**Abstract**

Usualmente hay dos inconvenientes que presenta el estimador de cuadrados minimos cuando la matriz de covarianza esta mal condicionada: (a) *error de predicción alto*, ya que es insesgado pero con alta varianza y (b) *interpretabilidad*, ya que todos los coeficientes son diferentes de cero. El estimador LASSO propuesto por Tibshirani(1996), al imponer una restricción a los coeficientes, resuelve estos problemas obteniendo: un estimador sesgado pero con menor varianza, de manera de lograr un menor error de predicción, Además el hecho que este estimador pueda asignar coeficientes nulos mejora su interpretabilidad, permitiendo eliminar variables innecesarias. En este trabajo se propone una versión robusta del estimador LASSO, evaluando su desempeño tanto en el caso que lo datos no contienen outliers como cuando los contienen.

# Estimadores Lasso de Tipo M

Virgilio L. Foglia

November 14, 2012

## Contents

# 1 Modelo Lineal

Sea $\mathbf{y} = (y_1, ...y_n)'$ el vector de respuestas, $\mathbf{X} = [\mathbf{j}, \mathbf{X}_r]$ donde $\mathbf{j} \in \mathbb{R}^n$ y tiene todas sus componentes iguales a 1 y $\mathbf{X}_r = (x_{ij})$, $1 \leq i \leq n$, $1 \leq j \leq p$ es la matriz de diseño, siendo el valor $x_{ij}$ la $i$-esima observación de la variable $j$. Consideremos el modelo lineal (ML)

$$\mathbf{y} = \boldsymbol{\mu}_0 + \boldsymbol{\varepsilon}\text{: } \boldsymbol{\mu}_0 = \mathbf{X}\boldsymbol{\beta}_0$$

donde $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, ..., \varepsilon_n)' \in \mathbb{R}^n$ es un vector de variables aleatorias independientes con $E(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = \sigma^2$ para $1 \leq i \leq n$ y $\boldsymbol{\beta}_0 \in \mathbb{R}^{p+1}$ es el vector de

coeficientes de regresión. Supondremos que el rango($\mathbf{X}$) $= p+1$ lo que permite que $\boldsymbol{\beta}_0$ sea identificable. Los elementos de la matriz $\mathbf{X}_r$ pueden ser números fijos ( no aleatorios) o variables aleatorias. En este ultimo caso supondremos que la matriz $\mathbf{X}_r$ es independiente del vector $\boldsymbol{\varepsilon}$.

### 1.1 Estimador de cuadrados mínimos

Notar que $\boldsymbol{\mu}_0$=$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}_0 \in V$ donde $V$ es el subespacio de $\mathbb{R}^n$ generado por las columnas de $\mathbf{X}$. El estimador de mínimos cuadrados se define como

$$\widehat{\boldsymbol{\beta}}_{ls} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

y por lo tanto $\widehat{\boldsymbol{\mu}}_{ls} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{ls}$ satisface

$$\widehat{\boldsymbol{\mu}}_{ls} = \arg\min_{\mu \in V} \|\mathbf{y} - \boldsymbol{\mu}\|^2 ,$$

donde $V = \{\boldsymbol{\mu} : \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}\}$.

Utilizando el teorema de Gauss-Markov se demuestra que tanto $\widehat{\boldsymbol{\beta}}_{ls}$ como $\widehat{\boldsymbol{\mu}}_{ls}$, son los estimadores de menor varianza en la familia de estimadores lineales e insesgados. Se designan entonces con el nombre de estimadores BLUE(Best Linear Unbiased Estimator ). Ademas, si se incorpora la hipótesis de que los $\varepsilon_i$ tienen distribución normal, tanto $\widehat{\boldsymbol{\beta}}_{ls}$ como $\widehat{\boldsymbol{\mu}}_{ls}$, son los estimadores de menor varianza uniformemente en la familia más amplia de los estimadores insesgados (estimadores INVU).

Sin embargo como veremos más adelante, cuando la matriz $\mathbf{X}$ está mal condicionada, estimadores sesgados pueden ser preferibles.

Entre estos están los que surgen de la selección de variables, y los que imponen una contracción("shrinkage") a los coeficientes, como el estimador ridge o el estimador LASSO introducido por Tibshirani (1996).

Las demostraciones de los Teoremas y Lemas se encuentrán en la Sección 11.

## 2 Estimador LASSO

Tishbirani (1996) definió el estimado LASSO (Least Absolute Shrinkage and Selection Operator) de la siguiente manera. Dado $t \geq 0$ el estimador LASSO de $\boldsymbol{\beta}_0$ se define por

$$\widehat{\boldsymbol{\beta}}_L = \arg\min_{\boldsymbol{\beta} \in B_t} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \ . \tag{1}$$

donde

$$B_t = \left\{\boldsymbol{\beta} = (\beta_0,\beta_1,...,\beta_p)' : \beta_0 \in \mathbb{R},\ |\beta_1| + ... + |\beta_p| \leq t\right\}$$

El correspondiente estimador de $\boldsymbol{\mu}_0$ está dado por

$$\widehat{\boldsymbol{\mu}}_L = \mathbf{X}\widehat{\boldsymbol{\beta}}_L$$

o equivalentemente

$$\widehat{\boldsymbol{\mu}}_L = \arg \min_{\boldsymbol{\mu} \in \mathbf{V}_t} ||\mathbf{y} - \boldsymbol{\mu}||^2,$$

donde

$$V_t = \{\boldsymbol{\mu} : \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta},\, \boldsymbol{\beta} \in B_t \}$$

**Theorem 1** *$V_t$ es un conjunto convexo.*

Cuando se usa el estimador de cuadrados minimos el residuo $\widehat{\boldsymbol{\varepsilon}}_{ls} = \mathbf{y} - \widehat{\boldsymbol{\mu}}_{ls}$ es siempre ortogonal a $V$. Sin embargo con el estimador LASSO esto no ocurre.

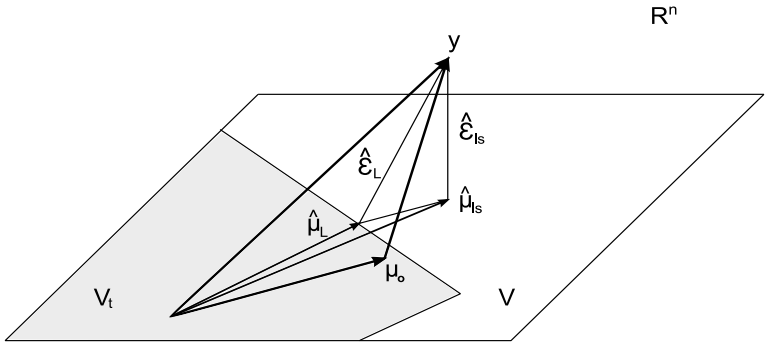


Figura 1: Cuadrados minimos y Lasso

**Remark 1** *El valor de $t \geq 0$ define el tamaño de la región $V_t$. Además, para cada posible valor de $t$ se obtendrá en general un estimador diferente, resultando $\widehat{\boldsymbol{\beta}}_L = \widehat{\boldsymbol{\beta}}_L(t)$. Mas adelante se analizará como elegir elvalor $t$, de manera de minimizar el error de predicción.*

## 2.1 Algunas propiedades del estimador LASSO

- Sea $\widehat{\boldsymbol{\mu}}_{ls} \in V$ el estimador de cuadrados mínimos y su correspondiente $\widehat{\boldsymbol{\beta}}_{ls}$. Se define

$$t_\infty = \sum_{j=1}^{p} \left| \widehat{\beta}_{jls} \right|.$$

  Notar que para una región de tamaño $t \geq t_\infty$, resultará $\widehat{\boldsymbol{\mu}}_{ls} \in V_t$. Como $\widehat{\boldsymbol{\mu}}_{ls}$ minimiza la distancia de $\mathbf{y}$ a todo $V$, en particular minimizará la distancia a $V_t$, resultando en este caso $\widehat{\boldsymbol{\mu}}_{ls} = \widehat{\boldsymbol{\mu}}_L$. Luego se tendrá que

$$t \geq t_\infty \Longrightarrow \widehat{\boldsymbol{\mu}}_L = \widehat{\boldsymbol{\mu}}_{ls},\ \widehat{\boldsymbol{\beta}}_L = \widehat{\boldsymbol{\beta}}_{ls}.$$

- Similarmente al caso de mínimos cuadrados, si $\widehat{\boldsymbol{\beta}}_{rL}$ es el estimador LASSO que corresponde a las variables predictoras de la matriz $\mathbf{X}_r$, el estimador del intercept será

$$\widehat{\beta}_{0L} = \overline{y} - \overline{\mathbf{x}}_r \widehat{\boldsymbol{\beta}}_{rL},$$

donde

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

y

$$\overline{\mathbf{x}}_r = (\overline{x}_r^{(1)}, ..., \overline{x}_r^{(p)}),$$

donde

$$\overline{x}_r^{(j)} = \frac{1}{n}\sum_{i=1}^{n} x_{rij}.$$

De aquí surge que si el estimador LASSO se aplica sobre una matriz estandarizada de manera que la media de cada columna sea 0, el estimador del intercept será siempre $\overline{y}$ sin importar la restricción del LASSO

- **Sesgo**: Para $t < t_\infty$, $\widehat{\boldsymbol{\mu}}_L$ y $\widehat{\boldsymbol{\beta}}_L$ son estimadores sesgados. Supóngase que para algún $t$, $\boldsymbol{\mu}_0 \notin V_t$. Como siempre $\widehat{\boldsymbol{\mu}}_L \in V_t$, y $V_t$ es convexo, resultará $E(\widehat{\boldsymbol{\mu}}_L) \in V_t$. Luego si $\widehat{\boldsymbol{\mu}}_L$ fuese insesgado $E(\widehat{\boldsymbol{\mu}}_L) = \boldsymbol{\mu}_0 \in V_t$ (contradicción). Con argumentos similares también resulta $\widehat{\boldsymbol{\beta}}_L$ sesgado.

Si bién el estimador LASSO es sesgado, eligiendo convenientemente $t$, se podrá obtener en muchos casos un estimador que a pesar de ser sesgado, tiene un el error cuadrático medio de predicción menor que el corresponde al estimador de mínimos cuadrados.

## 2.2 Estimador LASSO con matriz estandarizada

En lo que sigue, antes de aplicar el estimador LASSO se estandarizará la matriz $\mathbf{X}_r$, de manera que todas sus columnas tengan media 0 y varianza igual a 1. Llamemos $\mathbf{Z}_r$ a la matriz estandarizada de esta manera. O sea el modelo lineal sobre el cual se aplicará el LASSO será

$$\mathbf{y} = \mathbf{j}\beta_0^z + \mathbf{Z}_r\boldsymbol{\beta}_r^z + \boldsymbol{\varepsilon}.$$

De aquí se obtiene el estimador LASSO $\widehat{\boldsymbol{\beta}}_L^z = \left(\widehat{\beta}_{0L}^z, \widehat{\boldsymbol{\beta}}_{rL}^{z\prime}\right)' = \left(\overline{y}, \widehat{\boldsymbol{\beta}}_{rL}^{z\prime}\right)'$. Luego, si se designa $\overline{\mathbf{x}}_r = (\overline{x}_r^{(1)}, ..., \overline{x}_r^{(p)})$ y $\mathbf{S} \in \mathbb{R}^{pxp}$ diagonal, con $s_{jj} = sd(\mathbf{x}_r^{(j)})$, el estimador de $\boldsymbol{\beta}_L$ correspondientes a la matriz $\mathbf{X}$ se obtienen de la siguiente manera

$$\begin{array}{l} \widehat{\boldsymbol{\beta}}_{rL} = \mathbf{S}^{-1}\widehat{\boldsymbol{\beta}}_{rL}^z \\ \widehat{\beta}_{0L} = \overline{y} - \overline{\mathbf{x}}_r\widehat{\boldsymbol{\beta}}_{rL}^z \end{array} . \tag{2}$$

## 2.3 Interpretación geométrica

Para facilitar la interpretación se analizará el caso de dos predictores, es decir $\mathbf{X}_r = \left[\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right]$ y el ML

$$\mathbf{y} = \mathbf{j}\beta_0 + \mathbf{x}^{(1)}\beta_1 + \mathbf{x}^{(2)}\beta_2 + \boldsymbol{\varepsilon}.$$

El correspondiente modelo estandarizado será

$$\mathbf{y} = \mathbf{j}\beta_0^z + \mathbf{z}^{(1)}\beta_1^z + \mathbf{z}^{(2)}\beta_2^z + \boldsymbol{\varepsilon}.$$

Como cualquiera sea la restricción $t$ del LASSO, resulta siempre $\widehat{\beta}_{L0}^z = \overline{y}$, se considerará $\mathbf{y}^* = \mathbf{y} - \mathbf{j}\overline{y}$ y el modelo

$$\mathbf{y}^* = \mathbf{z}^{(1)}\beta_1^z + \mathbf{z}^{(2)}\beta_2^z + \boldsymbol{\varepsilon}.$$

Si $V_r$ es el espacio generado por $\mathbf{z}^{(1)}$ y $\mathbf{z}^{(2)}$, entonces para el LASSO con restricción $t$, la región $V_{rt}$ será

$$V_{rt} = \left\{ \boldsymbol{\mu} = \mathbf{z}^{(1)}\beta_1 + \mathbf{z}^{(2)}\beta_2 : |\beta_1| + |\beta_2| \leq t \right\} \subset V_r.$$

### 2.3.1 Región $V_{rt}$ cuando $\mathbf{z}^{(1)}$ es ortogonal a $\mathbf{z}^{(2)}$

En la Figura 2 se representa esta región para $t = 2$, junto con $\mathbf{z}^{(1)}$ y $\mathbf{z}^{(2)}$. Figuran además, $\boldsymbol{\mu}_r^z$ y el estimador de cuadrados mínimos $\widehat{\boldsymbol{\mu}}_{rls}^z = \Pr(\mathbf{y}^*/V_r)$. Notese que las coordenadas de $\widehat{\boldsymbol{\mu}}_{rls}^z$ en $\mathbf{z}^{(1)}$ y $\mathbf{z}^{(2)}$, son los coeficientes $\widehat{\beta}_{1ls}^z$ y $\widehat{\beta}_{2ls}^z$.

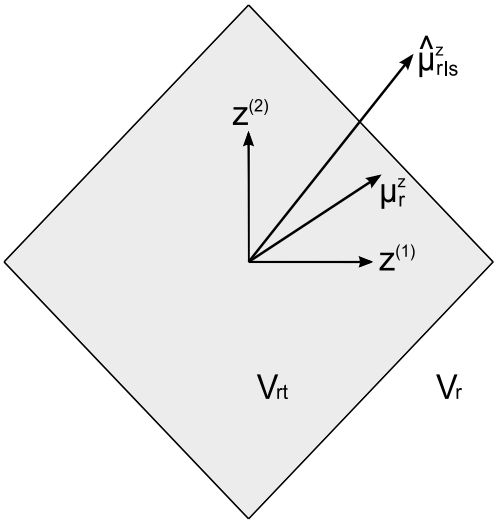


Figura 2: region $V_{rt}$ para t=2, y rho($z^{(1)}$;$z^{(2)}$) = 0

**Estimador de cuadrados mínimos y estimador LASSO: Caso 1** Cuando el estimador de cuadrados mínimos $\widehat{\boldsymbol{\mu}}^{z}_{rls}$ cae en una zona 1 como se representa en la Figura 3, el estimador LASSO $\widehat{\boldsymbol{\mu}}^{z}_{rL}$ correspondiente se encuentra reducido y coincide con el pié de la perpendicular de $\widehat{\boldsymbol{\mu}}^{z}_{rls}$ a la región $V_{rt}$. En este caso las coordenadas de $\widehat{\boldsymbol{\mu}}^{z}_{rL}$ en $\mathbf{z}^{(1)}$ y $\mathbf{z}^{(2)}$, son los coeficientes del estimador LASSO $\widehat{\beta}^{z}_{1L}$ y $\widehat{\beta}^{z}_{2L}$.

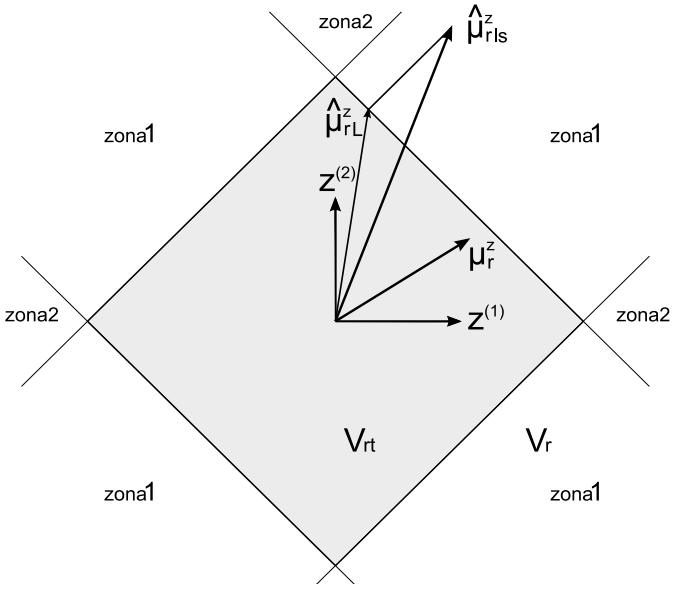


Figura 3: estimador Lasso en zona 1

Observando que $\widehat{\boldsymbol{\mu}}^{z}_{rls} - \widehat{\boldsymbol{\mu}}^{z}_{rL}$ está en la dirección de la bisectriz entre $\mathbf{z}^{(1)}$ y $\mathbf{z}^{(2)}$, entonces para cierto $\gamma > 0$ resultará $\widehat{\boldsymbol{\mu}}^{z}_{rls} - \widehat{\boldsymbol{\mu}}^{z}_{rL} = \gamma\mathbf{z}^{(1)} + \gamma\mathbf{z}^{(2)}$, y como $\widehat{\boldsymbol{\mu}}^{z}_{rls} = \mathbf{z}^{(1)}\widehat{\beta}^{z}_{1ls} + \mathbf{z}^{(2)}\widehat{\beta}^{z}_{2ls}$ y $\widehat{\boldsymbol{\mu}}^{z}_{rL} = \mathbf{z}^{(1)}\widehat{\beta}^{z}_{1L} + \mathbf{z}^{(2)}\widehat{\beta}^{z}_{2L}$ despejando resulta

$$\widehat{\boldsymbol{\mu}}^{z}_{rL} = \mathbf{z}^{(1)}(\widehat{\beta}^{z}_{1ls} - \gamma) + \mathbf{z}^{(2)}(\widehat{\beta}^{z}_{2ls} - \gamma),$$

de manera que

$$\begin{aligned} \widehat{\beta}^{z}_{1L} &= \widehat{\beta}^{z}_{1ls} - \gamma, \\ \widehat{\beta}^{z}_{2L} &= \widehat{\beta}^{z}_{2ls} - \gamma. \end{aligned}$$

Luego los coeficientes del estimador LASSO se encuentran disminuidos en una misma cantidad $\gamma > 0$ respecto de los obtenidos mediante cuadrados mínimos. Lo anterior se cumple si $\widehat{\boldsymbol{\mu}}^{z}_{rls}$ cae en la zona 1 del primer cuadrante. Pero si esta en otra zona 1 de otro cuadrante resultará $\widehat{\boldsymbol{\mu}}^{z}_{rls} - \widehat{\boldsymbol{\mu}}^{z}_{rL} = \gamma_1\mathbf{z}^{(1)} + \gamma_2\mathbf{z}^{(2)}$ con $|\gamma_1| = |\gamma_2| = \gamma > 0$, pero ahora $\gamma_1$ y $\gamma_2$ no son ambos positivos. En general para

cualquier cuadrante el signo del estimador LASSO coincide con el de cuadrados mínimos, resultando siempre $\left|\widehat{\beta}^z_{jls}\right| - \gamma \geq 0$, luego según Tibshirani

$$\widehat{\beta}^z_{jL} = \text{sg}(\widehat{\beta}^z_{jls})(\left|\widehat{\beta}^z_{jls}\right| - \gamma)^+.$$

Entonces $\left|\widehat{\beta}^z_{jL}\right| = \left|\widehat{\beta}^z_{jls}\right| - \gamma$, y si se suma resulta $t = \sum_{j=1}^{2} \left|\widehat{\beta}^z_{jL}\right| = \sum_{j=1}^{2} \left|\widehat{\beta}^z_{jls}\right| - 2\gamma = t^z_\infty - 2\gamma$. Luego

$$\gamma = \begin{cases} \frac{t^z_\infty - t}{2} & \text{si} \quad t < t^z_\infty \\ 0 & \text{si} \quad t \geq t^z_\infty \end{cases}$$

Notese que a medida que $t$ se reduce, disminuye la región $V_{rt}$ y aumenta $\gamma$, y en algún momento uno de los $\widehat{\beta}^z_{jL}$ será nulo. Pero entonces $\widehat{\boldsymbol{\mu}}^z_{rls}$ estará en zona 2, y cualquier ulterior reducción de $t$ mantendrá ese coeficiente LASSO en cero. Esto se analizará con el Caso 2.

**Remark 2** *Generalizando al caso de p predictores ortogonales, cuando* $\widehat{\boldsymbol{\mu}}^z_{rls} \notin V_{rt}$ *y cae en zonas equivalentes a la zona1 pero en p dimensiones, los estimadores LASSO de beta coinciden en signo con los de cuadrados mínimos, pero su magnitud se reduce en una misma cantidad aditiva* $\gamma > 0$ *que depende del* $t^z_\infty$ *de cuadrados mínimos y de la restricción t. Notar que al disminuir en una misma cantidad cada* $\left|\widehat{\beta}^z_{jls}\right|$*, en términos relativos el estimador LASSO penaliza más los* $\widehat{\beta}^z_{jls}$ *de menor valor absoluto. Esto difiere con respecto al estimador ridge en que cada coeficiente se reduce proporcionalmente. Por supuesto lo anterior en el caso planteado, ya que si el estimador de cuadrados mínimos* $\widehat{\boldsymbol{\mu}}^z_{rls}$ *cae en* $V_{rt}$*, en este caso* $\widehat{\boldsymbol{\mu}}^z_{rL} = \widehat{\boldsymbol{\mu}}^z_{rls}$*. Esto se verá con detalle en el Teorema 3.*

**Estimador de cuadrados mínimos y estimador LASSO: Caso 2** Cuando el estimador de cuadrados mínimos $\widehat{\boldsymbol{\mu}}^z_{rls}$ cae en una zona2 como se representa en la Figura 4, en el estimador LASSO, el valor de $\widehat{\boldsymbol{\mu}}^z_{rL}$ coincide con una semidiagonal de la región $V_{rt}$, siendo en este caso $\widehat{\boldsymbol{\mu}}^z_{rL} = \mathbf{z}^{(1)}t$, resultando entonces $\widehat{\beta}^z_{1L} = t$ y $\widehat{\beta}^z_{2L} = 0$.

Notar que si $\beta^z_2 = 0$, entonces con una adecuada restricción $t$, muy probablemente $\widehat{\boldsymbol{\mu}}^z_{rls}$ caiga en zona2, resultando $\widehat{\beta}^z_{2L} = 0$; es decir el estimador LASSO estimará exactamente el verdadero valor del parámetro. Esto no ocurre con cuadrados mínimos. Es justamente esta característica que hace útil al estimador LASSO como recurso para elegir un modelo más simple eliminando variables.

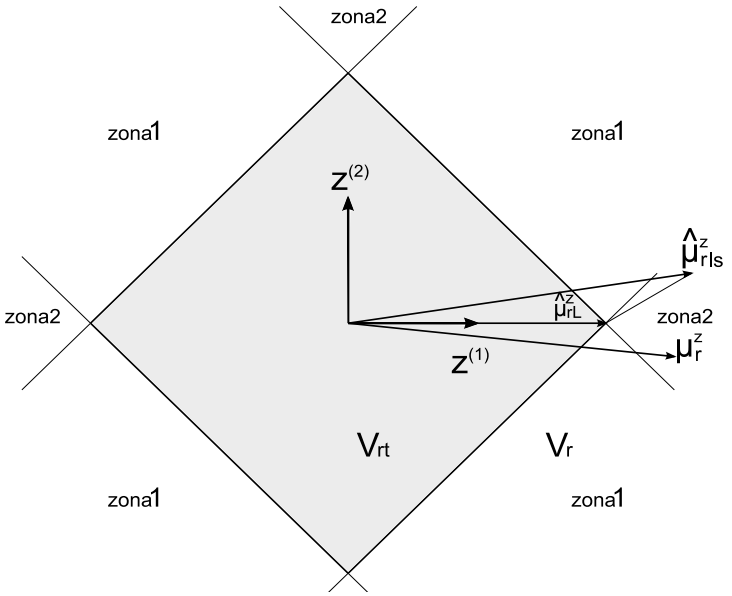


Figura 4: estimador Lasso en zona 2

### 2.3.2 Región $V_{rt}$ cuando hay correlación entre $\mathbf{z}^{(1)}$ y $\mathbf{z}^{(2)}$

En la Figura 5 se representa esta región suponiendo que $corr(\mathbf{z}^{(1)}; \mathbf{z}^{(2)}) = 0.8$ y la restricción $|\beta_1| + |\beta_2| \leq t$ con $t = 2$

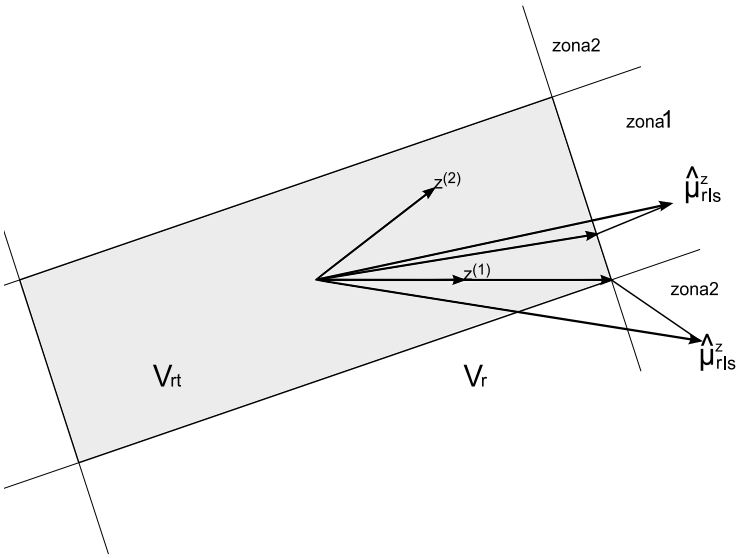


Figura 5: region $V_{rt}$ para t=2, y rho($z^{(1)}$;$z^{(2)}$) = 0.8

Cuando $\widehat{\boldsymbol{\mu}}^{z}_{rls}$ está en la zona1, también es cierto que $\widehat{\boldsymbol{\mu}}^{z}_{rls} - \widehat{\boldsymbol{\mu}}^{z}_{rL}$ está en la dirección de la bisectriz entre $\mathbf{z}^{(1)}$ y $\mathbf{z}^{(2)}$, y entonces también se cumplirá que los estimadores LASSO de beta coinciden en signo con los de cuadrados mínimos, con magnitud reducida en una misma cantidad aditiva $\gamma > 0$. Además cuando $\widehat{\boldsymbol{\mu}}^{z}_{rls}$ este en la zona2, alguno de los estimadores LASSO de los coeficientes será exactamente cero.

