UNIVERSIDAD NACIONAL DE COLOMBIA

Comparación de la Regresión GINI con la Regresión de Mínimos Cuadrados Ordinarios y otros Modelos de Regresión Lineal Robustos

Gloria Patricia Carmona Florez

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2014

Comparación de la Regresión GINI con la Regresión de Mínimos Cuadrados Ordinarios y otros Modelos de Regresión Lineal Robustos

Gloria Patricia Carmona Florez

Trabajo de grado presentado como requisito parcial para optar al título de:
Magister en Ciencias-Estadística

Director:
Ph.D., Juan Carlos Correa Morales

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2014

Resumen

El método de los Mínimos Cuadrados Ordinarios - OLS - es uno de los más usados para estimar la relación entre una variable dependiente (Y) e independientes (X). El modelo de regresión está dado por la relación $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$. Sin embargo, OLS es sensible a observaciones atípicas, las cuales podrían no ser de interés para el investigador, por lo cual es recomendable usar métodos robustos que superen las limitaciones del método OLS. La regresión de Gini es uno de los métodos que podría tener cierto grado de robustez según la literatura (Olkin y Yitzhaki, 1992) debido a la forma matemática como está planteada.

En este trabajo se compara la regresión de Gini (usando el enfoque no parametrico de promedios ponderados de pendientes, en lugar de usar el enfoque parámetrico) con la regresión OLS y otros métodos de regresión robustos, del tipo L (LAV, combinaciones lineales de estadísticos de orden), del tipo M (M de Huber, basado en el concepto de máxima verosimilitud) y del tipo MM (basado en la minimización de un estimador M).

La comparación de los métodos se realiza vía simulación bajo diferentes escenarios: Uno de normalidad de los errores (con $\mu = 0$ y $\sigma^2 = 1$) y tres escenarios de normalidad contaminada con un dato atípico, en los cuales se aumenta progresivamente la magnitud de la observación atípica (en 4σ, 8σ, 16σ). Además, se investiga el efecto del tamaño muestral ($n_1 = 10$, $n_2 = 30$ y $n_3 = 100$).

Como un indicador de la robustez de los métodos para estimar el coeficiente de regresión β=(β_0, β_1) en presencia de datos atípicos, se usa el Error Cuadrático Medio (MSE), el coeficiente de determinación (R^2) y el estadístico muestral χ^2_2 dado por:

$$\left(\hat{\beta} - \beta\right)' \frac{1}{\sigma^2} \left(X'X\right) \left(\hat{\beta} - \beta\right) \sim \chi^2_2 \tag{0-1}$$

Si el método es sensible a datos atípicos, entonces se espera que el estadístico muestral χ^2_2 se aleje de su valor esperado que es 2. Del mismo modo se espera que el MSE sea mayor en los métodos más robustos y consecuentemente el R^2 sea menor.
Los resultados encontrados vía simulación muestran mediante el análisis del MSE, el R^2 y el χ^2_2 que la regresión de Gini tiene un mayor grado de robustez en comparación con la regresíon OLS al estimar los coeficientes de regresión ante la presencia de datos atípicos, pero su robustez es menor que la de los métodos de estimación robustos LAV, M de Huber y MM.

Palabras clave: Mínimos Cuadrados Ordinarios, Regresión Gini, Modelos de Regresión Robustos, Eficiencia, Robustez, Datos atípicos.

Abstract

The method of Ordinary Least Squares-OLS- is one of the most used to estimate the relationship between a dependent (Y) and independent (X) variable. The regression model is given by the relationship $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. However, OLS is sensitive to outliers, which may not be of interest to the researcher, so it is advisable to use robust methods that overcome the limitations of the OLS method. Gini regression is one method that could have some degree of robustness according to the literature (Olkin y Yitzhaki, 1992) due to the mathematical form as it is raised. In this paper compares Gini regression (using the non-parametric approach of weighted average of slope, instead of the parametric approach) with OLS and other robust regression methods, the type L (LAV, linear combinations of order statistics), the type M (M Huber, based on the concept of maximum likelihood) and the type MM (based on the minimization of an estimator M).

The comparison of the methods is performed via simulation under different scenarios: One of normality of errors (with $\mu = 0$, and $\sigma^2 = 1$) and three of contaminated normal with an atypical data, in which the magnitude is gradually increased (4σ, 8σ, 16σ). Further, the effect of sample size ($n_1 = 10$, $n_2 = 30$ and $n_3 = 100$) is investigated.

As an indicator of the robustness of the methods to estimate the regression coefficient $\beta = (\beta_0, \beta_1)$ in the presence of atypicals, the Mean Square Error (MSE) is used, the determination coefficient R^2 and the sample statistic χ^2_2 given by:

$$\left(\hat{\beta} - \beta\right)' \frac{1}{\sigma^2} \left(X'X\right) \left(\hat{\beta} - \beta\right) \sim \chi^2_2 \tag{0-2}$$

If the method is sensitive to outliers, then it is expected that the sample statistic χ^2_2 away from its expected value that is 2. Similarly it is expected that MSE is greater in the most robust methods and consequently the R^2 is lesser.
The results show that the Gini regression has a higher degree of robustness compared with the OLS regression to estimate the regression coefficients in the presence of outliers, but their robustness is less than robust estimation methods LAV, M Huber and MM.

Keywords: Ordinary Least Square, Gini Regression, Robustness Regression, efficiency, Atypical.

Contenido

Resumen v

1. Introducción 2

2. Marco Teórico 4

2.1. **Análisis de Regresión Lineal** 4

2.1.1. Regresión por Mínimos Cuadrados Ordinarios 4

2.1.2. Estimación por Máxima Verosimilitud 11

2.1.3. Regresión Lineal Múltiple 13

2.1.4. Valores atípicos en Regresión 15

2.1.5. Técnicas de diagnóstico para la adecuación de modelos de regresión . 16

2.2. **Modelos de Regresión Robustos** 23

2.2.1. Estimadores M 24

2.2.2. Estimador S 30

2.2.3. Estimadores MM 32

2.2.4. Estimadores L 33

2.2.5. Propiedades de los Estimadores Robustos 37

3. Regresión Simple de Gini 39

3.1. Enfoque semi-paramétrico 40

3.2. Regresión de Mínimos Cuadrados Ordinarios bajo el enfoque semi-paramétrico 41

3.3. Enfoque semi-paramétrico de la regresión Gini 47

3.3.1. Comparación del esquema de ponderación de OLS y Gini 49

3.4. Enfoque de minimización de Gini 51

4. Metodología y Resultados 53

4.1. Escenarios de simulación para comparar los métodos de Regresión 53

4.2. Pasos para comparar los métodos de Regresión 54

4.3. Resultados de la simulación 56

4.3.1. Tablas de resultados. 63

5. Ejemplo de aplicación a datos reales 68

5.1. Tablas de resultados. 71

6. Conclusiones y Recomendaciones **73**
6.1. Conclusiones 73
6.2. Investigaciones futuras 74

Bibliografía **75**

7. Anexo: Código en R **80**

1 Introducción

El análisis de regresión es una técnica estadística para investigar y modelar la relación entre variables. Son numerosas las aplicaciones de la regresión, y las hay en casi cualquier campo, incluyendo la ingeniería, ciencias físicas y químicas, economía, administración, ciencias biológicas y ciencias sociales. De hecho, el análisis de regresión es la técnica estadística más usada. En casi todas las aplicaciones de regresión, la recta de regresión estimada solo es una apróximación a la verdadera relación funcional entre las variables de interés. Esas relaciones funcionales se basan, con frecuencia, en una teoría física, química o de otra disciplina científica o técnica; esto es, en el conocimeinto del mecanismo básico (Montgomery, Peck y Vining, 2002).
Considere el modelo de regresión lineal:

$$Y = X\beta + \varepsilon \tag{1-1}$$

donde $\mathbf{Y}$ es un vector de observaciones de $n \times 1$, $\mathbf{X}$ es una matriz con las variables independientes o regresoras del modelo de $n \times p$, β es el vector de coeficientes de regresión de $p \times 1$ y ε es el vector de errores aleatorios de $n \times 1$. El objetivo es estimar β. El término del error tiene vector de medias cero y matriz de varianzas y covarianzas $\sigma^2 I$, siendo I la matriz identidad.
Además, se suele suponer que los errores no están correlacionados. Esto quiere decir que el valor de un error no depende del valor de cualquier otro error. Las variables regresoras X están controladas por el analista de datos, y se puede medir con error despreciable, mientras que la variable respuesta Y es aleatoria. Con lo que hay una distribución de probabilidades de Y para cada valor posible de X.
Así, la media de Y es una función lineal de X, aunque la varianza de Y no depende del valor de X. Además, ya que el componente aleatorio del error no está correlacionado, las respuestas tampoco lo están.
A los parámetros β se les suele llamar coeficientes de regresión. Estos tienen una interpretación simple y frecuentemente útil. La pendiente β_j, para $j \geq 1$, es el cambio de la media de la distribución de Y producida por un cambio unitario en X. Si el intervalo de los datos incluye a $X = 0$, entonces la ordenada al origen, β_0, es la media de la distribución de la respuesta Y cuando $X = 0$. Si no incluye al cero, β_0 no tiene interpretación práctica.
Cuando hay observaciones atípicas, los coeficientes de regresión son influenciados por éstas, y muchas veces el investigador no está interesado en ellas porque de antemano sabe que estos no son propios del experimento o el estudio que conduce. Para superar estas limitaciones,

la literatura ofrece distintas formas de trabajar este tipo de situaciones (Montgomery, Peck y Vining, 2002). En el presente trabajo se mencionarán algunas de ellas, una es realizando pruebas de detección de datos atípicos (DFFITS, DFBETAS, R de Student, etc) para excluirlos del modelo de regresión, lo cual puede ser tedioso si la cantidad de valores atípicos es relativamente grande. La otra es usar estimadores robustos clásicos del tipo M (máxima verosimilitud), L (estadísticos de orden), R (se basan en rango o jerarquías) y combinaciones de ellos (MM), los cuales están construidos de tal manera que automáticamente le dan poca importancia a los valores atípicos mediante una función ponderadora. Otro método de regresión que promete superar las limitaciones de sensibilidad de OLS fue desarrollado por Gini en 1912 (Montanari y Monari 2008) usando una medida de dispersión alternativa denominada Diferencia Media de Gini (GMD), la cual en la actualidad no es muy conocida o utilizada y por ende no está implementada en los paquetes estadísticos como el R (R Core Team 2013). En este trabajo se pretende investigar cuán robusto son los estimadores de los coeficientes de regresión mediante la metodología de Gini, la cual tiene un enfoque paramétrico y uno no paramétrico. En este trabajo de tesis se trabaja el enfoque no paramétrico.

Para estudiar la robustez de la regresión Gini respecto a OLS se realizan simulaciones Montecarlo bajo diferentes condiciones: variando la distribución de los errores (normalidad y normalidad contaminada), aumentando progresivamente la magnitud del valor atípico y el tamaño muestral de las observaciones. Además, se investiga la robustez de los métodos clásicos M de Huber (un tipo de estimador M propuesto por Huber, 1973), MM (un tipo de estimador M propuesto por Yohai, 1984) y LAV (un tipo de estimador L conocido como estimador de Mínimo Valor Absoluto-LAV).

Este trabajo está distribuido de la siguiente manera.
Después de esta introducción se presenta el capítulo 2, el cual contiene una revisión de metodologías de regresión. Las primeras secciones del capítulo 2 son dedicadas a la Regresión OLS, luego se presentan los métodos de estimación robustos como MM, M de Huber y LAV. En el capítulo 3 se presenta la regresión de Gini, la cual se divide en dos secciones: estimación de los coeficientes mediante promedios ponderados de pendientes (enfoque no paramétrico) y estimación por minimización (enfoque paramétrico). En el capítulo 4 se presenta la metodología de simulación (la secuencia de pasos que se siguieron para comparar los métodos de regresión mencionados) y los resultados obtenidos. En el capítulo 5 se presenta un ejemplo aplicado a datos reales. En el capítulo 6 se presentan las conclusiones y recomendaciones para futuras investigaciones. Finalmente se presenta la Bibliografía y el Anexo con el código en el software R utilizado para la simulación Montecarlo.

2 Marco Teórico

2.1. Análisis de Regresión Lineal

Cuando se quiere ajustar el modelo de regresión lineal a un conjunto de datos hay varios métodos que se podrían usar para estimar los coeficientes de la regresión. El más común es el método de Mínimos Cuadrados Ordinarios (OLS). Este método, sin embargo, tiene debilidades y consecuentemente algunas veces se prefiere usar métodos de regresión alternativos (Birkes y Dodge, 1993). Una debilidad del método OLS es que permite que los datos atípicos tengan mucha influencia sobre la estimación de los coeficientes de regresión. Los métodos de regresión alternativa restringen la influencia de los datos atípicos. Entre los métodos de regresión alternativa más comunes están el Mínimo Valor Absoluto (LAV), la Regresión M, Regresión MM y Regresión S.
OLS se desempeña mejor si los errores tienen una distribución normal, de lo contrario (si hay presencia de datos atípicos) los estimadores de OLS son menos eficientes (en términos de la varianza de los estimadores de los coeficientes) que los obtenidos por los métodos de regresión robusta.

2.1.1. Regresión por Mínimos Cuadrados Ordinarios

La regresión de Mínimos Cuadrados Ordinarios -OLS- es el método más usado para estimar la relación entre variables dependientes (y) e independientes (x). Su principal objetivo consiste en ajustar una línea tan cercana como sea posible a los puntos, minimizando los errores (ε). El método OLS fue descubierto independientemente por Carl Friedrich Gauss en Alemania en el año 1795 y por Adrien Marie Legendre en Francia en 1805. Inicialmente las aplicaciones del método OLS fueron para datos astronómicos y geográficos (Birkes y Dodge, 1993).
Suponga que se realiza un experimento que intenta encontrar la relación entre dos variables, el cual se representa en el siguiente diagrama de dispersión. Figura **2-1**

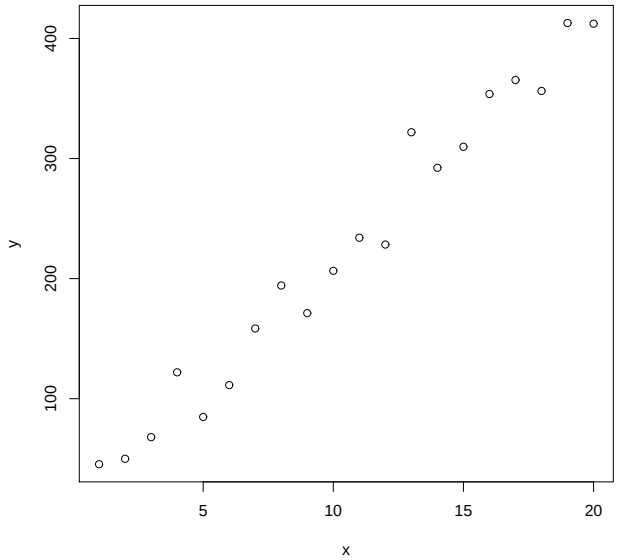


Figura 2-1: Diagrama de dispersión

El diagrama de dispersión parece indicar que hay una relación lineal entre las dos variables. Se podría representar esa relación lineal mediante una ecuación de línea recta (Montgomery, Peck y Vining, 2002)

$$y = \beta_0 + \beta_1 x. \tag{2-1}$$

Donde β_0 es la ordenada al origen y β_1 es la pendiente. Ahora bien, los datos no caen exactamente sobre una recta, por lo que se debe modificar la ecuación (2-1) para tomar en cuenta esto. Sea la diferencia entre el valor observado y y el de la línea $\beta_0 + \beta_1 x$ un error ε, el cual es un error estadístico, esto es, una variable aleatoria que explica por qué el modelo no ajusta exactamente los datos. Así, un modelo más plausible para los datos es

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2-2}$$

La ecuación (2-2) se llama modelo de regresión lineal, donde x es la variable explicativa y y es la variable respuesta. Ahora, si x está fija, la componente aleatoria ε determina las propiedades de y. Suponga que el promedio de ε es cero, $E(\varepsilon) = 0$, y que la varianza de ε es σ^2, es decir, $Var(\varepsilon) = \sigma^2$. Entonces, la respuesta media en cualquier valor de x es

$$E(y|x) = \mu_{y|x} = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x \tag{2-3}$$

La varianza de y para cualquier valor de x es

$$Var(y|x) = \sigma^2_{y|x} = Var(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2 \tag{2-4}$$

Así, el verdadero modelo de regresión $\mu_{y|x} = E(\beta_0 + \beta_1 x)$ es una línea recta de valores promedios, esto es, la altura de la línea de regresión en cualquier valor de x no es más que el valor esperado de y para esa x. Se puede interpretar la pendiente β_1 como el cambio de la media de y para un cambio unitario de x. Además, la variabilidad de y en algún valor particular de x queda determinada por la varianza de ε, σ^2. Esto implica que hay una distribución de valores de y en cada x, y que la varianza de esta distribución es igual en cada x. La siguiente figura **2-2** ilustra esta situación.

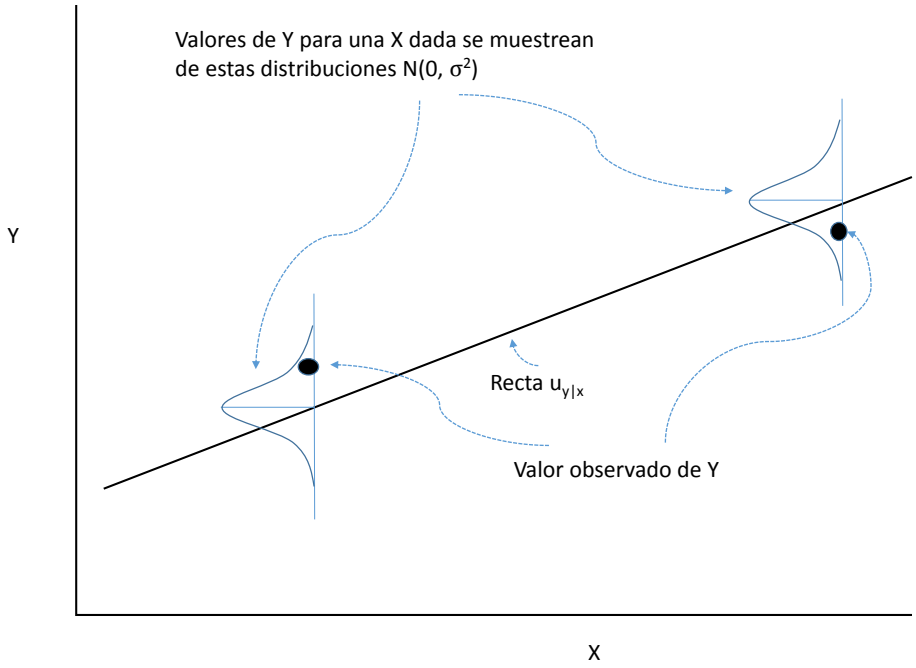


Figura 2-2: Cómo se generan las observaciones en la regresión lineal.

Obsérvese que se ha usado una distribución normal para describir la variación aleatoria de ε. Como y es la suma de la constante $\mu_{y|x} = E(\beta_0 + \beta_1 x)$ más ε, entonces y será una variable aleatoria que tradicionalmente se asume normal. Cuando σ^2 es pequeña los valores observados en y serán cercanos a la recta, y cuando σ^2 es grande, se pueden desviar bastante de la línea.

La idea básica es estimar la recta de regresión que mejor se ajuste a los datos. Para juzgar cuán bien se ajusta la recta de regresión $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ a los datos se puede hacer en términos del tamaño de los residuales $\varepsilon_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x)$ (Birkes y Dodge, 1993).

El residual ε_i es la distancia que hay entre la recta de regresión ajustada y los datos (x_i, y_i). Los supuestos estadísticos sobre el término de error ε_i son:

1. La media de ε es cero, es decir, $E(\varepsilon) = 0$.
2. La varianza de ε es una constante desconocida σ^2.
3. Son incorrelacionados, es decir, ε_i y ε_{i+1} son independientes.

Los residuales tienen un papel importante para investigar la adecuación del modelo de regresión ajustado, y para detectar diferencias respecto a las hipótesis básicas.
Se desea seleccionar β_0 y β_1 de tal manera que los residuales sean pequeños.
En OLS, el tamaño total de los residuales se mide por $\sum \varepsilon_i^2$. Los estimadores de β_0 y β_1 son aquellos valores $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimizan a $\sum \varepsilon_i^2$. Es decir, minimizar la siguiente función de dispersión

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2 \tag{2-5}$$

A continuación se presenta la expresión matemática de los estimadores de OLS.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2-6}$$

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} y_i x_i - \frac{\left(\sum\limits_{i=1}^{n} y_i\right)\left(\sum\limits_{i=1}^{n} x_i\right)}{n}}{\sum\limits_{i=1}^{n} x_i^2 - \frac{\left(\sum\limits_{i=1}^{n} x_i\right)^2}{n}} \tag{2-7}$$

en donde

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad y \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

son los promedios de y_i y x_i, respectivamente. Por consiguiente, $\hat{\beta}_0$ y $\hat{\beta}_1$ en las ecuaciones (2-6) y (2-7) son los estimadores por mínimos cuadrados de la ordenada al origen y la pendiente, respectivamente. El modelo ajustado de regresión lineal simple es, entonces,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2-8}$$

La ecuación 2-8 produce un estimador puntual, de la media de y para una determinada x. Como el denominador de la ecuación (2-7) es la suma corregida de cuadrados de las x_i y el numerador es la suma corregida de los productos cruzados de x_i y y_i, estas ecuaciones se pueden escribir en forma más compacta como sigue:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2-9}$$

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = \sum_{i=1}^{n} y_i (x_i - \bar{x}) \tag{2-10}$$

Entonces, una forma cómoda de escribir la ecuación (2-7) es

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{2-11}$$

A los parámetros β_0 y β_1 se les suele llamar **c**oeficientes de regresión. Estos tienen una interpretación simple y frecuentemente útil. La pendiente β_j es el cambio de la media de la distribución de y producido por un cambio unitario en x. Si el intervalo de los datos incluye a $x = 0$, entonces la ordenada al origen, β_0, es la media de la distribución de la respuesta y cuando $x = 0$. Si no incluye al cero, β_0 no tiene interpretación práctica.

Interpretación alternativa de los coeficientes de regresión.

Se puede demostrar que (2-10) se puede reescribir como

$$S_{xy} = \sum_{i=1}^{n} y_i (x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{2-12}$$

Entonces

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2-13}$$

La fórmula para $\hat{\beta}_1$ puede ser reescrita como

$$\hat{\beta}_1 = \sum_{i=1}^{n} w_i \left(\frac{y_i - \bar{y}}{x_i - \bar{x}}\right) \quad , \quad w_i = \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2-14}$$

Note que $\left(\frac{y_i-\bar{y}}{x_i-\bar{x}}\right)$ es la pendiente entre el punto central $(\bar{x}, \bar{y})$ y el punto de los datos (x_i, y_i). Entonces la pendiente $\hat{\beta}_1$ de la regresión estimada es un tipo de promedio de esas pendientes. Más precisamente, es un promedio ponderado. Note que los ponderadores w_i son no negativos y su suma es igual a uno, es decir, $\sum_{i=1}^{n} w_i = 1$. (Un promedio ordinario es un caso especial de un promedio ponderado en el que todos los pesos son iguales a $1/n$). Cada dato se pondera proporcionalmente al cuadrado de su distancia x desde el centro. Esto implica que los datos más alejados del centro de los datos tienen una mayor influencia en la estimación de la pendiente de la línea de regresión (Birkes y Dodge, 1993).

Algunas propiedades de los estimadores por Mínimos Cuadrados Ordinarios

Los estimadores obtenidos por OLS $\hat{\beta}_0$ y $\hat{\beta}_1$ tienen algunas propiedades importantes (Montgomery, Peck y Vining, 2002). Primero obsérvese que, según las ecuaciones (2-6) y (2-7) , $\hat{\beta}_0$, $\hat{\beta}_1$ son **combinaciones lineales** de las observaciones y_i. Por ejemplo

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i \tag{2-15}$$

donde $c_i = \frac{(x_i-\bar{x})}{S_{xx}}$, para $i = 1, 2, 3, ..., n$. Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ por OLS son **estimadores insesgados** de los parámetros β_0 y β_1 del modelo puesto que $E(\hat{\beta}_0) = \beta_0$ y $E(\hat{\beta}_1) = \beta_1$.
La varianza de ambos estimadores se define como sigue. Se ha supuesto que $Var(y_i) = \sigma^2$; en consecuencia,

$$Var(\hat{\beta}_1) = \sigma^2 \left(\sum_{i=1}^{n} c_i\right) = \frac{\sigma^2 \sum_{i=1}^{n} (x_i - \bar{x})}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}^2} \tag{2-16}$$

$$Var(\hat{\beta}_0) = Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \tag{2-17}$$

Puesto que la varianza de $\bar{y}$ no es más que $Var(\bar{y}) = \sigma^2/n$, y se puede demostrar que la covarianza entre $\bar{y}$ y $\hat{\beta}_1$ es cero. Otro resultado importante acerca de la calidad de los estimadores por OLS, $\hat{\beta}_0$ y $\hat{\beta}_1$, es el teorema de **teorema de Gauss Markov**, el cual establece que para el modelo de regresión (2-2) con la hipótesis $E(\varepsilon) = 0$ y $Var(\varepsilon) = \sigma^2$ y con errores no correlacionados, los estimadores por OLS son insesgados y tienen varianza mínima en comparación con todos los demás estimadores insesgados que sean combinaciones lineales de las y_i. Con frecuencia se dice que los estimadores por OLS son los **estimadores lineales insesgados óptimos**, donde "óptimos" implica que son de mínima varianza.

Algunas propiedades útiles del ajuste por OLS

1. La suma de los residuales en cualquier modelo de regresión que contenga una ordenada al origen β_0 siempre es igual a cero, esto es,

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}\varepsilon_i^2 = 0 \tag{2-18}$$

2. La suma de los valores observados y_i es igual a la suma de los valores ajustados $\hat{y}_i$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\hat{y}_i \tag{2-19}$$

3. La recta de regresión de OLS siempre pasa por el centroide de los datos, que es el punto $(\bar{x}, \bar{y})$.

4. La suma de los residuales, ponderados por el valor correspondiente de la variable explicativa, siempre es igual a cero:

$$\sum_{i=1}^{n} x_i\varepsilon_i = 0 \tag{2-20}$$

5. La suma de los residuales, ponderados por el valor ajustado correspondiente, siempre es igual a cero:

$$\sum_{i=1}^{n} \hat{y}_i\varepsilon_i = 0 \tag{2-21}$$

Estimación de σ^2

Además de estimar los parámetros β_0 y β_1, se requiere un estimador de σ^2 para probar la hipótesis y formar estimadores de intervalo pertinentes del modelo de regresión. En el caso ideal este estimador no debería depender de la adecuación del modelo ajustado.
La suma de cuadrados de residuales tiene $n-2$ grados de libertad, porque dos grados de libertad se asocian con los estimados $\hat{\beta}_0$ y $\hat{\beta}_1$ que se usan para obtener $\hat{y}_i$. Por lo que un estimador insesgado para σ^2 se define como sigue:

$$\hat{\sigma}^2 = \frac{SS_{\text{Res}}}{n-2} = MS_{\text{Res}} \tag{2-22}$$

La cantidad MS_{Res} se llama cuadrado medio residual. La raíz cuadrada de $\hat{\sigma}^2$ se llama, a veces, el error estándar de la regresión y tiene las mismas unidades que la variable respuesta y.
Ya que $\hat{\sigma}^2$ depende de la suma de cuadrados de los residuales, cualquier violación de la hipótesis sobre el error del modelo, o cualquier especificación equivocada de la forma del modelo pueden dañar gravemente la utilidad de $\hat{\sigma}^2$ como estimador de σ^2. Como σ^2 se calcula con los residuales del modelo de regresión, se dice que es un estimador de σ^2 dependiente del modelo.

Coeficiente de Determinación

Sea

$$SS_{Res} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2-23}$$

y

$$SS_T = \sum_{i=1}^{n} y_i^2 - n\bar{y} = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \tag{2-24}$$

luego

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{\text{Res}}}{SS_T} \tag{2-25}$$

R^2 es llamado coeficiente de determinación. Como SS_T es una medida de la variabilidad de y sin considerar el efecto de la variable explicativa x y SS_{Res} es una medida de la variabilidad de y que queda después de haber tenido en consideración a x, R^2 es nombrado, con frecuencia, la proporción de la variabilidad explicada por la variable explicativa x. Ya que $0 \leq SS_{\text{Res}} \leq SS_T$, entonces $0 \leq R^2 \leq 1$. Los valores de R^2 cercanos a 1 implican que la mayor parte de la variabilidad de y es explicada por el modelo de regresión.

$$E(R^2) = \frac{\hat{\beta}_1 S_{xx}}{\hat{\beta}_1 S_{xx} + \sigma^2} \tag{2-26}$$

2.1.2. Estimación por Máxima Verosimilitud

El método de OLS se puede aplicar para estimar los parámetros en un modelo de regresión lineal, independientemente de la forma de la distribución de los errores ε. Con el método

OLS se obtienen los mejores estimadores lineales insesgados de β_0 y β_1. Otros procedimientos estadísticos, como prueba de hipótesis y construcción de intervalos de confianza, suponen que los errores se distribuyen normalmente. Si se conoce la forma de la distribución de los errores, un método alternativo para estimar parámetros es el **método de máxima verosimilitud o máxima posibilidad**. Se tienen los datos $(y_i,\ x_i)$, $i = 1, 2, \ldots, n$. Si se supone que los errores en el modelo de regresión lineal son $NID(0, \sigma^2)$, las observaciones y_i en esa muestra son variables aleatorias normal e independiente distribuidas (NID) con promedio $\beta_0 + \beta_1 x_i$ y varianza σ^2. La función de verosimilitud o de posibilidad se determina con la distribución conjunta de las observaciones. Si se considera esta distribución conjunta con las observaciones dadas y los parámetros β_0, β_1 y σ^2 son constantes desconocidas, se tiene la función de verosimilitud. Para el modelo de regresión lineal simple con errores normales, la **función de verosimilitud** es

$$\begin{aligned} L(y_i, x_i, \beta_0, \beta_1, \sigma) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[-\tfrac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right] \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\tfrac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2.\right] \end{aligned} \tag{2-27}$$

Los estimadores de máxima verosimilitud son los valores de los parámetros, por ejemplo $\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2$, que maximizan a L o lo que es lo mismo, a $\ln L$. Así,

$$\ln L(y_i, x_i, \beta_0, \beta_1, \sigma) = -\left(\frac{n}{2}\right) \ln 2\pi - \left(\frac{n}{2}\right) ln\sigma^2 - \left(\frac{1}{2\sigma^2}\right) \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2-28}$$

A continuación se presenta la expresión matemática de los estimadores de máxima verosimilitud.

