UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2015, Vol. 8, No. 2, pp. 129–161

Comparación de la regresión GINI con la regresión de Mínimos cuadrados ordinarios y otros modelos de regresión lineal robustos[1]

Comparison of Gini Regression with OLS Regression and other Robust Linear Regression

Juan Carlos Correa Morales[a]
jccorrea@unal.edu.co

Gloria Patricia Carmona[b]
gpcarmonita@gmail.com

Resumen

En este trabajo se compara la regresión de Gini 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. Los resultados encontrados vía simulación muestran 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*:** Datos atípicos, eficiencia, míminos cuadrados ordinarios, modelos de regresión robustos, regresión Gini, robustez.

Abstract

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

[1]Correa, J.C., Carmona, G.P. (2015). Comparación de la Regresión GINI con la Regresión de Mínimos Cuadrados Ordinarios y otros Modelos de Regresión Lineal Robustos. *Comunicaciones en Estadística,* **8**(2), 129-161.

[a]Profesor asociado. Universidad Nacional de Colombia Facultad de Ciencias, Escuela de Estadística. Colombia.

[b]Estadística. Universidad Nacional de Colombia Facultad de Ciencias, Escuela de Estadística. Colombia

of the methods is performed via simulation under different scenarios. 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*:** Least Squares Estimation, Gini Regression, Robust Regression Model, Efficiency, Robustness, Outliers.

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, las ciencias físicas y químicas, la economía, la administración, las ciencias biológicas y las 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 aproximación 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 conocimiento del mecanismo básico (Montgomery, Peck & Vining, 2002).

Considere el modelo de regresión lineal:

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

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. 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. De esta manera 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_j. 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 estas, 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 & Vining, 2002). En el presente trabajo se mencionarán algunas de ellas. Una es la que se realiza mediante pruebas de detección de datos atípicos (DFFITS, DFBETAS, R de Student, etc.) para excluirlos del modelo de regresión, lo que 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), que 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 & Monari 2008) usando una medida de dispersión alternativa denominada Diferencia Media de Gini (GMD), que en la actualidad no es muy conocida o utilizada y por ende no está implementada en los paquetes estadísticos como R (R Core Team 2013). En este trabajo, se pretende investigar cuán robustos son los estimadores de los coeficientes de regresión mediante la metodología de Gini, que 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 en la sección 2 una revisión de metodologías de regresión, en concreto, se aborda la regresión OLS; luego se presentan los métodos de estimación robustos como MM, M de Huber y LAV. En la sección 3 se presenta la regresión de Gini, que se divide en dos partes: estimación de los coeficientes mediante promedios ponderados de pendientes (enfoque no paramétrico) y estimación por minimización (enfoque paramétrico). En la sección 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 la sección 5 se presenta un ejemplo aplicado a datos reales. Finalmente se presentan las conclusiones y recomendaciones para futuras investigaciones.

2. 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 1795 y por Adrien Marie Legendre en Francia en 1805. Inicialmente las aplicaciones del método OLS fueron para datos astronómicos y geográficos (Birkes & Dodge 1993).

Se podría representar esa relación lineal mediante una ecuación de línea recta (Montgomery, Peck & Vining 2002)

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

La expresión matemática de los estimadores de OLS es:

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

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

3. 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 & 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 aquel 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 & Hettmansperger (1978), Birkes & Dodge (1993), Huber (1973), Denby & Larsen (1977), Huber (1981), Hurdle (1981), Hettmansperger (1984), Lachan (1985), Hampel, Ronchetti, Rousseeuw & Stahel (1986), Ronald & Montgomery (1984), Staudte & Sheather (1990), Wilcox, (2005), Bianco, Garcia & 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 y serán comparados con el método Gini y OLS.

3.1. 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 $\rho(\varepsilon) = \frac{1}{2}\varepsilon^2$ con $-\infty < \varepsilon < \infty$. 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:

$$\mathbf{y_i} = \mathbf{x_i'}\beta + \varepsilon_\mathbf{i} \tag{3}$$

Esta regresión se da para la i-ésima observación de n valores independientes, pero en este caso la distribución de los errores puede ser de colas pesada y producir valores atípicos.

Dado un estimador $\hat{\beta}$ para β, el modelo de regresión ajustado es:

$$\hat{\mathbf{y}}_\mathbf{i} = \mathbf{x_i'}\hat{\beta} \tag{4}$$

Por su parte, los residuales están dados por:

$$\varepsilon_\mathbf{i} = \mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i} = \mathbf{y_i} - \mathbf{x_i'}\hat{\beta} \tag{5}$$

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

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

Se utilizan diferentes ponderaciones para ello, según la magnitud del error[1].

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 que 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 las figuras 1 y 2 se presentan algunas funciones de criterio robusto $\rho(z)$ de uso frecuente. El comportamiento de cada función ρ y de su derivada ψ correspondiente.

