

MAT-266: Métodos para regresión robusta, *M*-estimadores

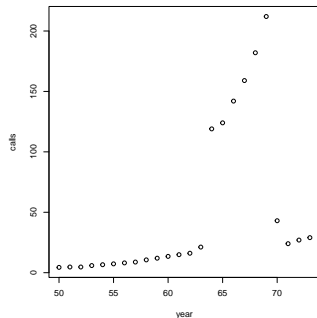
Felipe Osorio

fosorios.mat.utfsm.cl

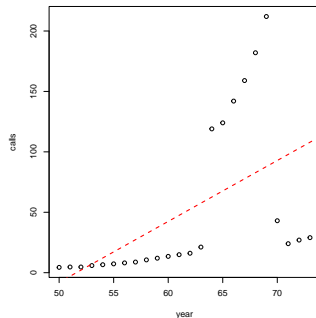
Departamento de Matemática, UTFSM



Llamadas telefónicas en Bélgica 1950-1973



(a) llamadas telefónicas



(b) recta ajustada

M-estimación

Para introducir ideas considere X_1, \dots, X_n variables aleatorias IID. Sabemos que la **media muestral** \bar{X} es solución del problema,

$$\min_{\theta} \sum_{i=1}^n (x_i - \theta)^2,$$

o análogamente,

$$\sum_{i=1}^n (x_i - \theta) = 0.$$

Mientras que, la **mediana** $\text{me}(\mathbf{x})$ que es **robusta contra outliers**, es solución del problema

$$\min_{\theta} \sum_{i=1}^n |x_i - \theta|,$$

es decir $\text{me}(\mathbf{x})$ es solución de la ecuación

$$\sum_{i=1}^n \{(-1)I_{(-\infty, 0)}(x_i - \theta) + I_{(0, \infty)}(x_i - \theta)\} = 0.$$



El MLE $\hat{\theta}_{\text{ML}}$ en un modelo paramétrico es el minimizador del negativo de la log-verosimilitud

$$\min_{\theta} \left\{ - \sum_{i=1}^n \log f(x_i; \theta) \right\},$$

y en el caso de que $\ell(\theta) = \sum_{i=1}^n \log f(x_i; \theta)$ sea diferenciable, $\hat{\theta}_{\text{ML}}$ es solución de

$$\sum_{i=1}^n \frac{\partial \log f(x_i; \theta)}{\partial \theta} = 0,$$

lo que lleva a la siguiente definición

Definición 1 (M-estimador):

Para X_1, \dots, X_n variables aleatorias IID. El **M-estimador** $\hat{\theta}_M$ con respecto a la función $\psi : \mathbb{R} \times \Theta \rightarrow \mathbb{R}$ se define como la solución de

$$\sum_{i=1}^n \psi(X_i; \hat{\theta}_M) = 0. \tag{1}$$



Usualmente (1) corresponde a la solución del problema

$$\min_{\theta} \sum_{i=1}^n \rho(x_i; \theta),$$

si ρ es diferenciable, entonces

$$\psi(x; \theta) = c \frac{\partial \rho(x; \theta)}{\partial \theta},$$

para alguna constante c .

Observación:

Para simplificar la notación podemos hacer

$$\rho(x; \theta) = \tilde{\rho}(x - \theta), \quad \psi(x; \theta) = \tilde{\psi}(x - \theta).$$

Ejemplo (Estimador LS)¹:

Sea $z = x - \theta$. Las funciones $\tilde{\rho}(z) = z^2$ y $\tilde{\psi}(z) = z$ llevan a la media muestral.

¹O bien, estimadores L_2 .