**Remark 3** *Como se ve, en el caso de dos predictores valen relaciones similares a las vistas en el caso de ortogonalidad. Sin embargo cuando $p > 2$ y hay correlación entre los predictores, no siempre es cierto que en la zona1 los estimadores LASSO se encuentren reducidos en magnitud en una misma cantidad aditiva $\gamma > 0$.*

## 2.4 El estimador LASSO para el caso de una matriz de diseño ortogonal

Considérese el modelo lineal estandarizado, donde $\mathbf{Z} \in \mathbb{R}^{n\times p}$ con $\mathbf{Z}'\mathbf{Z} = \mathbf{I}_p$ Luego

$$\mathbf{y} = \mathbf{j}\alpha_0 + \mathbf{Z}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}.$$

Como cualquiera sea la restricción $t$ del LASSO, resulta siempre $\widehat{\alpha}_{0L} = \widehat{\alpha}_{0ls} = \overline{y}$, se considerará $\mathbf{y}^* = \mathbf{y} - \mathbf{j}\overline{y}$ y entonces se cumple

$$\mathbf{y}^* = \mathbf{Z}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}.$$

En este caso el estimador LASSO de $\boldsymbol{\beta}_0$ está dado por

$$\widehat{\boldsymbol{\beta}}_L = \underset{|\beta_1|+\cdots+|\beta_p|\leq t}{\arg\min} \|\mathbf{y}^*-\mathbf{Z}\boldsymbol{\beta}\|^2 \ .$$

Sea $\widehat{\boldsymbol{\beta}}_{ls}$ es el estimador de cuadrados mínimos. Luego se tiene

$$\begin{aligned}
\|\mathbf{y}^*-\mathbf{Z}\boldsymbol{\beta}\|^2 &= \left\|\mathbf{y}^*-\mathbf{Z}\widehat{\boldsymbol{\beta}}_{ls}-\mathbf{Z}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\mathbf{ls}})\right\|^2 \\
&= \left\|\mathbf{y}^*-\mathbf{Z}\widehat{\boldsymbol{\beta}}_{ls}\right\|^2 + \left\|\mathbf{Z}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\mathbf{ls}})\right\|^2 \\
&= \left\|\mathbf{y}^*-\mathbf{Z}\widehat{\boldsymbol{\beta}}_{ls}\right\|^2 + (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\mathbf{ls}})'\mathbf{Z}'\mathbf{Z}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\mathbf{ls}}) \\
&= \left\|\mathbf{y}^*-\mathbf{Z}\widehat{\boldsymbol{\beta}}_{ls}\right\|^2 + \left\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\mathbf{ls}}\right\|^2,
\end{aligned}$$

donde la segunda igualdad se debe a la ortogonalidad de sus dos componentes. Según esto el LASSO se reduciría a

$$\widehat{\boldsymbol{\beta}}_L = \underset{|\beta_1|+\cdots+|\beta_p|\leq t}{\arg\min} \left\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\mathbf{ls}}\right\|^2 \ . \tag{3}$$

Se analizará otra expresión de la (3) para el caso en que opera la restricción, o sea cuando $t < t_\infty$. Se llamará $\delta = t_\infty - t$, y será siempre en este caso $\delta > 0$.

Se necesitarán los siguientes lemas

**Lemma 1** *Sea $\widehat{\boldsymbol{\beta}}_L$ solución de la (3) entonces*
***(a)*** *Sí $t \geq t_\infty \Longrightarrow \widehat{\boldsymbol{\beta}}_L = \widehat{\boldsymbol{\beta}}_{\mathbf{ls}}$*
***(b)*** *Sí $t = t_1 < t_\infty \Longrightarrow \widehat{\boldsymbol{\beta}}_L$ cumple $\left|\widehat{\beta}_{L1}\right| + \cdots + \left|\widehat{\beta}_{Lp}\right| = t_1$*

**Lemma 2** *Supongamos $Z'Z = I$. Sea $\widehat{\boldsymbol{\beta}}_L$ solución de la (3) entonces $\forall i$ con $1 \leq i \leq p$ se tiene*
***(a)*** *Sí $\widehat{\beta}_{lsi} = 0 \Longrightarrow \widehat{\beta}_{Li} = 0$*
***(b)*** *Sí $\widehat{\beta}_{lsi} \neq 0 \Longrightarrow \widehat{\beta}_{Li} = 0$ o $sg(\widehat{\beta}_{Li}) = sg(\widehat{\beta}_{lsi})$.*

**Lemma 3** *Supongamos $Z'Z = I$.Si $\widehat{\boldsymbol{\beta}}_L$ es solución de la (3) entonces $\forall i$ con $1 \leq i \leq p$*

$$\left|\widehat{\beta}_{Li}\right| \leq \left|\widehat{\beta}_{lsi}\right| \tag{4}$$

Se definen los $\delta_i = \left|\widehat{\beta}_{lsi}\right| - |\beta_i|$, que miden las reducciones en valor absoluto de los $\left|\widehat{\beta}_{lsi}\right|$.

Si se observa la (3) en $\left(\beta_1 - \widehat{\beta}_{ls1}\right)^2 + \cdots + \left(\beta_p - \widehat{\beta}_{lsp}\right)^2$ y ya que del Lema 2 en cualquier caso resultará $\widehat{\beta}_{Li} = 0$ o $\text{sg}(\widehat{\beta}_{Li}) = \text{sg}(\widehat{\beta}_{lsi})$ esto se puede escribir

$$(\left|\widehat{\beta}_{ls1}\right| - |\beta_1|)^2 + \cdots + (\left|\widehat{\beta}_{lsp}\right| - |\beta_p|)^2 = \delta_1^2 + \cdots + \delta_p^2,$$

o sea la minimización de (3) equivale a minimizar las reducciones.

Pero de acuerdo a (4) resultará $\delta_i \geq 0$, y del Lema 2 $\delta_i \leq \left|\widehat{\beta}_{lsi}\right|$. O sea estas reducciones cumplirán

$$0 \leq \delta_i \leq \left|\widehat{\beta}_{lsi}\right|$$

y además como

$$\sum_{i=1}^{p} \delta_i = \sum_{i=1}^{p} \left|\widehat{\beta}_{lsi}\right| - \sum_{i=1}^{p} |\beta_i| = t_\infty - t = \delta,$$

resultará también

$$\delta_1 + \cdot\cdot + \delta_p = \delta.$$

Luego el siguiente Teorema es inmediato

**Theorem 2** *Sea $Z'Z = I$. Luego encontrar la solución $\widehat{\boldsymbol{\beta}}_L$ equivale a encontrar las reducciones óptimas $\widehat{\boldsymbol{\delta}}_L$ definidas por*

$$\widehat{\boldsymbol{\delta}}_L = \underset{(\boldsymbol{\delta}_1,...,\delta_p)}{\arg\min} (\delta_1^2 + \cdot\cdot + \delta_p^2) \tag{5}$$

*donde $(\boldsymbol{\delta}_1, ..., \delta_p)$ satisface las siguientes restricciones*

$$\delta_1 + \cdot\cdot + \delta_p = \delta, \tag{6}$$

*y*

$$0 \leq \delta_i \leq \left|\widehat{\beta}_{lsi}\right|, \ 1 \leq i \leq p. \tag{7}$$

El siguiente teorema caracteriza el estimador LASSO en en este caso

**Theorem 3** *Sea $Z'Z = I$ y $t < t_\infty$. Supongamos que las observaciones están ordenadas de manera que $\left|\widehat{\beta}_{ls1}\right| < \left|\widehat{\beta}_{ls2}\right| < ... < \left|\widehat{\beta}_{lsp}\right|$. Luego*
***(a)*** *existe $j_0$, $1 \leq j_0 \leq p$, dependiente de $\delta$ tal que si definimos $\delta^*$ al valor dado por*

$$\delta^* = \frac{1}{(p - j_0 + 1)} (\delta - \sum_{i=1}^{j_0 - 1} \left|\widehat{\beta}_{lsi}\right|) \tag{8}$$

*entonces*

$$\widehat{\beta}_{Li} = \begin{cases} 0 & si \quad i < j_0 \\ sg(\widehat{\beta}_{lsi})(|\widehat{\beta}_{lsi}| - \delta^*) & si \quad i \geq j_0 \end{cases}.$$

***(b)*** *Se cumple que $\left|\widehat{\beta}_{ls,j_0-1}\right| \leq \delta^* < \left|\widehat{\beta}_{lsj_0}\right|$.*
***(c)*** *Sea $a_0 = 0$ y para $1 \leq j \leq p$ definamos*

$$a_j = \left|\widehat{\beta}_{lsj}\right| (p - j) + \sum_{i=1}^{j} \left|\widehat{\beta}_{lsi}\right|.$$

*Luego $a_p = t_\infty$ y $j_1 < j_2$ implica $a_{j_1} < a_{j_2}$ . Además el valor $j_0$ está univocamente determinado por $a_{j_0-1} \leq \delta < \alpha_{j_0}$.*

Este teorema muestra como aparecen los coeficientes iguales a cero en el estimador LASSO. Cuanto más chico sea $t$ (más grande $\delta$), mayor será el número de ceros. También se observa que los ceros estarán en aquellas coordenadas para las cuales el estimador de mínimos cuadrados tiene valores más chicos en valor absoluto.

**Ejemplo**. Para ejemplificar este Teorema *3* considérese el caso de $p = 4$, y con las componentes del estimador de mínimos cuadrados una vez ordenadas por valor absoluto dadas por $\left|\widehat{\beta}_{ls1}\right| = 2, \left|\widehat{\beta}_{ls2}\right| = 4, \left|\widehat{\beta}_{ls3}\right| = 8, \left|\widehat{\beta}_{ls4}\right| = 10$ o sea $t_\infty = 24$. Se quieren estudiar todas las soluciones cuando $0 < \delta \leq 24$. Observemos que en este caso $a_0 = 0$, $a_1 = 8$, $a_2 = 14$, $a_3 = 22$ y $a_4 = 24$. Entonces de acuerdo al Teorema 3 debemos considerar los suguientes cuatro casos.

**Caso1** Sea $0 \leq \delta < 8$, luego $j_0 = 1$, $\delta^* = \delta/4$ y

$$\beta_{Li} = \text{sg}(\widehat{\beta}_{lsi})(|\widehat{\beta}_{lsi}| - \delta/4)$$

**Caso 2** Sea $8 \leq \delta < 14$, luego $j_0 = 2$ , $\delta^* = (\delta - 2)/3$ y

$$\widehat{\beta}_{Li} = \begin{cases} 0 & \text{si} \quad i = 1 \\ \text{sg}(\widehat{\beta}_{lsi}) \left(|\widehat{\beta}_{lsi}| - \frac{\delta-2}{3}\right) & \text{si} \quad i > 1 \end{cases} .$$

**Caso 3** Sea $14 \leq \delta < 22$, luego $j_0 = 3$ , $\delta^* = (\delta - 6)/2$ y

$$\widehat{\beta}_{Li} = \begin{cases} 0 & \text{si} \quad i \leq 2 \\ \text{sg}(\widehat{\beta}_{lsi}) \left(|\widehat{\beta}_{lsi}| - \frac{\delta-6}{2}\right) & \text{si} \quad i > 2 \end{cases} .$$

**Caso 4** Sea $22 \leq \delta < 24$, luego $j_0 = 4$ , $\delta^* = \delta - 14$ y

$$\widehat{\beta}_{Li} = \begin{cases} 0 & \text{si} \quad i \leq 3 \\ \text{sg}(\widehat{\beta}_{lsi}) \left(|\widehat{\beta}_{lsi}| - (\delta - 14)\right) & \text{si} \quad i > 3 \end{cases} .$$

Usando estas fórmulas se obtiene la Figura 6 en donde se representa la variación

de los $\left|\widehat{\beta}_{Li}\right|$ en función de $\delta$

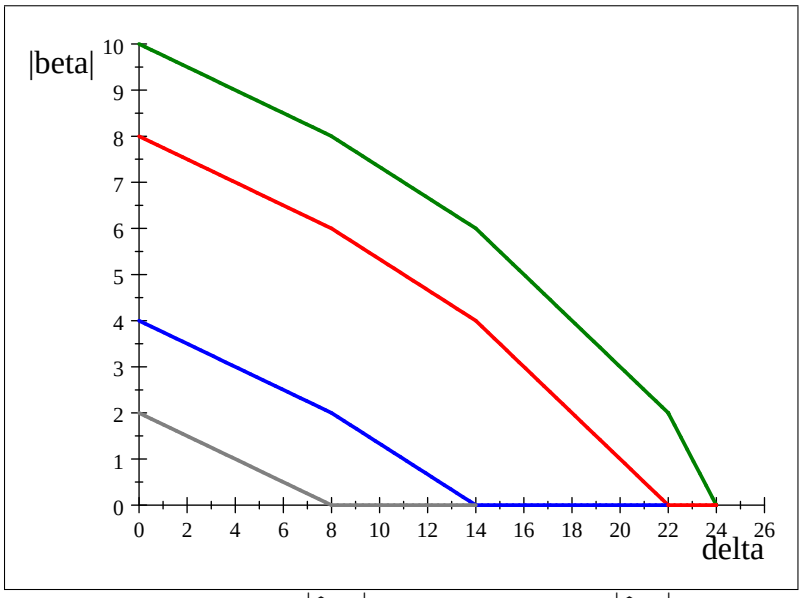


Figura 6: Variacion de $\left|\widehat{\beta}_{Li}\right|$ como funcion de $\delta$ : $\left|\widehat{\beta}_{L4}\right|$ verde, $\left|\widehat{\beta}_{L3}\right|$ rojo, $\left|\widehat{\beta}_{L2}\right|$ azul y $\left|\widehat{\beta}_{L1}\right|$ gris.

# 3 Regresión Ridge

Para un modelo lineal, y con el fin de comparar, se presentará ahora otro estimador de $\boldsymbol{\beta}_0$, también basado en una contracción de los coeficientes (Shrinkage). Dado $t > 0$ el estimador ridge de $\boldsymbol{\beta}_0$ se define por

$$\widehat{\boldsymbol{\beta}}_{RR} = \underset{||\boldsymbol{\beta}_r||_2^2 \leq t}{\arg\min} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 .$$

Si se la compara con (1), esta definición es similar a la del LASSO, cambiando solo la restricción: en el LASSO $\|\boldsymbol{\beta}_r\|_1 \leq t$, y con el estimador Ridge $\|\boldsymbol{\beta}_r\|_2 \leq \sqrt{t}$. El correspondiente estimador de $\boldsymbol{\mu}_0$ está dado por

$$\widehat{\boldsymbol{\mu}}_{RR} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{RR},$$

o equivalentemente

$$\widehat{\boldsymbol{\mu}}_{RR} = \arg \min_{\boldsymbol{\mu} \in \mathbf{V}_t} ||\mathbf{y} - \boldsymbol{\mu}||^2,$$

donde

$$V_t = \{\boldsymbol{\mu} : \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta},\ ||\boldsymbol{\beta}_r||^2 \leq t\}.$$

Sea $\widehat{\boldsymbol{\mu}}_{ls} \in V$ el estimador de cuadrados mínimos y su correspondiente $\widehat{\boldsymbol{\beta}}_{ls}$. Se define

$$t_{RR\infty} = \sum_{j=1}^{p} \widehat{\beta}_{jls}^{2} \, .$$

Debe notarse que si $t \geq t_{RR\infty}$, resultará $\widehat{\boldsymbol{\mu}}_{ls} \in V_t$. Como $\widehat{\boldsymbol{\mu}}_{ls}$ minimiza la distancia de $\mathbf{y}$ a todo $V$, en particular minimizará la distancia a $V_t$, resultando en este caso $\widehat{\boldsymbol{\mu}}_{ls} = \widehat{\boldsymbol{\mu}}_{RR}$. Luego se tendrá que

$$t \geq t_{RR\infty} \Longrightarrow \widehat{\boldsymbol{\mu}}_{RR} = \widehat{\boldsymbol{\mu}}_{ls}, \widehat{\boldsymbol{\beta}}_{RR} = \widehat{\boldsymbol{\beta}}_{ls}.$$

Además igual que con el LASSO, lo habitual es aplicar el el estimador ridge sobre la matriz $\mathbf{X}_r$ estandarizada, $\mathbf{Z}_r$, y luego utilizar expresiones similares a las (2) para des-estandarizar los estimadores.

### 3.0.1 Regressión Ridge con dos predictores estandarizados

Ahora se quiere estudiar el caso de solo dos predictores estandarizados, o sea el modelo

$$\mathbf{y} = \mathbf{j}\beta_0^z + \mathbf{z}^{(1)}\beta_1^z + \mathbf{z}^{(2)}\beta_2^z + \boldsymbol{\varepsilon}.$$

Como cualquiera sea el valor de $t$, resulta siempre $\widehat{\beta}^z_{RR0} = \overline{y}$, se considerará $\mathbf{y}^* = \mathbf{y} - \mathbf{j}\overline{y}$ y el modelo

$$\mathbf{y}^* = \mathbf{z}^{(1)}\beta_1^z + \mathbf{z}^{(2)}\beta_2^z + \boldsymbol{\varepsilon}.$$

Sea $V_r$ el subespacio generado pr $\mathbf{z}^{(1)}$ y $\mathbf{z}^{(2)}$, entonces para el estimador ridge la región $V_{rt}$ será

$$V_{rt} = \left\{ \boldsymbol{\mu} : \ \boldsymbol{\mu} = \mathbf{z}^{(1)}\beta_1 + \mathbf{z}^{(2)}\beta_2 \ , \ \beta_1^2 + \beta_2^2 \leq t \right\} \subset V_r.$$

**Región $V_{rt}$ cuando $\mathbf{z}^{(1)}$ es ortogonal a $\mathbf{z}^{(2)}$** En la Figura 7 se representa la región $V_{rt}$ para $t = 2$, cuando $\mathbf{z}^{(1)}$ es ortogonal a $\mathbf{z}^{(2)}$. En este caso la región resulta ser un círculo. Cuando $t = 2 < t^z_{RR\infty}$ el estimador de minimos cuadrados

es exterior a este círculo, o sea $\widehat{\boldsymbol{\mu}}^z_{rls} \notin V_{rt}$.

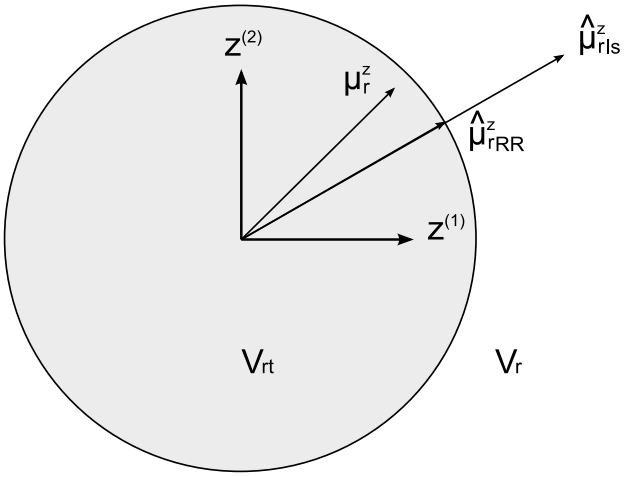


Figura 7: ridge-regression con rho($z^{(1)}$;$z^{(2)}$) = 0

Como el estimador ridge coincide con la mínima distancia de $\widehat{\boldsymbol{\mu}}^z_{rls}$ al círculo, resultará $\widehat{\boldsymbol{\mu}}^z_{rRR}$ colineal con el de mínimos cuadrados. Luego, como se demuestra en el siguiente teorema, los estimadores de los coeficientes $\widehat{\beta}^z_{1RR}$ y $\widehat{\beta}^z_{2RR}$ estarán reducidos respecto de $\widehat{\beta}^z_{1ls}$ y $\widehat{\beta}^z_{2ls}$ en una misma cantidad $k > 1$ multiplicativa, es decir:

$$\widehat{\beta}^z_{1RR} = \widehat{\beta}^z_{1ls}/k \ , \ \widehat{\beta}^z_{2RR} = \widehat{\beta}^z_{2ls}/k.$$

**Theorem 4** *Para p predictores ortogonales, con restricción t y* $\widehat{\boldsymbol{\mu}}^z_{rls} \notin V_{rt}$ *resulta que* $\widehat{\beta}^z_{iRR} = \widehat{\beta}^z_{ils}/k$ *con*

$$k = \frac{1}{\sqrt{t}} \left\| \widehat{\boldsymbol{\mu}}^z_{rls} \right\| \ .$$

Finalmente notese que el estimador ridge no soluciona el problema de interpretabilidad del modelo lineal ya que los coeficientes nunca se anulan(salvo el caso extremo $t = 0$). Además si $t \geq t^z_{RR\infty}$, resultará $\widehat{\boldsymbol{\mu}}^z_{rRR} = \widehat{\boldsymbol{\mu}}^z_{rls}$ y los estimadores Ridge coinciden con los de cuadrados mínimos, es decir $k = 1$. Por último, y siempre que $t < t^z_{RR\infty}$, aunque aumente el número de predictores la propiedad de proporcionalidad en la reducción de los coeficientes sigue siendo válida.

**Región $V_{rt}$ cuando hay correlación entre $\mathbf{z}^{(1)}$ y $\mathbf{z}^{(2)}$** En la Figura 8 se representa el caso en que $corr(\mathbf{z}^{(1)};\mathbf{z}^{(2)}) = 0.8$, y la región $V_{rt}$ para $t = 2$, suponiendo $t = 2 < t^{z}_{RR\infty}$ de manera que $\widehat{\boldsymbol{\mu}}^{z}_{rls} \notin V_{rt}$.

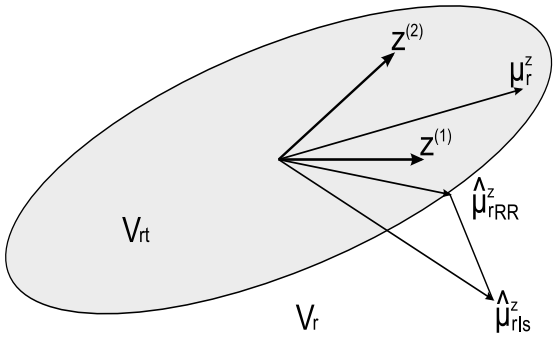


Figura 8: ridge-regression con rho($z^{(1)}$;$z^{(2)}$) = 0.8

Se observa que ahora, debido a la forma eliptica de la región $V_{rt}$, los coeficientes en general se reducen pero no en la misma proporción, pudiendo alguno incluso aumentar

# 4 El estimador LASSO como mínimos cuadrados penalizados

Según se vió en (1) y teniendo en cuenta que por el momento $t$ es un párametro fijo

$$\widehat{\boldsymbol{\beta}}_L = \underset{\sum_{j=1}^{p}|\beta_j|\leq t}{\arg\min} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \ ,$$

o equivalentemente

$$\widehat{\boldsymbol{\beta}}_L = \underset{\sum_{j=1}^{p}|\beta_j|\leq t}{\arg\min} \sum_{i=1}^{n}(y_i - \mathbf{x}_i\boldsymbol{\beta})^2, \tag{9}$$

donde $\mathbf{x}_i$ es la fila $i$ de la matriz $\mathbf{X}$. En el siguiente teorema se demuestra que (9) es equivalente a

$$\widehat{\boldsymbol{\beta}}_L = \arg\min_{\boldsymbol{\beta}} \left( \sum_{i=1}^{n} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \right) \tag{10}$$

para algún $\lambda$.Notar aquí que la suma de cuadrados de los residuos esta "penalizada" por el término $\lambda \sum_{j=1}^{p} \left|\beta_j\right|$.

**Theorem 5** *Si $\widehat{\boldsymbol{\beta}}^t_{L,1}$ es una solución de (9), y $\widehat{\boldsymbol{\beta}}^\lambda_{L,2}$ a una solución de (10), valen*

***(a)*** *Para todo $t \geq 0$ el estimador $\widehat{\boldsymbol{\beta}}^t_{L,1}$ existe y es único*

***(b)*** *Para todo $\lambda \geq 0$ el estimador $\widehat{\boldsymbol{\beta}}^\lambda_{L,2}$ existe y es único*

***(c)*** *Para todo $\lambda \geq 0$ existe $t \geq 0$ tal que $\widehat{\boldsymbol{\beta}}^t_{L,1} = \widehat{\boldsymbol{\beta}}^\lambda_{L,2}$*

***(d)*** *Para todo $t \geq 0$ existe $\lambda \geq 0$ tal que $\widehat{\boldsymbol{\beta}}^\lambda_{L,2} = \widehat{\boldsymbol{\beta}}^t_{L,1}$.*

Derivando (10) respecto de cada $\beta_j$ obtenemos

$$\begin{aligned}
&-2\sum_{i=1}^{n}(y_i - \mathbf{x}_i\boldsymbol{\beta}) = 0 \\
&-2\sum_{i=1}^{n}(y_i - \mathbf{x}_i\boldsymbol{\beta})x_{i1} + \lambda \mathrm{sg}(\beta_1) = 0 \\
&\cdots\cdots\cdots\cdots \\
&-2\sum_{i=1}^{n}(y_i - \mathbf{x}_i\boldsymbol{\beta})x_{ip} + \lambda \mathrm{sg}(\beta_p) = 0,
\end{aligned}$$

donde sg indica la función signo. Por lo tanto resulta

$$\begin{aligned}
&-2\mathbf{j}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \\
&-2\mathbf{x}_r^{(1)\prime}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \mathrm{sg}(\beta_1) = 0 \\
&\cdots\cdots\cdots\cdots \\
&-2\mathbf{x}_r^{(p)\prime}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \mathrm{sg}(\beta_p) = 0
\end{aligned}$$

donde $\mathbf{x}_r^{(j)}$ es la columna $j$ de $\mathbf{X}_r$. Por lo tanto tenemos

$$-2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \begin{bmatrix} 0 \\ \mathrm{sg}(\beta_1) \\ \ldots \\ \mathrm{sg}(\beta_p) \end{bmatrix} = \mathbf{0}, \tag{11}$$

y cuando no hay intercept ($\mathbf{X} = \mathbf{X}_r$), la ecuación será

$$\boxed{-2\mathbf{X}_r'(\mathbf{y} - \mathbf{X}_r\boldsymbol{\beta}) + \lambda \begin{bmatrix} \mathrm{sg}(\beta_1) \\ \ldots \\ \mathrm{sg}(\beta_p) \end{bmatrix} = \mathbf{0}} \tag{12}$$

Desarrollando la ecuación (11) se obtiene

$$-2\mathbf{j}'(\mathbf{y} - \mathbf{j}\beta_0 - \mathbf{X}_r\boldsymbol{\beta}_r) = 0$$

y

$$-2\mathbf{X}_r'(\mathbf{y} - \mathbf{j}\beta_0 - \mathbf{X}_r\boldsymbol{\beta}_r) + \lambda \begin{bmatrix} \text{sg}(\beta_1) \\ \cdot \\ \text{sg}(\beta_p) \end{bmatrix} = \mathbf{0}. \tag{13}$$

Pero si se reemplaza $\beta_0$ en la ecuación de (13) queda

$$\boxed{-2\mathbf{X}_r^{c\prime}(\mathbf{y}^c - \mathbf{X}_r^c\boldsymbol{\beta}_r) + \lambda \begin{bmatrix} \text{sg}(\beta_1) \\ \cdot \\ \text{sg}(\beta_p) \end{bmatrix} = \mathbf{0}}$$

donde $\mathbf{X}_r^c$ es la matriz $\mathbf{X}_r$ centrada, de modo que las columnas tienen media 0. Resolviendo esta ecuación se obtiene $\boldsymbol{\beta}_r$, y luego utilizando la parte de arriba de (13) resulta

$$\beta_0 = \overline{\mathbf{y}} - \overline{\mathbf{x}}_r\boldsymbol{\beta}_r.$$

# 5 Estimador LASSO Robusto

En la definición del estimador Lassso de Tibshirani se impone la restricción $\sum_{j=1}^{p} |\beta_j| \leq t$ a los coeficientes, y se busca el $\boldsymbol{\beta}$ que minimiza la suma de cuadrados de los residuos $r_i = y_i - \mathbf{x}_i\boldsymbol{\beta}$. Ahora bién, si hay outliers en los datos (ya sea en $\mathbf{y}$ y/o en $\mathbf{X}$), algunos residuos pueden ser muy grandes, y al elevarlos al cuadrado puede tener una alta influencia en el estimador. Se obtendría así un estimador $\widehat{\boldsymbol{\beta}}_L$ muy distorsionado.

Con el fin de resolver este problema, en lugar de minimizar $\sum_{i=1}^{n} r_i^2$, se utilizará otra función de los residuos,minimizando en su lugar

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

donde $\widehat{s}$ es un estimador de escala de los residuos, y $\rho : \rho : R \to R_{\geq 0}$ ($R_{\geq 0}$ son los reales no negativos) que satisface las siguientes propiedades:

**P1(i)** $\rho(u)$ es par y continua, (ii) $\rho(u)$ función no decreciente de $|u|$, (iii) $\rho(0) = 0$, (iv) $\lim_{u\to\infty} \rho(u) = 1$ y (v) $\rho(u)$ es estrictamente creciente si $u > 0$ y $\rho(u) < 1$.

