Anales Acad. Nac. de Cs. Ex., Fís. y Nat., tomo 53 (2001): 21-31.

MÉTODOS ESTADÍSTICOS ROBUSTOS

Víctor J. Yohai

Departamento de Matemáticas, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. Ciudad Universitaria, Pabellón 1, 1428 Buenos Aires. E-mail: vyohai@dm.uba.ar

Resumen

Los métodos estadísticos clásicos generalmente se obtienen optimizando su comportamiento cuando los errores tienen distribución normal. Sin embargo, pequeñas desviaciones de la normalidad o la presencia en la muestra de una pequeña proporción de puntos atípicos pueden afectar su comportamiento notablemente y llevar a conclusiones erróneas. Como alternativa, se proponen procedimientos robustos que tienen las siguientes propiedades: (i) son poco sensibles a las desviaciones mencionadas y (ii) tienen un comportamiento altamente eficiente cuando los errores son normales. En particular, se presentan estimadores robustos para el modelo de medición y regresión lineal.

Palabras clave: Métodos robustos, Regresión, Modelo de medición.

Abstract

In general, classical statistical methods are derived assuming normally distributed errors. However small deviations of normality or the presence in the sample of a small fraction of outliers may spoiled these procedures completely. As alternative, we propose robust statistical procedures which have the following properties (i) They are not much sensitive to deviations from normality or to the presence of a few outliers and (ii) they are highly efficient for normal errors. In particular, we present here robust estimates for the measurement model and linear regression

Key words: Robust methods, Measurement model, Linear regression.

1. Introducción

La mayor parte de los métodos estadísticos clásicos se basan en la hipótesis que los datos son normales.

La distribución normal, también llamada distribución de Gauss, está caracterizada por dos parámetros: la media μ y la varianza σ^2 y la simbolizaremos por $N(\mu, \sigma^2)$. Su función de densidad es

$$f^N_{\mu,\sigma^2}(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma}.$$

Conferencia pronunciada en su incorporación como Académico Titular, el 28 de abril de 2000.

En la Figura 1 se representa esta densidad para el caso $\mu = 0$ y $\sigma^2 = 1$. Su función de distribución está dada por

$$F^N_{\mu,\sigma^2}(x) = P(X \le x) = \int_{-\infty}^{x} f^N_{\mu,\sigma^2}(u)du,$$

donde P indica probabilidad y X denota a la correspondiente variable aleatoria.

El principal argumento que se utiliza para justificar en Estadística la hipótesis de normalidad es el Teorema Central del Límite. Un enunciado informal de este teorema es el siguiente: Sea U una variable aleatoria que se puede expresar como

$$U = Z_1 + Z_2 + \ldots + Z_n,$$

donde $Z_1, Z_2, \ldots, Z_n$ son variables aleatorias independientes y todas del "mismo orden", entonces la distribución de U se puede aproximar por la distribución normal. Debemos enfatizar dos aspectos de este enunciado.

1. Las variables deben ser todas del mismo orden.

2. La distribución de U no resulta exactamente normal sino sólo aproximadamente normal.

Una formalización de este enunciado es el Teorema de Lindeberg donde se precisa el concepto de sumandos del mismo orden, ver por ejemplo Billingsley [2].

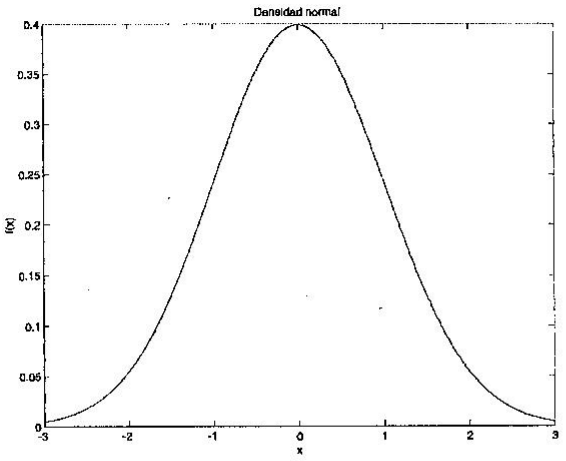

Figura 1. Densidad normal

Consideremos primero un modelo simple: el modelo de medición. Supongamos que se quiere medir una magnitud μ. En general, como los métodos de medición no están exentos de error, el valor medido X será

$$X = \mu + \varepsilon, \qquad (1)$$

donde ε es el error de medición. Por lo tanto, para poder estimar μ con precisión se requerirá repetir la medición varias veces. Supongamos que se realizan n mediciones y se obtienen los valores $X_1, X_2, \ldots, X_n$. Luego se tendrá

$$X_i = \mu + \varepsilon_i, \quad i = 1, \ldots, n. \qquad (2)$$

Las hipótesis habituales que se realizan sobre los errores de medición ε_i son

P1. $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ son variables aleatorias independientes.

P2. Todos los ε_i tienen una misma distribución F que es simétrica respecto de 0.

P3. La distribución F es normal.

Un estimador óptimo para μ bajo las hipótesis P1, P2 y P3 es el promedio o media muestral que está dado por

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

En efecto, si P1-P3 se cumplen, entonces la media muestral es el estimador insesgado de mínima varianza. Ver por ejemplo Bickel [1].