Ejemplo (Estimador LAD)²:

La mediana es un M -estimador con $\tilde{\rho}(z) = |z|$, y

$$\tilde{\psi}(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0. \end{cases}$$

Ejemplo (Media recortada):

La primera propuesta de Huber (1981) para reducir la influencia de outliers es

$$\tilde{\rho}(z) = \begin{cases} z^2, & |z| \leq k, \\ k^2, & |z| > k, \end{cases}$$

donde k es una constante de tuning, y

$$\tilde{\psi}(z) = \begin{cases} z, & |z| \leq k, \\ 0, & |z| > k. \end{cases}$$

²Estimadores mínimo desvío absoluto (LAD) corresponden a estimadores L_1 .

Ejemplo (Media Winsorizada)³:

Huber (1981) propuso un compromiso entre la media y la mediana, como:

$$\tilde{\rho}(z) = \begin{cases} \frac{1}{2}z^2, & |z| \leq k, \\ k|z| - \frac{1}{2}k^2, & |z| > k. \end{cases}$$

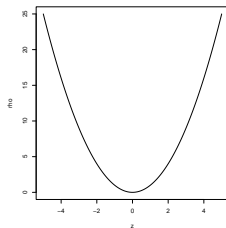
El estimador $\hat{\theta}_M$ es solución de:

$$\sum_{i=1}^n \tilde{\psi}(x_i - \theta) = 0,$$

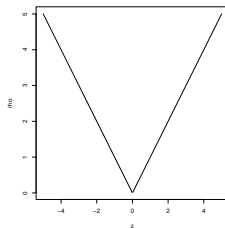
donde

$$\tilde{\psi}(z) = \begin{cases} -k, & z < -k, \\ z, & |z| \leq k, \\ k, & z > k. \end{cases}$$

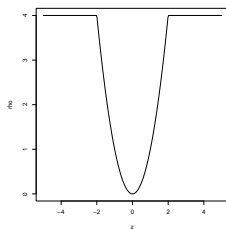
³Para $k \rightarrow 0$ obtenemos la mediana, cuando $k \rightarrow \infty$ lleva a la media, mientras que $k = 1.345$ tiene 95% de eficiencia bajo normalidad.



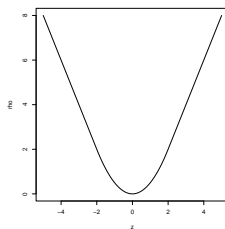
(a) media



(b) mediana



(c) media recortada



(d) media winsorizada

Sabemos que el estimador LS es solución del problema de estimación

$$\min_{\beta} \sum_{i=1}^n (Y_i - \mathbf{x}_i^{\top} \beta)^2. \quad (2)$$

Bajo normalidad, el MLE de β minimiza la función

$$-\sum_{i=1}^n \log f(Y_i; \beta) = \frac{n}{2} \log 2\pi\sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \mathbf{x}_i^{\top} \beta)^2. \quad (3)$$

Para obtener estimadores robustos Huber (1981) sugirió substituir el negativo de la función de log-verosimilitud en (3) por una función que permita disminuir el efecto de outliers.



De este modo Huber (1981) propone obtener estimadores tipo-ML, conocidos como *M*-estimadores resolviendo el problema

$$\min_{\beta} \sum_{i=1}^n \rho(Y_i - \mathbf{x}_i^{\top} \beta). \quad (4)$$

Es usual incorporar el parámetro de escala σ^2 definiendo,

$$z_i = \frac{Y_i - \mathbf{x}_i^{\top} \beta}{\sigma}, \quad i = 1, \dots, n,$$

y considere la función objetivo

$$Q_{\rho}(\beta) = \sum_{i=1}^n \rho(z_i).$$

De este modo,

$$\frac{\partial Q_{\rho}(\beta)}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \rho(z_i)}{\partial z_i} \frac{\partial z_i}{\partial \beta_j} = \sum_{i=1}^n W(z_i) z_i \frac{\partial z_i}{\partial \beta_j}, \quad W(z_i) = \frac{1}{z_i} \frac{\partial \rho(z_i)}{\partial z_i}.$$