De esta manera se obtendrá un estimador robusto de $\widehat{\boldsymbol{\beta}}_{LR}$, que se denominará un estimador MLASSO. La función $\rho(r)$ en principio debería ser "parecida" a la cuadrática cuando el residuo es pequeño, ya que cuando no hay outliers, es deseable que este estimador robusto se comporte como el estimador LASSO

definido por Tibshirani. Pero para residuos grandes, esta función debería acotarlos, de manera de reducir su impacto sobre el estimador. En este trabajo se utilizará la función bicuadrada de Tukey definida así

$$\rho_T(r,c) = \begin{cases} 1-(1-(\frac{r}{c})^2)^3 & \text{si} \quad |r| \leq c \\ 1 & \text{si} \quad |r| > c \end{cases} . \tag{14}$$

En la Figura 9 se representa esta función para $c = 3.44$

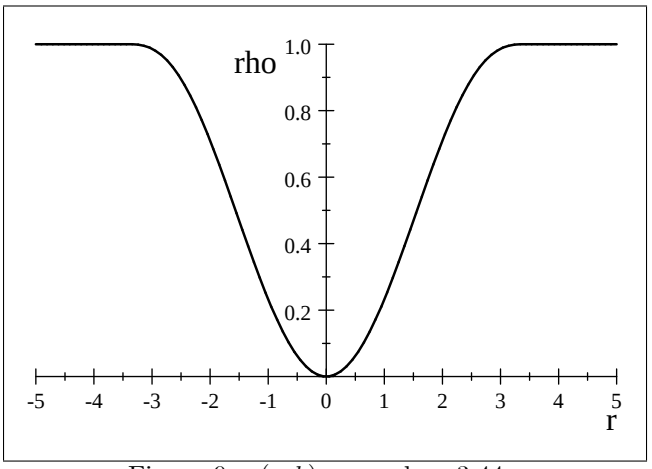

Figura 9: $\rho(r,k)_T$ para k = 3.44

## 5.1 Definición del estimador MLASSO

Definimos el estimador MLASSO de $\boldsymbol{\beta}_0$ mediante:

$$\widehat{\boldsymbol{\beta}}_{LR} = \underset{\sum_{j=1}^{p}|\beta_j|\leq t}{\arg\min} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{s}\right), \tag{15}$$

donde $s$ es un estimador de escala de los errores y la función $\rho$ satisface la propiedad **P1.**

Nuevamente esto es equivalente a que para algún $\lambda \geq 0$

$$\widehat{\boldsymbol{\beta}}_{LR} = \underset{\boldsymbol{\beta}}{\arg\min} \left( \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{s}\right) + \lambda \sum_{j=1}^{p} |\beta_j| \right). \tag{16}$$

Como estimador de escala de los residuos $\widehat{s}$ proponemos un S-estimador definido mediante

$$\widehat{s} = \min_{\boldsymbol{\beta}} s(\boldsymbol{\beta}),$$

donde $s(\boldsymbol{\beta})$ está definido por

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{y_i-\mathbf{x}_i\boldsymbol{\beta}}{c_0 s(\boldsymbol{\beta})}\right)=0.5$$

y $\rho_0$ satisface **P1** , con $c_0$ determinado de manera que cuando $y_i=\mathbf{x}_i\boldsymbol{\beta}+\varepsilon_i$ con los $\varepsilon_i$'s independientes con distribucion N(0,1) se tenga

$$E(\widehat{s})=1, \tag{17}$$

En este trabajo $\rho_0(u)$ se tomó igual a la función bicuadrada $\rho_{\mathrm{T}}(u,1)$ dada en (14). El valor $c_0$ depende de la matrix $\mathbf{X}_r$, y la forma de determinarlo se explica en la Sección 8.

## 5.2 El estimador MLASSO como mínimos cuadrados pesados penalizados

Derivando (16) respecto de $\boldsymbol{\beta}$, y llamando $\psi(u)=\rho'(u)$ resulta el siguiente sistema de ecuaciones

$$\begin{aligned}
&\sum_{i=1}^{n}\psi(\tfrac{y_i-\mathbf{x}_i\boldsymbol{\beta}}{s})\tfrac{1}{s}=0\\
&\sum_{i=1}^{n}\psi(\tfrac{y_i-\mathbf{x}_i\boldsymbol{\beta}}{s})\tfrac{x_{i1}}{s}+\lambda\mathrm{sg}(\beta_1)=0\\
&\cdots\cdots\cdots..\\
&\sum_{i=1}^{n}\psi(\tfrac{y_i-\mathbf{x}_i\boldsymbol{\beta}}{s})\tfrac{x_{ip}}{s}+\lambda\mathrm{sg}(\beta_p)=0.
\end{aligned} \tag{18}$$

Sean

$$r_i=y_i-\mathbf{x}\boldsymbol{\beta}, \tag{19}$$

$$w(u)=\psi(u)/u \tag{20}$$

y

$$\omega_i=w(r_i/s). \tag{21}$$

Luego el sistema (18) se puede escribir como

$$\begin{aligned}
&\sum_{i=1}^{n}\omega_i(y_i-\mathbf{x}_i\boldsymbol{\beta})\ =0\\
&\sum_{i=1}^{n}\omega_i(y_i-\mathbf{x}_i\boldsymbol{\beta})x_{i1}\ +\lambda s^2\mathrm{sg}(\beta_1)=0\\
&\cdots\cdots\cdots..\\
&\sum_{i=1}^{n}\omega_i(y_i-\mathbf{x}_i\boldsymbol{\beta})x_{ip}\ +\lambda s^2\mathrm{sg}(\beta_1)=0.
\end{aligned} \tag{22}$$

Observemos que los $\omega_i$ dependen de $\boldsymbol{\beta}$, y por lo tanto no son conocidos Primero resolveremos (22) cuando los $\omega_i$, $1 \leq i \leq n$ son conocidos y luego veremos como procedemos para resolver (18). El sistema (22) es equivalente a

$$\begin{aligned}
&\sum_{i=1}^{n}(\sqrt{\omega_i}y_i - \sqrt{\omega_i}\mathbf{x}_i\boldsymbol{\beta})\sqrt{\omega_i} = 0 \\
&\sum_{i=1}^{n}(\sqrt{\omega_i}y_i - \sqrt{\omega_i}\mathbf{x}_i\boldsymbol{\beta})\sqrt{\omega_i}x_{i1} + \lambda s^2 \mathrm{sg}(\beta_1) = 0 \\
&\cdots\cdots\cdots\cdots.. \\
&\sum_{i=1}^{n}(\sqrt{\omega_i}y_i - \sqrt{\omega_i}\mathbf{x}_i\boldsymbol{\beta})\sqrt{\omega_i}x_{ip} + \lambda s^2 \mathrm{sg}(\beta_p) = 0.
\end{aligned} \tag{23}$$

Denotemos ahora por

$$\mathbf{W} = \begin{bmatrix} \omega_1 & 0 & 0 & 0 \\ 0 & \omega_2 & 0 & 0 \\ 0 & 0 & \cdot & 0 \\ 0 & 0 & 0 & \omega_n \end{bmatrix}, \mathbf{y}^* = \mathbf{W}^{1/2}\mathbf{y}, \mathbf{X}^* = \mathbf{W}^{1/2}\mathbf{X} \tag{24}$$

y por $\mathbf{x}^{*(j)}$, $0 \leq j \leq p$ a la columna $j$ de $\mathbf{X}^*$. Luego (22) es equivalente a

$$\begin{aligned}
&\mathbf{x}^{*(0)\prime}(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}) = 0 \\
&\mathbf{x}^{*(1)\prime}(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}) + \lambda s^2 \mathrm{sg}(\beta_1) = 0 \\
&\cdots\cdots\cdots\cdots.. \\
&\mathbf{x}^{*(p)\prime}(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}) + \lambda s^2 \mathrm{sg}(\beta_p) = 0,
\end{aligned}$$

que también puede escribirse como

$$\mathbf{X}^{*\prime}(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}) + \lambda s^2 \begin{bmatrix} 0 \\ \mathrm{sg}(\beta_1) \\ \cdot \\ \mathrm{sg}(\beta_p) \end{bmatrix} = \mathbf{0}. \tag{25}$$

## 5.3 Algoritmo para computar el estimador MLASSO

El sistema de ecuaciones (25) es similar al sistema (11) cambiando $\mathbf{X}$ por $\mathbf{X}^*$, y $\mathbf{y}$ por $\mathbf{y}^*$, aunque difiere en que la primer columna de $\mathbf{X}^*$ es $\mathbf{j}^* = (\sqrt{\omega_1}, \sqrt{\omega_2}, \cdot\cdot, \sqrt{\omega_n})'$ en lugar de $(1, ....,1)'$. Por lo tanto habrá que transformar esta ecuación para poder resolverla utilizando una rutina de LASSO no robusta.

Observemos que $\mathbf{X}^* = [\mathbf{j}^*, \mathbf{X}_r^*]$, y que cada columna $\mathbf{x}_r^{*(j)}$ de $\mathbf{X}_r^*$ se puede escribir en 2 vectores: $\lambda_j\mathbf{j}^*$ en la dirección de $\mathbf{j}^*$, y $\mathbf{x}_r^{*\perp(j)} = \mathbf{x}_r^{*(j)} - \lambda_j\mathbf{j}^*$ en la dirección ortogonal a $\mathbf{j}^*$, donde

$$\lambda_j = \frac{\mathbf{j}^{*\prime}\mathbf{x}_r^{*(j)}}{\mathbf{j}^{*\prime}\mathbf{j}^*}.$$

Luego podemos escribir $\mathbf{x}_r^{*(j)} = \lambda_j \mathbf{j}^* + \mathbf{x}_r^{*\perp(j)}$, y entonces

$$\mathbf{X}^* = [\mathbf{j}^*, \mathbf{X}_r^*] = [\mathbf{j}^*, \lambda_1 \mathbf{j}^*, \cdots, \lambda_p \mathbf{j}^*] + [0, \mathbf{X}_r^{*\perp}]. \tag{26}$$

Por lo tanto resulta

$$\mathbf{X}^{*\prime}(\mathbf{y}^* - \mathbf{X}^* \boldsymbol{\beta}) = \mathbf{X}^{*\prime}(\mathbf{y}^* - (\beta_0 + \lambda_1 \beta_1 + \cdots + \lambda_p \beta_p)\mathbf{j}^* - \mathbf{X}_r^{*\perp} \boldsymbol{\beta}_r).$$

Notemos por $\varphi = \beta_0 + \lambda_1 \beta_1 + \cdots + \lambda_p \beta_p$, luego reemplazando en (26) se obtiene

$$\mathbf{X}^{*\prime}(\mathbf{y}^* - \mathbf{X}^* \boldsymbol{\beta}) = \left( \begin{bmatrix} \mathbf{j}^{*\prime} \\ \lambda_1 \mathbf{j}^{*\prime} \\ \cdot\cdot \\ \lambda_p \mathbf{j}^{*\prime} \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{X}_r^{*\perp\prime} \end{bmatrix} \right) (\mathbf{y}^* - \varphi \mathbf{j}^* - \mathbf{X}_r^{*\perp} \boldsymbol{\beta}_r). \tag{27}$$

Como

$$\mathbf{j}^{*\prime}(\mathbf{y}^* - \varphi \mathbf{j}^* - \mathbf{X}_r^{*\perp} \boldsymbol{\beta}_r) = \mathbf{j}^{*\prime} \mathbf{y}^* - \varphi \mathbf{j}^{*\prime} \mathbf{j}^*,$$

resulta

$$\mathbf{X}_r^{*\perp\prime}(\mathbf{y}^* - \varphi \mathbf{j}^* - \mathbf{X}_r^{*\perp} \boldsymbol{\beta}_r) = \mathbf{X}_r^{*\perp\prime}(\mathbf{y}^* - \mathbf{X}_r^{*\perp} \boldsymbol{\beta}_r),$$

y entonces queda

$$\mathbf{X}^{*\prime}(\mathbf{y}^* - \mathbf{X}^* \boldsymbol{\beta}) = \begin{bmatrix} \mathbf{j}^{*\prime} \mathbf{y}^* - \varphi \mathbf{j}^{*\prime} \mathbf{j}^* \\ \lambda_1(\mathbf{j}^{*\prime} \mathbf{y}^* - \varphi \mathbf{j}^{*\prime} \mathbf{j}^*) \\ \cdot\cdot \\ \lambda_p(\mathbf{j}^{*\prime} \mathbf{y}^* - \varphi \mathbf{j}^{*\prime} \mathbf{j}^*) \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{X}_r^{*\perp\prime}(\mathbf{y}^* - \mathbf{X}_r^{*\perp} \boldsymbol{\beta}_r) \end{bmatrix}.$$

Reemplazando en (25) se obtiene

$$\left( \begin{bmatrix} \mathbf{j}^{*\prime} \mathbf{y}^* - \varphi \mathbf{j}^{*\prime} \mathbf{j}^* \\ \lambda_1(\mathbf{j}^{*\prime} \mathbf{y}^* - \varphi \mathbf{j}^{*\prime} \mathbf{j}^*) \\ \cdot\cdot \\ \lambda_p(\mathbf{j}^{*\prime} \mathbf{y}^* - \varphi \mathbf{j}^{*\prime} \mathbf{j}^*) \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{X}_r^{*\perp\prime}(\mathbf{y}^* - \mathbf{X}_r^{*\perp} \boldsymbol{\beta}_r) \end{bmatrix} \right) + \lambda s^2 \begin{bmatrix} 0 \\ \operatorname{sg}(\beta_1) \\ \cdot \\ \operatorname{sg}(\beta_p) \end{bmatrix} = \mathbf{0},$$

pero como, según la primera fila $\mathbf{j}^{*\prime} \mathbf{y}^* - \varphi \mathbf{j}^{*\prime} \mathbf{j}^* = 0$, esta ecuación se simplifica a

$$\begin{bmatrix} \mathbf{j}^{*\prime} \mathbf{y}^* - \varphi \mathbf{j}^{*\prime} \mathbf{j}^* \\ \mathbf{X}_r^{*\perp\prime}(\mathbf{y}^* - \mathbf{X}_r^{*\perp} \boldsymbol{\beta}_r) \end{bmatrix} + \lambda s^2 \begin{bmatrix} 0 \\ \operatorname{sg}(\beta_1) \\ \cdot \\ \operatorname{sg}(\beta_p) \end{bmatrix} = \mathbf{0}.$$

O sea el estimador MLASSO satisface el siguiente sistema de ecuaciones

$$\boxed{\begin{array}{l} \mathbf{j}^{*\prime} \mathbf{y}^* - \varphi \mathbf{j}^{*\prime} \mathbf{j}^* = 0 \\ \mathbf{X}_r^{*\perp\prime}(\mathbf{y}^* - \mathbf{X}_r^{*\perp} \boldsymbol{\beta}_r) + \lambda s^2 \begin{bmatrix} \operatorname{sg}(\beta_1) \\ \cdot \\ \operatorname{sg}(\beta_p) \end{bmatrix} = \mathbf{0}. \end{array}} \tag{28}$$

Observar que si conocieramos $\mathbf{X}_r^{*\perp}$, $\mathbf{y}^*$ y $\mathbf{j}^{*\prime}$ el valor $\boldsymbol{\beta}_r$ lo obtendríamos resolviendo

$$\mathbf{X}_r^{*\perp\prime}(\mathbf{y}^* - \mathbf{X}_r^{*\perp}\boldsymbol{\beta}_r) + \lambda s^2 \begin{bmatrix} \text{sg}(\beta_1) \\ \cdot \\ \text{sg}(\beta_p) \end{bmatrix} = \mathbf{0}. \tag{29}$$

y esto se lograría usando un LASSO no robusto sin intercept. Finalmente $\beta_0$ se despejaría facilmente de

$$\mathbf{j}^{*\prime}\mathbf{y}^* - \varphi \mathbf{j}^{*\prime}\mathbf{j}^* = 0. \tag{30}$$

Sin embargo $y^*$, $\mathbf{X}_r^{*\perp}$ y $\mathbf{j}^*$ dependen de $\boldsymbol{\beta}_r$ y $\beta_0$. Entonces para resolver el sistema (28) podemos usar el siguiente algoritmo iterativo. Llamemos $(\beta_0^{(i)}, \boldsymbol{\beta}_r^{(i)})$ los valores calculados en el paso $i \geq 0$. Luego para obtener todos estos valores será suficiente definr quienes son $(\beta_0^{(0)}, \boldsymbol{\beta}_r^{(0)})$ y dar una regla para una vez conocido $(\beta_0^{(i)}, \boldsymbol{\beta}_r^{(i)})$ calcular $(\beta_0^{(i+1)}, \boldsymbol{\beta}_r^{(i+1)})$. Los valores iniciales $(\beta_0^{(0)}, \boldsymbol{\beta}_r^{(0)})$ pueden obtenerse usando un MM-estimador sin restriciones. El paso recursivo que se propone es el siguiente. Dados $(\beta_0^{(i)}, \boldsymbol{\beta}_r^{(i)})$ los valores $(\beta_0^{(i+1)}, \boldsymbol{\beta}_r^{(i+1)})$ se calculan haciendo los siguientes cinco pasos

1. Se calculan los pesos $w_i, 1 \leq i \leq n$ usando (19), (20) y (21) tomando como $\boldsymbol{\beta} = (\beta_0^{(i)}, \boldsymbol{\beta}_r^{(i)})$.

2. Usando estos valores de $w_i$ se obtienen $\mathbf{X}^* = (\mathbf{j}^*, \mathbf{X}_r^*)$ y $\mathbf{y}^*$ usando (24).

3. Se calculan $\lambda_j = \mathbf{j}^{*\prime}\mathbf{x}_r^{*(j)} \ / \ \mathbf{j}^{*\prime}\mathbf{j}^*$, $1 \leq j \leq p$ y se obtiene $\mathbf{X}_r^{*\perp}$ con columna $j$ igual a $\mathbf{x}_r^{*\perp(j)} = \mathbf{x}_r^{*(j)} - \lambda_j \mathbf{j}^*$.

4. Usando estos valores de $\mathbf{y}^*$ y $\mathbf{X}_r^{*\perp}$ y la (28), y usando una rutina LASSO no robusta sin intercept se resuelve **(29)** obteniendose $\boldsymbol{\beta}_r^{(i+1)}$.

5. De acuerdo a (30), el valor $\beta_0^{(i+1)}$ se obtiene resolviendo $\varphi = \beta_0 + \lambda_1\beta_1 + \cdots + \lambda_p\beta_p$ obteniendo

$$\beta_0^{(i+1)} = \frac{1}{\mathbf{j}^{*\prime}\mathbf{j}^*}\left[\mathbf{j}^{*\prime}\mathbf{y}^* - (\lambda_1\beta_{r1}^{(i+1)} + \ldots\lambda_p\beta_{rp}^{(i+1)})\mathbf{j}^{*\prime}\mathbf{j}^*\right].$$

Finalmente el algoritmo se para cuando

$$\frac{\left\|\boldsymbol{\beta}^{(i+1)} - \boldsymbol{\beta}^{(i)}\right\|}{\left\|\boldsymbol{\beta}^{(i)}\right\|} \leq \delta,$$

donde $\delta > 0$ determina la precision del valor del estimador.

# 6 Valor óptimo de $t$ determinante de la restricción

Ahora se analizará cuál es el valor más conveniente de $t$, de manera que el estimador obtenido permita predecir lo más correctamente posible. Si bién en este trabajo se utiliza tanto el estimador MLASSO como el LASSO propuesto por Tibshirani, el desarrollo y notación que siguen se refieren al primero, ya que el segundo es similar.

## 6.1 Minimización del error de predicción

Considerese el modelo lineal

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta_0} + \boldsymbol{\varepsilon},$$

y sea $\widehat{\boldsymbol{\beta}}_{LR}(\mathbf{X}, \mathbf{y}, t)$ el estimador MLASSO de $\boldsymbol{\beta}_0$, que obviamente depende del valor $t \geq 0$ elegido que determina la restricción. La intención ahora es elegir el valor de $t$ de manera de conseguir un buen comportamiento de las prediccióones realizadas con el estimador. Si se designa por $\widehat{\boldsymbol{\mu}}_{LR}(\mathbf{X}, \mathbf{y}, t)$ al estimador MLASSO de $\boldsymbol{\mu}_0$, un criterio podría ser el de minimizar el error cuadrático medio de este estimador, donde

$$ECM(t) = \frac{1}{n} E(\|\widehat{\boldsymbol{\mu}}_{LR}(\mathbf{X}, \mathbf{y}, t) - \boldsymbol{\mu}_0\|^2).$$

Llamando para abreviar $E(\widehat{\boldsymbol{\mu}}_{LR}(\mathbf{X}, \mathbf{y}, t)) = E(\widehat{\boldsymbol{\mu}}_{LR}(t)$ a la media del estimador de $\boldsymbol{\mu}_0$ se tendrá

$$\begin{array}{l} \text{ECM}(t) = \frac{1}{n} E(\|\widehat{\boldsymbol{\mu}}_{LR}(\mathbf{X}, \mathbf{y}, t) - E(\widehat{\boldsymbol{\mu}}_{LR}(t)) + E(\widehat{\boldsymbol{\mu}}_{LR}(t)) - \boldsymbol{\mu}_0\|^2) \\ = \frac{1}{n} E(\|E(\widehat{\boldsymbol{\mu}}_{LR}(t)) - \boldsymbol{\mu}_0\|^2) + E(\|\widehat{\boldsymbol{\mu}}_{LR}(\mathbf{X}, \mathbf{y}, t) - E(\widehat{\boldsymbol{\mu}}_{LR}(t))\|^2) \\ = \frac{1}{n} ||\text{sesgo}(\widehat{\boldsymbol{\mu}}_{LR}(\mathbf{X}, \mathbf{y}, t))||^2 + \frac{1}{n} \text{traza}(\text{COV}(\widehat{\boldsymbol{\mu}}_{LR}(\mathbf{X}, \mathbf{y}, t)). \end{array}$$

Cuando $t$ es grande, el estimador MLASSO se comporta como el MM-estimador, siendo el sesgo pequeño y las varianzas grandes. Por el otro lado, cuando $t$ es pequeño, el estimador está muy penalizado siendo grande el sesgo y pequeña la varianza. Entonces se podría elegir el valor de $t$ que minimice el ECM($t$). Como las expresiónes de $E(\widehat{\boldsymbol{\mu}}_{LR}(\mathbf{X}, \mathbf{y}, t))$ y traza(COV($\widehat{\boldsymbol{\mu}}_{LR}(\mathbf{X}, \mathbf{y}, t)$) son difíciles de obtener, se procederá de otra forma.

Supongamos que $\mathbf{y}^{\blacktriangle}$ cumpla el mismo modelo lineal $\mathbf{y}^{\blacktriangle} = \mathbf{X}\boldsymbol{\beta_0} + \boldsymbol{\varepsilon}^{\blacktriangle}$ con los $\boldsymbol{\varepsilon}^{\blacktriangle}$ independientes de los $\boldsymbol{\varepsilon}$. Luego el error cuadrático medio de predicción cuando se usa nuestro estimador para las nuevas observaciones será

$$\begin{array}{rcl} PSE(t) & = & \frac{1}{n} E(\left\|\mathbf{y}^{\blacktriangle} - \widehat{\boldsymbol{\mu}}_{LR}(\mathbf{X}, \mathbf{y}, t)\right\|^2) \\ & = & \frac{1}{n} E(\left\|\mathbf{y}^{\blacktriangle} - \boldsymbol{\mu}_0 + \boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_{LR}(\mathbf{X}, \mathbf{y}, t)\right\|^2) \\ & = & \frac{1}{n} E(\|\widehat{\boldsymbol{\mu}}_{LR}(\mathbf{X}, \mathbf{y}, t) - \boldsymbol{\mu}_0\|^2) + \frac{1}{n} E(\left\|\mathbf{y}^{\blacktriangle} - \boldsymbol{\mu}_0\right\|^2) \\ & = & ECM(t) + \sigma^2, \end{array}$$

donde $\sigma^2 = Var(\varepsilon_j) = Var(\varepsilon_j^{\blacktriangle})$. Como $\sigma^2$ es constante, será lo mismo minimizar ECM($t$) que $PSE(t)$, que es más simple.

## 6.2 Validación cruzada

De acuerdo a lo anterior, el valor óptimo de $t$, o sea $t_{opt}$, será el que minimice una medida del error de predicción del estimador. Para obtener un estimador insesgado de esta medida se recurrirá al método de validación cruzada.

Consideremos un estimador LASSO, $\widehat{\boldsymbol{\beta}}$, ya sea el clásico o el M-estimador. El método de validación cruzada consiste en lo siguiente: Para cada valor $t$ y cada observación $j$ llamemos $\widehat{\boldsymbol{\beta}}_{-j}^{(t)}$ el estimador usando la restricción correspondiente a $t$ y eliminando la observación $j$. Luego calculamos el error de predicción $\widehat{\epsilon}_{t,j}$ de la observación $j$ usando $\widehat{\boldsymbol{\beta}}_{-j}^{(t)}$ que viene dado por

$$\widehat{\epsilon}_{t,j} = y_j - \widehat{\boldsymbol{\beta}}_{-j}^{(t)\prime} \mathbf{x}_j.$$

Llamemos

$$\widehat{\boldsymbol{\epsilon}}_t = (\widehat{\epsilon}_{t,1}, \widehat{\epsilon}_{t,2}, \cdots, \widehat{\epsilon}_{t,n}),$$

El estimador del error de predicción para el el estimador LASSO cuando se usa la restricción correspondiente a $t$ , estará dado por un estimador de escala de $\widehat{\boldsymbol{\epsilon}}_t$ que llamaremos $\widehat{PSE}(t)$. Luego el valor de $t$ se elige como

$$t_{opt} = \arg\min_t \widehat{PSE}(t).$$

En el caso del LASSO clásico se usa como escala la raiz cuadrada del promedio cuadrático de los errores, es decir

$$PSE(t) = \left( \frac{\widehat{\boldsymbol{\epsilon}}_t' \widehat{\boldsymbol{\epsilon}}_t}{n} \right)^{1/2}.$$

Para el estimador MLASSO se utilizó un estimador de escala de tipo tau. Este estimador fue propuesto por Yohai y Zamar (1988) y están descripto en la Sección 7. Resultará entonces para

$$PSE^R(t) = \tau(\widehat{\epsilon}_{t1} ..., \widehat{\epsilon}_{tn}),$$

donde $\tau$ es una escala de tipo tau.

### 6.2.1 Tipo de estandarización

Siempre que se aplique la rutina LASSO, sea esta robusta o no, previamente debe estandarizarse la matriz $\mathbf{X}_r$ por columnas, utilizando apropiados estimadores de posición y escala. Posteriormente deberá des-estandarizarse el beta obtenido con con estos mismos estimadores.

En el caso del LASSO de Tibshirani se estandarizará con media y desvío estándar. En cambio en el caso del estimador MLASSO se usará un M-estimador de posición y un estimador de escala de tipo tau.

### 6.2.2 Determinación del intervalo de rastreo $[t_a, t_b]$

En principio este intervalo debería ser $[0, \infty)$. Sin embargo para la eficiencia del cálculo se tratará de acotarlo.

Para el estimador LASSO de Tibshirani si $t \geq t_\infty$ el estimador coincide con el de mínimos cuadrados, luego el intervalo debería ser $[0, t_\infty]$, donde

$$t_\infty = \sum_{j=1}^{p} \left| \widehat{\beta}^z_{jrls} \right|.$$

Para el M esimador LASSO si $t \geq t^R_\infty$ el estimador es un MM-estimador. Para hallar el $t^R_\infty$ en este caso, se aplica la rutina robusta con un valor de $t$ muy grande, por ejemplo 10000, y luego se suman los valores absolutos de los coeficientes o sea

$$\widehat{\boldsymbol{\beta}}^{z\infty}_{LR} \simeq \widehat{\boldsymbol{\beta}}^z_{LR}(\mathbf{X}^z, \mathbf{y}, 10000),$$

y luego se define

$$t^R_\infty = \sum_{i=1}^{p} \left| \widehat{\beta}^{z\infty}_{irLR} \right|.$$

El intervalo de rastreo debería ser $[0, t^R_\infty]$. Sin embargo en la validación cruzada se presenta otro problema, que se analizará aquí solo para el caso robusto ya que en el caso no robusto es similar.

