: $\widehat{\beta}_0 = \overline{Y} - \sum_{i=1}^{p} \widehat{\beta}_{-0}(c_1, c_2)(i) \cdot \overline{X}_i$

$c_1 = 1, c_2 = 0$ tenemos MCO $\qquad$ $c_1 = 0, c_2 = 1$ tenemos que $\widehat{\beta}_{-0}(0,1) = \widehat{\alpha}$

Similitud con el estimador cresta:

$$\begin{aligned} \widehat{\beta}_{-0}(c) &= \left(x^t x + c \cdot I_p\right)^{-1} \cdot \left(I_p + c \cdot V\right) \cdot x^t y, \quad c = \frac{c_2}{c_1},\ c_1 > 0 \\ \widehat{\beta}(k) &= \left(x^t x + k \cdot I_p\right)^{-1} x^t y, \quad k \geq 0 \end{aligned}$$

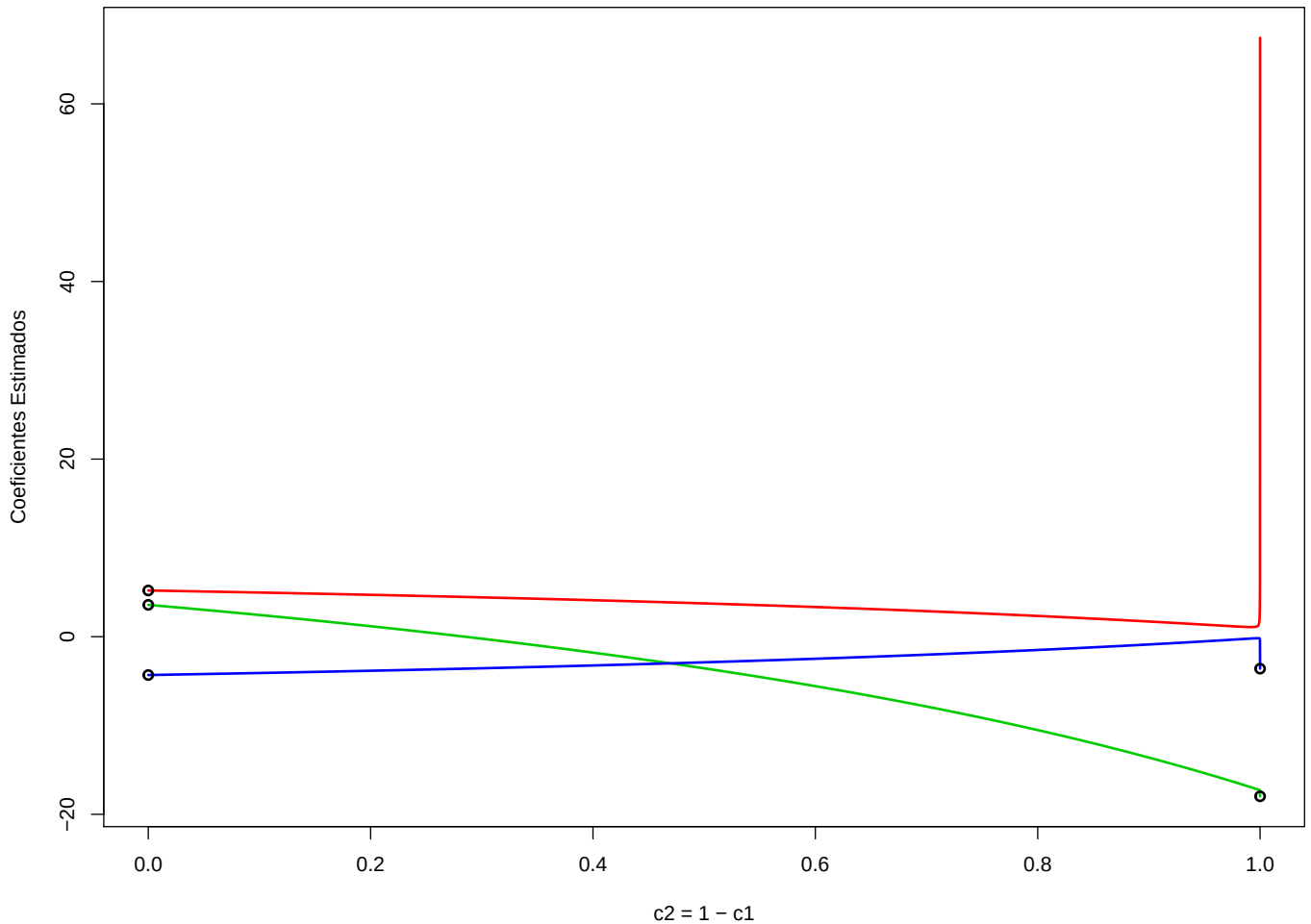
¿Cómo determinar los valores de c_1 y c_2?

Para simplificar interpretación $c_1 > 0$ y $c_1 + c_2 = 1 \rightarrow c_2 = 1 - c_1 < 1$

$c_1 = 0{,}2, c_2 = 0{,}8$: 20 % de importancia a SCR_{-0} y un 80 % a $||\widehat{\beta}_{-0} - \widehat{\alpha}||^2$

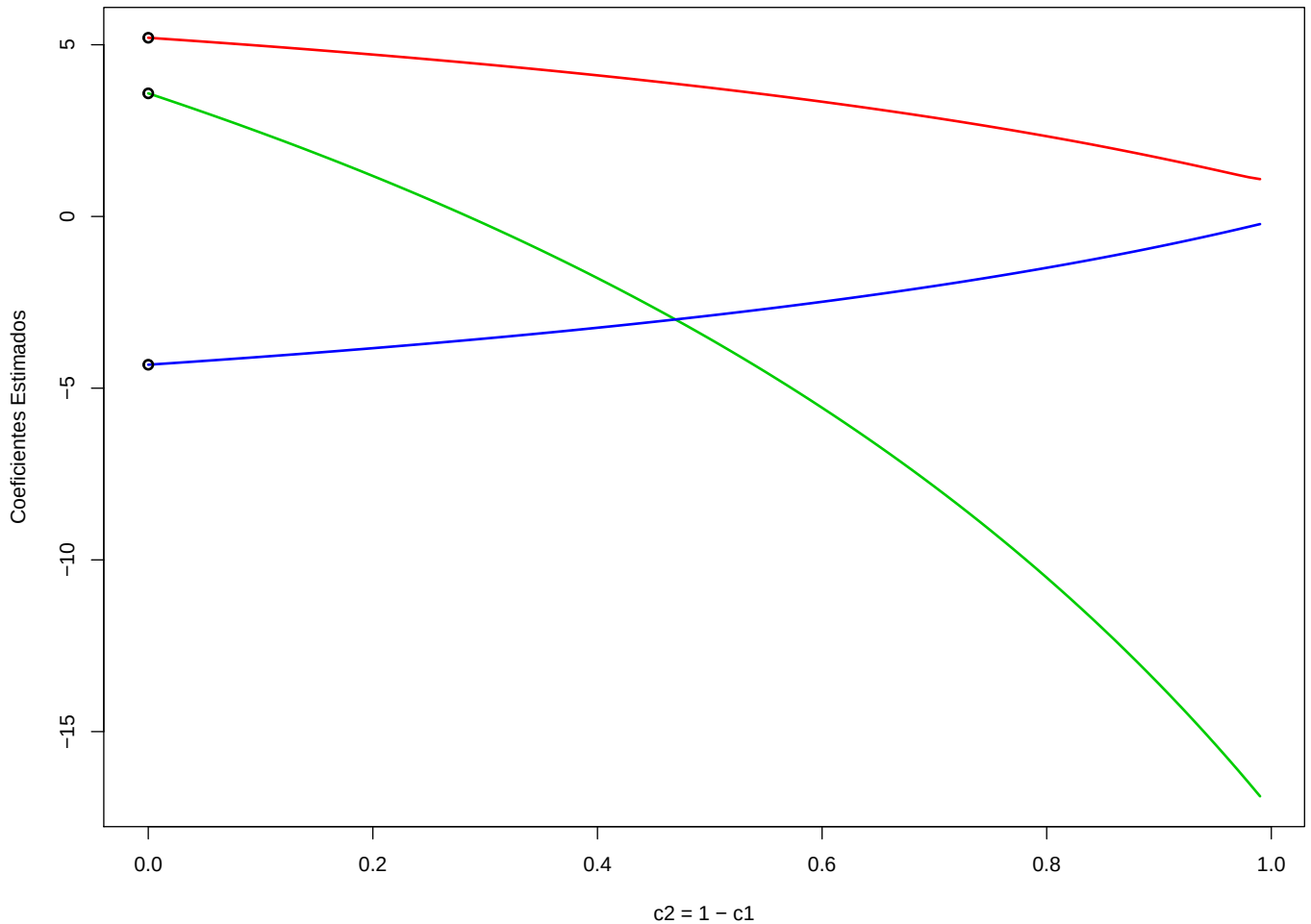
Estimaciones Inestables: elección de c_1 y c_2

Traza de los estimadores del ejemplo problemático: *cte* rojo, X verde, Z azul. ¿Las estimaciones se estabilizan para algún valor de c_1 y c_2?



