Garcia, María del Carmen
Servy. Elsa
Instituto de Investigaciones Teóricas y Aplicadas, de la Escuela de Estadística

REGRESIÓN ROBUSTA: UNA APLICACIÓN[1]

1.- INTRODUCCIÓN

Cuando las observaciones que se modelan mediante un modelo clásico de regresión lineal están aproximadamente distribuidas en forma normal, el método de los mínimos cuadrados es un buen procedimiento de estimación, pues produce estimadores de los parámetros que tienen buenas propiedades estadísticas. Sin embargo, hay casos en que es evidente que la distribución de la variable respuesta no es normal y/o hay observaciones atípicas que tienen efectos negativos sobre las propiedades del modelo de regresión estimado.

Las observaciones provienen, a veces, de distribuciones que tienen colas más gruesas o largas que la distribución normal. Estas distribuciones tienden a generar valores atípicos que pueden tener una gran influencia sobre los resultados arrojados por el método de mínimos cuadrados.

Una observación atípica puede ser eliminada en base al conocimiento del problema bajo estudio, pero esta práctica no es recomendable desde el punto de vista de la objetividad que se busca con el análisis estadístico. Además cuando se cuenta con un modelo con varios regresores y la muestra es de tamaño grande, resulta dificultoso identificar cuales elementos del modelo resultan distorsionados por esas observaciones aberrantes.

Un procedimiento de regresión robusta es aquel que amortigua el efecto de las observaciones que serian muy influyentes si se usaran los mínimos cuadrados como método de estimación, y tiende a dejar grandes los residuos asociados con valores atípicos, facilitando así la identificación de los puntos influyentes. Además, un procedimiento de regresión robusta debería producir los mismos resultados que los mínimos cuadrados cuando la distribución básica es normal, y cuando hay valores atípicos.

En este trabajo se presenta una aplicación de métodos robustos para el ajuste de modelos de regresión a los efectos de explicar el ingreso del jefe del hogar, a partir de la información suministrada por la Encuesta Permanente de Hogares, relevada por el INDEC, para el aglomerado Rosario en la segunda onda de 2002. El mismo se realiza en el marco del proyecto “Métodos no paramétricos y semiparamétricos para el análisis de regresión con datos univariados y multivariados”.

2.- Regresión robusta

En un conjunto de datos pueden existir observaciones que tienen una magnitud diferente que el resto, ya sea porque presentan valores grandes (o chicos) en la respuesta, o en las variables explicativas (o en ambas). Estas observaciones atípicas se denominan “outliers” o puntos de “leverage”, respectivamente.

Ante la presencia de este tipo de observaciones no resulta conveniente utilizar el método de estimación de mínimos cuadrados. Entre las alternativas al mismo se encuentran los mode-

[1] En este trabajo ha participado, como auxiliar de investigación, la alumna de la Licenciatura en Estadística Lorena Manno.

los lineales generalizados, que permiten el uso de otras distribuciones distintas a la normal, las regresiones noparamétricas, que evitan hacer supuestos distribucionales, y la regresión robusta, que es resistente a la falta de normalidad.

El principal propósito de la regresión robusta es proveer resultados resistentes en presencia de "outliers". Para lograr esta estabilidad la regresión robusta limita la influencia de los mismos.

La idea básica de los métodos robustos es calcular un estimador β_r, que minimice la siguiente función de los residuos (ε_i),

$$\boldsymbol{\psi}(\boldsymbol{\beta}) = \sum_{i=1}^{n} w(\varepsilon_i)\ \varepsilon_i^2 \quad ,$$

donde w(.) es una función de ponderación que se introduce para reducir, e incluso eliminar, el efecto de los residuos altos.

Los pesos $w(\varepsilon_i)$, por lo tanto, se definen de forma que tomen valores pequeños para los residuos grandes. El cálculo de los estimadores se hace iterativamente y la elección del punto inicial y del algoritmo iterativo es crucial para alcanzar rápidamente convergencia.

Para que un estimador robusto sea de utilidad práctica debe tener punto de quiebre y eficiencia relativa altos. El punto de quiebre de un estimador se debe tomar en cuenta al seleccionar un procedimiento de estimación robusto.

. Punto de quiebre

El punto de quiebre es la mínima fracción de datos anómalos que puede causar que el estimador no se útil. Este valor se puede usar como una medida de la robustez del estimador.

Si el punto de quiebre, para una muestra de tamaño n, es 1/n, equivale a decir que una sola observación puede distorsionar el estimador y hacer su utilidad práctica nula.

El punto de quiebre de los estimadores mínimo cuadráticos es 1/n. Esto tiene un impacto potencialmente grave sobre su uso práctico ya que se puede dificultar la determinación del grado de contaminación de la muestra por los datos anómalos.

Los investigadores consideran que, generalmente, la fracción de datos que contaminan un conjunto de observaciones se encuentra entre el 1 y 10 %. En consecuencia, se desea que el punto de quiebre de un estimador sea mayor que el 10 %.

Esto ha conducido al desarrollo de estimadores de punto de quiebre alto.

. Eficiencia

Cuando las observaciones provienen de una distribución normal y no hay observaciones atípicas es correcto utilizar estimadores mínimo cuadráticos ordinarios (MCO).

Se define la eficiencia de un estimador robusto como el cociente entre el cuadrado medio residual obtenido con los MCO y el cuadrado medio residual obtenido con el procedimiento robusto. Obviamente esa medida de eficiencia se debe aproximar a 1.