3.2. 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 $\rho(\varepsilon/s)$ de manera 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

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

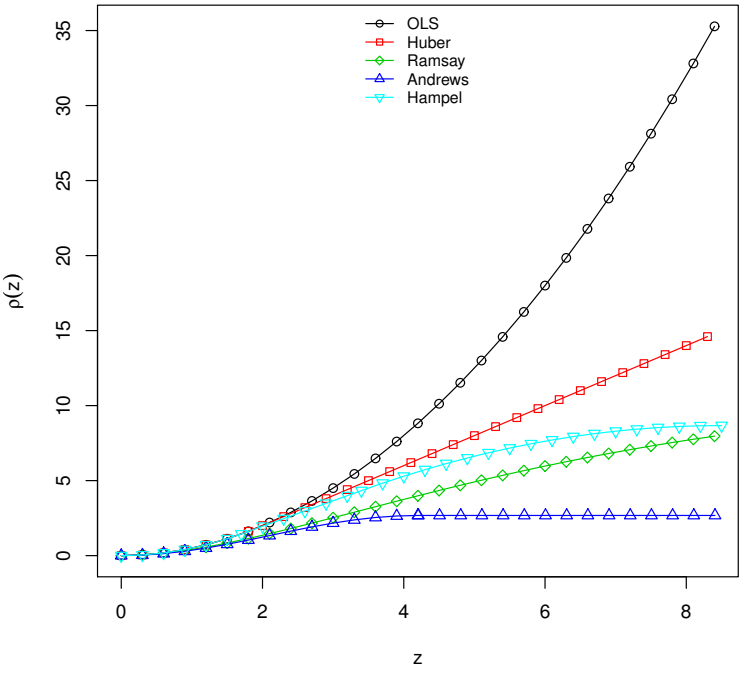


Figura 1: *Funciones de criterio robusto para ponderar datos atípicos. Fuente: Elaboración propia.*

los residuales, el método de estimación S minimiza la dispersión de los residuales $\underset{\beta}{Min}\, s[\varepsilon_1(\beta), \varepsilon_2(\beta), \ldots, \varepsilon_n(\beta)]$. Rousseeuw y Leroy (1987: 135-136) definen la función de dispersión, S, de los residuales $\varepsilon_i(\beta)$ y se determina esta solución

$$\sum_{i=1}^{n} \rho\left(\frac{\varepsilon_i}{s}\right) = k \tag{6}$$

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, $\frac{k}{\rho(c)} = \frac{1}{2}$.

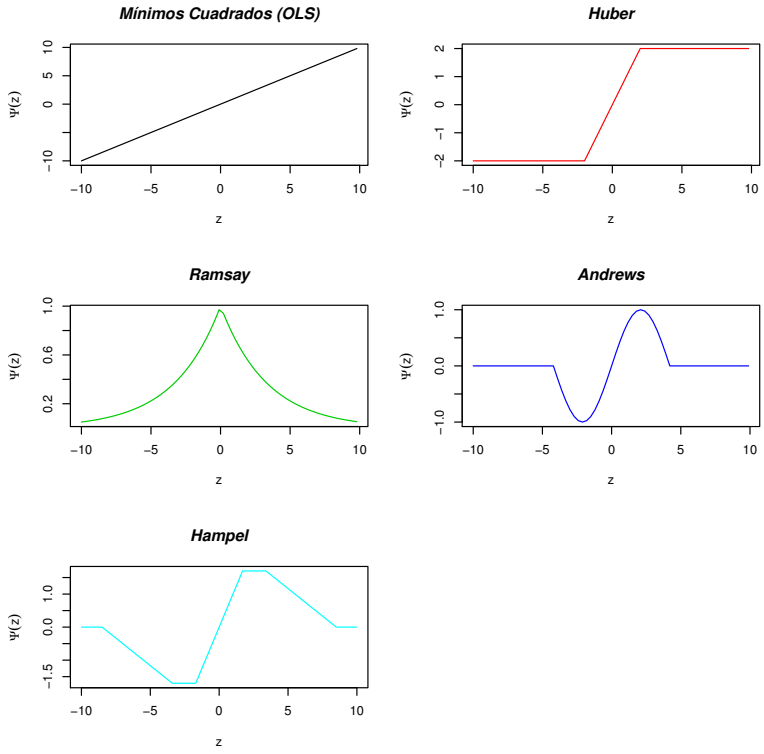


Figura 2: *Funciones robustas de influencia. Fuente: Elaboración propia.*

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

$$\rho(z) = \begin{cases} \frac{z^2}{2} - \frac{z^4}{2c^2} + \frac{z^6}{6c^4} & si \quad |z| \leq c \\ \frac{c^2}{6} & si \quad |z| > c \end{cases} \tag{7}$$

3.3. 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 & Ronchetti (2002) y 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, que 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 residuales $\varepsilon_i(\tilde{\beta}) = (y_i - x_i'\tilde{\beta})$.

2. En este segundo paso se calcula un estimado M con punto de quiebre del 50 %, usando los residuales del ajuste robusto (del paso 1) y la ecuación (6). 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, un estimador MM de $\hat{\beta}$ se define como una solución a esta ecuación:

$$\sum_{i=1}^{n} x_{ij} \psi_1 \left(\frac{y_i - x_i'\beta}{s_n} \right) = 0 \ , \quad j = 1, ..., k+1. \tag{8}$$

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

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

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

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

Una condición final que debe satisfacer la solución dada por la ecuación (8) es esta

$$\sum_{i=1}^{n} \rho_1 \left(\frac{y_i - x_i'\hat{\beta}}{s_n} \right) \leq \sum_{i=1}^{n} \rho_1 \left(\frac{y_i - x_i'\tilde{\beta}}{s_n} \right) \tag{9}$$

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 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`.