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x} \tag{2-29}$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2-30}$$

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2}{n} \tag{2-31}$$

Obsérvese que los estimadores de máxima verosimilitud de la ordenada al origen y la pendiente, $\tilde{\beta}_0, \tilde{\beta}_1$, son idénticos a los obtenidos con el método OLS. También, $\tilde{\sigma}^2$ es un estimador

sesgado de σ^2. El estimador sesgado se relaciona con el estimador insesgado $\hat{\sigma}^2$ mediante $\tilde{\sigma}^2 = ((n-1)/n)\hat{\sigma}^2$. El sesgo es pequeño cuando n es moderadamente grande, por lo general se usa el estimador insesgado $\hat{\sigma}^2$.
Casi siempre, los estimadores de máxima verosimilitud tienen mejores propiedades estadísticas que los estimadores por OLS (Montgomery, Peck y Vining, 2002). Los de máxima verosimilitud son insesgados (incluyendo a $\tilde{\sigma}^2$, que es asintóticamente insesgado o insesgado cuando n se hace grande) y tienen varianza mínima comparados con todos los demás estimadores insesgados. También son estimadores consistentes (consistencia es una propiedad de una muestra grande que indica que los estimadores difieren del valor verdadero del parámetro una cantidad muy pequeña), cuando n se hace grande. También, están basados en un conjunto de estadísticos suficientes muestral (esto implica que contienen toda la información de la muestra original, de tamaño n). Por otro lado, la estimación por máxima verosimilitud requiere hipótesis estadísticas más estrictas que las de los estimadores de mínimos cuadrados. Los estimadores de OLS requieren sólo hipótesis de segundo momento (que son hipótesis sobre el valor esperado, las varianzas y las covarianzas entre los errores aleatorios). Los estimadores de máxima verosimilitud requieren una hipótesis de distribución total, en este caso, son errores aleatorios que siguen una distribución normal con los mismos segundos momentos con los que se requirieron en las estimaciones por OLS.

2.1.3. Regresión Lineal Múltiple

Es más cómodo manejar modelos de regresión múltiple cuando se expresan en notación matricial. Esto permite presentar en forma muy compacta al modelo, los datos y los resultados (Staudte y Sheather, 1990).
En el caso de la regresión lineal múltiple, con k variables explicativas y $p = k+1$ parámetros, los cuales se estiman así

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & x_{11} & . & . & . & x_{1k} \\ 1 & x_{21} & . & . & . & x_{2k} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n1} & . & . & . & x_{nk} \end{bmatrix}; \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix} \ y \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} \tag{2-32}$$

Es decir,

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon} \tag{2-33}$$

donde

- $Y = [y_1, ..., y_n]'$ Es un vector de observaciones de $n \times 1$.

- $X = [x_{ij}]$ Es la matriz de las variables regresoras de $p \times n$.
- $\beta = [\beta_0, ..., \beta_p]'$ Es el vector de $p \times 1$ parámetros desconocidos.
- $\varepsilon = [\varepsilon_1, ..., \varepsilon_n]'$ Es el vector de errores estadísticos de $n \times 1$.
- $E(\varepsilon) = 0$ y $Cov(\varepsilon) = \sigma^2 I_n$.

La constante σ^2 es también un parámetro desconocido. El método de OLS selecciona un estimador $\hat{\beta}$ que minimiza la longitud de $\|\varepsilon\| = [\varepsilon'\varepsilon]^{1/2}$. El vector de residuales definido por

$$\varepsilon = Y - X\hat{\beta} \tag{2-34}$$

Derivando $\|\varepsilon\|^2$ e igualando a cero conduce a las ecuaciones normales

$$X'X\hat{\beta} = X'Y \tag{2-35}$$

y el estimador de OLS es

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2-36}$$

Bajo el supuesto de que $E(\varepsilon) = 0$ y $Cov(\varepsilon) = \sigma^2 I_n$ se puede verificar que $\hat{\beta}$ es insesgado y tiene matriz de covarianza proporcional a $(X'X)^{-1}$ (Staudte y Sheather, 1990):

$E(\hat{\beta}) = \beta$ y $Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$

Un estimador de σ^2 es el error cuadrático medio

$$s^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-p} = \frac{Y'Y - \hat{\beta}'X'Y}{n-p} \tag{2-37}$$

donde

$$\varepsilon \sim N_n(0, \sigma^2 I_n) \tag{2-38}$$

entonces

1. $\hat{\beta}$ es también el estimador de Máxima verosimilitud.
2. $\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$.

3. $(n-p)s^2/\sigma^2 \sim \chi^2_{n-p}$.

4. El error estándar de $\hat{\beta}_i$ es estimado por $s[(X'X)^{-1}]_{ii}^{1/2}$.

Intervalo de confianza para $\hat{\beta}$

Bajo el supueto que $\varepsilon \sim$ N $(0,\, \sigma^2 I_n)$, es fácil construir una región (elipsoide) de confianza para $\hat{\beta}$:

$$\left(\hat{\beta}-\beta\right)' \frac{1}{\sigma^2}\left(X'X\right)\left(\hat{\beta}-\beta\right) \sim \chi^2_2 \tag{2-39}$$

En términos de su matriz de covarianzas $s^2(X'X)^{-1}$, donde $s^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-p}$:

$$\left(\hat{\beta}-\beta\right)' \frac{1}{ps^2}\left(X'X\right)\left(\hat{\beta}-\beta\right) \sim F_{p,n-p} \tag{2-40}$$

Un intervalo de confianza simultáneo al $100(1-\alpha)\,\%$ para $\hat{\beta}_j$ está dado por

$$\hat{\beta}_i \pm \sqrt{\mathrm{var}(\hat{\beta}_i)}\sqrt{pF_{p,n-p}(\alpha)}, \quad i = 0, 1, ..., p \tag{2-41}$$

donde $\mathrm{var}(\hat{\beta}_i)$ son los elementos de la diagonal de $s^2(X'X)^{-1}$ correspondiente a $\hat{\beta}_i$ (Johnson y Wichern, 1992).

2.1.4. Valores atípicos en Regresión

Hay varias clases de valores atípicos que se presentan en el contexto de modelos de regresión. Una clasificación frecuente es la siguiente (Montgomery, Peck y Vining, 2002):

1. **Valor atípico de regresión**: Es un punto que se desvía de la recta de regresión que se determina con las $n-1$ observaciones restantes.

2. **Valor atípico residual**: Es un punto que tiene un residual estandarizado o estudentizado grande (mayor a 2) cuando el modelo ajustado usa las n observaciones. Nótese que un punto puede ser atípico residual sin necesidad de ser atípico de regresión, y viceversa.

3. **Valor atípico en el espacio de las X**: Es un punto alejado en el espacio de las X, que también puede ser atípico de regresión o atípico residual. Hay técnicas de estimación robustas que pueden ser afectadas por estos puntos atípicos.

4. **Valor atípico en el espacio de las** Y: Es un punto alejado en el espacio de las Y, que también puede ser atípico de regresión o atípico residual. El efecto que tiene sobre el modelo de regresión depende de su coordenada X, y de la disposición de las observaciones en la muestra. Cuando la coordenada X del punto está en los valores mínimos o máximos, sin ser atípicos en el espacio de las X necesariamente, se le conoce como atípicos influyentes.

5. **Valor atípico en el espacio** X **y** Y: Es un punto alejado en el espacio de las X y Y al mismo tiempo, que también puede ser atípico de regresión o atípico residual, pero que podría no tener efecto sobre el modelo de regresión. Se les conoce como atípico de apalancamiento.

Existen varios procedimientos de detección de valores atípicos como los Residuales estandarizados, Residuales estudentizados, la Distancia de Cook, los DFFITS y los DFBETAS, que se emplean según la clase de valor atípico (Montgomery, Peck y Vining, 2002). A continuación se realiza una breve descripción de algunas de ellas.

2.1.5. Técnicas de diagnóstico para la adecuación de modelos de regresión

Después de la estimación de los parámetros de un delo de regresión, es muy importante realizar su validación, para lo cual existe en la literatura un conjunto de pruebas que permiten realizar la validación y la comprobación de la adecuación del modelo estimado. En esta sección se presentará un resumen de algunas técnicas de utilidad para diagnosticar la adecuación del modelo de regresión estimado.

Análisis de residuales

El análisis de los residuales se realiza básicamente para diagnosticar posibles violaciones de los supuestos básicos. La relación entre la variable respuesta Y y las variables explicativas es lineal, el término del error ε tiene media cero, varianza σ^2 constante, no están correlacionados, y tienen distribución normal.
El no cumplimiento de los supuestos mencionados puede producir un modelo inestable, en el sentido que una muestra distinta podría conducir a un modelo totalmente diferente, y obtener conclusiones opuestas.
Los residuales de un modelo de regresión se definen como:

$$\varepsilon_i = Y_i - \hat{Y}_i \ , \qquad i = 1, 2, ..., n$$

Siendo Y_i una observación, y $\hat{Y}_i$ su valor ajustado correspondiente.

La varianza promedio aproximada de los residuales se estima con:

$$\frac{\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon}_i)}{n-p} = \frac{\sum_{i=1}^{n}(\varepsilon_i^2)}{n-p} = \frac{SS_{\text{Res}}}{n-p} = MS_{\text{Res}} \tag{2-42}$$

Sin embargo, los residuales no son independientes, ya que los n residuales sólo tienen $n-p$ grados de libertad asociados a ellos. Esta no independencia de los residuales tiene poco efecto en su aplicación para comprobar la adecuación del modelo, siempre y cuando n no sea pequeña en relación con la cantidad de parámetros p.

Métodos para escalar residuales

A continuación se presentarán algunos métodos frecuentes para escalar residuales. Los residuales escalados son útiles para determinar observaciones que sean atípicas o valores extremos, esto es, observaciones que en algún aspecto estén separadas del resto de los datos.

1. Residuales estandarizados:

   Como se muestra en la ecuación 2-42, la varianza aproximada de un residual puede ser estimada con MS_{Res}, el cuadrado medio de los residuales un escalamiento lógico de los residuales sería el de los residuales estandarizados

   $$d_i = \frac{\varepsilon_i}{\sqrt{MS_{\text{Res}}}} \quad , \quad i = 1, 2, ..., n$$

   Los residuales estandarizados tienen media cero y varianza aproximadamente de uno, en consecuencia, un residual estandarizado grande (por ejemplo $d_i > 3$) indica que se trata de un valor atípico potencial.

2. Residuales estudentizados:

   Si se usa el MS_{Res} como la varianza del i-ésimo residual ε_i sólo se tendrá una aproximación. Se puede mejorar el escalamiento de residuales dividiendo ε_i entre la desviación estándar exacta del i-ésimo residual.

   Las violaciones de los supuestos del modelo están, con más probabilidad, en los puntos atípicos, y pueden ser difíciles de detectar por inspección de los residuales ordinarios ε_i (o de los residuales estandarizados d_i), porque en general, sus residuales serán menores. Entonces, un procedimiento lógico es examinar los residuales estudentizados en lugar de ε_i o d_i.

   $$r_i = \frac{\varepsilon_i}{\sqrt{MS_{\text{Res}}(1-h_{ii})}} \quad , \quad i = 1, 2, ..., n$$

donde h_{ii} corresponde a los elementos de la diagonal de la matriz sombrero.

Cuando la forma del modelo es correcta estos residuales estudentizados tiene varianza constante $Var(r_i) = 1$, independiente del lugar de X_i. En muchos casos, la varianza de los residuales se estabiliza, en especial para grandes tamaños de muestra. En esos casos podrá haber poca diferencia entre los residuales estandarizados y los estudentizados.

Así, los residuales estandarizados y los estudentizados aportan con frecuencia información equivalente. Sin embargo, ya que cualquier punto con un residual grande y también una h_{ii} grande tiene una influencia potencial muy grande sobre el ajuste por mínimos cuadrados, se recomienda por lo general examinar los residuales estudentizados. En consecuencia, un residual estudentizado grande (por ejemplo $r_i > 2$) indica que se trata de un valor atípico.

3. Residuales *PRESS*:

   Los residuales estandarizados y los estudentizados son efectivos para detectar valores atípicos. Otro método para hacer que los residuales sean útiles en la determinación de valores atípicos consiste en examinar la cantidad que se calcula partiendo de $Y_i - \hat{Y}_i$, siendo $\hat{Y}_i$ el valor ajustado de la i-ésima respuesta, basado en todas las observaciones excepto esa i-ésima. La lógica de este método es que si la i-ésima observación Y_i realmente es atípica, el modelo de regresión basado en todas las observaciones estará demasiado afectado por esta observación. Esto puede conducir a un valor ajustado $\hat{Y}_i$ muy parecido al valor observado Y_i y en consecuencia el residual ordinario ε_i será pequeño. Por lo anterior, será difícil detectar el valor atípico. Sin embargo, si se elimina la i-ésima observación, entonces $\hat{Y}_i$ no puede estar influido por esa observación, así que el residual que resulte probablemente indique la presencia del valor atípico.

   Si se elimina la i-ésima observación, se ajusta el modelo de regresión a las $n-1$ observaciones restantes, y se calcula el valor predicho de Y_i correspondiente a la observación omitida, el error de predicción correspondiente es

$$\varepsilon_i = Y_i - \hat{Y}_i$$

   Este cálculo del error de predicción se repite para cada observación $i = 1, 2, ..., n$. A esos errores de predicción se les llama residuales *PRESS*

   El i-ésimo residual *PRESS* es

$$\varepsilon_{(i)} = \frac{\varepsilon_i}{1 - h_{ii}} \quad , \quad i = 1, 2, ..., n$$

   En la ecuación anterior el residual *PRESS* no es más que el residual ponderado por los elementos de la diagonal de la matriz sombrero h_{ii}. Los residuales asociados con

puntos para los que h_{ii} es grande tendrán residuales *PRESS* grandes. Esos puntos, en general, serán puntos de gran influencia. En general, una gran diferencia entre el residual y el residual *PRESS* indica un punto donde el modelo se ajusta bien a los datos, pero un modelo formado sin ese punto hace malas predicciones. La varianza del i-ésimo residual *PRESS* es

$$Var[\varepsilon_{(i)}] = Var\left[\frac{\varepsilon_i}{1-h_{ii}}\right] = \frac{1}{1-h_{ii}}[\sigma^2(1-h_{ii})] = \frac{\sigma^2}{1-h_{ii}}$$

por lo que un residual *PRESS* estandarizado es

$$\frac{\varepsilon_{(i)}}{\sqrt{Var[1-h_{ii}]}} = \frac{\varepsilon/(1-h_{ii})}{\sqrt{\sigma^2[1-h_{ii}]}} = \frac{\varepsilon}{\sqrt{\sigma^2[1-h_{ii}]}}$$

4. R de Student:

   El residual estudentizado r_i que se describió arriba se considera con frecuencia como diagnóstico para valores atípicos. Se acostumbra usar MS_{Res} como estimado de σ^2 al calcular r_i. A esto se le llama escalamiento del residual, porque MS_{Res} es un estimado de σ^2, que se obtiene al ajustar el modelo a las n observaciones. Otro método sería usar un estimado de σ^2 basado en un conjunto de datos con la i-ésima observación eliminada.

$$S^2_{(i)} = \frac{(n-p)MS_{Res} - \varepsilon_i/(1-h_{ii})}{n-p-1} \tag{2-43}$$

   El estimador de σ^2 en la ecuación 2-43 se usa en lugar de MS_{Res} para obtener un residual estudentizado, que se suele llamar R de Student o R-Student, y se define como:

$$t_i = \frac{\varepsilon_i}{\sqrt{S^2_{(i)}(1-h_{ii})}} \quad , \quad t = 1, 2, ..., n$$

   En muchos casos t_i será poco distinto del residual estudentizado r_i. Sin embargo, si la i-ésima observación es influyente, entonces $S^2_{(i)}$ puede diferir bastante de MS_{Res}, y entonces el estadístico R de Student será más sensible a este punto.

5. Datos de Balanceo:

   El lugar de los puntos en el espacio de X tiene importancia potencial en la determinación de las propiedades del modelo de regresión. En particular, los puntos alejados o atípicos tienen un impacto desproporcionado sobre las estimaciones de los parámetros, el error estándar, valores predichos y estadísticas de resumen del modelo. La matriz sombrero se define como

   $$H = X(X'X)^{-1}X'$$

   La matriz sombrero desempeña un papel importante en la identificación de observaciones influyentes. H determina las varianzas y covarianzas de $\hat{Y}_i$ y de ε, porque $Var(\hat{Y}_i) = \sigma^2 H$ y $Var(\varepsilon) = \sigma^2(I - H)$. Los elementos h_{ii} de la matriz H pueden ser vistos como la cantidad de datos de balanceo o palanqueo ejercido por la i-ésima observación Y_i sobre el i-ésimo valor ajustado $\hat{Y}_i$.

   Con frecuencia, la atención se dirige hacia los elementos h_{ii} de la diagonal de la matriz sombrero H, los cuales se pueden expresar como:

   $$h_{ii} = X_i'(X'X)^{-1}X_i$$

   Siendo X_i' la i-ésimo fila de la matriz X. La diagonal de la matriz sombrero es una medida estandarizada de la distancia de la i-ésima observación al centro (o al centroide) del espacio de X. Así, los elementos grandes en la diagonal indican observaciones que son potencialmente influyentes, por estar lejos, del resto de la muestra en el espacio de X. El tamaño promedio de los elementos de la diagonal es $\bar{h} = \frac{p}{n}$ (porque $\sum_{i=1}^{n} h_{ii} = rango(H) = rango(X) = p$), y se supone que toda observación para la cual la diagonal de la matriz sombrero es más del doble del promedio $2p/n$ está suficientemente alejada del resto de los datos como para considerarse un punto de balanceo.

   Nótese que todos los puntos de balanceo serán influyentes en los coeficientes de regresión. Como los elementos de la diagonal de la matriz sombrero sólo examinan el lugar de la observación en el espacio de X, algunos analistas prefieren examinar los residuales estudentizados, o los residuales R de Student junto con los h_{ii}. Las observaciones con elementos grandes en la diagonal y también con residuales grandes probablemente serán influyentes.

6. La distancia D de Cook:

   Al medir de observaciones atípicas es preferible tener en cuenta el lugar del punto en el espacio de x y también la variable respuesta. Cook ha sugerido una forma de hacerlo,

con una medida de la distancia, elevada al cuadrado, entre el estimado por mínimos cuadrados basado en los n puntos $\hat{\beta}$, y el estimado obtenido después de eliminar el i-ésimo punto, por ejemplo $\hat{\beta}_{(i)}$. Esta medida de la distancia se puede expresar como sigue

$$D(X'X, pMS_{Res}) = D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'X'X(\hat{\beta}_{(i)} - \hat{\beta})}{pMS_{Res}} \quad , \quad i = 1, 2, ..., n$$

Los puntos con grandes valores de D_i tienen gran influencia sobre el estimado de $\hat{\beta}$ por mínimos cuadrados. La magnitud de D_i se suele evaluar comparándola con $F_{\alpha,p,n-p}$. Si $D_i = F_{0,5,p,n-p}$, entonces al eliminar el punto i se movería $\hat{\beta}_{(i)}$ hacia la frontera de una región de confianza aproximada de 50 % para β basándose en el conjunto completo de datos. Es un desplazamiento grande e indica que el estimado por mínimos cuadrados es sensible al i-ésimo punto de datos. Como $F_{0,5,p,n-p} = 1$, se suelen considerar como puntos influyentes todos aquellos que: $D_i > 1$.

La estadística D_i se puede expresar como

$$D_i = \frac{r_i^2}{p}\frac{Var(\hat{Y}_i)}{Var(\varepsilon_i)} = \frac{r_i^2}{p}\frac{h_{ii}}{(1 - h_{ii})} \quad , \quad i = 1, 2, ..., n$$

Así se ve que, además de la constante p, la D_i es el producto del cuadrado del i-ésimo residual estudentizado por $h_{ii}/(1 - h_{ii})$. Así, D_i está formada por un componente que refleja lo bien que se ajusta el modelo a la i-ésima observación Y_i y un componente que mide lo alejado que el punto está del resto de los datos. Cualquiera de los componentes (o ambos), pueden contribuir a un valor grande de D_i. Así, en D_i se combina la magnitud del residual para la i-ésima observación y la ubicación de ese punto en el espacio de X, para evaluar su influencia.

Dado que $X\hat{\beta}_{(i)} - X\hat{\beta} = \hat{Y}_{(i)} - \hat{Y}$, otra forma de expresar la medida de distancia de Cook es

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})'(\hat{Y}_{(i)} - \hat{Y})}{pMS_{Res}}$$

Así, otra forma de interpretar la distancia de Cook es el cuadrado de la distancia euclidiana (sin considerar pMS_{Res}).

La distancia de Cook es un diagnóstico de eliminación; es decir, mide la influencia de la i-ésima observación si se eliminara de la muestra.

7. *DFBETAS*:

   Los *DFBETAS* también son utilizados para la eliminación de datos influyentes. Es un estadístico que indica cuánto cambia el coeficiente de regresión $\hat{\beta}_j$, en unidades de desviación estándar, si se omitiera la i-ésima observación y se calculan así:

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S^2_{(i)} C_{jj}}}$$

   siendo C_{jj} el j-ésimo elemento diagonal de $(X'X)^{-1}$ y $\hat{\beta}_{j(i)}$ el j-ésimo coeficiente de regresión, calculado sin usar la i-ésima observación. Un valor de magnitud grande de $DFBETAS_{j,i}$ indica que la observación i tiene gran influencia sobre el j-ésimo coeficiente de regresión. Obsérvese que $DFBETAS_{j,i}$ es una matriz de $n \times p$ que contiene información similar a la de la distancia de Cook.

   Otra forma de calcular los $DFBETAS$ es

$$DFBETAS_{j,i} = \frac{r_{j,i}}{\sqrt{r_j r_j}} \frac{\varepsilon_i}{S_{(i)}(1 - h_{ii})} = \frac{r_{j,i}}{\sqrt{r_j r_j}} \frac{t_i}{\sqrt{1 - h_{ii}}}$$

   en donde t_i es el residual de R de Student. Nótese que $DFBETAS_{j,i}$ mide tanto el balanceo ($\frac{r_{j,i}}{\sqrt{r_j r_j}}$ es una medida del impacto de la i-ésima observación sobre $\hat{\beta}_j$) como el efecto de un residual grande. El valor de corte sugerido es $\frac{2}{\sqrt{n}}$ para $DFBETAS_{j,i}$; esto es, si $|DFBETAS_{j,i}| > \frac{2}{\sqrt{n}}$ es necesario entonces examinar la i-ésima observación.

8. *DFFITS*:

   Los *DFFITS* son usados para investigar la influencia de la eliminación de la i-ésima observación sobre el valor predicho o ajustado

$$DFFIT_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{S^2_{(i)} h_{ii}}} \quad , \quad i = 1, 2, ..., n$$

   donde $\hat{Y}_{(i)}$ es el valor ajustado de Y_i, obtenido sin usar la i-ésima observación. El denominador es una estandarización, porque $Var(\hat{Y}_{(i)}) = \sigma^2(h_{ii})$. Así, $DFFITS_i$ es la cantidad de desviaciones estándar que cambia el valor ajustado $\hat{Y}_i$ si se elimina la observación i.

   Otra forma de calcular los *DFFITS*es

$$DFFIT_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} \frac{\varepsilon_i}{S_{(ii)}\sqrt{(1-h_{ii})}} = \sqrt{\left(\frac{h_{ii}}{1-h_{ii}}\right)} t_i$$

donde t_i es el residual R de Student. Así, $DFFITS_i$ es el valor del R de Student multiplicado por el balanceo de la i-ésima observación $\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$. Si el dato es atípico, el residual R de Student tendrá magnitud grande, mientras que si el dato tiene gran balanceo, h_{ii} se aproximará a la unidad. En cualquiera de esos casos, $DFFITS_i$ puede ser grande. Sin embargo, si $h_{ii} \cong 0$, el efecto del R de Student será moderado. En forma parecida, un residual R de Student cercano a cero, combinado con un punto de gran balanceo podría producir un valor pequeño de $DFFITS_i$. Así se ve que $DFFITS_i$ se puede afectar ya sea por datos de balanceo como por el error de predicción. Se sugiere investigar todas observaciones para la cuales $|DFFIT_i| > \frac{2}{\sqrt{p/n}}$.

En este trabajo de tesis se usará las técnicas *DFFITS* y los residuales estudentizados para identificar datos influyentes, y así obtener una recta mucho más ajustada.

2.2. Modelos de Regresión Robustos

Los métodos de regresión robustos son técnicas que en potencia se pueden usar cuando hay valores atípicos. Los estimadores robustos permiten ponderar aquellos valores que se encuentran más alejados de los centrales en una serie de datos. Cuando las observaciones Y en el modelo de regresión lineal $Y = X\beta + \varepsilon$ están normalmente distribuidas, el método OLS es un buen procedimiento de estimación de parámetros, porque produce un estimador del vector β de parámetros que tiene buenas propiedades estadísticas (Montgomery, Peck y Vining, 2002), sin embargo, hay muchos casos en los que hay evidencias de que la distribución de la variable respuesta tiene una distribución (considerablemente) no normal y/o hay valores atípicos que afectan al modelo de regresión. Un caso de mucho interés práctico es aquél en el que las observaciones tienen una distribución que tiene colas más largas o gruesas que la distribución normal. Esas distribuciones tienden a generar valores atípicos, que pueden tener una gran influencia sobre el método OLS en el sentido que tienden a "jalar" demasiado la ecuación de regresión en su dirección.

Una forma de manejar esta situación es eliminar los valores atípicos, así se obtiene una recta que pasa muy bien por el resto de los datos, mejor desde un punto de vista estadístico, sin embargo, lo que se está haciendo ahora es descartar observaciones tan solo porque es agradable desde el punto de vista de modelado estadístico, y por lo general esa práctica no es la ideal o la más adecuada para solucionar el problema. A veces, los datos se pueden eliminar (o modificar) con base en el conocimiento de la materia, pero cuando se hace eso con una

base puramente estadística, en general se pueden cometer errores; también se observa que en casos más complicados, donde intervienen muchas variables explicativas y la muestra es mayor, puede dificultarse identificar que el modelo de regresión se ha distorsionado por las observaciones atípicas.

Un procedimiento de regresión robusto es aquel que amortigua el efecto de las observaciones que serían muy influyentes si se usaran los mínimos cuadrados, lo que indica que un procedimiento robusto tiende a dejar grandes los residuales asociados con valores atípicos, facilitando así la identificación de puntos influyentes. Además de la insensibilidad a los valores atípicos, un procedimiento de estimación robusta debería producir, en esencia, los mismos resultados obtenidos por el método OLS cuando la distribución básica es normal, y cuando no hay valores atípicos.
Gracias a que los métodos de regresión robusta tienen propiedades de robustez y ofrecen mejores soluciones a problemas de regresión con datos atípicos, estos métodos han alcanzado gran importancia y difusión a partir del año 1973. Autores como Seber (1977), McKean y Hettmansperger (1978), Birkes y Dodge (1993), Huber (1973), Denby, y Larsen (1977), Huber (1981), Hurdle (1981), Hettmansperger (1984), Lachan (1985), Hampel, Ronchetti, Rousseeuw, y Stahel (1986), Ronald y Montgomery (1984), Staudte, y Sheather (1990), Wilcox, (2005), Bianco, Garcia y Yohai (2005), Montanari (2008), Aelst, Willems y Zamar (2013), han desarrollado importantes aportes a la estadística robusta.

A continuación se describen los procedimientos de estimación MM, M y L (LAV), que se emplean en este trabajo de tesis y serán comparados con el método Gini y OLS.

2.2.1. Estimadores M

El método general más común en la regresión robusta es el método de estimación M propuesto por Peter Huber (1973). Esta clase de estimadores pueden ser considerados como una generalización de los de Máxima Verosimilitud; de ahí la denominación "estimadores M", donde M quiere decir Máxima verosimilitud. En OLS los parámetros estimados $\hat{\beta}_0$ y $\hat{\beta}_1$ son seleccionados después de minimizar una suma de errores al cuadrado. En la estimación M, esta idea es generalizada y $\hat{\beta}_0$ y $\hat{\beta}_1$ son seleccionados después de minimizar una función ρ de residuales. La función ρ se relaciona con la función de máxima verosimilitud para una elección adecuada de la distribución del error. Por ejemplo, si se usa el método OLS la función a minimizar es $\rho(\varepsilon) = \frac{1}{2}\varepsilon^2 \quad para \quad -\infty < \varepsilon \ < \infty$. El estimador M de Huber es robusto frente a valores extremos en la dirección de y, pero no es robusto frente a valores extremos en la dirección x. Considere el modelo de regresión lineal

$$\mathbf{y_i} = \mathbf{x_i'}\beta + \varepsilon_\mathbf{i} \tag{2-44}$$

para la i-ésima observación de n valores independientes, pero en este caso la distribución de los errores puede ser de colas pesada, producir valores atípicos.
Dado un estimador $\hat{\beta}$ para β, el modelo de regresión ajustado es

$$\hat{\mathbf{y}}_\mathbf{i} = \mathbf{x}_\mathbf{i}'\hat{\beta} \tag{2-45}$$

y los residuales están dados por

$$\varepsilon_\mathbf{i} = \mathbf{y}_\mathbf{i} - \hat{\mathbf{y}}_\mathbf{i} = \mathbf{y}_\mathbf{i} - \mathbf{x}_\mathbf{i}'\hat{\beta} \tag{2-46}$$

El estimador M se obtiene después de minimizar la siguiente función suma ponderada de errores absolutos:

$$\sum_{i=1}^{n} w_i \left| y_i - x_i'\beta \right|$$

utilizando diferentes ponderaciones, según la magnitud del error [1]
Por ejemplo:

1. Fijando todas las ponderaciones iguales a 1, esto es $w_i = 1$, para todo i; en este caso el estimador que minimiza la suma de errores absolutos; el estimador se denomina "Desviación Absoluta Mínima":

   $$MAD = \sum_{i=1}^{n} \left| y_i - x_i'\beta \right|$$

2. Utilizando un esquema de ponderaciones que dependa de la magnitud del error en términos absolutos: si éste es menor a un valor arbitrario b , se lo toma como ponderación; en caso contrario, la ponderación se estabiliza en b:

   $$w_i = \begin{cases} \left| y_i - x_i'\beta \right|; & si \quad \left| y_i - x_i'\beta \right| < b \\ b * sig(y_i - x_i'\beta); & si \quad \left| y_i - x_i'\beta \right| \geq b \end{cases}$$

3. Aplicando un esquema similar al anterior, pero con valores w_i decrecientes hacia cero si $\left| y_i - x_i'\beta \right| \geq b$. Cuando $\left| y_i - x_i'\beta \right|$ alcance un valor arbitrario d, $w_i = 0$. Esta opción implica desechar observaciones que generen errores absolutos mayores que d .

[1] En este sentido, los estimadores el método OLS utiliza como ponderación la magnitud de cada error absoluto.