La justificación que se da para suponer P3 es el Teorema Central del Límite dado que generalmente es posible pensar que el error es debido a muchas causas independientes. Sin embargo $\overline{X}_n$ puede no ser un buen estimador de μ por las siguientes razones:

1. Podría haber una causa de error que predomina netamente sobre las otras, en cuyo caso el Teorema Central del Límite no se cumpliría, ya que no todos los errores serían del mismo orden.

2. Aun en el caso de que el Teorema Central del Límite valiese, la distribución de ε sería sólo aproximadamente normal y no exactamente normal.

El punto 2 lleva a preguntarse qué ocurre con $\overline{X}_n$ cuando la distribución de los

errores es sólo aproximadamente normal. ¿Seguirá siendo un buen estimador de μ? La respuesta es negativa ya que $\bar{X}_n$ es extremadamente sensible a la hipótesis de normalidad, especialmente a la presencia de observaciones atípicas conocidas en el lenguaje estadístico como "outliers". Estos "outliers" pueden ser debidos a varias causas:

- Errores al transcribir los datos
- Mal funcionamiento del instrumento de medición.
- Condiciones ambientales anormales, por ejemplo: humedad, temperatura, presión, etc.
- En el caso de variables económicas, éstas pueden estar afectadas por circunstancias institucionales anormales nacionales o internacionales.

Matemáticamente, estos "outliers" pueden modelarse por una distribución *normal contaminada*. En vez de suponer que la distribución F de los errores es normal, este modelo establece que F es de la forma

$$F = (1-\varepsilon)F^N_{0,\sigma^2} + \varepsilon H, \qquad (3)$$

donde ε es pequeño (por ejemplo 0.05) y H es una distribución simétrica arbitraria. Al conjunto de distribuciones de esta forma lo llamaremos v_ε. Una distribución F es de la forma (3) si con probabilidad $1-\varepsilon$ las observaciones tienen distribución normal con media 0 y varianza σ^2, F^N_{0,σ^2}, y con probabilidad ε son "outliers" que provienen de una distribución H generalmente desconocida. Cuando la distribución F es normal contaminada, el comportamiento de $\bar{X}_n$ deja de ser óptimo y puede dar resultados totalmente aberrantes. En efecto, si la muestra proviene del modelo (2) con errores ε_i con distribución F^N_{0,σ^2}, la varianza de $\bar{X}_n$ es σ^2/n. En cambio, si la muestra proviene de una F definida como en (3), la varianza puede ser infinita por más pequeño que sea ε.

Otra manera de ver la extrema sensibilidad de $\bar{X}_n$ frente a la presencia de "outliers" es la siguiente. Supongamos que las observaciones $X_1, ..., X_n$ provienen de una muestra normal y se agrega un "outlier" X_{n+1}. Entonces resulta

$$\bar{X}_{n+1} = \frac{n}{n+1}\bar{X}_n + \frac{1}{n+1}X_{n+1}.$$

Luego $\bar{X}_{n+1}$ puede hacerse tan grande como se quiera tomando X_{n+1} suficientemente grande. Esto muestra que un solo "outlier" puede arruinar completamente este estimador.

Este comportamiento del promedio, extremadamente sensible frente a pequeñas desviaciones de la normalidad, nos lleva a definir el concepto de estimador robusto. Un estimador $\hat{\mu}$ será robusto si

1. $\hat{\mu}$ es poco afectado por la presencia de un pequeño porcentaje de "outliers".
2. Es altamente eficiente bajo la hipótesis de normalidad. Por ejemplo se puede pedir que

$$\frac{\mathrm{var}(\bar{X}_n)}{\mathrm{var}(\hat{\mu})} \geq 0.95,$$

donde var indica varianza.

En la Sección 2 se presentan estimadores robustos para el modelo de medición. En la Sección 3 se introduce una medida de robustez: el punto de ruptura. En la Sección 4 se estudian estimadores robustos para el modelo de regresión lineal. Finalmente en la Sección 5 se mencionan otras áreas de la Estadística donde se aplican estimadores robustos.

2. Estimadores robustos para el modelo de medición

El estimador robusto más simple y antiguo para el modelo de medición (2) es la *mediana*. Dada una muestra $X_1, X_2, ..., X_n$, consideremos la muestra ordenada de menor a mayor $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$. La mediana se define como

$$\hat{\mu}_{MED} = \begin{cases} X_{(m+1)} & \text{si} \quad n = 2m+1 \\ \dfrac{X_{(m)} + X_{(m+1)}}{2} & \text{si} \quad n = 2m. \end{cases}$$

Es decir $\hat{\mu}_{MED}$ es el valor central si n es impar y el promedio de los dos valores centrales si n es par.

Veamos en un ejemplo cómo influye un "outlier" en la media y la mediana. Consideremos la muestra 2 < 2.05 < 2.06 < 2.08 < 2.09 < 2.11. Los valores centrales son 2.06 y 2.08, luego $\hat{\mu}_{MED} = 2.07$. Un simple cálculo muestra que $\bar{X} = 2.065$. Si se agrega a la muestra el "outlier" 10, el valor central resulta 2.08, y por lo tanto $\hat{\mu}_{MED} = 2.08$. En cambio el nuevo valor de la media es $\bar{X} = 3.20$. Si el "outlier" agregado es 100 en vez de 10 se obtiene $\hat{\mu}_{MED} = 2.08$ y $\bar{X} = 16.06$. En este ejemplo se observa como la mediana es mucho menos sensible que la media a la presencia de "outliers".

