

Los valores atípicos en econometría

1. El modelo lineal general

Aránzazu de Juan Fernández

Los valores atípicos en econometría

1. El modelo lineal general

Aránzazu de Juan Fernández

EDICIONES DE LA UNIVERSIDAD AUTÓNOMA
DE MADRID

Teléfono 397 42 33
28049 Madrid



© Ediciones de la Universidad Autónoma de Madrid.

Reservados todos los derechos. Está prohibido, bajo las sanciones penales y resarcimiento civil previsto en las leyes, reproducir, registrar o transmitir esta publicación, íntegra o parcialmente (salvo, en este último caso, para su cita expresa en un texto diferente, mencionando su procedencia), por cualquier sistema de recuperación y por cualquier medio, sea mecánico, electrónico, magnético, electroóptico, por fotocopia o cualquier otro, sin la autorización previa por escrito de Ediciones de la Universidad Autónoma de Madrid.

I.S.B.N.: 84-7477-628-7

Depósito Legal: B-42.163-1997

Impresión: Cardellach Còpies, SL (impresión digital)
Sant Pere, 40 - 08221 Terrassa (Barcelona)
Tel. (93) 786 47 09 - Fax (93) 786 48 09

ÍNDICE

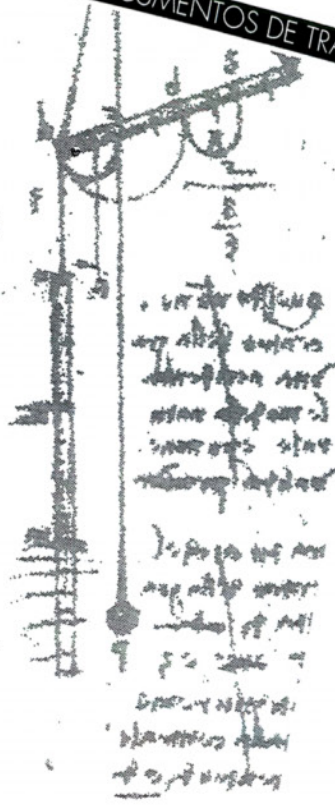
1.- Introducción	7
2.- Definiciones de valor atípico	11
2.1.- Definiciones	11
2.2.- Consideraciones sobre estas definiciones	12
2.3.- Medidas de influencia de los valores atípicos	13
2.4.- Curvas de influencia	15
2.5.- Clases de atípicos en el Modelo Lineal General	19
3.- Efectos de los valores atípicos sobre los estadísticos	23
4.- Detección de los valores atípicos	47
4.1.- Métodos basados en los residuos	48
4.2.- Métodos basados en la eliminación de una ó múltiples filas	51
4.2.1.- Eliminación de una única fila en X y en Y	52
4.2.2.- Eliminación de múltiples filas	58
4.3.- Métodos basados en la distancia de Mahalanobis	61
4.4.- Métodos Gráficos	70
5.- Modelización y corrección de valores atípicos	75
5.1.- Estimación robusta de los parámetros del modelo	76
5.1.1.- Estimador de desviaciones absolutas mínimas	78
5.1.2.- Estimador M de Huber	79
5.1.3.- Estimadores M generalizados ó estimadores GM	80
5.1.4.- Estimadores MM - Yohai (1980)	81
5.1.5.- Estimador de mínimos cuadrados ponderados	83
5.1.6.- Estimador LMS (Rousseeuw (1985))	84
5.1.7.- Estimador LTS (Rousseeuw (1983))	86
5.1.8.- Estimador S (Rousseeuw y Yohai (1984))	86
5.1.9.- Estimador de mediana repetida (Siegel (1982))	87
5.2.- Método Clásico - Variables ficticias	88
6.- Ejemplo de los distintos atípicos en el modelo de regresión	93
6.1.- Efectos de los valores atípicos sobre los estimadores y estadísticos	94
6.1.1.- Efecto de los valores atípicos en la variable endógena sobre los estimadores y estadísticos	95
6.1.2.- Efecto de los valores atípicos en la variable explicativa sobre los estimadores y los estadísticos	97
6.2.- Detección de los valores atípicos	98
6.2.1.- Detección de los valores atípicos sobre la variable endógena	99
6.2.2.- Detección de los valores atípicos sobre las variables explicativas	115
6.3.- Tratamiento de los valores atípicos	128
6.3.1.- Tratamiento de los valores atípicos en la variable endógena	129
6.3.2.- Tratamiento de los valores atípicos en la variable explicativa	131
6.4.- Conclusiones de los ejemplos	135
7.- Conclusiones y extensiones	141
Bibliografía	147

Los tres ejes de la estructura son
 ortogonales entre sí y se sitúan en
 los ejes de simetría de las vigas.
 Para el análisis se considerará la viga
 sometida a una carga distribuida
 que actúa perpendicularmente a
 uno de los ejes de simetría.
 Se considerará que la viga es
 homogénea e isotrópica y que
 se encuentra en un estado de
 equilibrio estático.
 Las hipótesis de la teoría de vigas
 se aplicarán a esta estructura.