El estimador M no necesariamente es invariante de escala. Para obtener una solución invariante de escala ante el error y obtener las estimaciones de $\hat{\beta}$, se resuelve la siguiente ecuación

$$\underset{\beta}{Min} \sum_{i=1}^{n} \rho\left(\frac{\varepsilon_i}{s}\right) = \underset{\beta}{Min} \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i\beta}{s}\right) \tag{2-47}$$

donde s es un estimador robusto de escala a menudo formado a partir de la combinación lineal de los residuales. La función ρ debe tener las siguientes propiedades:

1. Siempre es positiva, $\rho(\varepsilon) \geq 0$.
2. Igual a 0 cuando su argumento es 0, $\rho(0) = 0$.
3. Simétrica, $\rho(\varepsilon) = \rho(-\varepsilon)$.
4. Monotona en $|\varepsilon_i|$, $\rho(\varepsilon_i) \geq \rho(\varepsilon_{i'})$ para $|\varepsilon_i| > |\varepsilon_{i'}|$

Una elección muy frecuente de s es la mediana de la desviación absoluta de los errores, así como se muestra en la siguiente ecuación:

$$s = \frac{mediana|\varepsilon_i - mediana(\varepsilon_i)|}{0{,}6745} \tag{2-48}$$

La constante 0,6745 hace que s sea aproximadamente un estimador insesgado de σ si n es grande y la distribución del error es normal (Montgomery, Peck y Vining, 2002).
Para encontrar el estimador M, se debe minimizar (2-47) y se igualan a cero las primeras derivadas parciales de ρ con respecto a $\beta_j (j = 0, 1, ..., k)$, y así se obtiene una condición necesaria para encontrar el mínimo. De este modo se obtiene el sistema de $p = k + 1$ ecuaciones

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{y_i - x_i'\beta}{s}\right) = \sum_{i=1}^{n} x_{ij}\psi\left(\frac{\varepsilon_i}{s}\right) = 0 \quad , \quad j = 1, 2, ..., k+1 \tag{2-49}$$

donde $\psi(u) = \frac{\partial\rho}{\partial u}$ es la función "ponderadora". La función de pesos es definida como

$$w(u) = \frac{\psi(u)}{u} \tag{2-50}$$

haciendo $w_i = w\left(\frac{\varepsilon_i}{s}\right) \quad para \quad i = 1, ..., n$ con $w_i = 1$, si $\varepsilon_i = 0$. Sustituyendo esto en 2-49 se tiene

$$\begin{aligned}
&\sum_{i=1}^{n} x_{ij} w_i \tfrac{\varepsilon_i}{s} = \sum_{i=1}^{n} x_{ij} w_i \left(\tfrac{y_i - x_i'\beta}{s}\right) = 0 \ , \quad j = 1, ..., k+1 \\
&\Rightarrow \sum_{i=1}^{n} x_{ij} w_i \left(y_i - x_i'\beta\right) = 0 \ , \quad j = 1, ..., k+1 \\
&\Rightarrow \sum_{i=1}^{n} x_{ij} x_i w_i \beta = \sum_{i=1}^{n} x_{ij} w_i y_i = 0 \ , \quad j = 1, ..., k+1
\end{aligned} \tag{2-51}$$

dado que $s \neq 0$. La matriz de ponderación es definida como $W = diag(w_i : i = 1, ..., n)$

$$W = \begin{pmatrix} w_1 & & & & 0 \\ & w_2 & & & \\ & & . & & \\ & & & . & \\ 0 & & & & w_n \end{pmatrix} \tag{2-52}$$

así, la forma matricial de 2-51 es

$$\begin{aligned}
&X'WX\beta = X'WY \\
&\Rightarrow \hat{\beta} = \left(X'WX\right)^{-1} X'WY
\end{aligned} \tag{2-53}$$

Nótese que 2-53 tiene una forma muy similar a la solución dada por OLS, pero con la introducción de una matriz ponderadora cuyo objeto es reducir la influencia de los valores atípicos.

Los procedimientos de regresión robustos se pueden clasificar de acuerdo con el comportamiento de su función ψ. Esta función ψ controla el factor de ponderación que asigna el método a cada residual. Por ejemplo la función ψ para OLS no es acotada, por lo cual el método de estimación por OLS tiende a ser no robusto cuando se usan con datos procedentes de una distribución de colas gruesas. La función t de Huber tiene una función ψ monótona y no pondera residuales con tanta intensidad como el método OLS.

En la tabla **2.1** se presentan algunas funciones de criterio robusto $\rho(z)$ de uso frecuente. El comportamiento de cada función ρ y de su derivada ψ correspondiente se ilustra en las figuras **2-3**, **2-4** respectivamente.

Como ya se ha mencionado a lo largo de este documento los métodos de regresión robustos se pueden clasificar de acuerdo al comportamiento de su función ψ, la cual controla el factor de ponderación que se asigna a cada residual. A ψ se le suele llamar función de influencia en donde ψ es la derivada de la función ρ ($\psi = \rho'$) (Montgomery, Peck y Vining, 2002). Por ejemplo, la función ψ para los mínimos cuadrados no es acotada, por lo que los mínimos cuadrados tienden a ser no robustos cuando se usan con datos procedentes de una distribución

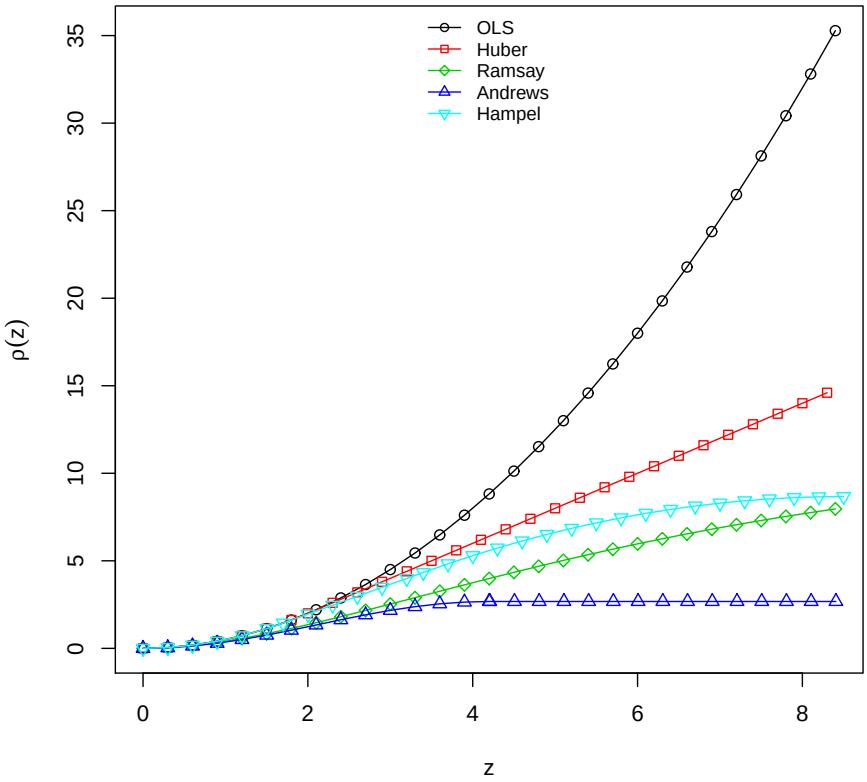


Figura 2-3: Funciones de Criterio Robusto para ponderar datos atípicos.

con colas gruesas. La función ψ de Huber es monótona (conserva un orden de crecimiento y de decrecimiento) y no pondera residuales grandes con tanta intensidad como los mínimos cuadrados. La función ψ es monótona redescendiente, dándole un peso menor a los residuales a medida que estos se hacen más grandes.

La tabla **2.1** muestra algunas funciones de criterio robusto, donde ρ es la función objetivo a minimizar (función de residuales), es decir que $\rho(z) = \rho(\varepsilon)$ definida por cada uno de los métodos robustos; la función ψ es la función de influencia y es expresada como la derivada de la función ρ ($\psi = \rho'$).

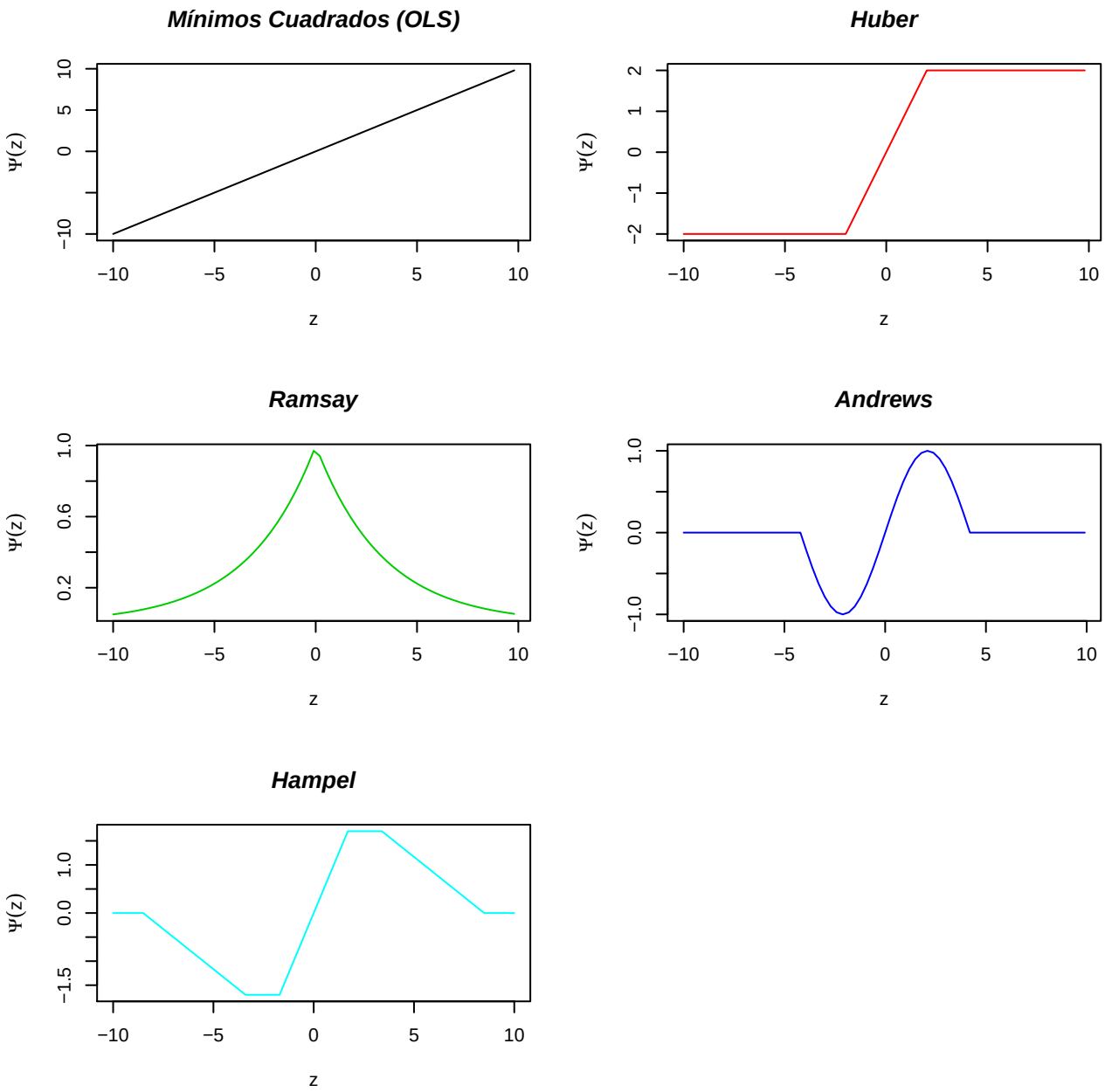


Figura 2-4: Funciones robustas de influencia

2.2.2. Estimador S

Los estimadores s se basan en estimaciones de escala, donde s quiere decir escala (scale). Los estimadores s forman una clase de estimadores de regresión de alto punto de quiebre, estos fueron desarrollados por Rousseeuw y Yohai (1984). Los estimadores s se obtienen luego de minimizar la función de dispersión de residuales $\rho(\varepsilon/s)$ tal que el factor de escala s debe ser aquel que produzca el β que minimice el error. El método de estimación por OLS minimiza la varianza de los residuales, el método de estimación s minimiza la dispersión de los residuales $\underset{\beta}{Min}\, s[\varepsilon_1(\beta), \varepsilon_2(\beta), ..., \varepsilon_n(\beta)]$. Rousseeuw y Leroy (1987: 135-136) definen la función de dispersión, s, de los residuales $\varepsilon_i(\beta)$ y se determina por la solución de

$$\sum_{i=1}^{n} \rho\left(\frac{\varepsilon_i}{s}\right) = k \tag{2-54}$$

donde k es una constante y la función objetivo ρ satisface las siguientes condiciones:

1. La función objetivo ρ es simétrica, continua, diferenciable, y $\rho(0) = 0$.

2. Existe un $c > 0$ tal que ρ es estrictamente creciente en $[0, c]$ y constante en el intervalo $[c, \infty]$.

3. El punto de quiebre de los estimadores s es de 0,5, es decir, $\frac{k}{\rho(c)} = \frac{1}{2}$.

Rousseeuw y Yohai (1984), han sugerido la siguiente función ρ

$$\rho(z) = \begin{cases} \frac{z^2}{2} - \frac{z^4}{2c^2} + \frac{z^6}{6c^4} & si \quad |z| \leq c \\ \frac{c^2}{6} & si \quad |z| > c \end{cases} \tag{2-55}$$

Para $z \sim N(0,1)$, la función de densidad de probabilidad es $f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$. Entonces con $c = 1{,}547$

$$
\begin{aligned}
K &= E_\Phi(\rho(z)) = \int_{-\infty}^{+\infty} \rho(z) f(z) dz \\
&= 2 \int_{0}^{+\infty} \rho(z) f(z) dz \ , \quad \rho(z) \quad y \quad f(z) \quad son \quad funciones \quad simétricas \\
&= 2 \int_{0}^{+\infty} \rho(z) \tfrac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{z^2}{2}\right) dz \\
&= 2 \int_{0}^{1{,}547} \left(\tfrac{z^2}{2} - \tfrac{z^4}{2*1{,}547^2} + \tfrac{z^6}{6*1{,}547^6}\right) \tfrac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{z^2}{2}\right) dz \\
&= 2 \int_{0}^{1{,}547} \left(\tfrac{1{,}547^2}{6}\right) \tfrac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{z^2}{2}\right) dz \\
&= \sqrt{\tfrac{2}{\pi}} \int_{0}^{1{,}547} \left(\tfrac{z^2}{2} - \tfrac{z^4}{4{,}786} + \tfrac{z^6}{34{,}365}\right) \exp\left(-\tfrac{z^2}{2}\right) dz + \tfrac{1{,}547^2}{3} \int_{1{,}547}^{+\infty} \tfrac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{z^2}{2}\right) dz \\
&= \sqrt{\tfrac{2}{\pi}} \int_{0}^{1{,}547} \left(\tfrac{z^2}{2} - \tfrac{z^4}{4{,}786} + \tfrac{z^6}{34{,}365}\right) \exp\left(-\tfrac{z^2}{2}\right) dz + 0{,}798 * (1 - \Phi(1{,}547)) \\
&= \sqrt{\tfrac{2}{\pi}} * 0{,}189 + 0{,}798 * (1 - 0{,}939) \\
&= 0{,}199
\end{aligned}
\tag{2-56}
$$

Ahora $\rho(1{,}547) = \frac{1{,}547^2}{2} = 0{,}399$, y por lo tanto, omitiendo el redondeo de los errores se llega a que

$$\frac{k}{\rho(c)} = \frac{E_\Phi(\rho(z))}{\rho(1{,}547)} = \frac{1}{2} = 0{,}5$$

Para conseguir un balance entre el punto de quiebre y la eficiencia asintótica se debe hacer una selección adecuada de los valores de las constantes c y k. Usualmente los valores asignados para estas dos constantes son: $c = 1{,}547$, $k = 0{,}1995$ para un punto de quiebre del 50 % y una eficiencia asintótica alrededor del 28 %, ver tabla **2.2**.

Finalmente el estimador, s, de escala es la desviación estándar de los residuales del ajuste que minimiza la dispersión de los residuales.

Punto de quiebre	Eficiencia	c	k
50 %	28.7 %	1.547	0.1995
40 %	46.2 %	1.988	0.2634
35 %	56.0 %	2.251	0.2957
30 %	66.1 %	2.560	0.3278
25 %	75.9 %	2.973	0.3593
20 %	84.7 %	3.420	0.3899
15 %	91.7 %	4.096	0.4194
10 %	96.6 %	5.182	0.4475

Tabla 2-1: Eficiencia relativa asintótica de estimadores s para diferentes puntos de quiebre.

2.2.3. Estimadores MM

El estimador MM es un tipo especial de estimador M y fue propuesto por Yohai (1987). Los estimadores MM son considerados como una generalización de los estimadores M (donde M quiere decir Máxima verosimilitud) dado que en el proceso de estimación ellos son obtenidos luego de aplicar consecutivamente el estimador M en las dos últimas etapas del proceso. La regresión MM ya ha sido trabajada por Yohai (1987), Welsh y Ronchetti (2002), Wibowo (2009).
El método de estimación MM es desarrollado con el objetivo de obtener simultáneamente un estimador de punto de quiebre alto y que mantenga una alta eficiencia. El estimador MM tiene simultáneamente las siguientes propiedades: es altamente eficiente cuando la distribución del error es normal y su punto de quiebre es de 0,5. Los estimadores MM son desarrollados básicamente en tres etapas o pasos, los cuales se mencionan a continuación:

1. En el primer paso se calcula un estimador de regresión con alto punto de quiebre, denotado por $\tilde{\beta}$ pero no necesariamente eficiente. Es decir, se calcula un estimador S. Usando este estimado se calculan los resuduales $\varepsilon_i(\tilde{\beta}) = (y_i - x_i'\tilde{\beta})$.

2. En este segundo paso se calcula un estimado M con puto de quiebre del 50 %, usando los residuales del ajuste robusto (del paso 1) y la ecuación 2-54. Estos $s(\varepsilon_1(\beta), \varepsilon_2(\beta), ..., \varepsilon_n(\beta))$ son denotados como s_n. La función objetivo usada en esta etapa es denotada ρ_0.

3. En este tercer paso el estimador MM es ahora definido como un estimador M de β usando una función redescendiente, $\psi_1(z) = \frac{\partial \rho_1(z)}{\partial z}$, y el estimado de escala s_n obtenido en el paso 2. Así que un estimador MM de $\hat{\beta}$ se define como una solución a

$$\sum_{i=1}^{n} x_{ij}\psi_1\left(\frac{y_i - x_i'\beta}{s_n}\right) = 0 \ , \quad j = 1, ..., k+1. \tag{2-57}$$

La función objetivo ρ_1 asociada con esta función redescendiente ψ no tiene que ser la misma que ρ_0, pero si debe satisfacer: premisas

a) La función objetivo ρ es simétrica y continuamente diferenciable, y $\rho(0) = 0$.

b) Existe un valor $c > 0$ tal que ρ es estrictamente creciente en el intervalo $[0, c]$ y constante en el intervalo $[c, \infty]$.

c) $\rho_1(z) \leq \rho_0(z)$.

Una condición final que debe satisfacer la solución dada por la ecuación 2-57 es que

$$\sum_{i=1}^{n} \rho_1 \left(\frac{y_i - x_i' \hat{\beta}}{s_n} \right) \leq \sum_{i=1}^{n} \rho_1 \left(\frac{y_i - x_i' \tilde{\beta}}{s_n} \right) \tag{2-58}$$

En una evaluación del desempeño de varios estimadores de regresión robusta, Simpson y Montgomery (1998), encontraron que los estimadores MM tienen una alta eficiencia y trabajan bien en la mayor parte de los escenarios de valores atípicos. Su único punto débil tiene que ver cuando hay grandes porcentajes de valores atípicos en el espacio de las X.
El algoritmo para estimadores MM está implementada en varios paquetes estadísticos como el R, la función utilizada en R para estimar las parámetros por MM es rlm, y se le debe especificar que estime por el método MM (rlm(Y $\sim$ X,method=MM) y se debe cargar la librería MASS.

2.2.4. Estimadores L

Los estimadores L [2] son combinaciones lineales de estadísticos de orden muestrales. Por ejemplo, supóngase que se desea estimar el parámetro de localización de una distribución a partir de una muestra aleatoria $x_1, x_2, ..., x_n$. Los estadísticos de orden para esta muestra son $x_1 \leq x_2 \leq, ..., \leq x_{n-1} \leq x_n$. La mediana de la muestra sería un estimador L para el problema de la localización (Dielman. T.E, 2005). Esta misma idea se extiende al problema de regresión de Mínimo Valor Absoluto, (Least Absolute Value-LAV).

Regresión cuantil

La regresión cuantil fue introducida por Koenker y Bassett en 1978. En la mayoría de los estudios sociales se modelan momentos condicionales como en el caso de la relación $y_i = x_i'\beta + \varepsilon_i$, donde la media de y se estima condicionada por las variables regresoras x. Sin embargo, existe actualmente un interés por algunos investigadores en métodos que permitan apreciar otros aspectos de la distribución de y. Por lo cual la regresión por cuantiles apunta a una descripción más detallada al modelizar, no solo el valor medio condicional, sino la mediana,

[2] "Least" por sus siglas en ingles

cuartiles, percentiles, etc; condicionados a los niveles de las variables explicativas. Hay casos en los que es preferible para el investigador utilizar la regresión cuantil que la regresión por OLS y una razón es que los modelos de regresión cuantil permiten una descripción más detallada del comportamiento de la variable respuesta, se adaptan a situaciones bajo condiciones más generales de la distribución del error, tienen propiedades de robustez y ofrecen mejores soluciones a problemas de regresión con datos atípicos. Es importante resaltar que aunque la regresión por OLS, ha alcanzado la mayor difusión y práctica en la Estadística del siglo XX, las ideas de regresión cuantil fueron anteriores a los procedimientos basados en OLS. Así, mientras que las primeras publicaciones de la regresión por OLS se pudieron dar alrededor del año 1805 por el trabajo de Legendre, a mediados del siglo XVIII Boscovich ya había ajustado datos sobre la elipticidad de la tierra mediante procedimientos de regresión cuantil (Dielman. T.E, 2005). Los métodos de regresión cuantil obtuvieron un gran desarrollo desde el surgimiento de la Estadística Robusta, que alcanzó una gran importancia a principios de los años 80. Huber (1981), Hampel (1986) han realizado grandes aportes a la Teoría de la Robustez, cuyos conceptos siguen siendo aplicados hoy en día en diferentes áreas de la ciencia.
Si se consideran cuantiles equiespaciados, por ejemplo, cada 5 % o 1 % de la población, es posible modelizar cada una de las posiciones predeterminadas de la distribución de y. Por tanto, la regresión por cuantiles amplifica notablemente el análisis de la distribución de la variable respuesta, condicionada al conjunto de variables explicativas. Un caso importante de la regresión por cuantiles es el estimador "Desviación Absoluta Mediada" (MAD), que corresponde al ajuste de la mediana condicional de la variable respuesta.

Regresion LAV

Los estimadores obtenidos por el método LAV se calculan minimizando el valor absoluto de los residuales

$$\underset{\beta}{Min} \sum_{i=1}^{n} |\varepsilon_i| = \underset{\beta}{Min} \sum_{i=1}^{n} |y_i - x_i\beta| \tag{2-59}$$

El método de regresión LAV puede ser visto como un caso especial de la Regresión Cuantil Generalizada, en este caso la función a minimizar es la siguiente

$$\sum_{i=1}^{n} \rho_\alpha(\varepsilon_i) \tag{2-60}$$

donde

$$\rho_\alpha(\varepsilon_i) = \begin{cases} \alpha\varepsilon_i, & si \quad \varepsilon_i \geq 0 \\ (\alpha - 1)\varepsilon_i, & si \quad \varepsilon_i < 0 \end{cases} \tag{2-61}$$

donde α es el cuantil a ser estimado. En general, los estimadores L tienen bajo punto de quiebre y baja eficiencia, y en algunos casos no es obvia una generalización clara de un procedimiento que pase de la regresión lineal simple a la múltiple (Mosteller y Tukey 1977).

La regresión LAV es también conocida por otros nombres diferentes tales como: norma de regresión L_1, regresión de mínima desviación absoluta, y regresión de mínimo error absoluto (Dielman. T.E, 2005). La regresión LAV fue propuesta con anterioridad al método de estimación por OLS y esta fue realizada por el matemático astrónomo y físico Jesuita Roger Boskovich, aproximadamente 50 años antes de la aparición del trabajo de Legendre sobre OLS en 1805.

Sea el modelo $y_i = x_i'\beta + \varepsilon_i$, donde ε_i es el error correspondiente a la observación i, resulta que:

1. El método OLS minimiza la función objetivo $\sum_i \varepsilon_i^2$

2. La regresión por la mediana (estimador MAD) minimiza $\sum_i |\varepsilon_i|$

3. La regresión por quintiles considera como función objetivo una suma que pondera asimétricamente los errores absolutos: otorga ponderación τ a errores por subestimación y ponderación $(1-\tau)$ para errores por sobreestimación.

En un análisis clásico de regresión lineal, formalizado mediante el modelo $y_i = x_i'\beta + \varepsilon_i$, se intenta determinar la incidencia de las variables explicativas x_i sobre el valor medio condicional de y; la incidencia resulta cuantificada al obtener las estimaciones $\hat{\beta}$.

En la regresión por cuantiles, en cambio, se analiza la incidencia de las variables explicativas sobre cada uno de los cuantiles de la distribución de y; se obtiene entonces un vector de estimaciones por cada cuantil, que se simbolizará como $\tilde{\beta}_\tau$.

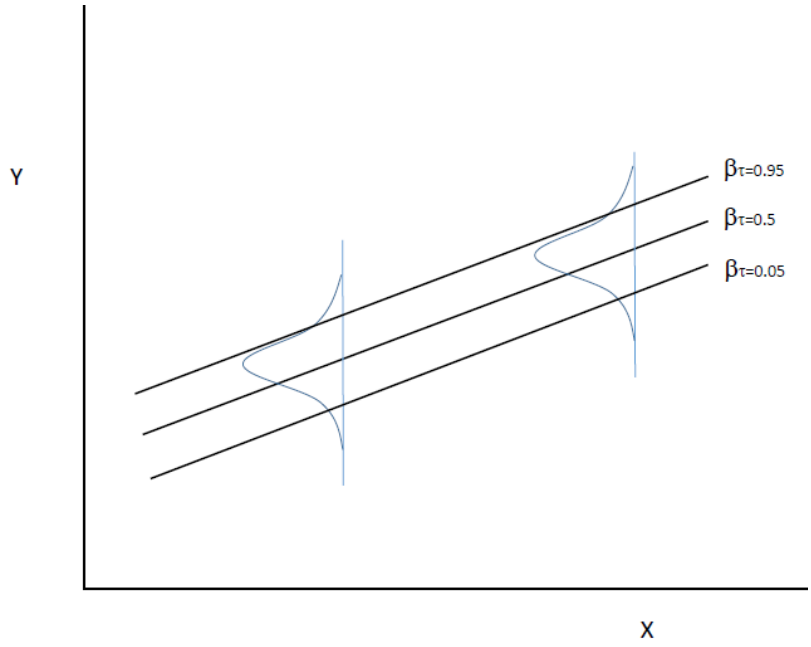


Figura 2-5: Regresión Cuantil

En el gráfico **2-5** se ha representado para dos valores de x la distribución condicionada de y que se supone en ambos casos normal pero con distinta media μ y varianza σ^2.
La recta central representa la recta de regresión de OLS, es decir, la esperanza condicional de y dado x, que coincide en este caso con la mediana 0,50. Las líneas punteadas son las regresiones cuantílicas Q_τ para $\tau = 0{,}05$ y $\tau = 0{,}95$.

$Q(\tau/x_i, \beta_\tau) = x_i'\beta$

siendo β_τ el vector de coeficientes asociado al τ cuantil.
La función objetivo presentada en la ecuación 2-61 se modifica de manera que el vector cuantílico $\tilde{\beta}_\tau$ resulta de minimizar, respecto a β_τ, la siguiente expresión:

$$\min_{\beta_\tau} \left\{ \sum_{i:y_i \geq x_i'\beta_\tau} \tau \left| y_i - x_i'\beta_\tau \right| + \sum_{i:y_i < x_i'\beta_\tau} (1-\tau) \left| y_i - x_i'\beta_\tau \right| \right\}$$

Obsérvese que se utiliza la notación β_τ y no β como en el caso de OLS, para destacar que diferentes elecciones de τ generan distintas estimaciones de β. Por otra parte, si τ es 0,90, por

ejemplo, la ponderación asimétrica opera con mayor intensidad para observaciones $y \geq x'\beta$ que para observaciones donde $y < x'\beta$.
La función objetivo no es diferenciable, por lo que se utilizan métodos de programación lineal para la búsqueda de la solución.
Es importante notar que este enfoque no requiere supuestos fuertes con respecto a la distribución de los errores, por lo cual es considerado como un método robusto para modelizar relaciones entre y y x. Los autores Koenker y Bassett demuestran además la consistencia y normalidad asintótica. Es importante aclarar, que si se tiene una variable y, donde, $y \sim N(\mu, \sigma^2)$, entonces la varianza de y por OLS es σ^2, mientras que por LAV es $\frac{\pi}{2}\sigma^2$. Esta aclaración es muy importante, puesto que ha permitido obtener los resultados de la simulación consistente y coherente con los esperados en este trabajo de tesis. En este trabajo de tesis se estimaron los coeficientes de regresión para el método LAV mediante el paquete estadístico R usando la función *rq*, para su uso se debe cargar el paquete *quantreg*.

2.2.5. Propiedades de los Estimadores Robustos

Tradicionalmente las tres principales propiedades de la teoría asintótica son listadas a continuación. De una secuencia de estimadores $\hat{\theta}$ de θ usualmente se espera que los resultados asintóticos (cuando n tiende a infinito) podrían describir el comportamiento de $\hat{\theta}$ para un tamaño muestral particular n de interés, o quizás, para un tamaño muestral moderadamente más grande (Staudte y Sheather, 1990).

1. **Consistencia.** Una secuencia de estimadores $\hat{\theta}$ es consistente para un parámetro θ si $\hat{\theta}$ converge en probabilidad a θ.

2. **Normalidad asintótica.** Se puede decir que $\hat{\theta}$ es asintóticamente normal con media θ y varianza v(θ) si $\sqrt{n}\left(\hat{\theta} - \theta\right)$ converge en distribución a N(θ, v(θ)).

3. **Eficiencia relativa.** Dadas dos secuencias de estimadores, dígase $\hat{\theta}_1$ y $\hat{\theta}_2$, los cuales son cada uno consistentes para θ y también asintóticamente normales, la eficiencia relativa del primero respecto al segundo está dada por la razón $v(\theta_2)/v(\theta_1)$, lo cual afortunadamente a menudo no depende del parámetro θ. Supóngase que $\hat{\theta}_1$ es el coeficiente de regresión estimado por OLS y $\hat{\theta}_2$ por algún método robusto. Nótese que entre más robusto sea el método, menor es la varianza de éste ante la presencia de datos atípicos, por lo cual la eficiencia de OLS respecto al método robusto es más pequeña.

4. **Punto de quiebre**

   El punto de quiebre con muestra finita es la mínima fracción de datos atípicos que puede hacer inútil al estimador. El mínimo punto de quiebre posible es $1/n$, lo que equivale a que una sola observación puede distorsionar al estimador como para hacer que su utilidad práctica sea nula, para quien forma el modelo. El punto de quiebre del

método OLS es $1/n$. Los estimadores M se pueden alterar debido a valores atípicos en el espacio de x, de manera idéntica que los de OLS, consecuentemente, el punto de quiebre de los estimadores M es $1/n$. Esto tiene un impacto potencialmente grave sobre su uso práctico, porque se puede dificultar la determinación del grado de contaminación de la muestra por datos atípicos. La mayoría de los analistas de datos con experiencia creen que la fracción de los datos que contaminan una muestra está comúnmente entre el 1 y el 10 %, por lo que se desea que el punto de quiebre de un estimador robusto sea mayor al 10 %. Esto ha conducido al desarrollo de estimadores robustos de punto de quiebre alto. Los estimadores M tienen puntos de quiebre bajos ($1/n$), mientras que los estimadores MM son de punto de quiebre alto.