Por tanto el M -estimador de β es solución del sistema de ecuaciones

$$\sum_{i=1}^n \psi\left(\frac{Y_i - \mathbf{x}_i^\top \beta}{\sigma}\right) \frac{x_{ij}}{\sigma} = 0, \quad j = 1, \dots, p,$$

que puede ser escrito en forma compacta como

$$\frac{1}{\sigma^2} \sum_{i=1}^n W_i (Y_i - \mathbf{x}_i^\top \beta) \mathbf{x}_i = \mathbf{0} \quad (5)$$

Sea $\mathbf{W} = \text{diag}(W_1, \dots, W_n)$ de este modo podemos escribir (5) como:

$$\mathbf{X}^\top \mathbf{W} (\mathbf{Y} - \mathbf{X} \beta) = \mathbf{0}.$$

Sin embargo, debemos resaltar que los pesos W_i dependen de z_i , los que a su vez dependen de β y por tanto se requiere métodos iterativos para obtener $\hat{\beta}_M$.



Usando una estimación inicial $\beta^{(0)}$ podemos usar la iteración

$$\beta^{(r+1)} = (\mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{Y}, \quad (6)$$

con

$$\mathbf{W}^{(r)} = \text{diag}(W_1^{(r)}, \dots, W_n^{(r)}), \quad W_i^{(r)} = \psi(e_i^{(r)}/\sigma)/(e_i^{(r)}/\sigma),$$

y

$$\mathbf{e}^{(r)} = \mathbf{Y} - \mathbf{X}\beta^{(r)}.$$

Desde el punto de vista computacional es preferible usar

$$\beta^{(r+1)} = \beta^{(r)} + \mathbf{p}_r,$$

con

$$\mathbf{p}_r = (\mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{e}^{(r)}.$$

Observación:

El procedimiento delineado en (6) es conocido como **mínimos cuadrados iterativamente ponderados (IRLS)**.

La convergencia de la iteración en (6) sólo es garantizada para funciones ρ **convexas** y para funciones de **redescenso**.



Existe una gran variedad de funciones ρ o ψ para definir *M*-estimadores, por ejemplo:

► Tukey's biweight

$$\psi(z) = z[1 - (t/k)]_+^2, \quad k = 4.685.$$

► Hampel's ψ

$$\psi(z) = \operatorname{sgn}(z) \begin{cases} |z|, & 0 < |z| \leq a, \\ a, & a < |z| \leq b, \\ a(c - |z|)/(c - b), & b < |z| \leq c, \\ 0, & c < |z|, \end{cases}$$

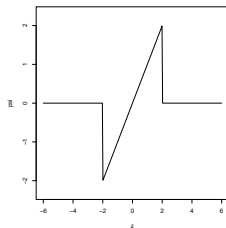
con $a = 1.645$, $b = 3$ y $c = 6.5$.

► Función seno de Andrews

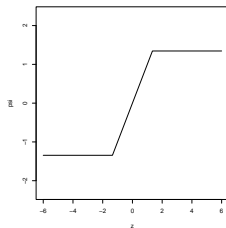
$$\psi(z) = \begin{cases} \sin(z/a), & |z| \leq \pi a, \\ 0, & |z| > \pi a, \end{cases}$$

donde $a = 1.339$.

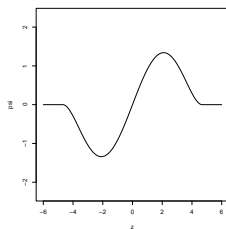




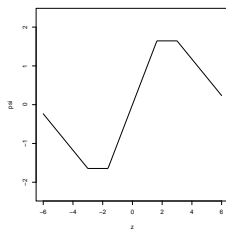
(a) media recortada



(b) Huber



(c) Tukey bisquare



(d) Hampel

Se ha sugerido usar el siguiente estimador para σ ,

$$\hat{\sigma}_{\text{rob}} = \frac{\text{MAD}(e)}{0.6745},$$

con

$$\text{MAD}(e) = \text{me}(|e - \text{me}(e)|),$$

que es conocido como **desviación mediana absoluta** de los residuos mínimos cuadrados $e = \mathbf{Y} - \mathbf{X}\hat{\beta}_{\text{LS}}$.