Si bien en muchas investigaciones recientes se estudia la eficiencia asintótica de los estimadores robustos, lo más importante desde un punto de vista practico, es la eficiencia para muestras finitas, es decir, cómo funciona determinado estimador respecto a los MCO, para tamaños de muestra moderados o chicos.

Para el análisis de regresión se han desarrollado varios métodos robustos. Sin embargo los más usados en las aplicaciones son la estimación M de Huber, la estimación con valores de ruptura alto y combinaciones de ambos métodos. Algunos de estos procedimientos se basan en la regresión L_1, sugerida por Edgeworth en 1887 al comprobar que los mínimos cua-

drados estaban influídos por valores atípicos.

2.1- Métodos de estimación robustos

Se describen los aspectos estadísticos y computacionales de algunos procedimientos robustos.

Sea el modelo de regresión lineal

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \ ,$$

donde, $X = \{x_{ij}\}$ es una matriz nxp, $\mathbf{Y}=(y_1,y_2,\ldots,y_n)'$ el vector de los n valores observados de la variable respuesta y $\boldsymbol{\beta} = (\beta_1,\beta_2,...,\beta_p)'$ el vector p x1 de los parámetros, cuyos elementos se desean estimar.

Si la estimación se obtiene mediante el método de mínimos cuadrados se minimiza

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 \ ,$$

donde, x_i' es la i-ésima fila de la matriz X.

Con el fin de evitar los problemas que causan los valores atípicos a este método, se propuso reemplazar el cuadrado de los errores por el valor absoluto de los mismos. Específicamente, la ecuación de regresión se obtendría minimizando $\sum | \varepsilon_i |$. Esta regresión es llamada **regresión L1** o **regresión de suma absoluta mínima.**

Sin embargo, el uso de este método es restringido por las siguientes razones:
a) El vector de coeficientes estimados no es único.
b) La regresión L_1 resiste la presencia de valores atípicos en la dirección de Y, pero es poco efectiva para valores anómalos en la dirección X.
c) La eficiencia del estimador disminuye a medida que aumenta el número de casos.
d) Para obtener las estimaciones del coeficiente de regresión hay que resolver un problema de programación lineal, el cual es computacionalmente muy lento.

En 1973, Huber propuso un nuevo método de regresión, que combina la regresión L_1 y la regresión por mínimos cuadrados ordinarios (MCO), la estimación M.

. Estimación M

Es el enfoque más simple tanto desde el punto de vista teórico como computacional. Aunque no es robusto a los puntos de leverage es muy usado en el análisis de datos para lo cual debe ser supuesto que la contaminación está principalmente en la dirección de la respuesta.

La idea básica de este método de regresión reside en dar a los residuos pequeños, un peso cuadrático y a los residuos grandes, un peso lineal.

Específicamente, esta clase de estimadores robustos, minimizan con respecto a $\boldsymbol{\beta}$, una función ρ de los residuos

$$\sum_{i=1}^{n} \rho(\varepsilon_i) = \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i' \boldsymbol{\beta}) \ .$$

Un estimador de este tipo se llama **estimador M**. La función ρ se relaciona con la función

de verosimilitud para una elección adecuada de la distribución del error.

El estimador M, no es necesariamente invariante con respecto a cambios de escala (es decir, si se multiplicaran los errores $(y_i - \mathbf{x}_i' \boldsymbol{\beta})$ por una constante, la nueva solución de la ecuación podría no ser igual que la anterior)

Para obtener una versión invariante en relación a la escala (o a la variabilidad) se busca,

$$\underset{\boldsymbol{\beta}}{\text{Minimizar}} \sum_{i=1}^{n} \boldsymbol{\rho}\left(\frac{\boldsymbol{\varepsilon}_i}{\boldsymbol{\sigma}}\right) = \underset{\boldsymbol{\beta}}{\text{Minimizar}} \sum_{i=1}^{n} \boldsymbol{\rho}\left(\frac{y_i - \mathbf{x}_i' \boldsymbol{\beta}}{\boldsymbol{\sigma}}\right) \qquad (1)$$

En donde, σ es parámetro de escala generalmente desconocido. Existen varios métodos para estimar σ. Una elección muy frecuentemente para obtener un estimador robusto de escala, s, es la mediana de la desviación absoluta (MAD) (la desviación absoluta con respecto a la mediana)

$$s = \text{mediana } |e_i - \text{mediana } (e_i)|/ 0.6745,$$

siendo, $e_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}$.

La constante de ajuste 0.6745 hace que s sea aproximadamente un estimador insesgado de σ si n es grande y la distribución de los errores es normal.

Para minimizar la ecuación (1) se deriva ρ con respecto a β_j, j=0, 1,…,p y se iguala a cero. De este modo se obtiene el sistema de p+1 ecuaciones

$$\sum_{i=1}^{n} \boldsymbol{\psi}\left(\frac{y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}}{\hat{\boldsymbol{\sigma}}}\right) x_{ij} = 0 \qquad j=0,1,\ldots,p. \qquad (2)$$

donde $\Psi = \rho'$, x_{ij} es la i-ésima observación del j-ésimo regresor y $x_{i0} = 1$.

En general la función ψ es no lineal y se debe resolver la ecuación (2) por métodos iterativos. Se mencionan tres propuestas para obtener las soluciones:

1) ***El metodo de Minimos Cuadrados iterativamente ponderados (IRLS)***

Este método se suele atribuir a Beaton y Tukey (1974). Para su aplicación se supone un estimador inicial β_0.