Sin embargo, el comportamiento de $\hat{\mu}_{MED}$ bajo datos normales es mucho menos eficiente que el de $\bar{X}_n$. En efecto, se puede demostrar que si se tiene una muestra aleatoria $X_1, X_2, \ldots, X_n$ del modelo de medición (2) donde los errores ε_i tienen distribución $N(\mu, \sigma^2)$, entonces $n^{1/2}(\bar{X}_n - \mu)$ tiene distribución $N(0, \sigma^2)$ independientemente de n. En cambio la distribución de $n^{1/2}(\hat{\mu}_{MED} - \mu)$ se aproxima a una $N(0, \pi\sigma^2/2)$ (ver por ejemplo Huber [11]). Por lo tanto

$$\frac{\text{varas}(\bar{X}_n)}{\text{varas}(\hat{\mu}_{MED})} = \frac{2}{\pi} = 0.636, \qquad (4)$$

donde varas indica la varianza asintótica, es decir la varianza de la aproximación normal. Como $\bar{X}_n$ es exactamente normal, varas($\bar{X}_n$) = var($\bar{X}_n$). La ecuación (4) muestra que la mediana no satisface el segundo requisito de un estimador robusto, es decir no es altamente eficiente bajo errores normales.

El próximo paso será buscar estimadores que satisfagan los dos requisitos de un estimador robusto. Para esto vamos a recordar caracterizaciones de $\bar{X}_n$ y $\hat{\mu}_{MED}$ que nos sugerirán nuevos estimadores.

Es bien conocido que $\bar{X}_n$ es el valor de μ que minimiza

$$\sum_{i=1}^{n}(X_i - \mu)^2$$

y que $\hat{\mu}_{MED}$ es el valor de $\hat{\mu}$ que minimiza

$$\sum_{i=1}^{n}|X_i - \mu|.$$

Ver por ejemplo Bickel [1].

Luego, los dos estimadores minimizan en distintas métricas las distancias del valor estimado a los elementos de la muestra. Las funciones que definen las distancias son

$$\rho_2(u) = u^2$$

en el caso de la media y

$$\rho_1(u) = |u|$$

en el caso de la mediana. La mayor robustez de $\hat{\mu}_{MED}$ se debe a que $\rho_1(u)$ crece más lentamente que $\rho_2(u)$ y por lo tanto le da menos peso a los "outliers". Esto sugiere definir una clase de estimadores más general ampliando la clase de funciones ρ.

En un trabajo fundacional de la teoría moderna de la robustez, Huber [9] define la clase de estimadores de tipo M (M-estimadores). Dada una función ρ, el M-estimador correspondiente es el valor μ que minimiza

$$\sum_{i=1}^{n}\rho(X_i - \mu). \qquad (5)$$

La función ρ debe satisfacer las siguientes propiedades:

R1. Es una función par, es decir $\rho(-u) = \rho(u)$.

R2. Si $|u_1| < |u_2|$ entonces $\rho(u_1) \leq \rho(u_2)$.

Huber [9] muestra que se puede obtener un estimador que herede simultáneamente las buenas propiedades de $\bar{X}_n$ y $\hat{\mu}_{MED}$ eligiendo una función ρ de modo que sea cuadrática como ρ_2 para valores pequeños de $|u|$ y lineal como ρ_1 para valores grandes de $|u|$. Más precisamente, Huber considera la familia de funciones

$$\rho_c^H(u) = \begin{cases} -2cu - c^2 & \text{si} \quad u < -c \\ u^2 & \text{si} \quad |u| \leq c \\ 2cu - c^2 & \text{si} \quad u > c. \end{cases} \qquad (6)$$

En los puntos $-c$ y c la función ρ_c^H cambia de cuadrática a lineal, y este cambio se hace

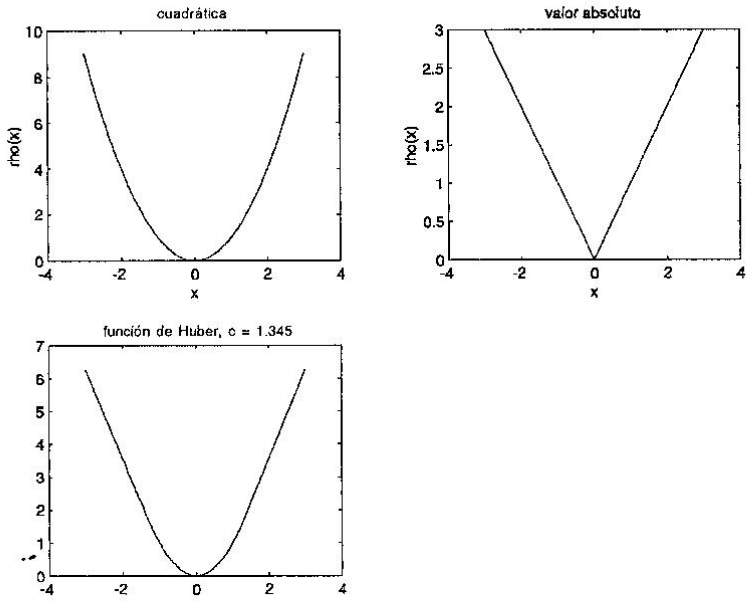

Figura 2. Funciones ρ

de manera que la función resulte derivable en esos puntos. Se puede observar que cuando $c \to \infty$ el M-estimador basado en ρ_c^H se acerca a la media y cuando $c \to 0$, se acerca a la mediana. En la Figura 2 se representan las funciones ρ correspondientes a la media muestral, la mediana y a la función $\rho_1^H(u)$.