Análogamente a β es posible obtener un M -estimador para σ resolviendo

$$\frac{1}{n-p} \sum_{i=1}^n \chi\left(\frac{Y_i - \mathbf{x}_i^{\top} \beta}{\sigma}\right) = \delta,$$

para δ una constante positiva.



Sea

$$\kappa = 1 + \frac{p}{n} \frac{\text{var}(\psi')}{\{\mathbb{E}(\psi')\}^2},$$

que es evaluado en la distribución de los errores ϵ (en la práctica deben ser estimados desde los residuos). Entonces la matriz de covarianza asintótica de $\hat{\beta}_M$ es dada por (ver Huber, 1981)

$$\kappa^2 \frac{\sum_i \psi^2(e_i)/(n-p)}{[\sum_i \psi'(e_i)/n]^2} (\mathbf{X}^\top \mathbf{X})^{-1}.$$



Podemos considerar la estimación norma- L_p en la clase de los M -estimadores, como la solución del problema

$$\min_{\beta} \left(\sum_{i=1}^n |Y_i - \mathbf{x}_i^{\top} \beta|^p \right)^{1/p}, \quad p \geq 1. \quad (7)$$

En efecto, basta considerar

$$\rho(z) = |z|^p, \quad \psi(z) = |z|^{p-2}, \quad W(z) = |z|^{p-2}.$$

Observación:

Considere Y_1, \dots, Y_n variables aleatorias provenientes de la función de densidad

$$f(y) = c \exp(-|y|^p), \quad p \geq 1,$$

con c una constante de normalización. La distribución Laplace es obtenida para $p = 1$, mientras que la distribución normal es recuperada para $p = 2$.



Para $1 < p < 2$ podemos usar IRLS como un método para aproximar la solución del problema en (7). En efecto, Osborne (1985)⁴ reestableció el problema de estimación norma- L_p como:

$$\min_{\beta} Q_p(\beta), \quad Q_p(\beta) = \sum_{i=1}^n |e_i|^p = \sum_{i=1}^n |e_i|^{p-2} e_i^2,$$

que puede ser interpretado como un problema de mínimos cuadrados ponderados,

$$\min_{\beta} \|W^{(p-2)/2}(Y - X\beta)\|_2^2, \quad W = \text{diag}(|e_1|, \dots, |e_n|).$$

Esto lleva al siguiente algoritmo.

⁴Finite Algorithms in Optimization and Data Analysis. Wiley, New York.



Algoritmo 1: IRLS para estimación L_p .

Entrada: Datos \mathbf{X} , \mathbf{y} , $p \in (1, 2)$ y estimación inicial $\beta^{(0)}$.

Salida : Aproximación del estimador L_p , $\hat{\beta}$.

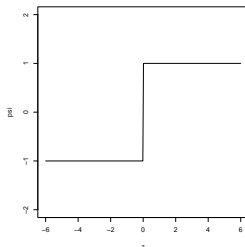
```
1 begin
2   for  $r = 0, 1, 2, \dots$  do
3      $\mathbf{e}^{(r)} = \mathbf{Y} - \mathbf{X}\beta^{(r)}$ 
4      $\mathbf{W}^{(r)} = \text{diag}(|e_1^{(r)}|^{(p-2)/2}, \dots, |e_n^{(r)}|^{(p-2)/2})$ 
5     Resolver  $\mathbf{p}_r$  desde
6        $\min_{\mathbf{p}_r} \|\mathbf{W}^{(r)}(\mathbf{e}^{(r)} - \mathbf{X}\mathbf{p}_r)\|_2^2$ 
7     Hacer
8        $\beta^{(r+1)} = \beta^{(r)} + \mathbf{p}_r$ 
9   end
10  return  $\hat{\beta} = \beta^{(*)}$ 
11 end
```