Las p+1 ecuaciones (2) se escriben en la forma

$$\sum_{i=1}^{n} x_{ij} w_{i0} (y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}) = 0 \qquad (3)$$

donde, la función de peso w(x) se define $w(x) = \frac{\boldsymbol{\psi}(x)}{x}$ o específicamente como

$$w_{i0} = \begin{cases} \boldsymbol{\Psi}\,[\,(y_i - \mathbf{x}_i' \boldsymbol{\beta}_0)/s\,] \;/\; (y_i - \mathbf{x}_i' \boldsymbol{\beta}_0)/s & , \text{si } y_i \neq \mathbf{x}_i' \boldsymbol{\beta}_0 \\ 1 & , \text{si } y_i = \mathbf{x}_i' \boldsymbol{\beta}_0 \end{cases}$$

La forma matricial de la ecuación (3),

$$\mathbf{X' W_0 X \beta = X' W_0 y,}$$

representa las ecuaciones normales de mínimos cuadrados iterativamente ponderados.

Sea la siguiente formula iterativa:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \sigma\,(\mathbf{X'}\,\mathbf{W}^{(m)}\,\mathbf{X})^{-1}\,\mathbf{X'}\,\mathbf{W}^{(m)}\,(\mathbf{Y} - \mathbf{X'}\,\boldsymbol{\beta}^{(m)}),$$

siendo, **W** una matriz diagonal, de dimensión n, de "pesos" con elementos diagonales w_{10}, $w_{20},\ldots, w_{n0}$ obtenidos por la ecuación (4).

Por lo general solo se requieren unas pocas iteraciones para alcanzar la convergencia.

El valor inicial del proceso iterativo depende de la forma de la función Ψ.

2) ***El método de Newton Raphson***: cuya formula de iteración es:

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + \hat{\sigma}\left[\sum \mathbf{x}_i' \psi'\left(\frac{y_i - \mathbf{x}_i'\,\hat{\boldsymbol{\beta}}^{(m)}}{\hat{\sigma}}\right)\mathbf{x}_i\right]^{-1} \mathbf{X}'\psi\left(\frac{y_i - \mathbf{x}_i'\,\hat{\boldsymbol{\beta}}^{(m)}}{\hat{\sigma}}\right)$$

3) ***El método H de Huber***: teniendo en cuenta que el estimador mínimo cuadrático puede ser escrito como

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X'X})^{-1}\,\mathbf{X}'\mathbf{e},$$

entonces, se puede escribir el siguiente proceso iterativo

$$\hat{\boldsymbol{\beta}}^{(m'+1)} = \hat{\boldsymbol{\beta}}^{(m')} + (\mathbf{X'X})^{-1}\,\mathbf{X}'\,\hat{e}^{(m)}.$$

Huber sugirió el siguiente proceso iterativo

$$\hat{\boldsymbol{\beta}}^{(m'+1)} = \hat{\boldsymbol{\beta}}^{(m')} + \hat{\sigma}(\mathbf{X'X})^{-1}\,\mathbf{X}'\,\psi\left(\frac{y_i - \mathbf{x}_i'\,\hat{\boldsymbol{\beta}}^{(m)}}{\hat{\sigma}}\right)$$

La idea aquí es reemplazar en cada iteración el residuo $e_i = y_i - x_i'\,\hat{\boldsymbol{\beta}}$ por el residuo modificado $\boldsymbol{\psi}\left(\frac{y_i - \mathbf{x}_i'\,\hat{\boldsymbol{\beta}}}{\hat{\boldsymbol{\sigma}}}\right)$.

Existen varias propuestas para la función peso Ψ las principales se presentan en el cuadro siguiente

Cuadro 1 Funciones de peso

Criterio	Expresión
Función t de Huber a= 0.6745	$\Psi(z) = \begin{cases} z & si \mid z \mid \leq a \\ a.sign(z) & en\ otro\ caso \end{cases}$
E_a de Ramsay a=0.3	$\Psi(z) = \begin{cases} z.exp(-a\lvert z\rvert) & si\ \lvert z\rvert \leq \infty \\ 0 & en\ otro\ caso \end{cases}$
Función de onda de Andrews A a=1.339	$\Psi(z) = \begin{cases} a.sen(z/a) & si\ \lvert z\rvert \leq \pi.a \\ 0 & en\ otro\ caso \end{cases}$
Función 17ª de Hampel a=1.7, b=3.4 c=8.5	$\Psi(z) = \begin{cases} z & si\ \lvert z\rvert \leq a \\ a.sign(z) & si\ a < \lvert z\rvert \leq b \\ a.(c- \lvert z\rvert)/ (c-b).sign(z) & si\ b < \lvert z\rvert \leq c \\ 0 & en\ otro\ caso \end{cases}$

Los procedimientos de regresión robusta se pueden clasificar de acuerdo con el comportamiento de su función de peso Ψ. Esta función controla el factor de ponderación que se asigna a cada residuo (además de una constante de proporcionalidad) y a veces se la llama **función de influencia**.

La función Ψ para los mínimos cuadrados no es acotada, por lo que los estimadores no son robustos.

La función Ψ para la t de Huber es monótona y no pondera los residuos grandes con tanta intensidad como los mínimos cuadrados.

Las tres últimas funciones de influencia decaen a medida que el residuo se hace más grande. La función E_a de Ramsay decae suavemente, esto es, la función Ψ es asintótica a cero para | z | grande. La función de onda de Andrews y la función 17 A de Hampel decaen rápidamente, es decir, la función Ψ es igual a cero cuando | z | es suficientemente grande.

El proceso robusto requiere, además, la especificación de ciertas "constantes de ajuste" para las funciones Ψ.

