Regresión Robusta

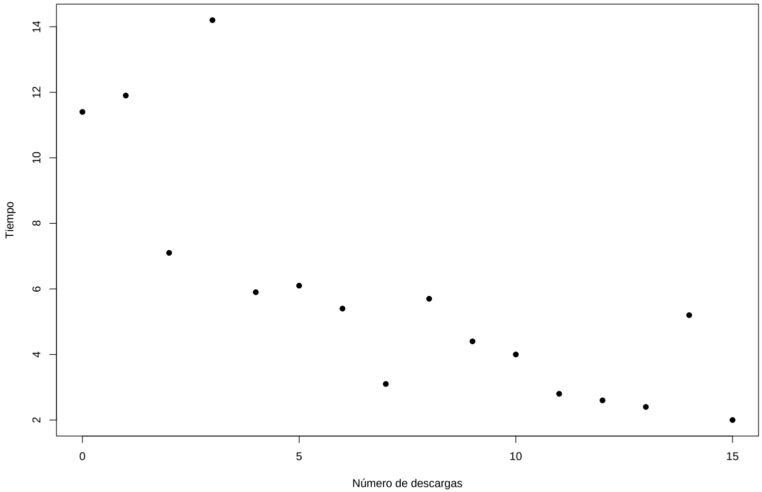
Roberto C. Duarte

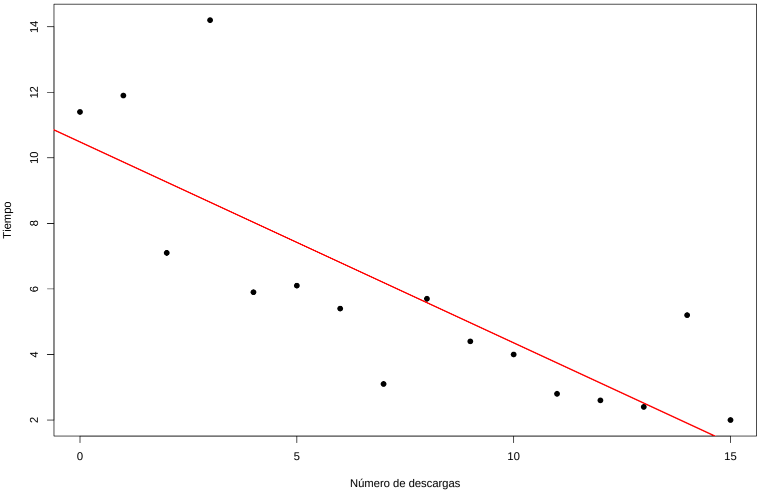
Universidad de Buenos Aires

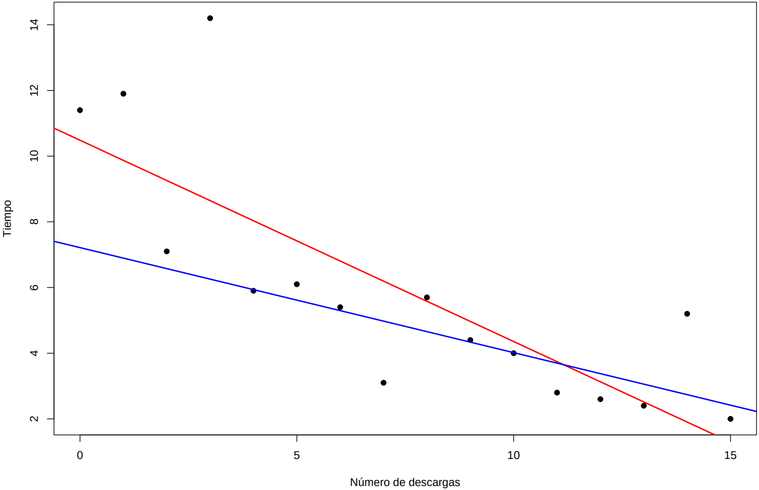
Noviembre 2022

Ejemplo (**Ejemplo 1**)

*En un experimento sobre la velocidad de aprendizaje de las ratas, se registraron los tiempos que tardan en pasar por una **shuttlebox** en varios intentos. Si el tiempo superaba los 5 segundos la rata recibía una descarga eléctrica durante el siguiente intento. Los datos son el número de descargas recibidas y el tiempo medio de todos los intentos entre descargas.*