Estimación norma- L_1

Para $p = 1$, tenemos

$$\psi(z) = z/|z|.$$

con



De este modo, el procedimiento de estimación norma- L_1 es robusto, con pesos

$$W_i = 1/|e_i|, \quad i = 1, \dots, n.$$



Schlossmacher (1973)⁵ propuso usar IRLS para estimación norma- L_1 en regresión basado en la ecuación

$$\sum_{i=1}^n \frac{Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{|Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}|} \mathbf{x}_i = \mathbf{0}.$$

Sin embargo este procedimiento ha sido fuertemente criticado por una gran cantidad de autores (ver discusión en Capítulo 4 de Björk, 1996)⁶.

⁵Journal of the American Statistical Association 68, 857-859.

⁶Numerical Methods for Least Squares Problems. SIAM, Philadelphia.



Charnes, Cooper y Ferguson (1955)⁷ mostraron que problema de regresión L_1

$$\min_{\beta} Q_1(\beta), \quad Q_1(\beta) = \sum_{i=1}^n |Y_i - \mathbf{x}_i^\top \beta|,$$

es equivalente a resolver el siguiente problema de **programación lineal (LP)**

$$\begin{aligned} \min \quad & \sum_{i=1}^n (\epsilon_i^+ + \epsilon_i^-), \\ \text{sujeto a: } \quad & \epsilon_i^+ \geq 0, \quad \epsilon_i^- \geq 0, \\ & \epsilon_i^+ - \epsilon_i^- = \epsilon_i = Y_i - \mathbf{x}_i^\top \beta, \end{aligned}$$

para $i = 1, \dots, n$, con ϵ_i^+ y ϵ_i^- variables no negativas.

Observación:

Barrodale y Roberts (1973)⁸ presentan un algoritmo de propósito especial para resolver este problema modificando el **método simplex** y la **estructura de datos** requerida.

⁷Management Science 1, 138-151

⁸SIAM Journal on Numerical Analysis 10, 839-848.



Llamadas telefónicas en Bélgica 1950-1973

```
# carga datos de llamadas
> library(MASS)
> data(phones)

# Ajuste mínimos cuadrados
> f0 <- lm(calls ~ year, data = phones, x = TRUE, y = TRUE)

# Salida
> summary(f0)
```

```
Call:
lm(formula = calls ~ year, data = phones)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-78.97 -33.52 -12.04  23.38 124.20
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -260.059    102.607   -2.535   0.0189 *
year           5.041      1.658    3.041   0.0060 **
```

```
Residual standard error: 56.22 on 22 degrees of freedom
Multiple R-squared:  0.2959,    Adjusted R-squared:  0.2639
F-statistic: 9.247 on 1 and 22 DF,  p-value: 0.005998
```



```
# carga biblioteca L1pack (requiere fastmatrix)
> library(L1pack)

# extrae 'x' e 'y'
> x <- f0$x
> y <- f0$y

# ajuste usando regresión L1
> f1 <- l1fit(x, y, intercept = FALSE)
Warning message:
In l1fit(x, y, intercept = FALSE) : Non-unique solution possible
```



Llamadas telefónicas en Bélgica 1950-1973

```
# Salida
```

```
> f1
```

```
$coefficients
```

```
(Intercept)      year  
      -75.19       1.53
```

```
$minimum
```

```
[1] 844
```

```
$fitted.values
```

1	2	3	4	5	6	7	8	9	10	11
1.31	2.84	4.37	5.90	7.43	8.96	10.49	12.02	13.55	15.08	16.61
12	13	14	15	16	17	18	19	20	21	22
18.14	19.67	21.20	22.73	24.26	25.79	27.32	28.85	30.38	31.91	33.44
23	24									
34.97	36.50									

```
$residuals
```