Los estimadores M se pueden alterar debido a valores atípicos en el espacio de x (valores con "leverage" alto), siendo su punto de quiebre, justamente, 1/n.

2.2.- Otros estimadores robustos

Como los estimadores M funcionan mal con respecto al punto de quiebre se han desarrollado métodos alternativos. En esta sección se presentan algunos métodos propuestos y sus ventajas y desventajas.

Para paliar la desventaja que presentan los estimadores M y MCO con respecto al punto de quiebre se ha tratado de desarrollar otros estimadores con punto de quiebre alto.

Es conveniente que los estimadores posean un punto de quiebre de alrededor del 50%.

Mínimos cuadrados recortados (LTS)

La estimación LTS es un método con valores de ruptura alto introducido por Rousseeauw (1984). El valor de ruptura es una medida de la proporción de contaminación que un procedimiento puede resistir y mantener su robustez.

Los estimadores se calculan resolviendo

$$\underset{\boldsymbol{\beta}}{\text{Minimizar}} \sum_{i=1}^{h} \varepsilon_{(i)}^2 ,$$

siendo, $\varepsilon_{(1)}^2 < \varepsilon_{(2)}^2 < .... < \varepsilon_{(n)}^2$ los residuos ordenados al cuadrado y h, a determinar, es definido en el rango $\frac{n}{2}+1 \le h \le \frac{3n+p+1}{4}$.

Las mejores propiedades de robustez se obtienen cuando h=n/2 aproximadamente, en cuyo caso se alcanza un punto de quiebre de 50%.

La función objetivo de estos estimadores es suave haciendo que el estimador LTS sea menos sensible a efectos locales que el MCO.

La versión robusta del coeficiente de determinación para la estimación LTS se define como

$$R_{LTS}^2 = 1 - \frac{S_{LTS}^2(\mathbf{X},\mathbf{Y})}{S_{LTS}^2(\mathbf{1},\mathbf{Y})},$$

siendo,

$S_{LTS}^2(\mathbf{X},\mathbf{Y})$ es el estimador LTS robusto del parámetro de escala en el modelo completo y

$S_{LTS}^2(\mathbf{1},\mathbf{Y})$ es el estimador LTS robusto del parámetro de escala en el modelo que sólo toma en cuenta la ordenada .

Estimadores S

La estimación S es un método con valores de ruptura alto introducido por Rousseauw y Yohai (1984). Con el mismo valor de ruptura, tiene una eficiencia estadística más alta que el LTS.

El estimador se obtiene a partir de,

$$\underset{\boldsymbol{\beta}}{\text{Minimizar}}\ S[e_1(\boldsymbol{\beta}), e_2(\boldsymbol{\beta}), ..., e_n(\boldsymbol{\beta})],$$

siendo, $e_i(\boldsymbol{\beta})$ los residuos para una probable solución para β y $S[e_1(\boldsymbol{\beta}), e_2(\boldsymbol{\beta}), ..., e_n(\boldsymbol{\beta})]$ se determina como solución de

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{\varepsilon_i}{\sigma}\right)=k\,.$$

Rousseeuw y Yohai (1984) sugieren la función ρ

$$\rho(z)=\begin{cases}\dfrac{z^2}{2}-\dfrac{z^4}{2c^2}-\dfrac{z^6}{6c^4} & |z|\le c\\[2ex] \dfrac{c^2}{6} & |z|>c\end{cases}$$

La versión robusta del coeficiente de determinación para la estimación S se define como

$$R_S^2=1-\frac{(n-p)S_p^2}{(n-1)S_\mu^2},$$

siendo,

S_p^2 es el estimador S robusto del parámetro de escala en el modelo completo y

S_μ^2 es el estimador S robusto del parámetro de escala en el modelo considera la ordenada solamente.

Los estimadores S tienen un punto de quiebre alto (50%) y mayor eficiencia asintótica que los Mínimos cuadrados recortados (LTS).

Estimadores MM

Estimación MM introducido por Yohai (1987) combina la estimación M con valores de ruptura alto.

Su objetivo fue producir un estimador de punto de quiebre alto, que mantuviera una buena eficiencia. Este estimador tiene tres etapas:

1. El estimador inicial es un estimador S, por lo que tiene punto de quiebre alto.

2. En la segunda etapa se calcula un estimador M de la desviación estándar del error, con los residuos del estimador S inicial.

3. El último paso consiste en calcular un estimador M para los parámetros de regresión mediante una función ψ que decaiga rápidamente, esto es, que asigne un peso igual a cero a los residuos suficientemente grandes.

Para estimar los parámetros, β, se busca el mínimo de

$$Q_{MM}=\sum_{i=1}^{n}\rho\left(\frac{y_i-\mathbf{x}_i'\boldsymbol{\beta}}{\hat{\sigma}}\right).$$

El valor $\hat{\sigma}$ es el estimador robusto de escala obtenido en el segundo paso. Las funciones ρ que se pueden seleccionar son mencionadas anteriormente en el procedimiento S.

La versión robusta del coeficiente de determinación es

$$R^2 = \frac{\sum \rho\left(\frac{y_i - \hat{\boldsymbol{\mu}}}{s}\right) - \rho\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{s}\right)}{\sum \rho\left(\frac{y_i - \hat{\boldsymbol{\mu}}}{s}\right)}$$

$\hat{\boldsymbol{\mu}}$ es el estimador MM de tendencia central y s el estimador robusto de escala en el modelo completo.