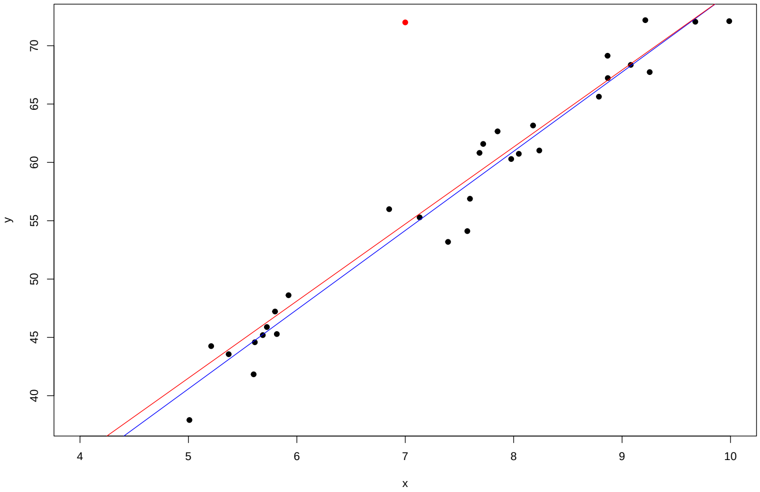

Tiempo
2
4
6
8
10
12
14
0
5
10
15
Número de descargas


Tiempo
Número de descargas
2
4
6
8
10
12
14
0
5
10
15

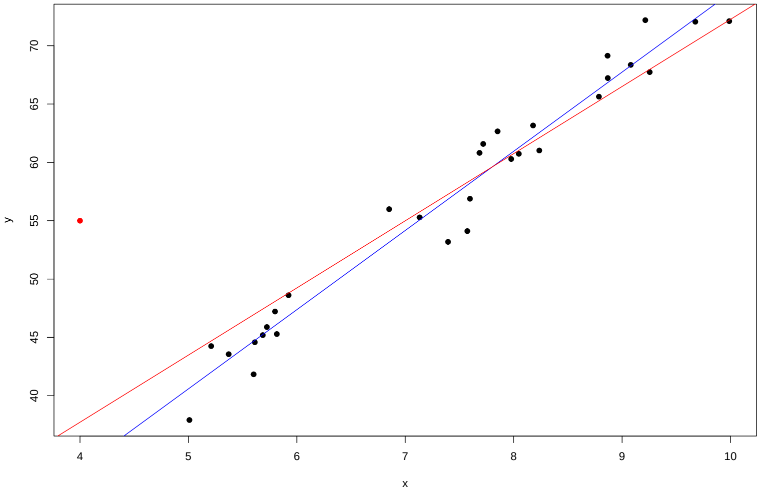

Tiempo
Número de descargas
2
4
6
8
10
12
14
0
5
10
15

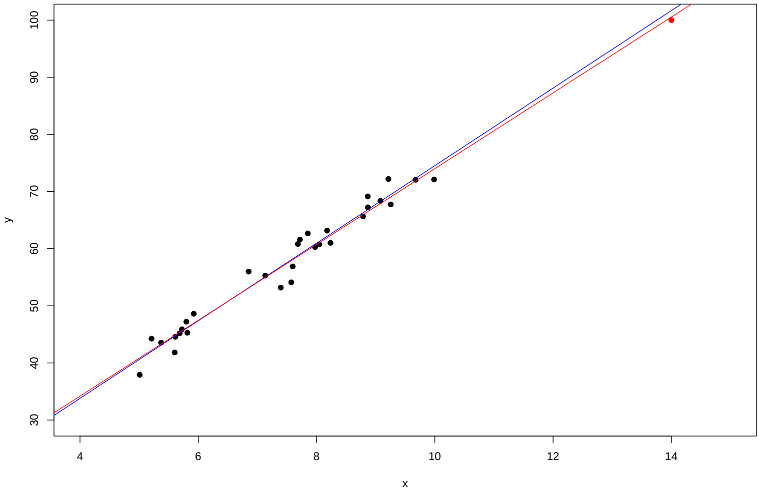
Outlier y punto influyente

- **Valor atípico:** Es una observación que es numéricamente distante del resto de los datos.
- **Punto influyente:** Punto que tiene impacto en las estimaciones del modelo.

