INVESTIGACIÓN & DESARROLLO, Vol. **20**, No. **2**: 41 – 56 (2020)
ISSN 2518-4431

# TÉCNICAS ROBUSTAS Y NO ROBUSTAS PARA IDENTIFICAR OUTLIERS EN EL ANÁLISIS DE REGRESIÓN

# ROBUST AND NONROBUST TECHNIQUES FOR IDENTIFYING OUTLIERS IN REGRESION ANALYSIS

**Darwin Ugarte Ontiveros y Ruth Marcela Aparicio de Guzman**[1]
*Centro de Investigaciones Económicas y Empresariales*
[1]*Universidad Privada Boliviana*
darwinugarte@upb.edu
(Recibido el 28 de noviembre 2019, aceptado para publicación el 02 de octubre 2020)

## RESUMEN

Verificar si los resultados de un modelo de regresión reflejan el patrón de los datos, o si los mismos se deben a unas cuantas observaciones atípicas (outliers) es un paso importante en el proceso de investigación empírica. Para este propósito resulta aún común apoyarse en procedimientos (estándares) que no son eficaces para este propósito, al sufrir del denominado "masking effect", algunos de ellos sugeridos incluso en los libros tradicionales de econometría. El presente trabajo pretende alertar a la comunidad académica sobre el peligro de implementar estas técnicas estándares, mostrando el pésimo desempeño de las mismas. Asimismo, se sugiere aplicar otras técnicas más idóneas sugeridas en la literatura sobre "estadística robusta" para identificar outliers en el análisis multivariado. Para facilitar la aplicación de las mismas, el trabajo pone a disposición de la comunidad académica un programa en Stata del tipo do-file para identificar y categorizar outliers basado en el trabajo de [1]. Simulaciones de Monte Carlo dan evidencia de la aplicabilidad de la misma.

**Palabras Clave:** Outliers, Estadística Robusta, Análisis de Regresión, Stata.

## ABSTRACT

Checking whether the results of a regression model describe properly the data, or whether they are influenced by few outliers is an important step in the empirical research process. For this purpose, it is still common to rely on procedures which are not effective, as they suffer from the so-called "masking effect", some of them even suggested in traditional econometrics books. This work aims to warn about the danger of implementing these standard techniques, as they have poor performance. Likewise, we suggest applying more suitable techniques suggested in the literature on "robust statistics" to identify outliers in multivariate analysis. To facilitate their application, we present a Stata program (do-file type) to identify and categorize outliers based on the work of [1]. Monte Carlo simulations provide evidence of its applicability.

**Keywords:** Outliers, Robust Regression, Regression Analysis, Stata.

## I. INTRODUCCIÓN

El principal objetivo de la econometría es confrontar la teoría económica con la realidad. Para ello los economistas estiman modelos estadísticos de regresión con el afán de cuantificar el nivel de relación entre las variables. El objetivo de este análisis es identificar cómo una variable dependiente ($Y_{nx1}$) se encuentra relacionada con un conjunto ($p$) de variables explicativas ($X_{nxp}$), estimando el siguiente modelo $Y=X\beta+\varepsilon$ donde $\varepsilon_{nx1}$ es el vector que contiene el término de error y $\beta_{px1}$ es el vector que contiene los parámetros de interés a ser estimados.

Uno de los problemas del análisis de regresión es su excesiva vulnerabilidad a observaciones con valores extremos diferentes a la mayoría de los datos, los llamados outliers. En la práctica unos pocos outliers fácilmente pueden distorsionar las estimaciones de una regresión, obteniéndose parámetros que no reflejen la verdadera relación entre las variables. Una ilustración de este fenómeno se visualiza en la Figura 1, para el caso de un regresor: siendo la relación entre las variables negativa, unos cuantos outliers pueden distorsionar la estimación y puede concluirse erróneamente que la relación es positiva.

Esta fragilidad en la estimación de $\beta$ es independiente del método de estimación que se utilice: (i) si se minimiza la suma del cuadrado de los residuos ($r_{nx1}$), Mínimos Cuadrados Ordinarios ($MCO$), es decir, $\hat{\beta}_{MCO} = \arg\min_{\beta} r'r$, donde, $r = Y - X\hat{\beta}$, entonces una sola observación outlier producirá un residuo al cuadrado con valor extremo que sobredimensionará la medida de error agregado a minimizarse; (ii) De igual manera, si se estimaría por Máxima


**DOI:** 10.23881/idupbo.020.2-3e

Verosimilitud (*MV*), al buscar los parámetros que con mayor probabilidad han sido generados por los datos, en presencia de outliers, va a maximizarse una función de probabilidad conjunta distorsionada y además con una distribución no-Normal; se debe recordar que la función de probabilidad conjunta es la suma, en logaritmos, de las probabilidades de realización de cada residuo; (iii) por otro lado, el Método de Momentos (*MGM*) al estimar los parámetros que cumplan con las condiciones de momentos de la muestra, en presencia de outliers, va a minimizar funciones objetivo basadas en momentos muestrales distorsionados por los outliers.

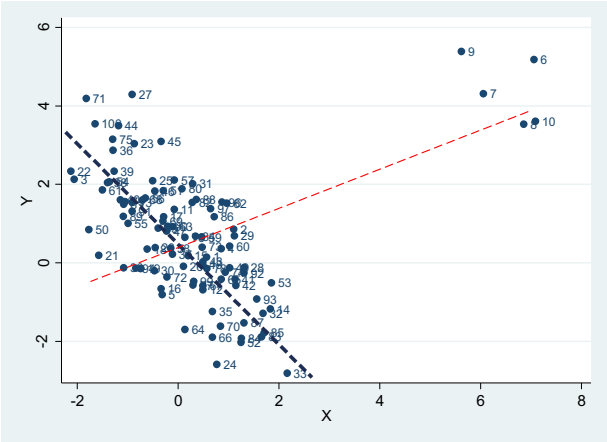


**Figura 1:** El efecto de los outliers en el análisis de regresión.
**Fuente:** Elaboración propia.

Identificar outliers en el análisis multivariado no es fácil, el análisis en más de dos dimensiones plantea desafíos que no ocurren con datos univariados. Los puntos que se considerarían valores atípicos en un espacio bivariado podrían no ser atípicos en ninguno de los dos subconjuntos univariados. Los puntos que podrían considerarse como valores atípicos en un espacio bivariado pueden no ser atípicos en el espacio multivariado (un outlier en dos dimensiones puede ser absorbido por un tercer regresor). De igual manera, los valores extremos en todas las dimensiones no necesariamente sesgan la regresión (los puntos de apalancamiento buenos o good leverage points).

Para evitar estos problemas en la estimación de las relaciones, en textos de econometría es común encontrar sugerencias para identificar el nivel de influencia de cada observación en el análisis de regresión, utilizando herramientas como *la diagonal de la matriz de predicciones (el apalancamiento o leverage), los residuos estudentizados, las distancias de Cook,* etc. Así, las observaciones podrían ser categorizadas como outliers o no de acuerdo con criterios establecidos por estas medidas. Sin embargo, existe una rama de la literatura en estadística, denominada "estadística robusta" que cuestiona la aplicabilidad de estas técnicas para identificar outliers debido a que sufren del problema denominado *masking effect*. Siguiendo a [1], y a [2], este fenómeno resalta el hecho de que la medida para identificar outliers se encuentra distorsionada por la existencia de estos outliers, y por tanto o no podrá identificar cabalmente a los mismos. Asimismo, esta literatura propone una serie de métodos denominados "robustos" que no sufren de estos problemas

Para comprender la magnitud de la deficiencia en la literatura económica en lo que al tratamiento de valores extremos se refiere, nótese que [3], [4] sugiere un test para identificar outliers multivariados basado en lo que denominaremos "técnicas no robustas". El reporte de Google Scholar sugiere que este test ha sido aplicado en 1596 trabajos empíricos hasta la fecha (según reporte de citación). Asimismo, considérese que en la literatura de trabajos empíricos en economía, resulta común la ausencia de algún tipo de diagnóstico sobre la robustez de sus resultados a valores extremos en la muestra.

El presente trabajo tiene como objeto: (i) Realizar una revisión teórica de las características de los métodos tradicionales y los métodos robustos para evaluar cuan eficientes o deficientes son en la tarea de identificación de valores atípicos en la muestra. (ii) Demostrar, mediante simulaciones que las medidas tradicionales para identificar outliers no son útiles al sufrir del denominado "masking effect", es decir son técnicas no robustas. (iii) Proponer unos códigos en Stata (do-file) para identificar y categorizar outliers eficazmente basado en métodos sugeridos en la literatura de estadística robusta.