5. **Eficiencia en términos del MSE**

   Cuando las observaciones son normales y no hay valores atípicos, la mejor técnica de estimación de los coeficientes de regresión es el método OLS, porque produce el menor Error Cuadrático Medio MSE (varianza estimada de los residuales). Cuando hay valores atípicos el MSE es más grande en los modelos de estimación robustos que en OLS, ya que por OLS la recta de regresión es jalada en su dirección, disminuyendo la magnitud del residual.

   La eficiencia se define como:

$$Eficiencia = \frac{MSE_{OLS}}{MSE_{Robusto}} \tag{2-62}$$

   Bajo la normalidad de las observaciones y la ausencia de valores atípicos, la eficiencia de un buen estimador robusto debería ser cercana al 1, indicando que produce, en esencia, los mismos resultados que OLS. Obsérvese que a medida que el estimador es más robusto, el MSE es más grande en presencia de datos atípicos y por lo tanto la eficiencia es más pequeña. Entonces, el método más robusto será aquel que tenga la menor eficiencia relativa a OLS en términos del MSE.

3 Regresión Simple de Gini

La regresión de Gini es considerada como un método robusto para la estimación de parámetros en aquellos casos donde los supuestos de los errores no se cumplen (Olkin y Yitzhaki, 1992). El bloque básico de construcción de regresión es la covarianza entre la variable dependiente Y y la variable explicativa X. La regresión de Gini considera dos enfoques para la construcción de un modelo de regresión y ambos pueden ser calculados con base en la Diferencia Media de Gini (GMD).

Diferencia Media de Gini-GMD

Olkin y Yitzhaki (1992) proponen una expresión para calcular la GMD. Sean X_1 y X_2 dos variables aleatorias idénticamente distribuidas. La GMD de X_1 y X_2 se define como

$$G_X = E\left|X_1 - X_2\right|.$$

La varianza de X_1 y X_2 puede ser escrita como

$$\sigma^2 = \frac{1}{2}E[X_1 - X_2]^2.$$

La GMD se puede definir en una variedad de formas (Olkin y Yitzhaki, 1992), algunas de las cuales son

$$G_X = 4Cov(X, F_X(X))$$

y

$$G_X = 2\int F_X(t)[1 - F_X(X)]dt$$

donde F es la función de distribución acumulada.

Cuando X tiene una distribución normal la GMD puede calcularse como

$$G_X = \frac{2\sigma_X}{\sqrt{\pi}}.$$

Así la GMD y la varianza son medidas de dispersión equivalentes.

Como se mencionó antes, la regresión de Gini puede ser calculada desde dos enfoques.

1. El primer enfoque es considerado como un procedimiento **semi-paramétrico**, el cual está basado en expresar la covarianza de Gini entre la variable dependiente y la variable explicativa en función de una suma de promedios de pendientes ponderadas. El enfoque semi-paramétrico de la regresión Gini es similar en su estructura al método OLS.

2. El segundo enfoque está basado en la minimización de la GMD de los residuales.

3.1. Enfoque semi-paramétrico

Sea (Y, X) variables aleatorias bivariadas que sigue una distribución continua con el primer y segundo momento finito (media y varianza). Nótese que en este punto no se imponen supuestos específicos. En particular, no se asume que X es fija o que hay alguna relación lineal entre las dos variables Y y X.
Cuando el objetivo del investigador es construir un predictor lineal de Y basado en los datos que contiene la variable X, el predictor lineal teórico se denota por

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{3-1}$$

donde $\hat{\beta}_0$ y $\hat{\beta}_1$ son constantes arbitrarias estimadas con los datos. Los residuales se definen como

$$\varepsilon = Y - \hat{Y} = Y - \hat{\beta}_0 - \hat{\beta}_1 X. \tag{3-2}$$

Note que (3-2) es una identidad y no se han impuesto supuestos sobre los residuales ε. Todas sus propiedades se derivan de las propiedades de (Y, X). Usando las propiedades de la covarianza se obtiene

$$Cov(Y, X) = Cov(\beta_0 + \beta_1 X + \varepsilon, X) = \beta_1 Cov(X, X) + Cov(\varepsilon, X). \tag{3-3}$$

Ahora se adiciona un supuesto: imponer la restricción de ortogonalidad, $\mathrm{Cov}(\varepsilon, X) = 0$, lo cual permite resolver para $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\mathrm{Cov}(Y, X)}{\mathrm{Cov}(X, X)} \tag{3-4}$$

El $\hat{\beta}_1$ de la ecuación 3-4 es equivalente en su estructura al $\hat{\beta}_1$ obtenido para OLS con las ecuaciones normales, una de las cuales es $Cov(\varepsilon, X) = 0$. En OLS la restricción de que $Cov(\varepsilon, X) = 0$ es derivada de la minimización de la varianza del término del error. A partir de la ecuación 3-4 y remplazando cada término por el por el término equivalente del método

de Gini, $Cov(Y, F(X))$ y $Cov(X, F(X))$ reemplaza la $Cov(Y, X)$ y $Cov(X, X)$ conduciendo al parámetro equivalente para la regresión semi-paramétrico de Gini:

$$\hat{\beta}_{GN} = \frac{\mathrm{Cov}(Y, F(X))}{\mathrm{Cov}(X, F(X))}. \tag{3-5}$$

Teniendo en cuenta las propiedades de la covarianza se tiene la siguiente restricción

$$\mathrm{Cov}(\varepsilon_{GN}, F(X)) = 0 \tag{3-6}$$

donde ε_{GN} son lo residuales de la regresión semi-paramétrica de Gini (Yitzhaki y Schechtman, 2013). La ecuación 3-6 es equivalente a la ecuación normal obtenida por el método OLS. Una vez que se determina la pendiente asociada al predictor lineal de la ecuación 3-1, se puede utilizar una restricción adicional para determinar β_0. Si se quiere que la línea de regresión pase a través de la media de las variables, entonces β_0 puede ser determinado como una solución de 3-7

$$\mu_y = \hat{\beta}_0 + \hat{\beta}_{GN}\mu_x. \tag{3-7}$$

Sin embargo, también se pueden usar otros criterios para determinar a β_0, como la minimización de la suma de las desviaciones absolutas de los residuales de una constante, en cuyo caso β_0 causa la línea de regresión que pasa a través de la mediana o algún cuantil de la distribución residual, como en la regresión cuantil (con algunas modificaciones). El punto importante aquí es que se puede separar entre el criterio que se usa para determinar la pendiente y el criterio que se usa para determinar la constante, el cual es determinado como un parámetro de localización (Yitzhaki y Schechtman, 2013). La regresión de Gini ya ha sido trabajada y aplicada por algunos autores como: David (1968), Meintanis, y Donatos,(1997), Mosteller y Tukey (1983), Yitzhaki (1998), Schechtman y Yitshaki (2003); Yitzhaki (2003), Yitzhaki y Schechtman (2004), Yitzhaki y Schechtman (2005), Borroni, Zenga (2006), Borroni, Cazzaro (2006), Edna y Yitzhaki (2007), Edna, Yitzhaki, y, Artsev, (2008), Edna y Yitzhaki (2011), Edna, Yitzhaki y Pudalov (2013), Yitzhaki y Schechtman (2013). A continuación se muestra como el coeficiente de regresión de Gini también puede ser expresado como una suma ponderada de pendientes de la recta de regresión (Yitzhaki y Schechtman, 2013). En este capítulo se presenta el enfoque semi-paramétrico de la regresión OLS, el enfoque semi-paramétrico de la regresión Gini, el enfoque de minimización de la regresión Gini y finalmente se presenta una comparación gráfica de los ponderadores de Gini y OLS.

3.2. Regresión de Mínimos Cuadrados Ordinarios bajo el enfoque semi-paramétrico

El propósito de esta sección es interpretar el coeficiente de regresión de OLS como un promedio ponderado de pendientes de la recta de regresión. Esta interpretación es útil debido a

que éste recalca el hecho de que los coeficientes de regresión de la ecuación 3-1 no depende de la distribución subyacente de la variable dependiente y que el supuesto de linealidad entre las variables del modelo de regresión no juega un rol en la obtención del coeficiente (Yitzhaki y Schechtman, 2013). Algunos autores que han desarrollado importantes trabajos sobre la regresión no paramétrica son: Jurecková (1971), Jaeckel (1972), Stone (1977), Tukey (1977), Randles, y Wolfe(1979), Boente y Fraiman (1989), Wolfe y Hollander (1999), Yitzhaki y Schechtman (2009).
Sea $(Y,\ X)$ una variable aleatoria bivariada con función de densidad $f(y,x)$. Sea f_x, F_x, μ_x, y σ_x^2 la densidad, la distribución acumulada, el valor esperado y la varianza de x, respectivamente. Sea $g(x) = E[Y|X = x]$ la recta de regresión, donde su derivada, $g'(x)$, es la pendiente de la recta de regresión denotada como

$$g'(x) = \frac{\partial E[Y|X = x]}{\partial x}.$$

Sea $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ el predictor lineal de Y dado X. Entonces $\hat{\beta}_1$ de OLS puede ser expresado como un promedio ponderado de pendientes:

$$\hat{\beta}_{1OLS} = \int w(x)g'(x)dx.$$

donde $w(x) > 0$ denota los pesos en x y $\int w(x)dx = 1$. Los pesos son

$$w(x) = \frac{1}{\sigma_x^2}[\mu_x F_x(x) - ALCV(x)] = \frac{F_x(x)}{\sigma_x^2}(\mu_x - E[X|X \leq x]). \tag{3-8}$$

donde

$$ALCV(x) = \int_{-\infty}^{x} t f_X(t)dt = F_x(x)E[X|X \leq x].$$

Demostración:

$$ALCV(x) = \int_{-\infty}^{x} t f_X(t)dt = F_x(x)E[X|X \leq x].$$

Para ello se tienen en cuenta la definición de $LAD(q)$ (Yitzhaki y Schechtman (2013)), cap 5, pag 77.

$$LAD(q) = \int_{-\infty}^{q} (q - x)dF(x) = F(q)E[(q - X)|X < q]$$

Así,

$$
\begin{aligned}
&LAD(q) = \textstyle\int_{-\infty}^{q} (q-x)dF(x) = F(q)E[(q-X)|X<q] \\
&\Leftrightarrow \textstyle\int_{-\infty}^{q} qF(x)dx - \int_{-\infty}^{q} xF(x)dx = F(q)E[(q-X)|X<q] \\
&\Leftrightarrow \textstyle\int_{-\infty}^{q} xF(x)dx = q\int_{-\infty}^{q} F(x)dx - F(q)E[(q-X)|X<q] \\
&\Leftrightarrow qF(q) - F(q)[E(q|X<q) - E(X|X<q)] \\
&\Leftrightarrow qF(q) - qF(q) + F(q)E(X|X<q) \\
&\Rightarrow \textstyle\int_{-\infty}^{q} xF(x)dx = F(q)E(X|X<q) \\
&\Rightarrow \textstyle\int_{-\infty}^{x} tF(t)dt = F_x(x)E(X|X<x) \\
&\Rightarrow \textstyle\int_{-\infty}^{x} tf_x(t)dt = F_x(x)E(X|X<x)
\end{aligned}
$$

$ALCV(x)$ es la curva de Lorenz absoluta como una función de x, mientras $[\mu_x F_x(X)]$ es la línea transformada de independencia (LOI).
En Yitzhaki (1996) demuestran que

$$\hat{\beta}_{1OLS} = \frac{\mathrm{Cov}(Y,X)}{\mathrm{Cov}(X,X)} = \frac{1}{\sigma_x^2}\int_{-\propto}^{\propto} [\mu_x F_x(x) - ALC(x)]g'(x)dx = \int w(x)g'(x)dx.$$

Demostración:

$$\hat{\beta}_{1OLS} = \frac{Cov(Y,X)}{Cov(X,X)}$$

donde $Cov(Y,X)$ se puede expresar como

$$
\begin{aligned}
&Cov(Y,X) = E_X E_Y[(Y-\mu_Y)(X-\mu_X)] = E_X E_Y[(X-\mu_X)Y] \\
&= E_X[(X-\mu_X)(Y|X=x)] \\
&= \textstyle\int (X-\mu_X)g(x)f_X(x)dx
\end{aligned}
$$

donde $g(x) = E(Y|X=x)$
Aplicando integración por partes, se tiene que

$$
\begin{aligned}
v'(x) &= \textstyle\int (X-\mu_X)f_X(x)dx, & v(x) &= \textstyle\int_{-\infty}^{x} (t-\mu_X)f_X(t)dt \\
u(x) &= g(x), & u'(x)dx &= g'(x)dx
\end{aligned}
$$

Luego,

$$Cov(Y,X) = \left\{\int_{-\infty}^{\infty} (t-\mu_X)f_X(t)dt\right\} g(x)|_{-\infty}^{\infty} + \int_{-\infty}^{\infty}\left[\int_{-\infty}^{x} (\mu_X - t)f_X(t)dt\right] g'(x)dx.$$

Dado que el segundo momento existe, el primer término $\left\{\int_{-\infty}^{\infty} (t-\mu_X)f_X(t)dt\right\} g(x)|_{-\infty}^{\infty} = 0$ así

$$Cov(Y,X) = \int_{-\infty}^{\infty} [\mu_X F(x) - ALC(x)] g'(x) dx$$

por lo tanto

$$\hat{\beta}_{1OLS} = \frac{Cov(Y,X)}{Cov(X,X)} = \frac{1}{\sigma_X^2} \int_{-\infty}^{\infty} [\mu_X F(x) - ALCV(x)]\, g'(x) dx = \int w(x) g'(x) dx.$$

Para mostrar que la suma de los pesos es igual a uno se aplica el mismo procedimiento para el denominador del coeficiente de regresión de OLS con $g'(x) = 1$. Es decir,

$$\sigma_X^2 = \mathrm{Cov}(X,X) = \int_{-\infty}^{\infty} [\mu_X F_X(x) - ALCV(x)] dx \tag{3-9}$$

y se obtiene que

$$\int w(x) dx = \frac{1}{\sigma_X^2} \int [\mu_X F_X(x) - ALCV(x)] dx = 1.$$

El esquema de ponderación depende de la distribución de la variable explicativa X. Con el fin de ilustrar el efecto que tiene el tipo de distribución de la variable explicativa en el esquema de ponderación se consideran tres ejemplos específicos: cuando X se distribuye uniforme, normal y lognormal (Yitzhaki y Schechtman, 2013).

1. **X se distribuye uniforme en el intervalo $[a, b]$.**

   Si X se distribuye uniforme o igualmente espaciada ($f(x) = \frac{1}{b-a}$), el esquema ponderado de pendientes permanece inalterado independiente de si éste es visto como una función de X o F_X, implicando que el esquema ponderado de OLS y de Gini son idénticos, es decir, los coeficientes de regresión son los mismos independientemente de la forma de la recta de regresión. De hecho, el ponderador dado por OLS y Gini en el caso de una distribución uniforme de X, en el intervalo $[a,\, b]$ aplicándolo a (3-8), es:

   $$w(x) = \frac{6(b-x)(x-a)}{(b-a)^3}.$$

   De aquí se deduce que los coeficientes de regresión estimados con Gini serían tan sensibles a datos atípicos como los estimados con OLS cuando la distribución de X es uniforme. Por otro lado, cuando la función de distribución de X no es igualmente espaciada el esquema de ponderación de OLS y de Gini son diferentes (Yitzhaki y

Schechtman, 2013). Este esquema de ponderación, que representa un caso en el que las observaciones de X están igualmente espaciadas, es simétrica alrededor de la mediana, y entre más cerca este la observación de la mediana, mayor es el peso que se le da a la observación. Una característica interesante de este esquema de ponderación es que su forma se mantiene sin cambios, independientemente de si se ve como una función de X o F_X, lo que implica que los sistemas de ponderación de las regresiones OLS y Gini son idénticos. Es decir, siempre que la distribución de la variable explicativa es uniforme (o equidistante), entonces las regresiones OLS y Gini dan el mismo coeficiente de regresión, independientemente de la forma de la curva de regresión.

2. **X se distribuye normal estándar**

   Sea X una variable que sigue una distribución normal estándar. Aplicándolo a 3-8 el ponderador sería:

$$w(x) = \frac{1}{\sqrt{2\pi}} \int_{-\propto}^{x} te^{-t^2/2} dt = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

   El peso en este caso es igual al valor de la densidad de la distribución normal estándar en X. Por lo tanto, cada percentil de la población recibe un peso igual. Este resultado ofrece una explicación intuitiva para el sistema de ponderación de OLS. Este es precisamente el caso en que la variable explicativa tiene una distribución normal, lo que explica su eficiencia en este caso.

3. **X tiene una distribución lognormal**

   Sea X una variable que tiene una distribución lognormal con parámetros μ y σ^2 (es decir, $ln(X)$ es normal con media μ y σ^2). La media de X es $\mu_X = e^{\mu + \frac{1}{2}\sigma^2}$ y la varianza de X es $\sigma_X^2 = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$.

   Es conveniente reescribir el lado derecho de la ecuación 3-8 como

$$w(x) = \frac{\mu_X}{\sigma_X^2} \left[F_X(x) - F_1(x)\right] = \frac{F_X(x)}{\sigma_X^2} \left[\mu_X - E(X|X \leq x)\right]$$

   donde $F_1(x) = \left(\frac{1}{\mu_X}\right) \int_0^x t dF_X(t)$ es la distribución acumulada del primer momento de X. La distribución acumulada del primer momento es también lognormal, con parámetros $\mu + \sigma^2$ y σ^2.

Por lo tanto, el peso en X es la diferencia entre dos distribuciones. Así que el peso para una variable X que se distribuye lognormal se puede escribir como

$$w(x) = \frac{\mu_X}{\sigma_X^2}\left[\Phi\left(\frac{\ln(x)-\mu}{\sigma}\right) - \Phi\left(\frac{\ln(x)-\mu-\sigma^2}{\sigma}\right)\right]$$

donde $\Phi()$ es la distribución acumulada normal estándar (Yitzhaki y Schechtman, 2013).

Para el caso en que X **sea una variable discreta**, la expresión 3-8 para el coeficiente de regresión de OLS se puede definir también como una suma ponderada de pendientes para cuando las observaciones en X son ordenadas de forma adyacentes.

$$\hat{\beta}_{10LS} = \sum_{i=1}^{n-1} w_i b_i, \tag{3-10}$$

donde $w_i > 0$, $\sum w_i = 1$, $b_i = \frac{\Delta y_i}{\Delta x_i}$, $\Delta x_i = x_{i+1} - x_i$ y las observaciones son ordenadas en orden creciente de X. Los pesos o ponderadores son dados por

$$w_i = \frac{(\sum\limits_{j=1}^{n-1} i(n-j)\Delta x_j + \sum\limits_{j=1}^{i-1} j(n-i)\Delta x_j)\Delta x_i}{\sum\limits_{k=1}^{n-1}(\sum\limits_{j=k}^{n-1} k(n-j)\Delta x_j + \sum\limits_{j=1}^{k-1} j(n-k)\Delta x_j)\Delta x_k} \tag{3-11}$$

donde $\sum\limits_{j=1}^{0} j(n-1)\Delta x_j = 0$.
cuando w_i se expresa en términos del ALCV el i-ésimo peso puede ser expresado como:

$$w_i = \frac{i}{n}\frac{(\bar{x}_n - \bar{x}_i)\Delta x_i}{\hat{\sigma}_X^2}, \tag{3-12}$$

donde $\bar{x}_i$ es la media de las i-ésimas observaciones más pequeñas de X, $\hat{\sigma}_x^2$ es la varianza estimada de X.
Note que $\frac{i}{n}\bar{x}_i$ sería el equivalente discreto de $ALCV$ y $\frac{i}{n}\bar{x}_n$ sería el de LOI, por lo tanto (3-12) sería la diferencia entre $ALCV$ y LOI, como en el caso continuo (Yitzhaki y Schechtman, 2013).
Los dos tipos de componentes en 3-10 son la pendiente b_i y el peso w_i. La pendiente b_i es determinada por las observaciones Y, mientras que los pesos w_i son determinados por la distribución de la variable explicativa X solamente, similar al caso continuo.

Por lo tanto se puede interpretar que la diferencia entre los estimadores de OLS y Gini se origina en el esquema de ponderación empleado. La contribución de cada observación a la estimación del coeficiente de regresión se descompone en dos partes: el efecto del esquema de ponderación w_i sobre la observación y la pendiente definida por cada par de observaciones adyacentes.
La ecuación 3-10 revela que los pesos dependen tanto del rango de la observación (i) como de la distancia en X entre observaciones adyacentes (ΔX). Similar al caso continuo, los pesos se incrementan cuando la observación se acerca a la mediana y cuando la distancia entre observaciones crece. Para controlar el efecto de (ΔX), considere el caso donde la observaciones son igualmente espaciadas, esto es, $\Delta X = c$ para todo i. Entonces (3-11) se puede simplificar como

$$w_i = \ Ki(n-i), \quad donde \quad K = \frac{6}{[n(n-1)(n+1)]}.$$

Tomando la derivada de w_i e igualándola a cero para encontrar el máximo, se encuentra que su máximo está en $i = n/2$. Esto permite ver que el esquema ponderado es una parábola simétrica alrededor de la mediana, y que las observaciones más cercanas a la mediana obtienen los mayores pesos (Yitzhaki y Schechtman, 2013).

3.3. Enfoque semi-paramétrico de la regresión Gini

El enfoque semi-paramétrico de la regresión Gini se basa en el hecho de que los coeficientes de regresión pueden ser presentados como una suma ponderada de pendientes de la recta de regresión.
Los métodos de regresión semi-paramétrico de OLS y de Gini no requieren la especificación de una forma funcional del modelo, estos pueden ser usados cuando el investigador esté interesado en estimar promedios de pendientes ponderadas sin requerir la especificación de una función formal que describa el comportamiento entre la variable dependiente y las independientes.
Los estimadores de la regresión Gini que son obtenidos bajo el enfoque semi-paramétrico serán denotados por el subíndice GN. Se hace referencia a este enfoque como semi-paramétrico porque no depende del supuesto de linealidad entre las variables ni el de la distribución de los errores.
Los coeficientes de regresión tanto en OLS como en Gini pueden ser expresados en términos de sumas ponderadas de pendientes. La única diferencia entre ambos métodos tiene que ver con las funciones ponderadoras que utiliza cada método para ponderar la pendiente (Yitzhaki y Schechtman, 2013).
A continuación se muestra como el coeficiente de regresión de Gini también puede ser expresado como una suma ponderada de pendientes de la recta de regresión (Yitzhaki y Schechtman,

2013).

$$\hat{\beta}_{GN} = \int v(x)g'(x)dx \tag{3-13}$$

donde $v(x) > 0$ denota los pesos en X, $v(x) > 0$ y $\int v(x)g'(x)dx = 1$, donde el esquema de ponderación v_i depende de la distribución de las variables explicativas X únicamente, lo cual aplica para el caso discreto y continuo.

$$v(x) = \frac{[1 - F_X(x)]F_X(x)}{\int_{-\propto}^{\propto} [1 - F_X(t)]F_X(t)dt} \tag{3-14}$$

$$\begin{aligned}&\mathrm{Cov}(Y, F_X(X)) = E_X E_Y((Y - \mu_Y)(F_X(X) - 1/2)\\ &= E_X(F_X(X) - 1/2)g(x)) = \int (F_X(x) - 1/2)g(x)f_X(x)dx\end{aligned}$$

Usando integración por partes, se tiene

$$\begin{aligned}v' &= F_X(x) - 1/2)f_X(x)dx; \quad v = -(\tfrac{1}{2})\,(1 - F_X(x) - [1 - F_X(x)]^2)\\ u &= g(x) \qquad\qquad\qquad\qquad\qquad\qquad u' = g'(x)\end{aligned}$$

Luego

$$2Cov(Y, F_X(X)) = -(\frac{1}{2})([1 - F_X(x)] - [1 - F_X(x)]^2 g(x)|_{-\infty}^{\infty} + \int (1 - F_X(x))F_X(x)g'(x)dx.$$

Para una función $g(x)$ acotada, el primer término $-(\frac{1}{2})([1-F_X(x)]-[1-F_X(x)]^2 g(x)|_{-\infty}^{\infty} = 0$. Por lo tanto,

$$2Cov(Y, F_X(X)) = \int (1 - F_X(x))F_X(x)g'(x)dx. \tag{3-15}$$

Aplicando el mismo procedimiento para el denominador con $g'(x) = 1$ se completa la demostración de la expresión 3-13

La expresión 3-13 también puede ser extendida para el caso en que X **es una variable discreta**

$$\hat{\beta}_{GN} = \sum_{i=1}^{n-1} v_i b_i$$

donde $v_i > 0$, $\sum v_i = 1$, $b_i = \frac{\Delta y_i}{\Delta X_i}$, $\Delta X_i = X_{i+1} - X_i$ y las observaciones son ordenadas en orden creciente de acuerdo a X. Los pesos asignados a cada pendiente b_i están dados por la siguiente expresión

$$v_i = \frac{(n-i)i\Delta x_i}{\sum\limits_{k=1}^{n-1} [(n-k)k]\Delta x_k}.$$

3.3.1. Comparación del esquema de ponderación de OLS y Gini

Como se ha visto hasta este momento los estimadores para ambos métodos de regresión Gini y OLS pueden ser expresados como suma de promedios de pendientes ponderados, debido a que las pendientes entre observaciones adyacentes pueden ser determinadas por los datos. En esta sección se hace un análisis del esquema de ponderación de OLS y Gini, el cual depende de dos factores a saber:

1. El primer factor que afecta el esquema de ponderación por Gini y OLS tiene que ver con el rango de la observación (i), el máximo peso es asignado a las observaciones que están ubicadas alrededor de la mediana de la variable explicativa, y entonces el peso asignado a cada observación disminuye simétricamente a medida que la observación se aleja de la mediana. Esta propiedad es compartida por el esquema de ponderación de OLS (3-9) y de Gini (3-14). Por tanto en el método de estimación por Gini como en OLS el efecto de ponderación para cada observación i es el mismo: el mayor peso lo obtienen las observaciones cercanas a la mediana de X.

2. El segundo factor que afecta el esquema de ponderación pora Gini y OLS tiene que ver con la distancia entre las observaciones adyacentes, la cual es denotada por ΔX. La diferencia entre el esquema de ponderación de ambos método OLS y Gini tiene que ver con el peso que le asigna a cada valor atípico con ΔX. Aunque el peso en la regresión de Gini se basa en la misma distancia ΔX, el peso en OLS se basa en $(\Delta X)^2$. Esta diferencia explica el hecho de que OLS es más sensible a los valores atípicos que el método de estimación de Gini. Nótese que el supuesto de linealidad no juega ningún rol en la estimación semi-paramétrica de los coeficientes de regresión de OLS y Gini debido a que son expresados como promedio ponderado de observaciones adyacentes (Yitzhaki y Schechtman, 2013).

A continuación se presenta un análisis gráfico, las gráficas son obtenidas vía simulación, cuyo objetivo es comparar los ponderadores de Gini con los de OLS para cuando X tiene una distribución uniforme.
La figura **3-1** compara las funciones ponderadoras w_x de OLS y v_x de Gini, en función del rango de X, (i), para $\Delta X = c = 1$.

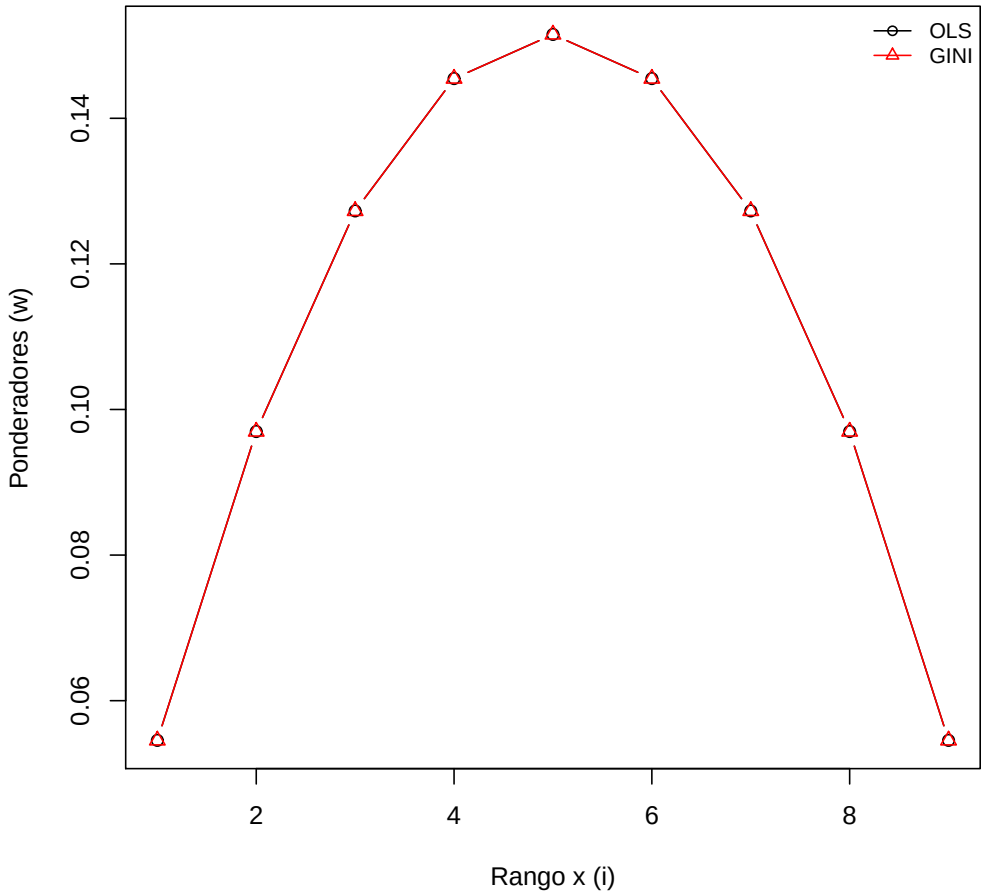


Figura 3-1: Ponderadores de OLS y Gini para X con distribución uniforme ($\Delta X = c = 1$).

De la figura **3-1** se observa que los pesos asignados por Gini y OLS son los mismos, haciendo que ambos métodos sean igualmente sensibles a datos atípicos (Yitzhaki y Schechtman, 2013).

La figura **3-2** compara las funciones ponderadoras w_x de OLS y v_x de Gini en función del diferencial en X, ΔX.

Se observa que para un par de datos adyacentes(X_i, X_{i+1}) pero alejados entre sí (ΔX grande), la ponderación dada a dicha pendiente (b_i) por el método Gini es menor al de OLS, lo cual permite confirmar que el método de estimación de Gini es más robusto a datos atípicos que OLS cuando la distribución de X no es uniforme.