Los estimadores MM tienen alta eficiencia y trabajan bien en la mayor parte de los escenarios de valores atípicos. Su único punto débil aparece en casos en donde hay un gran porcentaje de valores atípicos en el espacio x que tengan residuos de tamaño moderado.

En la estimación por MM se manejan mucho mejor los valores atípicos grandes, aun cuando sean puntos de gran influencia.

2.3.- Diagnósticos

Distancia robusta y "leverage"

La Distancia robusta se define como

$$RD(x_i) = \left[\left(x_i - T(X)\right)' C(X)^{-1} \left(x_i - T(X)\right)\right]^{1/2},$$

donde, T(X) y C(X) son el vector de tendencia central y la matriz de dispersión robustos.

Esta distancia se utiliza para comprobar si la observación es extrema con respecto a la variable X o dicho de otra forma si es un punto de "leverage".

La variable "leverage" se define como

$$LEVERAGE = \begin{cases} 0 & \text{si} \quad RD(x_i) \le C(p) \\ 1 & \text{en otro caso} \end{cases}.$$

El punto de corte es $C(p) = \sqrt{\chi^2_{p;1-\alpha}}$, siendo p el número de parámetros del modelo.

"Outliers"

Los residuos e_i, i=1,…,n, basados sobre los estimadores robustos descriptos se utilizan para detectar observaciones atípicas com respecto a la respuesta o "outliers".

La variable "outliers" se define como

$$OUTLIER = \begin{cases} 0 & \text{si } |e_i| \le k\boldsymbol{\sigma} \\ 1 & \text{en otro caso} \end{cases}$$

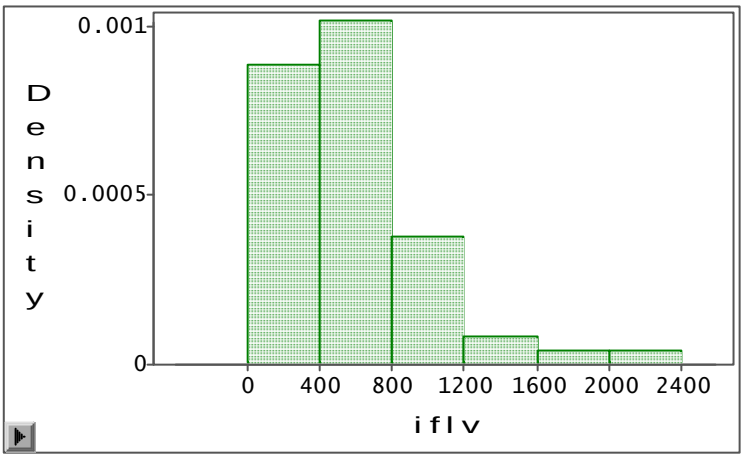
3.- Aplicación

La información relativa a los individuos que se releva mediante la Encuesta Permanente de Hogares (EPH), que lleva a cabo el Instituto Nacional de Estadísticas y Censos (INDEC), ofrece importantes ventajas para el análisis empírico de variables registradas en la misma. Variables demográficas obtenidas de la encuesta se pueden incorporar a ecuaciones de regresión, conjuntamente con otras características propias de los individuos.
Las variables registradas en la EPH utilizadas para explicar el ingreso de fuente laboral (Y), denominadas variables explicativas (X), se definen a continuación. Para el caso de los varones se agrega a cada variable el sufijo V y para las mujeres se agrega el sufijo M (Blaconá y otros, 2002).

- X_1 (edad): variable continua calculada a partir de la fecha de nacimiento declarada por el encuestado.
- X_2: promedio de los años de escolaridad declarados por el encuestado.
- X_3: cantidad de hijos menores de 6 años.
- X_4: cantidad de hijos entre 6 y 18 años.

La estimación de modelos de regresión que explican los ingresos individuales presenta algunas dificultades, entre las que se pueden destacar la no normalidad de la variable respuesta.
En el gráfico 1 se visualiza la distribución de la variable ingreso

Gráfico 1 Distribución de variable ingreso



La distribución es asimétrica, lo que muestra que la variable no tiene distribución normal: Una distribución de este tipo es, generalmente, generadora de puntos atípicos.
Se aplica un procedimiento de selección de variables mediante el cual se obtiene que las variables que explican mejor el ingreso son los años de escolaridad del varón y la mujer.
Para el análisis de los datos se utilizan los procedimientos REG y ROBUSTREG de SAS.

3.1.- Estimación mínimo cuadrática del modelo de regresión para la variable ingreso
Utilizando las variables años de escolaridad del hombre y la mujer se realiza la estimación mínimo cuadrática de la función de regresión, la cual viene dada por

$$\hat{y}_i = -10.583 + 30.060\, x_{2Vi} + 39.358\, x_{2Mi}$$
$$\qquad\qquad (18.249) \qquad (16.709)$$

siendo los valores entre paréntesis los errores estándares.

La variable años de escolaridad de la mujer ayuda a explicar la respuesta (p=0.022), en el sentido que es mejor incluirla en el modelo que no incluirla, mientras que los años de escolaridad del varón no resulta significativa (p=0.1051).