3.4. Estimadores L

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

3.4.1. Regresion LAV

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

$$\underset{\beta}{Min} \sum_{i=1}^{n} |\varepsilon_i| = \underset{\beta}{Min} \sum_{i=1}^{n} |y_i - x_i\beta| \tag{10}$$

El método de regresión LAV puede ser visto como un caso especial de la regresión cuantil generalizada. En este caso, la función por minimizar es esta

$$\sum_{i=1}^{n} \rho_\alpha(\varepsilon_i) \tag{11}$$

Donde

$$\rho_\alpha(\varepsilon_i) = \begin{cases} \alpha\varepsilon_i, & si \quad \varepsilon_i \geq 0 \\ (\alpha - 1)\varepsilon_i, & si \quad \varepsilon_i < 0 \end{cases} \tag{12}$$

Donde α es el cuantil que debe 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).

4. Regresión 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 & 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).

[2] "Least" por sus siglas en ingles

4.1. Diferencia media de Gini-GMD

Olkin & 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 & Yitzhaki, 1992), algunas de las cuales son estas:

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

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

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 Gini puede ser calculada desde dos enfoques:

1. El primer enfoque es considerado como un procedimiento *semiparamétrico*, que 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 semiparamétrico 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.

4.2. Enfoque semiparamétrico

Sean (Y, X) variables aleatorias bivariadas que sigue una distribución continua con primer y segundo momento finito. 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 así:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{13}$$

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

$$\varepsilon = Y - \hat{Y} = Y - \hat{\beta}_0 - \hat{\beta}_1 X. \tag{14}$$

Note que (14) 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:

$$Cov(Y, X) = Cov(\beta_0 + \beta_1 X + \varepsilon, X) = \beta_1 Cov(X, X) + Cov(\varepsilon, X). \tag{15}$$

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

$$\hat{\beta}_1 = \frac{\text{Cov}(Y, X)}{\text{Cov}(X, X)} \tag{16}$$

El $\hat{\beta}_1$ de la ecuación (16) 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 (16) y reemplazando 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 semiparamétrico de Gini:

$$\hat{\beta}_{GN} = \frac{\text{Cov}(Y, F(X))}{\text{Cov}(X, F(X))}. \tag{17}$$

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

$$\text{Cov}(\varepsilon_{GN}, F(X)) = 0 \tag{18}$$

Donde ε_{GN} son lo residuales de la regresión semiparamétrica de Gini (Yitzhaki & Schechtman, 2013). La ecuación (18) 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 (13), 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 (19):

$$\mu_y = \hat{\beta}_0 + \hat{\beta}_{GN}\mu_x. \tag{19}$$

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, que es determinado como un parámetro de localización (Yitzhaki & Schechtman, 2013). La regresión de Gini ya ha sido trabajada y aplicada por algunos autores, como David (1968), Meintanis, & Donatos,(1997), Mosteller & Tukey (1983), Yitzhaki (1998), Schechtman & Yitshaki (2003), Yitzhaki (2003), Yitzhaki & Schechtman (2004), Yitzhaki & Schechtman (2005), Borroni & Zenga (2006), Borroni & Cazzaro (2006), Edna & Yitzhaki (2007), Edna, Yitzhaki, & Artsev (2008), Edna & Yitzhaki (2011), Edna, Yitzhaki & Pudalov (2013), Yitzhaki y Schechtman (2013).

4.3. Enfoque semiparamétrico de la regresión Gini

El enfoque semiparamétrico 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 semiparamétrico de OLS y de Gini no requieren la especificación de una forma funcional del modelo; estos pueden ser usados cuando el investigador está 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 semiparamétrico serán denotados por el subíndice GN. Se hace referencia a este enfoque como semiparamétrico porque no depende del supuesto de linealidad entre las variables ni del supuesto 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 & Schechtman, 2013).

A continuación se muestra cómo el coeficiente de regresión de Gini también puede ser expresado como una suma ponderada de pendientes de la recta de regresión (Yitzhaki & Schechtman, 2013):

$$\hat{\beta}_{GN} = \int v(x)g'(x)dx \tag{20}$$

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 que aplica para el caso discreto y continuo:

$$v(x) = \frac{[1 - F_X(x)]F_X(x)}{\int_{-\infty}^{\infty}[1 - F_X(t)]F_X(t)dt} \tag{21}$$

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

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

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 con X. Los pesos asignados a cada pendiente b_i están dados por la siguiente expresión

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

4.4. 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, que depende de dos factores:

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 y de Gini. 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 que se denota 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 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 semiparamétrica de los coeficientes de regresión de OLS y Gini debido a que son expresados como promedio ponderado de observaciones adyacentes (Yitzhaki & 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** compara las funciones ponderadoras w_x de OLS y v_x de Gini, en función del rango de X, (i), X con distribución uniforme ($\Delta X = c = 1$) y para una distribución de X no uniforme.