El documento se encuentra estructurado de la siguiente manera: en la siguiente sección se describen los métodos tradicionales sugeridos, así como otros métodos más robustos y se realiza un análisis de las características de cada uno para evaluar su eficacia para la detección de outliers. La tercera sección propone unos códigos en Stata para identificar y categorizar los outliers de acuerdo con el esquema sugerido por [1]. La cuarta sección presenta evidencia basada en simulaciones de Monte Carlo sobre el desempeño de todas las técnicas mencionadas. En la quinta sección se concluye.

## 2. MÉTODOS ROBUSTOS Y NO ROBUSTOS PARA IDENTIFICAR OUTLIERS

Siguiendo a [5], los outliers, o valores extremos se clasifican en "valores extremos verticales", "valores extremos de influencia buenos" (good leverage points) y "valores extremos de influencia malos" (bad leverage points).

Como se observa en la Figura 2, si se toma en cuenta observaciones en dos dimensiones ($X$,$Y$), los valores extremos verticales son aquellos donde los valores de $Y$ están lejos de la mayor parte de los datos en la dimensión-$Y$, es decir, son valores atípicos en la variable dependiente, pero, los valores de X tienen el mismo comportamiento de las observaciones de la muestra en la dimensión-$X$. Este tipo de observaciones afectan el valor del intercepto.

Los puntos de influencia buenos (good leverage points) son observaciones cuyos valores de X están lejos de la mayor parte de los datos en la dimensión-$X$, se trata de valores atípicos en los regresores, pero que se encuentran cercanos a la línea de regresión. Estas observaciones no afectan a los estimadores, pero pueden afectar a la inferencia e inducir al fácil rechazo de la hipótesis nula de no significancia del coeficiente estimado [9].

Los puntos de influencia malos comprenden observaciones que tienen dos características: los valores de $X$ se encuentran lejos de la mayor parte de los datos en la dimensión-$X$ y las observaciones se encuentran lejos de la línea de regresión. Estos puntos de influencia malos afectan tanto al intercepto como a la pendiente.

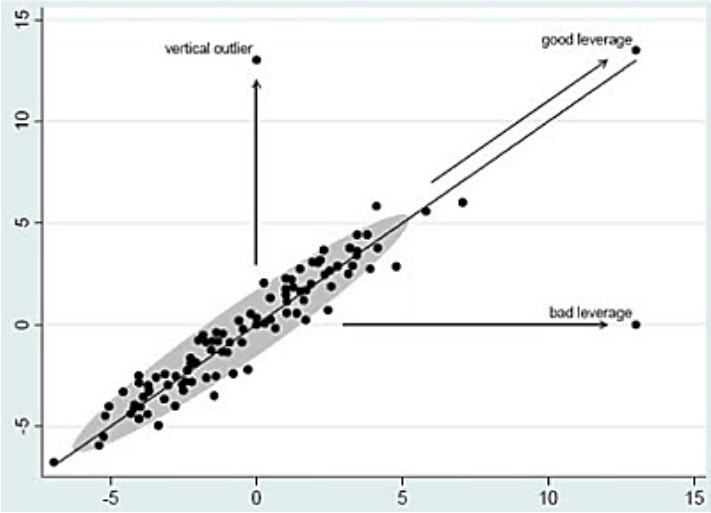


**Figura 2:** Tipos de outliers en el análisis de regresión.
**Fuente:** [5].

Identificar outliers en el análisis multivariado no es fácil, el análisis en más de dos dimensiones plantea desafíos que no ocurren con datos univariados. Los puntos que se considerarían valores atípicos en un espacio bivariado pueden no ser atípicos en ninguno de los dos subconjuntos univariados. Los puntos que podrían considerarse como valores atípicos en un espacio bivariado pueden no ser atípicos en el espacio multivariado (un outlier en dos dimensiones puede ser absorbido por un tercer regresor). Los valores extremos en todas las dimensiones no necesariamente sesgan la regresión, este tipo de outliers, como dijimos, son los puntos de apalancamiento buenos (good leverage points).

Por ello, al emprenderse un análisis multivariado, es necesario apoyarse en técnicas específicas enfocadas a identificar valores extremos de manera multivariada. En la Tabla 1 se presenta un resumen de las técnicas que se sugieren para este propósito en los principales textos de econometría.

Como se demostrará en la siguiente sección, y acorde con la literatura en estadística robusta, estas técnicas pueden caracterizarse por poseer bajo punto de quiebre (low breakdown point)[1], es decir son técnicas no resistentes a los outliers, técnicas no robustas [6].

El motivo de la "no robustez" de estas técnicas es que sufren del denominado "*masking effect*" [1]. Estos fenómenos resaltan el hecho de que la medida para identificar outliers se encuentra distorsionada por la existencia de estos outliers, y por tanto o no podrá identificar cabalmente a los mismos (masking effect) o terminará identificando como outliers observaciones que en realidad no lo son (swamping effect).

[1] En la literatura sobre estadística robusta, una medida del nivel de resistencia de los estimadores a los outliers se denomina punto de quiebre (breakdown point). Este indicador consiste en el menor nivel de contaminación que el estimador puede soportar antes de reportar resultados sesgados. Así por ejemplo, MCO tiene un punto de quiebre de 0%, es decir que una sola observación puede cambiar el estimador al punto que éste no describa de manera significativa el conjunto de datos. El mayor nivel de punto de quiebre de un estimador es del 50%.

**TABLA 1 - EL TRATAMIENTO DE OUTLIERS EN LOS LIBROS DE ECONOMETRÍA**

| Libro | Páginas | Métodos descritos para identificar outiliers |
|---|---|---|
| Perachi, F. (2011) "Econometrics". John Wiley and Sons. 1ra edición [10] | 303-305 | Matriz de predicción como medida de apalancamiento, Distancias de Cook |
| | 527-532 | Regresión robusta: estimadores M, S, MM |
| Hayashi, F. (2000) "Econometrics" Princeton University Press. 1ra edición. [11] | 21-23 | Matriz de predicción como medida de apalancamiento |
| Davidson y Mckinnon (2004) "Econometrics Theory and Methods" Oxford Unversity Press. 1ra edición. [8] | 76-81 | Matriz de predicción como medida de apalancamiento, Distancias de Cook |
| Greene, W. (2012) "Econometric Anaysis" Prentice Hall. 7ma edición [12] | 99-102 | Distancias de Cook, Residuos estudentizados |
| Hansen, B. (2017) "Econometrics" University of Wisconsin [13] | 77-79 | Matriz de predicción como medida de apalancamiento, DFITS, DFBETAs, Distancias de Cook |
| Ruud, P. (2000) "An introduction to Classical Econometrics" Oxford University Press. 1ra edición. [14] | 251-255 | Regresión no robusta: LAD |
| Verbeck, M. (2012) "A guide to modern Econometrics" John Wiley and Sons. 4ta edición. [15] | 47-49 | Residuos estudentizados, Regresión no robusta LAD, LTS |
| Baltagi, B. (2011) "Econometrics" Springer. 5ta edición. [7] | 179-187 | Residuos estudentizados, Matriz de predicción como medida de apalancamiento, DFITS, DFBETAs, Distancias de Cook. |
| Kennedy, P. (2008) "A guide to Econometrics" Blackwell Publishing. 6ta edición. [16] | 346-349 | Regresión robusta: estimador M. Regresión no robusta: LAD, LTS |
| Wooldridge, J. (2013) "Introductory Econometrics: A Modern Approach" Centage Learning. 5ta edición. [17] | 326-334 | Rediduos estudentizados, Regresión no robusta: LAD |
| Gujarati, D. y Porter, D. (2009) "Basic Econometrics" McGraw Hill. 5ta edición. [18] | 496-498 | Breve discusión sobre la importancia del análisis de outliers. |

**Fuente:** Elaboración propia.

Para comprender la magnitud de la deficiencia en la literatura económica en lo que al tratamiento de valores extremos se refiere, nótese que [3], [4] sugieren un test para identificar outliers multivariados basado en lo que denominaremos "técnicas no robustas". El reporte de Google Scholar sugiere que este test ha sido aplicado en 1596 trabajos empíricos hasta la fecha (según reporte de citación). Asimismo, considérese que en la literatura de trabajos empíricos en economía, resulta común la ausencia de algún tipo de diagnóstico sobre la robustez de sus resultados a valores extremos en la muestra.