Huber prueba que las funciones de la familia ρ_c^H tienen propiedades de optimalidad utilizando un enfoque minimax que describimos a continuación.

Sea $X_1, \ldots, X_n$ una muestra aleatoria del modelo (2) donde los ε_i tienen distribución F y sea $\hat{\mu}_n$ el M-estimador definido por la minimización de (5). Entonces Huber [9] prueba que bajo condiciones suficientemente generales

$$n^{1/2}(\hat{\mu}_n - \mu) \to_D N(0, V(\rho, F)),$$

donde $\to_D$ indica convergencia en distribución. Esto significa que para n grande la distribución de $n^{1/2}(\hat{\mu}_n - \mu)$ es aproximadamente N(0, $V(\rho, F)$). La varianza asintótica $V(\rho, F)$ viene dada por

$$V(\rho, F) = \frac{E_F(\psi^2(X))}{E_F^2(\psi'(X))},$$

donde $\psi = \rho'$.

Si se conoce F, el estimador que debería usarse es el M-estimador correspondiente a una función ρ que minimice $V(\rho, F)$. Este estimador es el estimador de máxima verosimilitud y la correspondiente ρ es la función

$$\rho_F(u) = -\log f(u)),$$

donde $f = F'$. Si F es normal

$$\rho_F(u) = \frac{u^2}{2\sigma^2} + \frac{\log(2\pi\sigma^2)}{2},$$

y esto es equivalente a elegir $\rho(u) = u^2$. Por lo tanto en este caso el estimador óptimo es la media muestral.

Sin embargo, la situación más común es que F no sea conocida exactamente. Una hipótesis razonable es que F pertenezca a ν_ε, el conjunto de distribuciones normales contaminadas. En este caso, si se usa el M-estimador definido por una función ρ, el peor comportamiento de su varianza asintótica estará dado por

$$V^*(\rho) = \max_{F \in V_\varepsilon} V(\rho, F),$$

donde max indica máximo. El criterio minimax consiste en elegir la función ρ^* que minimiza $V^*(\rho)$. Huber [9] mostró que la función minimax ρ^* pertenece a la familia $\rho_c^H(u)$ definida en (6), donde el valor c depende de ε. El valor de c decrece con ε, por lo tanto al aumentar la proporción de "outliers", el M-estimador minimax se acerca a la mediana.

Un problema para la elección de la constante c que define la función $\rho_c^H(u)$ es que ε en general no es conocido. Un criterio muy utilizado en la práctica es elegir c de manera que para muestras normales el estimador tenga una eficiencia, en términos de varianza asintótica, igual a 0.95. De esta manera se cumple la segunda propiedad que le pedimos al estimador robusto: que fuera altamente eficiente para muestras normales. Este valor es $c = 1.345$.

Ejemplo 1. En la Tabla I presentamos los valores de 24 determinaciones del contenido de cobre en harina integral (en partes por millón) ordenados de menor a mayor.

A primera vista el valor 28.95 se destaca de los restantes y puede ser considerado un "outlier". El valor del promedio es $\bar{X} = 4.28$. Si se elimina este "outlier" se obtiene $\bar{X} = 3.21$. En cambio $\hat{\mu}_{MED} = 3.385$ y si se quita el "outlier" resulta $\hat{\mu}_{MED} = 3.37$. Usando el M-estimador correspondiente a $\rho_{1.345}^H$ se obtiene $\hat{\mu}_M = 3.21$ cuando se usan todas las observaciones y $\hat{\mu}_M = 3.18$ cuando se quita el "outlier". Por lo tanto, como era de esperar, $\bar{X}$ es mucho más sensible que $\hat{\mu}_{MED}$ y $\hat{\mu}_M$ a la presencia del "outlier".

Tabla I.– Contenido en cobre

2.20	2.20	2.40	2.80	2.50	2.70	2.80	2.90
3.00	3.03	3.10	3.37	3.40	3.40	3.40	3.50
3.60	3.70	3.70	3.70	3.70	3.77	5.28	28.95

3. Punto de ruptura: una medida de robustez

Una medida de la robustez de un estimador introducida por Hampel [5] es el *punto de ruptura*. Heurísticamente, el punto de ruptura es la mínima proporción de "outliers" que puede hacer que el estimador tome valores arbitrariamente grandes (positivos o negativos). El concepto introducido por Hampel es de carácter asintótico. Aquí nos referiremos a una versión del punto de ruptura introducida por Donoho y Huber [3] para muestras finitas.

Consideremos un estimador $\hat{\mu}$ y supongamos que se tiene un conjunto de n observaciones $x_1, x_2, \ldots, x_n$. Para cada k entero positivo, llamemos A_k al máximo valor absoluto que el estimador puede tomar cuando se reemplazan k de las observaciones dadas por valores arbitrarios. Sea ahora k^* el mínimo valor k que hace $A_k = \infty$. El punto de ruptura del estimador $\hat{\mu}$ (que simbolizaremos por PR), es la proporción del número total de observaciones que k^* representa. Es decir

$$PR(\hat{\mu}) = \frac{k^*}{n}.$$

Informalmente podemos decir que el punto de ruptura de un estimador es la mínima proporción de "outliers" que puede hacer el estimador igual a más o menos infinito.