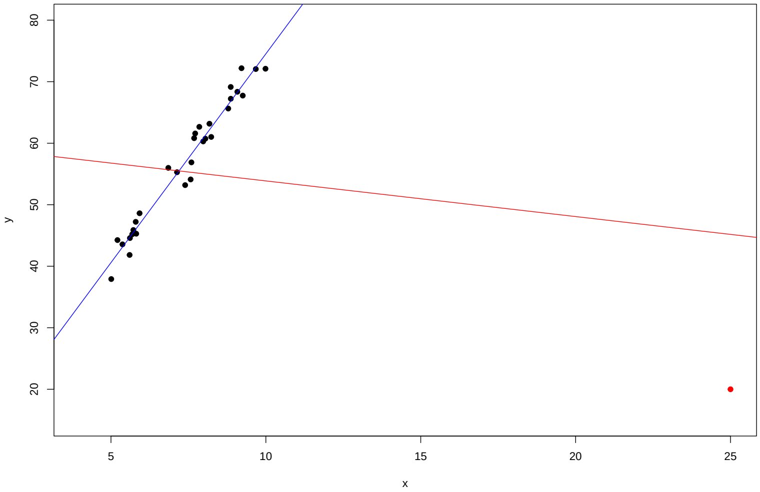


y
70
65
60
55
50
45
40
x
4
5
6
7
8
9
10


y
x
40
45
50
55
60
65
70
4
5
6
7
8
9
10



y
x
30
40
50
60
70
80
90
100
4
6
8
10
12
14


80
70
60
50
40
30
20
y
5
10
15
20
25
x

Definición (**Leverage**)

Sea $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ *la matriz de proyección, llamamos "leverage"(apalancamiento) de la observacion* $\boldsymbol{x}_i$ *al elemento* $h_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$ *de la matriz de proyección.*

$$0 \leq h_{ii} \leq 1$$

$$\sum_{i=1}^{n} h_{ii} = Rango(\mathbf{X})$$

$$h_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$$

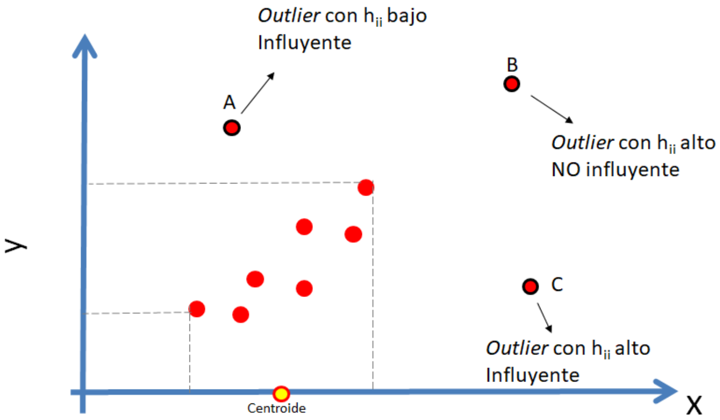
$\bar{\mathbf{x}}$ es el vector de promedios de los $\mathbf{x}_i$ y $\mathbf{X}^*$ es la matriz de dimensión $n \times (p-1)$ cuya i-esima fila es el vector $(\mathbf{x}_i - \bar{\mathbf{x}})'$.

$$h_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$$

$\bar{\mathbf{x}}$ es el vector de promedios de los $\mathbf{x}_i$ y $\mathbf{X}^*$ es la matriz de dimensión $n \times (p-1)$ cuya i-esima fila es el vector $(\mathbf{x}_i - \bar{\mathbf{x}})'$.

El **leverage** mide la distancia entre la i-esima observación y el vector de promedios.