### 2.1 Medidas de robustez

Son dos las medidas que se utilizan en estadística para caracterizar su "resistencia a los outliers"

(i) La Función de Influencia (FI)

Siguiendo a ([19]), consiste esencialmente en la primera derivada de un estadístico (un estimador) considerado como funcional de algunas distribuciones de probabilidad, que permite utilizar aproximaciones de Taylor, para analizar el comportamiento local del estimador ante ligeros cambios en la distribución de los datos (contaminación). En otras palabras, la FI proporciona una aproximación lineal del estimador en distribuciones contaminadas y así nos dice cómo una proporción infinitesimal de contaminación afecta la estimación en muestras grandes.

El interés se encuentra en conocer si la FI es acotado o suave. Cuando no tiene límites, el efecto de un valor atípico en el estimador puede ser arbitrariamente grande. Esto implica que el estimador no es robusto a valores atípicos. Cuando la IF es suave, un pequeño cambio en un punto de datos tiene solo un pequeño efecto sobre el estimador.

Dado que la FI refleja el sesgo asintótico causado por los valores atípicos en los datos, se pueden derivar varias medidas de él, ver [20] como la sensibilidad al error bruto (que mide la robustez local) o la varianza asintótica (que mide la eficiencia local o la "bondad" en el modelo ideal). Sin embargo, como se indica en [21], no todos los estimadores poseen una función de influencia.

(ii) Punto de ruptura (breakdown point, bp)

Otra medida de robustez utilizada es el punto de ruptura. Siguiendo a [22], el *bp* da la fracción más pequeña de contaminación (valores atípicos o datos agrupados en el extremo de una cola) tolerada antes de que el estadístico "se rompa" y se vuelva totalmente poco confiable. El punto de ruptura *bp* es, por tanto, una medida de solidez global (de fiabilidad).

Al ser esta medida de mayor uso, veamos en detalle su definición. Sea cualquier muestra de *n* puntos de datos,

$$Z = \{(x_{i1},...,x_{ip},y_1),...,(x_{n1},...,x_{np},y_n)\} \tag{1}$$

Y sea *T* un estimador de la regresión de manera que,

$$T(Z) = \hat{\theta} \tag{2}$$

Ahora considere todas las posibles muestras corruptas *Z** que se obtienen reemplazando cualquier m de los puntos de datos originales por valores arbitrarios (valores atípicos incorrectos). Sea el sesgo máximo que puede causar dicha contaminación:

$$sesgo(m;T;Z) = \sup_{Z^*} \left\| T(Z^*) - T(Z) \right\| \tag{3}$$

Si el sesgo (*m*; *T*; *Z*) es infinito, esto significa que m valores atípicos pueden tener un gran efecto arbitrario en T, entonces el estimador se "rompe". Por lo tanto, el punto de ruptura de la muestra finita del estimador T en la muestra Z se define como:

$$bp(T,Z) = \min\left\{ \frac{m}{n}; sesgo(m;T;Z)\ es\ finito \right\} \tag{4}$$

[23] introdujo el punto de ruptura como un concepto asintótico, y [22] dieron la correspondiente noción de muestra finita. El *bp* es a menudo la primera y más importante medida robusta que se debe analizar antes de entrar en detalles sobre las propiedades de robustez local. Para la media aritmética su BP es 0, para la mediana éste es 1/2, lo que significa que un poco menos de la mitad de los datos pueden moverse al infinito y la mediana todavía permanecerá no afectada. La desviación estándar y la desviación media tienen un BP igual a 0. Y la desviación mediana (absoluta) o "DMA", ver [19], que es la mediana de las diferencias absolutas de los datos respecto a su mediana, tiene una BP igual a 1/2.

Siguiendo a [21], el punto de ruptura asintótico máximo posible de una estimación de regresión equivariante es 1/2, y el punto de ruptura muestral finito máximo posible es [$n - p + 1$] / 2n. Para mínimos cuadrados, una observación inusual es suficiente para influir en las estimaciones de los coeficientes. Por tanto, su punto de ruptura viene dado por:

$$bp^{MCO}(T,Z) = 1/n \tag{5}$$

A medida que n aumenta, 1/*n* tiende a 0, lo que significa que el punto de ruptura de MCO es 0%.

### 2.2 Técnicas no robustas para identificar outliers

La no robustez de las siguientes técnicas para el análisis de regresión se explica en tanto las mismas se apoyan en medidas basadas en MCO, como sus valores predichos o residuos. Y como se explicó anteriormente MCO tiene un punto de ruptura de 0%.

**(i) La diagonal de la matriz de predicciones (apalancamiento o leverage)**

Se utiliza para detectar observaciones que tienen un gran impacto en los valores predichos por el modelo ($\hat{Y}$). En el marco del modelo $Y = X\beta + \varepsilon$:

$$\begin{cases} \hat{\beta} = (X'X)^{-1}X'Y \\ X\hat{\beta} = \underbrace{X(X'X)^{-1}X'}_{Hat\ Matrix}Y \end{cases} \tag{6}$$

En esta metodología se utilizan los elementos de la diagonal de la matriz sombrero. Cada valor predicho es una combinación lineal de los valores observados de $Y_i$. Si $h_{ii}$ (el i-ésimo elemento diagonal de la matriz $H$) es grande, entonces la observación $i$-ésima es influyente sobre $\hat{Y}_i$. Una expresión que demuestra los factores que determinan esta influencia está dada por: $h_{ii} = (1/n) + (x_i^2 / \sum_{i=1}^{n} x_i^2)$, donde $x_i = X_i - \bar{X}$. Nótese que $h_{ii}$ es una función sólo de los valores $X_i$, también que es una medida proporcional a la distancia entre los valores $X$ de la $i$-esima observación y su media sobre todas las $n$ observaciones. De esta manera, puede interpretarse que un valor elevado de $h_{ii}$ implicará que la observación $i$-ésima está distante del centro de las observaciones. El punto crítico a partir del cual se considerará a una observación como influyente es $h_{ii} > 2p/n$, es decir mayor a dos veces la media de $h_{ii}$ ($traza(H) = \sum_{i=1}^{n} h_{ii} = p$)). Una discusión detallada sobre el papel de matriz de predicción $H$ en la identificación de observaciones influentes (leverage points) se puede encontrar en [24].

Es importante destacar que la no robustez de este enfoque reside en el hecho que la medida $h_{ii}$ sufre del "masking effect" porque depende de $x_i$, las desviaciones respecto al promedio ($\bar{X}$), esta última, se conoce que es altamente vulnerable a los valores extremos. Así la distribución de $x_i$ puede considerarse como distorsionada en presencia de outliers y no necesariamente detectará los outliers.

**(ii) Residuos estudentizados**

En el análisis de regresión una observación con residuo diferente al resto de la muestra puede implicar valores atípicos de la misma. Los residuos estudentizados consisten en la división de cada residuo i-ésimo dividido por la desviación estándar de todos los residuos exceptuando el i-ésimo.

$$r_i^{student} = \frac{r_i}{\sigma^2(r)_{-i}} \tag{7}$$

El motivo de este cambio en el denominador se debe a que $\sigma^2(r)_i = \sigma^2(1 - h_{ii})$. Así, a mayores valores de $h_{ii}$ menor será la varianza del residuo $r_i$. Los residuos estudentizados cuantifican qué tan grandes son los residuos en unidades de desviaciones estándar. Observaciones con $r_i^{student}$ mayor a 2 en valor absoluto se considerarán valores atípicos en la dimensión-$Y$.

Este enfoque es no robusto debido a que los residuos estimados en el numerador, $r = Y - X\hat{\beta}$, son estimaciones basadas en coeficientes β que yá se encuentran distorsionados por la presencia de outliers.