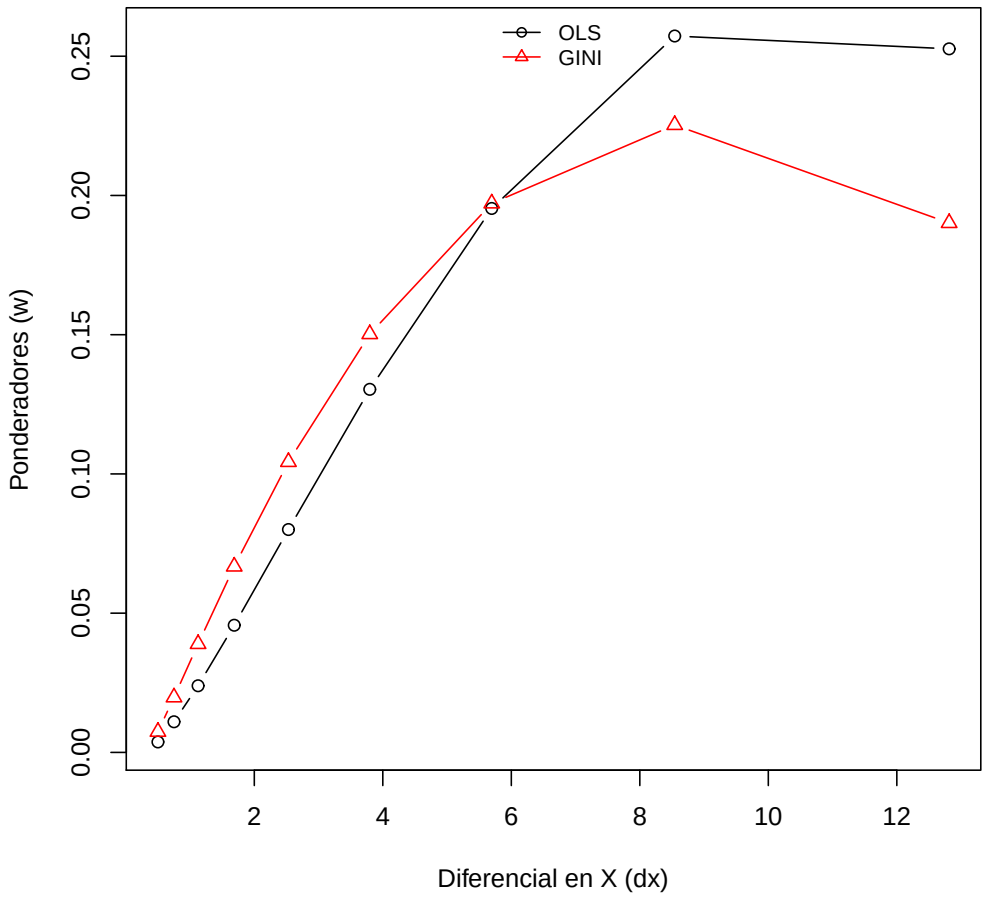


Figura 3-2: Ponderadores de OLS y Gini para una distribución de X no uniforme.

3.4. Enfoque de minimización de Gini

El enfoque de minimización de Gini se basa en la minimización de la GMD de los residuales. Para poder optimizar se tiene que especificar el modelo y la función objetivo, lo cual significa en este caso asumir que el modelo es lineal. El enfoque es similar a la regresión de desviación mínima absoluta (LAV) o la regresión cuantil (Bassett y Koenker, 1978). Aquí, en vez de minimizar la suma de desviaciones absoluta de los residuales o la suma ponderada de las desviaciones absolutas de un cuantil de los residuales, se minimiza la GMD de los residuales (la media de las diferencias absolutas entre todos los pares de residuales). Los parámetros y estimadores derivados siguiendo el enfoque de minimización serán denotados por el sub índice GM. (Este tipo de regresiones han sido desarrollados por Jureckova(1969, 1971); Jaeckel (1972); McKean y Hettmansperger (1978); y Hettmansperger (1984))). Hettmansperger hace

referencia a este método como la regresión R debido a que se minimiza la suma de los productos de los rangos de los residuales y los residuales.
Considere el modelo

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Note que no se ha impuesto ningún supuesto sobre ε.
El β_1 que minimiza la GMD de los residuales forma la ecuación normal del tipo $\text{cov}(X, F(\varepsilon)) = 0$. Considere el residual ordenado por $\varepsilon_1 \leq \varepsilon_2 \leq ... \leq \varepsilon_n$. Entonces

$$G_\varepsilon = \sum_{i,j} |(y_j - y_i) - \beta_1(x_j - x_i)| = 2\sum_{i<j} [(y_j - y_i) - \beta_1(x_j - x_i)].$$

Del cual se obtiene que

$$\frac{\partial G_\varepsilon}{\partial \beta_1} = -2\sum_{i<j}(x_j - x_i) = 4\sum_{i=1}^{n} x_i \left[i - \frac{n+1}{2}\right] = 4Cov\left(x, \frac{R_\varepsilon}{n}\right)$$

donde R_ε es el rango de ε (aquí R_ε/n es la función de distribución empírica acumulada de ε). Recuerde que los residuales son ordenados. Aquí el rango de ε_i es i. En el mínimo la derivada (si existe) es igual a cero, lo cual se cumple si $\text{cov}\left(x, \frac{R_\varepsilon}{n}\right) = 0$.
Esto es una restricción de ortogonalidad (no hay correlación entre el ε y x, $cov(x, F(\varepsilon)) = 0$).
Similar a la regresión cuantil el estimador $\hat{\beta}$ no puede ser expresado explícitamente y debe ser calculado con métodos numéricos.

4 Metodología y Resultados

4.1. Escenarios de simulación para comparar los métodos de Regresión

De acuerdo al enfoque no paramétrico (o semi-paramétrico) de Gini para calcular los coeficientes de regresión, estos son influenciados por observaciones atípicas tanto en el eje de las X como en el de las Y. En consecuencia, en este trabajo de tesis se van a simular cuatro escenarios, como se ilustra en la figura **4-1**, con el fin de influenciar la estimación de los coeficientes de regresión aumentando sistemáticamente la magnitud del dato atípico, el cual es atípico simultáneamente en X y en Y.
Para el modelo de regresión lineal simple $Y{=}\beta_0{+}\beta_1 X{+}\varepsilon$, donde los errores ε se generan así:

1. Escenario 0: En este escenario se genera n errores normales con media cero y varianza constante, es decir, $\varepsilon \sim N(\mu = 0, \sigma^2 = 1)$.

2. Normalidad de los errores pero contaminando la última observación bajo diferentes magnitudes, a saber:

   2.1. Escenario 1: En este escenario se toman los mismos errores que fueron generados en el escenario 0, los cuales fueron generados bajo normalidad con media cero y varianza constante, pero se contamina la última componente del vector Y a 4σ desde la recta ajustada. Ver gráfico **4-1**.

   2.2. Escenario 2: En este escenario se toman los mismos errores que fueron generados en el escenario 0, los cuales fueron generados bajo normalidad con media cero y varianza constante, pero se contamina la última componente del vector Y a 8σ desde la recta ajustada. Ver gráfico **4-1**.

   2.3. Escenario 3: En este escenario se toman los mismos errores que fueron generados en el escenario 0, los cuales fueron generados bajo normalidad con media cero y varianza constante, pero se contamina la última componente del vector Y a 16σ desde la recta ajustada. Ver gráfico **4-1**.

En cuanto a la distribución de la variable explicativa X, ésta se genera igualmente espaciada o equidistante (tomando diferenciales de 1, es decir, $\Delta X = 1$) para las primeras $n-1$ observaciones y la última observación (n) se aleja ($n + (n/5)$) unidades de la penúltima

$(n-1)$, con el objetivo de resaltar la diferencia entre el ponderador de Gini y el de OLS; esto se realiza así dado que Gini es más robusto que OLS cuando hay datos atípicos en el eje de la X.

4.2. Pasos para comparar los métodos de Regresión

1. **Para el escenario de normalidad (Escenario 0) se siguen los siguientes pasos:**

   1.1. Generar las observaciones de la variable explicativa X de tal manera que esté igualmente espaciada o equidistante para las primeras $n-1$ observaciones, es decir, a un diferencial $\Delta X = 1$, y la última observación se aleja $n/5$. Es decir, $X = 1, 2, 3, ..., (n-1), (n+n/5)$.

   1.2. Generar n errores normales independientes con media $\mu = 0$ y varianza $\sigma^2 = 1$, es decir, $\varepsilon \sim N(\mu, \sigma^2)$.

   1.3. Generar la variable respuesta Y con el modelo lineal teórico $Y = \beta_0 + \beta_1 X + \varepsilon$, donde $\beta_0 = 5$ y $\beta_1 = 20$; estos valores son tomados a criterio propio.

   1.4. Estimar los 5 parámetros: coeficientes de regresión β_0 y β_1, la varianza (la cual es estimada con el MSE), el coeficiente de determinación (R^2) y el estadístico muestral χ^2 para cada uno de los métodos de regresión (*Gini*, *OLS*, *LAV*, *M de Huber*, *MM*) bajo cada uno de los escenrarios (normalidad y normalidad contaminada).

   1.5. Repetir los pasos del 1,1 a 1,4 para $N = 1000$ veces y almacenar cada resultado en una matriz de $N \times 5$ por cada uno de los métodos de regresión, es decir, en 5 matrices.

   1.6. Comparar la distribución empírica del estadístico χ^2 con la distribución teórica χ^2_2 usando la prueba de bondad de ajuste de Kolmogorov- Smirnov.

   1.7. Promediar por columnas los resultados almacenados en la matriz de $N \times 5$ para cada uno de los métodos de regresión.

   1.8. Ejecutar el escenario de normalidad (Escenario 0) para n_1=10, n_2=30 y n_3=100. Estos valores son elegidos a criterio propio.

2. **Para los escenarios de normalidad contaminando con datos atípicos se siguen los mismos pasos del paso 1, pero con las siguientes modificaciones:**

   2.1. Escenario 1: En este escenario se toman los mismos errores que fueron generados en el escenario 0, los cuales fueron generados bajo normalidad con media cero y varianza constante, pero se contamina la última componente del vector Y a 4σ.

   2.2. Escenario 2: En este escenario se toman los mismos errores que fueron generados en el escenario 0, los cuales fueron generados bajo normalidad con media cero y varianza constante, pero se contamina la última componente del vector Y a 8σ.

2.3. Escenario 3: En este escenario se toman los mismos errores que fueron generados en el escenario 0, los cuales fueron generados bajo normalidad con media cero y varianza constante, pero se contamina la última componente del vector Y a 16σ.

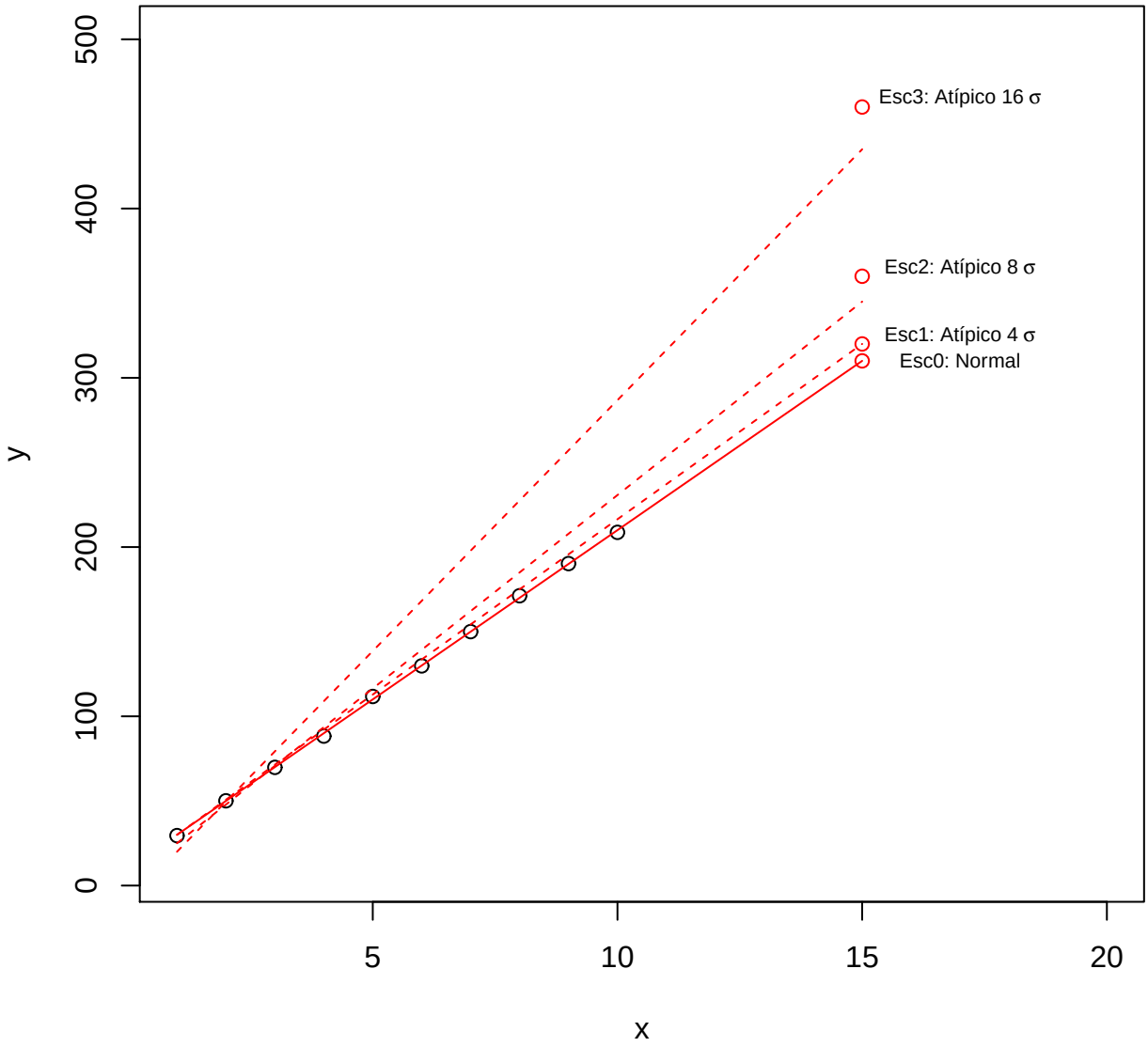


Figura 4-1: Escenarios de simulación

La robustez de los estimadores β_0 y β_1 ante la presencia de datos atípicos obtenidos por cada método se mide mediante el siguiente estadístico (Johnson y Wichern, 1992):

$$\left(\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}\right)' \Sigma_{\underline{\hat{\beta}}}^{-1} \left(\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}\right) \sim \chi_2^2$$

otra forma de escribirlo es:

$$\left(\hat{\beta} - \beta\right)' \frac{1}{\sigma^2} \left(X'X\right) \left(\hat{\beta} - \beta\right) \sim \chi_2^2 \tag{4-1}$$

donde $\hat{\beta}$=($\hat{\beta}_0$, $\hat{\beta}_1$) es el estimador del vector de coeficientes teórico β=(β_0, β_1).

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}; \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix} \quad y \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

y $\sigma^2 = 1$.

Se espera que el estadístico muestral calculado (χ^2) se aleje de su valor esperado $E(\chi_2^2) = 2$ a medida que aumenta la magnitud del dato atípico, indicando la sensibilidad del método.

4.3. Resultados de la simulación

A continuación se presentan las tablas de resultados obtenidas mediante simulación Montecarlo descrita en la sección anterior. Cada tabla contiene los parámetros estimados en cada escenario (escenario de normalidad y normalidad con datos contaminados), cada método de regresión y cada tamaño muestral. Para el siguiente análisis es importante recordar que el modelo teórico bajo el cual se realizaron las simulaciones para la comparación de los métodos es: Y=β_0+$\beta_1 X$+ε; con $\beta_0 = 5$, $\beta_1 = 20$ y $\varepsilon \sim N(\mu, \sigma^2)$.

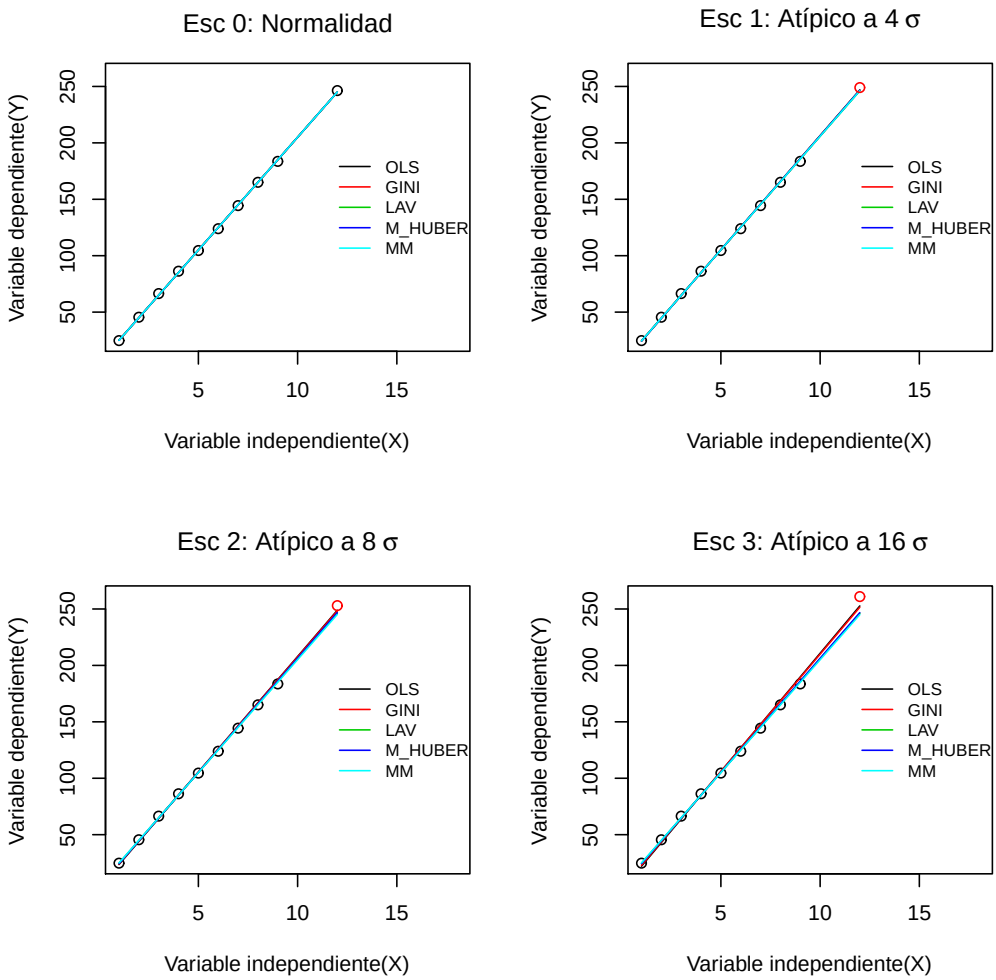


Figura 4-2: Modelo de regresión estimado por cada método en cada escenario para $n = 10$.

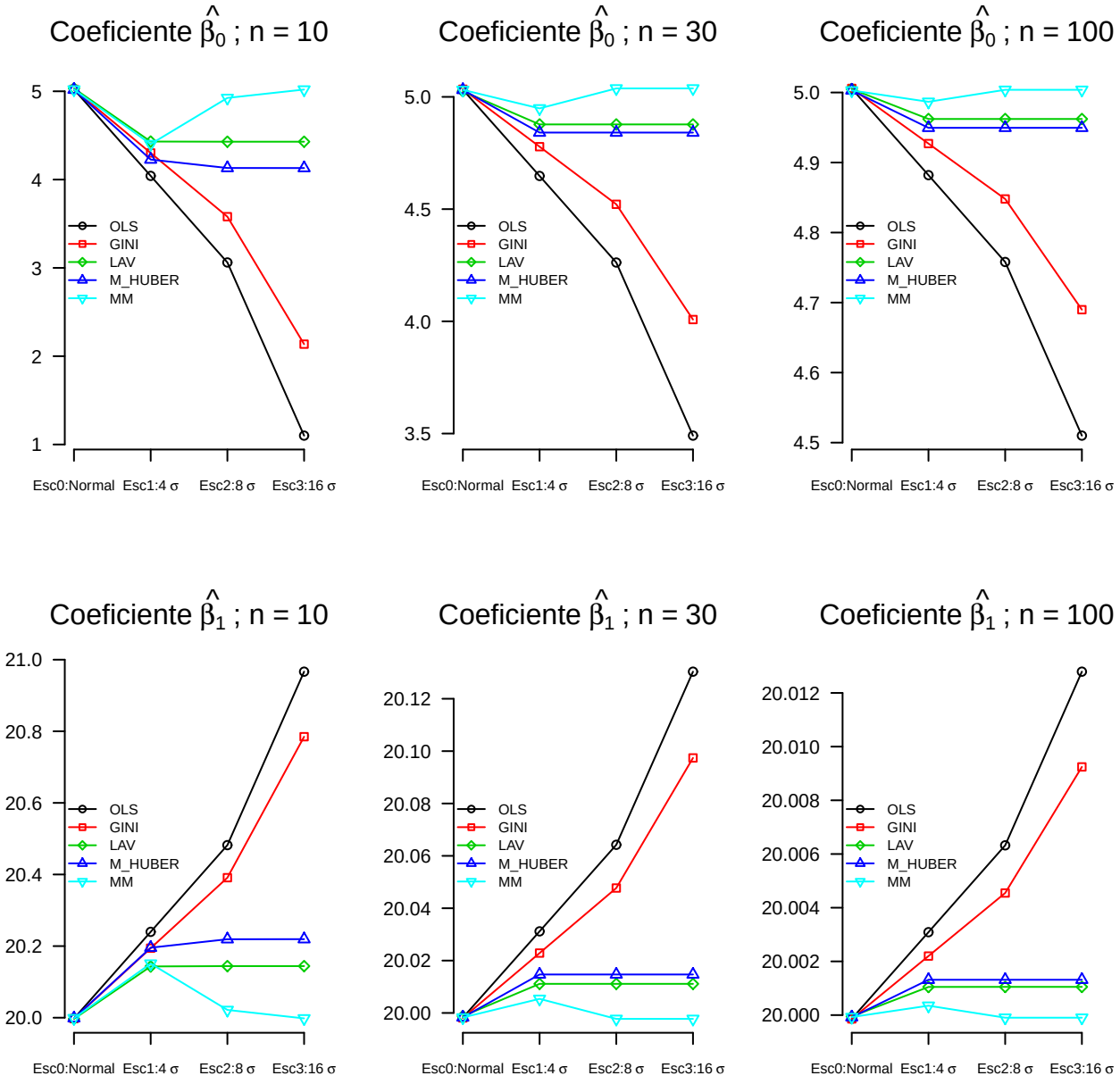


Figura 4-3: Parámetros β_0 y β_1 estimados por cada método con diferentes tamaños de muestra en cada escenario.

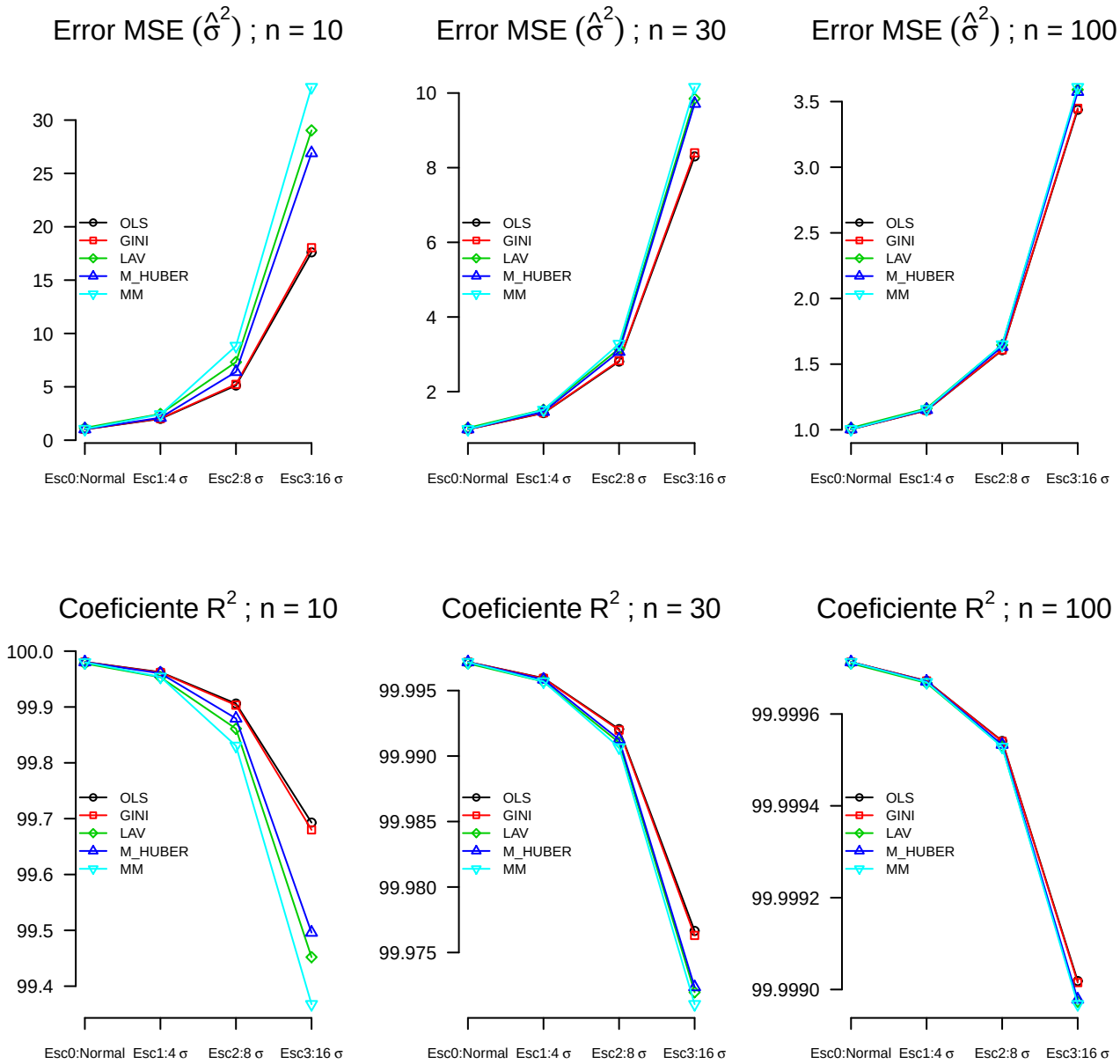


Figura 4-4: Parámetros MSE y R^2 estimados por cada método con diferentes tamaños de muestra en cada escenario.

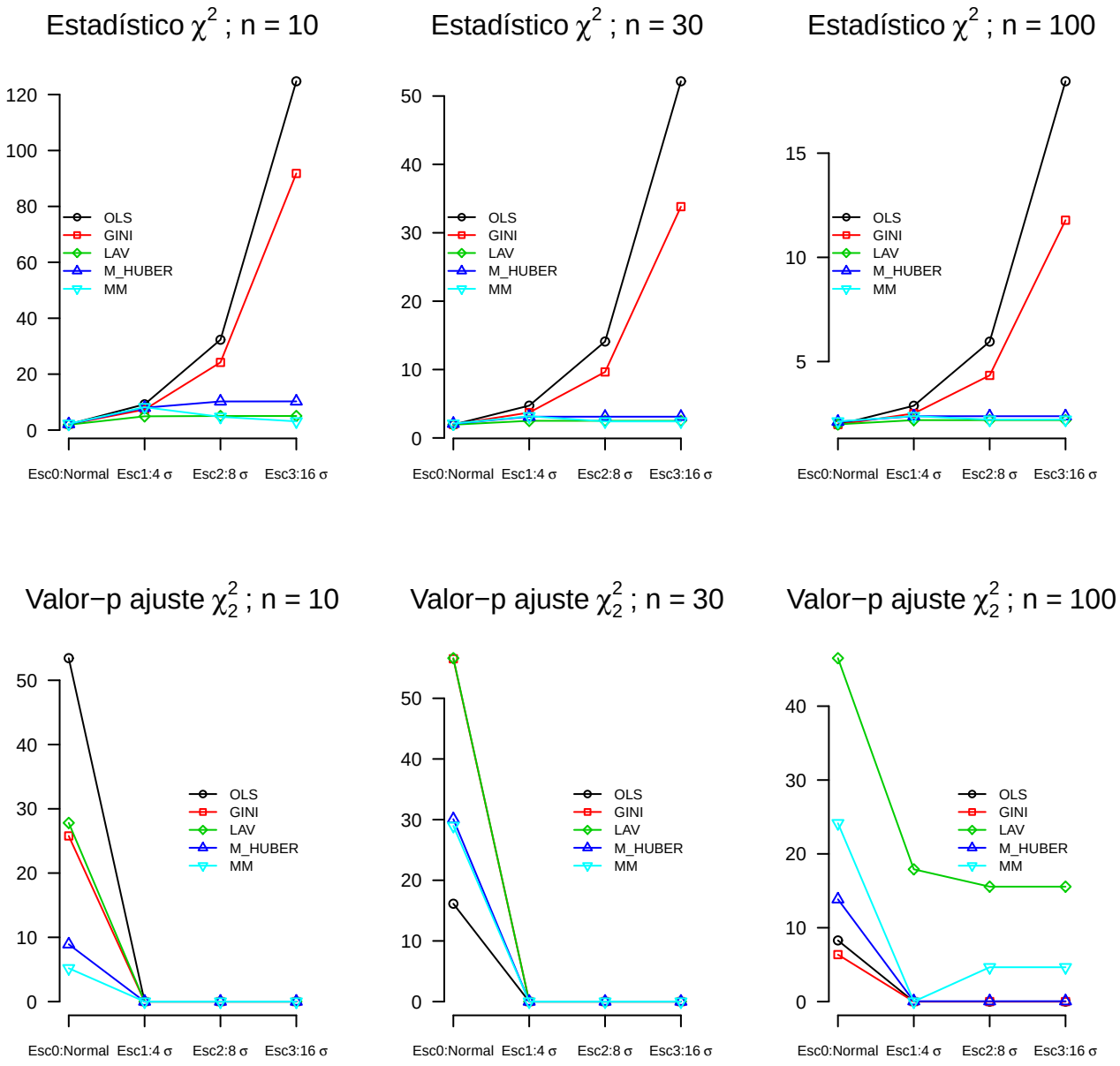


Figura 4-5: Estadístico χ^2 y Vp (del ajuste de la distribución χ^2_2) obtenidos por cada método con diferentes tamaños de muestra en cada escenario.

- Para el escenario de normalidad del error, todos los métodos robustos MM, M Huber, LAV y Gini son tan buenos como OLS para estimar los parámetros de la regresión (β_0, β_1) y los estadísticos MSE y R^2, como era de esperarse por teoría. Ver escenario 0 de normalidad en las gráficas **4-3** y **4-4**.

- Para el escenario de normalidad contaminada con un dato atípico a 4σ, 8σ y 16σ, las

estimaciones de β_0 y β_1 realizadas por OLS y Gini fueron influenciadas proporcionalmente a la magnitud del atípico, alejándose de su valor teórico de $\beta_0 = 5$ y $\beta_1 = 20$. Pero Gini es afectado en menor proporción que OLS, mientras que por LAV, M Huber y MM el estimador solamente fue influenciado al principio, luego sigue produciendo las mismas estimaciones sin importar que tanto se aumente la magnitud del atípico. Ver gráfica **4-3**.