En general, se tiene

$$\frac{1}{n} \le PR(\hat{\mu}) \le 0.50.$$

En particular para el promedio y la mediana se obtiene

$$PR(\bar{X}) = \frac{1}{n}, \quad PR(\hat{\mu}_{MED}) = 0.50$$

El punto de ruptura de los M-estimadores definidos por la minimización de (5) depende de $\psi = \rho'$. Si ψ es acotada entonces $PR = 0.50$ para cualquier conjunto de datos. Si ψ no es acotada, $PR = 1/n$. Esto significa que los M-estimadores son robustos si y sólo si ψ es acotada.

La función ρ_c^H definida en (6) tiene como derivada

$$\psi_c^H(u) = \begin{cases} -2c & \text{si} \quad u < -c \\ 2u & \text{si} \quad |u| < c \\ 2c & \text{si} \quad u > c. \end{cases}$$

Esto implica que $\max|\psi_c^H(u)| = 2c$ y luego $\psi_c^H(u)$ resulta acotada. Por lo tanto los correspondientes M-estimadores tienen el máximo posible punto de ruptura 0.50.

4. Modelo de Regresión Lineal

Uno de los problemas comunes a todas las ramas de la Ciencia y de la Tecnología es explicar el comportamiento de una variable en función de otras. Llamemos a la variable a explicar Y (variable dependiente) y llamemos X_1, ..., X_k a las variables utilizadas para explicar Y (variables independientes). Un modelo de regresión establece que

$$Y = g(X_1, ..., X_k) + \varepsilon,$$

donde ε es un error que puede reflejar

1. error de medición de Y
2. la existencia de otras variables que afectan a Y y que no estamos considerando por desconocimiento de cuáles son o porque no hay registros de las mismas.

Generalmente se supone que $E(\varepsilon) = 0$ (E indica esperanza o valor medio). El modelo más simple de regresión es el lineal en el cual

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_x X_k + \varepsilon.$$

El coeficiente β_i representa el incremento medio de la variable Y cuando la variable X_i aumenta en una unidad, y β_0 el valor medio de la variable Y cuando todas las variables son 0.

Para poder estimar el vector de coeficientes $\beta = (\beta_0, \beta_1, ..., \beta_k)$ se requiere haber realizado varias observaciones del fenómeno. Supongamos que hemos realizado n observaciones

$$\begin{aligned} \mathbf{Z}_1 &= (Y_1, X_{11}, \cdots, X_{1k}) \\ \mathbf{Z}_2 &= (Y_2, X_{21}, \cdots, X_{2k}) \\ &\cdots\cdots\cdots\cdots \\ &\cdots\cdots\cdots\cdots \\ \mathbf{Z}_n &= (Y_n, X_{n1}, \cdots, X_{nk}) \end{aligned}$$

Si los errores son independientes, normales y tienen la misma varianza, entonces el estimador óptimo, propuesto por primera vez por Gauss, es el estimador de mínimos cuadrados (EMC), que se define de la manera siguiente. Dado un posible estimador del vector β, $\mathbf{b} = (b_0, b_1, ..., b_k)$, se determinan los errores que corresponderían a cada una de las observaciones

$$\begin{aligned} e_1(\mathbf{b}) &= Y_1 - b_0 - b_1 X_{11} - ... - b_k X_{1k} \\ e_2(\mathbf{b}) &= Y_2 - b_0 - b_1 X_{21} - ... - b_k X_{2k} \\ ..... &\quad ... \quad ..... \\ ..... &\quad ... \quad ..... \\ e_n(\mathbf{b}) &= Y_n - b_0 - b_1 X_{n1} - ... - b_k X_{nk}. \end{aligned}$$

Entonces el estimador de *mínimos cuadrados* es el vector $\mathbf{b}$ que minimiza

$$C_2(\mathbf{b}) = \sum_{i=1}^{n} e_i^2(\mathbf{b}).$$

Nuevamente este procedimiento es muy sensible a la presencia de "outliers". Un "outlier" en este caso es un punto (Y^*, X_1^*, ..., X_k^*) para el cual $|Y^* - \beta_0 - \beta_1 X_1^* - ... - \beta_x X_k^*|$ es grande en referencia a los otros errores. Nuevamente este procedimiento es muy sensible a la presencia de "outliers". Uno solo puede provocar cambios ilimitados en los coeficientes estimados por mínimos cuadrados. Por lo tanto su punto de ruptura es $1/n$.

Con el objetivo de obtener estimadores robustos, Huber [10] propuso extender los M-estimadores para el modelo de regresión. Estos se definen como el vector $\mathbf{b}$ que minimiza

$$C_\rho(\mathbf{b}) = \sum_{i=1}^{n} \rho(e_i(\mathbf{b})),$$

donde $\psi = \rho'$ es acotada. Una posible elección sería tomar $\rho(u) = |u|$ o las funciones $\rho_c^H(u)$ definidas en (6). En el primer caso se tiene el estimador de *mínimos valores absolutos* (EMVA) y corresponde a minimizar

$$C_1(\mathbf{b}) = \sum_{i=1}^{n} |e_i(\mathbf{b})|.$$

A pesar de que los M-estimadores son menos sensibles a "outliers" que el EMC, no son totalmente robustos. Su punto de ruptura continúa siendo $1/n$. Es decir, un solo "outlier" puede provocar que un M-estimador tome valores arbitrariamente grandes.