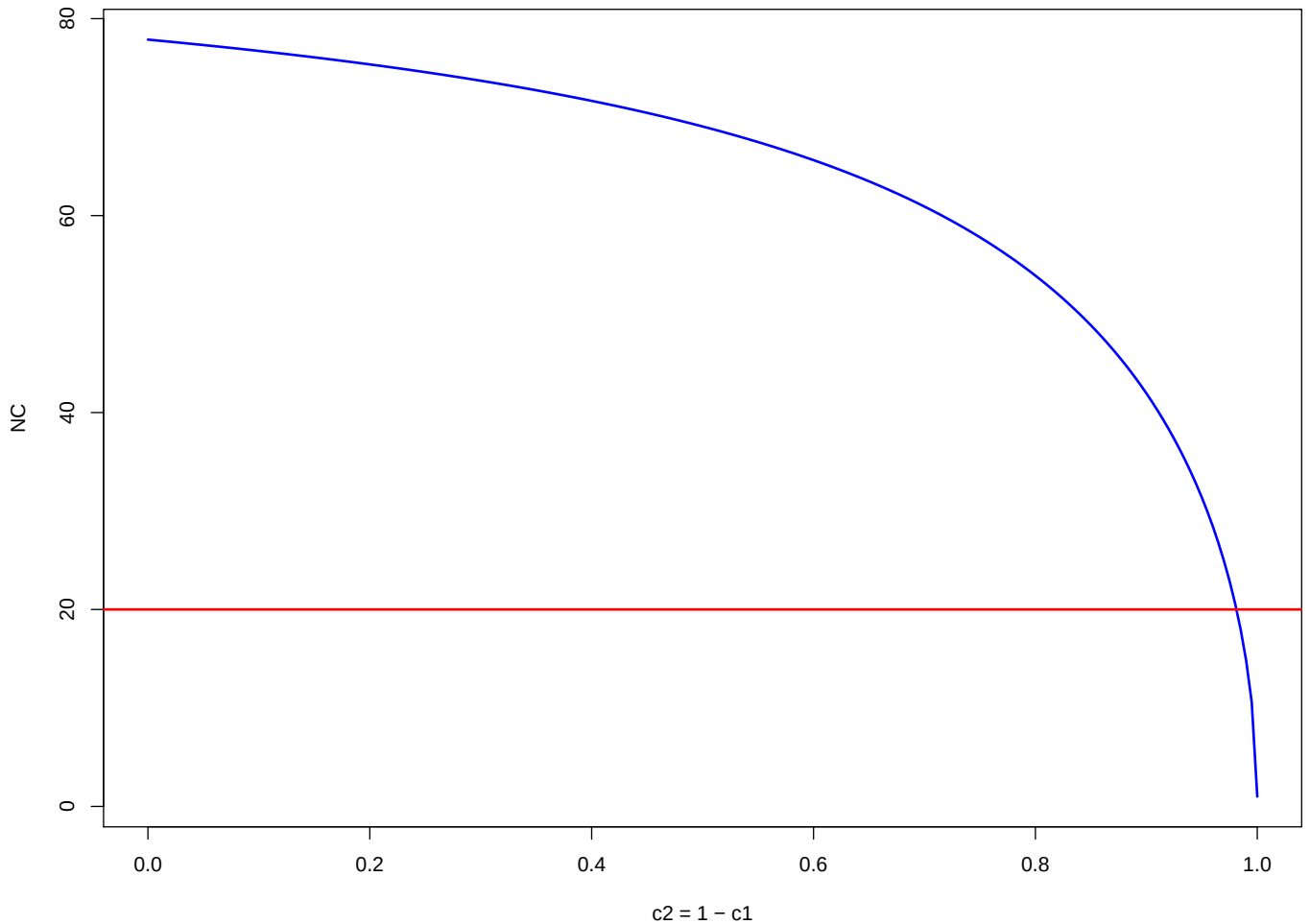
Estimaciones Inestables: elección de c_1 y c_2

Traza de los estimadores del ejemplo problemático: *cte* rojo, X verde, Z azul. ¿Las estimaciones se estabilizan para algún valor de c_1 y c_2?



Estimaciones Inestables: elección de c_1 y c_2

Analizar condicionamiento de la matriz $c_1 \cdot x^t x + c_2 I_p$. El número de condición es menor que 20 para $c_1 = 0{,}015$ y $c_2 = 0{,}985$.



Estimaciones Inestables: $c_1 = 0{,}015$ y $c_2 = 0{,}985$

Ejemplo problemático:

$$\begin{aligned} Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z + u \quad &\rightarrow \quad \widehat{\beta}_0 = 1{,}1106, \widehat{\beta}_1 = -16{,}6883, \\ &\quad\ \widehat{\beta}_2 = -0{,}2605, \\ Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z^* + u \quad &\rightarrow \quad \widehat{\beta}_0 = 1{,}177855, \widehat{\beta}_1 = -16{,}436553, \\ &\quad\ \widehat{\beta}_2 = -0{,}311766, \end{aligned}$$

Estimaciones Inestables: $c_1 = 0{,}015$ y $c_2 = 0{,}985$

Ejemplo problemático:

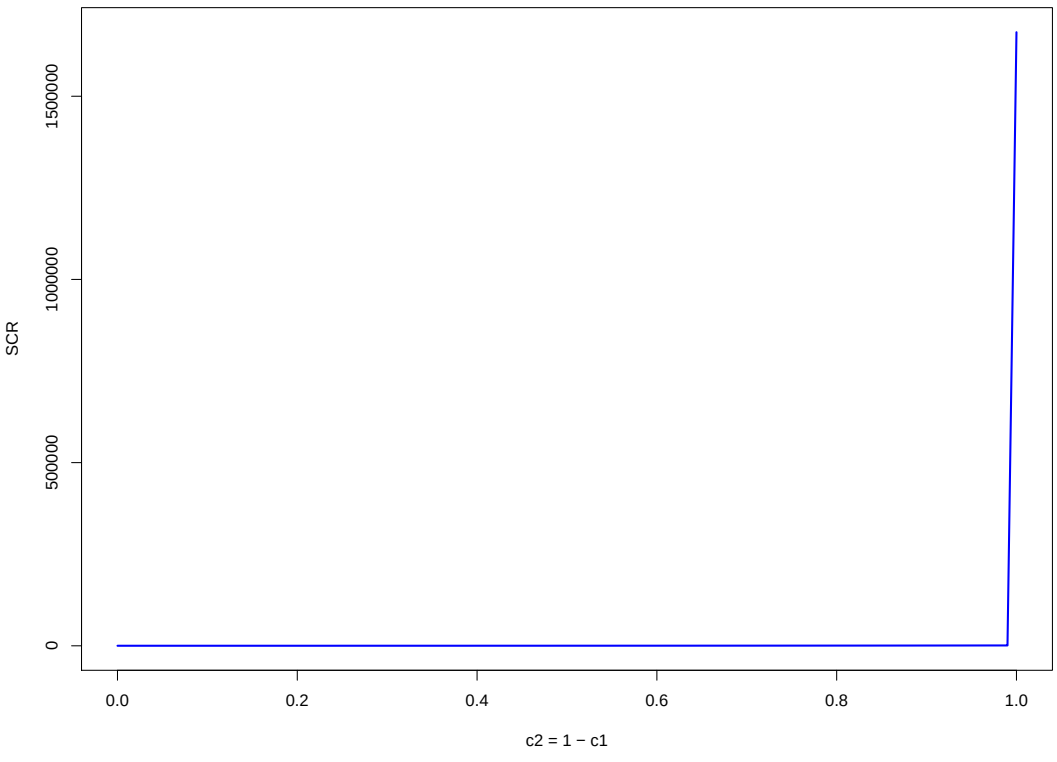
$$\begin{aligned} Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z + u \quad &\rightarrow \quad \widehat{\beta}_0 = 1{,}1106, \widehat{\beta}_1 = -16{,}6883, \\ &\quad \widehat{\beta}_2 = -0{,}2605, \ R^2 = 0{,}9994121 \\ Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z^* + u \quad &\rightarrow \quad \widehat{\beta}_0 = 1{,}177855, \widehat{\beta}_1 = -16{,}436553, \\ &\quad \widehat{\beta}_2 = -0{,}311766, \ R^2 = 0{,}9994239 \end{aligned}$$