[1]	3.09	1.86	0.33	0.00	-0.83	-1.66	-2.39	-3.22	-2.95
[10]	-3.08	-3.11	-3.24	-3.57	0.00	96.27	99.74	116.21	131.68
[19]	153.15	181.62	11.09	-9.44	-7.97	-7.50			

```
$rank
```

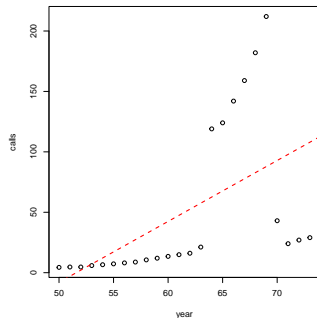
```
[1] 2
```

```
$numIter
```

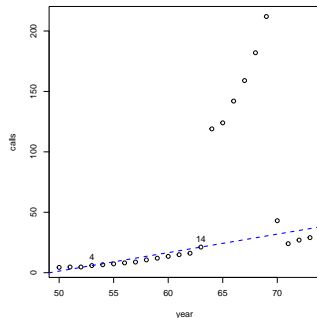
```
[1] 2
```



Llamadas telefónicas en Bélgica 1950-1973



(a) estimación LS



(b) estimación L_1



Llamadas telefónicas en Bélgica 1950-1973

```
# Huber, k = 1.345
> f2 <- rlm(calls ~ year, data = phones, psi = psi.huber, maxit = 50)
> f2
Call:
rlm(formula = calls ~ year, data = phones, psi = psi.huber, maxit = 50)
Converged in 33 iterations
```

```
Coefficients:
(Intercept)      year
-102.62220      2.04135
```

```
Degrees of freedom: 24 total; 22 residual
Scale estimate: 9.03
```

```
# Tukey's bisquare, k = 4.685
> f3 <- rlm(calls ~ year, data = phones, psi = psi.bisquare)
> f3
Call:
rlm(formula = calls ~ year, data = phones, psi = psi.bisquare)
Converged in 10 iterations
```

```
Coefficients:
(Intercept)      year
-52.302456      1.098041
```

```
Degrees of freedom: 24 total; 22 residual
Scale estimate: 1.65
```



Llamadas telefónicas en Bélgica 1950-1973

```
# Salida
> summary(f3)

Call: rlm(formula = calls ~ year, data = phones, psi = psi.bisquare)
Residuals:
    Min       1Q   Median       3Q      Max
-1.6585  -0.4143   0.2837  39.0866 188.5376

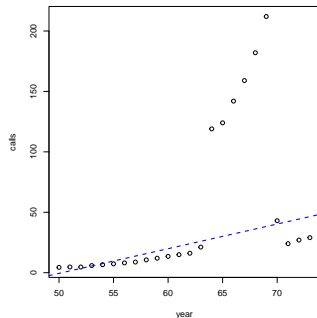
Coefficients:
            Value      Std. Error t value
(Intercept) -52.3025     2.7530  -18.9985
year          1.0980     0.0445   24.6846

Residual standard error: 1.654 on 22 degrees of freedom

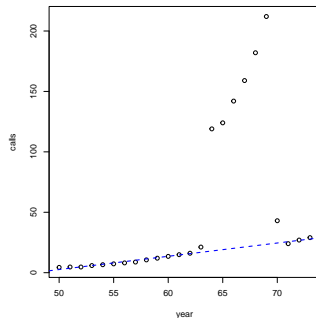
# 'pesos' estimados
> f3$w
 [1] 0.8948 0.9667 0.9997 0.9999 0.9949 0.9794 0.9610 0.9280 0.9797
[10] 0.9923 0.9998 0.9983 0.9965 0.4739 0.0000 0.0000 0.0000 0.0000
[20] 0.0000 0.0000 0.0000 0.9105 0.9980 0.9568
```



Llamadas telefónicas en Bélgica 1950-1973



(a) Huber, $k = 1.345$



(b) bisquare, $k = 4.685$