Se considerará que la viga es
 homogénea e isotrópica y que
 se encuentra en un estado de
 equilibrio estático.
 Las hipótesis de la teoría de vigas
 se aplicarán a esta estructura.

$$I_x = \frac{b h^3}{12}$$

$$I_y = \frac{h b^3}{12}$$



Se considerará que la viga es
 homogénea e isotrópica y que
 se encuentra en un estado de
 equilibrio estático.

1.- INTRODUCCIÓN

Se considerará que la viga es
 homogénea e isotrópica y que
 se encuentra en un estado de
 equilibrio estático.
 Las hipótesis de la teoría de vigas
 se aplicarán a esta estructura.

1.- INTRODUCCIÓN

Los numerosos trabajos publicados sobre los valores atípicos que se presentan en los datos de las variables económicas ponen de manifiesto la importancia de éstos en el campo estadístico. De hecho, la existencia de valores atípicos no corregidos en las variables que se analizan en el campo econométrico puede generar efectos bastante negativos, llevando, por ejemplo, a identificar modelos que podríamos llamar "espurios" además de obtener estimadores sesgados que no reflejan los valores reales de los coeficientes, y en muchos casos dejando en entredicho el significado económico de las relaciones objeto del estudio.

En la extensa bibliografía que se ha desarrollado sobre los valores atípicos, en el campo econométrico, se aprecian dos tendencias según el tipo de relación que se especifique. Por un lado, la consideración de estas observaciones como una parte importante del análisis del modelo de regresión lineal (MLR) y por otro, el análisis de intervención de los atípicos en el campo de las series temporales, tendencias ambas que se diferencian en la consideración de una relación generalmente estática en el modelo de regresión mientras que la posibilidad de que un valor atípico expanda su efecto a lo largo de varios períodos es la que se considera en el enfoque de series temporales.

En cada una de estas dos tendencias, se han desarrollado métodos de detección de estos valores atípicos, métodos de corrección así como medidas de la influencia que estas observaciones puedan tener sobre el análisis, siendo también importante el desarrollo teórico sobre los efectos que estos valores pueden tener en cada uno de los casos. Se puede señalar que ha sido tan importante el desarrollo teórico en torno a estas observaciones que los programas de software poco a poco han ido incorporando estos métodos tanto de detección como de corrección de los valores atípicos.

A nuestro juicio, la detección de un supuesto valor atípico debe ir acompañada de un ejercicio de búsqueda de las causas que provocan la aparición de éste. Debido a que principalmente el objeto del análisis econométrico se centra en las variables de tipo económico, pensamos que no debe ser difícil hallar las causas de estos valores. De esta forma, se puede evitar aceptar un modelo como bueno en el que la modelización de múltiples valores atípicos, no captados por la estructura del modelo, sin causa

aparente. Es decir, la justificación económica de la aparición de un valor atípico, ya sea por intervención de la autoridad económica, ya sea por la ocurrencia de un hecho que modifica la estructura de la variable, evitará la introducción o modelización de valores atípicos sin causa justificada, pudiendo explicarse estas observaciones por otro tipo de variables económicas que realmente sí que se encuentran relacionadas con la realidad económica. Por este motivo, creemos que a las definiciones dadas sobre valores atípicos habría que añadirles alguna consideración sobre la justificación de su aparición.

En otro orden de cosas, las soluciones que se han propuesto a este problema son muy distintas, yendo desde la eliminación de estos valores hasta su modelización mediante la introducción de variables ficticias para captar este comportamiento anómalo. Sin embargo, en este campo, también las soluciones son distintas dependiendo del tipo de análisis que se realice.