Bondad de ajuste:

$$R^2 = 1 - \frac{(y - x \cdot \widehat{\beta}(c_1, c_2))^t \cdot (y - x \cdot \widehat{\beta}(c_1, c_2))}{SCT}$$

¡Ojo! no tiene por qué verificarse que $R^2 > 0$

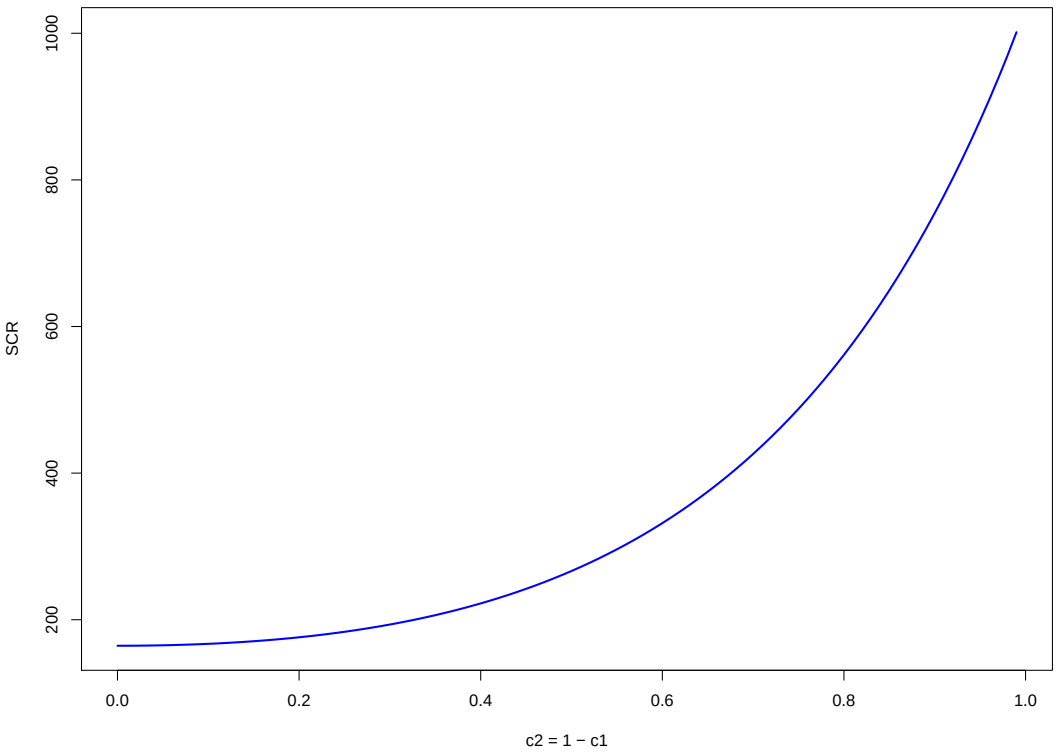
Estimaciones Inestables: bondad de ajuste (SCR)



$c_2 = 1 - c_1$	SCR	$\|\|\widehat{\beta}_{-0} - \widehat{\alpha}\|\|$
0	164.5073	21.569
0.1	167.1106	20.421
0.2	176.0527	19.153
0.3	193.4705	17.748
0.4	222.2768	16.181
0.5	266.4943	14.426
0.6	331.7636	12.452
0.7	426.1329	10.229
0.8	561.3199	7.747
0.9	754.7928	5.141
1	1674812.7258	$3.66 \cdot 10^{-15}$

$SCT = 1675458$

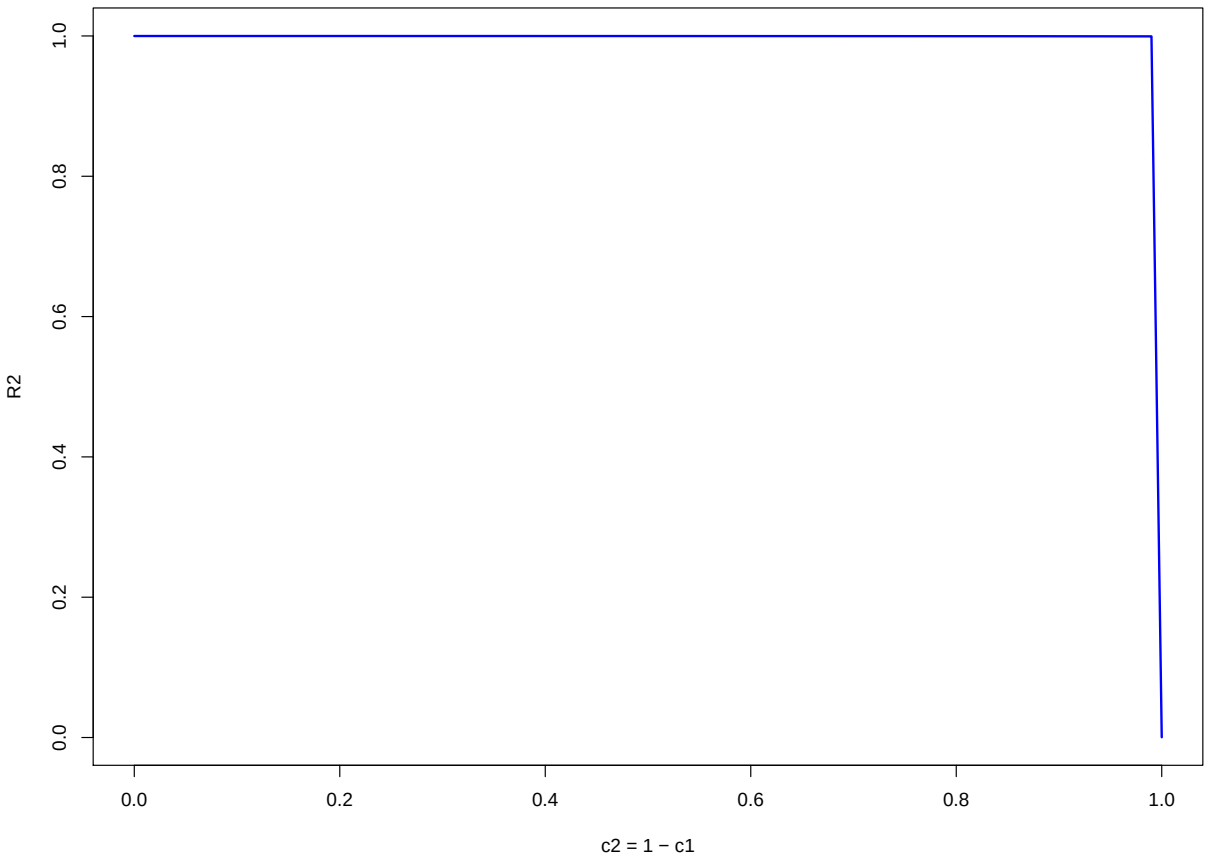
Estimaciones Inestables: bondad de ajuste (SCR)



$c_2 = 1 - c_1$	SCR	$\|\|\widehat{\beta}_{-0} - \widehat{\alpha}\|\|$
0	164.5073	21.569
0.1	167.1106	20.421
0.2	176.0527	19.153
0.3	193.4705	17.748
0.4	222.2768	16.181
0.5	266.4943	14.426
0.6	331.7636	12.452
0.7	426.1329	10.229
0.8	561.3199	7.747
0.9	754.7928	5.141
1	1674812.7258	$3.66 \cdot 10^{-15}$