En cada una de las $n$ etapas en que se particiona la matriz $\mathbf{X}$, se obtienen las matrices $\mathbf{X}^z_{(-i)}$. Pero los $t^R_{i\infty}$ que corresponden a estas matrices son diferentes, e incluso algunos pueden ser mayores que el de la matriz total $\mathbf{X}^z$. Entonces antes, deberán calcularse los $t^R_{i\infty}$ correspondientes a cada sub-matriz. Luego el extremo superior del intervalo será

$$T^R_\infty = \max_{1 \leq i \leq n} \{t^R_{i\infty}\},$$

y el intervalo de rastreo sería $[0, T^R_\infty]$.

**Remark 4** *La necesidad de utilizar un intervalo más amplio (con límite superior $T^R_\infty$ en lugar de $t^R_\infty$) se presenta en general cuando hay outliers con alto Leverage. En esos casos la curva $PSE(t)$ a veces presenta*

1. en el intervalo $[0, t^R_\infty]$, el mínimo en $t_a$ con $t_a < t^R_\infty$.
2. en el intervalo $[0, T^R_\infty]$, el mínimo en $t_b$ con $t_b > t^R_\infty$.

Entonces si se utiliza la opción 1, el óptimo estaría en $t^R_{opt} = t_a$, y actuaría la restricción del LASSO. Sin embargo con la opción 2, al estar el mínimo en $t_b > t^R_\infty$, el óptimo sería $t^R_{opt} = t^R_\infty$, ya que arriba de este valor el estimador es el mismo.

En definitiva cuando se obtenga el $t^R_{opt} \in [0, T^R_\infty]$, como podría suceder que $t^R_{opt} > t^R_\infty$ se tomará como $t^R_{opt}$ a

$$t^R_{opt} = \min\{t^R_{opt}, t^R_\infty\}.$$

Para ejemplificar en la Figura 10 siguiente se representa una curva $PSE(t)$ para un intervalo de rastreo $[0, T_\infty^R]$, (aunque en el gráfico, para mayor claridad se restringe la curva para valores de $t$ en $[18, T_\infty^R]$ ).

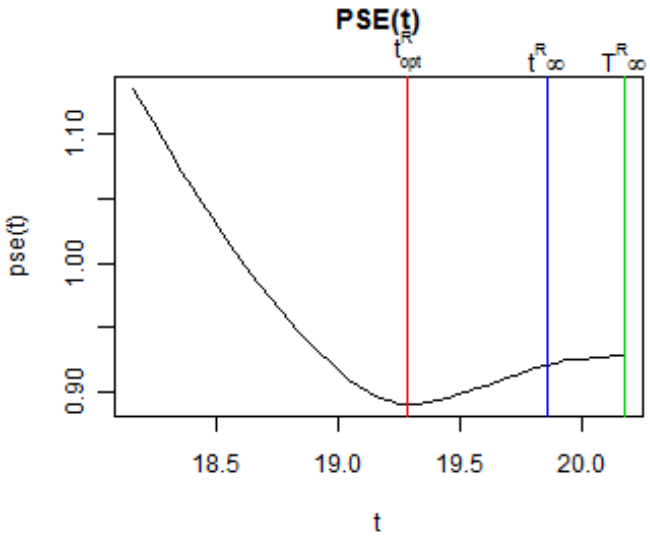


Figura 10: curva $PSE(t)$

Desde un punto de vista teórico utilizar el intervalo $[0, T_\infty^R]$ sería correcto ya que siempre el mínimo se encontrará ahí. A continuación este intervalo debe subdividirse en $m$ puntos, cuanto más cercanos mejor para lograr una buena estimación de $t_{opt}^R$. Pero como en cada uno de estos puntos la rutina de validación cruzada demora mucho, convendrá, si es posible, acotar mas este intervalo. Esto es posible en las simulaciones, ya que por ejemplo, si el beta con que se generan los datos tiene todos sus $\beta_i \neq 0$, el $t_{opt}^R$ estará casi siempre más cerca de $T_\infty^R$, lo que hace posible tomar un intervalo $[t_a, T_\infty^R]$ con $t_a$ alejado de cero, que es mucho menor. Por otro lado si los datos se generan con muchos $\beta_i = 0$, el $t_{opt}^R$ estará frecuentemente al principio del intervalo, y entonces un intervalo del tipo $[0, t_b]$ con $t_b$ menor que $T_\infty^R$ puede ser apropiado. Por eso en las simulaciones se incluyó una rutina que partiendo en las primeras 10 simulaciones con un intervalo $[0, T_\infty^R]$, lo va ajustando automaticamente según los $t_{opt}^R$ obtenidos.

Vamos a ejemplificar el comportamiento de la rutina de rastreo para el estimador MLASSO sin outliers, con $N = 200$ simulaciones, $p = 10$, el beta con que se generaron los datos tiene 8 ceros iniciales, y con correlación entre predictores 0.8. En las primeras 10 simulaciones se tomo como intervalo de rastreo a el máximo $[0, T_\infty^R]$, y en las siguientes este intervalo se ajusta a $[f_1 T_\infty^R, f_2 T_\infty^R]$, donde $f_1$ y $f_2$ son números entre 0 y 1. En la Figura 11 se representan en

azul estas fracciones para las simulaciones 11 a 200. En verde figuran los $t^R_{opt}$ obtenidos pero relativos a $T^R_\infty$, o sea $t^R_{opt}/T^R_\infty$. Nótese como se van ajustando las fracciones de manera que solo en unas pocas simulaciones (en este caso 7 sobre 200) el óptimo cae en los extremos del intervalo (puntos en rojo). Por último, aunque no está graficado aquí, cuando el beta con que se generan los datos tiene ninguno o pocos ceros, la fracción superior del intervalo es casi siempre igual a 1, y la inferior también más cercana a 1.

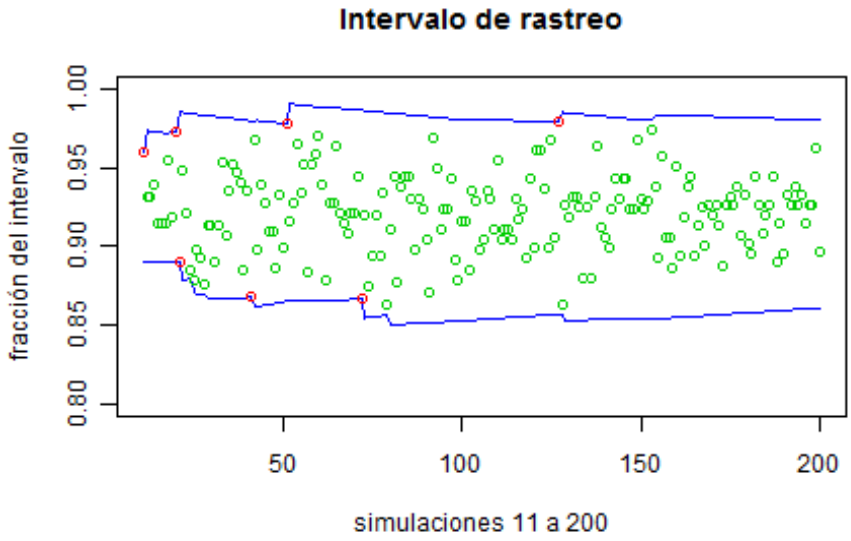


Figura11: Intervalo de rastreo

# 7 Estimador de escala tau

En esta sección describiremos los estimadores de escala tau que se usaron para la validación cruzada.Dada una muestra $\mathbf{x} = (x_1, x_2, \cdot\cdot, x_n)$, en general un M-estimador de escala $\widehat{\sigma}(\mathbf{x})$con punto de ruptura 0.5 está dado por el valor $\sigma$ que satisface la ecuación

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{x_i}{\sigma}\right) = 0.5, \tag{31}$$

donde $\rho_0$ es una función que verifica las propiedades **P1**.

Un inconveniente con los M-estimadores de escala es que no se puede lograr simultaneamente alta eficiencia y alto punto de ruptura. Para superar este problema Yohai y Zamar (1988), desarrollaron los estimadores de escala de

tipo tau, con el que sí se pueden lograr que se cumplan estas dos propiedades simultaneamente. Se define el estimador de escala de tipo tau $\tau(\mathbf{x})$ mediante

$$\tau^2(\mathbf{x}) = \widehat{\sigma}^2(\mathbf{x}) \frac{1}{n} \sum_{i=1}^{n} \rho_1 \left( \frac{x_i}{\widehat{\sigma}(\mathbf{x})} \right), \tag{32}$$

donde $\rho_1$ es una función que también satisface **P1**.

En este trabajo se utiliza para el M-estimador de escala de (31)

$$\rho_0(u) = I(|u| > 1) = \begin{cases} 0 & \text{si } |u| \leq 1 \\ 1 & \text{si } |u| > 1 \end{cases} \quad \text{y } \delta = 0.5.$$

Como se desea que el estimador obtenido $\widehat{\sigma}(\mathbf{x})$ tienda asintóticamente al $SD(\epsilon)$ en el caso que los $x_j$ tengan distribución normal, se deberá resolver en $c$, $E_\Phi \rho_0(u/c) = 0.5$ para $u \tilde{} N(0;1)$, lo que da $c = \Phi^{-1}(3/4) = 0.675$. Finalmente el estimador queda

$$\widehat{\sigma}(\mathbf{x}) = \frac{Med(|\epsilon_i|)}{c} = \frac{Med(|x_i|)}{0.675}.$$

Para el tau-estimador de (32) se utilizó una particular función $\rho$ definida mediante polinomios

$$\rho(u) = \begin{cases} \frac{u^2}{2} & \text{si } 0 \leq |u| < 2 \\ 1.792 - 0.972u^2 + 0.432u^4 - 0.052u^6 + 0.002u^8 & \text{si } 2 \leq |u| < 3 \\ 3.25 & \text{si } 3 \leq |u| \end{cases}. \tag{33}$$

En la Figura 12 se representa en negro sólido esta función (normalizada de manera que $\lim_{u \longrightarrow \pm\infty} \rho(u) = 1$), y para comparar: en verde $\rho(u) = u^2$, que da la desviación standard que es no robusto $\widehat{\sigma}(\epsilon) = SD(\epsilon)$ pero muy eficiente para distribuciones normales. En rojo punteado la función rho de una bicuadrada con c= 3.

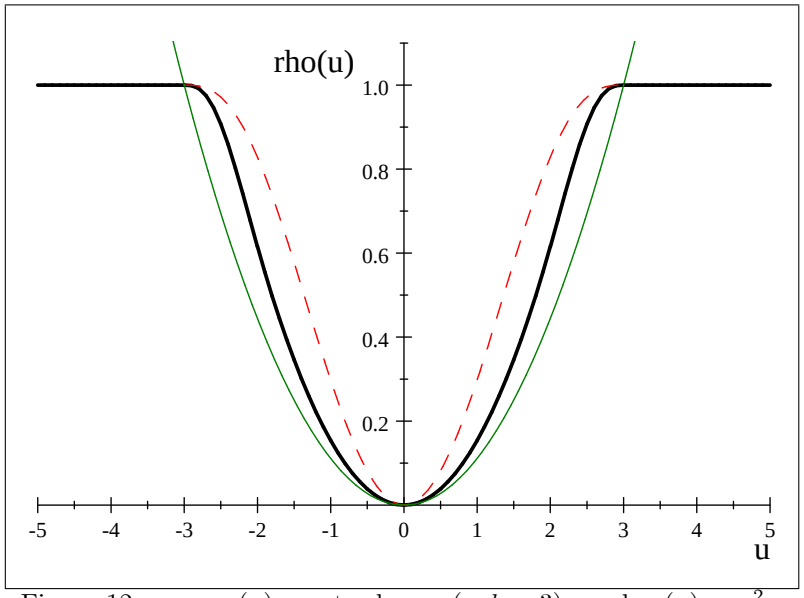

Figura 12: negro $\rho(u)$; punteado $\rho_{BS}(u, k = 3)$; verde $\rho(u) = u^2$

Como se aprecia en la figura, la función rho usada en (33), para $|u| \leq 3$ es intermedia entre las dos, y con ella el estimador tau mantiene una alta eficiencia, y tiene *punto de ruptura aproximadamente del 50%*.

# 8 Estudio de simulación por Montecarlo

## 8.1 Descripción de los modelos simulados

En todos los casos se generan inicialmente datos de un modelo lineal

$$\mathbf{y} = \mathbf{j}\beta_0 + \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon},$$

donde:

- $\beta_0 = 0$ y $\boldsymbol{\beta}_r = [\beta_1, \beta_2, \cdot\cdot, \beta_p]'$ con los primeros *cer* coeficientes nulos y el resto con valor 10
- $\mathbf{X}_r \in \mathbb{R}^{n\times p}$ se genera con $N(0; 1)$ en dos variantes: con correlación $r = 0$ entre columnas y con correlación $r = 0.8$ entre columnas
- Las componentes de $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ se generan como $n$ variables independientes con distribución $N(0; 1)$
- $\mathbf{y} \in \mathbb{R}^n$ se obtiene mediante $\mathbf{y} = \mathbf{j}\beta_0 + \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}$.

Luego los parámetros que se irán variando en las salidas serán: *cer*, la cantidad de ceros iniciales en $\boldsymbol{\beta}_r$; y $r$ que mide la presencia o no de correlación entre predictores.

## 8.2 Indicadores de eficiencia y cantidad de ceros encontrados

Se efectuan $N = 200$ simulaciones, y con los datos generados en cada una se obtienen:

- el estimador LASSO de Tibshirani (óptimo por validación cruzada) $\widehat{\boldsymbol{\beta}}_L$.
- el estimador de cuadrados mínimos $\widehat{\boldsymbol{\beta}}_{ls}$.
- el estimador MLASSO (óptimo por validación cruzada), $\widehat{\boldsymbol{\beta}}_{LR}$.
- el MM-estimador de regresión $\widehat{\boldsymbol{\beta}}_M$.

Para cada muestra se calculan los estimadores mencionados. Luego para cada estimador se estiman el error cuadrático y el número de ceros en en las posiciones en que $\boldsymbol{\beta}_r$ tiene ceros, o sea:

$$\begin{array}{ll} EC_L = \left\|\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}\right\|^2 & EC_{LR} = \left\|\widehat{\boldsymbol{\beta}}_{LR} - \boldsymbol{\beta}\right\|^2 \\ EC_{ls} = \left\|\widehat{\boldsymbol{\beta}}_{ls} - \boldsymbol{\beta}\right\|^2 \quad , & EC_M = \left\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}\right\|^2 \\ C_L = \#ceros_L & C_{LR} = \#ceros_{LR}. \end{array}$$

Luego de finalizadas $N$ simulaciones se calculan las medias de estos seis estimadores, obteniendo estimaciones de los errores cuadráticos medios, y el número medio de ceros de los estimadores LASSO, o sea:

$$\begin{array}{l} \mathbf{ecm}_L = \widehat{E}(\left\|\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}\right\|^2) \\ \mathbf{ecm}_{ls} = \widehat{E}(\left\|\widehat{\boldsymbol{\beta}}_{ls} - \boldsymbol{\beta}\right\|^2) \\ \mathbf{mc}_L = \widehat{E}(\#ceros_L) \\ \mathbf{ecm}_{LR} = \widehat{E}(\left\|\widehat{\boldsymbol{\beta}}_{LR} - \boldsymbol{\beta}\right\|^2) \\ \mathbf{ecm}_M = \widehat{E}(\left\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}\right\|^2) \\ \mathbf{mc}_{LR} = \widehat{E}(\#ceros_{LR}). \end{array}$$

Por último se calculan varios indicadores de eficiencia:

- eficiencia del LASSO de Tibshirani respecto de cuadrados mínimos

$$\mathbf{eff}_L = eff_{L/ls} = \frac{\mathbf{ecm}_{ls}}{\mathbf{ecm}_L},$$

- eficiencia del estimador MLASSO respecto del MM-estimador

$$\mathbf{eff}_{LR} = eff_{LR/M} = \frac{\mathbf{ecm}_M}{\mathbf{ecm}_{LR}},$$

- eficiencia del estimador MLASSO respecto del LASSO de Tibshirani

$$\mathbf{EFF} = eff_{LR/L} = \frac{\mathbf{ecm}_L}{\mathbf{ecm}_{LR}},$$

- eficiencia del MM-estimador respecto de cuadrados mínimos

$$\mathbf{eff} = eff_{M/ls} = \frac{\mathbf{ecm}_{ls}}{\mathbf{ecm}_M},$$

- eficiencia del estimador MLASSO respecto de cuadrados mínimos

$$\mathbf{eff}_{LR/ls} = \frac{\mathbf{ecm}_{ls}}{\mathbf{ecm}_{LR}}.$$

**Remark 5** *Notar que entre las eficiencias vale la relación*

$$\boxed{\mathbf{EFF} = \frac{\mathbf{eff}_{LR}}{\mathbf{eff}_L}\mathbf{eff}.}$$

*Luego, si puede cumplirse aproximadamente* $\mathbf{eff}_{LR} \approx \mathbf{eff}_L$*, resultará entonces* $\mathbf{EFF} \approx \mathbf{eff}$.

## 8.3 Caso de datos con outliers

Para analizar el comportamiento de los estimadores LASSO (robusto y no robusto) cuando hay outliers, generaremos muestras con 10% de outliers. Se consideraran dos casos: muestras con outliers de bajo leverage y muestras con outliers de alto leverage.

Debido a la forma de la función de peso bicuadrada que interviene en la rutina Robusta, cuando los outliers son leves, el error cuadrático del estimador es usualmente pequeño; y cuando los outliers son muy grandes, también el error cuadrático es pequeño, ya que el peso de los residuos de estas observaciones se hace cero, y no influyen en el estimador.

Sin embargo para niveles intermedios de outliers el error cuadrático del estimador puede ser significativo y nos interesará saber cual será su valor máximo.

Por eso los datos obtenidos por simulación deberán parametrizarse según el tamaño de los outliers que contienen.

### 8.3.1 Datos con outliers de bajo leverage

Si $\mathbf{X}, \mathbf{y}$ son los datos obtenidos por simulación, en lo que sigue se definirán 9 tamaños de outliers.

Si se designa $\mathbf{X}^{LL}, \mathbf{y}^{LL}$ a los datos con outliers de bajo leverage, $q = [0.1n]$ la cantidad de outliers, $\kappa \in \mathbb{R}$ un parámetro para cuantificar los outliers, y $1 \leq j \leq 9$ un indice para parametrizar su tamaño se define

$$\mathbf{X}^{LL} = \mathbf{X},$$

$$y_{ij}^{LL} = \begin{cases} y_i + (j-1)\kappa & \text{si} \quad i \leq q \\ y_i & \text{si} \quad i > q \end{cases}.$$

Se tendrán entonces 9 conjuntos de datos. Para $j = 1$ son datos sin outliers, y a medida que aumenta $j$ se incrementa el tamaño de los mismos. Respecto del parámetro $\kappa$ habrá que calibrarlo de manera que para $j = 2$ y $j = 9$ el error cuadrático sea bajo, y se incremente para valores intermedios. En este trabajo se utilizó $k = 1.25$.

### 8.3.2 Datos con outliers de alto leverage

Designemos con $\mathbf{X}^{HL}, \mathbf{y}^{HL}$ a los datos con outliers de alto leverage, con $q = [0.1n]$ la cantidad de ellos, con $\kappa \in \mathbb{R}$ un parámetro para cuantificar los outliers, $1 \leq j \leq 9$ un indice para parametrizar su tamaño, y $\boldsymbol{\beta}$ el valor verdadero de los coeficientes de regresión, se define

$$\mathbf{X}^{HL} = \mathbf{X} + V,$$

donde $V = (v_{ir})$ y sea

$$v_{ir} = \begin{cases} 0 & \text{si} \quad i \leq q, \ r = 1 \\ 10 & \text{si} \quad i \leq q, \ r > 1 \\ 0 & \text{si} \quad i > q \end{cases}.$$

En cuanto el valor de $y^{HL}$ nuevamente se considerarán 9 casos según el tamaño del outlier, que los indicaremos con $1 \leq j \leq 9$. Sea $\boldsymbol{\beta}$ el valor verdadero de los coeficientes de regresión y sea $y_i$ la $i$-esima coordenada de $\mathbf{y}$,luego para $1 \leq j \leq 9$ el vector contaminado $\mathbf{y}_j^{HL} = (y_{1j}^{HL}, ..., y_{nj}^{HL})'$ se define por

$$y_{ij}^{HL} = \begin{cases} y_i + 10 \sum_{k=1}^{p} \beta_j + (j-1)\kappa & \text{si} \quad i \leq q \\ y_i & \text{si} \quad i > q \end{cases}.$$

donde $\kappa$ se elije de manera que los casos considerados cubran el tamaño del outlier que provoca el mayor error cuadrático medio. Para $j = 1$ son datos con leverage pero sin outliers, y a medida que aumenta $j$ se incrementa el nivel de los mismos. En las simulacione se utilizó $k = 5$.

**Remark 6** *En lo que sigue, tanto para $\mathbf{X}^{LL}, \mathbf{y}^{LL}$ o $\mathbf{X}^{HL}, \mathbf{y}^{HL}$, se utilizará la notación unificada $\mathbf{X}^{O}, \mathbf{y}^{O}$, y se indizarán estos datos según el nivel de outliers $1 \leq j \leq 9$; por ejemplo $\mathbf{X}^{O}[j]$ y $\mathbf{y}^{O}[j]$.*

## 8.4 Indicadores de comportamiento frente a outliers

Aquí se efectuan $N = 120$ simulaciones, y con los datos generados en cada simulación $\mathbf{X}^O[j]$ y $\mathbf{y}^O[j]$ para cada nivel de outliers $1 \leq j \leq 9$ se obtienen:

- el estimador LASSO de Tibshirani (óptimo por validación cruzada), $\widehat{\boldsymbol{\beta}}_L^O[j]$
- el estimador MLASSO (óptimo por validación cruzada), $\widehat{\boldsymbol{\beta}}_{LR}^O[j]$
- el MM-estimador de regresión $\widehat{\boldsymbol{\beta}}_M^O[j]$

y con ellos se calculan los respectivos errores cuadráticos(respecto del $\boldsymbol{\beta}$ con que fueron generados los datos sin outliers)

$$\begin{aligned} EC_L^O[j] &= \left\| \widehat{\boldsymbol{\beta}}_L^O[j] - \boldsymbol{\beta} \right\|^2 \\ EC_{LR}^O[j] &= \left\| \widehat{\boldsymbol{\beta}}_{LR}^O[j] - \boldsymbol{\beta} \right\|^2 \\ EC_M^O[j] &= \left\| \widehat{\boldsymbol{\beta}}_M^O[j] - \boldsymbol{\beta} \right\|^2 . \end{aligned}$$

Además para $\widehat{\boldsymbol{\beta}}_L^O[j]$ y $\widehat{\boldsymbol{\beta}}_{LR}^O[j]$, la cantidad de ceros que tienen en las posiciones en que $\boldsymbol{\beta}_r$ tiene ceros, o sea:

$$\begin{aligned} C_L^O[j] &= \#ceros_L^O \\ C_{LR}^O[j] &= \#ceros_{LR}^O. \end{aligned}$$

Posteriormente, luego de finalizadas $N$ simulaciones se calculan el número medio de ceros de los estimadores LASSO

$$\begin{aligned} \mathbf{mc}_L^O[j] &= \widehat{E}(C_L^O[j]) \\ \mathbf{mc}_{LR}^O[j] &= \widehat{E}(C_{LR}^O[j]) \end{aligned}$$

y se grafican boxplots de:

**Errror cuadrático del Lasso no robusto** representando

$$EC_L^O[j],\ 1 \leq j \leq 9.$$

**Errror cuadrático del Lasso Robusto** representando

$$EC_{LR}^O[j],\ 1 \leq j \leq 9$$

**Errror cuadrático del MM-estimador** representando

$$EC_M^O[j],\ 1 \leq j \leq 9$$

**Indicador de respuesta a outliers** Por último se definirá un indicador que mide el desempeño frente a outliers del estimador MLASSO respecto del MM-estimador. Para ello, al finalizar las $N$ simulaciones, se calculan las medias truncadas al 10% superior de los errores cuadráticos, para cada nivel de outliers $2 \leq j \leq 9$

$$\begin{aligned} Tmean^{O}_{LR}[j] &= \underset{1\leq\ r\ \leq N}{Tmean}\{EC^{O}_{LR}[j]\} \\ Tmean^{O}_{M}[j] &= \underset{1\leq\ r\ \leq N}{Tmean}\{EC^{O}_{M}[j]\} \end{aligned}$$

que estiman el error cuadrático entre los 8 niveles de outliers, para el estimador MLASSO y para el MM-estimador. Luego para cada uno se calcula el máximo de estas 8 estimaciones

$$\begin{aligned} máxTmean^{O}_{LR} &= \underset{2\leq\ j\ \leq 9}{máx}\{Tmean^{O}_{LR}[j]\} \\ máxTmean^{O}_{M} &= \underset{2\leq\ j\ \leq 9}{máx}\{Tmean^{O}_{M}[j]\}, \end{aligned}$$

y se obtiene el indicador

$$I^{O} = \frac{máxTmean^{O}_{LR}}{máxTmean^{O}_{M}}. \tag{34}$$

## 8.5 Corrección de los estimadores por grados de libertad

**Obtención de las constantes $c_0$ y $c$ para el MM-estimador** Para obtener los valores $c_0$ y $c$ utilizador por el estimador MM descripto en Sección 5.1 se efectuaron $N = 10000$ simulaciones, generando datos con errores $N(0; 1)$, y con matrices $X_r$ con la misma distribución que las usadas en El Monte Carlo. El valor de $c_0$ se ajustó de manera que la escala de los residuos tuviera esperanza 1 y $c$ de manera que la eficiencia del MM estimador con respecto del estimador de cuadrados mínimos fuese del 85%.

Para $n = 50$ y $n = 100$ (usados en este trabajo), y diferentes $(p + 1)/n$ se obtuvieron las tablas:

| $(p+1)/n$ | 0.1 | 0.12 | 0.2 | 0.32 |
|---|---|---|---|---|
| $c_0$ | 1.71 | 1.75 | 1.92 | 2.25 |
| $c$ | 3.63 | 3.67 | 3.83 | 4.09 |

Tabla 1: Valores de $c_0$ y $c$ ajustados para $n = 50$

| $(p+1)/n$ | 0.1 | 0.11 | 0.2 | 0.32 |
|---|---|---|---|---|
| $c_0$ | 1.70 | 1.72 | 1.91 | 2.24 |
| $c$ | 3.56 | 3.58 | 3.80 | 4.20 |

Tabla 2: Valores de $c_0$ y $c$ ajustados para $n = 100$

Puede observarse que estos valores difieren de los asintóticos $c_0 = 1.55$ y $c = 3.44$.Notar que a medida que aumenta $p/n$ se incrementan tanto $c_0$ como $c$.

## 8.6 Resultados del estudio de simulación

En este trabajo se consideraron $n = 50$ y 100. El número de coeficientes igual cero en cada caso fue el siguiente

- con $n = 50$, $p = 5$ y cantidad de ceros en $\boldsymbol{\beta}_r$ : $cer = 0, 2, 4$
- con $n = 100$, $p = 10$ y cantidad de ceros en $\boldsymbol{\beta}_r$ : $cer = 0, 5, 8$

Además, en cada caso se varió la correlación entre columnas de la matriz $\mathbf{X}_r$ : $rho = 0$ y $rho = 0.8$.