Gráfico 2 Residuos vs valores ajustados

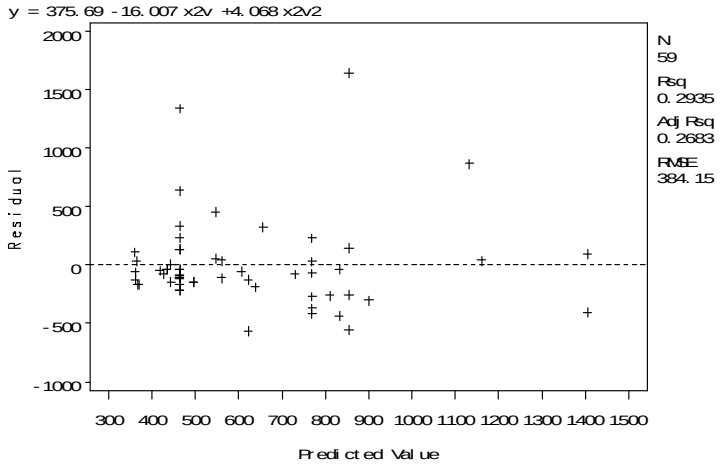
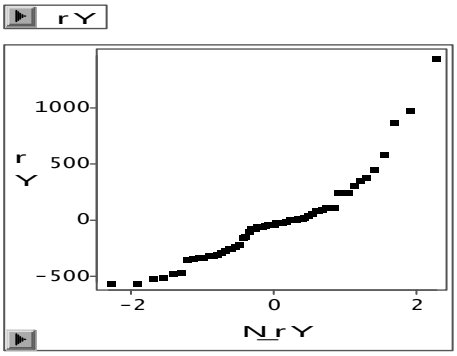


Gráfico 3 Gráfico probabilístico normal de los residuos



En los gráficos de residuos y de normalidad se visualizan observaciones con residuos grandes.
Mediante un análisis de influencia se confirma que de esas observaciones atípicas, 6 presentan “leverage” alto, es decir, extremas con respecto a las variables explicativas y 3 son “outliers”, inusuales con respecto a la respuesta.
En este ajuste resulta llamativo que el coeficiente de la variable escolaridad del varón no sea significativo cuando realmente tendría que serlo.

3.2.- Estimación robusta del modelo de regresión para la variable ingreso

Se aplican los procedimientos M, LTS, S y MM de regresión robusta.

. Regresión M

Los parámetros estimados y un resumen del tipo de observaciones aberrantes se presentan en la tabla 1 y 2.

Tabla 1 Estimaciones de los parámetros y pruebas de hipótesis

Parámetro	Estimador	Error Estándar	Limites de Confianza	Chi- Cuadrado	Pr > ChiSq
Ordenada	161.73	76.96	10.8;312.6	4.42	0.0356
x2v	41.92	10.78	20.8; 63.1	15.11	0.0001
x2m	-3.27	9.87	-22.6;16.1	0.11	0.7405
Escala	157.77				

Tabla 2 Resumen de los diagnósticos

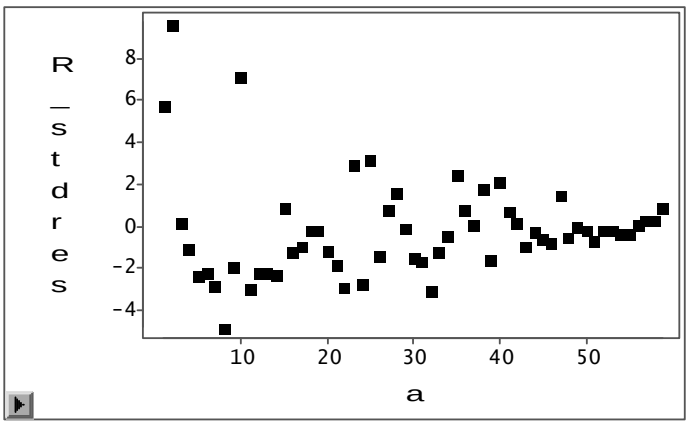
Tipo de Observación	Proporción	Punto Corte	Número
Outlier	0.1186	3.0000	7
Leverage	0.0508	2.7162	3

$R^2 = 0.1636$

Si bien la variable de interés resulta significativa el conjunto de datos contiene algunos puntos de leverage alto, por lo que el procedimiento no resulta robusto. Sería recomendable usar otros métodos.

El gráfico siguiente muestra los residuos estandarizados de la regresión anterior. En el mismo se observan los outliers mencionados anteriormente.

Gráfico 4 Residuos estandarizados de la regresión M



. Regresión LTS

Se utiliza este procedimiento pues tiene un punto de quiebre mayor que el anterior. Estos estimadores se usan principalmente para detectar los "outliers" en el conjunto de datos. Luego, para estimar los parámetros finales se usa una regresión mínimo cuadrática ponderada adjudicando menos peso a los puntos detectados en el paso anterior.

Tabla3 Estimaciones LTS de los parámetros

Parámetro	Estimación
Ordenada	150.44
x2v	39.28
x2m	-7.66
Escala (sLTS)	137.85
Escale (Wscale)	145.17

La tabla 3 muestra los parámetros de escala y de las covariables estimados. Se presentan dos estimadores de escala robustos, siendo más eficiente el estimador ponderado (Wscale). La estimación LTS produce resultados consistentes con los de la estimación M.

La tabla 4 muestra los diagnósticos con los puntos de leverage y outliers basados sobre los estimadores LTS.