Valores atípicos



¿Por qué puede influir tanto una observación?

$$
\begin{aligned}
E(r_i) &= 0 \\
Var(r_i) &= \sigma^2(1 - h_{ii})
\end{aligned}
$$

Cuanto mayor sea h_{ii} menor es la varianza de r_i

Definición (**Distancia de Cook**)

Llamemos $\widehat{\boldsymbol{\beta}}$ *y* $\widehat{\boldsymbol{\beta}}_{(i)}$ *a los estimadores OLS usando los datos completos y los datos sin la i-esima observación. Sean*

$$\begin{aligned} \widehat{\boldsymbol{y}} &= \boldsymbol{X}\widehat{\beta} \\ \widehat{\boldsymbol{y}}_{(i)} &= \boldsymbol{X}\widehat{\beta}_{(i)} \end{aligned}$$

Definimos la distancia de Cook de la i-ésima observación por

$$D_i = \frac{1}{pS^2}||\widehat{\boldsymbol{y}}_{(i)} - \widehat{\boldsymbol{y}}||^2 = \frac{r_i^2}{S^2 p}\frac{h_{ii}}{(1-h_{ii})^2}$$

Criterios clásicos para la distancia de Cook

Observación influyente

- $D_i \geq F(p, n-p, 0{,}50)$
- $D_i > 1$

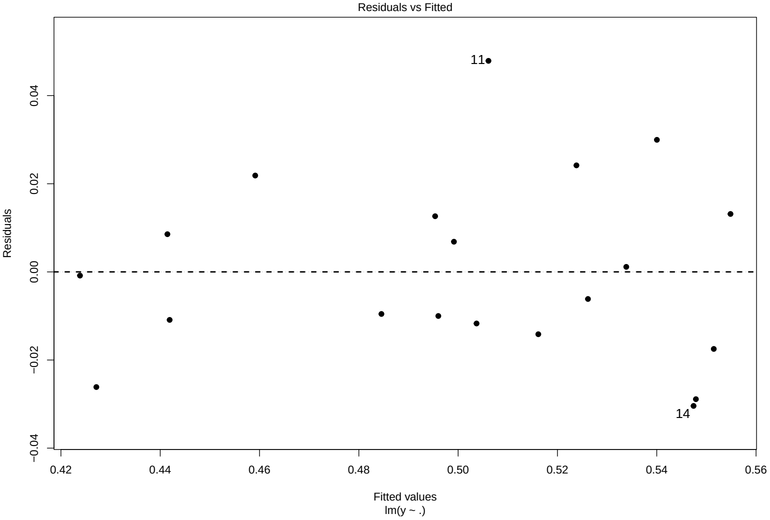
Observación con alto leverage

- $h_{ii} > 2\bar{h} = \frac{2}{n} \sum h_{jj} = \frac{2p}{n}$ (ajustado por la cantidad de covariables)
- $h_{ii} > 0{,}5$ (sin ajustar por la cantidad de covariables)

Problema: El leverage está basado en promedios, que es una medida poco robusta.

Ejemplo (**Ejemplo 5.2 de [Maronna et al., 2019]**)

Este ejemplo se basa en datos modificados de peso especifico de madera, los datos cuentan con 20 casos con 5 variables explicativas donde las observaciones 4, 6, 8 y 19 fueron reemplazadas por valores atípicos (dataset wood).


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ .)
0.04
0.02
0.00
−0.02
−0.04
0.42
0.44
0.46
0.48
0.50
0.52
0.54
0.56
11
14

M estimador de regresión

Definimos el M estimador de regresión como:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}\right)$$

diferenciando la ecuación anterior tenemos

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\widehat{\boldsymbol{\beta}})}{\widehat{\sigma}}\right)\mathbf{x}_i = \mathbf{0}$$

Definición (función ρ)

- *Si* $0 \leq u \leq v \Rightarrow \rho(u) \leq \rho(v)$.
- $\rho(u) = \rho(-u)$.
- ρ *es derivable.*
- ρ *debe ser acotada.*

Definición (función ψ)

Es una función que es derivada de una función ρ.