### 8.6.1 Caso n=50 p=5

### 8.6.2 Comportamiento sin Outliers

Se realizaron N=200 simulaciones. La explicación de las cantidades que figuran en las siguientes tablas se detallan en (8). Pero a continuación se hace una breve descripción de algunas:

- **effL:** eficiencia del LASSO de Tibshirani respecto de cuadrados mínimos
- $\text{ECM}_{\mathbf{L}}$ : error cuadrático medio del LASSO de Tibshirani
- $\mathbf{mc_L}$**:** número medio de ceros detectados con el LASSO de Tibshirani
- **effLR**: eficiencia del estimador MLASSO respecto del MM-estimador
- $\text{ECM}_{\mathbf{LR}}$ : error cuadrático medio del estimador MLASSO
- $\mathbf{mc_{LR}}$**:** número medio de ceros detectados con el estimador MLASSO
- **EFF**: eficiencia del estimador MLASSO respecto del LASSO de Tibshirani=$\text{ECM}_{\mathbf{L}}/\text{ECM}_{\mathbf{LR}}$
- **eff**: eficiencia del MM-estimador respecto de cuadrados mínimos

| No. de ceros | **effL** | $\mathbf{mc_L}$ | **EFF** | **effLR** | **eff** | $\mathbf{mc_{LR}}$ |
|---|---|---|---|---|---|---|
| **0** | 0.94 | 0 | 0.77 | 0.87 | 0.83 | 0 |
| **2** | 1.02 | 0.76 | 0.77 | 0.94 | 0.83 | 0.72 |
| **4** | 1.49 | 2.58 | 0.78 | 1.40 | 0.83 | 2.54 |

Tabla 3: Resultados sin outliers cuando $\rho = 0$

| No. de ceros | **effL** | $\mathbf{mc_L}$ | **EFF** | **effLR** | **eff** | $\mathbf{mc_{LR}}$ |
|---|---|---|---|---|---|---|
| **0** | 0.99 | 0 | 0.78 | 0.94 | 0.83 | 0 |
| **2** | 1.15 | 0.58 | 0.79 | 1.10 | 0.83 | 0.53 |
| **4** | 1.88 | 2.38 | 0.80 | 1.82 | 0.83 | 2.22 |

Tabla 4: Resultados sin outliers cuando $\rho = 0.8$

En los gráficos de las Figuras 13 y 14 que siguen se representa simultaneamente la información de estas dos tablas. Por eso en los ejes de abcisas se encuentran repetidos la cantidad de ceros en $\boldsymbol{\beta}_r$: la primera vez, para la tabla con $\rho = 0$, y la segunda para cuando $\rho = 0.8$. Se explicarán estas figuras.

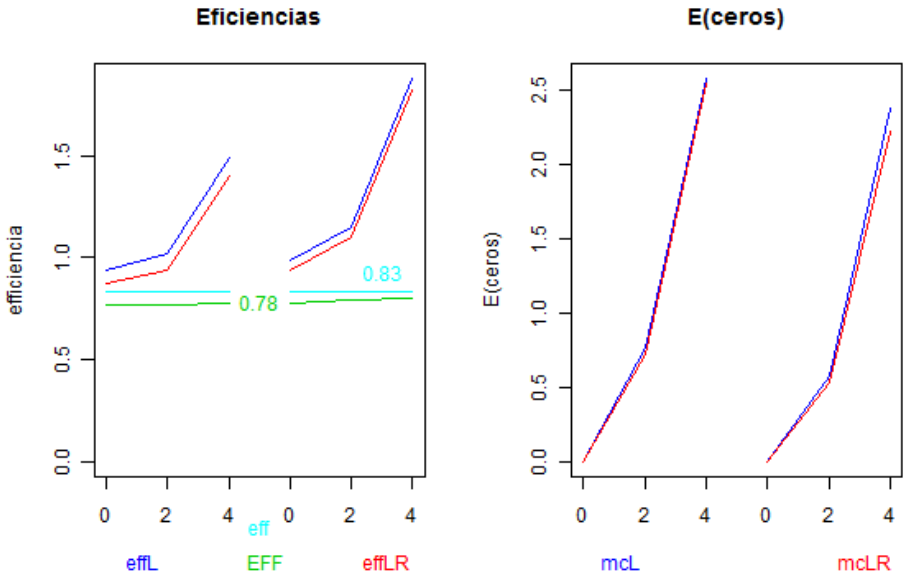


Figura13: Eficiencias y E(ceros) para n=50 y p=5

En la Figura 13:

- gráfico de Eficiencias: en azul la **effL**, en rojo la **effLR,** en verde la **EFF**, y en turquesa la **eff,** todos en función de la cantidad de ceros.
- gráfico de E(ceros): en azul $\mathbf{mc_L}$, en rojo $\mathbf{mc_{LR}}$, en función de la cantidad de ceros.

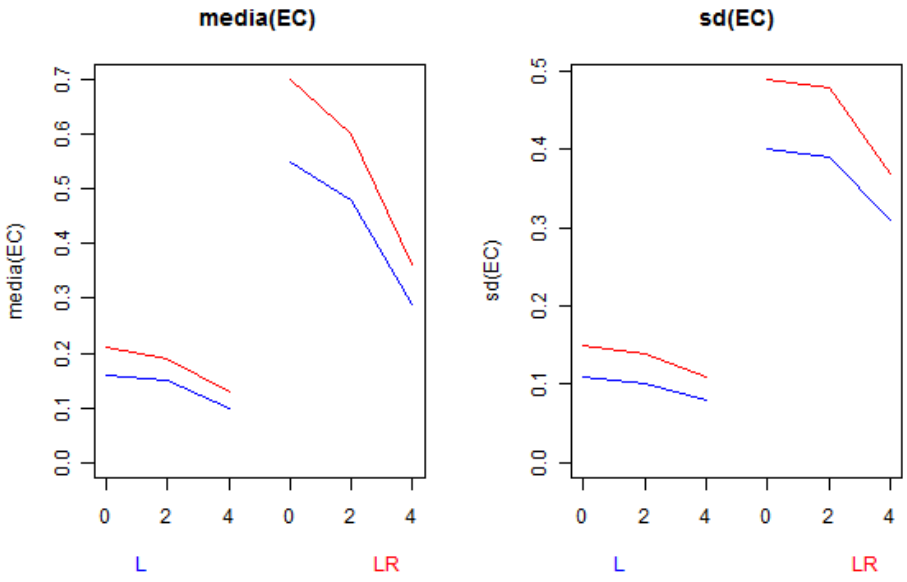


Figura14: media y desvio del EC para n=50 y p=5

En la Figura 14:

- gráfico de media(EC): en azul la $\mu(\mathrm{EC}_{\mathbf{L}})$, en rojo la $\mu(\mathrm{EC}_{\mathbf{LR}})$ en función de la cantidad de ceros.
- gráfico de sd(EC): en azul la $\sigma(\mathrm{EC}_{\mathbf{L}})$, en rojo la $\sigma(\mathrm{EC}_{\mathbf{LR}})$ en función de la cantidad de ceros.

### 8.6.3 Comportamiento con Outliers

A continuación, y para cada combinación de los parámetros (cer=0,2,4 y $\rho = 0$, 0.8) se representan dos figuras en función del nivel de outliers. En la primera estan los boxplots del error cuadratico para el Lasso(de Tibshirani), el MLasso(robusto) y el MM estimador de regresión; en la parte superior para el caso de outliers de bajo leverage, y en la inferior cuando son de alto leverage. En la segunda se comparan las curvas de error cuadratico medio del MLaso vs MMestimador, también para los casos de bajo leverage(BL) y alto leverage(AL). Se presenta aquí el indicador de respuesta a outliers $I^O$ definido en (34).

**Boxplots y curvas de error cuadrático medio cuando cer=0 y $\rho = 0$**

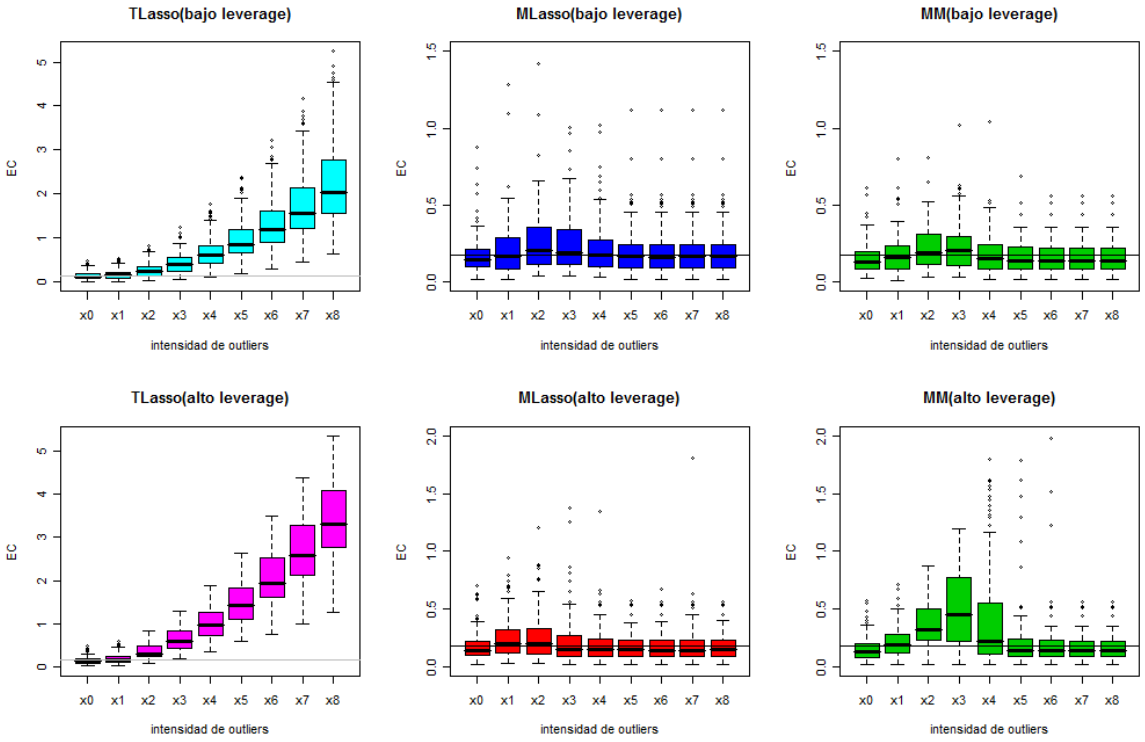


Figura 15: Boxplots del error cuadratico cuando cer=0 $\rho$=0

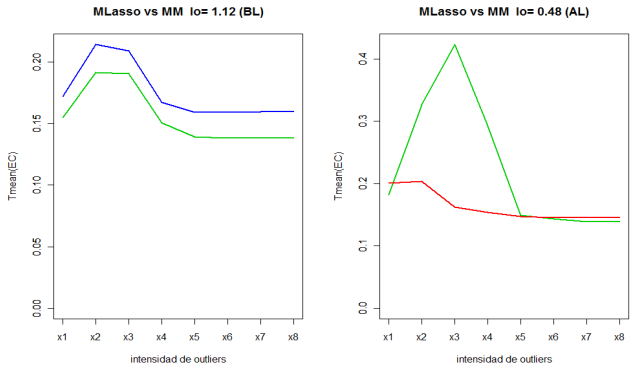


Figura 16: Curvas de error cuadratico medio cer=0 $\rho$=0

**Boxplots y curvas de error cuadrático medio cuando cer=0 y** $\rho = 0.8$

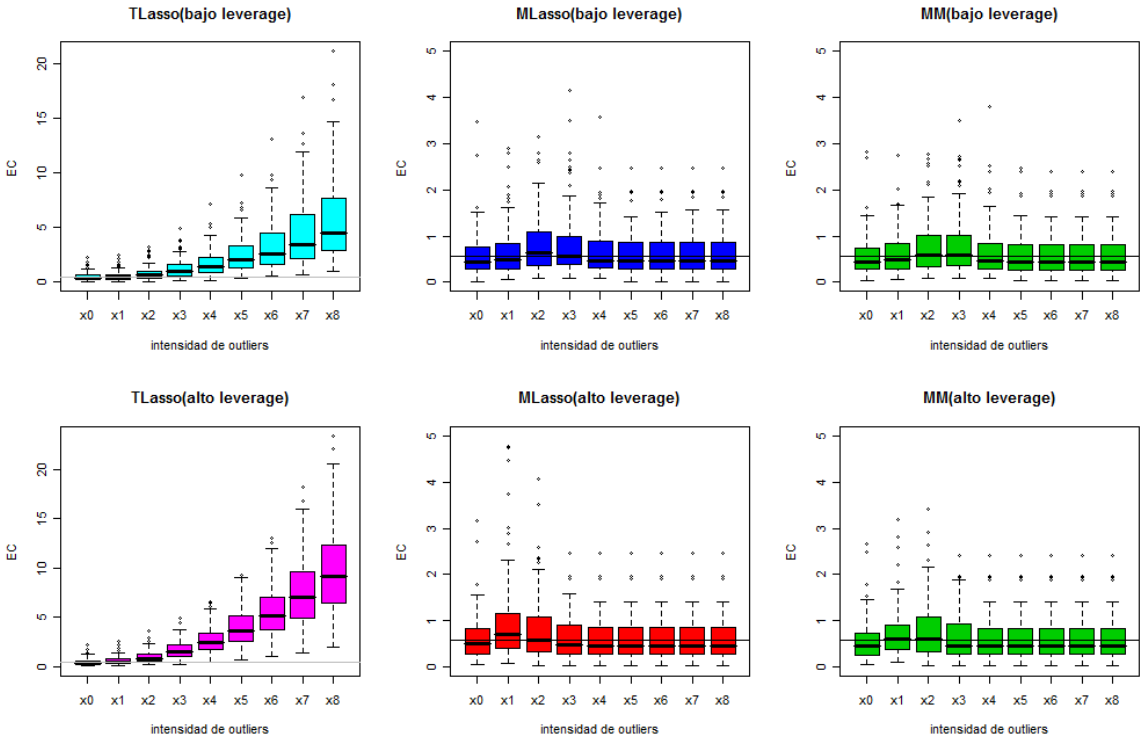


Figura 17: Boxplots del error cuadratico cuando cer=0 $\rho$=0.8

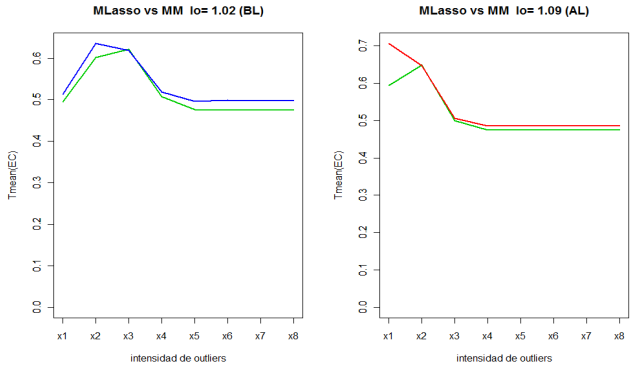


Figura 18: Curvas de error cuadratico medio cer=0 $\rho$=0.8

**Boxplots y curvas de error cuadrático medio cuando cer=2 y $\rho = 0$**

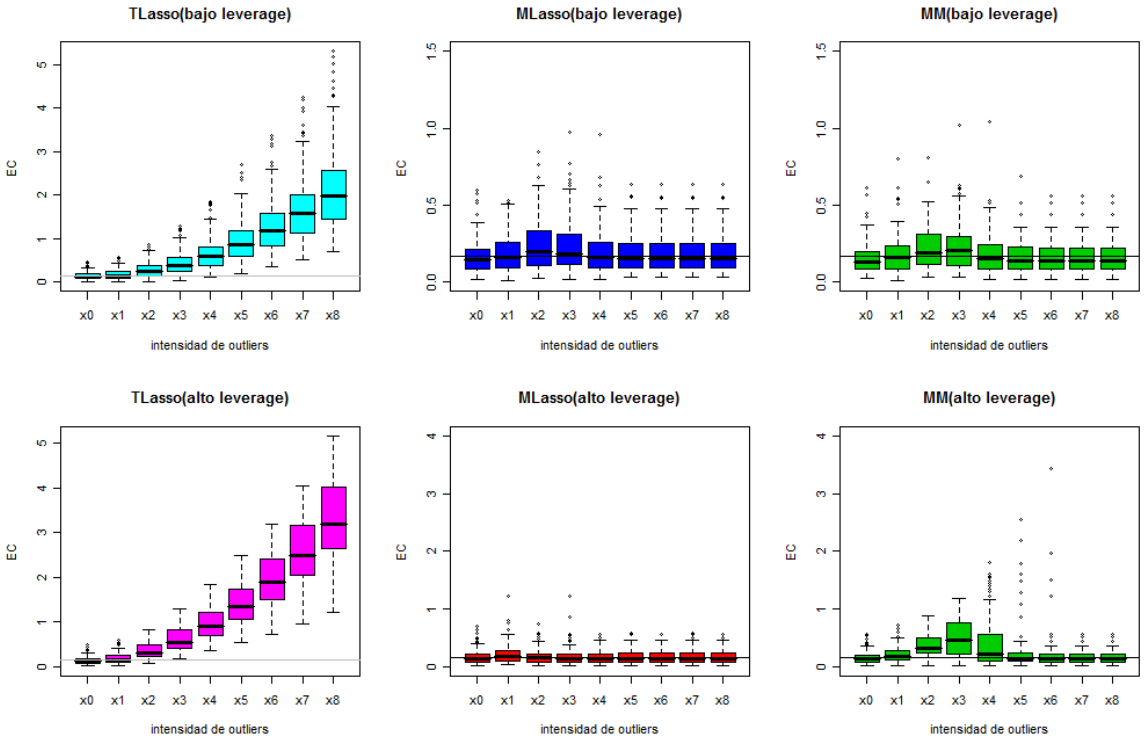


Figura 19: Boxplots del error cuadratico cuando cer=2 $\rho$=0

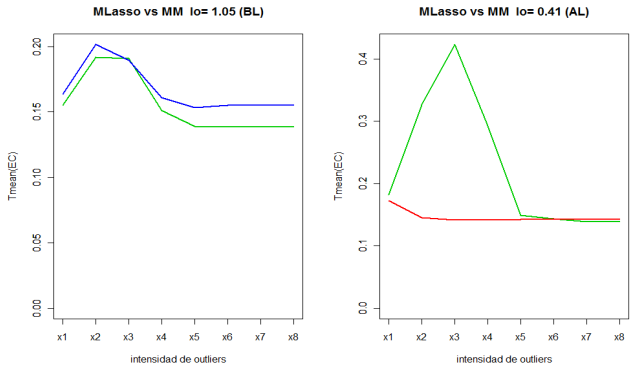


Figura 20: Curvas de error cuadratico medio cer=2 $\rho$=0

**Boxplots y curvas de error cuadrático medio cuando cer=2 y** $\rho = 0.8$

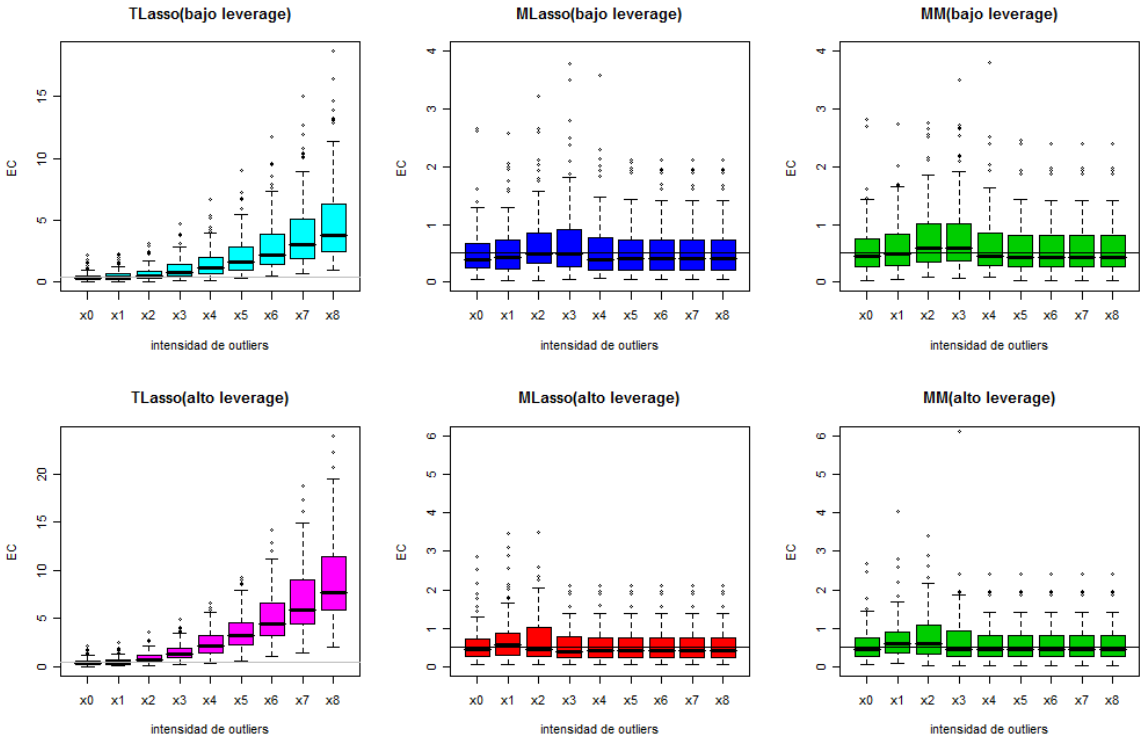


Figura 21: Boxplots del error cuadratico cuando cer=2 $\rho$=0.8

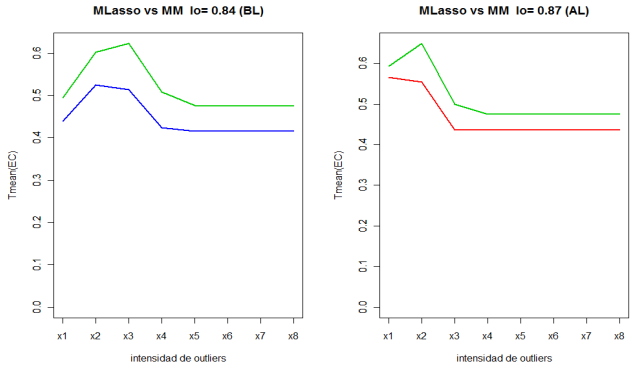


Figura 22: Curvas de error cuadratico medio cer=2 $\rho$=0.8

**Boxplots y curvas de error cuadrático medio cuando cer=4 y $\rho = 0$**

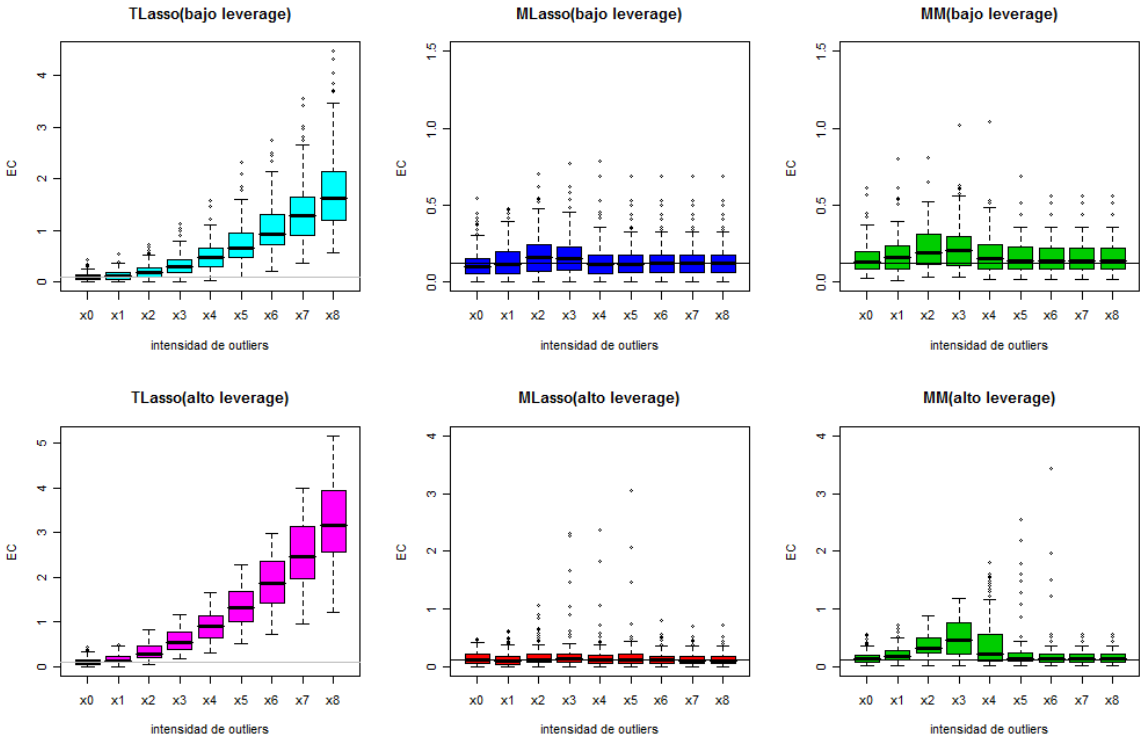


Figura 23: Boxplots del error cuadratico cuando cer=4 $\rho$=0

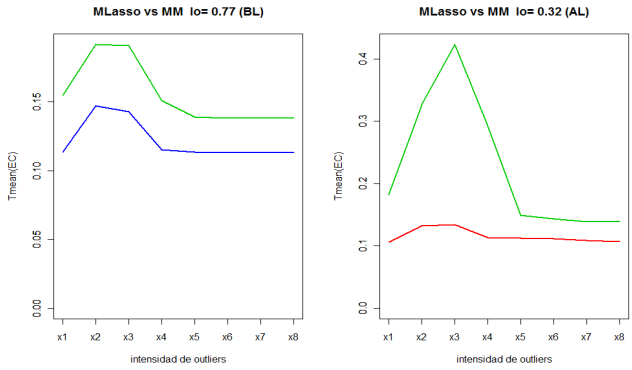


Figura 24: Curvas de error cuadratico medio cer=4 $\rho$=0

**Boxplots y curvas de error cuadrático medio cuando cer=4 y $\rho = 0.8$**

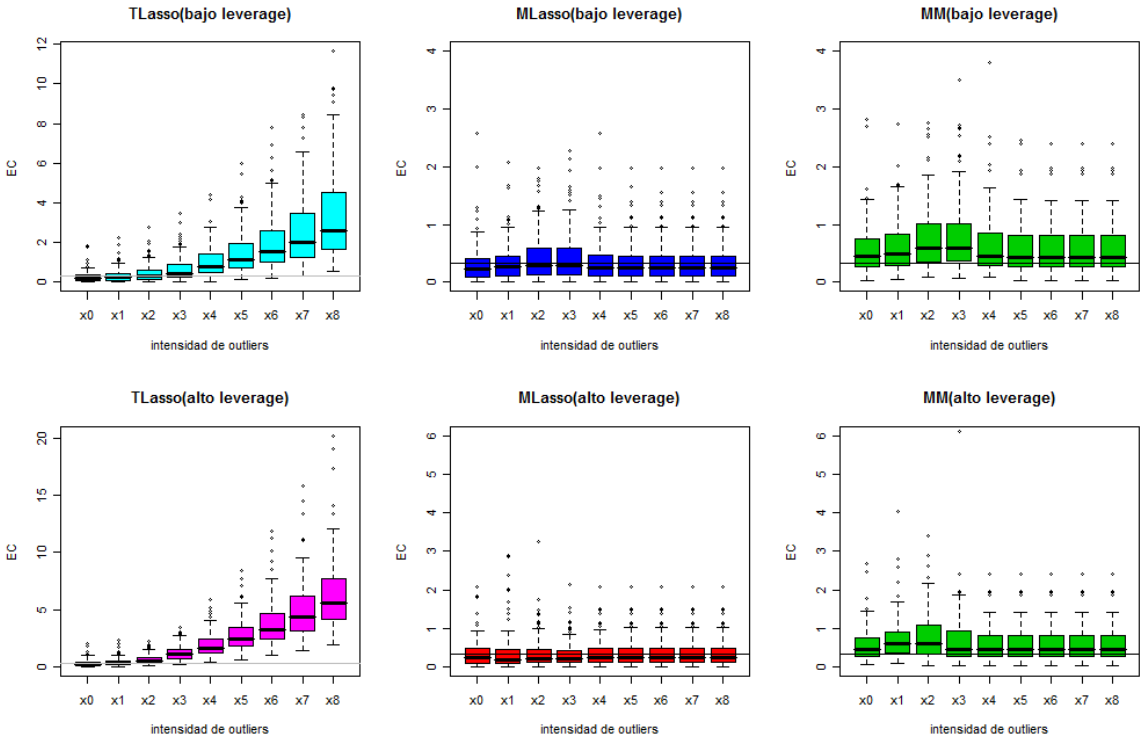


Figura 25: Boxplots del error cuadratico cuando cer=4 $\rho$=0.8

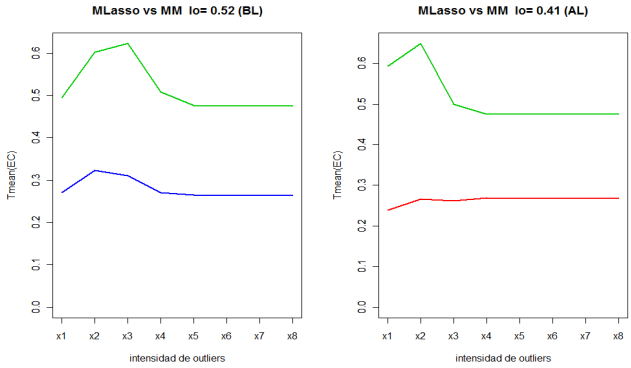


Figura 26: Curvas de error cuadratico medio cer=4 $\rho$=0.8

### 8.6.4 Caso n=100 y p=10

### 8.6.5 Comportamiento sin Outliers

Se realizaron N=200 simulaciones. La explicación de las cantidades que figuran en las siguientes tablas se detallan en (8).

| $\boldsymbol{\rho = 0}$ | **effL** | $\mathbf{mc_L}$ | **EFF** | **effLR** | **eff** | $\mathbf{mc_{LR}}$ |
|---|---|---|---|---|---|---|
| **cer=0** | 0.96 | 0 | 0.79 | 0.90 | 0.84 | 0 |
| **cer=5** | 1.13 | 2.19 | 0.82 | 1.09 | 0.85 | 2.19 |
| **cer=8** | 1.60 | 5.34 | 0.82 | 1.61 | 0.82 | 5.23 |

Tabla 5: Resultados sin outliers cuando $\rho = 0$

| $\boldsymbol{\rho = 0.8}$ | **effL** | $\mathbf{mc_L}$ | **EFF** | **effLR** | **eff** | $\mathbf{mc_{LR}}$ |
|---|---|---|---|---|---|---|
| **cer=0** | 1 | 0 | 0.84 | 0.96 | 0.87 | 0 |
| **cer=5** | 1.30 | 1.97 | 0.85 | 1.30 | 0.84 | 1.9 |
| **cer=8** | 2.04 | 4.62 | 0.85 | 2.03 | 0.85 | 4.76 |

Tabla 6: Resultados sin outliers cuando $\rho = 0.8$

En los gráficos de las Figuras 27 y 28 que siguen se representa simultaneamente la información de estas dos tablas. Por eso en los ejes de abcisas se encuentran repetidos la cantidad de ceros en $\boldsymbol{\beta}_r$: la primera vez, para la tabla con $\rho = 0$, y la segunda para cuando $\rho = 0.8$. Se explicarán estas Figuras.