Para simplificar el estudio de las propiedades de robustez de los M-estimadores vamos a considerar el modelo con una sola variable. Consideremos n observaciones (Y_1, X_1), (Y_2, X_2), ..., (Y_n, X_n), y supongamos que se agrega una observación (Y^*, X^*) que es un "outlier". Se debe distinguir dos casos

1. X^* es un "outlier" entre X_1, ..., X_n, es decir está alejado del centro de estos datos. En este caso (Y^*, X^*) puede tener una influencia no acotada en el M-estimador.

2. X^* no es un "outlier" entre X_1, ..., X_n. En este caso (Y^*, X^*) tiene poca influencia en el M-estimador.

Los M-estimadores son robustos solamente respecto a "outliers" de tipo 2.

El primer método de estimación de los coeficientes de regresión con punto de ruptura 0.50 con respecto a todo tipo de "outliers" es el estimador de *mínima mediana de cuadrados (EMMC)*. Este estimador fue propuesto por Hampel [6] y luego desarrollado por Rousseeuw [12] quien elaboró un algoritmo de cálculo. El EMMC se define como el vector $\mathbf{b}$ que minimiza

$$C_{MED}(\mathbf{b}) = \text{MEDIANA } (e_1^2(\mathbf{b}), \ldots, e_n^2(\mathbf{b})).$$

Como contrapartida a su alta robustez, la eficiencia del EMMC es extremadamente pobre para muestras normales. Mientras el EMC y los M-estimadores se aproximan a los valores verdaderos de los coeficientes de regresión con orden $n^{-1/2}$ (es decir la diferencia entre valor estimado y valor verdadero es de orden $n^{-1/2}$) el EMMC se aproxima con orden $n^{-1/3}$. Otra dificultad para hacer inferencia estadística con el EMMC es que su distribución asintótica no es normal. Esto hace difícil realizar tests e intervalos de confianza basados en este estimador.

El próximo paso será considerar estimadores robustos con punto de ruptura 0.50 pero con mayor eficiencia que el EMMC y con distribución aproximadamente normal.

Dada una muestra de residuos

$$\mathbf{u} = (u_1, \ldots, u_n),$$

podemos definir diferentes medidas que indiquen cuan lejos de 0 se encuentran. Por ejemplo, tenemos las siguientes medidas

$$S_2(u_1, \cdots u_n) = \left(\frac{1}{n} \sum_{i=1}^{n} u_i^2 \right)^{1/2},$$

$$S_1(u_1, \cdots u_n) = \frac{1}{n} \sum_{i=1}^{n} |u_i|$$

y

$$S_{MED}(u_1, \cdots, u_n) = \text{MEDIANA } (|u_1|, |u_2|, \cdots, |u_n|).$$

Obsérvese que el EMC minimiza $S_2(e_1(\mathbf{b}), \ldots, e_n(\mathbf{b}))$, el EMVA $S_1(e_1(\mathbf{b}), \ldots, e_n(\mathbf{b}))$ y el EMMC $S_{MED}(e_1(\mathbf{b}), \ldots, e_n(\mathbf{b}))$.

En general, una función S es llamada un estimador de escala si tiene las siguientes propiedades

1. $S(u_1, \ldots, u_n) \geq 0$.
2. $S(\lambda u_1, \ldots, \lambda u_n) = |\lambda| S(u_1, \ldots, u_n)$.
3. Es invariante por permutaciones del orden de la muestra.
4. Si $|u_i^*| \geq |u_i|$ para todo i, entonces $S(u_1^*, \ldots, u_n^*) \geq S(u_1, \ldots, u_n)$.

Es fácil probar que S_1, S_2 y S_{MED} verifican estas propiedades y por lo tanto son estimadores de escala.

Un M-estimador de escala $S(u_1, \ldots, u_n)$ se define como el valor s que satisface

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

donde ρ satisface las propiedades R1-R2 establecidas anteriormente y

R3 $\rho(0) = 0$

Rousseeuw y Yohai [13] introducen los S-estimadores para el modelo de regresión lineal como el valor **b** que minimiza

$$S(e_1(\mathbf{b}), \ldots, e_n(\mathbf{b})),$$

donde S es un M-estimador de escala.

Rousseeuw y Yohai mostraron que si ρ satisface R1-R3 y además es acotada, entonces el punto de ruptura del correspondiente S-estimador de regresión es

$$PR = \min\left(\frac{a}{c}, 1-\frac{a}{c}\right),$$

donde

$$a = \max \rho.$$

Luego si $c = a/2$, el punto de ruptura de un S-estimador es 0.50.

En cuanto a su eficiencia, Rousseeuw y Yohai demostraron que si ρ es dos veces derivable, los S-estimadores se aproximan al valor verdadero con orden $n^{-1/2}$. Por lo tanto son más eficientes que el EMMC. Además la distribución de estos estimadores se aproxima a la normal cuando el tamaño de la muestra aumenta.