**(iii) Distancias de Cook**

Es una medida que combina la información de los anteriores dos criterios, apalancamiento y residuos. El concepto de influencia se sustenta en el efecto que conlleva la eliminación de la observación bajo consideración sobre las conclusiones del análisis. Existen diferentes maneras equivalentes de reflejar esta idea:

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{-i})^T (X^T X)(\hat{\beta} - \hat{\beta}_{-i})}{p\sigma_{-i}^2} \tag{8}$$

Esta es la expresión inicialmente planteada por [25], de ésta se desprende que la distancia de Cook mide el cambio agregado en los coeficientes estimados cuando cada observación es alejada de la estimación. Valores altos de $D_i$ indicarán que los puntos asociados tienen una gran influencia en la estimación de $\beta$.[2] Equivalentemente, estas distancias pueden expresarse como:

[2] A su vez, considerando que $\hat{Y} = X\hat{\beta}$ :

$$D_i = \frac{(\hat{Y}_i - \hat{Y}_{-i})^T (\hat{Y}_i - \hat{Y}_{-i})}{p\sigma_{-i}^2}$$

Entonces la distancia de Cook para la i-ésima observación se basa en las diferencias entre las respuestas pronosticadas del modelo construido a partir de todos los datos y las respuestas pronosticadas del modelo construido al omitir la i-ésima observación.

$$D_i = \frac{1}{p}\left(r_i^{student}\right)^2\left(\frac{h_{ii}}{1-h_{ii}}\right) \tag{9}$$

De esta manera, $D_i$ será mayor si los residuos estudentizados son grandes, es decir si existen valores atípicos en la dimensión-*Y*, o si el nivel de apalancamiento o influencia de las observaciones son grandes, recuérdese que $h_{ii}$ depende únicamente de los valores en la dimensión-*X*. Valores de $D_i$ mayores a *p/n* se consideran como influyentes.

Al ser esta medida una combinación de los dos indicadores anteriores hereda la no-robustez de estas por las razones anteriormente mencionadas.

**2.2 Técnicas robustas para identificar outliers**

En virtud de que los métodos tradicionales no pueden detectar los outliers porque son afectados por las observaciones que ellos deben justamente identificar como outliers, la literatura en estadística propone otro tipo de estimadores para detectar valores atípicos multivariados en la muestra que no sufren del masking effect.

Identificar outliers multivariados no es fácil. Los outliers univariados, o valores extremos en una variable, son fácilmente identificables a través de un simple histograma. Los outliers bivariados también pueden ser identificados visualmente. Sin embargo, para el caso de más de dos variables, puede suceder que una observación no sea outlier en dos dimensiones pero si en tres o cuatro, lo que ya no es identificable visualmente. Un estimador común para detectar valores extremos multivariados es la Distancia de Mahalanobis (DM), $DM_i=\sqrt{(X_i-\mu)\Sigma^{-1}(X_i-\mu)}$, donde μ es el llamado vector de ubicación o centralidad (location vector) que no es más que el vector de medias de las variables, $\Sigma^{-1}$ es la matriz de covarianzas y $X_i$ es la fila i de la matriz de observaciones X. Las DM miden la distancia de las observaciones respecto al centro de los datos (μ) considerando la forma de los mismos (Σ); así las observaciones con valores de DM extremos pueden ser considerados outliers multivariados (nótese que las DM tienen una distribución $\chi^2_{\#variables:\ p}$). El problema con esta medida, denominado "masking effect", es que μ y Σ a su vez pueden ser distorsionados por los outliers, haciendo de MD una medida no representativa de la mayoría de los datos. Por ello, en la literatura sobre Estadística Robusta muchos estimadores robustos de μ y Σ han sido propuestos. Los tres más importantes son el Determinante de Covarianza Mínima (Minimum Covariance Determinant, MCD), el estimador S de ubicación y dispersión multivariada (S-estimator of location and scatter), y el enfoque de Stahel-Donoho (SD). En general estos tres estimadores tienen la propiedad de equivarianza afín, es decir, que se comportan adecuadamente ante transformaciones afines en los datos; también tienen un punto de quiebre del 50%, es decir que son altamente resilientes a observaciones extremas. Para mayores detalles sobre sus algoritmos véase [2].

**(i) El Determinante de Covarianza Mínima (MCD)**

Considérese los datos multivariados $X_n = \{x_1, \ldots, x_n\}$ con *n* observaciones $x_i = (x_{i1},\ldots,x_{ip})^T$ donde *i*=1,...,*n* en *p* dimensiones. El método *MCD* busca las h observaciones (de la muestra n) cuya matriz de varianza covarianza (Σ) tenga el menor determinante posible (det(Σ)). Nótese que el determinante de (Σ) es una medida unidimensional de la variabilidad multivariada en una muestra. Esto se puede observar más fácilmente para el caso de dos variables:

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{yx}^2 & \sigma_y^2 \end{bmatrix};\ \det(\Sigma) = \sigma_x^2\sigma_y^2 - \sigma_{xy}^2 \tag{10}$$

que es la diferencia entre la dispersión univariada conjunta de las variables menos la dispersión debida a las covariaciones entre ellas. Así, el método *MCD* busca el grupo de h observaciones con la menor varianza generalizada multivariada.

La implementación del estimador *MCD* es un tanto problemática ya que su algoritmo obtiene resultados inestables en múltiples replicaciones y además los mismos son sensibles al valor inicial elegido en la minimización. En Stata este estimador puede ser implemento utilizando el comando mcd, en R usando las funciones fastmcd o covMcd, en *SAS* utilizando la función *MCD*.

**(ii) El estimador S-multivariado de ubicación y dispersión**

Para explicar este estimador, es necesario comenzar recordando que un parámetro de ubicación es una medida de la centralidad de una distribución: es el punto alrededor del cual la dispersión de las observaciones es la más baja. En el caso univariado, éste es la media aritmética, y puede obtenerse resolviendo el problema siguiente:

$$\hat{\mu} = \arg\min_{\mu} \sigma^2 \tag{11}$$

donde,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{12}$$

$$1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 \tag{13}$$

Generalizando este enfoque al caso multivariado, uno puede aplicar la misma lógica utilizando la distancia de Mahalanobis como la medida generalizada de variabilidad, así:

$$(\hat{\mu}, \hat{\Sigma}) = \arg\min_{\{\mu,\ \Sigma\}} \det(\Sigma) \tag{14}$$

donde,

$$p = \frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{(X_i - \mu)\Sigma^{-1}(X_i - \mu)}\right)^2 \tag{15}$$

En esta expresión, $p$ es igual a los grados de libertad de la distribución $\chi_p^2$ de las distancias de Mahalanobis, asumiendo que éstas son generadas por una distribución normal multivariada. Como se mencionó anteriormente, la distancia de Mahalanobis muestra qué tan lejos está la observación $X_i$ respecto al centro de los datos y sufre del problema de "masking effect". Asimismo, el hecho de que las distancias de Mahalanobis sean elevadas al cuadrado nos replantea el problema original: aquellas observaciones que están alejadas del centro de la distribución, los outliers, ejercen una influencia por demás importante en la estimación de los parámetros.

Al respecto, el estimador S-multivariado propone remplazar la función cuadrática por una función $\rho(\cdot)$ que sea: no decreciente en valores positivos de su argumento y que se incremente menos drásticamente que la función cuadrática. Existen múltiples candidatos para la función $\rho(\cdot)$, la más utilizada es la función Tukey Biweight, la misma se muestra en la Figura 3.

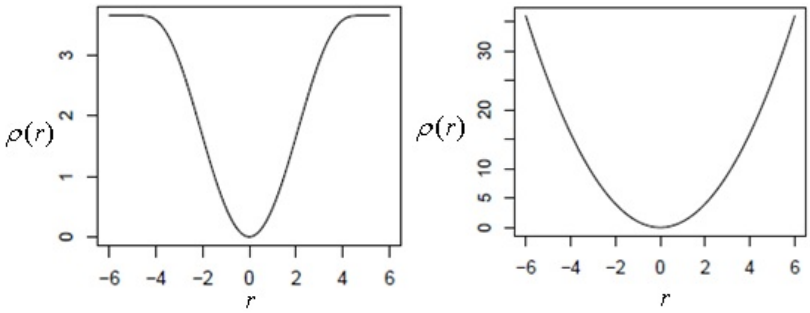


**Figura 3:** La función Tukey Biweight y la función cuadrática.
**Fuente:** Elaboración propia.

Entonces, el problema puede ser reescrito de la siguiente manera:

$$(\hat{\mu}_S, \hat{\Sigma}_S) = \arg\min_{\{\mu,\ \Sigma\}} \det(\Sigma) \tag{16}$$

donde,

$$b = \frac{1}{n}\sum_{i=1}^{n}\rho\left(\sqrt{(X_i - \mu)\Sigma^{-1}(X_i - \mu)}\right) \tag{17}$$

$b = E[\rho(u)]$ y $u$ se supone que posee una distribución normal estándar; esto garantiza la consistencia gaussiana del estimador. El estimador S-mutivariado $(\hat{\mu}_S, \hat{\Sigma}_S)$ consiste en estimar los parámetros $\mu$ y $\Sigma$ de manera simultánea en este sistema.