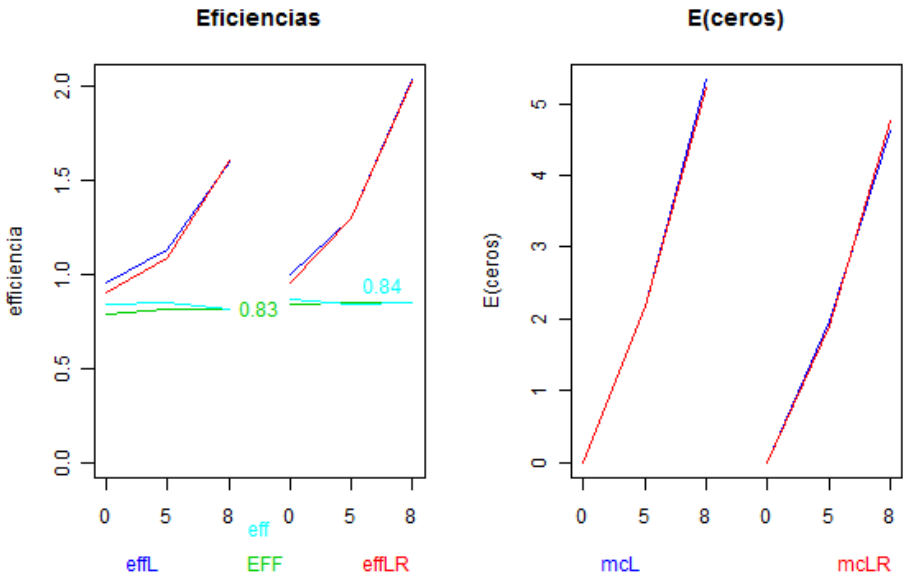


Figura 27: Eficiencias y E(ceros) para n=100 y p=10

En la Figura 27:

- gráfico de Eficiencias: en azul la **effL**, en rojo la **effLR,** en verde la **EFF**, y en turquesa la **eff,** todos en función de la cantidad de ceros.
- gráfico de E(ceros): en azul $\mathbf{mc_L}$, en rojo $\mathbf{mc_{LR}}$, en función de la cantidad de ceros.

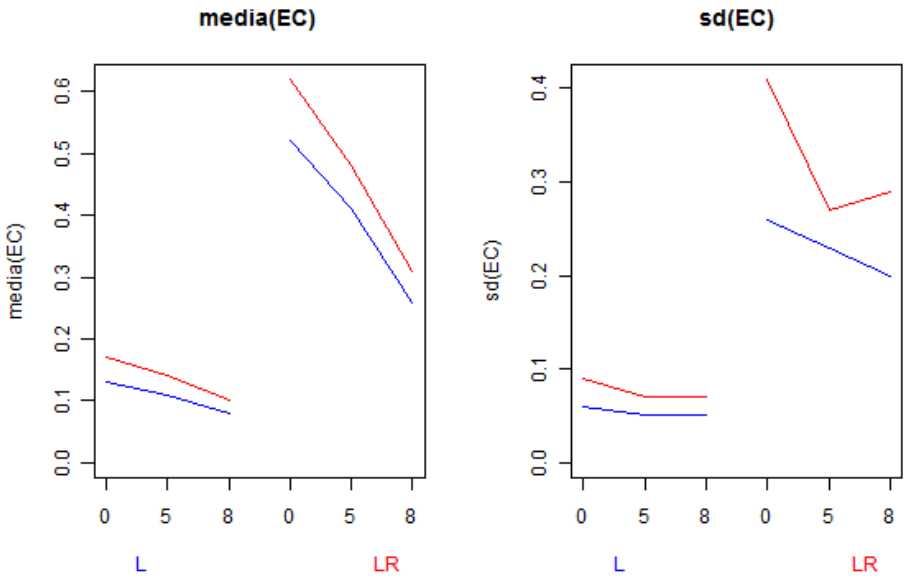


Figura 28: media y desvio del EC para n=100 y p=10

En la Figura 28:

- gráfico de media(EC): en azul la $\mu(\mathrm{EC}_{\mathbf{L}})$, en rojo la $\mu(\mathrm{EC}_{\mathbf{LR}})$ en función de la cantidad de ceros.
- gráfico de sd(EC): en azul la $\sigma(\mathrm{EC}_{\mathbf{L}})$, en rojo la $\sigma(\mathrm{EC}_{\mathbf{LR}})$ en función de la cantidad de ceros.

### 8.6.6 Comportamiento con Outliers

A continuación, y para cada combinación de los parámetros (cer=0,5,8 y $\rho = 0$, 0.8) se representan dos figuras en función del nivel de outliers. En la primera estan los boxplots del error cuadratico para el Lasso(de Tibshirani), el MLasso(robusto) y el MM estimador de regresión; en la parte superior para el caso de outliers de bajo leverage, y en la inferior cuando son de alto leverage. En la segunda se comparan las curvas de error cuadratico medio del MLaso vs MMestimador, también para los casos de bajo leverage(BL) y alto leverage(AL). Se presenta aquí el indicador de respuesta a outliers $I^O$ definido en (34).

**Boxplots y curvas de error cuadrático medio cuando cer=0** $\rho = 0$

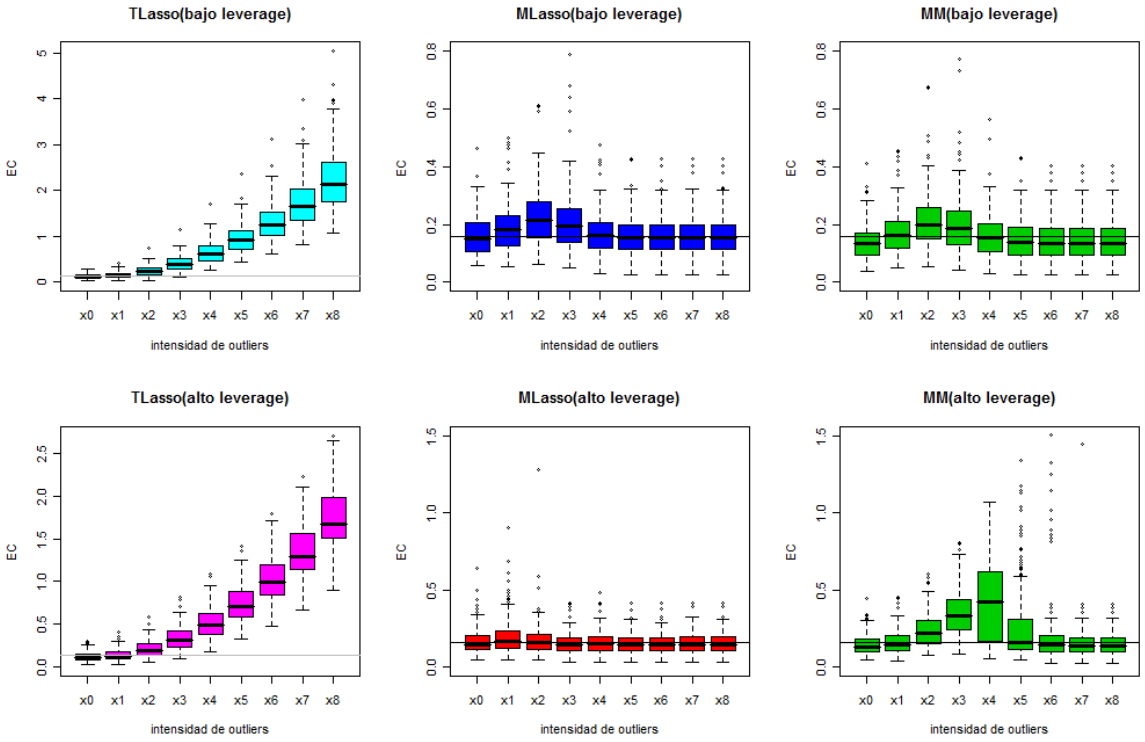


Figura 29: Boxplots del error cuadratico cuando cer=0 $\rho$=0

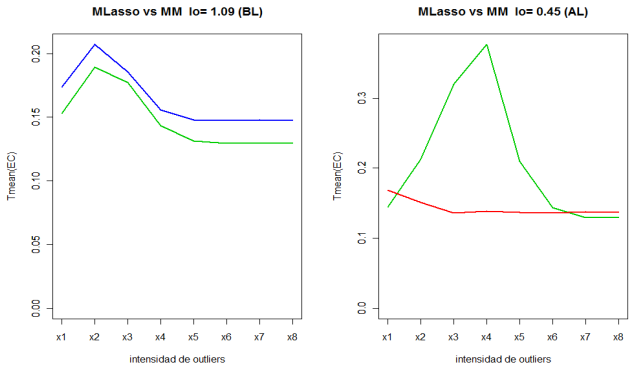


Figura 30: Curvas de error cuadratico medio cer=0 $\rho$=0

**Boxplots y curvas de error cuadrático medio cuando cer=0 y $\rho = 0.8$**

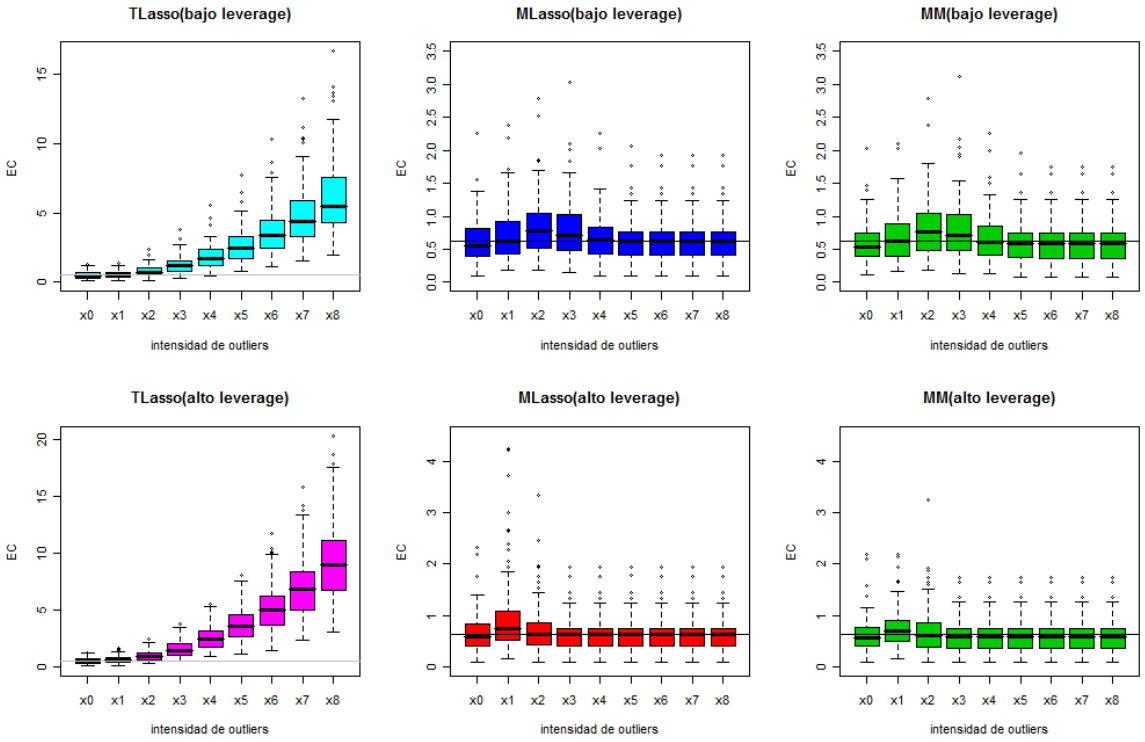


Figura 31: Boxplots del error cuadratico cuando cer=0 $\rho$=0.8

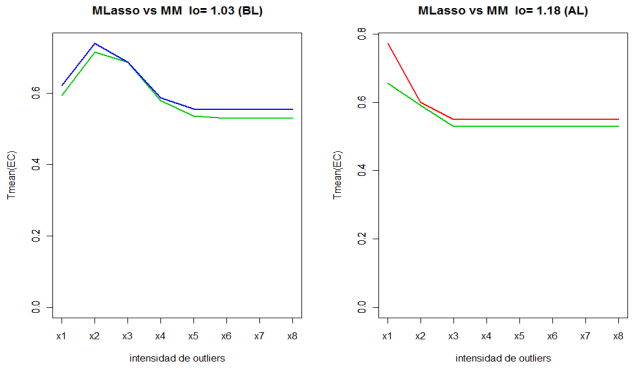


Figura 32: Curvas de error cuadratico medio cer=0 $\rho$=0.8

**Boxplots y curvas de error cuadrático medio cuando cer=5 y $\rho = 0$**

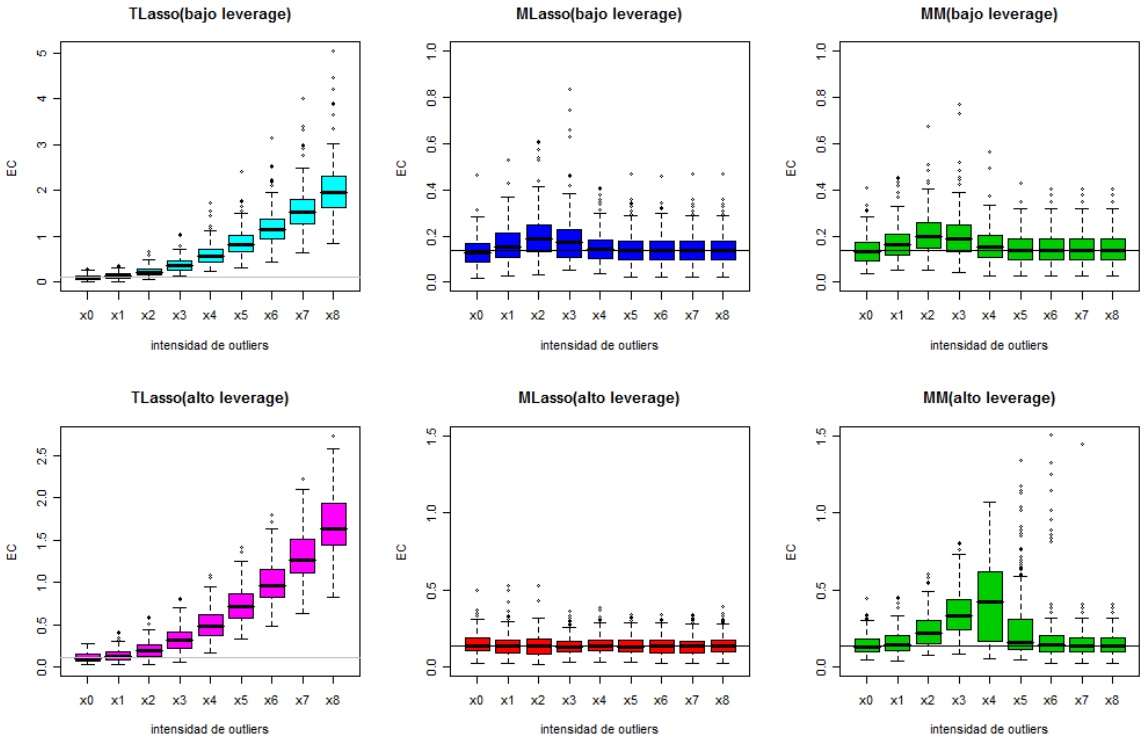


Figura 33: Boxplots del error cuadratico cuando cer=5 $\rho$=0

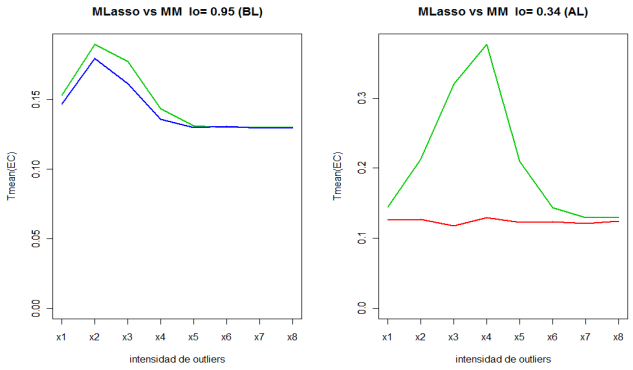


Figura 34: Curvas de error cuadratico medio cer=5 $\rho$=0

**Boxplots y curvas de error cuadrático medio cuando cer=5 y $\rho = 0.8$**

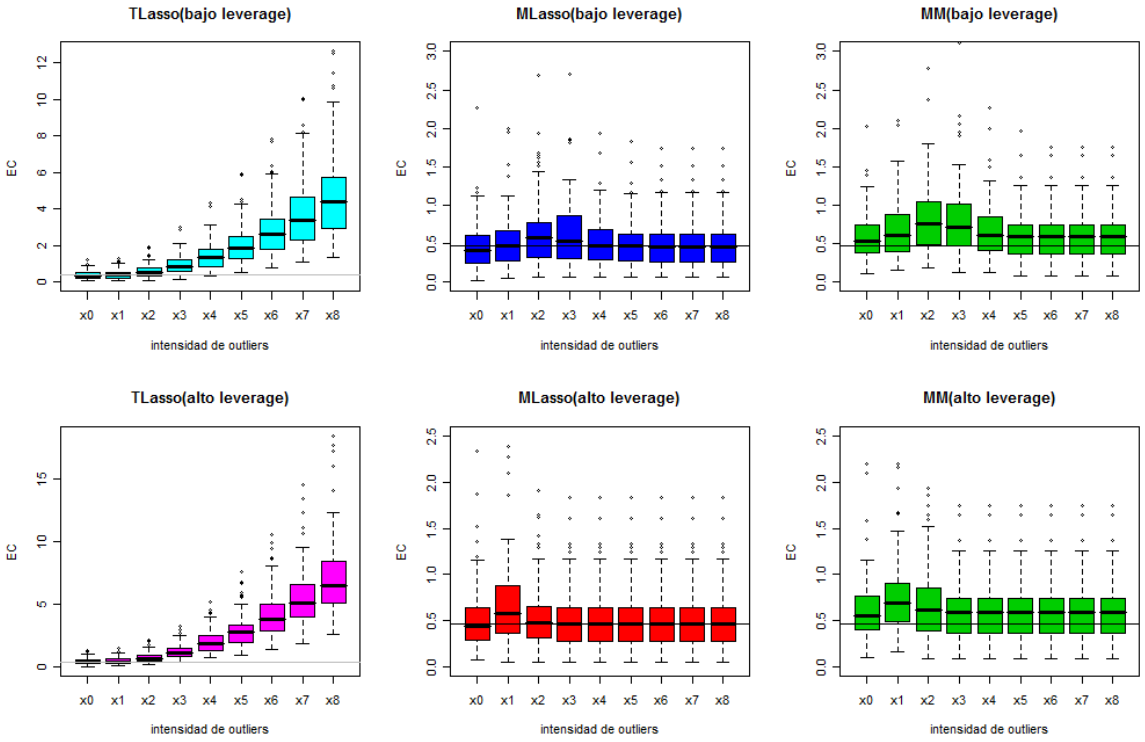


Figura 35: Boxplots del error cuadratico cuando cer=5 $\rho$=0.8

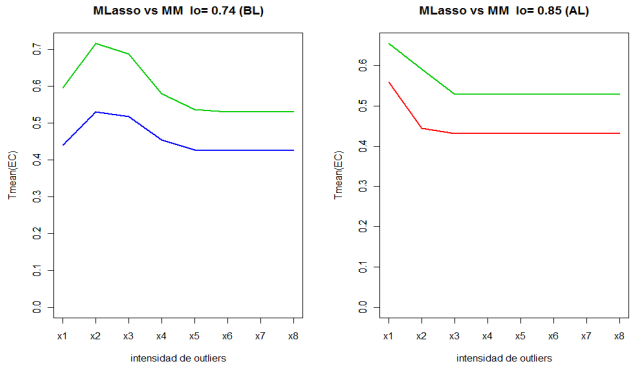


Figura 36: Curvas de error cuadratico medio cer=5 $\rho$=0.8

**Boxplots y curvas de error cuadrático medio cuando cer=8 y $\rho = 0$**

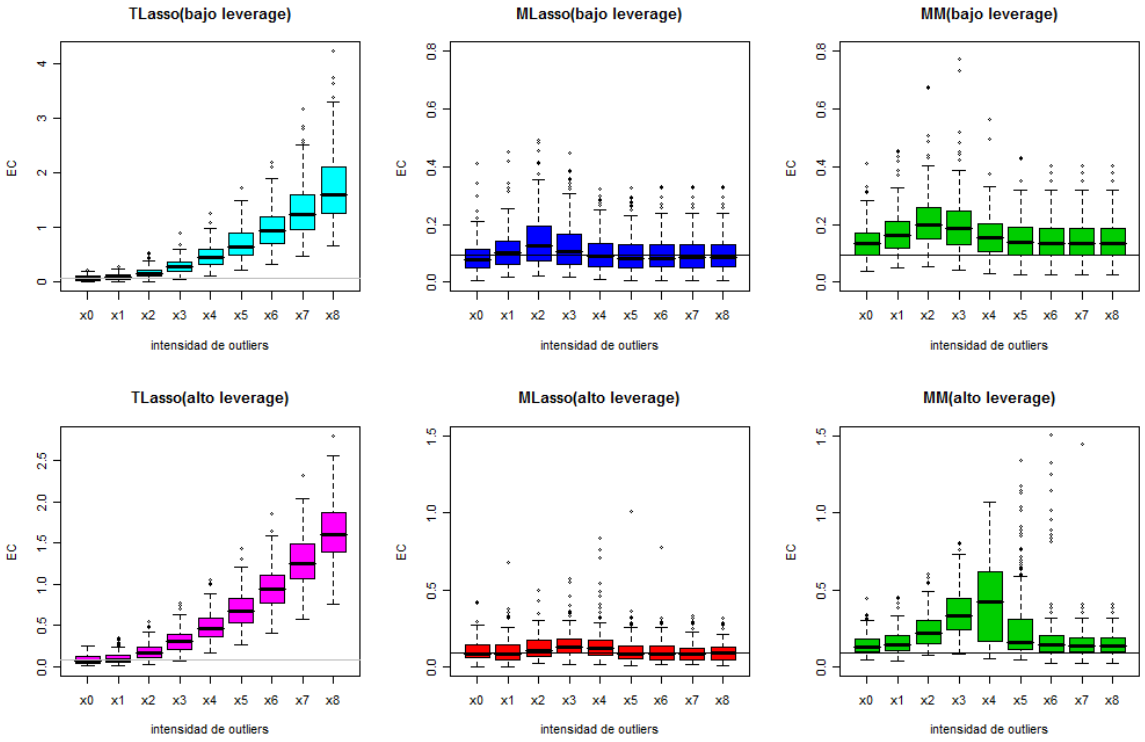


Figura 37: Boxplots del error cuadratico cuando cer=8 $\rho$=0

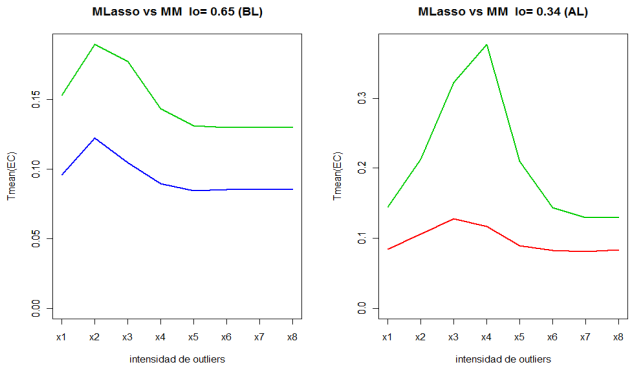


Figura 38: Curvas de error cuadratico medio cer=8 $\rho$=0

**Boxplots y curvas de error cuadrático medio cuando cer=8 y $\rho = 0.8$**

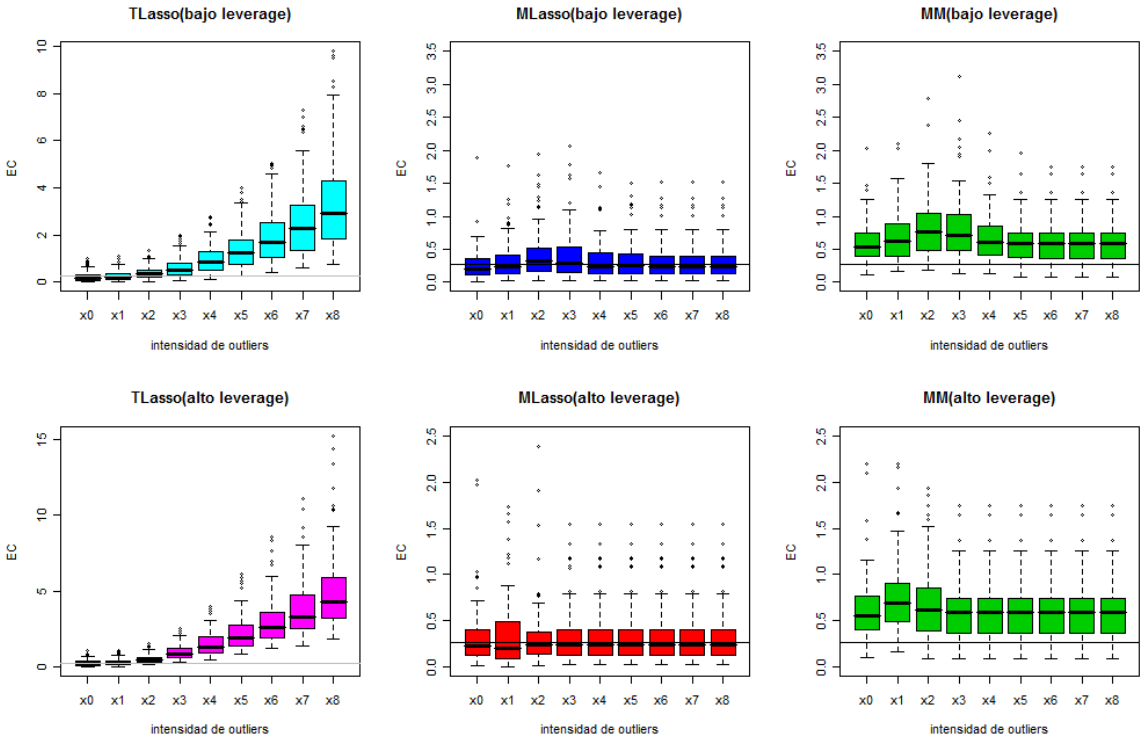


Figura 39: Boxplots del error cuadratico cuando cer=8 $\rho$=0.8

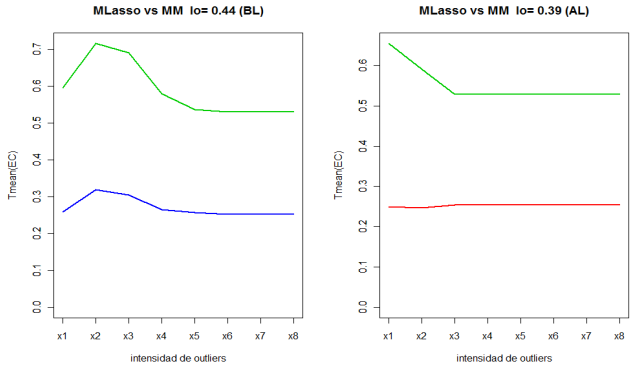


Figura 40: Curvas de error cuadratico medio cer=8 $\rho$=0.8

# 9 Análisis de los resultados de la simulación

## 9.1 Caso sin outliers

(El análisis que sigue se aplica a las Figuras 13, 14, 27 y 28).