Hossjer [8] resolvió el problema de encontrar el S-estimador con $PR = 0.50$ con mínima varianza asintótica bajo errores normales. El resultado es que si llamamos a este estimador $\hat{\beta}^{SO} = \left(\hat{\beta}_1^{SO}, \hat{\beta}_2^{SO}, \cdots, \hat{\beta}_k^{SO}\right)$ y $\hat{\beta}^{MC} = \left(\hat{\beta}_1^{MC}, \hat{\beta}_2^{MC}, \cdots, \hat{\beta}_k^{MC}\right)$ al estimador de mínimos cuadrados

$$\frac{\text{varas}\left(\hat{\beta}_i^{MC}\right)}{\text{varas}\left(\hat{\beta}_i^{SO}\right)} = 0.329, \quad 1 \le i \le k.$$

Por lo tanto, aunque los S-estimadores son mucho más eficientes que el EMMC, su eficiencia es baja relativa al EMC.

Yohai y Zamar [16] definen un nuevo tipo de escalas, llamadas escalas de tipo τ, que permiten encontrar estimadores que simultáneamente tienen alto punto de ruptura y son eficientes. Estas escalas se basan en 2 funciones ρ_1 y ρ_2 que tienen las propiedades R1-R3 y que además son acotadas.

Usando ρ_1 se calcula un M-estimador de escala inicial S_M

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1\left(\frac{u_i}{S_M}\right) = c.$$

El τ-estimador de escala S_τ se define de la siguiente manera

$$S_\tau^2(u_1, \cdots, u_n) = S_M^2(u_1, \cdots, u_n)\frac{1}{n}\sum_{i=1}^{n}\rho_2\left(\frac{u_i}{S_M}\right),$$

y los τ-estimadores de regresión se definen minimizando

$$S_\tau(e_1(\mathbf{b}), \ldots, e_n(\mathbf{b})).$$

Obsérvese que si $\rho_2(u) = u^2$ entonces

$$S_\tau(u_1, \cdots, u_n) = \frac{1}{n}\sum_{i=1}^{n}u_i^2$$

independientemente de ρ_1. Es decir, se obtiene la escala que se usa para mínimos cuadrados. Luego, para obtener un estimador que sea altamente eficiente bajo normalidad bastará elegir como ρ_2 una función parecida a u^2 pero acotada.

Yohai y Zamar [16] mostraron que se pueden elegir ρ_1 y ρ_2 de manera que el correspondiente τ-estimador de regresión tenga punto de ruptura 0.50 y eficiencia tan alta como se quiera cuando los errores son normales. Por ejemplo, si llamamos $\left(\hat{\beta}_1^{\tau}, \cdots, \hat{\beta}_k^{\tau}\right)$ a los τ-estimadores, se puede conseguir que

$$\frac{\text{varas}\left(\hat{\beta}_i^{MC}\right)}{\text{varas}\left(\hat{\beta}_i^{\tau}\right)} = 0.95, \quad 1 \le i \le k.$$

Finalmente Gervini y Yohai [2002] obtienen estimadores que tienen simultá-

neamente las siguientes propiedades: (i) punto de ruptura igual 0.50 y (ii) eficiencia asintótica igual a la del EMC. Más precisamente, si llamamos a estos estimadores $\left(\hat{\beta}_1^{EF}, \cdots, \hat{\beta}_k^{EF}\right)$, se tendrá

$$\frac{\text{varas}\left(\hat{\beta}_i^{LS}\right)}{\text{varas}\left(\hat{\beta}_i^{EF}\right)} = 1.00.$$

Estos estimadores son obtenidos en los siguientes pasos:

1. Se encuentra un S-estimador con punto de ruptura 0.50.

2. Se obtienen los residuos a partir del S-estimador obtenido en el paso 1.

3. Se compara la distribución de los residuos con la distribución normal, y se elimina una cierta cantidad de observaciones correspondientes a los residuos más grandes en valor absoluto. Esto se hace de manera que la cola de la distribución de los residuos restantes se "parezca" a la correspondiente a la normal.

4. Con las observaciones restantes se calcula el estimador de mínimos cuadrados.

Ejemplo 2. El primer ejemplo de regresión tiene como variable dependiente Y el número de llamadas internacionales realizadas desde Bélgica anualmente y como variable independiente X el año. Los datos, que fueron obtenidos de Rousseeuw y Leroy [14], corresponden al período comprendido entre los años 1950 y 1973. En la Figura 3 graficamos esos datos, pudiéndose observar que aquellos correspondientes a los años 1964-1969 pueden considerarse "outliers". En la Figura 3 también se grafican las rectas correspondientes al EMC y a un S-estimador (SE) con punto de ruptura 0.50. Se observa que el EMC es afectado sensiblemente por "outliers", pero no así el S-estimador. Los datos de este ejemplo fueron obtenidos del "Belgian Statistical Survey". En un Boletín posterior se reconoció que los datos de los años 1964-1969 correspondían a gasto y no a número de llamadas.