Como destacan [26]el estimador S-multivariado muestra una mayor estabilidad que el estimador Minimum Covariance Determinant (MCD) bajo múltiples replicaciones, y su uso debe ser privilegiado al diagnosticar la presencia de outliers. El comando en Stata para implementar el estimador S-multivariado es smultiv, en R la función es CovSest.

**(iii) El estimador Stahel-Donoho (SD)**

Este método identifica outliers analizando todas las proyecciones univariadas de un conjunto de datos. Es decir, dada la dirección $a \in R^p$ con $\|a\| = 1$, dada por $a'X=\{a'X_1, \ldots, a'X_n\}$ que representa la proyección de los datos $X$ en la dirección $a$. Una observación es outlier si es identificado con valores extremos de estas proyecciones en diferentes direcciones $a$: $\Theta_i = \max_{\|a\|=1} \frac{|x'a - \hat{\mu}(a'X)|}{\hat{\sigma}(a'X)}$. Entonces, a partir de esta medida se definen los pesos $w_i = f(\Theta_i)$ de manera que las observaciones con mayores distancias en las direcciones $a$ reciben menores pesos. En este marco, las medidas de centralidad y dispersión "robustas" propuestas por [27] y [28], son medidas ponderadas por la función decreciente de $\Theta_i$ :

$$\mu_{SD} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} \qquad \Sigma_{SD} = \frac{\sum_{i=1}^{n} w_i (X_i - \mu_{SD})(X_i - \mu_{SD})^T}{\sum_{i=1}^{n} w_i} \tag{18}$$

En otras palabra, el estimador de [27] y [28] de centralidad y dispersión multivariada (*SD*) consiste en calcular las distancias de cada punto proyectando los datos unidimensionalmente en todas las posibles direcciones y estimando las distancias de cada observación al centro "robusto" de cada proyección. Las distancias se definen con la distancia de Mahalanobis, $DM(X_i) \equiv \left[(X_i - \mu_{SD})\Sigma_{SD}^{-1}(X_i - \mu_{SD})\right]^{1/2}$, que como se mencionó anteriormente está distribuida como $\sqrt{\chi_p^2}$ . Así, se puede considerar como outlier a aquella observación con distancia *DM* superior a su percentil 95. En Stata el commando *robmv* puede ser usado para aplicar este estimador, en *R* las funciones son *CovSde* y *outlyingness*.

Las características de este enfoque, sus fundamentos en el análisis de proyecciones unidiminesiones de múltiples variables, permiten aplicar idóneamente la misma para identificar outliers en modelos con múltiples variables categóricas, como en [29] y en muestras con distribuciones no gausianas (asimétricas) como en [30].

**2.3 Regresión robusta**

En la anterior sección se resaltó la importancia de considerar el efecto de las observaciones atípicas (outliers) en el análisis econométrico multivariado, los tipos de outliers y la manera correcta de detectarlos. Así, una vez identificados los outliers se podría excluirlos de la muestra o darles una menor ponderación en la regresión. Un segundo enfoque para lidiar con el problema de los outliers es utilizar directamente métodos econométricos que no sean vulnerables a las observaciones atípicas, lo que se conoce como Regresión Robusta.

(i) El estimador *M*

Este estimador es una modificación de la función objetivo de Mínimos Cuadrados Ordinarios (MCO). Considerando que la vulnerabilidad de MCO proviene del mayor peso que se otorga a los valores extremos por elevar al cuadrado los residuos a ser minimizados, este estimador propone minimizar en su lugar otra función "$\rho$" que asigne menor peso a los residuos extremos:

$$\hat{\beta}^M = \arg\min_{\beta} \sum_{i=1}^{n} \rho(r_i(\beta)) \tag{19}$$

En este marco *MCO* puede ser entendido como un estimador *M* con $\rho=()^2$. En los métodos M las funciones ρ tienen que cumplir ciertas propiedades (no decreciente, simétrica, tener un único mínimo en cero, y ser menos creciente que la función cuadrática), las mismas pueden ser monótonas (si son enteramente convexas) o redescendientes (si tienen un límite establecido a partir de un valor k, denominado punto de quiebre), éstas últimas son las que dan robustez a la estimación, la función redescendiente mayormente utilizada es la Tukey Biweight (TB).

La estimación del modelo *M* es un problema de Mínimos Cuadrados Ponderados Iterados con los pesos definidos como $w_i = \rho(r_i/\sigma)/r_i^2$. Como se aprecia los residuos son estandarizados por una medida de dispersión σ para garantizar la

propiedad de equivarianza de escala, es decir, la independencia con respecto a las unidades de medida de las variables. Entonces, en la práctica se estima: $\hat{\beta}^M = \arg\min_{\beta} \sum_{i=1}^{n} \rho(r_i(\beta))$. La necesidad de iterar surge por el hecho que los pesos $w_i$ dependen de los residuos, los residuos dependen de los coeficientes estimados, y los coeficientes estimados dependen de los pesos; así se necesita un punto de comienzo, en [31] se encuentra un resumen del algoritmo de este proceso.

En Stata el estimador M con la función Tukey Biweight puede ser implementado con el comando rreg o mregress; en R se puede usar la función rlm en el paquete robustbase . Sin embargo, debido al enfoque iterativo en su estimación, este método no tiene las propiedades de robustez deseadas, ya que sólo es resistente a los outliers verticales.

**(ii) El estimador S**

Un estimador más robusto puede ser obtenido enfocando el análisis desde otra perspectiva también interesante. Así, inicialmente es necesario recordar que MCO no es más que la minimización de n veces la varianza de los residuos, ya que: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2(\beta))$. Expresión que puede ser re-escrita como: $1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i(\beta))}{\hat{\sigma}}\right)^2$.

En este marco, con el fin de aumentar la robustez, en el espíritu del estimador *M*, la función cuadrática puede ser sustituida por otra función que conceda menos importancia a los residuos grandes:

$$1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i(\theta))}{\hat{\sigma}^S}\right) \tag{20}$$

De esta manera, el estimador S minimizará la varianza $\hat{\sigma}^S$ que satisfaga la siguiente expresión:

$$\hat{\beta}^S = \arg\min_{\beta} \sum_{i=1}^{n} \rho(r_i(\beta)) \tag{21}$$

condicional a que,

$$\Psi = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i(\beta))}{\hat{\sigma}}\right) \tag{22}$$

donde $\Psi = E[\rho(Z)]$ y $Z \sim N(0,1)$, es una corrección que restringe la condición a minimizar para garantizar Normalidad y ρ es la función Tukey Biweight. El algoritmo para su estimación, denominado fast-S algorithm corresponde a [32]. El estimador S es resistente a los outliers en las dimensiones *Y* y *X*, sin embargo existe un trade-off entre su grado de robustez y eficiencia. En Stata el comando para su implementación es sregress. En *R* la función es lmrob en el paquete robustbase.

**(iii) El estimador *MM***

Este es un estimador robusto y a su vez eficiente. Se lo puede describir como un estimador *M* con varianza S. Es decir, el estimador MM resulta de la combinación de los dos métodos anteriores,

$$\hat{\beta}^{MM} = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{r_i(\beta)}{\hat{\sigma}^S}\right) \tag{23}$$

en una primera parte se implementa el estimador robusto pero de baja eficiencia *S* y de la misma se calcula la varianza $\hat{\sigma}^S$, la misma que es utilizada en una segunda parte en la estimación de un modelo redescendiente M vía Mínimos Cuadrados Ponderados Iterados pero con un valor robusto como punto de inicio en las iteraciones $\hat{\beta}_0^S$, con lo que se adquiere mayor estabilidad y eficiencia; en ambas etapas la función ρ es la Tukey Biweight. El algoritmo para este método puede encontrarse en [2]. Para implementar el método en Stata el comando correspondiente es mmregress. En R la función es lmrob en el paquete robustbase.

**(iv) El estimador *MS***

Los anteriores estimadores, sin embargo, tienen problemas en sus algoritmos en presencia de variables explicativas dicotómicas. Para subsanar ello, [2]proponen dividir las variables explicativas en dos grupos, las continuas y las

dicotómicas, $Y_i = \alpha + \beta X_i + \delta D_i + u_i$ e implementar alternando un modelo *M* para obtener $\delta$, asumiendo que se conoce $\beta$, y un modelos S para estimar $\beta$, asumiendo que se conoce $\delta$, hasta alcanzar la convergencia. La idea es aprovechar las propiedades de ambos métodos, el modelo *M* es resistente a los outliers verticales, como los creados por las variables dicotómicas, y el modelo *S* es resistente a los outliers en *Y* y *X* pero vulnerable en su algoritmo a la presencia de variables categóricas. El algortimo para la estimación del modelo *MS* se encuentra en [2]. En Stata el comando para aplicar este estimador es msregress, en R la función lmrob.