En general, tanto para el LASSO de Tibshirani como para el estimador MLASSO, al aumentar el número de ceros en el $\boldsymbol{\beta}$ con que se generaron los datos, disminuye el error cuadrático medio del estimador, y aumenta el número medio de ceros detectados. Sin embargo estos efectos no son lineales, ya que son leves al principio, y mas marcados cuando se incrementa el número de ceros en $\boldsymbol{\beta}$.

Lo anterior es razonable ya que al aumentar los ceros en $\boldsymbol{\beta}$, es cuando el LASSO exhibe su propiedad de interpretabilidad, proporcionando un modelo más simple, con menos predictores, y disminuyendo el error cuadrático medio. Lo mismo sucede cuando se aplica el método de selección de variables a un modelo lineal.

Además la disminución del error cuadrático medio del estimador se hace más importante cuando los predictores están correlacionados (con $\rho = 0.8$). Y el aumento del número medio de ceros detectados se hace levemente mas importante cuando $\rho = 0$.

En síntesis cuando aumentan ceros en $\boldsymbol{\beta}$ el ECM baja (especialmente cuando $\rho = 0.8$) y la esperanza del numero de ceros aumenta (especialmente cuando $\rho = 0$) Lo anterior se refleja en el comportamiento de las eficiencias, tanto para $n = 50$ como para $n = 100$:

- La eficiencia del estimador LASSO de Tibshirani respecto de cuadrados mínimos (**effL**), cuando $\rho = 0$, parte de aproximadamente 0.95 cuando no hay ceros en $\boldsymbol{\beta}$, y aumenta con los ceros en $\boldsymbol{\beta}$, hasta aproximadamente 1.55 (para 80% de ceros). En el caso de $\rho = 0.8$, estas cifras mejoran un poco, variando de 1 a aproximadamente 1.95.

- La eficiencia del estimador MLASSO respecto del MM-estimador de regresión (**effLR**), cuando $\rho = 0$, parte de aproximadamente 0.9 cuando no hay ceros en $\boldsymbol{\beta}$, y aumenta con los ceros en $\boldsymbol{\beta}$, hasta aproximadamente 1.50 (para 80% de ceros). En el caso de $\rho = 0.8$, estas cifras mejoran un poco, variando de 0.95 a 1.90.

Como la variación de estas dos eficiencias es aproximadamente proporcional (**effL** $\propto$ **effLR**), y recordando la expresión

$$\mathbf{EFF} = \frac{\mathbf{effLR}}{\mathbf{effL}}\mathbf{eff}$$

resulta que

- La eficiencia del estimador MLASSO respecto del de Tibshirani (**EFF**) se mantiene aproximadamente constante en alrededor de 0.83 (fué un poco menor cuando n=50, p=5, ya que en este caso la eficiencia del MM.estimador respecto de cuadrados mínimos(**eff**) fué un poco menor).

- El número medio de ceros detectados es similar para ambos estimadores, aumentando con la cantidad de ceros en $\boldsymbol{\beta}$, llegando a detectar aproximadamente el 65% (para 80% de ceros generados). Las cifras de detección son un poco menores cuando $\rho = 0.8$.

## 9.2 Caso con outliers

Aquí se analiza el comportamiento del estimador MLASSO con niveles crecientes de outliers (bajo/alto Leverage), y se lo compara con el LASSO de Tibshirani (no robusto), y con el MM-estimador de regresión, en las mismas condiciones. Estos análisis se repiten variando el número de ceros en $\boldsymbol{\beta}$ y la presencia o no de correlación entre predictores ($\rho = 0$ o $\rho = 0.8$).

Para comparar el estimador MLASSO con el MM-estimador en presencia de outliers con bajo/alto Leverage**,** se utilizó el indicador definido en (34). Para cada combinación ($cer$ ; $\rho$) se representaron las curvas de error cuadrático medio del MLASSO y del MM estimador en función de la intensidad de outliers(ver por ejemplo las Figuras 16, 18, etc) . El indicador de respuesta a outliers se definió

$$I^O = \frac{\max(\mathbf{ecm}_{MLASSO})}{\max(\mathbf{ecm}_{MM})}$$

y se tendrán dos indicadores $I^O_{BL}$ y $I^O_{AL}$ según los outliers sean con bajo/alto leverage(BL o AL).

Este indicador cuando vale 1, significa que el estimador MLASSO tiene un comportamiento similar al del MM-estimador; si toma un valor menor que 1, su comportamiento es mejor, y si es mayor que 1 es peor.

Se presenta a continuación una tabla con este indicador para las dos situaciones analizadas (la información de estas tablas surgen de las salidas de 8.6.3 y 8.6.6, que corresponden a las curvas de error cuadrático medio):

**Indicador de respuesta a outliers MLASSO vs MM-estimador**

| **n=50 p=5** | $I^O_{BL}$ | $I^O_{AL}$ |
|---|---|---|
| cer=0 $\rho$=0 | 1.12 | 0.48 |
| cer=0 $\rho$=0.8 | 1.02 | 1.09 |
| cer=2 $\rho$=0 | 1.05 | 0.41 |
| cer=2 $\rho$=0.8 | 0.84 | 0.87 |
| cer=4 $\rho$=0 | 0.77 | 0.32 |
| cer=4 $\rho$=0.8 | 0.52 | 0.41 |

| **n=100 p=10** | $I^O_{BL}$ | $I^O_{AL}$ |
|---|---|---|
| cer=0 $\rho$=0 | 1.09 | 0.45 |
| cer=0 $\rho$=0.8 | 1.03 | 1.18 |
| cer=5 $\rho$=0 | 0.95 | 0.34 |
| cer=5 $\rho$=0.8 | 0.74 | 0.85 |
| cer=8 $\rho$=0 | 0.65 | 0.34 |
| cer=8 $\rho$=0.8 | 0.44 | 0.39 |

Tabla 7: Indicadores de respuesta a outliers $I^O_{BL}$ y $I^O_{AL}$

Notese que los indicadores son similares tanto para $n = 50$ como para $n = 100$. Se observa en algunos casos un comportamiento levemente peor del estimador MLASSO respecto del MM-estimador, y esto se presenta cuando no hay ceros en los predictores, que es cuando en general el MLASSO responde peor. A medida que aumentan los ceros, ambos indicadores $I^O_{BL}$ y $I^O_{AL}$ mejoran.

Además cuando hay bajo Leverage, la presencia de correlación siempre hace mejorar el indicador $I^O_{BL}$. Lo contrario ocurre cuando hay alto Leverage, ya que $I^O_{AL}$ aumenta.

Se presenta un aspecto interesante en las salidas 8.6.3 y 8.6.6 si se presta atención al número medio de ceros detectados para cada nivel de outliers, tanto para el LASSO de Tibshirani, como para el estimador MLASSO. Como en el caso de bajo leverage, el número medio de ceros detectados varia poco respecto del nivel de outliers, se calculó como indicador el promedio para los niveles $2 \leq j \leq 9$ obteniendo $\overline{E(cer)}_L$ y $\overline{E(cer)}_{ML}$. Se tienen entonces las tablas:

**Media de ceros detectados LASSO vs MLASSO (bajo leverage)**

| **n=50 p=5** | $\overline{E(cer)}_L$ | $\overline{E(cer)}_{ML}$ | **n=100 p=10** | $\overline{E(cer)}_L$ | $\overline{E(cer)}_{ML}$ |
|---|---|---|---|---|---|
| cer=0 $\rho$=0 | 0 | 0 | cer=0 $\rho$=0 | 0 | 0 |
| cer=0 $\rho$=0.8 | 0 | 0 | cer=0 $\rho$=0.8 | 0 | 0 |
| cer=2 $\rho$=0 | 0.93 | 0.99 | cer=5 $\rho$=0 | 2.13 | 2.45 |
| cer=2 $\rho$=0.8 | 0.71 | 0.70 | cer=5 $\rho$=0.8 | 2.26 | 2.04 |
| cer=4 $\rho$=0 | 2.8 | 2.83 | cer=8 $\rho$=0 | 5.14 | 5.36 |
| cer=4 $\rho$=0.8 | 2.31 | 2.39 | cer=8 $\rho$=0.8 | 5.14 | 4.71 |

Tabla 8: Evaluación de los ceros detectados del LASSO vs MLASSO (bajo leverage)

O sea, frente a outliers de bajo Leverage, el LASSO de Tibshirani detecta en promedio aproximadamente la misma cantidad de ceros que el estimador MLASSO. Por lo tanto, en este aspecto el LASSO de Tibshirani tendría un comportamiento robusto.

Si se analizan estas mismas tablas cuando los outliers están en presencia de alto Leverage se tiene:

**Media de ceros detectados LASSO vs MLASSO (alto leverage)**

| **n=50 p=5** | $\overline{E(cer)}_L$ | $\overline{E(cer)}_{ML}$ | **n=100 p=10** | $\overline{E(cer)}_L$ | $\overline{E(cer)}_{ML}$ |
|---|---|---|---|---|---|
| cer=0 $\rho$=0 | 0 | 0 | cer=0 $\rho$=0 | 0 | 0 |
| cer=0 $\rho$=0.8 | 0 | 0 | cer=0 $\rho$=0.8 | 0 | 0 |
| cer=2 $\rho$=0 | 0 | 0.6 | cer=5 $\rho$=0 | 0.08 | 1.43 |
| cer=2 $\rho$=0.8 | 0.59 | 0.55 | cer=5 $\rho$=0.8 | 2.36 | 1.54 |
| cer=4 $\rho$=0 | 0.21 | 1.95 | cer=8 $\rho$=0 | 0.74 | 3.6 |
| cer=4 $\rho$=0.8 | 1.58 | 2.21 | cer=8 $\rho$=0.8 | 4.11 | 5.6 |

Tabla 9: Evaluación de los ceros detectados del LASSO vs MLASSO (bajo leverage)

Nótese aquí que el estimador MLASSO responde en general mejor que el de Tibshirani(siendo esta mejora mas marcada cuando $\rho$=0).

## 9.3 Demoras de las rutinas

A continuación se presentan los tiempos en minutos por cada realización de las dos rutinas mencionadas (incluyendo en todos los casos la validación cruzada,

ajustes del intervalo de rastreo, etc). Estas tablas se construyeron simulando datos sin correlación entre predictores y sin ceros en $\boldsymbol{\beta}$. Cuando la correlación es 0.8 los tiempos informados se incrementan menos del 10%; y si hay ceros en $\boldsymbol{\beta}$, casi no varían.

**Respuesta sin outliers(min)**

| Estimador | **n=50 p=5** | **n=100 p=10** |
|---|---|---|
| **LASSO** | 0.21 | 0.71 |
| **MLASSO** | 1.65 | 7.82 |

Tabla 10: Demoras cuando no hay outliers

**Respuesta con outliers(min)**

| Estimador | **n=50 p=5** | **n=100 p=10** |
|---|---|---|
| **LASSO** | 0.19 | 0.67 |
| **MLASSO** | 1.56 | 7.90 |

Tabla 11: Demoras cuando hay outliers

Debe tenerse en cuenta que en las simulaciones que corresponden a respuesta sin outliers se efectuaron 200 veces, y 120 para las de respuesta con outliers, ya que debido a la parametrización según el tamaño de los outliers demoran mucho mas.

El sistema utilizado para realizar el estudio de Monte Carlo fue el siguiente:

- procesador: Intel Core i7 (4 núcleos) CPU 920 @ 2.67GHz 2.66GHz
- memoria (RAM): 4GB
- disco rígido: 1000GB
- Sistema operativo: Windows Vista ultimate

# 10 Análisis de un ejemplo con datos reales

En un trabajo de Janssens, K., Deraedt, I., Freddy, A., and Veekman, J. (1998), se analizaron 180 vasos de vidrio de los siglos 15-17, que surgieron de excavaciones arqueológicas realizadas en Antwerp, Belgica. En cada uno se registro la composición de 13 diferentes sustancias, y se efectuó un análisis de espectro EPXMA (electron probe X-ray microanalysis) sobre 486 frecuencias.

Para simplificar, en este ejemplo se utilizó el contenido $y_i$ de solo una de las substancias de cada vaso, resultando $\mathbf{y} \in \mathbb{R}^{180}$, y como variables predictoras se consideraron las intensidades del espectro en 30 frecuencias, resultando entonces

la matriz $\mathbf{X} \in \mathbb{R}^{180x30}$.

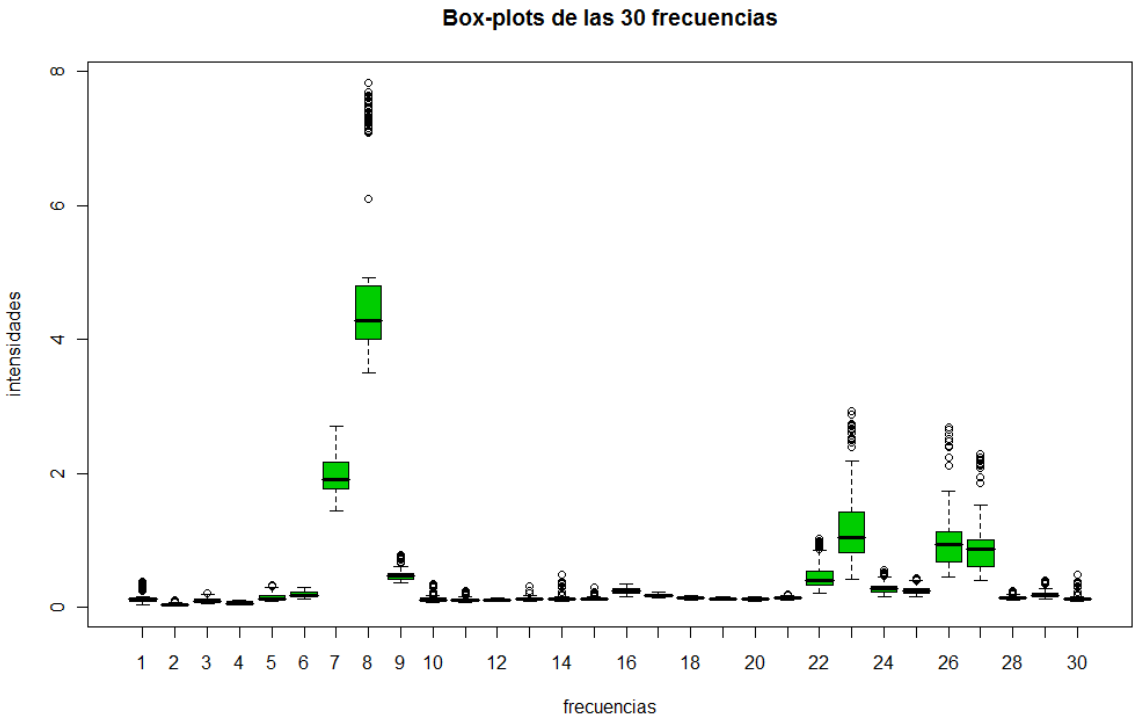


Figura 41: Boxplots para las 30 frecuencias

En la Figura 41 se presentan los boxplots de las intensidades de espectro en todas las frecuencias analizadas.

Se aplicó sobre los datos $(\mathbf{y}, \mathbf{X})$ los estimadores Lasso, MLasso y el MM-estimador, obteniendo respectivamente $\mathbf{b}_L$, $\mathbf{b}_{ML}$ y $\mathbf{b}_{MM}$ según el formato:

$$\mathbf{b} = \begin{array}{c|cccc} & b_1 & b_2 & \cdots & b_{10} \\ b_0 & b_{11} & b_{12} & \cdots & b_{20} \\ & b_{21} & b_{22} & \cdots & b_{30} \end{array}$$

$$\mathbf{b}_L = \begin{array}{c|cccccccccc} & \text{-0.50} & \text{-4.58} & \text{-2.92} & 1.70 & 1.31 & 0.25 & \text{-0.36} & \text{-0.03} & 0.70 & 0.17 \\ 0.30 & 0 & \text{-0.58} & 6.81 & 13.74 & 4.19 & 0.95 & \text{-2.62} & \text{-3.93} & \text{-7.16} & \text{-0.68} \\ & \text{-2.33} & 0 & 0.27 & 0 & \text{-3.08} & 0 & 0 & 1.08 & \text{-1.75} & 1.96 \end{array}$$

$$\mathbf{b}_{ML} = \begin{array}{c|cccccccccc} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \text{-0.01} & 0 & 0 \\ \text{-0.04} & 0 & 0 & 0 & 2.21 & 0 & 0 & \text{-0.24} & 0 & 0 & 0 \\ & 0 & 0 & 0 & 0 & 0.01 & 0 & 0 & 0.04 & 0 & 0.03 \end{array} \quad (35)$$

$$\mathbf{b}_{MM} = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|} \hline & -0.15 & -1.89 & -0.04 & -1.03 & 0.42 & 0.97 & -0.17 & 0.00 & 0.21 & 1.16 \\ \hline 0.32 & -2.20 & 0.63 & 3.11 & 3.83 & 1.40 & 1.09 & -2.00 & -1.63 & -1.09 & -1.89 \\ \hline & -2.74 & 0.00 & 0.08 & 0.16 & -0.08 & -0.08 & 0.14 & 0.62 & -1.80 & 1.10 \\ \hline \end{array}$$

Se observan solo 6 coeficientes diferentes de cero en el MLasso, contra 25 en el Lasso(no robusto), siendo esto último explicable por la presencia de outliers en los datos. Por otro lado, en la Figura 42 se representa la curva de error de predicción del MLasso obtenida mediante cross-validación. El valor óptimo de la restricción fué $t^R_{opt} = 0.057$, mucho menor que $t^R_\infty = 0.865$ que corresponde al MLasso "sin restricción", o sea a un MM-estimador de regresion. Esto explica la cantidad de ceros obtenida.

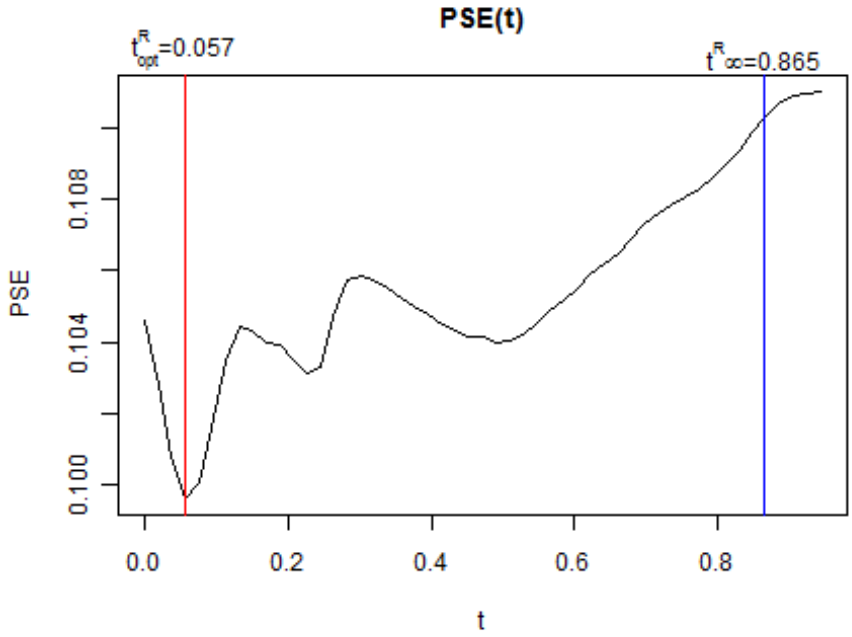


Figura 42: Curva de error de prediccion del MLasso

Con el fin de evaluar los tres estimadores(los dos estimadores Lasso con su correspondiente restricción óptima, y el MM-estimador), se calcularon los residuos mediante cross-validación, y se midió su escala mediante un estimador tau. En la Tabla 14 figuran estos indicadores.

| **estimador** | **escala-tau** |
|---|---|
| Lasso | 0.144 |
| MLasso | 0.100 |
| MM-est. | 0.114 |

Tabla 12: Escala-tau de los residuos

Aquí el MLasso es el que tiene menor escala, levemente inferior al del MM-estimador, pero con una expresión mucho mas simple que permite una mayor interpretablidad.

A continuación, en el MLasso se eliminaron las observaciones cuyo residuo en valor absoluto fué por lo menos 4 veces la escala de la Tabla 12. Luego de eliminar 14 outliers y quedando 166 observaciones, se aplicó el estimador Lasso de Tibshirani, resultando

$$\mathbf{b}_L^* = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|}\hline & 0 & 0 & 0 & 0 & 0 & 0 & -0.08 & 0 & 0 & 0 \\ \hline 0.07 & -0.31 & 0 & 0.48 & 3.46 & 0 & 0 & -0.97 & -0.11 & 0 & 0 \\ \hline & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline \end{array}$$

Notese que aunque esta solución es muy simple(solo 6 coeficientes diferentes de 0), no es muy parecida a la obtenida con el MLassso sobre el total de observaciones (35).

Para finalizar se eliminaron las variables que tienen coeficiente 0 según el MLasso de (35), y se ejecutó el MM-estimador para evaluar la significación, resultando

| | **Estimate** | **Std. Error** | **t value** | **Pr(>\|t\|)** |
|---|---|---|---|---|
| $alfa$ | -0.097 | 0.085 | -1.130 | 0.260 |
| $x_8$ | -0.026 | 0.010 | -2.755 | **0.007** |
| $x_{14}$ | 4.366 | 1.112 | 3.925 | **0.0001** |
| $x_{17}$ | -0.785 | 0.377 | -2.083 | 0.039 |
| $x_{25}$ | -0.170 | 0.205 | -0.829 | 0.408 |
| $x_{28}$ | 0.056 | 0.859 | 0.065 | 0.948 |
| $x_{30}$ | 0.030 | 0.391 | 0.076 | 0.939 |

Tabla 13: Significación con el MM-estimador de las variables seleccionadas con el MLasso

Notese que la variable mas fuertemente significativa es $x_{14}$, que es la que siempre tuvo coeficiente distinto de cero, positivo, y mayor en todos los estimadores analizados en este ejemplo.

# 11 Demostraciones

*Demostración del Teorema 1*

Si $\boldsymbol{\mu}_a \in V_t$ y $\boldsymbol{\mu}_b \in V_t$, habría que probar que

$$\forall \theta \in \mathbb{R} \text{ con } 0 \leq \theta \leq 1 \Longrightarrow \boldsymbol{\mu} = (1-\theta)\boldsymbol{\mu}_a + \theta\boldsymbol{\mu}_b \in V_t.$$

Como $\boldsymbol{\mu}_a = \mathbf{X}\boldsymbol{\beta}_a$ con $\sum_{i=1}^{p} |\beta_{ai}| \leq t$ y $\boldsymbol{\mu}_b = \mathbf{X}\boldsymbol{\beta}_b$ con $\sum_{i=1}^{p} |\beta_{bi}| \leq t$ resulta

$$\boldsymbol{\mu} = (1-\theta)\mathbf{X}\boldsymbol{\beta}_a + \theta\mathbf{X}\boldsymbol{\beta}_b = \mathbf{X}\left[(1-\theta)\boldsymbol{\beta}_a + \theta\boldsymbol{\beta}_b\right],$$

donde

$$\begin{aligned}\sum_{i=1}^{p}|(1-\theta)\beta_{ai}+\theta\beta_{bi}| &\leq (1-\theta)\sum_{i=1}^{p}|\beta_{ai}|+\theta\sum_{i=1}^{p}|\beta_{bi}| \\ &\leq (1-\theta)t+\theta t=t.\end{aligned}$$

Luego $\boldsymbol{\mu}\in V_t$.

*Demostración del Teorema 4*

En este caso la región $V_{rt}$ es la esfera en $\mathbb{R}^p$

$$V_{rt}=\left\{\boldsymbol{\mu}_r^z\ :\ \boldsymbol{\mu}_r^z=\mathbf{z}^{(1)}\beta_1^z+\cdots+\mathbf{z}^{(p)}\beta_p^z,\ (\beta_1^z)^2+\cdots+\left(\beta_p^z\right)^2\leq t\right\}$$

y $\widehat{\boldsymbol{\mu}}_{rRR}^z$ está en el contorno de esta esfera, resultando

$$\widehat{\boldsymbol{\mu}}_{rRR}^z=\mathbf{z}^{(1)}\widehat{\beta}_{1RR}^z+\cdots+\mathbf{z}^{(p)}\widehat{\beta}_{pRR}^z \text{ con } (\widehat{\beta}_{1RR}^z)^2+\cdots+\left(\widehat{\beta}_{pRR}^z\right)^2=t.$$

Luego debido a la ortogonalidad de los $\mathbf{z}^{(i)}$

$$\begin{aligned}\|\widehat{\boldsymbol{\mu}}_{rRR}^z\|^2 &= \left\|\mathbf{z}^{(1)}\widehat{\beta}_{1RR}^z\right\|^2+\cdots+\left\|\mathbf{z}^{(p)}\widehat{\beta}_{pRR}^z\right\|^2 \\ &= (\widehat{\beta}_{1RR}^z)^2+\cdots+(\widehat{\beta}_{pRR}^z)^2=t,\end{aligned}$$

y entonces $\|\widehat{\boldsymbol{\mu}}_{rRR}^z\|=\sqrt{t}$. Finalmente, como en el caso de perfecta ortogonalidad la región $V_{rt}$ es esférica, entonces $\widehat{\boldsymbol{\mu}}_{rRR}^z$ y $\widehat{\boldsymbol{\mu}}_{rls}^z$ serán colineales y entonces resultará $k=\|\widehat{\boldsymbol{\mu}}_{rls}^z\|/\|\widehat{\boldsymbol{\mu}}_{rRR}^z\|=\|\widehat{\boldsymbol{\mu}}_{rls}^z\|/\sqrt{t}$.