En este trabajo, haremos un breve resumen de las tendencias sobre valores atípicos que se han desarrollado, centrándonos principalmente en los avances más recientes. Pero hasta llegar a este punto pasaremos por una sección en la que se definirán los valores atípicos, viéndose en una tercera sección los efectos que estas observaciones plantean sobre los principales estadísticos en el modelo de regresión lineal. En la sección 4 se resumirán los principales métodos de detección de estas observaciones. La sección 5 se dedicará a la corrección de los valores atípicos ilustrando tanto los métodos de detección como las consecuencias de la falta de tratamiento en la sección 6 mediante un modelo de regresión sencillo. Finalmente en la sección 7 se presentarán las principales conclusiones así como las líneas de investigación futuras en este campo.

Los tres tipos de
 estructura de trabajo
 en la economía de
 los países en desarrollo
 son: la estructura de
 los países en desarrollo
 que se caracteriza por
 la presencia de una
 gran masa de
 mano de obra
 no calificada,
 la estructura de
 los países en desarrollo
 que se caracteriza por
 la presencia de una
 gran masa de
 mano de obra
 no calificada,
 la estructura de
 los países en desarrollo
 que se caracteriza por
 la presencia de una
 gran masa de
 mano de obra
 no calificada.

La estructura de
 los países en desarrollo
 que se caracteriza por
 la presencia de una
 gran masa de
 mano de obra
 no calificada,
 la estructura de
 los países en desarrollo
 que se caracteriza por
 la presencia de una
 gran masa de
 mano de obra
 no calificada,
 la estructura de
 los países en desarrollo
 que se caracteriza por
 la presencia de una
 gran masa de
 mano de obra
 no calificada.

3 6 8 12

2.- DEFINICIONES DE VALOR ATÍPICO

El valor atípico es el valor
 que se encuentra en la
 estructura de los países
 en desarrollo que se
 caracteriza por la
 presencia de una gran
 masa de mano de obra
 no calificada. El valor
 atípico es el valor que
 se encuentra en la
 estructura de los países
 en desarrollo que se
 caracteriza por la
 presencia de una gran
 masa de mano de obra
 no calificada. El valor
 atípico es el valor que
 se encuentra en la
 estructura de los países
 en desarrollo que se
 caracteriza por la
 presencia de una gran
 masa de mano de obra
 no calificada.

2.- DEFINICIONES DE VALOR ATÍPICO

2.1. DEFINICIONES

En los numerosos artículos y libros escritos acerca del problema de los valores atípicos se han dado diferentes definiciones de valores atípicos, si bien todas ellas se pueden englobar en la siguiente: Un valor atípico es aquella observación que se distancia claramente del resto de las observaciones, es decir, sería una observación que es inconsistente con el resto de los datos, (Barnett y Lewis (1984)).

Como ya se dijo, ésta podría considerarse como una definición global. Cabe señalar que se han propuesto definiciones atendiendo a la función de probabilidad que siguen los datos, siendo la más común la distribución normal. En este caso, entonces, se consideraría una observación como atípica, asumiendo que los datos siguen una $N(0,1)$, cuando el valor de dicha observación supere a las bandas $\pm 1.96\sigma$, siendo estos límites los marcados por una distribución normal de ∞ observaciones, con un 95% de probabilidad. De esta forma, para una muestra de tamaño infinito con una $N(0,1)$, habrá un 95% de probabilidad de que los valores se encuentren entre esos dos límites, siendo un atípico todo valor que se encuentre en el 5% restante. Esta es la definición de valor atípico que más ampliamente se utiliza en el análisis econométrico, si bien no debe considerarse como única, siendo imprescindible la verificación de la hipótesis de normalidad en estas observaciones para poder utilizar esta definición de valor atípico.

Por otro lado, y tal y como ponen de manifiesto Barnett y Lewis (1984), se debe tener una especial precaución en distinguir entre observaciones extremas, atípicos y contaminantes. Las **observaciones extremas** son aquellas que aparecen en primer y último lugar cuando se ha ordenado la muestra de menor a mayor o viceversa. Cualquiera de estas dos observaciones puede ser un **atípico** atendiendo a la distribución de probabilidad que siga la muestra. Las **observaciones contaminantes** son aquellas que, aún apareciendo en la muestra, pueden venir generadas por una distribución de probabilidad distinta. De esta forma, los atípicos pueden o no ser contaminantes y los contaminantes pueden o no ser atípicos.

Finalmente y a modo de resumen incluimos las definiciones de atípicos más comúnmente utilizadas en las investigaciones.

*** Definición de Ferguson**

En una muestra de tamaño moderado tomada de una población aparecen uno o varios valores que sorprendentemente se encuentran lejos del grupo principal de datos. Ferguson considera que estos valores atípicos deben ser rechazados porque existe una probabilidad muy pequeña de que las observaciones atípicas ocurran y además porque se obtienen estimaciones mejores con una muestra en la que se han eliminado dichos valores que en muestras completas que tienen en cuenta los atípicos.