De la figura anterior, se observa que para el esquema de ponderación por Gini y OLS, en el caso de que X se distribuye uniforme ($\Delta X = c = 1$), los pesos asignados por Gini y OLS son los mismos; por ello, ambos métodos son igualmente sensibles a datos atípicos (Yitzhaki & Schechtman, 2013). Asimismo se observa que la ponderación de ambos métodos difiere en el caso de que la X no es uniforme, siendo más sensible OLS que Gini, como era de esperarse.

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 que 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.

4.5. 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 que 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 & 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 subíndice GM. Este tipo de regresiones han sido desarrollados por Jureckova(1969, 1971), Jaeckel (1972), McKean & 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

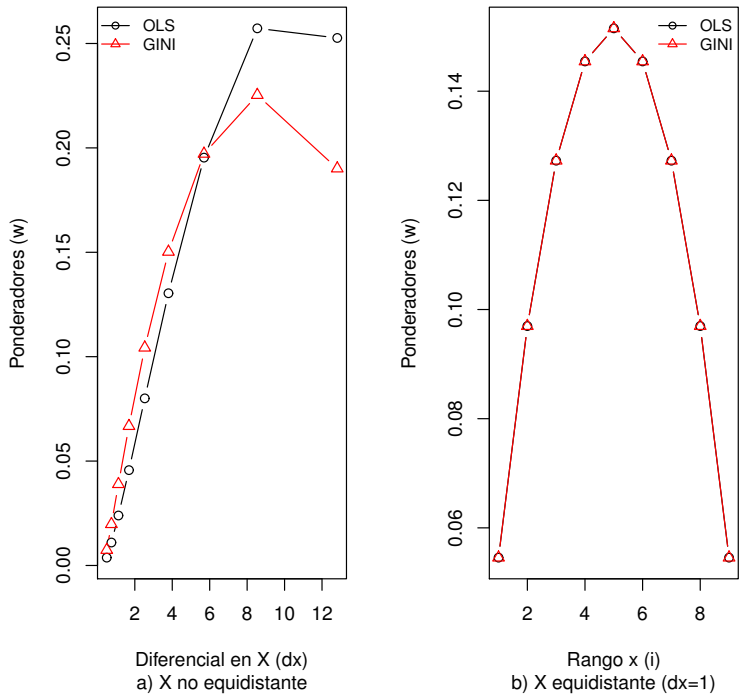


Figura 3: *Ponderadores de OLS y Gini para una X uniforme y no uniforme. Fuente: Elaboración propia.*

los residuales y los residuales.

Considere el siguiente 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 $\varepsilon_1 \leq \varepsilon_2 \leq \cdots \leq \varepsilon_n$. Entonces se tiene:

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

Del que se obtiene:

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

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, que se cumple si $\text{cov}\left(x, \frac{R_\varepsilon}{n}\right) = 0$.

Esto es una restricción de ortogonalidad (no hay correlación entre ε y x, es decir, $cov(x, F(\varepsilon)) = 0$)

Similar a la regresión cuantil el estimador $\hat{\beta}$ no puede ser expresado explícitamente y debe ser calculado con métodos numéricos.

5. Metodología y resultados

5.1. Escenarios de simulación para comparar los métodos de regresión

De acuerdo con el enfoque no paramétrico (o semiparamétrico) 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, 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: se generan 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: 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 figura 4.
   - 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 figura 4.

2.3. Escenario 3: 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 figura 4.

En cuanto a la distribución de la variable explicativa X, esta 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.

5.2. 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 está 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 (que 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 escenarios (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: 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: 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: 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.

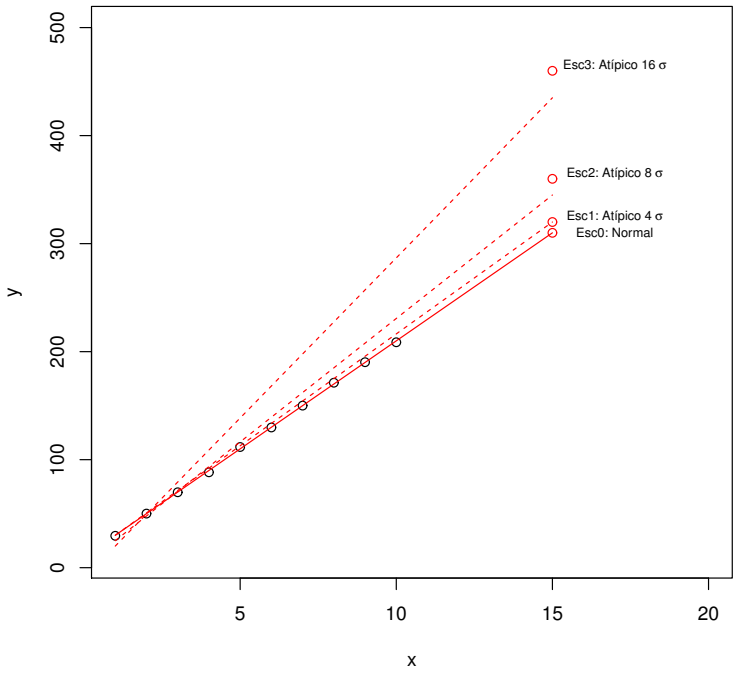


Figura 4: *Escenarios de simulación. Fuente: Elaboración propia.*

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 & Wichern, 1992):

$$\left(\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}-\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}\right)' \Sigma^{-1}_{\underline{\hat{\beta}}} \left(\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}-\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}\right) \sim \chi^2_2$$