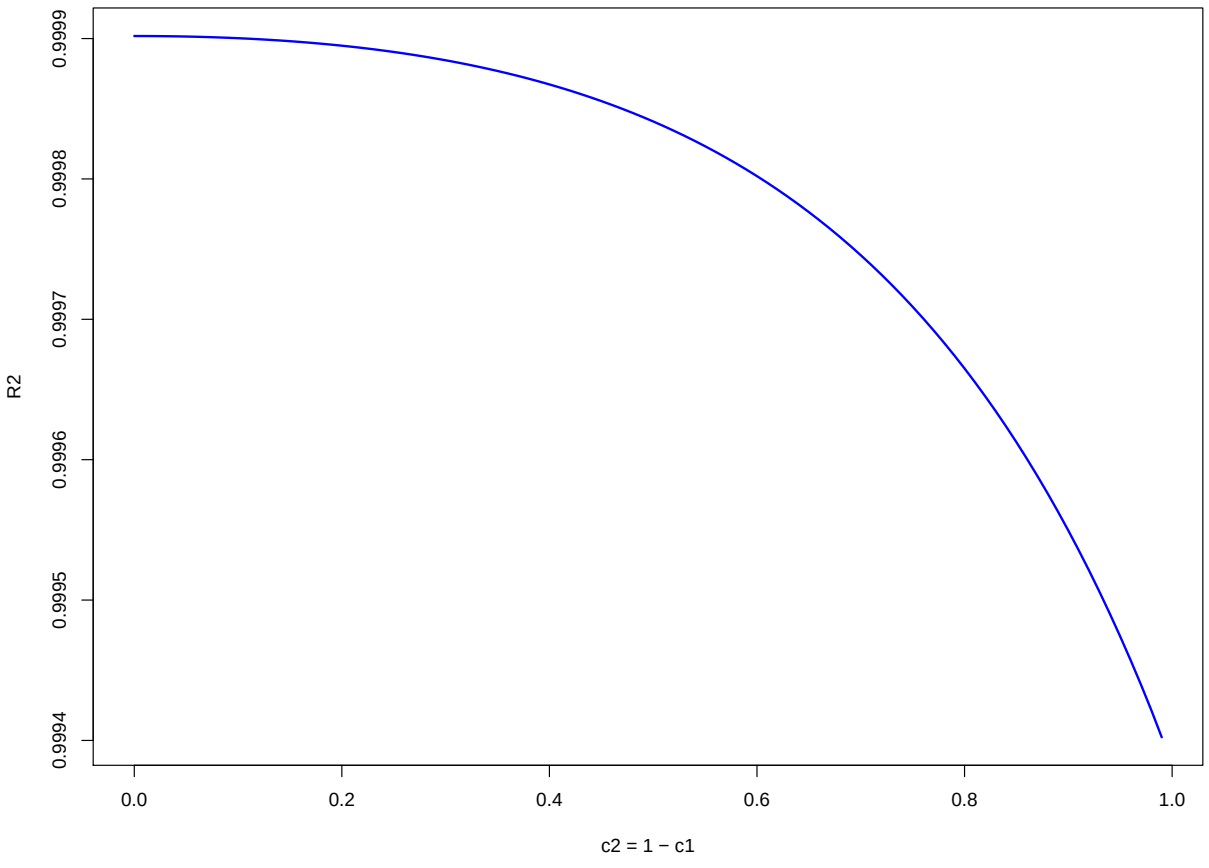
$SCT = 1675458$

Estimaciones Inestables: bondad de ajuste (R^2)



$c_2 = 1 - c_1$	R^2
0	0.9999018135
0.1	0.9999002597
0.2	0.9998949226
0.3	0.9998845268
0.4	0.9998673337
0.5	0.9998409424
0.6	0.9998019863
0.7	0.9997456618
0.8	0.9996649752
0.9	0.9995495006
1	0.0003849483

Estimaciones Inestables: bondad de ajuste (R^2)



$c_2 = 1 - c_1$	R^2
0	0.9999018135
0.1	0.9999002597
0.2	0.9998949226
0.3	0.9998845268
0.4	0.9998673337
0.5	0.9998409424
0.6	0.9998019863
0.7	0.9997456618
0.8	0.9996649752
0.9	0.9995495006
1	0.0003849483

Índice

Hipótesis de normalidad

Llamando $H = \left(c_1 \cdot x^t x + c_2 I_p\right)^{-1} \cdot (c_1 \cdot I_p + c_2 \cdot V)$ se verifica que:

$$\begin{aligned}
\widehat{\beta}_{-0}(c_1, c_2) &= H \cdot x^t y = H \cdot x^t x \cdot \widehat{\beta} \\
&\hookrightarrow E[\widehat{\beta}_{-0}(c_1, c_2)] = H \cdot x^t x \cdot \beta \\
&\hookrightarrow var(\widehat{\beta}_{-0}(c_1, c_2)) = \sigma^2 \cdot H \cdot x^t x \cdot H^t \\
\widehat{\beta}_{-0}(c_1, c_2) &= H \cdot x^t y = H \cdot x^t x \cdot \beta + H \cdot x^t \cdot u
\end{aligned}$$

Hipótesis de normalidad

Llamando $H = \left(c_1 \cdot x^t x + c_2 I_p\right)^{-1} \cdot (c_1 \cdot I_p + c_2 \cdot V)$ se verifica que:

$$\begin{aligned}
\widehat{\beta}_{-0}(c_1, c_2) &= H \cdot x^t y = H \cdot x^t x \cdot \widehat{\beta} \\
&\hookrightarrow E[\widehat{\beta}_{-0}(c_1, c_2)] = H \cdot x^t x \cdot \beta \\
&\hookrightarrow var(\widehat{\beta}_{-0}(c_1, c_2)) = \sigma^2 \cdot H \cdot x^t x \cdot H^t \\
\widehat{\beta}_{-0}(c_1, c_2) &= H \cdot x^t y = H \cdot x^t x \cdot \beta + H \cdot x^t \cdot u
\end{aligned}$$

Considerando que $u \sim N(0, \sigma^2 I)$:

$$\widehat{\beta}_{-0}(c_1, c_2) \sim N\left(H \cdot x^t x \cdot \beta, \sigma^2 \cdot H \cdot x^t x \cdot H^t\right)$$

Ahora bien, no es válido para hacer inferencia ya que depende del parámetro desconocido β.

Hipótesis de normalidad

Llamando $H = \left(c_1 \cdot x^t x + c_2 I_p\right)^{-1} \cdot (c_1 \cdot I_p + c_2 \cdot V)$ se verifica que:

$$\begin{aligned}
\widehat{\beta}_{-0}(c_1, c_2) &= H \cdot x^t y = H \cdot x^t x \cdot \widehat{\beta} \\
&\hookrightarrow E[\widehat{\beta}_{-0}(c_1, c_2)] = H \cdot x^t x \cdot \beta \\
&\hookrightarrow var(\widehat{\beta}_{-0}(c_1, c_2)) = \sigma^2 \cdot H \cdot x^t x \cdot H^t \\
\widehat{\beta}_{-0}(c_1, c_2) &= H \cdot x^t y = H \cdot x^t x \cdot \beta + H \cdot x^t \cdot u
\end{aligned}$$

Considerando que $u \sim N(0, \sigma^2 I)$:

$$\widehat{\beta}_{-0}(c_1, c_2) \sim N\left(H \cdot x^t x \cdot \beta, \sigma^2 \cdot H \cdot x^t x \cdot H^t\right)$$

Ahora bien, no es válido para hacer inferencia ya que depende del parámetro desconocido β.

Otra opción: Para $y_A = \begin{pmatrix} \sqrt{c_1} \cdot y \\ \sqrt{c_2} \cdot V \cdot x^t y \end{pmatrix}$ y $x_A = \begin{pmatrix} \sqrt{c_1} \cdot x \\ \sqrt{c_2} \cdot I_p \end{pmatrix}$ entonces $y_A = x_A \beta_A + u_A$ verifica $\widehat{\beta}_A = \widehat{\beta}_{-0}(c_1, c_2)$ si bien $var(\widehat{\beta}_A) \neq var(\widehat{\beta}_{-0}(c_1, c_2))$.

Hipótesis de normalidad

Llamando $H = \left(c_1 \cdot x^t x + c_2 I_p\right)^{-1} \cdot (c_1 \cdot I_p + c_2 \cdot V)$ se verifica que:

$$\begin{aligned}
\widehat{\beta}_{-0}(c_1, c_2) &= H \cdot x^t y = H \cdot x^t x \cdot \widehat{\beta} \\
&\hookrightarrow E[\widehat{\beta}_{-0}(c_1, c_2)] = H \cdot x^t x \cdot \beta \\
&\hookrightarrow var(\widehat{\beta}_{-0}(c_1, c_2)) = \sigma^2 \cdot H \cdot x^t x \cdot H^t \\
\widehat{\beta}_{-0}(c_1, c_2) &= H \cdot x^t y = H \cdot x^t x \cdot \beta + H \cdot x^t \cdot u
\end{aligned}$$

Considerando que $u \sim N(0, \sigma^2 I)$:

$$\widehat{\beta}_{-0}(c_1, c_2) \sim N\left(H \cdot x^t x \cdot \beta, \sigma^2 \cdot H \cdot x^t x \cdot H^t\right)$$

Ahora bien, no es válido para hacer inferencia ya que depende del parámetro desconocido β.