## 3. UN PROGRAMA EN STATA PARA IDENTIFICAR Y CATEGORIZAR OUTLIERS MULTIVARIADOS

Uno de los factores que pueden explicar la omisión del análisis sobre el efecto de valores extremos en la investigación empírica, puede ser la ausencia de instrumentos disponibles en los softwares convencionales que implementen estas técnicas robustas. En esta sección se presenta un programa en Stata del tipo do-file para identificar y categorizar outliers multivariados para datos de corte transversal, implementando la herramienta sugerida por [1], la misma que no sufre del mencionado "masking effect".

### 3.3.1 La técnica de Rousseauw y vanZomeren (1990) y su código en Stata

El mismo consiste en graficar en el eje de las ordenadas los residuos estandarizados y en el eje de las abscisas las distancias de Mahalanobis. Para mayores detalles y evidencia sobre su "robustez", véase [1], [2].

Los residuos estandarizados son la medida utilizada para identificar valores extremos en la dimensión-*y*, éstas consisten en el cociente de los residuos sobre su desviación estándar, $r_i / \sigma(r)_i$. Para que ésta sea resistente a outliers (robusto), en el numerador se utilizan los residuos de la regresión *S*, y en el denominador, como medida de dispersión se utiliza la desviación absoluta mediana normalizada: $MAD(r) = Med|r_i - Med(r)|/0.6745$. Valores de estos residuos mayores a 2.25 en valor absoluto requieren atención ya que pueden ser o "good leverage points" (si a su vez estas observaciones son valores extremos en *x*), o "vertical outliers" (si no son outliers en *x*).

Para medir valores extremos en la dimensión-*X*, para cada observación se calcula la distancia de Mahalanobis, $MD_i = \sqrt{(x_i - \mu)\Sigma^{-1}(x_i - \mu)^T}$ , donde $\mu$ representa el vector de medias y Σ la matriz de covarianzas; las *MDi* pueden entenderse como la distancia estandarizada de cada observación al centro de los datos. Al tener una distribución chi-2, observaciones con valores mayor a $MD_i > \sqrt{\chi^2_{\#\text{var}iables,0.975}}$ pueden definirse como valores extremos en la dimensión-*X*, y éstas pueden ser o "good leverage points" (si a su vez estas observaciones son valores extremos en *y*), o "bad leverage points" (si no son outliers en *X*). Asimismo, para que las $MD_i$ sean robustas, $\mu$ y Σ se calculan a través del estimador *S*-multivariado; este último detalle define la contribución del presente programa.

A continuación se presenta el programa en Stata para obtener este gráfico.

```
*****************************************
program define outid
syntax varlist , [dummies(varlist)]
local dv: word 1 of `varlist'
local expl: list varlist - dv
local ndum: word count `dummies'
local nvar: word count `varlist'
local p=`ndum'+`nvar'
local b=sqrt(invchi2(`p'),0.975)
capture drop outS rdS id
capture qui smultiv `expl', gen(outS rdS) dummies(`dummies')
label var rdS “Robust_distance_S”
gen id=_n
if `ndum'==0 {
capture drop S_outlier S_stdres
capture qui mmregress `dv' `expl', outlier
capture drop stdres
rename S_stdres stdres
}
else {
capture drop MS_outlier MS_stdres
capture qui mmregress `dv' `expl', outlier dummies(`dummies')
capture drop stdres
```

```
rename MS_stdres stdres
}
label var stdres “Robust standardized residuals”
twoway (scatter stdres rdS if abs(stdres)<4&rdS<sqrt(2)*`b') (scatter stdres rdS if abs(stdres)>=4|rdS>=2*`b',
mlabel(id) msymbol(circle_hollow)), xline(`b') yline(2.25) yline(-2.25) legend(off)
capture drop vo glp blp
gen vo=(abs(stdres)>2 & rdS<`b')
gen glp=(abs(stdres)<2 & rdS>`b')
gen blp=(abs(stdres)>2 & rdS>`b')
label var vo “Vertical outliers”
label var glp “Good leverage points”
label var blp “Bad leverage points”
edit id vo glp blp stdres rdS `varlist' `dummies' if vo==1 | glp==1 | blp==1
end
**************************
```

El programa hace uso de los comandos smultiv, sregress, msregress de [9] es muy importante que el lector instale los mismos en su computadora. Puede hacerlo utilizando los comandos findit o ssc install.

Para implementar estos códigos en Stata, el usuario debe copiar los mismos en un do-file, correr el programa y luego aplicar el mismo a su modelo de regresión de acuerdo a la siguiente sintaxis:

**outid** Variable_dependiente Variables_explicativas _continuas, dummies(Variables_explicativas_categóricas)

Un ejemplo de aplicación se presenta a continuación. En un do-file se puede escribir:

```
clear
set obs 300
set seed 1010
drawnorm x1-x5 e
gen i= _n
gen d1=(x4>0.7)
gen d2=(x5<-0.9)
gen y=x1+x2+x3+d1+d2+e
replace x1=invnorm(uniform())+10 in 1/20
replace y=invnorm(uniform())+10 in 15/30
scatter y x1, mlabel(i)
```

En la Figura 4, las observaciones 1 al 14 son outliers del tipo bad leverage points, las observaciones 15 a 20 son good leverage points, mientras que las observaciones 20 a 30 son vertical outliers. En este escenario, para identificar y clasificar los outliers se puede implementar el programa descrito mediante la siguiente sintaxis en Stata:

**outid** y x1 x2 x3, dummies(d1 d2)

Los resultados que se obtienen al implementarse el programa outid son los siguientes: la Figura 5, propuesto por [1], y la representación en la base de datos de las observaciones identificadas como outliers.

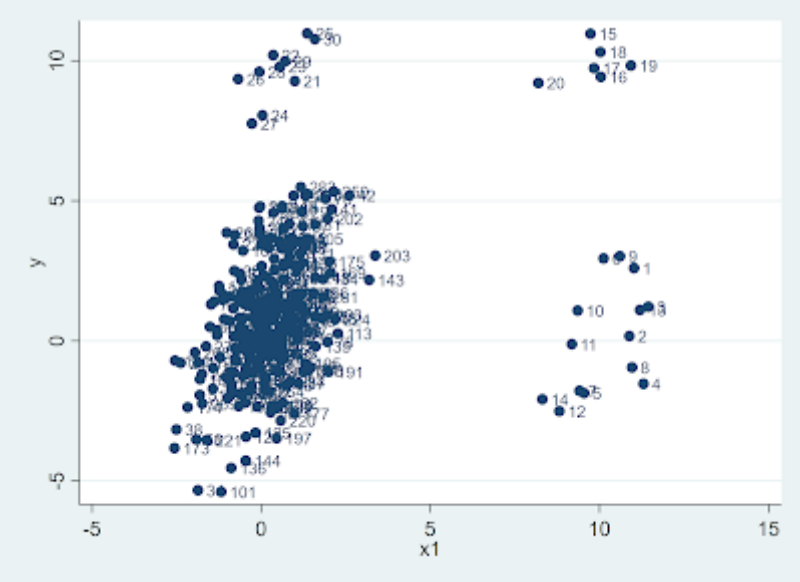


**Figura 4:** Identificación y clasificación de outliers.
**Fuente:** Elaboración propia.

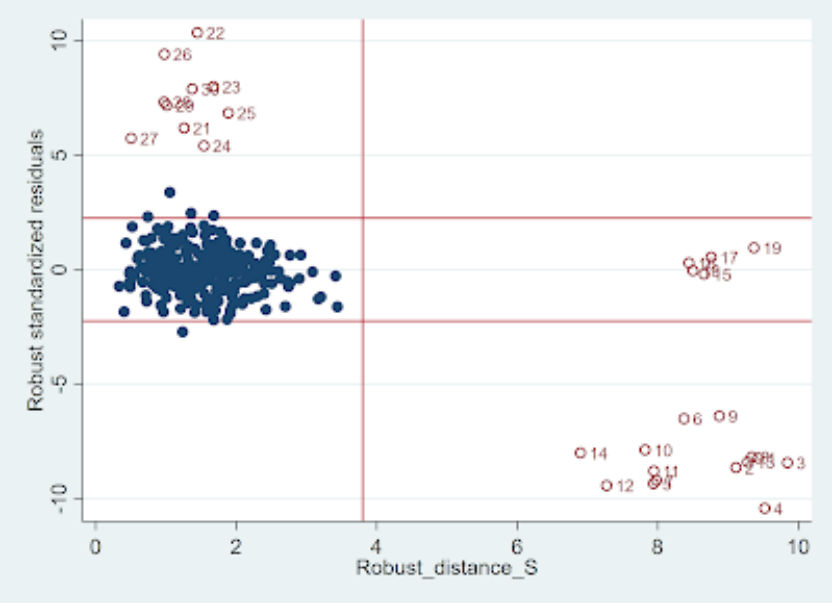


**Figura 5:** Identificación de outliers.
**Fuente:** Elaboración propia.

El diagrama de la Figura 6 es útil para fines de interpretación.