Otra forma de escribirlo es:

$$\left(\hat{\beta}-\beta\right)' \frac{1}{\sigma^2}\left(X'X\right)\left(\hat{\beta}-\beta\right) \sim \chi^2_2 \tag{22}$$

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.

5.3. Resultados de la simulación

A continuación se presentan las tablas de resultados obtenidas mediante la simulación Montecarlo descrita anteriormente. 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 que se realizaron las simulaciones para la comparación de los métodos es: $Y=\beta_0+\beta_1 X+\varepsilon$; con $\beta_0=5$, $\beta_1=20$ y $\varepsilon\sim N(\mu,\sigma^2)$.

- 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 figuras 6 y 7.
- 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 figura 6.

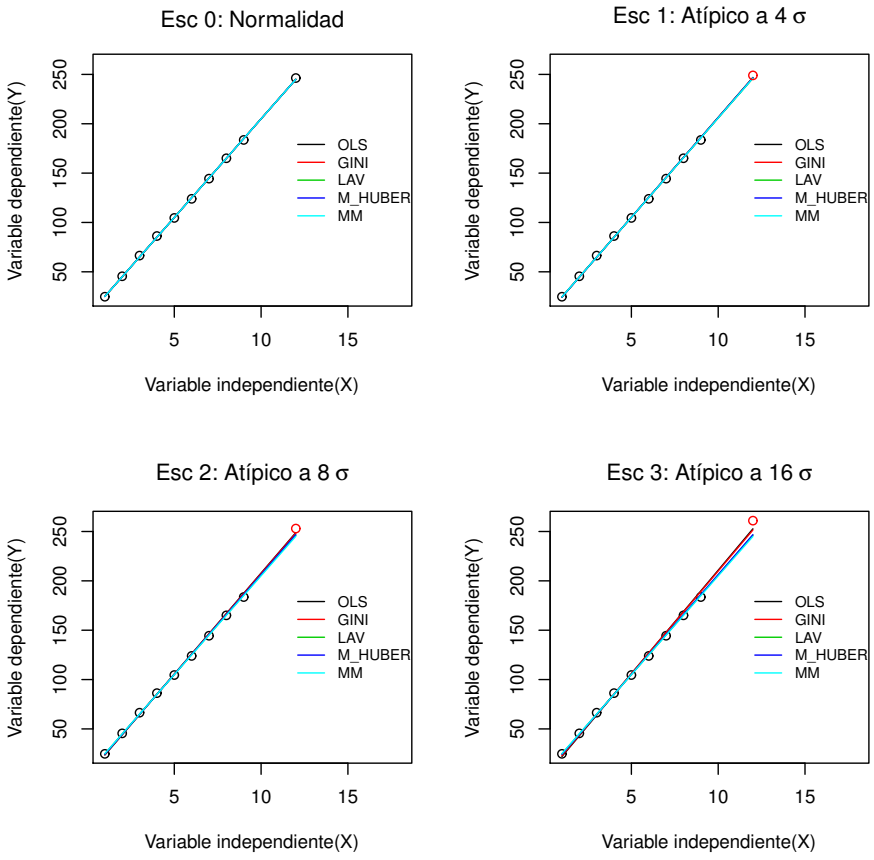


Figura 5: *Modelo de regresión estimado por cada método en cada escenario para* $n = 10$. *Fuente: Elaboración propia.*

- Un análisis sencillo de cuán robusto son los métodos se presenta en las tablas 19 y 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:

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

  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

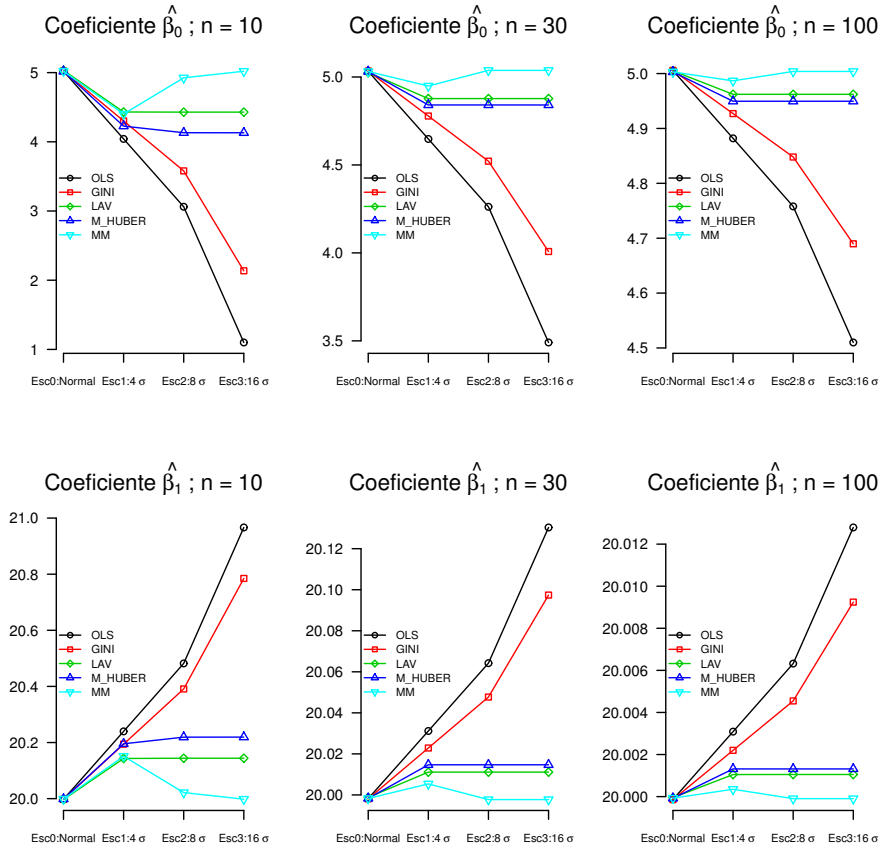


Figura 6: *Parámetros β_0 y β_1 estimados por cada método con diferentes tamaños de muestra en cada escenario. Fuente: Elaboración propia.*

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 figura 7).

- 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 siguiente forma:

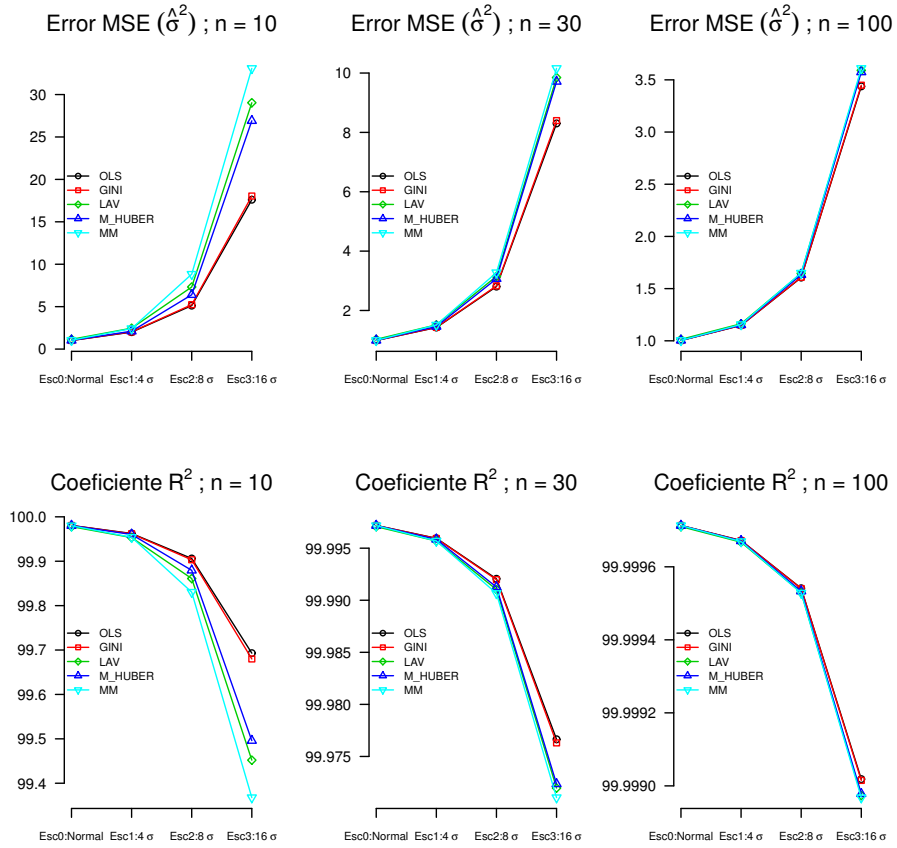


Figura 7: *Parámetros MSE y R^2 estimados por cada método con diferentes tamaños de muestra en cada escenario. Fuente: Elaboración propia.*

$$\left(\hat{\beta}-\beta\right)' \frac{1}{\sigma^2}\left(X'X\right)\left(\hat{\beta}-\beta\right) \sim \chi^2_2 \tag{23}$$

Esta 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.