- Un análisis sencillo de cuán robusto son los métodos se presenta en la Tabla **4-19** y **4-20**, donde se calcula la variación que sufre el intercepto $\hat{\beta}_0$ y la pendiente $\hat{\beta}_1$ con respecto al escenario de normalidad, calculada de la siguiente forma:

$$\%Variación = \frac{\hat{\beta}_{Normalcontaminada} - \hat{\beta}_{Normal}}{\hat{\beta}_{Normal}}$$

  Es así que $\hat{\beta}_0$ y $\hat{\beta}_1$ sufren una variación de menor a mayor al ser estimados con MM, LAV, M Huber, Gini y OLS, respectivamente.

- El Error Cuadrático Medio, MSE, se hace más grande a medida que aumenta la magnitud del atípico, pero es mayor en los métodos que presenta la mayor robustez ante la presencia de atípicos, indicando que la recta ajustada por estos métodos tiende a permanecer inmóvil. Consecuentemente sucede lo contrario con el coeficiente de determinación R^2, el cual se hace más pequeño a medida que aumenta la magnitud del atípico, pero es menor en los métodos más robustos. Con el análisis de MSE y del R^2 se puede empezar a cuantificar la robustez de los métodos y establecer un orden. Es así que se puede clasificar de mayor a menor robustez a MM, LAV, M Huber, Gini y OLS, respectivamente. Ver gráfica **4-4**.

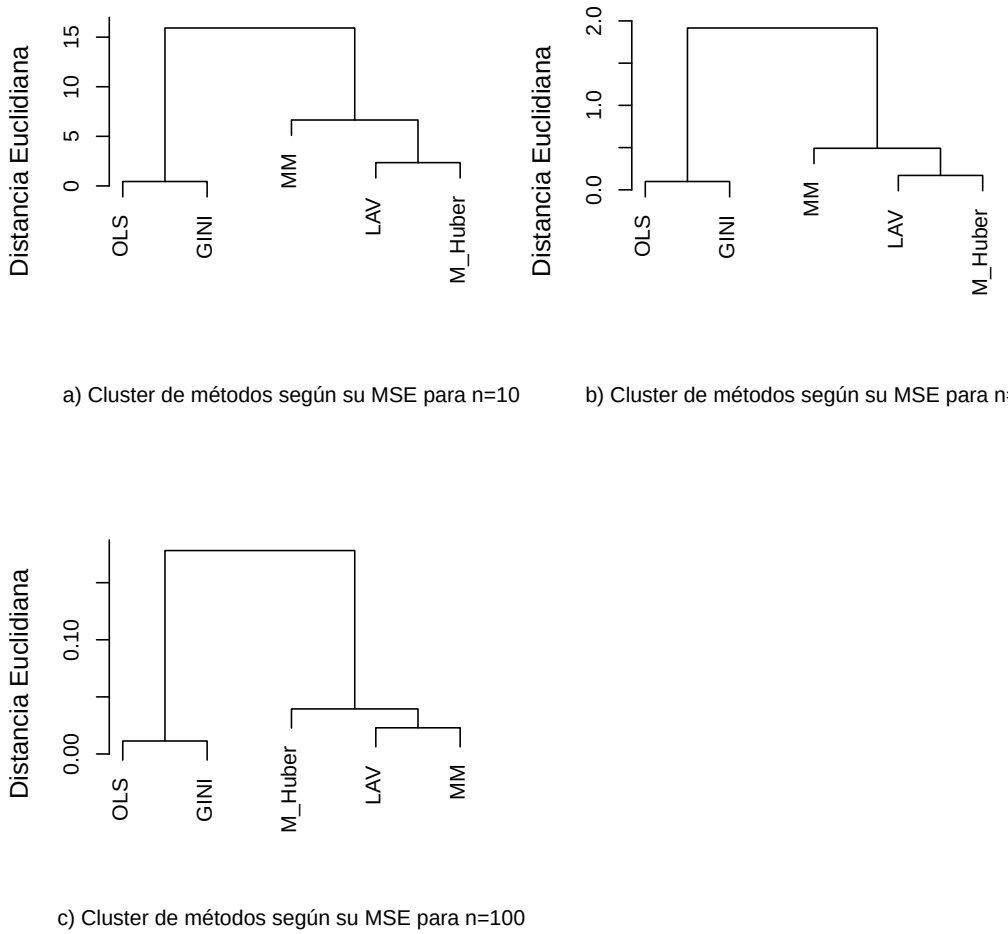
- Otro criterio propuesto en este trabajo para medir la robustez es comparar la distribución de los estimadores de los coeficientes de regresión $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ con la distribución χ^2_2 de la sigueinte forma:

$$\left(\hat{\beta} - \beta\right)' \frac{1}{\sigma^2} \left(X'X\right) \left(\hat{\beta} - \beta\right) \sim \chi^2_2 \tag{4-2}$$

  la cual es cierta cuando los errores se distribuyen normales. Los resultados demuestran que en el escenario de normalidad de los errores, el estadístico calculado efectivamente tiende a 2 y la prueba de bondad de ajuste no rechaza la hipótesis de que estos proviene de una distribución χ^2_2, según lo corrobora el valor p, el cual es mayor a un $\alpha = 5\,\%$. Pero en los escenarios de normalidad contaminada con un atípico, el estadístico calculado se aleja de su valor esperado que es 2, pero se aleja mucho más en OLS y Gini que en los demás métodos. Este criterio permite establecer el mismo orden de robustez que el establecido con el MSE y el R^2: MM, LAV, M Huber, Gini y OLS.

- Al aumentar el tamaño de la muestra a 10, 30 y 100 observaciones, la influencia del atípico se hace menor y las estimaciones se acercan más a los valores teóricos.

- Con el fin de establecer cuáles métodos son más similares entre sí al momento de estimar los coeficientes de regresión en los diferentes escenarios, se propone agruparlos por su MSE producido tanto en el escenario de normalidad como en el de normalidad contaminada con atípicos. Los resultados muestran que OLS y Gini producirán resultados similares cuando los errores sean normales o cuando estén contaminados, tanto para muestras pequeñas como para grandes, mientras que LAV y M Huber serán más parecidos para muestras pequeñas, pero para muestras grandes LAV se asemejará más a MM. Ver gráfica **4-6**



a) Cluster de métodos según su MSE para n=10

b) Cluster de métodos según su MSE para n=30

c) Cluster de métodos según su MSE para n=100

Figura 4-6: Cluter de los Métodos según su MSE para $n = 10$, $n = 30$,$n = 100$.

4.3.1. Tablas de resultados.

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	5,01	4,88	4,76	4,51
GINI	5,01	4,93	4,85	4,69
LAV	5,00	4,96	4,96	4,96
MHuber	5,00	4,95	4,95	4,95
MM	5,00	4,99	5,00	5,00

Tabla 4-1: Estimación de β_0 con $n = 100$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	20,00	20,00	20,01	20,01
GINI	20,00	20,00	20,00	20,01
LAV	20,00	20,00	20,00	20,00
MHuber	20,00	20,00	20,00	20,00
MM	20,00	20,00	20,00	20,00

Tabla 4-2: Estimación de β_1 con $n = 100$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	1,00	1,15	1,61	3,44
GINI	1,00	1,15	1,61	3,45
LAV	1,01	1,16	1,65	3,59
MHuber	1,00	1,15	1,63	3,57
MM	1,00	1,16	1,65	3,61

Tabla 4-3: Estimación del MSE con $n = 100$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	100,00	100,00	100,00	100,00
GINI	100,00	100,00	100,00	100,00
LAV	100,00	100,00	100,00	100,00
MHuber	100,00	100,00	100,00	100,00
MM	100,00	100,00	100,00	100,00

Tabla 4-4: Estimación del R^2 con $n = 100$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	8,28	0,00	0,00	0,00
GINI	6,36	0,00	0,00	0,00
LAV	46,50	17,91	15,56	15,56
MHuber	13,85	0,05	0,05	0,05
MM	24,15	0,01	4,65	4,65

Tabla 4-5: Valor P con $n = 100$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	1,97	2,88	5,96	18,45
GINI	1,98	2,51	4,33	11,78
LAV	1,99	2,19	2,19	2,19
MHuber	2,11	2,38	2,38	2,38
MM	2,13	2,38	2,21	2,21

Tabla 4-6: Estadístico calculado χ^2_2 con $n = 100$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	5,03	4,65	4,26	3,49
GINI	5,03	4,78	4,52	4,01
LAV	5,03	4,88	4,88	4,88
MHuber	5,03	4,84	4,84	4,84
MM	5,03	4,95	5,04	5,04

Tabla 4-7: Estimación de β_0 con $n = 30$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	20,00	20,03	20,06	20,13
GINI	20,00	20,02	20,05	20,10
LAV	20,00	20,01	20,01	20,01
MHuber	20,00	20,01	20,01	20,01
MM	20,00	20,01	20,00	20,00

Tabla 4-8: Estimación de β_1 con $n = 30$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	0,99	1,42	2,80	8,30
GINI	0,99	1,43	2,83	8,40
LAV	1,03	1,52	3,15	9,85
MHuber	0,99	1,46	3,07	9,71
MM	0,99	1,51	3,28	10,16

Tabla 4-9: Estimación del MSE con $n = 30$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	100,00	100,00	99,99	99,98
GINI	100,00	100,00	99,99	99,99
LAV	100,00	100,00	99,99	99,97
MHuber	100,00	100,00	99,99	99,97
MM	100,00	100,00	99,99	99,97

Tabla 4-10: Estimación del R^2 con $n = 30$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	16,13	0,00	0,00	0,00
GINI	56,52	0,00	0,00	0,00
LAV	56,62	0,00	0,00	0,00
MHuber	30,08	0,00	0,00	0,00
MM	28,97	0,00	0,00	0,00

Tabla 4-11: Valor P con $n = 30$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	1,95	4,72	14,10	52,16
GINI	1,97	3,71	9,65	33,82
LAV	1,94	2,50	2,50	2,50
MHuber	2,08	3,10	3,10	3,10
MM	2,10	3,17	2,41	2,41

Tabla 4-12: Estadístico calculado χ^2_2 con $n = 30$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	5,02	4,04	3,06	1,10
GINI	5,02	4,30	3,58	2,14
LAV	5,03	4,43	4,43	4,43
MHuber	5,02	4,23	4,13	4,13
MM	5,02	4,40	4,92	5,02

Tabla 4-13: Estimación de β_0 con $n = 10$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	20,00	20,24	20,48	20,97
GINI	20,00	20,19	20,39	20,78
LAV	20,00	20,14	20,14	20,14
MHuber	20,00	20,20	20,22	20,22
MM	20,00	20,15	20,02	20,00

Tabla 4-14: Estimación de β_1 con $n = 10$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	1,01	2,00	5,12	17,61
GINI	1,02	2,03	5,23	18,04
LAV	1,15	2,46	7,32	29,04
MHuber	1,03	2,10	6,40	26,91
MM	1,04	2,42	8,83	33,09

Tabla 4-15: Estimación del MSE con $n = 10$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	99,98	99,96	99,91	99,69
GINI	99,98	99,96	99,92	99,74
LAV	99,98	99,95	99,86	99,45
MHuber	99,98	99,96	99,88	99,50
MM	99,98	99,95	99,83	99,37

Tabla 4-16: Estimación del R^2 con $n = 10$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	53,44	0,00	0,00	0,00
GINI	25,78	0,00	0,00	0,00
LAV	27,81	0,00	0,00	0,00
MHuber	8,91	0,00	0,00	0,00
MM	5,19	0,00	0,00	0,00

Tabla 4-17: Valor P con $n = 10$

Métodos	Normalidad	4*σ	8*σ	16*σ
OLS	2,06	9,22	32,31	124,76
GINI	2,10	7,32	24,20	91,78
LAV	1,91	4,92	5,03	5,03
MHuber	2,13	7,97	10,23	10,25
MM	2,23	8,18	4,74	3,10

Tabla 4-18: Estadístico calculado χ^2_2 con $n = 10$

Métodos	4*σ	8*σ	16*σ
OLS	-19,5 %	-39,0 %	-78,1 %
GINI	-14,4 %	-28,7 %	-57,5 %
LAV	-12,0 %	-12,0 %	-12,0 %
MHuber	-15,8 %	-17,7 %	-17,7 %
MM	-12,3 %	-1,9 %	0,0 %

Tabla 4-19: Variación de la estimación de β_0 con $n = 10$

Métodos	4*σ	8*σ	16*σ
OLS	1,2 %	2,4 %	4,8 %
GINI	1,0 %	2,0 %	3,9 %
LAV	0,7 %	0,7 %	0,7 %
MHuber	1,0 %	1,1 %	1,1 %
MM	0,8 %	0,1 %	0,0 %

Tabla 4-20: Variación de la estimación de β_1 con $n = 10$

5 Ejemplo de aplicación a datos reales

Con el objetivo de aplicar cada uno de los métodos de regresión MM, M Huber, LAV, Gini y OLS, se analiza el precio de venta (Y) de vehículos usados Renault Twingo en función de los kilómetros recorridos (X), cuyos modelos oscilan entre el año 2003 y el 2014[1]. Luego de depurar los datos originales (mostrado en el escenario 1 de la figura **5-1**), y quitar los valores atípicos, el modelo de regresión lineal ajustado por OLS es (ver escenario 0 de la figura **5-1**): $\hat{Y}$=18,57-0,063X, es decir que $\hat{\beta}_0 = 18{,}57$ y $\hat{\beta}_1 = -0{,}063$, lo cual significa que por cada mil kilómetros recorridos por un Automóvil Twingo, éste se deprecia en promedio 63 mil pesos. Para efectos de comparación de la robustez de los métodos, se modelaron dos escenarios más en los cuales el dato se desplaza a lo largo del rango de X: en el escenario 2 el dato atípico en Y es desplazado hasta el valor máximo de X con el fin de volverlo dato influyente; y en el escenario 3 el dato atípico se desplaza hasta un valor atípico de X con el fin de evaluar la robustez de Gini con respecto a OLS y a los otros métodos debido a que Gini teóricamente presentaría una mayor robustez si el dato es atípico en Y y en X.

[1]fuente: http://www.tucarro.com.co, *tomados el 15 de agosto del año 2014*

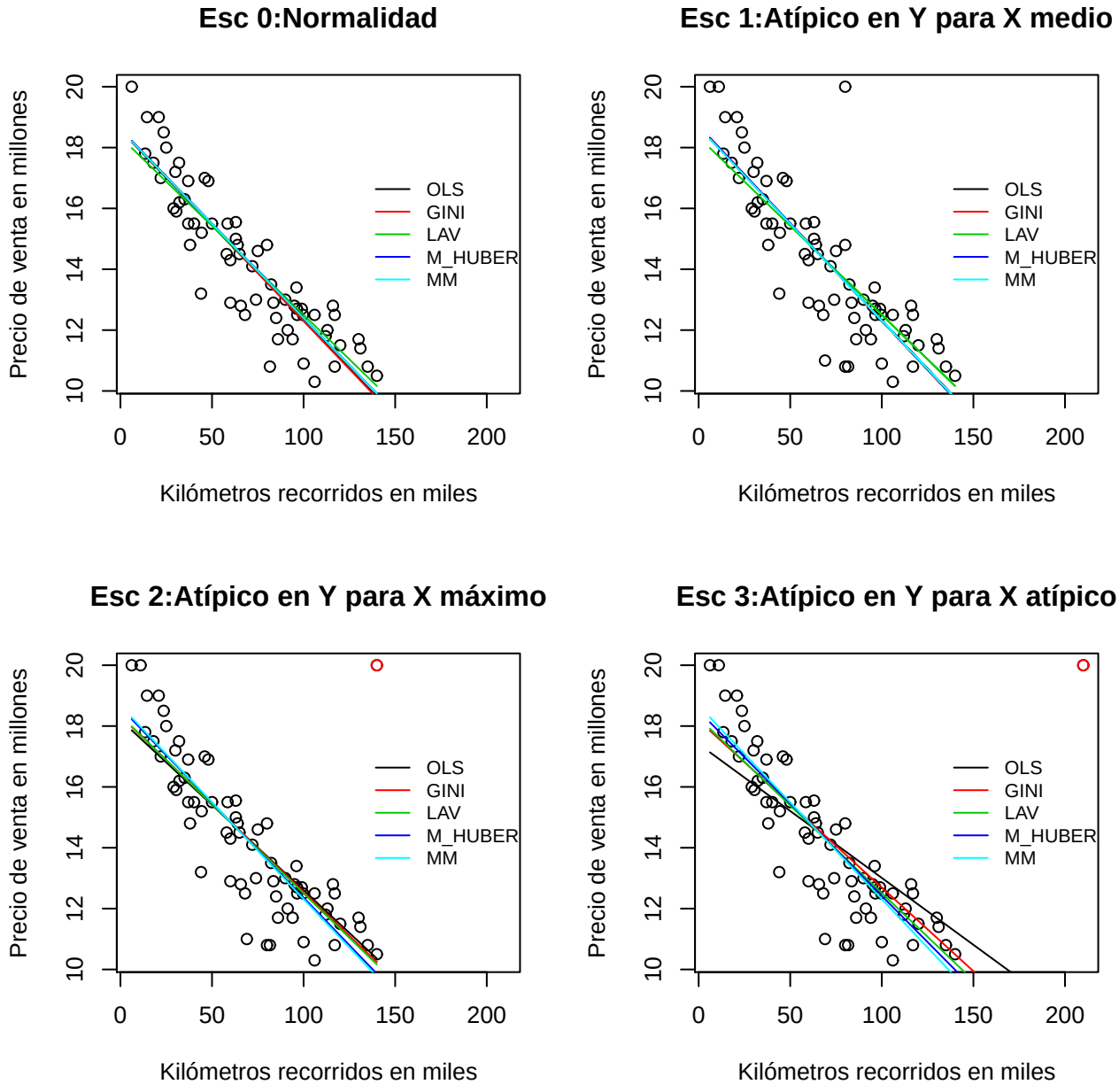


Figura 5-1: Análisis de regresión para precios de venta de Automóviles Twingo por diferentes métodos Robustos.

A continuación se muestra los parámetros $\hat{\beta}_0$, $\hat{\beta}_1$, y los estadísticos MSE y el R^2 calculados por cada uno de los métodos de regresión estudiados en este trabajo de tesis, para cada escenario y para un tamaño de muestra de $n = 68$ datos.

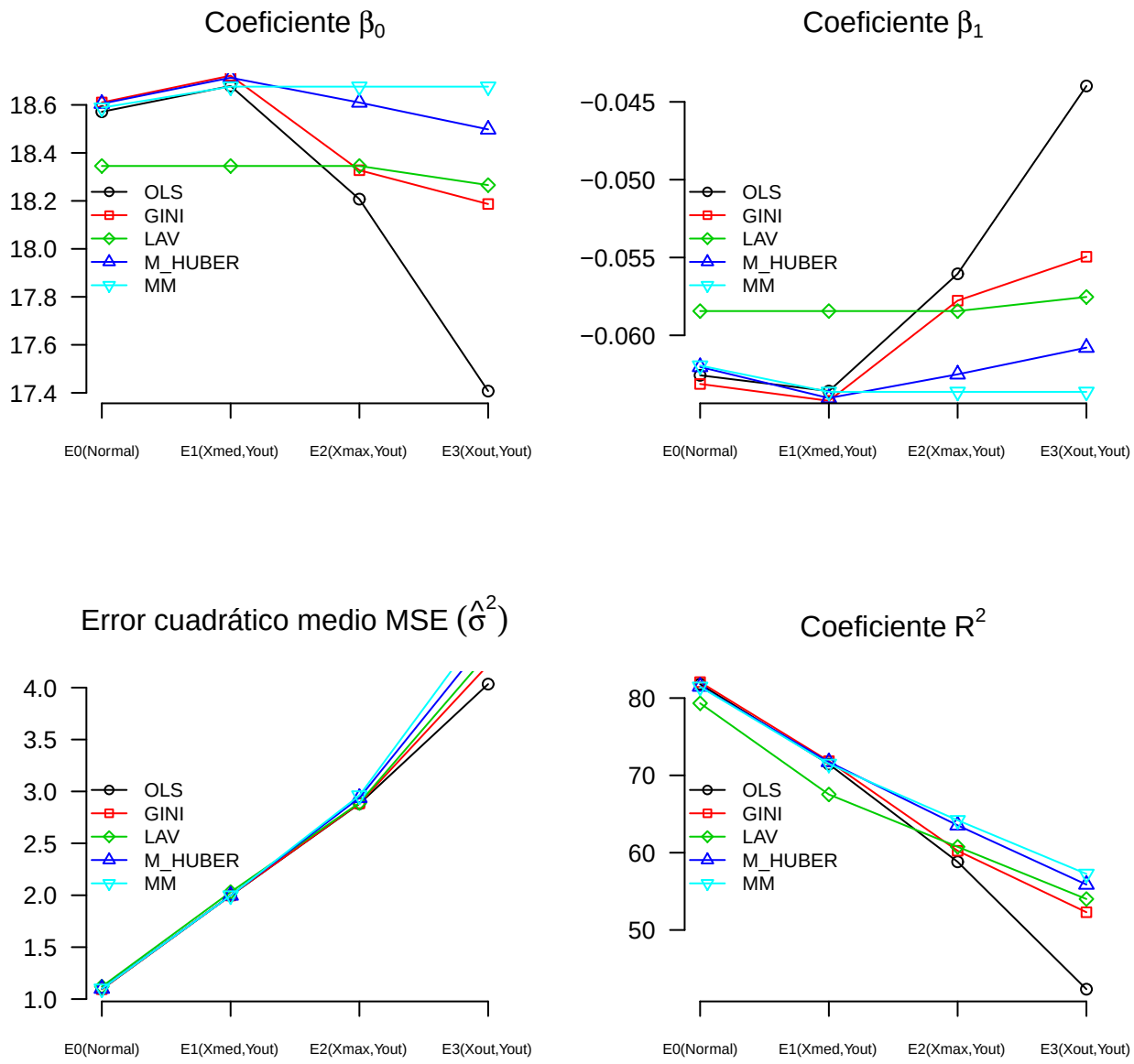


Figura 5-2: Tendencia de los parámetros estimados por cada método según el escenario.

- Para el escenario de normalidad del error, todos los métodos robustos MM, M Huber, LAV y Gini son tan buenos como OLS para estimar los parámetros de la regresión ($\hat{\beta}_0$, $\hat{\beta}_1$), y para calcular los estadísticos MSE y R^2, como era de esperarse por teoría. Ver escenario 0 de normalidad en la gráfica **5-2**.

- En los escenarios de normalidad contaminada, en los cuales el dato atípico se aleja a lo largo de la X, las estimaciones de $\hat{\beta}_0$ y $\hat{\beta}_1$ realizadas por OLS y Gini fueron influenciadas

proporcionalmente, alejándose de su valor teórico de $\hat{\beta}_0 = 18{,}57$ y $\hat{\beta}_1 = -0{,}063$. Pero Gini es afectado en menor proporción que OLS, sobre todo cuando se tiene un atípico simultáneamente en X y en Y (escenario 3). Mientras que por LAV, M Huber y MM el estimador solamente fue influenciado al principio, luego sigue produciendo las mismas estimaciones.

- Un análisis sencillo de cuán robusto son los métodos se presenta en la Tabla **5-5** y **5-6**, donde se calcula la variación que sufre el intercepto $\hat{\beta}_0$ y la pendiente $\hat{\beta}_1$ con respecto al escenario de normalidad, calculada de la siguiente forma:

$$\%Variacion = \frac{\hat{\beta}_{Normalcontaminada} - \hat{\beta}_{Normal}}{\hat{\beta}_{Normal}}$$

- El error cuadrático medio-MSE y el coeficiente de determinación R^2 presentados en las tablas **5-3** y **5-4** y en la gráfica **5-2** muestran que el MSE aumenta, y consecuentemente el R^2 disminuye, a medida que se aumenta la magnitud del dato atípico en el eje de la X. Pero este comportamiento es más notorio en los métodos robustos que en OLS y Gini, lo cual es de esperarse puesto que entre más robusto es el método, este produce una recta de regresión más estable ante la presencia de datos atípicos.

5.1. Tablas de resultados.

Métodos	Normalidad	Escenario1	Escenario2	Escenario3
OLS	18,57	18,68	18,21	17,41
GINI	18,61	18,72	18,33	18,19
LAV	18,35	18,35	18,35	18,27
MHuber	18,61	18,71	18,61	18,50
MM	18,59	18,68	18,68	18,68

Tabla 5-1: Estimación de $\hat{\beta}_0$ con $n = 68$

Métodos	Normalidad	Escenario1	Escenario2	Escenario3
OLS	-0,063	-0,064	-0,056	-0,044
GINI	-0,063	-0,064	-0,058	-0,055
LAV	-0,058	-0,058	-0,058	-0,058
MHuber	-0,062	-0,064	-0,063	-0,061
MM	-0,062	-0,064	-0,064	-0,064

Tabla 5-2: Estimación de $\hat{\beta}_1$ con $n = 68$

Métodos	Normalidad	Escenario1	Escenario2	Escenario3
OLS	1,09	2,00	2,88	4,03
GINI	1,10	2,00	2,88	4,22
LAV	1,12	2,03	2,89	4,33
MHuber	1,10	2,00	2,94	4,48
MM	1,10	2,00	2,96	4,64

Tabla 5-3: Estimación de MSE con $n = 68$

Métodos	Normalidad	Escenario1	Escenario2	Escenario3
OLS	81,81	71,45	58,82	42,34
GINI	82,18	76,08	72,10	67,29
LAV	79,33	67,53	60,76	54,01
MHuber	81,49	71,73	63,54	55,88
MM	81,46	71,48	64,20	57,25

Tabla 5-4: Estimación de R^2 con $n = 68$

Métodos	Escenario1	Escenario2	Escenario3
OLS	0,6 %	-2,0 %	-6,3 %
GINI	0,6 %	-1,5 %	-2,3 %
LAV	0,0 %	0,0 %	-0,4 %
MHuber	0,6 %	0,0 %	-0,6 %
MM	0,5 %	0,5 %	0,5 %

Tabla 5-5: Variación de la estimación de $\hat{\beta}_0$ con $n = 10$

Métodos	Escenario1	Escenario2	Escenario3
OLS	1,6 %	-10,4 %	-29,7 %
GINI	1,7 %	-8,5 %	-13,0 %
LAV	0,0 %	0,0 %	-1,6 %
MHuber	3,2 %	0,8 %	-2,0 %
MM	2,7 %	2,7 %	2,7 %

Tabla 5-6: Variación de la estimación de $\hat{\beta}_1$ con $n = 10$

6 Conclusiones y Recomendaciones

6.1. Conclusiones

Según los resultados obtenidos se pueden tener las siguientes conclusiones:

1. Los coeficientes de regresión estimados por el método de Gini son más robustos ante la presencia de observaciones atípicas en comparación a los estimadores obtenidos por el método OLS. La robustez de Gini frente a OLS es más evidente para datos atípicos en el eje y que tengan una correspondiente coordenada atípica en el eje x, debido a que la función ponderadora de pendientes de Gini le da una importancia menor a las observaciones atípicas en x, mientras que la función ponderadora de OLS le da una importancia mayor. Las diferencias en las estimaciones obtenidas por Gini y OLS sólo son evidentes cuando la variable explicativa x no se distribuye uniforme o no es igualmente espaciada.

2. Al aumentar el tamaño muestral de las observaciones, se tiende a eliminar el efecto de las observaciones atípicas al momento de estimar los coeficientes de regresión por Gini, OLS y los métodos MM, M de Huber y LAV.

3. Aunque de antemano en este trabajo de tesis se sabía que los métodos MM, M de Huber y LAV eran muchísimo más robustos que Gini, se prosiguió a compararlo sólo con el fin de mostrar cuán parecido o diferente era Gini frente a OLS en relación a estos otros métodos desarrollados matemáticamente para ser robustos.

4. Los métodos MM, M de Huber y LAV fueron similares entre sí y diferentes a Gini y OLS en la estimación de los parámetros arrojados en la regresión lineal (β_0 y β_1) y en los estadísticos (MSE, R^2), debido a que fueron desarrollados matemáticamente (funciones de influencia acotadas) para ser robustos. Los métodos Gini y OLS fueron parecidos entre sí porque su naturaleza matemática radica en que la función de dispersión de los errores es lineal en Gini y cuadrática en OLS.

5. Se recomienda usar la regresión Gini cuando la muestra tiene observaciones atípicas simultáneamente en x y en y, pero si éstas fuesen solo atípicas en y las estimaciones obtenidas por OLS y Gini son equivalentes. Sin embargo, existen métodos más robustos como MM, M Huber y LAV que producen mejores resultados para cualquier tipo de atípico.

6.2. Investigaciones futuras

Para futuras investigaciones se recomienda simular un modelo lineal múltiple con variables explicativas x que tengan diferentes grados de correlación, con el fin de estudiar el efecto de la multicolinealidad, la cual se sabe que afecta la estimación de los parámetros en métodos sensibles como OLS, pero sería interesante saber que tanto afecta a métodos levemente robustos como Gini y fuertemete robustos como MM, M de Huber y LAV.

Bibliografía

[1] Aelst, V. S., Willems, G., y, Zamar, H. R. (2013). Robust and efficient estimation of the residual scale in linear regression. *Journal of Multivariate Analysis*, Vol. 116, pp.278-296.

[2] Bassett, G., Jr., y Koenker, R. (1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, Vol. 73, pp. 618-622.

[3] Bianco, M. A., Garcia, M. B. y, Yohai, J.V. (2005). Robust estimation for linear regression with asymmetric errors. *The Canadian Journal of Statistics*, Vol. 33, No. 4, pp. 511-528.

[4] Birkes, D., y Dodge, Y. (1993). *Alternative Methods of Regression, New York*, John Wiley & Sons Inc.

[5] Boente, G. y, Fraiman, R. (1989). Robust Nonparametric Regression Estimation. *Journal of Multivariate Analysis*, Vol.29, pp. 180-198.

[6] Borroni, C. G, Cazzaro, M. (2006). Some developments about a new nonparametric test based on Gini's mean difference. *Statistica & Applicazioni*, Vol. 3, No. 1, pp. 29-44.

[7] Borroni, C. Zenga, M. (2006). A test of concordance based on Gini's mean difference. *Communications in Statistics - Theory and Methods*, Vol.16, No.1, pp. 289-308.

[8] Boscovich, R.J., (1757). De literaria expeditione per pontificiam ditioned. et. synopsis amplioris opens, ac habeniur plura ejus ex exemplaria etiam sensorum impressa. *Bononiensi Scientiarum el Anum Instituto Atque Academia Commentarii*, Vol.4. pp.353-396.

[9] Casella, G., y, Berger, R. L. (2002). Statistical Inference. Ed Duxbury.

[10] David, H. A. (1968). Gini's Mean Difference Rediscovered. *Biometrika*, Vol. 55, No. 3, pp. 573-575.

[11] Denby, L.,y, Larsen, W. A.,(1977). Robust Regression Estimators Compared via Monte Carlo. *Communications in Statistics-Theory and Methods*, Vol.6, pp.335-362.

[12] Dielman T.E. (2005). Least Absolute Value Regression: Recent Contributions. *Journal of Statistical Computation and Simulation*, Vol. 75, No. 4. Pp. 263-286

[13] Douglas, C., Mongomery., E A., Peck, G. Geffrey Vining. (2002). *Introducción al análisis de regresión lineal*, 1a Edición, México. John Wiley & Sons, Inc.