**Figura 6: Tipos de outliers en la regresión**
**Fuente:** Elaboración propia

Una vez identificados los outliers se debería analizar cada caso, es decir, si corresponden a la muestra, de que tipo son, o si tal vez se trata de información errónea del proceso de levantamiento de los datos. Asumiendo que todos corresponden a la muestra pero no representan a la misma, una estrategia racional es correr las regresiones con y sin

outliers, y comparar los coeficientes para ver el nivel de importancia de estas observaciones sobre las estimaciones (se podría usar un test de Haussman para este propósito). Por ejemplo ver Tabla 2.

Se debe notar que la primera regresión es con la muestra contaminada por los outliers, mientras que la segunda es con la muestra sin los outliers identificados con el programa **outid**.

Como se expresó anteriormente, la contribución de este programa radica en el uso del estimador S-multivariado para calcular el centro y dispersión de los datos de manera robusta. En Stata, los comandos de [5] permiten obtener la representación de [1], utilizando la opción graph, pero utilizando el estimador Determinante de Covarianza Mínima (Minimum Covariance Determinant, MCD) para identificar los outliers en ladimensión X. Sin embargo, como demuestran [26], el MCD es inestable, presenta baja eficiencia, es computacionalmente pesado, y es superado en propiedades por el estimador S-multivariado, el mismo que se implementa en **outid**.

## 4. ROBUSTEZ DE LOS ESTIMADORES PARA IDENTIFICAR OUTLIERS MULTIVARIADOS

En esta sección, mediante simulaciones de Monte Carlo se evalúa la robustez de los estimadores anteriormente presentados para identificar outliers multivariados, incluyendo el desempeño del programa **outid** planteado en este capítulo.

Para ello, inicialmente se crea un conjunto de datos de tamaño $n$ generando aleatoriamente 5 variables explicativas, continuas independientes con distribución normal, media cero y varianza unitaria. Este conjunto de datos se llama la muestra limpia. Posteriormente, se reemplaza aleatoriamente $k\%$ de las observaciones de la primera variable con valores aleatorios extraídos de una distribución gaussiana con una media de 5 y una desviación estándar de 0.1. Este conjunto de datos se llama la muestra contaminada. Dos variables se manejan en este setup para crear diferentes escenarios, el tamaño de muestra que tomará valores $n$=100 (muestra pequeña) y $n$=1000 (muestra grande); y el nivel de contaminación, donde $k$ tomará valores para representar muestras leve e intensamente contaminadas por outliers: $k = 3\%$ y $30\%$.

**TABLA 2 - REGRESIONES CON Y SIN OUTLIERS**

```
. reg y x1 x2 x3 d1 d2
```

| Source | SS | df | MS | | |
|---|---|---|---|---|---|
| | | | | Number of obs | = 300 |
| | | | | F(5, 294) | = 35.74 |
| Model | 956.931893 | 5 | 191.386379 | Prob > F | = 0.0000 |
| Residual | 1574.21844 | 294 | 5.3544845 | R-squared | = 0.3781 |
| | | | | Adj R-squared | = 0.3675 |
| Total | 2531.15034 | 299 | 8.46538574 | Root MSE | = 2.314 |

| y | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| x1 | .3636589 | .0503998 | 7.22 | 0.000 | .2644689 | .462849 |
| x2 | .8642697 | .1360916 | 6.35 | 0.000 | .5964325 | 1.132107 |
| x3 | .9463285 | .1277357 | 7.41 | 0.000 | .6949363 | 1.197721 |
| d1 | 1.265538 | .2988224 | 4.24 | 0.000 | .6774363 | 1.85364 |
| d2 | 1.40641 | .3436175 | 4.09 | 0.000 | .7301481 | 2.082672 |
| _cons | .1018344 | .176152 | 0.58 | 0.564 | -.2448442 | .4485131 |

```
. reg y x1 x2 x3 d1 d2 if vo==0 & glp==0 & blp==0
```

| Source | SS | df | MS | | |
|---|---|---|---|---|---|
| | | | | Number of obs | = 263 |
| | | | | F(5, 257) | = 186.90 |
| Model | 865.231558 | 5 | 173.046312 | Prob > F | = 0.0000 |
| Residual | 237.952298 | 257 | .925884428 | R-squared | = 0.7843 |
| | | | | Adj R-squared | = 0.7801 |
| Total | 1103.18386 | 262 | 4.2106254 | Root MSE | = .96223 |

| y | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| x1 | .9225287 | .0586668 | 15.72 | 0.000 | .8069998 | 1.038057 |
| x2 | .9761833 | .0593947 | 16.44 | 0.000 | .859221 | 1.093146 |
| x3 | .9477619 | .0565375 | 16.76 | 0.000 | .8364262 | 1.059098 |
| d1 | .9276139 | .1320016 | 7.03 | 0.000 | .6676715 | 1.187556 |
| d2 | .9931294 | .1511577 | 6.57 | 0.000 | .6954641 | 1.290795 |
| _cons | -.0601213 | .0764363 | -0.79 | 0.432 | -.2106425 | .0903999 |

La Figura 7 exhibe la relación entre $Y$ y $X_1$ para una muestra aleatoria de la simulación, una realización para cada uno de los cuatro casos considerados.

Los resultados de 1500 simulaciones para los diferentes escenarios se presentan en la siguiente tabla. La misma mide el denominado "*masking effect*" como el porcentaje de observaciones que sí son outliers pero que no fueron identificados como tal por el método bajo consideración. La columna uno describe el escenario considerado, el tipo de outliers se describe en la columna dos, mientras que el resto de las columnas muestra los 4 métodos considerados, la última con el programa **outid** sugerido en este documento.

En general, la conclusión que emana de la Tabla 3 es que las medidas tradicionales sugeridas en los libros de econometría para identificar outliers multivariados no son eficaces para tales fines. En los diferentes escenarios considerados, la diagonal de la matriz de predicciones en promedio no identificó el 78% de los outliers, las distancias de Cook el 77% y los residuos estudentizados no identificaron el 42% de los outliers generados; resultados que dan cuenta del fenómeno conocido como "masking effect". A su vez, si se consideran los estimadores obtenidos después de implementar el programa **outid**, se tiene que, en los diferentes escenarios considerados, el porcentaje promedio de outliers no identificados fueron de 0.05% y 1% con ambos enfoques, respectivamente, es decir que el mismo sí pueden identificar adecuadamente los valores extremos multivariados.