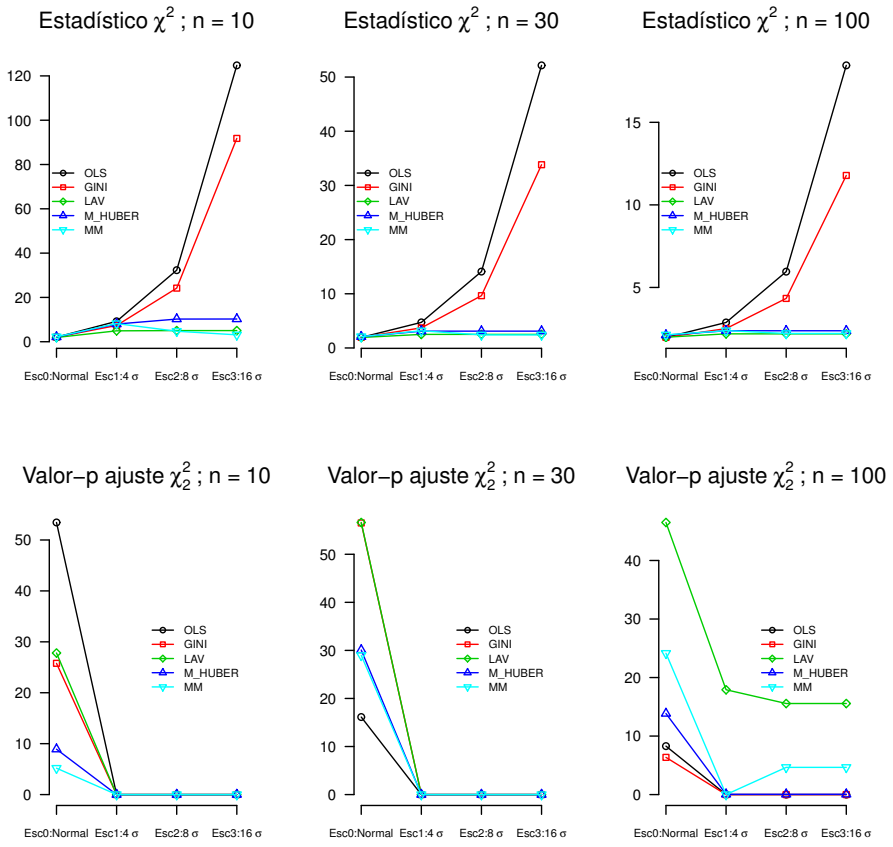
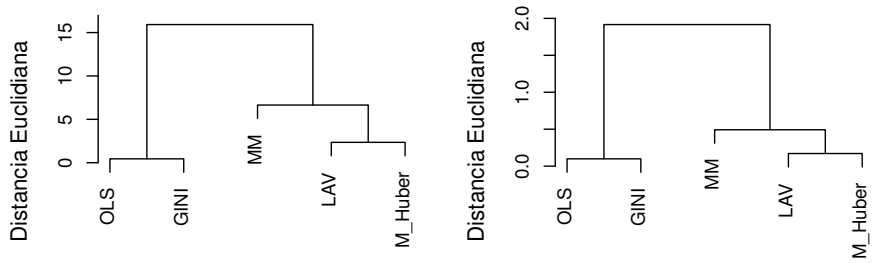


Figura 8: *Estadístico χ^2 y Vp (del ajuste de la distribución χ^2_2) obtenidos por cada método con diferentes tamaños de muestra en cada escenario. Fuente: Elaboración propia.*

- 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.

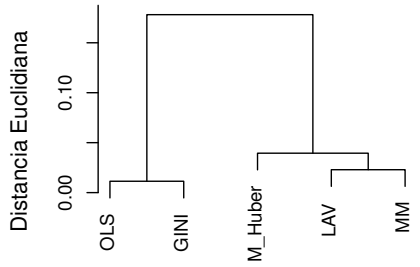
- Con el fin de establecer cuáles métodos son más similares entre sí al momento de estimar los coeficientes de regresión en los diferentes escenarios, se propone agruparlos por su MSE producido tanto en el escenario de normalidad como en el de normalidad contaminada con atípicos. Los resultados muestran que OLS y Gini producirán resultados similares cuando los errores sean normales o cuando están 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 mas a

MM.



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 9: *Cluster de los Métodos según su MSE para* $n = 10$, $n = 30$,$n = 100$. *Fuente: Elaboración propia.*

5.4. Tablas de resultados

Tabla 1: *Estimación de* β_0 *con* $n = 100$

Métodos	Normalidad	4σ	8σ	16σ
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 2: *Estimación de* β_1 *con* $n = 100$

Métodos	Normalidad	4σ	8σ	16σ
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 3: *Estimación del* MSE *con* $n = 100$

Métodos	Normalidad	4σ	8σ	16σ
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: *Estimación del* R^2 *con* $n = 100$

Métodos	Normalidad	4σ	8σ	16σ
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 5: *Valor P con* $n = 100$

Métodos	Normalidad	4σ	8σ	16σ
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 6: *Estadístico calculado* χ^2_2 *con* $n = 100$

Métodos	Normalidad	4σ	8σ	16σ
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 7: *Estimacion de* β_0 *con* $n = 30$

Métodos	Normalidad	4σ	8σ	16σ
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 8: *Estimación de* β_1 *con* $n = 30$

Métodos	Normalidad	4σ	8σ	16σ
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 9: *Estimación del MSE con $n = 30$*

Métodos	Normalidad	4σ	8σ	16σ
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 10: *Estimación del R^2 con $n = 30$*

Métodos	Normalidad	4σ	8σ	16σ
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 11: *Valor P con $n = 30$*