[14] E. Schechtman, E., Yitshaki, S. The Hebrew University, Jerusalem, Israel (2003). A Family of Correlation Coefficients Based on Extended Gini. *Journal of Economic Inequality*, Vol.1, No.2, pp. 3083-3088.

[15] Edna, S. y, Yitzhaki, S. (2007). A Measure Of Association Based On Gin's Mean Difference. *Communications in Statistics-Theory and Methods*, Vol.16, No.1, pp. 207-231.

[16] Edna, S. y, Yitzhaki, S. (2013). Gini's Multiple Regressions. *Springer Series in Statistics*,Vol.1, pp. 4614-4720, New York.

[17] Edna, S., Yitzhaki, S. y, Pudalov, T. (2011). Gini's multiple regressions: two approaches and their Interaction. *International Journal of Statistics*, Vol. 69, No. 3, pp. 67-99.

[18] Edna, S., Yitzhaki, S. y, Artsev, Y. (2008). Who Does Not Respond in the Household Expenditure Survey: An Exercise in Extended Gini Regressions. *American Statistical Association Journal of Business & Economic Statistics*, Vol. 26, No. 3, pp. 329-344.

[19] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. y Stahel, W. A. (1986). Robust statistics: The approach based on inuence functions. Wiley.

[20] Hettmansperger, T. P. (1984). Statistical inference based on ranks. New York: John Wiley &Sons.

[21] Huber, P, J. (1973). Regression: Asymptotics, Conjectures, and Monte Carlo. *Annals of Statistics*, Vol.1, pp.799-821.

[22] Huber, P. J. (1981). Robust statistics. Wiley.

[23] Hurdle, W. (1984). Robust Regression Function Estimation. *Journal of Multivariate Analysis*, Vol. 14, pp. 169-180.

[24] Iachan, R. (1985). Robust Designs for Ratio and Regression Estimation. *Journal of Statistical Planning and Inference*, Vol. 11, pp. 149-161.

[25] Jaeckel, A. L (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Annals of Mathematical Statistics*, Vol. 43, No. 5, pp. 1449-1458.

[26] Jurecková ,J. (1971). Nonparametric estimates of regression coefficients. *Annals of Mathematical Statistics*, Vol.42, No. 4, pp. 1328-1338.

[27] Koenker, R. y Bassett, G. (1978). Regression quantiles. Econometrica, Vol.46 , pp.33-50.

[28] Koenker, R. y K.F.Hallock (2001): Quantile Regression. *Journal of Economic Perspectives*. Vol.15, N°4,pp.143-156.

[29] Kordzakhia, N., Mishra, D.G, y, Reiersolmoen, L. (2001). Robust estimation in the logistic regression model. *Journal of Statistical Planning and Inference*, Vol. 98, pp. 211-223.

[30] Li, J., y, Zheng, M. (2009). Robust estimation of multivariate regression model. *Statistical Planning and Inference* & *Springer*, Vol. 50, pp. 81-100.

[31] Meintanis, S.G., y, Donatos, S.G. (1997). A comparative study of some robust methods for coefficient-estimation in linear regression. *Computational Statistics* & *Data Analysis*, Vol. 23, pp. 525-540.

[32] Montanari,A.,Monari, P.(2008). Gini's ideas: new perspectives for modern multivariate statistical analysis *Statistica*, Vol. 3, N0. 4. Moody, Ivy, and Tukey, Paul A. (1979). *An Exploratory Analysis of Data on Lead Intoxication*. Bell Laboratories memorandum.

[33] Mosteller, F. y, Tukey, J, W. (1983). *Understanding Robust and Exploratory Data Analysis*. New York. John Wiley & Sons, Inc. Mussard, S. Souissi, F. (2013). Gini-PLS Regressions.pp. 1-20.

[34] McKean, J. W., y Hettmansperger, T. P. (1978). A robust analysis of the general linear model based on one step R-estimates. *Biometrika*, Vol.65, pp.571-579

[35] Olkin, I. y, Yitzhaki, S. (1992). Gini Regression Analysis. *International Statistical*, Vol. 60, No. 2, pp. 185-196.

[36] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

[37] Richard A. Johnson A.D y Dean W. Wichern W.D (1992).Applied Multivariate Statistical Analysis, 6th Edition.

[38] Randles, R. H., y, Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*, John Wiley & Sons, Inc.

[39] Ronald G. Askin, R. G., y,. Montgomery, D. C. (1984). An Analysis of Constrained Robust Regression Estimators. *Naval Research Logistics Quarterly*, Vol. 31, pp. 283-296.

[40] Rousseeuw, P. J. Y V. Yohai. (1984). Robust regression by means of S-estimators, *Springer, Berlín*, pp. 256-272.

[41] Seber, G. A. F. (1977). *Linear Regressión Analysis*. New York. John Wiley & Sons, Inc.

[42] Simpson, J. R. Y D. C. Montgomery. (1998). A performance-based assessment of robust regression methods, *Commun. Stat. Simul. Comput.*, Vol. 27. pp. 1031-1099.

[43] Staudte, R. G., y S . J. Sheather (1990). *Robust Estimation and Testing.* Wiley, New York.

[44] Stewart, J. (1999). *Calculo de varias variables: Conceptos y Contextos.* International Thomson Editores, S.A. de C.V

[45] Stone, Charles J. (1977). Consistent Nonparametric Regression. *Annals of Statistics*, Vol. 5, 595-620.

[46] Tukey, John W. (1977). *Exploratory Data Analysis.* Reading, Mass. Addison-Wesley.

[47] Welsh, H. A, Ronchetti, E. (2002). A journey in single steps: robust one-step M-estimation in linear regression. *Journal of Statistical Planning and Inference*, Vol. 103 pp. 287-310.

[48] Wibowo, A. (2009). Robust Kernel Ridge Regression Based on M-Estimation. *Computational Mathematics and Modeling & Springer Science+Business Media,Inc* Vol. 20, No. 4, pp. 438-446.

[49] Wilcox, R. R. (2005). Introduction to robust estimation and hypothes testing. 2a Edición. San Diego, CA: Academic Press.

[50] Wolfe, D. A y Hollander, M.(1999). *Nonparametric Stastical Methods*, Second Edition, New York. John Wiley & Sons, Inc.

[51] Yitzhaki, S. (1998). More Than a Dozen Alternative Ways of Spelling Gini. *Research on Economic Inequality*, Vol.8, pp.13-30.

[52] Yitzhaki, S. (2003). Gini's Mean difference: a superior measure of variability for non-normal distributions. *International Journal of Statistics*, Vol. 61, No. 2, pp. 285-316.

[53] Yitzhaki, S. y, Schechtman, E. (2004). The Gini Instrumental Variable, or The Double Instrumental Variable Estimator. *International Journal of Statistics*, vol. 62, No. 3, pp. 287-313.

[54] Yitzhaki, S. y, Schechtman, E. (2013). *The Gini Methodology A Primer on a Statistical Methodology.* Springer Series in Statistics, New York. Heidelberg Dordrecht London.

[55] Yitzhaki, S. y, Schechtman, E. (2005). The properties of the extended Gini measures of variability and inequality. *International Journal of Statistics*, Vol. 63, No. 3, pp. 401-433.

[56] Yitzhaki, S., y, Schechtman, E. (2009). Identifying monotonic and non-monotonic relationships. Development of Progress in Applied Statistical Modelling for the Economic Sciences Using Non-paramtric.

[57] Yohai, V. J. (1987). High breakdown-point and high eefficiency estimates for regression. *The Annals of Statistics*, Vol.15, pp.642-656.