Ejemplo 3. Este ejemplo está basado en datos astronómicos. Los datos, obtenidos de Rousseeuw y Leroy [14], corresponden al diagrama de Hertzprung-Russell de 47 estrellas del conglomerado CYG. La variable dependiente Y es el logaritmo de la intensidad de la luz y la variable independiente X el logaritmo de la temperatura. En la Figura 4 están graficados los datos y las rectas correspondientes al EMC y un S-estimador (SE) con punto de ruptura 0.50. Se pueden observar cuatro claros "outliers" con temperaturas extremadamente bajas. Estos "outliers" provocan que la recta correspon-

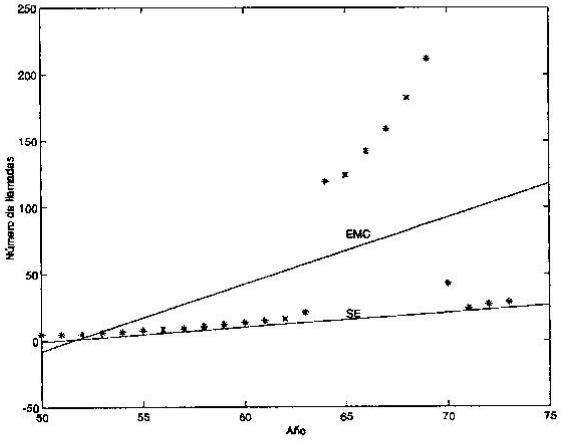

Figura 3. Datos de llamadas internacionales

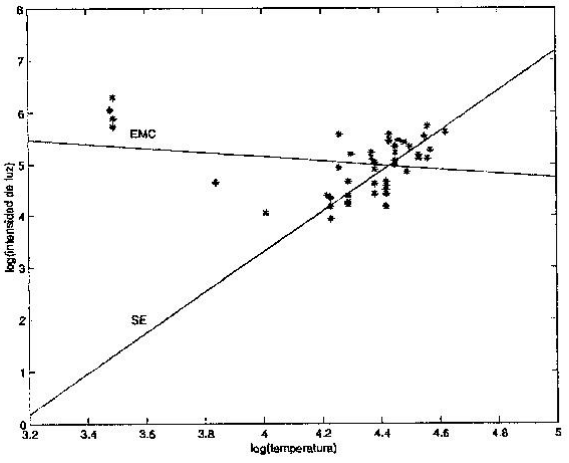

Figura 4. Diagrama de Hertzprung-Rusell

diente al EMC esté alejada de la mayoría de los datos. En cambio, la recta correspondiente al S-estimador es poco influida por los "outliers" y representa adecuadamente la relación lineal entre ambas variables para la mayoría de los puntos.

5. Métodos robustos en otros modelos

También se han desarrollado métodos robustos para otros modelos estadísticos además de los considerados en este trabajo. Podemos mencionar entre otras áreas donde se aplican métodos robustos: Análisis Multivariado, Series de Tiempo, Modelos Lineales Generalizados, Modelos Econométricos, Datos Direccionales, Regresión no Lineal, etc.

Los aspectos teóricos de los métodos robustos pueden consultarse en los libros de Huber [11] y de Hampel et al. [7]. Un enfoque más aplicado puede encontrarse en el libro de Rousseeuw y Leroy [1987].

Referencias

[1] Bickel, P.J. y Docksum, K.A. *Mathematical Statistics: basic ideas and selected topics,* Prentice Hall, Upper Saddle River, (2000).

[2] Billingsley, P. *Probability and Measure.* 3rd Edition, Wiley, New York, (1995).

[3] Donoho, D.L. y Huber, P.J. The notion of breakdown point, in *A Festschrift for Erich Lehmann,* P.J. Bickel, K. Doksum and J.L. Hodges, Jr., eds., Wadsworth, Belmont, 157, (1983).

[4] Gervini, D. y Yohai, V.J. A class of fully efficient regression estimates. Aparecerá en el *Ann. Statist.,* (2002).

[5] Hampel, F.R. A general definition of qualitative robustness. *Ann. Math. Statist.,* **42**, 1887, (1971).

[6] Hampel, F.R. Beyond location parameters: robustness concepts and methods. *Bulletin of the International Statistical Institute,* **46**, 375, (1975).

[7] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. y Stahel, W.A. *Robust Statistics. The Approach Based on Influence Functions,* Wiley, New York, (1986).

[8] Hossjer, O. On the optimality of S-estimators, *Statist. and Probability Letters,* **12**, 413, (1992).

[9] Huber, P.J. Robust estimation of a location parameter, *Ann. Math. Statist.,* **35**, 73, (1964).

[10] Huber, P.J. Robust regression: Asymptotics, conjectures and Monte Carlo, *Ann. Statist.,* **1**, 799, (1973).

[11] Huber, P.J. *Robust Statistics,* Wiley, New York, (1981).

[12] Rousseeuw, P.J. Least median of squares regression, *J. Amer. Statist. Assoc,* **79**, 871, (1984).

[13] Rousseeuw P.J. y Yohai, V.J. Robust regression by means of S-estimators, in *Robust and Nonlinear Time Series Analysis,* J. Franke, W. Hardle, and R.D. Martin (eds.), Lecture Notes in Statistics, **26**, Springer, New York, 256, (1984).

[14] Rousseeuw, P.J. y Leroy, A.M. *Robust regression and outlier detection,* Wiley, New York, (1987).

[15] Yohai, V.J. High breakdown point and high efficiency robust estimates for regression, *Ann. Statist,* **15**, 642, (1987).

[16] Yohai, V.J. y Zamar, R.H. High breakdown-point estimates of regression by means of the minimization of an efficient scale, *J. Amer. Statist. Assoc.,* **83**, 406, (1988).

Manuscrito recibido y aceptado en abril de 2002.