Otra opción: Para $y_A = \begin{pmatrix} \sqrt{c_1} \cdot y \\ \sqrt{c_2} \cdot V \cdot x^t y \end{pmatrix}$ y $x_A = \begin{pmatrix} \sqrt{c_1} \cdot x \\ \sqrt{c_2} \cdot I_p \end{pmatrix}$ entonces $y_A = x_A \beta_A + u_A$ verifica $\widehat{\beta}_A = \widehat{\beta}_{-0}(c_1, c_2)$ si bien $var(\widehat{\beta}_A) \neq var(\widehat{\beta}_{-0}(c_1, c_2))$.

Mientras que se avanza por este camino, ¿qué hacer?

Bootstrap

Bootstrap según Wikipedia

El bootstrapping (o bootstrap) es un método de remuestreo propuesto por Bradley Efron en 1979. Se utiliza para aproximar la distribución en el muestreo de un estadístico. Se usa frecuentemente para aproximar el sesgo o la varianza de un análisis estadístico, así como para construir intervalos de confianza.....

$$1, 2, 5, 4, 9, 3 \longrightarrow (1{,}031748, 6{,}968252)$$

$$\begin{aligned}
2, 5, 3, 1, 1, 4 &\rightarrow media = 2{,}666667 \\
1, 5, 1, 2, 4, 9 &\rightarrow media = 3{,}666667 \\
4, 1, 3, 5, 1, 2 &\rightarrow media = 2{,}666667 \\
9, 5, 5, 3, 3, 5 &\rightarrow media = 5 \\
3, 5, 4, 3, 3, 1 &\rightarrow media = 3{,}166667 \\
9, 5, 2, 1, 1, 1 &\rightarrow media = 3{,}166667 \\
&\vdots
\end{aligned}$$

Bootstrap

Bootstrap según Wikipedia

El bootstrapping (o bootstrap) es un método de remuestreo propuesto por Bradley Efron en 1979. Se utiliza para aproximar la distribución en el muestreo de un estadístico. Se usa frecuentemente para aproximar el sesgo o la varianza de un análisis estadístico, así como para construir intervalos de confianza.....

$$1, 2, 5, 4, 9, 3 \longrightarrow (1{,}031748, 6{,}968252)$$

$$\begin{array}{lcl} 2, 5, 3, 1, 1, 4 & \rightarrow & media = 2{,}666667 \\ 1, 5, 1, 2, 4, 9 & \rightarrow & media = 3{,}666667 \\ 4, 1, 3, 5, 1, 2 & \rightarrow & media = 2{,}666667 \\ 9, 5, 5, 3, 3, 5 & \rightarrow & media = 5 \\ 3, 5, 4, 3, 3, 1 & \rightarrow & media = 3{,}166667 \\ 9, 5, 2, 1, 1, 1 & \rightarrow & media = 3{,}166667 \\ & \vdots & \end{array}$$

10000 repeticiones

$$(P_{2,5},\ P_{97,5}) = (2{,}166667,\ 6{,}333333)$$

Bootstrap

Bootstrap según Wikipedia

El bootstrapping (o bootstrap) es un método de remuestreo propuesto por Bradley Efron en 1979. Se utiliza para aproximar la distribución en el muestreo de un estadístico. Se usa frecuentemente para aproximar el sesgo o la varianza de un análisis estadístico, así como para construir intervalos de confianza.....

$$1, 2, 5, 4, 9, 3 \longrightarrow (1{,}031748, 6{,}968252)$$

$$\begin{array}{lcl} 2, 5, 3, 1, 1, 4 & \rightarrow & media = 2{,}666667 \\ 1, 5, 1, 2, 4, 9 & \rightarrow & media = 3{,}666667 \\ 4, 1, 3, 5, 1, 2 & \rightarrow & media = 2{,}666667 \\ 9, 5, 5, 3, 3, 5 & \rightarrow & media = 5 \\ 3, 5, 4, 3, 3, 1 & \rightarrow & media = 3{,}166667 \\ 9, 5, 2, 1, 1, 1 & \rightarrow & media = 3{,}166667 \\ \vdots & & \end{array}$$

10000 repeticiones

$$(P_{2,5},\ P_{97,5}) = (2{,}166667,\ 6{,}333333)$$

Teorema Central del Límite

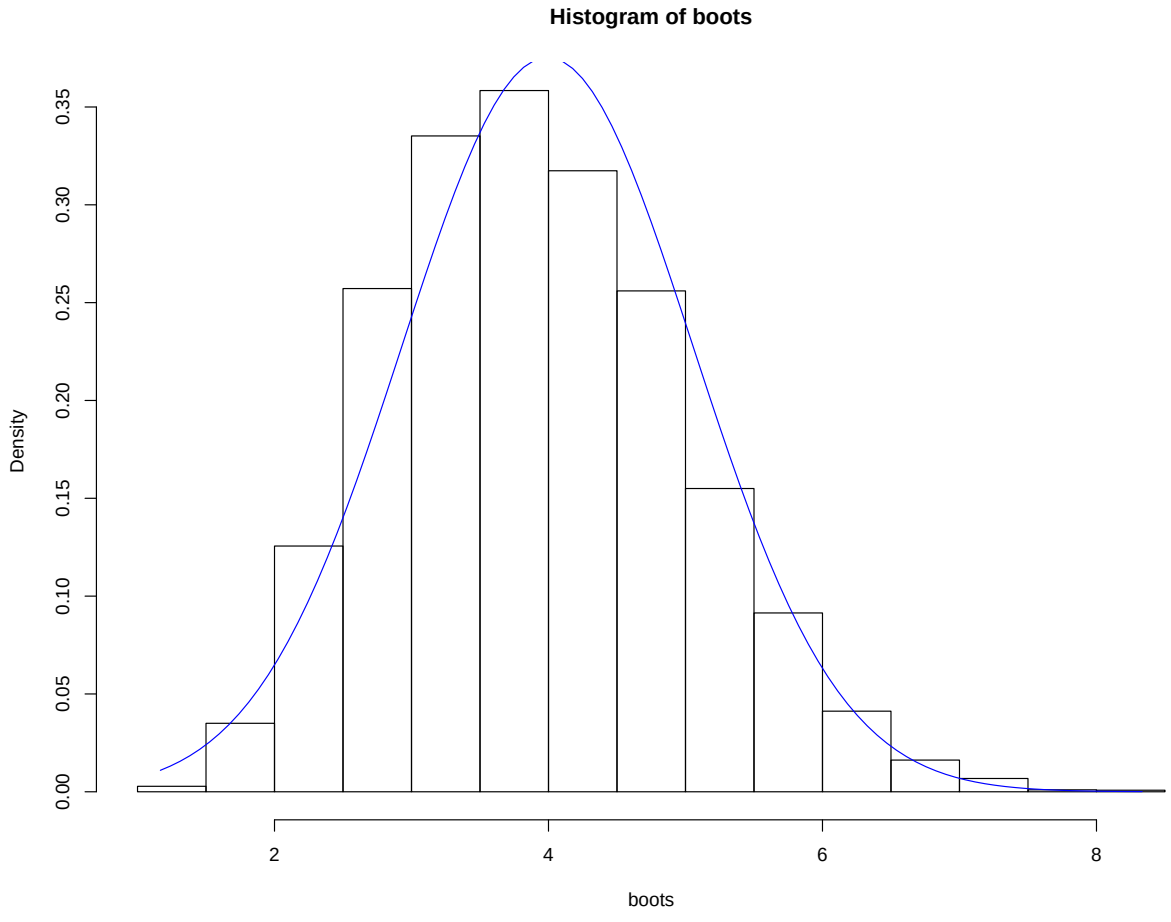
$$media_{rep} \pm 1{,}96 \cdot desv.tip_{rep} =$$

$$= (1{,}907400, 6{,}088767)$$

Son incluso más finos!!! (9)

Bootstrap

¡ojo! Kolmogorov-Smirnoff, H_0 : normalidad, $p - value < 2{,}2e - 16$
Shapiro no funciona en R para muestras superiores a 5000 (dudas de la utilidad de estas pruebas para tamaños muestrales tan elevados)



Volviendo a nuestro ejemplo problemático

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z + u$$

10000 repeticiones = 7 segundos aproximadamente

estimador	desv.tip estimada	intervalo
1.1411141	0.67056703	(-0.1731973, 2.4554255)
-16.7141945	0.07182645	(-16.8549743, -16.5734147)
-0.2560827	0.00872478	(-0.2731833, -0.2389821)

La constante no es significativamente distinta de cero

$(P_{2,5},\ P_{97,5})$
(-0.1443570, 2.4673888)
(-16.8781300, -16.5963682)
(-0.2725277, -0.2381248)

Volviendo a nuestro ejemplo problemático

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z + u$$

10000 repeticiones = 7 segundos aproximadamente

estimador	desv.tip estimada	intervalo
1.1411141	0.67056703	(-0.1731973, 2.4554255)
-16.7141945	0.07182645	(-16.8549743, -16.5734147)
-0.2560827	0.00872478	(-0.2731833, -0.2389821)

La constante no es significativamente distinta de cero

$(P_{2,5},\ P_{97,5})$
(-0.1443570, 2.4673888)
(-16.8781300, -16.5963682)
(-0.2725277, -0.2381248)

R^2

$$(0{,}9991456, 0{,}9996612)$$

$$(P_{2,5},\ P_{97,5}) = (0{,}9991075, 0{,}9996142)$$

Bootstrap y estimaciones inestables

¿Cómo saber si tenemos estimaciones inestables?