7 Anexo: Código en R

```
#INSTALAR PAQUETES Y CARGAR LIBRERÍAS*****************************************.
#install.packages("MASS", dependencies = TRUE)#regresiones, funciones, distribuciones.
#install.packages("quantreg", dependencies = TRUE)#regresión cuantil.

library(MASS)# Soporta funciones y conjuntos de datos: regresion (rlm,lm), distribuciones,etc.
library(SparseM) # complemento del paquete "quantreg" para algebra lineal y matricial
library(quantreg)#regresión cuantil (rq)

#INICIALIZACIÓN DE PARÁMETROS,VARIABLES, VECTORES Y MATRICES*******************.

N=1000#número de iteraciones.
nm=c(100,30,10)#número de observaciones.
#se fijan los parámetros del modelo lineal teórico.
b0=5# intercepto.
b1=20 # pendiente de x1.

#matrices para guardar los parámetros estimados por cada método de regresión (b0,b1,MSE,R2).
np=4# número de parámetros directos a estimar: b0,b1,MSE,R2
npi=1 # número de parámetros indirectos a estimar: Estadistico chi calculado (EC)
nv=np+npi# número de variables a calcular (b0,b1,MSE,R2,EC)
#párametros OLS bajo normalidad y contaminando con un atípico de varias desviaciones estandares.
v0matrixRegOLS=matrix(ncol=nv,nrow=N)#normalidad sin atípico.
v1matrixRegOLS=matrix(ncol=nv,nrow=N)#atípico de 5 desviaciones estandares.
v2matrixRegOLS=matrix(ncol=nv,nrow=N)#atípico de 50 desviaciones estandares.
v3matrixRegOLS=matrix(ncol=nv,nrow=N)#atípico de 150 desviaciones estandares.

#párametros GINI bajo normalidad y contaminando con un atípico de varias desviaciones estandares.
v0matrixRegGINI =matrix(ncol=nv,nrow=N)#normalidad sin atípico.
v1matrixRegGINI =matrix(ncol=nv,nrow=N)#atípico de 5 desviaciones estandares.
v2matrixRegGINI =matrix(ncol=nv,nrow=N)#atípico de 50 desviaciones estandares.
v3matrixRegGINI =matrix(ncol=nv,nrow=N)#atípico de 150 desviaciones estandares.

#párametros LAV bajo normalidad y contaminando con un atípico de varias desviaciones estandares.
v0matrixRegLAV =matrix(ncol=nv,nrow=N)#normalidad sin atípico.
v1matrixRegLAV =matrix(ncol=nv,nrow=N)#atípico de 5 desviaciones estandares.
v2matrixRegLAV =matrix(ncol=nv,nrow=N)#atípico de 50 desviaciones estandares.
v3matrixRegLAV =matrix(ncol=nv,nrow=N)#atípico de 150 desviaciones estandares.

#párametros MHUBER bajo normalidad y contaminando con un atípico de varias desviaciones estandares.
v0matrixRegMHuber =matrix(ncol=nv,nrow=N)#normalidad sin atípico.
v1matrixRegMHuber =matrix(ncol=nv,nrow=N)#atípico de 5 desviaciones estandares.
v2matrixRegMHuber =matrix(ncol=nv,nrow=N)#atípico de 50 desviaciones estandares.
v3matrixRegMHuber =matrix(ncol=nv,nrow=N)#atípico de 150 desviaciones estandares.

#párametros MM bajo normalidad y contaminando con un atípico de varias desviaciones estandares.
v0matrixRegMM =matrix(ncol=nv,nrow=N)#normalidad sin atípico.
v1matrixRegMM =matrix(ncol=nv,nrow=N)#atípico de 5 desviaciones estandares.
v2matrixRegMM =matrix(ncol=nv,nrow=N)#atípico de 50 desviaciones estandares.
v3matrixRegMM =matrix(ncol=nv,nrow=N)#atípico de 150 desviaciones estandares.
```

```
#Tabla de resultados indice robustez para los métodos según los escenerios de normalidad.
nmet=5  #número de métodos a comparar: OLS, GINI, LAV, MM, M Huber
ne=4# número de escenarios a estudiar: Esc 0 (normalidad), Esc 1 (5Sd), Esc2 0 (50Sd), Esc3 (150Sd).

#Tabla de resultados  del intercepto para los métodos según los escenerios de normalidad.
tablaB0=matrix(ncol=ne,nrow=nmet)
colnames(tablaB0)=c("Esc0","Esc1","Esc2","Esc3")
rownames(tablaB0)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados de la pendiente x1 para los métodos según los escenerios de normalidad.
tablaB1=matrix(ncol=ne,nrow=nmet)
colnames(tablaB1)=c("Esc0","Esc1","Esc2","Esc3")
rownames(tablaB1)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del MSE para los métodos según los escenerios de normalidad.
tablaMSE=matrix(ncol=ne,nrow=nmet)
colnames(tablaMSE)=c("Esc0","Esc1","Esc2","Esc3")
rownames(tablaMSE)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del R^2 para los métodos según los escenerios de normalidad.
tablaR2=matrix(ncol=ne,nrow=nmet)
colnames(tablaR2)=c("Esc0","Esc1","Esc2","Esc3")
rownames(tablaR2)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del valor p de la  prueba de bondad de ajuste con Kolmogorov S.
tablaVP=matrix(ncol=ne,nrow=nmet)
colnames(tablaVP)=c("Esc0","Esc1","Esc2","Esc3")
rownames(tablaVP)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del Estadístico muestral de la prueba de bondad de ajuste con Kolmogorov-S.
tablaEC=matrix(ncol=ne,nrow=nmet)
colnames(tablaEC)=c("Esc0","Esc1","Esc2","Esc3")
rownames(tablaEC)=c("OLS","GINI","LAV","M_Huber","MM")

# Para el primer tamaño muestral n1
#Tabla de resultados  del intercepto para los métodos según los escenerios de normalidad.
n1tablaB0=matrix(ncol=ne,nrow=nmet)
colnames(n1tablaB0)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n1tablaB0)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados de la pendiente x1 para los métodos según los escenerios de normalidad.
n1tablaB1=matrix(ncol=ne,nrow=nmet)
colnames(n1tablaB1)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n1tablaB1)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del MSE para los métodos según los escenerios de normalidad.
n1tablaMSE=matrix(ncol=ne,nrow=nmet)
colnames(n1tablaMSE)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n1tablaMSE)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del R^2 para los métodos según los escenerios de normalidad.
n1tablaR2=matrix(ncol=ne,nrow=nmet)
colnames(n1tablaR2)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n1tablaR2)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del valor p de la  prueba de bondad de ajuste con Kolmogorov S.
n1tablaVP=matrix(ncol=ne,nrow=nmet)
colnames(n1tablaVP)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n1tablaVP)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del Estadístico muestral de la prueba de bondad de ajuste con Kolmogorov-S.
```

```
n1tablaEC=matrix(ncol=ne,nrow=nmet)
colnames(n1tablaEC)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n1tablaEC)=c("OLS","GINI","LAV","M_Huber","MM")

#Para el segundo tamaño muestral n2
#Tabla de resultados  del intercepto para los métodos según los escenerios de normalidad.
n2tablaB0=matrix(ncol=ne,nrow=nmet)
colnames(n2tablaB0)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n2tablaB0)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados de la pendiente x1 para los métodos según los escenerios de normalidad.
n2tablaB1=matrix(ncol=ne,nrow=nmet)
colnames(n2tablaB1)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n2tablaB1)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del MSE para los métodos según los escenerios de normalidad.
n2tablaMSE=matrix(ncol=ne,nrow=nmet)
colnames(n2tablaMSE)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n2tablaMSE)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del R^2 para los métodos según los escenerios de normalidad.
n2tablaR2=matrix(ncol=ne,nrow=nmet)
colnames(n2tablaR2)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n2tablaR2)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del valor p de la  prueba de bondad de ajuste con Kolmogorov S.
n2tablaVP=matrix(ncol=ne,nrow=nmet)
colnames(n2tablaVP)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n2tablaVP)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del Estadístico muestral de la prueba de bondad de ajuste con Kolmogorov-S.
n2tablaEC=matrix(ncol=ne,nrow=nmet)
colnames(n2tablaEC)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n2tablaEC)=c("OLS","GINI","LAV","M_Huber","MM")

# Para el tercer tamaño muestral n3
#Tabla de resultados  del intercepto para los métodos según los escenerios de normalidad.
n3tablaB0=matrix(ncol=ne,nrow=nmet)
colnames(n3tablaB0)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n3tablaB0)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados de la pendiente x1 para los métodos según los escenerios de normalidad.
n3tablaB1=matrix(ncol=ne,nrow=nmet)
colnames(n3tablaB1)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n3tablaB1)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del MSE para los métodos según los escenerios de normalidad.
n3tablaMSE=matrix(ncol=ne,nrow=nmet)
colnames(n3tablaMSE)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n3tablaMSE)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del R^2 para los métodos según los escenerios de normalidad.
n3tablaR2=matrix(ncol=ne,nrow=nmet)
colnames(n3tablaR2)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n3tablaR2)=c("OLS","GINI","LAV","M_Huber","MM")

#Tabla de resultados del valor p de la  prueba de bondad de ajuste con Kolmogorov S.
n3tablaVP=matrix(ncol=ne,nrow=nmet)
colnames(n3tablaVP)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n3tablaVP)=c("OLS","GINI","LAV","M_Huber","MM")
```

```
#Tabla de resultados del Estadístico muestral de la prueba de bondad de ajuste con Kolmogorov-S.
n3tablaEC=matrix(ncol=ne,nrow=nmet)
colnames(n3tablaEC)=c("Esc0","Esc1","Esc2","Esc3")
rownames(n3tablaEC)=c("OLS","GINI","LAV","M_Huber","MM")

#paso 3: ESTIMAR LOS PARÁMETROS DEL MODELO CON LOS MÉTODOS DE REGRESIÓN********.

#paso 3.1 Regresión por minimos Cuadrados Ordinarios (OLS).
RegOLS=function(Y,X){
mod=lm(Y~X)#función LM para estimar por OLS.
b0e=mod$coefficients[1]# Intercepto estimado.
b1e=mod$coefficients[2] #Pendiente x1 estimado.
msee=anova(mod)$Mean[2] # Error cuadratico medio estimado.
R2e=100*(summary(mod)$r.squared) # R cuadrado estimado.
rta=matrix(c(b0e,b1e,msee,R2e),ncol=np)# valores de retorno.
colnames(rta)=c("Beta0","Beta1","MSE","R2")
return(rta)#Retorna los parametros de la regresión.
}# fin de OLS

#paso 3.2 Regresión de Gini
RegGINI=function(Y,X){
p=ncol(X)#Número de regresoras X.
n=nrow(X)#Número de observaciones.
b1e=cov(Y,rank(X))/cov(X,rank(X))
Xprom=apply(X,2,mean)# Promedio de las regresoras X.
Yprom=mean(Y)#Promedio de la variable independiente Y.
b0e=Yprom-b1e%*%Xprom #intercepto de GINI estimado
Yg= t(rep(b0e,n)+ b1e%*%t(X))#Modelo ajustado de GINI.
eg=Y-Yg#Residuales de GINI.
msee=(1/(n-p-1))*sum((eg)^2 )# Error cuadratico Medio de GINI.
#R2e = 100*(1-(cov(eg,rank(eg))/cov(Y,rank(Y)))^2)#Coeficiente de determinación.

SSE=sum(eg^2)#Suma de cuadrados de los errores.
SSM=sum((Yg-mean(Y))^2 )#Suma de cuadrados de la regresión.
SST=SSM+SSE#Suma de cuadrados totales del modelo.
R2e=100*(1-(SSE/SST))#coeficiente de determinación R^2

rta=matrix(c(b0e,b1e,msee,R2e),ncol=np)# valores de retorno.
colnames(rta)=c("Beta0","Beta1","MSE","R2")
return(rta)#Retorna los parametros de la regresión.
}# fin de GINI

###paso 3.3 Regresión por LAV
RegLAV=function(Y,X){
p=ncol(X)#Número de regresoras X.
n=nrow(X)#Número de observaciones.

mod=rq(Y~X)#función rq para estimar por LAV.
b0e=mod$coefficients[1]# Intercepto.
b1e=mod$coefficients[2] #Pendiente X1.

msee=(1/(n-p-1))*sum((residuals(mod))^2) # Error cuadratico medio.
Ymod=t(b0e+b1e%*%t(X))#Modelo ajustado de LAV.
SSE=sum((residuals(mod))^2 )#Suma de cuadrados de los errores.
SSM=sum((Ymod-mean(Y))^2 )#Suma de cuadrados de la regresión.
SST=SSM+SSE#Suma de cuadrados totales del modelo.
R2e=100*(1-(SSE/SST))#coeficiente de determinación R^2
rta=matrix(c(b0e,b1e,msee,R2e),ncol=np)# valores de retorno.
colnames(rta)=c("Beta0","Beta1","MSE","R2")
return(rta)#Retorna los parametros de la regresión.
}# fin de LAV
```

```
###paso 3.3 Regresión por M de Hubbert
RegMHuber=function(Y,X){
p=ncol(X)#Número de regresoras X.
n=nrow(X)#Número de observaciones.
mod=rlm(Y~X,method="M",maxit=500)#función rlm para estimar por MHuber.
b0e=mod$coefficients[1]# Intercepto
b1e=mod$coefficients[2] #Pendiente x1
msee=(1/(n-p-1))*sum((residuals(mod))^2) # Error cuadratico medio
Ymod=t(b0e+b1e%*%t(X))#Modelo ajustado por MHUBER
SSE=sum((residuals(mod))^2 )#Suma de cuadrados de los errores.
SSM=sum((Ymod-mean(Y))^2 )#Suma de cuadrados de la regresión.
SST=SSM+SSE #Suma de cuadrados totales del modelo.
R2e=100*(1-(SSE/SST))#coeficiente de determinación R^2
rta=matrix(c(b0e,b1e,msee,R2e),ncol=np)# valores de retorno.
colnames(rta)=c("Beta0","Beta1","MSE","R2")
return(rta)#Retorna los parametros de la regresión.
}# fin de M de Hubbert

###paso 3.3 Regresión por MM
RegMM=function(Y,X){
p=ncol(X)#Número de regresoras X.
n=nrow(X)#Número de observaciones.
mod=rlm(Y~X,method="MM",maxit=500)#función rlm para estimar por MM.
b0e=mod$coefficients[1]# Intercepto.
b1e=mod$coefficients[2] #Pendiente x1.
msee=(1/(n-p-1))*sum((residuals(mod))^2) # Error cuadratico medio.
Ymod=t(b0e+b1e%*%t(X))#Modelo ajustado por MM.
SSE=sum((residuals(mod))^2 )#Suma de cuadrados de los errores.
SSM=sum((Ymod-mean(Y))^2 )#Suma de cuadrados de la regresión.
SST=SSM+SSE #Suma de cuadrados totales del modelo.
R2e=100*(1-(SSE/SST))#coeficiente de determinación R^2
rta=matrix(c(b0e,b1e,msee,R2e),ncol=np)# valores de retorno.
colnames(rta)=c("Beta0","Beta1","MSE","R2")
return(rta)#Retorna los parametros de la regresión.
}# fin de MM

# REALIZAR LAS N ITERACIONES PARA CADA TABLA DE RESULTADOS.

for(t in 1:length(nm)){#recorre el vector de observaciones.

# EJECUTAR LAS N ITERACIONES PARA CADA UNO DE LOS ESECNARIOS PARA ENCONTRAR LA TENDENCIA CENTRAL
for(j in 1:N){

n=nm[t];n#vector de observaciones o tamañao de muestra.
#paso 1 : FIJAR EL VECTOR DE REGRESORES X1 ************************************.
X=NULL
X=c(seq(1,n-1),n+n/5)# la variable regresora x1 ordenada, con el último dato alejado.
X=as.matrix(X);X# matricialmente
#Paso 2:DEFINIR RELACION LINEAL TEORICO ENTRE LAS REGRESORAS Y LA VBLE RTA Y ****.
p=ncol(X);p#Número de regresoras X.
Y0=NULL
Y0=t(b0+b1%*%t(X))# modelo teórico.
Y0=as.matrix(Y0);Y0
#*******************************************************************************

#Paso 3:GENERAR ERRORES NORMALES Y DESVIADOS DE LA NORMALIDAD******************.
s=1# desvacion estandar
u=0# media
e=NULL
e=rnorm(n,u,s) ;e#generar n errores normales con media=0 y var=1.
```

```
Y=NULL
Y=Y0+e #modelo observado.
Y=as.matrix(Y) ;Y

#Paso 2:DEFINIR RELACION LINEL TEORICO ENTRE LAS REGRESORAS Y LA VBLE RTA

#****************************
#Paso 4: ESTIMAR LOS PARÁMETROS B0,B1, MSE Y R2, Rob, Efi CON CADA UNO DE LOS MÉTODOS DE REGRESIÓN.
v0matrixRegOLS[j,1:4]=RegOLS(Y,X)#parámetros de OLS
v0matrixRegGINI[j,1:4]=RegGINI(Y,X)#parámetros de GINI
v0matrixRegLAV[j,1:4]=RegLAV(Y,X)#parámetros de LAV
v0matrixRegMHuber[j,1:4]=RegMHuber(Y,X)#parámetros de M Huber
v0matrixRegMM[j,1:4]=RegMM(Y,X)#parámetros de MM

# Valor P de la prueba de bondad de ajuste
n=nrow(X)
unos=rep(1,n)#crea un vector de unos para poder tener una matriz de 2x2
Xm=cbind(unos,X)# forma una matriz de nx2 con X y el vector de unos
Xc=t(Xm)%*%(Xm)
# se calcula el estadistico muestral.
B_B0OLS=t(as.matrix(c(v0matrixRegOLS[j,1],v0matrixRegOLS[j,2])-c(b0,b1)))
B_B0GINI=t(as.matrix(c(v0matrixRegGINI[j,1],v0matrixRegGINI[j,2])-c(b0,b1)))
B_B0LAV=t(as.matrix(c(v0matrixRegLAV[j,1],v0matrixRegLAV[j,2])-c(b0,b1)))
B_B0MHuber=t(as.matrix(c(v0matrixRegMHuber[j,1],v0matrixRegMHuber[j,2])-c(b0,b1)))
B_B0MM=t(as.matrix(c(v0matrixRegMM[j,1],v0matrixRegMM[j,2])-c(b0,b1)))

# estadistico muestral calculado

EstcalOLS = B_B0OLS%*%Xc%*%t(B_B0OLS)/s^2
EstcalGINI=B_B0GINI%*%Xc%*%t(B_B0GINI)/s^2
EstcalLAV=B_B0LAV%*%Xc%*%t(B_B0LAV)/((pi/2)*s^2)
EstcalMHuber=B_B0MHuber%*%Xc%*%t(B_B0MHuber)/s^2
EstcalMM=B_B0MM%*%Xc%*%t(B_B0MM)/s^2

# se agrega la última columna: estadistico muestral
v0matrixRegOLS[j,5]=EstcalOLS
v0matrixRegGINI[j,5]=EstcalGINI
v0matrixRegLAV[j,5]=EstcalLAV
v0matrixRegMHuber[j,5]=EstcalMHuber
v0matrixRegMM[j,5]=EstcalMM

#*********************************************************************************
# Escenario 1 de no normalidad y observaciones influyentes.
out=4*s#magnitud del atípico a 2 desviación estándar.
#Dado que X está ordenado, entonces la Y por defecto tambien lo está, por ende
#...la última comoponente de Y corresponde al valor mayor.
Y[n]=Y0[n]+out#contamina la última observación del modelo observado.

dfit=abs(dffits(lm(Y~X)))#halla la diferencia entre el modelo observado y el ajustado.
restudent=abs(rstudent(lm(Y~X)))#residuales estudentizados del modelo ajustado por OLS.
while(restudent[n]<2 & dfit[n]<2*sqrt((p+1)/n)){##verifica que haya observaciones influyentes.
out=out+s#aumenta la magnitud del atípico hasta que se vuelve influyente.
Y[n]=Y0[n]+out#actualiza la observación.
dfit=abs(dffits(lm(Y~X)))#vuelve a calcular los ddfits.
restudent=abs(rstudent(lm(Y~X)))
}#fin de while.

#*********************************************************************************
#Paso 4: ESTIMAR LOS PARÁMETROS  B0,B1, MSE Y R2, Rob, Efi CON CADA UNO DE LOS MÉTODOS DE REGRESIÓN.

v1matrixRegOLS[j,1:4]=RegOLS(Y,X)#parámetros de OLS
```

```
v1matrixRegGINI[j,1:4]=RegGINI(Y,X)#parámetros de GINI
v1matrixRegLAV[j,1:4]=RegLAV(Y,X)#parámetros de LAV
v1matrixRegMHuber[j,1:4]=RegMHuber(Y,X)#parámetros de M Huber
v1matrixRegMM[j,1:4]=RegMM(Y,X)#parámetros de MM

# se calcula el estadistico muestral.
B_BOOLS=t(as.matrix(c(v1matrixRegOLS[j,1],v1matrixRegOLS[j,2])-c(b0,b1)))
B_BOGINI=t(as.matrix(c(v1matrixRegGINI[j,1],v1matrixRegGINI[j,2])-c(b0,b1)))
B_BOLAV=t(as.matrix(c(v1matrixRegLAV[j,1],v1matrixRegLAV[j,2])-c(b0,b1)))
B_BOMHuber=t(as.matrix(c(v1matrixRegMHuber[j,1],v1matrixRegMHuber[j,2])-c(b0,b1)))
B_BOMM=t(as.matrix(c(v1matrixRegMM[j,1],v1matrixRegMM[j,2])-c(b0,b1)))

# estadistico muestral calculado

EstcalOLS = B_BOOLS%*%Xc%*%t(B_BOOLS)/s^2
EstcalGINI=B_BOGINI%*%Xc%*%t(B_BOGINI)/s^2
EstcalLAV=B_BOLAV%*%Xc%*%t(B_BOLAV)/((pi/2)*s^2)
EstcalMHuber=B_BOMHuber%*%Xc%*%t(B_BOMHuber)/s^2
EstcalMM=B_BOMM%*%Xc%*%t(B_BOMM)/s^2

# se agregan la última columna: estadistico muestral
v1matrixRegOLS[j,5]=EstcalOLS
v1matrixRegGINI[j,5]=EstcalGINI
v1matrixRegLAV[j,5]=EstcalLAV
v1matrixRegMHuber[j,5]=EstcalMHuber
v1matrixRegMM[j,5]=EstcalMM

#*******************************************************************************.
# Escenario 2 de no normalidad y observaciones influyentes.

#Dado que X está ordenado, entonces la Y por defecto tambien lo está, por ende
#...la última comoponente de Y corresponde al valor mayor.
Y[n]=Y0[n]+2*out#contamina la última observación del modelo observado.

#*******************************************************************************
#Paso 4: ESTIMAR LOS PARÁMETROS  B0,B1, MSE Y R2, Rob, Efi CON CADA UNO DE LOS MÉTODOS DE REGRESIÓN.

v2matrixRegOLS[j,1:4]=RegOLS(Y,X)#parámetros de OLS
v2matrixRegGINI[j,1:4]=RegGINI(Y,X)#parámetros de GINI
v2matrixRegLAV[j,1:4]=RegLAV(Y,X)#parámetros de LAV
v2matrixRegMHuber[j,1:4]=RegMHuber(Y,X)#parámetros de M Huber
v2matrixRegMM[j,1:4]=RegMM(Y,X)#parámetros de MM

# se calcula el estadistico muestral.
B_BOOLS=t(as.matrix(c(v2matrixRegOLS[j,1],v2matrixRegOLS[j,2])-c(b0,b1)))
B_BOGINI=t(as.matrix(c(v2matrixRegGINI[j,1],v2matrixRegGINI[j,2])-c(b0,b1)))
B_BOLAV=t(as.matrix(c(v2matrixRegLAV[j,1],v2matrixRegLAV[j,2])-c(b0,b1)))
B_BOMHuber=t(as.matrix(c(v2matrixRegMHuber[j,1],v2matrixRegMHuber[j,2])-c(b0,b1)))
B_BOMM=t(as.matrix(c(v2matrixRegMM[j,1],v2matrixRegMM[j,2])-c(b0,b1)))

# estadistico muestral calculado

EstcalOLS = B_BOOLS%*%Xc%*%t(B_BOOLS)/s^2
EstcalGINI=B_BOGINI%*%Xc%*%t(B_BOGINI)/s^2
EstcalLAV=B_BOLAV%*%Xc%*%t(B_BOLAV)/((pi/2)*s^2)
EstcalMHuber=B_BOMHuber%*%Xc%*%t(B_BOMHuber)/s^2
EstcalMM=B_BOMM%*%Xc%*%t(B_BOMM)/s^2

# se agregan la última columna: estadistico muestral
v2matrixRegOLS[j,5]=EstcalOLS
v2matrixRegGINI[j,5]=EstcalGINI
```

```
v2matrixRegLAV[j,5]=EstcalLAV
v2matrixRegMHuber[j,5]=EstcalMHuber
v2matrixRegMM[j,5]=EstcalMM

#*******************************************************************************

# Escenario 3 de no normalidad y observaciones influyentes.

#Dado que X está ordenado, entonces la Y por defecto tambien lo está, por ende
#...la última comoponente de Y corresponde al valor mayor.
Y[n]=Y0[n]+4*out#contamina la última observación del modelo observado.
#*******************************************************************************
#Paso 4: ESTIMAR LOS PARÁMETROS B0, B1, MSE Y R2, Rob, Efi CON CADA UNO DE LOS MÉTODOS DE REGRESIÓN.

v3matrixRegOLS[j,1:4]=RegOLS(Y,X)#parámetros de OLS
v3matrixRegGINI[j,1:4]=RegGINI(Y,X)#parámetros de GINI
v3matrixRegLAV[j,1:4]=RegLAV(Y,X)#parámetros de LAV
v3matrixRegMHuber[j,1:4]=RegMHuber(Y,X)#parámetros de M Huber
v3matrixRegMM[j,1:4]=RegMM(Y,X)#parámetros de MM

# se calcula el estadistico muestral.
B_B0OLS=t(as.matrix(c(v3matrixRegOLS[j,1],v3matrixRegOLS[j,2])-c(b0,b1)))
B_B0GINI=t(as.matrix(c(v3matrixRegGINI[j,1],v3matrixRegGINI[j,2])-c(b0,b1)))
B_B0LAV=t(as.matrix(c(v3matrixRegLAV[j,1],v3matrixRegLAV[j,2])-c(b0,b1)))
B_B0MHuber=t(as.matrix(c(v3matrixRegMHuber[j,1],v3matrixRegMHuber[j,2])-c(b0,b1)))
B_B0MM=t(as.matrix(c(v3matrixRegMM[j,1],v3matrixRegMM[j,2])-c(b0,b1)))

# estadistico muestral calculado

EstcalOLS = B_B0OLS%*%Xc%*%t(B_B0OLS)/s^2
EstcalGINI=B_B0GINI%*%Xc%*%t(B_B0GINI)/s^2
EstcalLAV=B_B0LAV%*%Xc%*%t(B_B0LAV)/((pi/2)*s^2)
EstcalMHuber=B_B0MHuber%*%Xc%*%t(B_B0MHuber)/s^2
EstcalMM=B_B0MM%*%Xc%*%t(B_B0MM)/s^2

# se agregan la última columna: estadistico muestral
v3matrixRegOLS[j,5]=EstcalOLS
v3matrixRegGINI[j,5]=EstcalGINI
v3matrixRegLAV[j,5]=EstcalLAV
v3matrixRegMHuber[j,5]=EstcalMHuber
v3matrixRegMM[j,5]=EstcalMM
#*******************************************************************************
}# fin for para N iteraciones

#*******************************************************************************
# Paso 5: CALCULAR LA TENDENCIA CENTRAL (PROMEDIO) DE LOS PARÁMETROS EN CADA ESCENARIO
# Escenario 0 de normalidad
v0ParOLS=apply(v0matrixRegOLS,2,mean)#Valor medio de cada parametro estim con OLS sin atipicos
v0ParGINI=apply(v0matrixRegGINI,2,mean)#Valor medio de cada parametro estim con GINI sin atipicos
v0ParLAV=apply(v0matrixRegLAV,2,mean)#Valor medio de cada parametro estim con LAV sin atipicos
v0ParMHuber=apply(v0matrixRegMHuber,2,mean)#Valor medio de cada parametro estim con M Huber sin atipicos
v0ParMM=apply(v0matrixRegMM,2,mean)#Valor medio de cada parametro estim con MM sin atipicos
# Escenario 1 de no normalidad
v1ParOLS=apply(v1matrixRegOLS,2,mean)#Valor medio de cada parametro estim con OLS
v1ParGINI=apply(v1matrixRegGINI,2,mean)#Valor medio de cada parametro estim con GINI
v1ParLAV=apply(v1matrixRegLAV,2,mean)#Valor medio de cada parametro estim con LAV
v1ParMHuber=apply(v1matrixRegMHuber,2,mean)#Valor medio de cada parametro estim con M Huber
v1ParMM=apply(v1matrixRegMM,2,mean)#Valor medio de cada parametro estim con MM
# Escenario 2 de no normalidad
v2ParOLS=apply(v2matrixRegOLS,2,mean)#Valor medio de cada parametro estim con OLS
v2ParGINI=apply(v2matrixRegGINI,2,mean)#Valor medio de cada parametro estim con GINI
```

```
v2ParLAV=apply(v2matrixRegLAV,2,mean)#Valor medio de cada parametro estim con LAV
v2ParMHuber=apply(v2matrixRegMHuber,2,mean)#Valor medio de cada parametro estim con M Huber
v2ParMM=apply(v2matrixRegMM,2,mean)#Valor medio de cada parametro estim con MM
# Escenario 3 de no normalidad
v3ParOLS=apply(v3matrixRegOLS,2,mean)#Valor medio de cada parametro estim con OLS
v3ParGINI=apply(v3matrixRegGINI,2,mean)#Valor medio de cada parametro estim con GINI
v3ParLAV=apply(v3matrixRegLAV,2,mean)#Valor medio de cada parametro estim con LAV
v3ParMHuber=apply(v3matrixRegMHuber,2,mean)#Valor medio de cada parametro estim con M Huber
v3ParMM=apply(v3matrixRegMM,2,mean)#Valor medio de cada parametro estim con MM

#calculo del valor p en la prueba de bondad de ajuste.
w=100# factor porcentual
v0valorpOLS=w*ks.test(v0matrixRegOLS[,5],"pchisq",p+1)$p.value
v0valorpGINI=w*ks.test(v0matrixRegGINI[,5],"pchisq",p+1)$p.value
v0valorpLAV=w*ks.test(v0matrixRegLAV[,5],"pchisq",p+1)$p.value
v0valorpMHuber=w*ks.test(v0matrixRegMHuber[,5],"pchisq",p+1)$p.value
v0valorpMM=w*ks.test(v0matrixRegMM[,5],"pchisq",p+1)$p.value

v1valorpOLS=w*ks.test(v1matrixRegOLS[,5],"pchisq",p+1)$p.value
v1valorpGINI=w*ks.test(v1matrixRegGINI[,5],"pchisq",p+1)$p.value
v1valorpLAV=w*ks.test(v1matrixRegLAV[,5],"pchisq",p+1)$p.value
v1valorpMHuber=w*ks.test(v1matrixRegMHuber[,5],"pchisq",p+1)$p.value
v1valorpMM=w*ks.test(v1matrixRegMM[,5],"pchisq",p+1)$p.value

v2valorpOLS=w*ks.test(v2matrixRegOLS[,5],"pchisq",p+1)$p.value
v2valorpGINI=w*ks.test(v2matrixRegGINI[,5],"pchisq",p+1)$p.value
v2valorpLAV=w*ks.test(v2matrixRegLAV[,5],"pchisq",p+1)$p.value
v2valorpMHuber=w*ks.test(v2matrixRegMHuber[,5],"pchisq",p+1)$p.value
v2valorpMM=w*ks.test(v2matrixRegMM[,5],"pchisq",p+1)$p.value

v3valorpOLS=w*ks.test(v3matrixRegOLS[,5],"pchisq",p+1)$p.value
v3valorpGINI=w*ks.test(v3matrixRegGINI[,5],"pchisq",p+1)$p.value
v3valorpLAV=w*ks.test(v3matrixRegLAV[,5],"pchisq",p+1)$p.value
v3valorpMHuber=w*ks.test(v3matrixRegMHuber[,5],"pchisq",p+1)$p.value
v3valorpMM=w*ks.test(v3matrixRegMM[,5],"pchisq",p+1)$p.value

#*********************************************************************************

# tabla de resultados de intercepto B0
tablaB0[1,1]=v0ParOLS[1];tablaB0[1,2]=v1ParOLS[1];tablaB0[1,3]=v2ParOLS[1];tablaB0[1,4]=v3ParOLS[1]
tablaB0[2,1]=v0ParGINI[1];tablaB0[2,2]=v1ParGINI[1];tablaB0[2,3]=v2ParGINI[1];tablaB0[2,4]=v3ParGINI[1]
tablaB0[3,1]=v0ParLAV[1];tablaB0[3,2]=v1ParLAV[1];tablaB0[3,3]=v2ParLAV[1];tablaB0[3,4]=v3ParLAV[1]
tablaB0[4,1]=v0ParMHuber[1];tablaB0[4,2]=v1ParMHuber[1];tablaB0[4,3]=v2ParMHuber[1];tablaB0[4,4]=v3ParMHuber[1]
tablaB0[5,1]=v0ParMM[1];tablaB0[5,2]=v1ParMM[1];tablaB0[5,3]=v2ParMM[1];tablaB0[5,4]=v3ParMM[1]
# tabla de resultados de coeficiente x1
tablaB1[1,1]=v0ParOLS[2];tablaB1[1,2]=v1ParOLS[2];tablaB1[1,3]=v2ParOLS[2];tablaB1[1,4]=v3ParOLS[2]
tablaB1[2,1]=v0ParGINI[2];tablaB1[2,2]=v1ParGINI[2];tablaB1[2,3]=v2ParGINI[2];tablaB1[2,4]=v3ParGINI[2]
tablaB1[3,1]=v0ParLAV[2];tablaB1[3,2]=v1ParLAV[2];tablaB1[3,3]=v2ParLAV[2];tablaB1[3,4]=v3ParLAV[2]
tablaB1[4,1]=v0ParMHuber[2];tablaB1[4,2]=v1ParMHuber[2];tablaB1[4,3]=v2ParMHuber[2];tablaB1[4,4]=v3ParMHuber[2]
tablaB1[5,1]=v0ParMM[2];tablaB1[5,2]=v1ParMM[2];tablaB1[5,3]=v2ParMM[2];tablaB1[5,4]=v3ParMM[2]
# tabla de resultados de MSE
tablaMSE[1,1]=v0ParOLS[3];tablaMSE[1,2]=v1ParOLS[3];tablaMSE[1,3]=v2ParOLS[3];tablaMSE[1,4]=v3ParOLS[3]
tablaMSE[2,1]=v0ParGINI[3];tablaMSE[2,2]=v1ParGINI[3];tablaMSE[2,3]=v2ParGINI[3];tablaMSE[2,4]=v3ParGINI[3]
tablaMSE[3,1]=v0ParLAV[3];tablaMSE[3,2]=v1ParLAV[3];tablaMSE[3,3]=v2ParLAV[3];tablaMSE[3,4]=v3ParLAV[3]
tablaMSE[4,1]=v0ParMHuber[3];tablaMSE[4,2]=v1ParMHuber[3];tablaMSE[4,3]=v2ParMHuber[3];tablaMSE[4,4]=v3ParMHuber[3]
tablaMSE[5,1]=v0ParMM[3];tablaMSE[5,2]=v1ParMM[3];tablaMSE[5,3]=v2ParMM[3];tablaMSE[5,4]=v3ParMM[3]
# tabla de resultados de R2
tablaR2[1,1]=v0ParOLS[4];tablaR2[1,2]=v1ParOLS[4];tablaR2[1,3]=v2ParOLS[4];tablaR2[1,4]=v3ParOLS[4]
tablaR2[2,1]=v0ParGINI[4];tablaR2[2,2]=v1ParGINI[4];tablaR2[2,3]=v2ParGINI[4];tablaR2[2,4]=v3ParGINI[4]
tablaR2[3,1]=v0ParLAV[4];tablaR2[3,2]=v1ParLAV[4];tablaR2[3,3]=v2ParLAV[4];tablaR2[3,4]=v3ParLAV[4]
tablaR2[4,1]=v0ParMHuber[4];tablaR2[4,2]=v1ParMHuber[4];tablaR2[4,3]=v2ParMHuber[4];tablaR2[4,4]=v3ParMHuber[4]
```

```
tablaR2[5,1]=v0ParMM[4];tablaR2[5,2]=v1ParMM[4];tablaR2[5,3]=v2ParMM[4];tablaR2[5,4]=v3ParMM[4]

tablaVP[1,1]=v0valorpOLS;tablaVP[1,2]=v1valorpOLS;tablaVP[1,3]=v2valorpOLS;tablaVP[1,4]=v3valorpOLS
tablaVP[2,1]=v0valorpGINI;tablaVP[2,2]=v1valorpGINI;tablaVP[2,3]=v2valorpGINI;tablaVP[2,4]=v3valorpGINI
tablaVP[3,1]=v0valorpLAV;tablaVP[3,2]=v1valorpLAV;tablaVP[3,3]=v2valorpLAV;tablaVP[3,4]=v3valorpLAV
tablaVP[4,1]=v0valorpMHuber;tablaVP[4,2]=v1valorpMHuber;tablaVP[4,3]=v2valorpMHuber;tablaVP[4,4]=v3valorpMHuber
tablaVP[5,1]=v0valorpMM;tablaVP[5,2]=v1valorpMM;tablaVP[5,3]=v2valorpMM;tablaVP[5,4]=v3valorpMM

tablaEC[1,1]=v0ParOLS[5];tablaEC[1,2]=v1ParOLS[5];tablaEC[1,3]=v2ParOLS[5];tablaEC[1,4]=v3ParOLS[5]
tablaEC[2,1]=v0ParGINI[5];tablaEC[2,2]=v1ParGINI[5];tablaEC[2,3]=v2ParGINI[5];tablaEC[2,4]=v3ParGINI[5]
tablaEC[3,1]=v0ParLAV[5];tablaEC[3,2]=v1ParLAV[5];tablaEC[3,3]=v2ParLAV[5];tablaEC[3,4]=v3ParLAV[5]
tablaEC[4,1]=v0ParMHuber[5];tablaEC[4,2]=v1ParMHuber[5];tablaEC[4,3]=v2ParMHuber[5];tablaEC[4,4]=v3ParMHuber[5]
tablaEC[5,1]=v0ParMM[5];tablaEC[5,2]=v1ParMM[5];tablaEC[5,3]=v2ParMM[5];tablaEC[5,4]=v3ParMM[5]

if(n==10){
n1tablaB0=tablaB0
n1tablaB1=tablaB1
n1tablaMSE=tablaMSE
n1tablaR2=tablaR2
n1tablaEC=tablaEC
n1tablaVP=tablaVP
}
if(n==30){
n2tablaB0=tablaB0
n2tablaB1=tablaB1
n2tablaMSE=tablaMSE
n2tablaR2=tablaR2
n2tablaEC=tablaEC
n2tablaVP=tablaVP
}
if(n==100){
n3tablaB0=tablaB0
n3tablaB1=tablaB1
n3tablaMSE=tablaMSE
n3tablaR2=tablaR2
n3tablaEC=tablaEC
n3tablaVP=tablaVP
}

print(c("obs=", n))
print("Intercepto")
print(tablaB0)
print("Coeficiente x1")
print(tablaB1)
print("MSE")
print(tablaMSE)
print("R2")
print(tablaR2)
print("Valor P")
print(tablaVP)
print("Estadístico Muestral")
print(tablaEC)
}# fin de for para n (observaciones)

#Gráficas de los escenarios*************

par(mfrow=c(2,2))
plot(X,Y0+e,xlim=c(min(X),1.5*max(X)),ylim=c(min(Y),max(Y)),
xlab="Variable independiente(X)",ylab="Variable dependiente(Y)")
mtext(side = 3,expression("Esc 0: Normalidad"), line = 1,cex.axis=0.8)
# normalidad
```

```
Yols=n1tablaB0[1,1]+n1tablaB1[1,1]*X
Ygin=n1tablaB0[2,1]+n1tablaB1[2,1]*X
Ylav=n1tablaB0[3,1]+n1tablaB1[3,1]*X
Ymh=n1tablaB0[4,1]+n1tablaB1[4,1]*X
Ym=n1tablaB0[5,1]+n1tablaB1[5,1]*X
points(X,Yols,type="l",col=1)
points(X,Ygin,type="l",col=2)
points(X,Ylav,type="l",col=3)
points(X,Ymh,type="l",col=4)
points(X,Ym,type="l",col=5)
legend("right",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), lty=1,bty="n")

plot(X[1:(n-1)],Y[1:(n-1)],xlim=c(min(X),1.5*max(X)),ylim=c(min(Y),max(Y)),
xlab="Variable independiente(X)",ylab="Variable dependiente(Y)")

mtext(side = 3,expression("Esc 1: Atípico a 4" ~sigma), line = 1,cex.axis=0.8)
Yols=n1tablaB0[1,2]+n1tablaB1[1,2]*X
Ygin=n1tablaB0[2,2]+n1tablaB1[2,2]*X
Ylav=n1tablaB0[3,2]+n1tablaB1[3,2]*X
Ymh=n1tablaB0[4,2]+n1tablaB1[4,2]*X
Ym=n1tablaB0[5,2]+n1tablaB1[5,2]*X
points(X,Yols,type="l",col=1)
points(X,Ygin,type="l",col=2)
points(X,Ylav,type="l",col=3)
points(X,Ymh,type="l",col=4)
points(X,Ym,type="l",col=5)
legend("right",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), lty=1,bty="n")
points(X[n],Y0[n]+out,col="red")

plot(X[1:(n-1)],Y[1:(n-1)],xlim=c(min(X),1.5*max(X)),ylim=c(min(Y),max(Y)),
xlab="Variable independiente(X)",ylab="Variable dependiente(Y)")
mtext(side = 3,expression("Esc 2: Atípico a 8" ~sigma), line = 1,cex.axis=0.8)
Yols=n1tablaB0[1,3]+n1tablaB1[1,3]*X
Ygin=n1tablaB0[2,3]+n1tablaB1[2,3]*X
Ylav=n1tablaB0[3,3]+n1tablaB1[3,3]*X
Ymh=n1tablaB0[4,3]+n1tablaB1[4,3]*X
Ym=n1tablaB0[5,3]+n1tablaB1[5,3]*X
points(X,Yols,type="l",col=1)
points(X,Ygin,type="l",col=2)
points(X,Ylav,type="l",col=3)
points(X,Ymh,type="l",col=4)
points(X,Ym,type="l",col=5)
legend("right",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), lty=1,bty="n")
points(X[n],Y0[n]+2*out,col="red")

plot(X[1:(n-1)],Y[1:(n-1)],xlim=c(min(X),1.5*max(X)),ylim=c(min(Y),max(Y)),
xlab="Variable independiente(X)",ylab="Variable dependiente(Y)")
mtext(side = 3,expression("Esc 3: Atípico a 16" ~sigma), line = 1,cex.axis=0.8)
Yols=n1tablaB0[1,4]+n1tablaB1[1,4]*X
Ygin=n1tablaB0[2,4]+n1tablaB1[2,4]*X
Ylav=n1tablaB0[3,4]+n1tablaB1[3,4]*X
Ymh=n1tablaB0[4,4]+n1tablaB1[4,4]*X
Ym=n1tablaB0[5,4]+n1tablaB1[5,4]*X
points(X,Yols,type="l",col=1)
points(X,Ygin,type="l",col=2)
points(X,Ylav,type="l",col=3)
points(X,Ymh,type="l",col=4)
points(X,Ym,type="l",col=5)
legend("right",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), lty=1,bty="n")
points(X[n],Y0[n]+4*out,col="red")
```

```
#### Gráficas de los coeficientes"

par(mfrow=c(2,3))
ejx=c(1,2,3,4)
# beta0 para n1
plot(ejx,ylim=c(min(n1tablaB0),max(n1tablaB0)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n1tablaB0[1,],type="o",pch=21,col=1)
points(ejx,n1tablaB0[2,],type="o",pch=22,col=2)
points(ejx,n1tablaB0[3,],type="o",pch=23,col=3)
points(ejx,n1tablaB0[4,],type="o",pch=24,col=4)
points(ejx,n1tablaB0[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3,expression("Coeficiente"~hat(beta[0])~"; n ="~10),line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")
# beta0 para n2
plot(ejx,ylim=c(min(n2tablaB0),max(n2tablaB0)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n2tablaB0[1,],type="o",pch=21,col=1)
points(ejx,n2tablaB0[2,],type="o",pch=22,col=2)
points(ejx,n2tablaB0[3,],type="o",pch=23,col=3)
points(ejx,n2tablaB0[4,],type="o",pch=24,col=4)
points(ejx,n2tablaB0[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3,expression("Coeficiente"~hat(beta[0])~"; n ="~30), line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")
# beta0 para n3
plot(ejx,ylim=c(min(n3tablaB0),max(n3tablaB0)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n3tablaB0[1,],type="o",pch=21,col=1)
points(ejx,n3tablaB0[2,],type="o",pch=22,col=2)
points(ejx,n3tablaB0[3,],type="o",pch=23,col=3)
points(ejx,n3tablaB0[4,],type="o",pch=24,col=4)
points(ejx,n3tablaB0[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression
("Esc2:8"~sigma),expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3,expression("Coeficiente"~hat(beta[0])~"; n ="~100),
line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5),
pch=21:25, lty=1,bty="n")

# beta1 para n1
plot(ejx,ylim=c(min(n1tablaB1),max(n1tablaB1)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n1tablaB1[1,],type="o",pch=21,col=1)
points(ejx,n1tablaB1[2,],type="o",pch=22,col=2)
points(ejx,n1tablaB1[3,],type="o",pch=23,col=3)
points(ejx,n1tablaB1[4,],type="o",pch=24,col=4)
points(ejx,n1tablaB1[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3,expression("Coeficiente"~hat(beta[1])~"; n ="~10), line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")

# beta1 para n2
plot(ejx,ylim=c(min(n2tablaB1),max(n2tablaB1)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n2tablaB1[1,],type="o",pch=21,col=1)
```

```
points(ejx,n2tablaB1[2,],type="o",pch=22,col=2)
points(ejx,n2tablaB1[3,],type="o",pch=23,col=3)
points(ejx,n2tablaB1[4,],type="o",pch=24,col=4)
points(ejx,n2tablaB1[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3,expression("Coeficiente"~hat(beta[1])~"; n ="~30), line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")

# beta1 para n3
plot(ejx,ylim=c(min(n3tablaB1),max(n3tablaB1)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n3tablaB1[1,],type="o",pch=21,col=1)
points(ejx,n3tablaB1[2,],type="o",pch=22,col=2)
points(ejx,n3tablaB1[3,],type="o",pch=23,col=3)
points(ejx,n3tablaB1[4,],type="o",pch=24,col=4)
points(ejx,n3tablaB1[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3,expression("Coeficiente"~hat(beta[1])~"; n ="~100), line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")

par(mfrow=c(2,3))
# mse para n1
plot(ejx,ylim=c(min(n1tablaMSE),max(n1tablaMSE)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n1tablaMSE[1,],type="o",pch=21,col=1)
points(ejx,n1tablaMSE[2,],type="o",pch=22,col=2)
points(ejx,n1tablaMSE[3,],type="o",pch=23,col=3)
points(ejx,n1tablaMSE[4,],type="o",pch=24,col=4)
points(ejx,n1tablaMSE[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3, expression("Error MSE"~(hat(sigma)^2)~"; n ="~10), line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")

# mse para n2
plot(ejx,ylim=c(min(n2tablaMSE),max(n2tablaMSE)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n2tablaMSE[1,],type="o",pch=21,col=1)
points(ejx,n2tablaMSE[2,],type="o",pch=22,col=2)
points(ejx,n2tablaMSE[3,],type="o",pch=23,col=3)
points(ejx,n2tablaMSE[4,],type="o",pch=24,col=4)
points(ejx,n2tablaMSE[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3, expression("Error MSE"~(hat(sigma)^2)~"; n ="~30), line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")

# mse para n3
plot(ejx,ylim=c(min(n3tablaMSE),max(n3tablaMSE)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n3tablaMSE[1,],type="o",pch=21,col=1)
points(ejx,n3tablaMSE[2,],type="o",pch=22,col=2)
points(ejx,n3tablaMSE[3,],type="o",pch=23,col=3)
points(ejx,n3tablaMSE[4,],type="o",pch=24,col=4)
points(ejx,n3tablaMSE[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3, expression("Error MSE"~(hat(sigma)^2)~"; n ="~100), line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")
```

```
# R2 para n1

plot(ejx,ylim=c(min(n1tablaR2),max(n1tablaR2)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n1tablaR2[1,],type="o",pch=21,col=1)
points(ejx,n1tablaR2[2,],type="o",pch=22,col=2)
points(ejx,n1tablaR2[3,],type="o",pch=23,col=3)
points(ejx,n1tablaR2[4,],type="o",pch=24,col=4)
points(ejx,n1tablaR2[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3, expression("Coeficiente"~R^2~"; n ="~10), line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")
# R2 para n2
plot(ejx,ylim=c(min(n2tablaR2),max(n2tablaR2)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n2tablaR2[1,],type="o",pch=21,col=1)
points(ejx,n2tablaR2[2,],type="o",pch=22,col=2)
points(ejx,n2tablaR2[3,],type="o",pch=23,col=3)
points(ejx,n2tablaR2[4,],type="o",pch=24,col=4)
points(ejx,n2tablaR2[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3, expression("Coeficiente"~R^2~"; n ="~30), line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")
# R2 para n3
plot(ejx,ylim=c(min(n3tablaR2),max(n3tablaR2)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n3tablaR2[1,],type="o",pch=21,col=1)
points(ejx,n3tablaR2[2,],type="o",pch=22,col=2)
points(ejx,n3tablaR2[3,],type="o",pch=23,col=3)
points(ejx,n3tablaR2[4,],type="o",pch=24,col=4)
points(ejx,n3tablaR2[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3, expression("Coeficiente"~R^2~"; n ="~100), line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")

par(mfrow=c(2,3))
# EC para n1
plot(ejx,ylim=c(min(n1tablaEC),max(n1tablaEC)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n1tablaEC[1,],type="o",pch=21,col=1)
points(ejx,n1tablaEC[2,],type="o",pch=22,col=2)
points(ejx,n1tablaEC[3,],type="o",pch=23,col=3)
points(ejx,n1tablaEC[4,],type="o",pch=24,col=4)
points(ejx,n1tablaEC[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3, expression("Estadístico"~chi^2~"; n ="~10), line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")

# EC para n2
plot(ejx,ylim=c(min(n2tablaEC),max(n2tablaEC)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n2tablaEC[1,],type="o",pch=21,col=1)
points(ejx,n2tablaEC[2,],type="o",pch=22,col=2)
points(ejx,n2tablaEC[3,],type="o",pch=23,col=3)
points(ejx,n2tablaEC[4,],type="o",pch=24,col=4)
points(ejx,n2tablaEC[5,],type="o",pch=25,col=5)
```

```
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3, expression("Estadístico"~chi^2~"; n ="~30), line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")
# EC para n3
plot(ejx,ylim=c(min(n3tablaEC),max(n3tablaEC)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n3tablaEC[1,],type="o",pch=21,col=1)
points(ejx,n3tablaEC[2,],type="o",pch=22,col=2)
points(ejx,n3tablaEC[3,],type="o",pch=23,col=3)
points(ejx,n3tablaEC[4,],type="o",pch=24,col=4)
points(ejx,n3tablaEC[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3, expression("Estadístico"~chi^2~"; n ="~100), line = 1,cex.axis=0.8)
legend("left",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")

# Vp para n1
plot(ejx,ylim=c(min(n1tablaVP),max(n1tablaVP)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n1tablaVP[1,],type="o",pch=21,col=1)
points(ejx,n1tablaVP[2,],type="o",pch=22,col=2)
points(ejx,n1tablaVP[3,],type="o",pch=23,col=3)
points(ejx,n1tablaVP[4,],type="o",pch=24,col=4)
points(ejx,n1tablaVP[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3, expression("Valor-p ajuste"~chi[2]^2~"; n ="~10), line = 1,cex.axis=0.8)
legend("right",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")
# Vp para n2
plot(ejx,ylim=c(min(n2tablaVP),max(n2tablaVP)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n2tablaVP[1,],type="o",pch=21,col=1)
points(ejx,n2tablaVP[2,],type="o",pch=22,col=2)
points(ejx,n2tablaVP[3,],type="o",pch=23,col=3)
points(ejx,n2tablaVP[4,],type="o",pch=24,col=4)
points(ejx,n2tablaVP[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3, expression("Valor-p ajuste"~chi[2]^2~"; n ="~30), line = 1,cex.axis=0.8)
legend("right",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")
# Vp para n3

plot(ejx,ylim=c(min(n3tablaVP),max(n3tablaVP)),type="n",axes=F,ann=F)
axis(side=2,las=1)
points(ejx,n3tablaVP[1,],type="o",pch=21,col=1)
points(ejx,n3tablaVP[2,],type="o",pch=22,col=2)
points(ejx,n3tablaVP[3,],type="o",pch=23,col=3)
points(ejx,n3tablaVP[4,],type="o",pch=24,col=4)
points(ejx,n3tablaVP[5,],type="o",pch=25,col=5)
axis(1,at=ejx,lab=c("Esc0:Normal",expression("Esc1:4"~sigma),expression("Esc2:8"~sigma),
expression("Esc3:16"~sigma)),cex.axis=0.8,las=1)
mtext(side = 3, expression("Valor-p ajuste"~chi[2]^2~"; n ="~100), line = 1,cex.axis=0.8)
legend("right",c("OLS","GINI","LAV","M_HUBER","MM"),cex=0.8,col=c(1:5), pch=21:25, lty=1,bty="n")

# dendograma con cluster del MSE
require(graphics)


par(mfrow=c(2,2))
hc<-hclust(dist(n1tablaMSE,method="euclidean"))
```

```
plot(hc,labels=rownames(n1tablaMSE), ann = F)
mtext(side = 2, "Distancia Euclidiana", line = 3)
title(xlab="a) Clúster de métodos según su MSE para n=10")

hc<-hclust(dist(n2tablaMSE,method="euclidean"))
plot(hc,labels=rownames(n2tablaMSE), ann = F)
mtext(side = 2, "Distancia Euclidiana", line = 3)
title(xlab="b) Clúster de métodos según su MSE para n=30")

hc<-hclust(dist(n3tablaMSE,method="euclidean"))
plot(hc,labels=rownames(n3tablaMSE), ann = F)
mtext(side = 2, "Distancia Euclidiana", line = 3)
title(xlab="c) Clúster de métodos según su MSE para n=100")
```