Métodos	Normalidad	4σ	8σ	16σ
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 12: *Estadístico calculado χ^2_2 con $n = 30$*

Métodos	Normalidad	4σ	8σ	16σ
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 13: *Estimación de* β_0 *con* $n = 10$

Métodos	Normalidad	4σ	8σ	16σ
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 14: *Estimación de* β_1 *con* $n = 10$

Métodos	Normalidad	4σ	8σ	16σ
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 15: *Estimación del* MSE *con* $n = 10$

Métodos	Normalidad	4σ	8σ	16σ
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 16: *Estimación del* R^2 *con* $n = 10$

Métodos	Normalidad	4σ	8σ	16σ
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 17: *Valor P con* $n = 10$

Métodos	Normalidad	4σ	8σ	16σ
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 18: *Estadístico calculado* χ^2_2 *con* $n = 10$

Métodos	Normalidad	4σ	8σ	16σ
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 19: *Variación de la estimación de* β_0 *con* $n = 10$

Métodos	4σ	8σ	16σ
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 20: *Variación de la estimación de* β_1 *con* $n = 10$

Métodos	4σ	8σ	16σ
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 %

6. 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 solo 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 se sabía que los métodos MM, M de Huber y LAV eran muchísimo más robustos que Gini, se prosiguió a compararlo solo con el fin de mostrar cuán parecido o diferente era Gini frente a OLS en relación con 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 estas 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.

Recibido: 28 de julio del 2014
Aceptado: 9 de julio del 2015

Referencias

Aelst, V. S., W. G. & Zamar, H. R. (2013), 'Robust and efficient estimation of the residual scale in linear regression.', *Journal of Multivariate Analysis* **116**, 278–296.

Bassett, G., J. & Koenker, R. (1978), 'Asymptotic theory of least absolute error regression.', *Journal of the American Statistical Association* **73**, 618–622.

Bianco, M. A., G. M. B. y. Y. J. (2005), 'Robust estimation for linear regression with asymmetric errors.', *The Canadian Journal of Statistics* **33**, 511–528.

Birkes, D. & Dodge, Y. (1993), *Alternative Methods of Regression*, 1 edn, John Wiley & Sons Inc., New York.

Boente, G. & Fraiman, R. (1989), 'Robust Nonparametric Regression Estimation.', *Journal of Multivariate Analysis* **29**, 180–198.

Borroni, C. G. & Cazzaro, M. (2006), 'Some developments about a new nonparametric test based on Gini's mean difference.', *Statistica & Applicazioni* **3**, 29–44.

Borroni, C. & Zenga, M. (2006), 'A test of concordance based on Gini's mean difference.', *Communications in Statistics - Theory and Methods* **16**, 289–308.

Boscovich, R. (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* **4**, 353–396.

Casella, G. & Berger, R. L. (2002), *Statistical Inference*, 2 edn, Duxbury, New York.

David, H. A. (1968), 'Gini's Mean Difference Rediscovered.', *Biometrika* **55**, 573–575.

Denby, L., a. L. W. A. (1977), 'Robust Regression Estimators Compared via Monte Carlo.', *Communications in Statistics-Theory and Methods* **6**, 335–362.

Douglas, C., M. E. A. P. & Vining, G. G. (2002), *Introduccion al análisis de regresión lineal*, 2 edn, John Wiley & Sons, New York.

E. Schechtman, E., Y. S. (2003), 'A Family of Correlation Coefficients Based on Extended Gini.', *Journal of Economic Inequality* **1**, 3083–3088.

Edna, S., Y. S. & 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* **26**, 329–344.

Edna, S., Y. S. & Pudalov, T. (2011), 'Gini's multiple regressions: two approaches and their Interaction.', *International Journal of Statistics* **69**, 67–99.

Edna, S. & Yitzhaki, S. (2007), ' A Measure Of Association Based On Gini's Mean Difference.', *Communications in Statistics-Theory and Methods* **16**, 207–231.

Edna, S. & Yitzhaki, S. (2013), *Gini's Multiple Regressions*, 1 edn, John Wiley & Sons, New York.

Hampel, F. R., R. E. M. R. P. J. & Stahel, W. A. (1986), *Robust statistics: The approach based on inuence functions.*, 1 edn, Wiley, New York.

Hettmansperger, T. P. (1984), *Statistical inference based on ranks.*, 1 edn, John Wiley & Sons, New York.

Huber, P, J. (1973), ' Regression: Asymptotics, Conjectures, and Monte Carlo.', *Annals of Statistics* **1**, 799–821.

Huber, P. J. (1981), *Robust statistics.*, 1 edn, John Wiley & Sons, New York.

Hurdle, W. (1984), 'Robust Regression Function Estimation.', *Journal of Multivariate Analysis* **14**, 169–180.

Iachan, R. (1985), 'Robust Designs for Ratio and Regression Estimation.', *Journal of Statistical Planning and Inference* **11**, 149–161.

T.E., D. (2005), 'Least Absolute Value Regression: Recent Contributions.', *Journal of Statistical Computation and Simulation* **75**, 263–286.