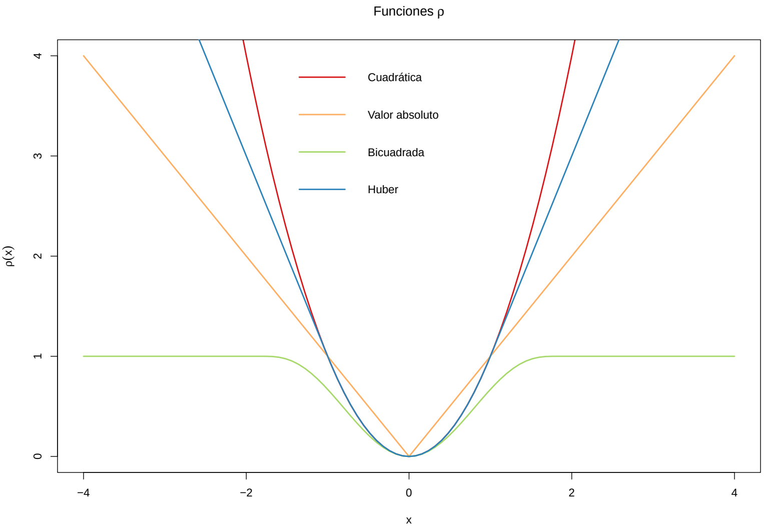
- ψ *es impar y* $\psi(x) \geq 0 \forall x \geq 0$.

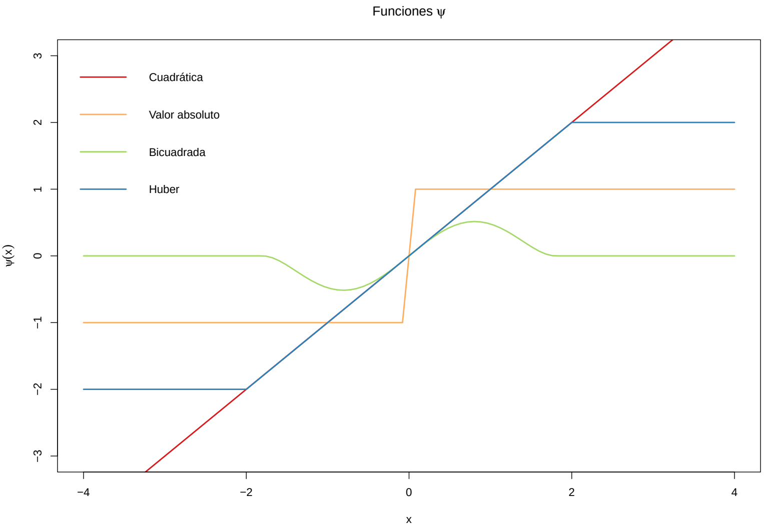
- ρ de Huber

$$\rho_k(x) = \begin{cases} x^2 & si \quad |x| \leq k \\ 2k|x| - k^2 & si \quad |x| > k \end{cases}$$

- ρ bicuadrada

$$\rho_k(x) = \begin{cases} 1 - \left[1 - \left(\frac{x}{k}\right)^2\right]^3 & si \quad |x| \leq k \\ 1 & si \quad |x| > k \end{cases}$$


Funciones ρ
Cuadrática
Valor absoluto
Bicuadrada
Huber
ρ(x)
x
−4
−2
0
2
4
0
1
2
3
4


Funciones ψ
Cuadrática
Valor absoluto
Bicuadrada
Huber
ψ(x)
x
−4
−2
0
2
4
−3
−2
−1
0
1
2
3

- Los M- estimadores correspondientes a funciones ψ monótonas se llaman M- estimadores monótonos.
- Los M- estimadores correspondientes a funciones ψ que tienden a cero en $\pm\infty$ se llaman **M-estimadores redescendientes.**

Para protegerse de los outliers de alto leverage y del efecto de enmascaramiento resulta imprescindible usar M - estimadores redescendientes.

¿ Cómo calcular los M-estimadores ?

1. Calcular un estimador inicial $\widehat{\boldsymbol{\beta}}_0$ robusto que no necesite un estimador de escala previo.
2. Calcular los residuos $r_i(\widehat{\boldsymbol{\beta}}_0)$ correspondientes al ajuste anterior y estimar $\widehat{\sigma}$ con un estimador robusto.
3. Encontrar la solución $\widehat{\boldsymbol{\beta}}$ de

$$\sum_{i=1}^{n} \psi\left(\frac{r_i}{\widehat{\sigma}}\right) \mathbf{x}_i = \mathbf{0}$$

   usando un método iterativo con $\widehat{\boldsymbol{\beta}}_0$ como estimador inicial.