*Demostración del Teorema 5.*

Sea para cierto $t\geq 0$, $\widehat{\boldsymbol{\beta}}_{L,1}^t$ solución del LASSO

$$\widehat{\boldsymbol{\beta}}_{L,1}^t=\underset{\sum_{j=1}^{p}|\beta_j|\leq t}{\arg\min}\sum_{i=1}^{n}(y_i-\mathbf{x}_i\boldsymbol{\beta})^2\ , \tag{36}$$

y para cierto $\lambda\geq 0$, $\widehat{\boldsymbol{\beta}}_{L}^{\lambda}$ solución de

$$\widehat{\boldsymbol{\beta}}_{L,2}^{\lambda}=\underset{\boldsymbol{\beta}}{\arg\min}\left(\sum_{i=1}^{n}(y_i-\mathbf{x}_i\boldsymbol{\beta})^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right). \tag{37}$$

Llamemos $A(\boldsymbol{\beta})=\sum_{i=1}^{n}(y_i-\mathbf{x}_i\boldsymbol{\beta})^2$ y $B(\boldsymbol{\beta})=\sum_{j=1}^{p}\left|\beta_j\right|$.

**(a)** Como $A(\boldsymbol{\beta})$ es función continua y $B(\boldsymbol{\beta})\leq t$ es compacto (por ser cerrado y acotado), mediante el teorema de Weierstrass resulta que existe al menos un mínimo global en la (36). Supongamos ahora que si para $t\geq 0$ existiesen $\boldsymbol{\beta}_1$ y $\boldsymbol{\beta}_2$ soluciones de (36) con $\boldsymbol{\beta}_1\neq\boldsymbol{\beta}_2$. Entonces

$$A(\boldsymbol{\beta}_1)=A(\boldsymbol{\beta}_2),$$

ya que si no se cumpliese la igualdad, no serían ambas soluciones. Llamese ahora $\mathbf{X}\boldsymbol{\beta}_1 = \boldsymbol{\mu}_1$ y $\mathbf{X}\boldsymbol{\beta}_2 = \boldsymbol{\mu}_2$. Si $\mathbf{X}$ es de rango completo tenemos también que $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. Entonces

$$A(\boldsymbol{\beta}_1) = \|\mathbf{y} - \boldsymbol{\mu}_1\|^2 = \|\mathbf{y} - \boldsymbol{\mu}_2\|^2 = A(\boldsymbol{\beta}_2).$$

Como $\boldsymbol{\mu}_1 \in V_t$, $\boldsymbol{\mu}_2 \in V_t$ y $V_t$ es convexo, resultará también

$$\boldsymbol{\mu}^* = 0.5\boldsymbol{\mu}_1 + 0.5\boldsymbol{\mu}_2 \in V_t.$$

Además valen las igualdades

$$(\mathbf{y} - \boldsymbol{\mu}^*) + \frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{2} = \mathbf{y} - \boldsymbol{\mu}_2 \tag{38}$$

$$(\mathbf{y} - \boldsymbol{\mu}^*) - \frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{2} = \mathbf{y} - \boldsymbol{\mu}_1, \tag{39}$$

y como las $\|\cdot\|^2$ de la derecha son iguales, resultará también

$$\begin{aligned} & \|\mathbf{y} - \boldsymbol{\mu}^*\|^2 + 2(\mathbf{y} - \boldsymbol{\mu}^*)' \frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{2} + \left\| \frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{2} \right\|^2 \\ = \quad & \|\mathbf{y} - \boldsymbol{\mu}^*\|^2 - 2(\mathbf{y} - \boldsymbol{\mu}^*)' \frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{2} + \left\| \frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{2} \right\|^2, \end{aligned}$$

o sea

$$(\mathbf{y} - \boldsymbol{\mu}^*)' \frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{2} = 0$$

y esto quiere decir que $\mathbf{y} - \boldsymbol{\mu}^*$ es ortogonal a $\frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{2}$. Luego usando la primera de las ecuaciones en (38) resulta

$$\|\mathbf{y} - \boldsymbol{\mu}^*\|^2 + \left\| \frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{2} \right\|^2 = \|\mathbf{y} - \boldsymbol{\mu}_2\|^2,$$

y como $\left\| \frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{2} \right\|^2 > 0$ se cumplirá también que

$$\|\mathbf{y} - \boldsymbol{\mu}^*\|^2 < \|\mathbf{y} - \boldsymbol{\mu}_2\|^2 = \|\mathbf{y} - \boldsymbol{\mu}_1\|^2.$$

Entonces habríamos encontrado otra solución $\boldsymbol{\mu}^* = \mathbf{X}\boldsymbol{\beta}^*$ de la (36) en que

$$A(\boldsymbol{\beta}^*) < A(\boldsymbol{\beta}_1) = A(\boldsymbol{\beta}_2),$$

y esto es un absurdo. Luego debe cumplirse que $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ (unicidad).

(**b**) Para la existencia se utilizará el siguiente Lema

**Lemma 4** *Si* $\mathbf{X}'\mathbf{X}$ *es definida positiva vale*

$$\forall M \quad \exists k > 0 : \forall \boldsymbol{\beta} \ con \ \|\boldsymbol{\beta}\| > k \Longrightarrow A(\boldsymbol{\beta}) + \lambda B(\boldsymbol{\beta}) > M.$$

**Proof.** Como $\lambda \geq 0$ y $B(\boldsymbol{\beta}) \geq 0$, resultará $A(\boldsymbol{\beta}) + \lambda B(\boldsymbol{\beta}) \geq A(\boldsymbol{\beta})$. Además, debido a la ortogonalidad

$$\begin{aligned} A(\boldsymbol{\beta}) &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{ls} - \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_{ls})\|^2 \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{ls}\|^2 + \|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_{ls})\|^2 \geq \|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_{ls})\|^2 \end{aligned}$$

Como $\mathbf{X}'\mathbf{X}$ es definida positiva, su menor autovalor resultará $\theta_{p+1} > 0$.Luego $\|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_{ls})\|^2 \geq \theta_{p+1} \|\boldsymbol{\beta} - \boldsymbol{\beta}_{ls}\|^2$. Luego se tiene

$$A(\boldsymbol{\beta}) + \lambda B(\boldsymbol{\beta}) \geq \theta_{p+1} \|\boldsymbol{\beta} - \boldsymbol{\beta}_{ls}\|^2$$

y eligiendo u $k > \|\boldsymbol{\beta}_{ls}\|$, entonces $\forall \boldsymbol{\beta}$ con $\|\boldsymbol{\beta}\| > k$ resultará $\|\boldsymbol{\beta} - \boldsymbol{\beta}_{ls}\| > k - \|\boldsymbol{\beta}_{ls}\|$, quedando la desigualdad anterior

$$\forall \boldsymbol{\beta} \text{ con } \|\boldsymbol{\beta}\| > k \Longrightarrow A(\boldsymbol{\beta}) + \lambda B(\boldsymbol{\beta}) > \theta_{p+1}(k - \|\boldsymbol{\beta}_{ls}\|)^2$$

Finalmente, eligiendo un $k$ que cumpla también $\theta_{p+1}(k - \|\boldsymbol{\beta}_{ls}\|)^2 > M$ quedará probado el Lemma.

Para la existencia, como $A(\boldsymbol{\beta}) + \lambda B(\boldsymbol{\beta})$ es función continua, y para $k_0 > 0, \|\boldsymbol{\beta}\| \leq k_0$ es un compacto(por ser cerrado y acotado), mediante el teorema de Weierstrass resulta que en $\|\boldsymbol{\beta}\| \leq k_0$ existe al menos un mínimo global de la (37). Se designará $M_0$ a uno de estos mínimos. Luego utilizando el Lema(4), resulta que para $M_0$ existirá un $k > 0$, tal que para $\|\boldsymbol{\beta}\| > k \Longrightarrow A(\boldsymbol{\beta}) + \lambda B(\boldsymbol{\beta}) > M_0$. Si ahora se considera la región $\|\boldsymbol{\beta}\| \leq k$, que es un compacto, nuevamente por el teorema de Weierstrass resulta que existirá al menos un mínimo global en la (37), que se designará $M$. Se demostrará que cualesquiera sea $k$, siempre $M \leq M_0$.

En efecto notar que si $k_0 \leq k$, como vale la inclusión $\|\boldsymbol{\beta}\| \leq k_0 \subseteq \|\boldsymbol{\beta}\| \leq k$, resultará $M \leq M_0$. Por otro lado si $k_0 > k$, como para $\|\boldsymbol{\beta}\| > k$ siempre $A(\boldsymbol{\beta}) + \lambda B(\boldsymbol{\beta}) > M_0$, resultará en este caso $M = M_0$. Luego en definitiva:

$$\begin{aligned} \|\boldsymbol{\beta}\| &\leq k \Longrightarrow M \leq M_0 \\ \|\boldsymbol{\beta}\| &> k \Longrightarrow A(\boldsymbol{\beta}) + \lambda B(\boldsymbol{\beta}) > M_0 \end{aligned}$$

Esto quiere decir que el mínimo de la (37) estará en el compacto $\|\boldsymbol{\beta}\| \leq k$. Queda probada la existencia. ■

Para la unicidad notar que en la (37) $A(\boldsymbol{\beta})$ es una función estrictamente convexa, y $B(\boldsymbol{\beta})$ convexa. Como $\lambda \geq 0$, resultará también $A(\boldsymbol{\beta}) + \lambda B(\boldsymbol{\beta})$ estrictamente convexa. Luego el mínimo en $\mathbb{R}^{p+1}$ será único.

**(c)** Si para $\lambda \geq 0$ resulta $\widehat{\boldsymbol{\beta}}_{L,2}^{\lambda}$ la solución de (37) esto quiere decir que $\forall \boldsymbol{\beta} \in \mathbb{R}^{p+1}$

$$\begin{aligned} A(\boldsymbol{\beta}) + \lambda B(\boldsymbol{\beta}) &\geq A(\widehat{\boldsymbol{\beta}}_{L,2}^{\lambda}) + \lambda B(\widehat{\boldsymbol{\beta}}_{L,2}^{\lambda}) \\ A(\boldsymbol{\beta}) - A(\widehat{\boldsymbol{\beta}}_{L,2}^{\lambda}) &\geq \lambda(B(\widehat{\boldsymbol{\beta}}_{L,2}^{\lambda}) - B(\boldsymbol{\beta})). \end{aligned}$$

Luego si se designa $B(\widehat{\boldsymbol{\beta}}^{\lambda}_{L,2}) = t \geq 0$, entonces en la segunda ecuación, si se cumple $B(\boldsymbol{\beta}) \leq t$ y se tendrá

$$\forall \boldsymbol{\beta} : B(\boldsymbol{\beta}) \leq t \implies A(\boldsymbol{\beta}) - A(\widehat{\boldsymbol{\beta}}^{\lambda}_{L,2}) \geq \lambda(t - B(\boldsymbol{\beta})) \geq 0,$$

o sea

$$\forall \boldsymbol{\beta} : B(\boldsymbol{\beta}) \leq t \implies A(\boldsymbol{\beta}) \geq A(\widehat{\boldsymbol{\beta}}^{\lambda}_{L,2}).$$

En particular para $\widehat{\boldsymbol{\beta}}^{t}_{L,1}$ resultará $A(\widehat{\boldsymbol{\beta}}^{t}_{L,1}) \geq A(\widehat{\boldsymbol{\beta}}^{\lambda}_{L,2})$. Pero como $B(\widehat{\boldsymbol{\beta}}^{\lambda}_{L,2}) = t$, y $\widehat{\boldsymbol{\beta}}^{t}_{L,1}$ es solución de la (36), resultará también $A(\widehat{\boldsymbol{\beta}}^{\lambda}_{L,2}) \geq A(\widehat{\boldsymbol{\beta}}^{t}_{L,1})$. En definitiva se cumplirá $A(\widehat{\boldsymbol{\beta}}^{t}_{L.1}) = A(\widehat{\boldsymbol{\beta}}^{\lambda}_{L,2})$, y por la unicidad en (36) esto implica $\widehat{\boldsymbol{\beta}}^{t}_{L,1} = \widehat{\boldsymbol{\beta}}^{\lambda}_{L,2}$

**(d)** Primero vamos a mostrar que $C(\boldsymbol{\lambda}) = B(\widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}}_{L,2})$ es continua. en $\boldsymbol{\lambda}$. Como $B(\boldsymbol{\beta})$ es continua basta mostrar que $\mathbf{b}(\boldsymbol{\lambda}) = \widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}}_{L,2}$ es continua en $\boldsymbol{\lambda}$. Supongamos que no lo sea. Luego existe una sucesion $\boldsymbol{\lambda}_1, .., \boldsymbol{\lambda}_n$ tal que $\boldsymbol{\lambda}_n \to \boldsymbol{\lambda}_0$ y $\mathbf{b}(\boldsymbol{\lambda}_n) = \widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}_n}_{L,2} \nrightarrow \widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}_0}_{L,2}$. Como $\widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}_n}_{L,2}$ estan en un compacto, sin pérdida de generalidad podemos suponer que $\widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}_n}_{L,2} \to \boldsymbol{\beta}^* \neq \widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}_0}_{L,2}$.

Como $\widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}_0}_{L,2}$ es único tenemos que

$$A(\boldsymbol{\beta}^*) + \boldsymbol{\lambda}_0 B(\boldsymbol{\beta}^*) > A(\widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}_0}_{L,2}) + \boldsymbol{\lambda}_0 B(\widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}_0}_{L,2}).$$

Usando el hecho que $A(\boldsymbol{\beta})$ y $B(\boldsymbol{\beta})$ son continuas existe $\boldsymbol{\beta}^{**} \in V_{\boldsymbol{\lambda}^*}$ con $\boldsymbol{\lambda}^* < \boldsymbol{\lambda}_0$ tal que

$$A(\boldsymbol{\beta}^*) + \boldsymbol{\lambda}_0 B(\boldsymbol{\beta}^*) > A(\boldsymbol{\beta}^{**}) + \boldsymbol{\lambda}_0 B(\boldsymbol{\beta}^{**}).$$

Por otro lado tenemos usando nuevamente la continudad de $A(\boldsymbol{\beta})$ y $B(\boldsymbol{\beta})$ resulta

$$\begin{aligned} A(\widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}_n}_{L,2}) + \boldsymbol{\lambda}_n B(\widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}_n}_{L,2}) &\to A(\boldsymbol{\beta}^*) + \boldsymbol{\lambda}_0 B(\boldsymbol{\beta}^*), \\ A(\boldsymbol{\beta}^{**}) + \boldsymbol{\lambda}_n B(\boldsymbol{\beta}^{**}) &\to A(\boldsymbol{\beta}^{**}) + \boldsymbol{\lambda}_0 B(\boldsymbol{\beta}^{**}). \end{aligned}$$

Por lo tanto existe un $n_0$ tal que $\boldsymbol{\lambda}_{n_0} < \boldsymbol{\lambda}^*$ y

$$A(\widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}_n}_{L,2}) + \boldsymbol{\lambda}_n B(\widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}_n}_{L,2}) > A(\boldsymbol{\beta}^{**}) + \boldsymbol{\lambda}_n B(\boldsymbol{\beta}^{**}).$$

Esto contradice la definición de $\widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}_n}_{L,2}$. Por lo tanto $\mathbf{b}(\boldsymbol{\lambda})$ es continua y $C(\boldsymbol{\lambda}) = B(\widehat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}}_{L,2})$ también es continua.

Ya hemos visto que

$$C(0) = t_\infty, \tag{40}$$

ahora vamos a mostrar que

$$\lim_{\boldsymbol{\lambda} \to \infty} C(\boldsymbol{\lambda}) = 0. \tag{41}$$

Supongamos que (41) no sea cierto, luego existe $\varepsilon > 0$ y una sucesión $\boldsymbol{\lambda}_n \to \infty$ tal que $C(\boldsymbol{\lambda}_n) > \varepsilon$

Pero entonces

$$A(\widehat{\boldsymbol{\beta}}_{L,2}^{\boldsymbol{\lambda}_n}) + \boldsymbol{\lambda}_n B(\widehat{\boldsymbol{\beta}}_{L,2}^{\boldsymbol{\lambda}_n}) \geq \varepsilon \boldsymbol{\lambda}_n$$

y por lo tanto

$$\lim_{n\to\infty} (A(\widehat{\boldsymbol{\beta}}_{L,2}^{\boldsymbol{\lambda}_n}) + \boldsymbol{\lambda}_n B(\widehat{\boldsymbol{\beta}}_{L,2}^{\boldsymbol{\lambda}_n})) = \infty. \tag{42}$$

En cambio

$$\lim_{n\to\infty} (A(\mathbf{0}) + \boldsymbol{\lambda}_n B(\mathbf{0})) = A(0) \tag{43}$$

Por lo tanto existe $n_0$ tal que

$$A(\mathbf{0}) + \boldsymbol{\lambda}_n B(\mathbf{0}) < A(\widehat{\boldsymbol{\beta}}_{L,2}^{\boldsymbol{\lambda}_n}) + \boldsymbol{\lambda}_n B(\widehat{\boldsymbol{\beta}}_{L,2}^{\boldsymbol{\lambda}_n})$$

contradiciendo la definición de $\widehat{\boldsymbol{\beta}}_{L,2}^{\boldsymbol{\lambda}_n}$.

Sea ahora $t$ tal que $0 < t < t_\infty$. Como $C(\boldsymbol{\lambda})$ es continua, usando el teorema del punto intermedio, (40) y (41) resulta que existe $\lambda$ tal que $\mathbf{C}(\boldsymbol{\lambda}) = \mathbf{t}$. Vamos a mostrar que

$$\widehat{\boldsymbol{\beta}}_{L,2}^{\lambda} = \widehat{\boldsymbol{\beta}}_{L,1}^{t}.$$

Supongamos que esto no es cierto, luego existe $\boldsymbol{\beta}^* \in V_t$ tal que

$$A(\boldsymbol{\beta}^*) < A(\widehat{\boldsymbol{\beta}}_{L,2}^{\lambda}).$$

Como además se tiene

$$B(\boldsymbol{\beta}^*) \leq t = B(\widehat{\boldsymbol{\beta}}_{L,2}^{\lambda})$$

resulta

$$A(\boldsymbol{\beta}^*) + \lambda B(\boldsymbol{\beta}^*) < A(\widehat{\boldsymbol{\beta}}_{L,2}^{\lambda}) + \lambda B(\widehat{\boldsymbol{\beta}}_{L,2}^{\lambda}),$$

y esto contradice la definición de $\widehat{\boldsymbol{\beta}}_{L,2}^{\lambda}$.

*Demostración del Lema 1*

**(a)** pues $\widehat{\boldsymbol{\beta}}_L = \widehat{\boldsymbol{\beta}}_{\mathbf{ls}}$ satisface trivialmente la (3).

**(b)** pues si existiera una solución $\widehat{\boldsymbol{\beta}}_L^*$ que cumpliese $\left|\widehat{\beta}_{L1}^*\right| + \cdots + \left|\widehat{\beta}_{Lp}^*\right| = t_* < t_1$, y como la región $|\beta_1| + \cdots + |\beta_p| \leq t_1$ es convexa, siempre existirá en el segmento de $\mathbb{R}^p$ determinado por $\widehat{\boldsymbol{\beta}}_L^*$ y $\widehat{\boldsymbol{\beta}}_{\mathbf{ls}}$, un punto intermedio $\widehat{\boldsymbol{\beta}}_L^{**}$, que pertenece a la región. Pero con $\left\|\widehat{\boldsymbol{\beta}}_L^{**} - \widehat{\boldsymbol{\beta}}_{\mathbf{ls}}\right\|^2 < \left\|\widehat{\boldsymbol{\beta}}_L^{*} - \widehat{\boldsymbol{\beta}}_{\mathbf{ls}}\right\|^2$ (contradicción).

*Demostración del Lema 2*

Se analizará solo el caso en que $t = t_1 < t_\infty$ ya que el otro es trivial.

**(a)** Supongase que $\widehat{\boldsymbol{\beta}}_L$ solución tiene $\widehat{\beta}_{Li} = b \neq 0$ con $\widehat{\beta}_{lsi} = 0$. Si se reemplaza $\widehat{\beta}_{Li}$ por cero, el mínimo en (3) sería menor, cumpliéndose $\left|\widehat{\beta}_{L1}\right| + \cdot + 0 + \cdot + \left|\widehat{\beta}_{Lp}\right| = t_* < t_1$ lo cual esta en contradicción con el lema anterior.

**(b)** Si $\widehat{\boldsymbol{\beta}}_L$ es solución, y $\widehat{\beta}_{Li} = 0$ ya está demostrado. Cuando $\widehat{\beta}_{Li} \neq 0$, si $\mathrm{sg}(\widehat{\beta}_{Li}) \neq \mathrm{sg}(\widehat{\beta}_{lsi}) \neq 0$, entonces cambiando en $\widehat{\boldsymbol{\beta}}_L$ el signo de $\widehat{\beta}_{Li}$, se obtendría un mínimo menor y cumpliendo la restricción. Luego deberá ser $\mathrm{sg}(\widehat{\beta}_{Li}) = \mathrm{sg}(\widehat{\beta}_{lsi})$.

*Demostración del Lema 3* :

Se analizará solo el caso en que $t = t_1 < t_\infty$ ya que el otro es trivial. Supóngase que en $\widehat{\boldsymbol{\beta}}_L$, existe $\widehat{\beta}_{Li}$ con $\left|\widehat{\beta}_{Li}\right| > \left|\widehat{\beta}_{lsi}\right|$ . Notar que no podrá ser que $\forall j \neq i$, $\left|\widehat{\beta}_{Lj}\right| \geq \left|\widehat{\beta}_{lsj}\right|$ pues entonces resultaría $t_1 > t_\infty$ (contradicción). Luego debe $\exists k \neq i$ que cumpla $\left|\widehat{\beta}_{Lk}\right| < \left|\widehat{\beta}_{lsk}\right|$. Entonces para cierto $\theta > 0$, definiendo $\widehat{\beta}^*_{Li}$ =sg$(\widehat{\beta}_{lsi})(\left|\widehat{\beta}_{Li}\right| - \theta)$, y $\widehat{\beta}^*_{Lk}$ =sg$(\widehat{\beta}_{lsk})(\left|\widehat{\beta}_{Lk}\right| + \theta)$, resultará que en la (3) se seguirá cumpliendo que $\left|\widehat{\beta}_{L1}\right| + \cdot + \left|\widehat{\beta}^*_{Li}\right| + \cdot + \left|\widehat{\beta}^*_{Lk}\right| + \cdot + \left|\widehat{\beta}_{Lp}\right| = t_1$ pero ahora el mínimo será menor(contradicción).

Los Lemas 5-7 que se dan a continuación serán usados en la demostración del Teorema 3.

**Lemma 5** *Supongamos $Z'Z = I$. Si $\delta_i < \left|\widehat{\beta}_{lsi}\right|$ y $\delta_j < \left|\widehat{\beta}_{lsj}\right|$ entonces $\delta_i = \delta_j$.*

*Demostración.* Sin pérdida de generalidad podemos suponer que $i = 1$ y $j = 2$ con $\delta_1 \leq \delta_2$.. Supongamos que $\delta_1 < \delta_2$. Entonces podemos encontrar $\varepsilon > 0$ tal que $\delta_1 + \varepsilon < \delta_2$, $\delta_1 + \varepsilon < \left|\widehat{\beta}_{ls1}\right|$. Pongamos $\delta^*_1 = \delta_1 + \varepsilon$ y $\delta^*_2 = \delta_2 - \varepsilon$. Luego $(\delta^*_1, \delta^*_2, \delta_2, ..., \delta_p)$ también satisface las restricciones del LASSO. Por otro lado

$$\delta^{*2}_1 + \delta^{*2}_2 = \delta^2_1 + \delta^2_2 + 2\varepsilon^2 + 2\varepsilon(\delta_1 - \delta_2)$$

y como $\delta_1 - \delta_2 < -\varepsilon$ se tiene $\delta^{*2}_1 + \delta^{*2}_2 < \delta^2_1 + \delta^2_2$. Luego

$$\delta^{*2}_1 + \delta^{*2}_2 + \sum_{i=3}^{p} \delta^2_i < \delta^2_1 + \delta^2_2 + \sum_{i=3}^{p} \delta^2_i$$

y esto contradice la optimalidad de $(\delta_1, ...\delta_p)$.

**Lemma 6** *Supongamos $Z'Z = I$. Si $t < t_\infty$ entonces si $\delta_i = \left|\widehat{\beta}_{lsi}\right|$ y $\left|\widehat{\beta}_{lsj}\right| < \left|\widehat{\beta}_{lsi}\right|$ entonces $\delta_j = \left|\widehat{\beta}_{lsj}\right|$.*

*Demostración.* Sin pérdida de generalidad podemos suponer que $j = 1$ y $i = 2$ .. Supongamos que $\delta_1 < \left|\widehat{\beta}_{ls1}\right|$.. Entonces podemos encontrar $\varepsilon > 0$ tal que $\delta_1 + \varepsilon < \delta_2$, $\delta_1 + \varepsilon < \left|\widehat{\beta}_{ls1}\right|$. Pongamos $\delta^*_1 = \delta_1 + \varepsilon$ y $\delta^*_2 = \delta_2 - \varepsilon$. Luego la demostración sigue como en el Lema anterior.

**Lemma 7** *Supongamos $Z'Z = I$. Si Si $t < t_\infty$ entonces si $\delta_i = \left|\widehat{\beta}_{lsi}\right|$ y $\left|\widehat{\beta}_{lsj}\right| > \left|\widehat{\beta}_{lsi}\right|$ entonces $\delta_j \geq \delta_i$.*

*Demostración.*Sin pérdida de generalidad podemos suponer que $i = 1$ y $j = 2$.Supongamos que $\delta_2 < \delta_1$. Entonces se podrá encontrar $\varepsilon > 0$ tal que

$\delta_2 + \varepsilon < \delta_1, \delta_2 + \varepsilon < \left|\widehat{\beta}_{ls2}\right|$. Pongamos $\delta_2^* = \delta_2 + \varepsilon$ y $\delta_1^* = \delta_1 - \epsilon$. Luego la demostración sigue como en el Lema 5.

*Demostración del Teorema 3*

**(a)** Sea $j_0$ el menor valor de $i$ tal que $\delta_i < \left|\widehat{\beta}_{lsi}\right|$. Por el lema 6 $\delta_i = \left|\widehat{\beta}_{lsi}\right|$ para $i < j_0$ y por lo tanto $\widehat{\beta}_{Li} = 0$. Por el lema 5 todos los $\delta_i$ son iguales para $i \geq j_0$ y deben ser iguales a $\delta^*$ para que $\sum_{i=1}^{p} \delta_i = \delta$.

**(b)** $\delta^* < \left|\widehat{\beta}_{lsj_0}\right|$ por definición de $j_0$ en la demostración de (a). El lema 7 implica $\left|\widehat{\beta}_{ls,j_0-1}\right| \leq \delta^*$.

(c) Observemos que (8) implica que que $\widehat{\beta}_{lsj_0-1} \leq \delta^* < \widehat{\beta}_{lsj_0-1}$ es equivalente a $a_{j_0-1} \leq \delta < a_{j_0}$.

# 12 Rutinas en R

## 12.1 Función R para calcular el estimador MLASSO

## 12.2 Comentarios sobre la función Mlasso

## 12.3 Rutinas de estandarización y auxiliares

## 12.4 Estimador de escala tau (rutina tauscale1)

## 12.5 Estimador de escala para el estimador MLASSO (rutina scaleR)

## 12.6 Análisis sin outliers

## 12.7 Análisis con outliers

# 13 Bibliografía


1. *Tibshirani R. (1996)* Regression shrinkage and selection via the LASSO. J. Royal Statistic Soc. B, **58**, 267-288.

2. *Hastie T. y Tibshirani R. y Friedman J. (2001)*: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag. 763 pages.

3. *Miller A. (2002)*:Subset Selection in Regression (Chapman & Hall/CRC Monographs on Statistics & Applied Probability).

4. *Maronna R.A. y Douglas Martin R. y Yohai V.J.(2006)*: Robust Statistics, Theory and Methods. John Wiley.

5. *Yohai V. J. y Maronna R. A.(2010)*: Correcting MM estimates for "fat" data sets. Elsevier.

6. *Yohai V.J. y Zamar R.H. (1988)*: High breakdown estimates of regression by means of the minimization of an efficient scale, Journal of the American Statistical Association, **83**, 406-413.

7. *Yohai V.J. (1987):* High breakdown-point and high efficiency estimates for regression, The Annals of Statistics, **15**, 642-656.

8. *Maronna Ricardo A(2010)*: Robust Ridge Regression for High-Dimensional Data. American Statistical Association and American Society for Quality TECHNOMETRICS, Accepted for publication DOI 10.1198/TECH.2010.09114.