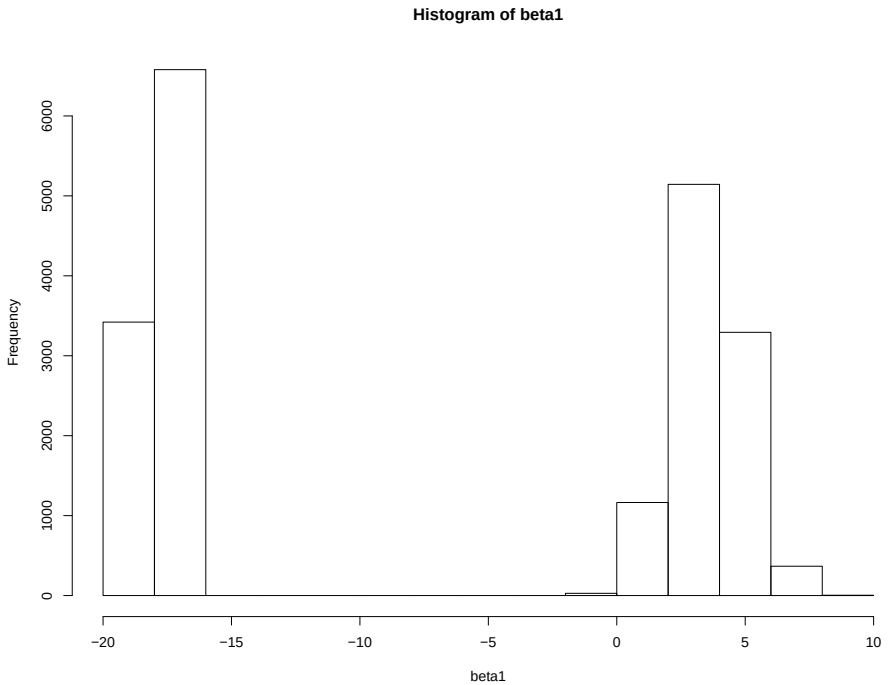
$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z + u, \quad Y = \beta_0 + \beta_1 \cdot X + u, \quad Y = \beta_0 + \beta_2 Z + u$$

Bootstrap y estimaciones inestables

¿Cómo saber si tenemos estimaciones inestables?

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z + u, \quad Y = \beta_0 + \beta_1 \cdot X + u, \quad Y = \beta_0 + \beta_2 Z + u$$

$$(-18{,}093, 5{,}795) = (P_{2,5}, P_{97,5})$$

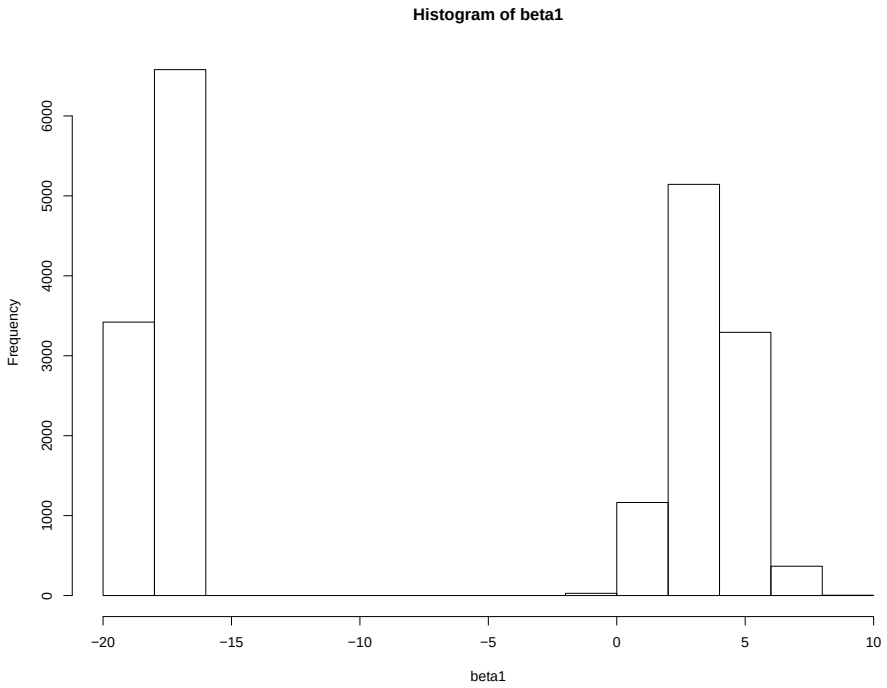


Bootstrap y estimaciones inestables

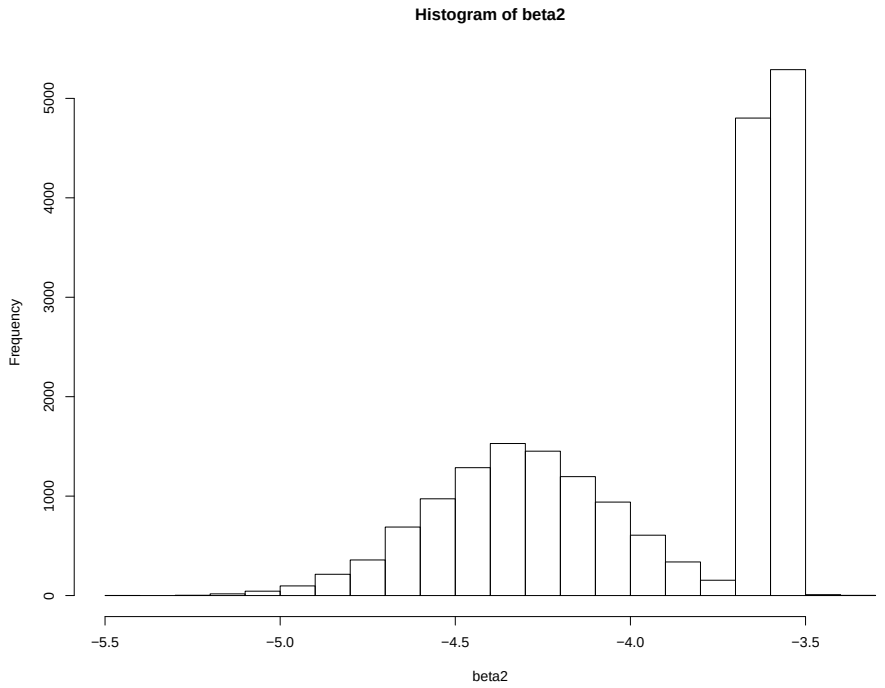
¿Cómo saber si tenemos estimaciones inestables?