Consideremos las ecuaciones de estimación:

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{x}_i'\beta}{\widehat{\sigma}}\right)\mathbf{x}_i = \mathbf{0}$$

Consideremos las ecuaciones de estimación:

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{x}_i'\beta}{\widehat{\sigma}}\right)\mathbf{x}_i = \mathbf{0}$$

$$\sum_{i=1}^{n} \underbrace{\frac{\psi\left(\frac{y_i - \mathbf{x}_i'\beta}{\widehat{\sigma}}\right)}{\frac{y_i - \mathbf{x}_i'\beta}{\widehat{\sigma}}}}_{W_i}\left(\frac{y_i - \mathbf{x}_i'\beta}{\widehat{\sigma}}\right)\mathbf{x}_i = \mathbf{0}$$

Consideremos las ecuaciones de estimación:

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{x}_i'\beta}{\widehat{\sigma}}\right)\mathbf{x}_i = \mathbf{0}$$

$$\sum_{i=1}^{n} \underbrace{\frac{\psi\left(\frac{y_i - \mathbf{x}_i'\beta}{\widehat{\sigma}}\right)}{\frac{y_i - \mathbf{x}_i'\beta}{\widehat{\sigma}}}}_{W_i} \left(\frac{y_i - \mathbf{x}_i'\beta}{\widehat{\sigma}}\right)\mathbf{x}_i = \mathbf{0}$$

$$\sum_{i=1}^{n} W_i \left(\frac{y_i - \mathbf{x}_i'\beta}{\widehat{\sigma}}\right)\mathbf{x}_i = \mathbf{0}$$

MM – estimador

Yohai (1987) propone un método de estimación para obtener una robustez y eficiencia óptimas. Usar un S estimador para la estimación inicial $\widehat{\beta}_0$ y usar un M– estimador de escala para la estimación de $\widehat{\sigma}$, los estimadores resultantes con este proceso se conocen como MM – estimadores.

Definición (**S-estimador**)

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{r_i}{\widehat{\sigma}}\right)=\delta$$

$\delta \in (0,1)$, *usualmente usamos* $\delta = 0{,}5$

Modelo de regresión logistica

$$\begin{aligned} y_i &\sim B(1, p_i) \\ p_i &= F(\boldsymbol{\beta}'\mathbf{x}) \end{aligned}$$

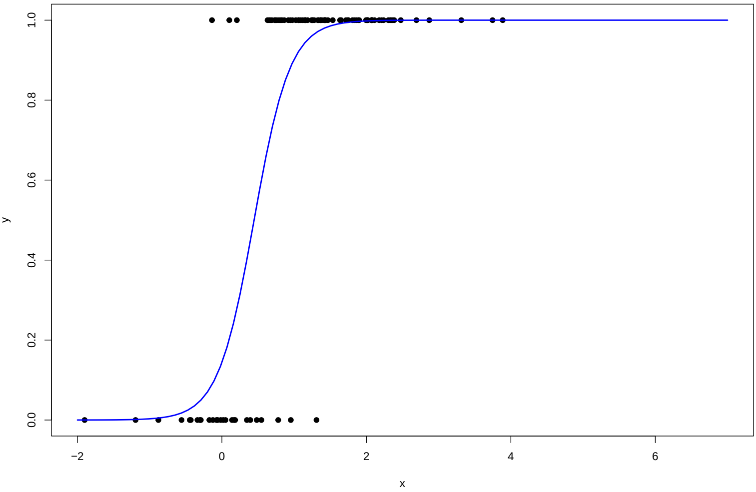
dónde

$$F(y) = \begin{cases} \frac{e^y}{1+e^y} & \textit{Modelo logistico} \\ \\ \Phi(y) & \textit{Modelo probit} \end{cases}$$