Tabla 4 Resumen de los Diagnósticos

Tipo de Observación	Proporción	Punto Corte	Número
Outlier	0.1525	3.0000	9
Leverage	0.0508	2.7162	3

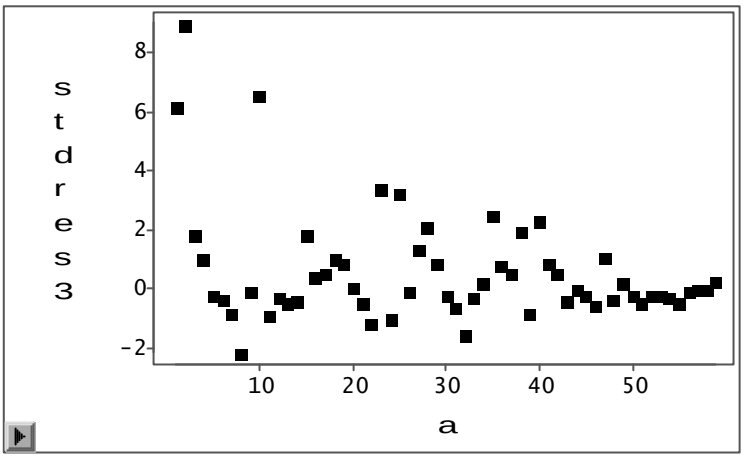
$R^2 = 0.2847$

Los estimadores que se presentan a continuación son los estimadores mínimo cuadráticos calculados después de detectar los outliers.

Tabla 5 Estimaciones de los parámetros del ajuste final por mínimos cuadrados ponderados

Parámetro	Estimador	Error Estándar	Limites de Confianza	Chi- Cuadrado	Pr > ChiSq
Ordenada	210.94	65.11	83.31;338.57	10.49	0.0012
x2v	30.59	9.25	12.45; 48.73	10.92	0.0009
x2m	-0.32	8.54	-17.05; 16.42	0.00	0.9705
Escala	168.23				

Gráfico 5 Residuos estandarizados de la regresión



Este gráfico es bastante similar al obtenido en la regresión M.

. Regresión S

Estos estimadores poseen un valor de ruptura más alto que los anteriores. Las tablas 6 y 7 presentan las estimaciones y los diferentes tipos de observaciones inusuales.

Tabla 6 Estimaciones de los parámetros y pruebas de hipótesis

Parámetro	Estimador	Error Estándar	Limites de Confianza	Chi-Cuadra-do	Pr > ChiSq
Ordenada	161.61	77.91	8.89;314.33	4.30	0.0381
x2v	41.95	11.28	19.85;64.05	13.84	0.0002
x2m	-3.27	10.37	-23.58;17.04	0.10	0.75
Escala	252.05				

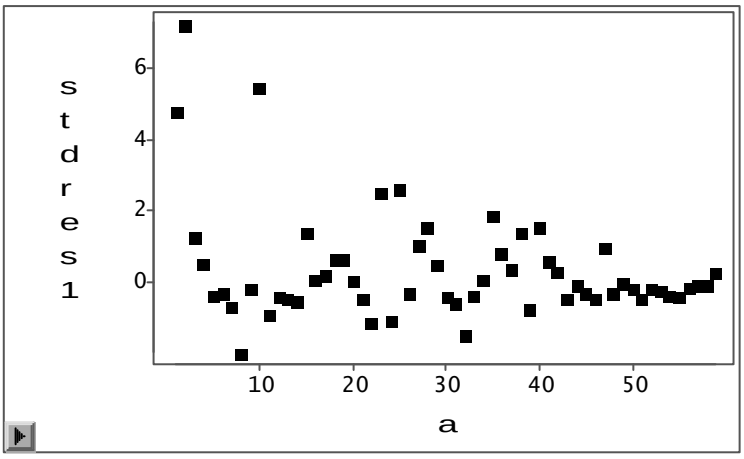
Tabla 7 Resumen de los Diagnósticos

	Tipo de Observación	Proporción	Punto Corte	Número
Outlier	0.0508	3.0000	3	
	Leverage	0.0508	2.7162	3

$R^2 = 0.2874$

Las magnitudes de las estimaciones son similares a las producidas por el método M.

Gráfico 6 Residuos estandarizados de la regresión



Si bien este gráfico muestra los mismos outliers que los destacados por los otros métodos sus magnitudes son inferiores.

. Regresión MM

Estos estimadores poseen una ruptura alta y la eficiencia es más alta que la estimación S.

Tabla 8 Estimaciones de los parámetros y pruebas de hipótesis

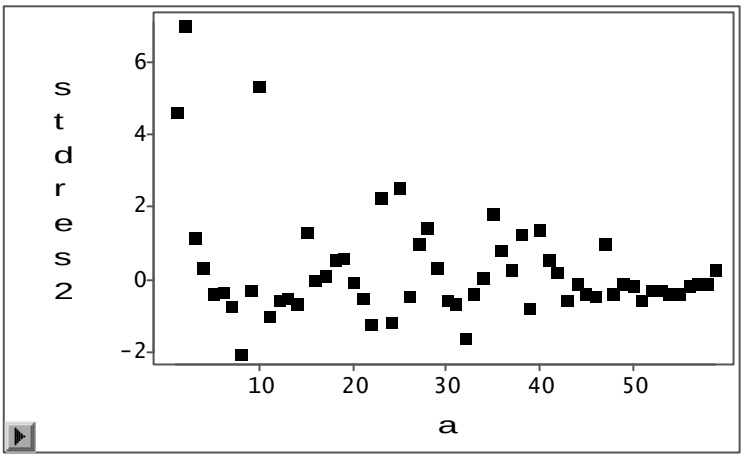
Parámetro	Estimador	Error Estándar	Limites de Confianza	Chi-Cua-drado	Pr > ChiSq
Ordenada	148.79	79.20	-6.45; 304.02	3.53	0.0603
x2v	44.65	11.40	22.29;67.00	15.33	<.0001
x2m	-3.16	10.63	-23.98; 17.67	0.09	0.7664
Escala	255.57				

Tabla 9 Resumen de los Diagnósticos

Tipo de Observación	Proporción	Punto Corte	Número
Outlier	0.0508	3.00	3
Leverage	0.0508	2.7162	3

$R^2 = 0.1805$

Gráfico 7 Residuos estandarizados de la regresión



Los dos ajustes robustos S y M presentan resultados similares. La ventaja de ellos es que tienen punto de quiebre alto que nos dice que resisten un alto porcentaje de punto anómalos sin que el modelo de regresión se torne inútil.