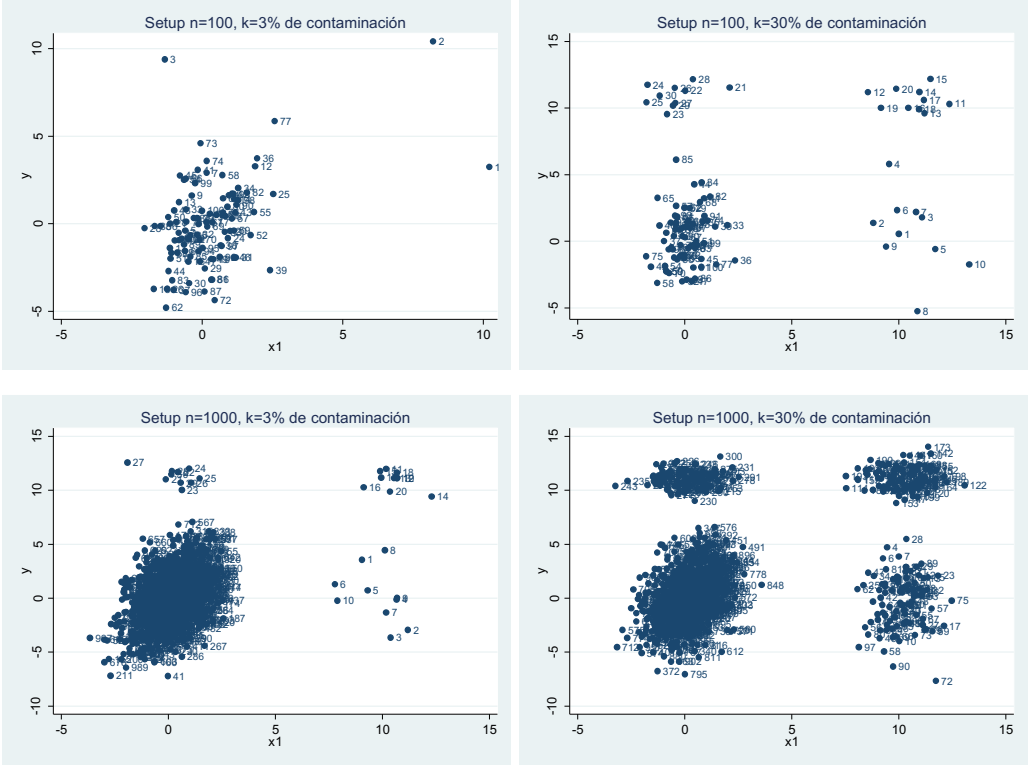


**Figura 7:** Una realización de los esquemas de la simulación.
**Fuente:** Elaboración propia.

**TABLA 3 - RESULTADOS DE LAS SIMULACIONES**
(Porcentaje de observaciones que si son outliers pero no fueron identificados)

| MASKING EFFECT | | | | | |
|---|---|---|---|---|---|
| Escenario | Tipo de outliers | Distancia matriz predicciones (%) | Distancias de Cook (%) | Residuos estudentizados (%) | Regresión sin outliers aplicando el programa "outlid" (%) |
| n=100 k=3% | BLP | 76.1 | 44.3 | | 0.0 |
| | GLP | 75.0 | 86.2 | | 0.2 |
| | VO | | | 34.2 | 0.3 |
| n=100 k=30% | BLP | 80.8 | 96.6 | | 0.0 |
| | GLP | 90.5 | 98.2 | | 0.1 |
| | VO | | | 58.6 | 0.2 |
| n=1000 k=3% | BLP | 45.0 | 17.1 | | 0.0 |
| | GLP | 75.0 | 76.0 | | 0.0 |
| | VO | | | 25.2 | 0.1 |
| n=1000 k=30% | BLP | 85.4 | 99.4 | | 0.0 |
| | GLP | 97.1 | 99.7 | | 0.1 |
| | VO | | | 48.5 | 0.1 |

**Fuente:** Elaboración propia.

## 5. CONCLUSIONES

Las observaciones atípicas distorsionan la muestra y por lo tanto los estimadores que resultan no reflejan la verdadera relación entre las variables de toda la muestra. La motivación del presente trabajo es crear conciencia en la comunidad académica sobre la manera adecuada de tratar este problema en el análisis estadístico.
A través de simulaciones se logra demostrar que las medidas estándares utilizadas como la diagonal de la matriz de predicciones, los residuos estudentizados, y las distancias de Cook, no logran identificar en un porcentaje elevado estos valores atípicos, porque sufren del denominado "masking effect".

En este trabajo se plantea un programa del tipo do-file que permite identificar outliers de manera robusta, en la medida que se basa en el estimador S-multivariado de ubicación y dispersión. Simulaciones de Monte Carlo muestran la aplicabilidad del programa sugerido para identificar adecuadamente los valores extremos en un escenario multivariado.

## 6. REFERENCIAS


[1] B. C. Rousseeuw, Peter J and Van Zomeren, "Points, Unmasking multivariate outliers and leverage," *J. Am. Stat. Assoc.*, vol. 85, pp. 633–639, 1990.

[2] V. J. Maronna, Ricardo A and Yohai, "Robust regression with both continuous and categorical predictors," *J. Stat. Plan. Inference*, vol. 89, pp. 197–214, 2000.

[3] A. S. Hadi, "Identifying multiple outliers in multivariate data," *J. R. Stat. Soc. Ser. B (Methodological*, vol. 54, pp. 761–774, 1992.

[4] A. S. Hadi, "A modification of a method for the detection of outliers in multivariate samples," J. R. Stat. Soc. Ser. B, vol. 56, pp. 393–396, 1994.

[5] C. Verardi, Vincenzo and Croux, "Robust regression in Stata," *Stata J.*, vol. 9, no. SAGE Publications Sage CA: Los Angeles, CA, pp. 439–453, 2009.

[6] V. J. and others Maronna, Ricardo A and Martin, R Douglas and Yohai, *Robust statistics: theory and methods (with R)*. John Wiley \& Sons, 2019.

[7] Baltagi, *Econometrics*. 2011.

[8] R. Davidson and J. G. MacKinnon, Instructor ' s Manual to Accompany Econometric Theory and Methods. 2004.

[9] C. Dehon, M. Gassner, and V. Verardi, "Beware of 'Good' outliers and overoptimistic conclusions," *Oxf. Bull. Econ. Stat.*, vol. 71, no. 3, pp. 437–452, 2009, doi: 10.1111/j.1468-0084.2009.00543.x.

[10] F. Perachi, *Econometrics*. 2011.

[11] F. Hayashi, Econometrics,‖ Princeton University Press: Princeton. 2000.

[12] W. H. Greene, "Econometric analysis, 71e," *Stern Sch. Business, New York Univ.*, 2012.

[13] B. Hansen, *Econometrics. US*. University of Wisconsin Press, 2017.

[14] P. A. Ruud, *An introduction to classical econometric theory*. Oxford University Press, 2000.

[15] M. Verbeek, *A guide to modern econometrics*. John Wiley \& Sons, 2008.

[16] P. Kennedy, A guide to modern econometrics. 2008.

[17] J. M. Wooldridge, Introductory econometrics: A modern approach. 2016.

[18] D. N. Gujarati, *Basic econometrics*. Tata McGraw-Hill Education, 2009.

[19] Hampel and F. R., "The influence curve and its role in robust estimation," *J. Am. Stat. Assoc.*, vol. 69, pp. 383–393, 1974.

[20] W. A. Hampel, Frank R and Ronchetti, Elvezio M and Rousseeuw, Peter J and Stahel, "Robust statistics: the approach based on influence functions," vol. 196, no. John Wiley& Sons, 2011.

[21] A. M. Rousseeuw, Peter J and Leroy, "Robust regression and outlier detection," *New York John Wiley Sons*, vol. 589, 2005.

[22] D. L. Donoho and P. J. Huber, "The notion of breakdown point," *A festschrift Erich L. Lehmann*, vol. 157184, 1983.

[23] Hampel and F. R, "A general qualitative definition of robustness," *Ann. Math. Stat.*, pp. 1887–1896, 1971.

[24] R. E. Hoaglin, David C and Welsch, "The hat matrix in regression and ANOVA," *Am. Stat.*, vol. 32, pp. 17–22, 1978.

[25] R. D. Cook, "Influential Observations in Linear Regression," *J. Am. Stat. Assoc.*, vol. 74, pp. 169–174, 1977.

[26] A. Verardi, Vincenzo and McCathie, "The S-estimator of multivariate location and scatter in Stata," *Stata J.*, vol. 12, pp. 299–307, 2012.

[27] W. A. Stahel and Werner A, "Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen.," *ETH Zurich*, 1981.

[28] Donoho and D. L., "Breakdown properties of multivariate location estimators," *Tech. report, Harvard Univ. Boston. URL http//www-stat. stanford~…*, 1982.

[29] D. and others Verardi, Vincenzo and Gassner, Marjorie and Ugarte, "Robustness for dummies," *ECARES Work. Pap.*, 2012.

[30] C. Verardi, Vincenzo and Vermandele, "Univariate and multivariate outlier identification for skewed or heavy-tailed distributions," *Stata J.*, vol. 18, pp. 517–532, 2018.

[31] S. John and Weisberg, An R companion to applied regression. 2018.

[32] Y. V. Salibi.an-Barrera M., "A fast algorithm for S-regression estimates," *J. Comput. Graph. Stat.*, vol. 15, pp. 414–427, 2006.