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z + u, \quad Y = \beta_0 + \beta_1 \cdot X + u, \quad Y = \beta_0 + \beta_2 Z + u$$

$(-18{,}093, 5{,}795) = (P_{2,5}, P_{97,5})$



$(P_{2,5}, P_{97,5}) = (-4{,}762, -3{,}5906)$

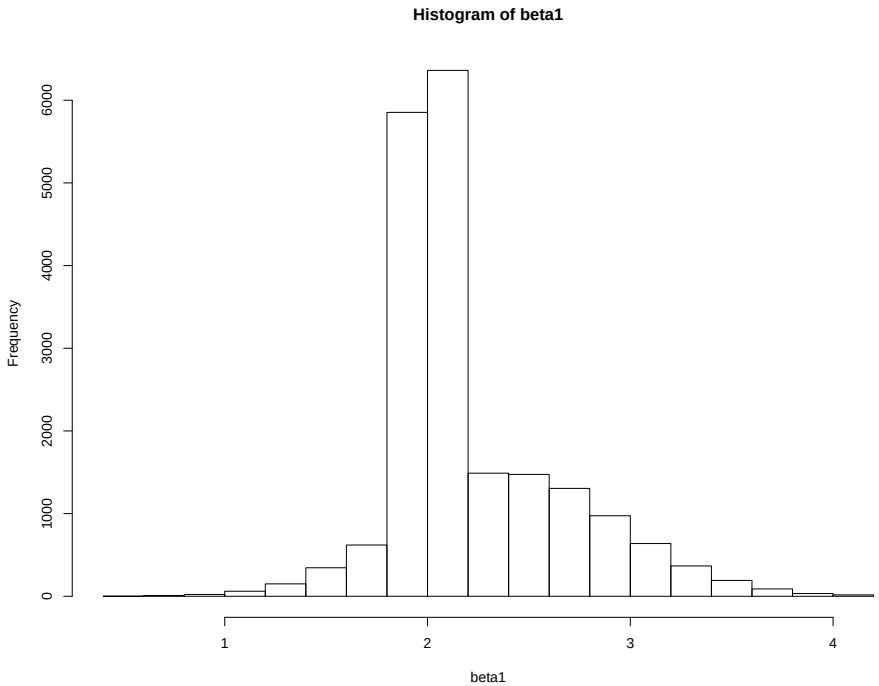


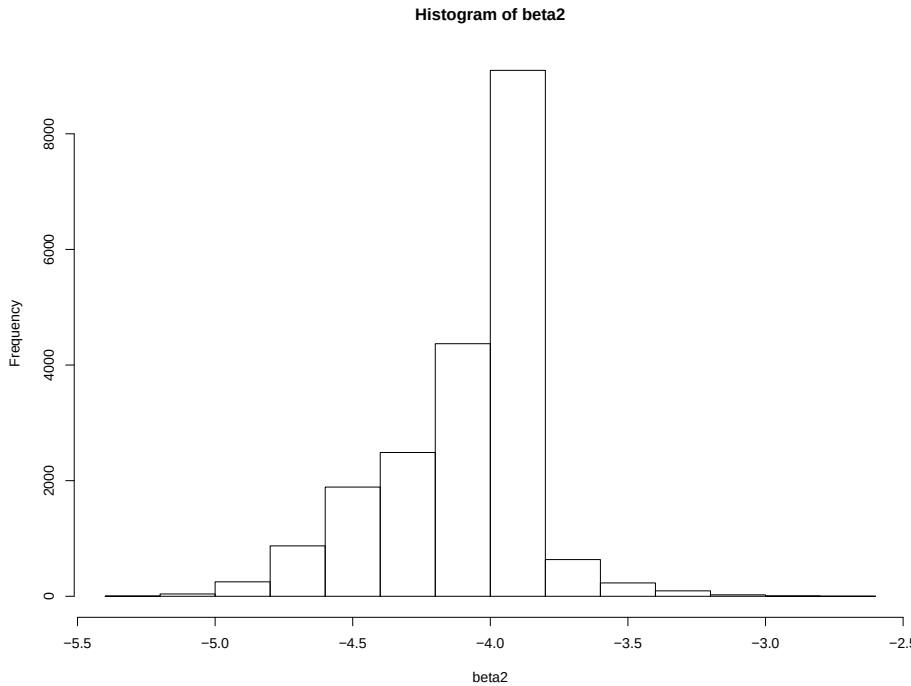
Bootstrap y estimaciones inestables

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 Z + u, \quad Y = \beta_0 + \beta_1 \cdot X + u, \quad Y = \beta_0 + \beta_2 Z + u$$

$$(1{,}568, 3{,}292) = (P_{2,5}, P_{97,5})$$

$$(P_{2,5}, P_{97,5}) = (-4{,}728 - 3{,}664)$$





Índice

Manpower data

The hospital manpower data, taken from Myers (1990), are a well-known example of highly collinear data to which ridge regression and various shrinkage and selection methods are often applied.
The data consist of measures taken at 17 U.S. Naval Hospitals and the goal is to predict the required monthly man hours for staffing purposes.

$$H = \beta_0 + \beta_1 \cdot L + \beta_2 \cdot X + \beta_3 \cdot B + \beta_4 \cdot A + \beta_5 \cdot S + u$$

Hours: horas mensuales por hombre (monthly man hours).

Load: carga diaria promedio del paciente (average daily patient load).

Xray: exposiciones mensuales a rayos X (monthly X-ray exposures).

BedDays: habitaciones diarias ocupadas mensualmente (monthly occupied bed days).

AreaPop: población elegible en el área (en miles) (eligible population in the area in thousands).

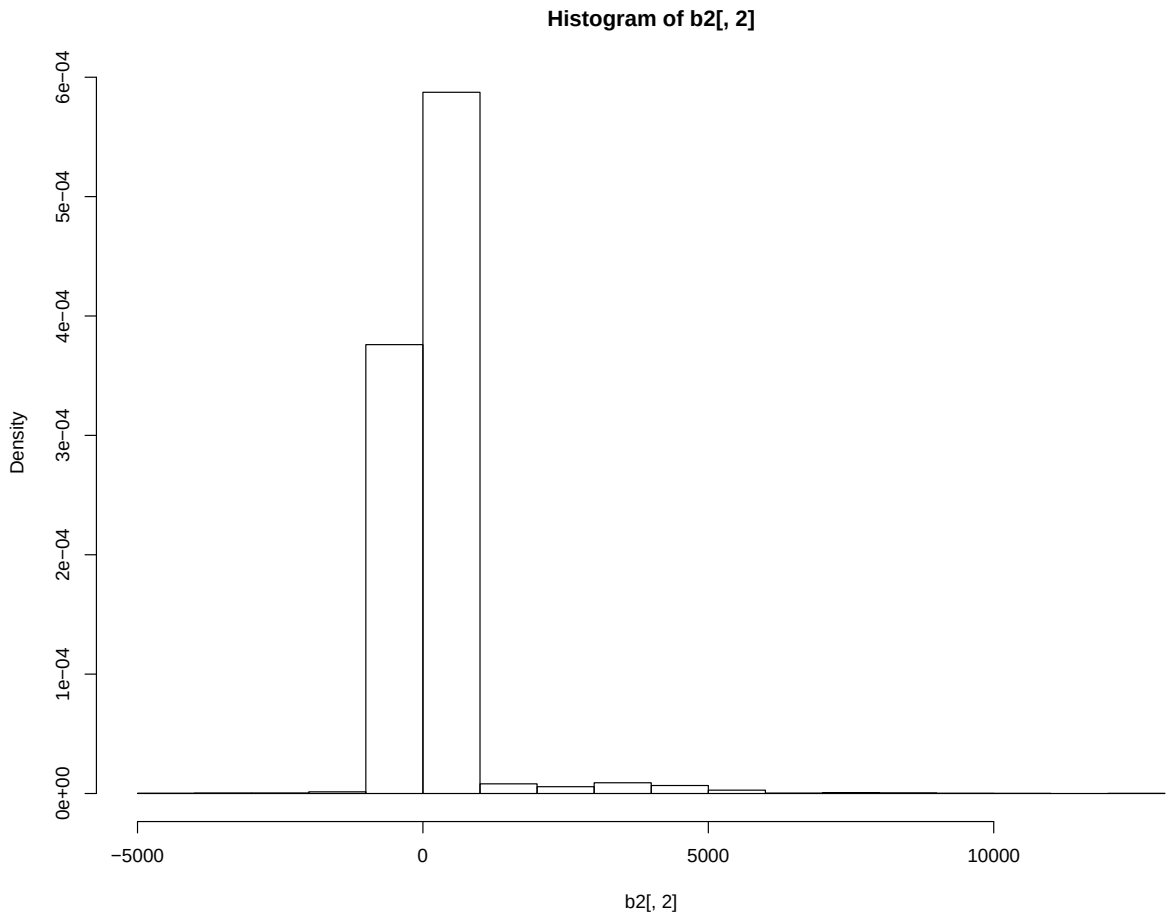
Stay: duración promedio de la estancia del paciente en días (average length of patient's stay in days).