4. Discusión

Los métodos de regresión robusta ofrecen al analista de datos una gran ayuda para tratar puntos atípicos y observaciones muy influyentes.
Se puede considerar que los métodos de regresión robusta son procedimientos para aislar puntos anormalmente influyentes, de manera de poder estudiarlos más profundamente.

El análisis robusto se puede utilizar como confirmatorio de mínimos cuadrados. Siempre que se hace un análisis por mínimos cuadrados, seria conveniente también hacer un ajuste robusto. Si concuerdan, en forma sustancial, los resultados de los dos procedimientos se deben usar los resultados de los mínimos cuadrados, sin embargo, si difieren, se deben identificar las razones de tales diferencias.

La estimación M introducida por Huber (1973) es el enfoque más simple tanto computacional como teóricamente. Aunque no es robusto a los puntos de leverage es muy usado en el análisis de datos, cuando se puede suponer que la contaminación está principalmente en la dirección de la respuesta.

Los estimadores M poseen punto de quiebre bajo, en consecuencia, una sola observación anómala puede destruir la utilidad del estimador.

La estimación LTS (Mínimos cuadrados recortados) es un método con valores de ruptura alto. El valor de ruptura es una medida de la proporción de contaminación que un procedimiento puede resistir y mantener su robustez.

La estimación S, que es un método con valores de ruptura alto, tiene una eficiencia estadística más alta que el LTS.

La estimación MM combina la estimación M con valores de ruptura alto. Posee la propiedad de ruptura alta y eficiencia más alta que la estimación S.

En la aplicación se observa que en todos los análisis robustos hay coincidencia sobre la significación de los coeficientes de las variables explicativas. Los métodos robustos muestran que la variable años de escolaridad del varón es mejor predictor del ingreso del hogar que el de la mujer, contrastando con los resultados obtenidos usando el método de mínimos

cuadrados.

El número de observaciones atípicas detectadas por los métodos S y MM resultan inferiores a las informadas por los otros procedimientos.

REFERENCIAS BIBLIOGRÁFICAS

Cohen, R. A. (2002), "SASâ Meets Big Iron:High Performance Computing in SAS Analytic Procedures", *Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

De Long, J.B., Summers, L.H. (1991), "Equipment investment and economic growth". *Quarterly Journal of Economics*, **106**, 445-501.

Hampel, F. R., Ronchetti, E.M., Rousseeuw, P.J.and Stahel, W.A. (1986), *Robust Statistics, The Approach Based on Influence Functions*, John Wiley & Sons, New York.

Hawkins, D. M., Bradu, D. and Kass, G. V. (1984), "Location of several outliers in multiple regression data using elemental sets," *Technometrics* , **26**, 197-208.12

Holland, P. and Welsch, R. (1977), "Robust regression using interatively reweighted least-squares," Commun. Statist. Theor. Meth. **6**, 813-827.

Huber, P.J. (1973), "Robust regression: Asymptotics, conjectures and Monte Carlo," *Ann. Stat.,* **1**, 799-821.

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

Ronchetti, E. (1985). "Robust Model Selection in Regression" *Statistics and Probability Letters*, **3**, 21-23.

Rousseeuw, P.J. (1984). "Least Median of Squares Regression," *Journal of the American Statistical Association*, **79**, 871-880.

Rousseeuw, P.J. and Hubert, M. (1996). "Recent Development in PROGRESS" C*omputational Statistics and Data Analysis,* **21**, 67-85.

Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection.* Wiley-Interscience, New York (Series in Applied Probability and Statistics), 329 pages. ISBN 0-471-85233-3.

Rousseeuw, P.J. and Van Driessen, K. (1999). "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics* , **41**, 212-223.

Rousseeuw, P.J. and Van Driessen, K. (1998). "Computing LTS Regression for Large Data Sets," Technical Report, University of Antwerp.

Rousseeuw, P.J. and Yohai, V. (1984), "Robust Regression by Means of S estimators", in in Robust and Nonlinear Time Series Analysis, edited by J. Franke, W. Härdle, and R.D. Martin, Lecture Notes in Statistics 26, Springer Verlag, New York, 256-274.

Ruppert, D. (1992), "Computing S Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics*, **1**, 253-270.

Yohai V.J. (1987), "High Breakdown Point and High Efficiency Robust Estimates for Regression," *Annals of Statistics*, **15**, 642-656.

Yohai V.J., Stahel, W.A. and Zamar, R.H. (1991), "A Procedure for Robust Estimation and

Inference in Linear Regression," in Stahel, W.A. and Weisberg, S.W., Eds., *Directions in Robust Statistics and Diagnostics, Part II,* Springer-Verlag, New Work.

Yohai, V.J. and Zamar, R.H. (1997), "Optimal locally robust M estimate of regression". Journal of Statist. Planning and Inference, **64**, 309-323.

Zaman, A., Rousseeuw, P.J., Orhan, M. (2001), "Econometric applications of high-breakdown robust regression techniques", *Econometrics Letters*, **71**, 1-8.