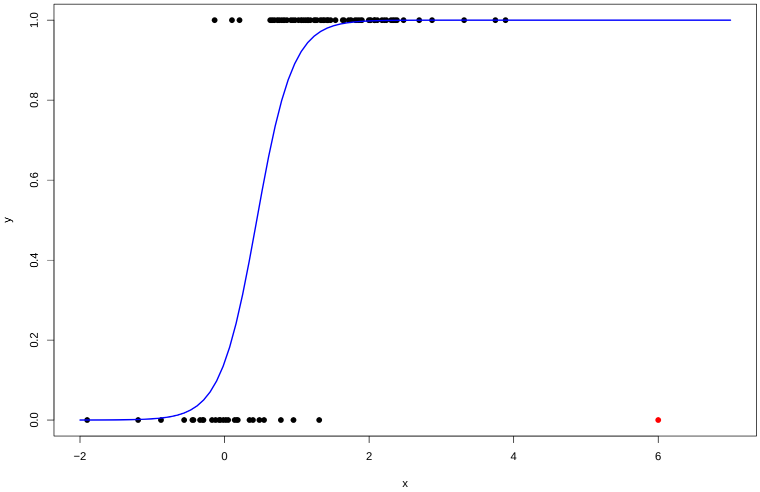
F^{-1}: Función de enlace

Estimador de máxima verosimilitud

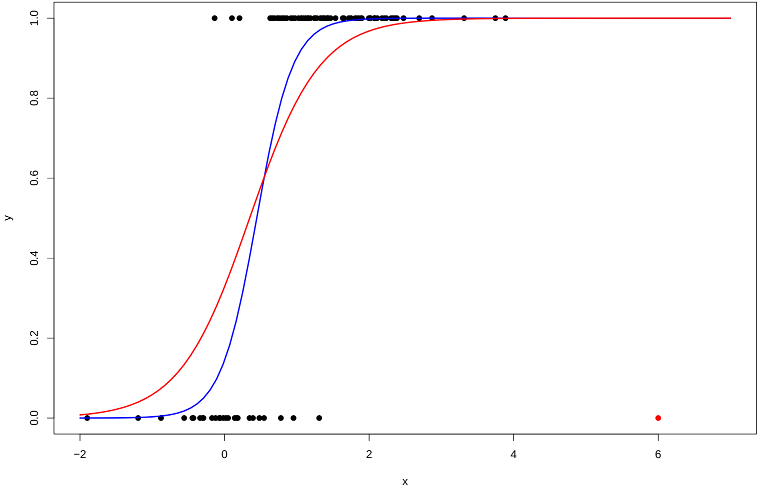
$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \arg\max \sum_{i=1}^{n} \log(p_i^{y_i}(1-p_i)^{1-y_i}) \\
&= \arg\max \sum_{i=1}^{n} y_i \log(p_i) + (1-y_i)\log(1-p_i) \\
&= \arg\min \sum_{i=1}^{n} -y_i \log(p_i) - (1-y_i)\log(1-p_i) \\
&= \arg\min \sum d(y_i, p_i)
\end{aligned}
$$

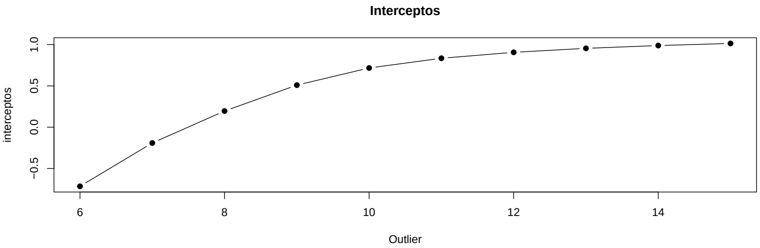


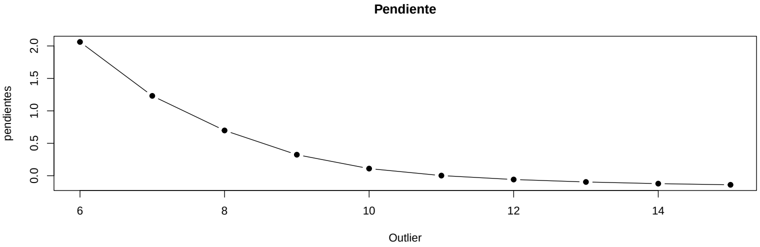
1.0
0.8
0.6
0.4
0.2
0.0
y
−2
0
2
4
6
x



1.0
0.8
0.6
0.4
0.2
0.0
y
−2
0
2
4
6
x


1.0
0.8
0.6
0.4
0.2
0.0
y
−2
0
2
4
6
x


Interceptos
interceptos
1.0
0.5
0.0
−0.5
6
8
10
12
14
Outlier


Pendiente
pendientes
2.0
1.5
1.0
0.5
0.0
6
8
10
12
14
Outlier

Mínimos cuadrados pesados

Carroll and Pederson (1993) porponen el estimador

$$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n} w_i \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right]$$

con $w_i = W(h_n(\mathbf{x}_i))$ donde W es una función no decreciente tal que $W(u)u$ es acotada.