Manpower data

¿Inestabilidad en las estimaciones? Bootstrap para β_1 en:

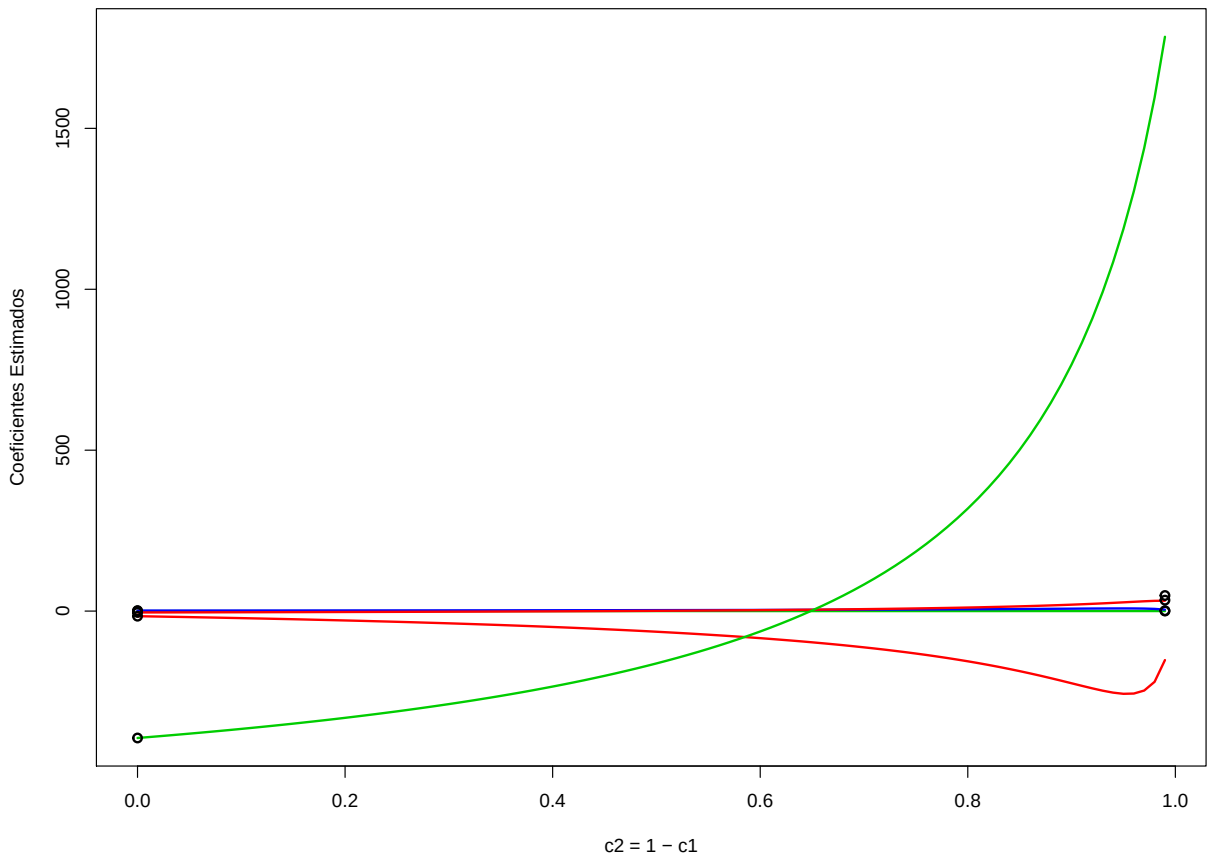
$$H = \beta_0 + \beta_1 \cdot L + \beta_2 \cdot X + \beta_3 \cdot B + \beta_4 \cdot A + \beta_5 \cdot S + u$$

$(P_{2,5}, P_{97,5}) = (-330{,}0086, 2185{,}8051)$, $min = -4980{,}894$, $max = 12833{,}13$



Manpower data

Traza de los estimadores: $\widehat{\beta}_1(c_1, c_2)$ rojo, $\widehat{\beta}_6(c_1, c_2)$ verde.



Manpower data

MCO $NC = 278{,}8721$

	$\widehat{\beta}$	$\sqrt{var(\widehat{\beta})}$
	1962.94	1071.361
carga paciente	-15.851	97.652
expo. rayos X	0.055	0.021
hab. ocupadas	1.589	3.092
población	-4.218	7.176
estancia media	-394.314	209.639

$$||\widehat{\beta}_0 - \widehat{\alpha}||^2 = 2426{,}373$$

$$R^2 = 0{,}990833$$

Metodología alternativa $c_2 = 0{,}99988$

$$NC = 19{,}65125$$

$\widehat{\beta}$	$(P_{2,5},\ P_{97,5})$
-11986.72	(-25283.007, -6141.665)
31.211	(24.921, 38.091)
0.157	(-0.222, 0.511)
-1.712	(-3.649, -0.2901)
30.793	(-11.051, 68.284)
2358.27	(1316.001, 5113.263)

$$||\widehat{\beta}_0 - \widehat{\alpha}||^2 = 14{,}79423$$

$$R^2 = 0{,}8423752$$

Manpower data

MCO $NC = 278{,}8721$

	$\widehat{\beta}$	$\sqrt{var(\widehat{\beta})}$
	1962.94	1071.361
carga paciente	-15.851	97.652
expo. rayos X	0.055	0.021
hab. ocupadas	1.589	3.092
población	-4.218	7.176
estancia media	-394.314	209.639

$$||\widehat{\beta}_0 - \widehat{\alpha}||^2 = 2426{,}373$$

$$R^2 = 0{,}990833$$

Metodología alternativa $c_2 = 0{,}99988$

$$NC = 19{,}65125$$

$\widehat{\beta}$	$(P_{2,5},\ P_{97,5})$
-11986.72	(-25283.007, -6141.665)
31.211	(24.921, 38.091)
0.157	(-0.222, 0.511)
-1.712	(-3.649, -0.2901)
30.793	(-11.051, 68.284)
2358.27	(1316.001, 5113.263)

$$||\widehat{\beta}_0 - \widehat{\alpha}||^2 = 14{,}79423$$

$$R^2 = 0{,}8423752$$

$$\widehat{\alpha} = (34'033, 0'246, 1'117, 48'436, 2030'974)$$

$$c = 0{,}9999999 \rightarrow \widehat{\beta}_4 = -1{,}236346\ (-1'993,\ 0'087), \quad NC = 4{,}366702$$

Manpower data

¿Como afecta la disminución del R^2 (de 0.990833 a 0.8423752) a la predicción?

Obs.	$\widehat{H}$ según MCO	$\widehat{H}$ según alternativa	H
1	775.0251	-902.6030	566.52
2	740.6702	3205.9981	696.82
3	1103.9234	-1301.1400	1033.15
4	1240.4956	-821.9650	1603.62
5	1564.4217	1690.1287	1611.37
6	2151.2717	401.0485	1613.27
7	1689.7004	2091.1631	1854.17
8	1736.2355	1473.4803	2160.55
9	2736.9890	4123.2179	2305.58
10	3681.8534	9098.7043	3503.93
11	3239.2889	3597.9314	3571.89
12	4353.3330	2127.7420	3741.40
13	4257.0884	4864.7385	4026.52
14	8767.7485	9682.8557	10343.81
15	12237.0274	14563.4461	1732.17
16	15038.3909	12309.7002	15414.94
17	19320.6970	18429.7132	18854.45

Evidentemente peor ajuste, quizás si nuestro objetivo es la predicción mejor no hacer nada.

Manpower data

Si queremos evitar estimaciones/predicciones negativas se podría plantear:

minimizar

$$c_1 \cdot SCR_{-0} + c_2 \cdot ||\widehat{\beta}_{-0} - \widehat{\alpha}||^2$$

sujeto a

$$\widehat{y}_i = \overline{Y} + x_{i\cdot} \cdot \widehat{\beta}_{-0} \geq 0, \quad i = 1, \cdots, p$$

$$SCR_{-0} = Y^t Y - 2 \cdot \widehat{\beta}_{-0}^t X^t Y + \widehat{\beta}_{-0}^t X^t X \widehat{\beta}_{-0}$$

$$||\widehat{\beta}_{-0} - \widehat{\alpha}||^2 = \widehat{\beta}_{-0}^t \widehat{\beta}_{-0} - 2 \cdot \widehat{\beta}_{-0}^t \widehat{\alpha} + \widehat{\alpha}^t \widehat{\alpha}$$

$$\widehat{\alpha} = (\widehat{\alpha}_i)_{i=1,\ldots,p} = \left(\frac{\sum_{t=1}^{n} x_{it} y_t}{\sum_{t=1}^{n} x_{it}^2} \right)_{i=1,\ldots,p}$$

Estimaciones inestables y ausencia de normalidad en el modelo de regresión lineal

Román Salmerón Gómez
romansg@ugr.es
Víctor Blanco Izquierdo
vblanco@ugr.es

Universidad de Granada
Departamento de Métodos Cuantitativos para la Economía y la Empresa

I Jornadas de Análisis Cuantitativo en Economía
9 y 10 de Mayo 2018