Pregibon (1981) propone el estimador

$$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n} \rho(d(y_i, p_i))$$

Bianco and Yohai (1996) observaron que en algunos casos este estimador no cumple la condición de consistencia de Fisher y proponen el siguiente estimador

Estimador Bianco-Yohai

$$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n} \left[\rho(d(p_i, y_i)) + q(p_i)\right]$$

$$q(u) = \nu(u) + \nu(1-u)$$

$$\nu(u) = 2\int_0^u \psi(-2\log t)dt$$

$q(p_i)$ es una corrección para garantizar la consistencia de Fisher.

Croux and Haesbroeck (2003) describen condiciones suficientes para en la función ρ para garantizar el mínimo

$$\psi_c^{CH} = \exp\left(-\sqrt{max(u,c)}\right)$$

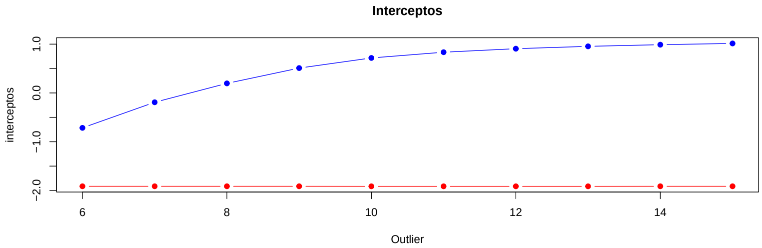
Luego la ecuación de estimación resultante es

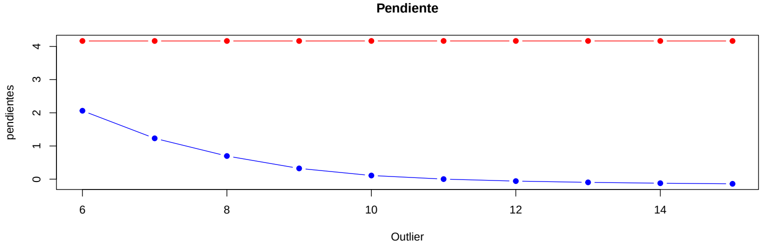
$$\sum_{i=1}^{n} p_i(\boldsymbol{\beta})(1-p_i(\boldsymbol{\beta}))\left[\psi(-2\log p_i(\boldsymbol{\beta})) - \psi(-2\log(1-p_i(\boldsymbol{\beta})))\right]\mathbf{x}_i = \mathbf{0}$$

- El resultado anterior garantiza que el estimador cumpla la condición de consistencia de Fisher.
- Bianco and Yohai (1996) probaron que el estimador es asintoticamente normal.
- La función Ψ resulta no ser acotada, **Croux and Haesbroeck (2003)** propone reducir la ponderación de las observaciones con alto leverage.

M estimador redescendiente ponderado

$$\boldsymbol{\beta} = \arg\min \sum_{i=1}^{n} w_i \left[\rho(d(p_i(\boldsymbol{\beta}), y_i)) + q(p_i(\boldsymbol{\beta}))\right]$$


Interceptos
interceptos
1.0
0.0
−1.0
−2.0
6
8
10
12
14
Outlier


Pendiente
pendientes
4
3
2
1
0
6
8
10
12
14
Outlier

Ejemplos

Ejemplo

Datos correspondientes a 33 pacientes que padecen leucemia, la variable respuesta vale 1 cuando el paciente sobrevive más de 52 semanas y las dos covariables observadas son:
WBC: *Número de glóbulos blancos.*
AG: *Presencia o ausencia de cierta característica morfológica en los glóbulos.*

Bibliografia

Bianco, A. M. (1990).
M-estimadores robustos para regresión binominal y multinominal.
PhD thesis, Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019).
Robust statistics: theory and methods (with R).
John Wiley & Sons.