*** Definición de Maddala**

Un valor atípico no es más que una observación que se comporta de forma diferente al resto de las observaciones. Según Maddala, en caso de que existan valores atípicos, lo usual es omitirlos y volver a estimar la regresión presentando tanto los resultados con los atípicos como los resultados sin ellos, analizando las posibles diferencias.

2.2 - CONSIDERACIONES SOBRE ESTAS DEFINICIONES

La principal consideración que cabe hacer sobre esta definición es que nunca se hace alusión a la existencia de algún motivo que cause la aparición de este valor atípico. Todas las definiciones se establecen sobre bases teóricas, desde el punto de vista estadístico, sin hacer alusión a que la existencia de los valores atípicos en variables de tipo económico debe venir ocasionada por algún hecho que provoca la aparición de una observación que en valor absoluto supera claramente al resto.

Estas consideraciones se suelen introducir en lo que los autores denominan orígenes de los valores atípicos, atribuyendo su aparición a errores de medida (principalmente errores en la recolección de los datos o en la introducción de los datos), a errores en la ejecución, refiriéndose en este caso Barnett y Lewis a la introducción entre la muestra de observaciones que no pertenecen a dicha muestra y a la variabilidad inherente a la propia variable (variación natural de una muestra).

A nuestro entender, conocer la causa de la aparición de una observación atípica en una muestra es fundamental, siendo el primer ejercicio que debe realizarse al ser detectada. De esta forma, evitaremos un problema importante en el análisis econométrico: la utilización de métodos de corrección de tales valores de forma indiscriminada, es decir, puede que el modelo estimado no sea el adecuado y la forma de solucionar este problema desde luego no es corregir valores atípicos mediante cualquier método. Sin embargo, conociendo la causa que provoca la aparición de estas observaciones, se sabrá con seguridad que la corrección de los atípicos no se hace de forma indiscriminada y que éstos no aparecen como consecuencia de una mala especificación del modelo.

Asimismo, las consecuencias de la no detección de los valores atípicos pueden provocar también importantes distorsiones en los modelos. De hecho, se pueden producir situaciones en las que si no se corrigen estas observaciones, se intente explicar una variable mediante una serie de sofisticados métodos que carecen totalmente de explicación, sobre todo en el campo económico, mientras que al corregirlos, los modelos econométricos suelen presentar sencillos esquemas, con una interpretación económica fácil.

A partir de estas consideraciones, se puede decir que los valores atípicos pueden provocar serias alteraciones de los modelos pero que estas dependen de la influencia que éstos tengan sobre las estimaciones y los modelos considerados.

2.3 - MEDIDAS DE INFLUENCIA DE LOS VALORES ATÍPICOS

Resulta ampliamente reconocido en las investigaciones que 1) algunos atípicos tienen una influencia muy pequeña sobre los parámetros estimados del modelo ajustado a una variable y 2) las observaciones que se han detectado como atípicas mediante contrastes normales tipo estándar pueden tener una influencia muy fuerte en la estimación de los parámetros (Peña (1987)). Asimismo, Peña (1987) señala que los diferentes estimadores en un modelo, no tienen porqué verse afectados de la misma manera que los otros por lo que propone diferentes medidas de influencia de los valores atípicos que recogeremos aquí.

Antes de exponer dichas medidas, hay que señalar que Peña deriva un método para detectar la influencia de los valores atípicos válido para todo tipo de situaciones, incluyendo

como casos particulares el modelo de regresión lineal, los modelo ARIMA y los modelos de función de transferencia siendo además muy fáciles de implementar.

Considera un modelo, en el que se pueden incluir todas las situaciones anteriormente descritas, del tipo:

$$y_t = \mu + b'x_t + \phi'y_{t-1} + \theta'a_{t-1} + a_t$$

donde μ es un parámetro de nivel global, b , ϕ y θ son vectores de parámetros, x_t representa el vector de variables explicativas, y_{t-1} es el vector de valores retardados de la respuesta y a_{t-1} es un vector iid $N(0, \sigma^2)$, incluyendo de esta forma un parámetro de nivel μ , un parámetro de escala σ^2 y un vector de parámetros estructurales $\beta' = (b', \phi', \theta')$, parámetros todos ellos que en la estimación pueden verse afectados por la presencia de algún valor atípico. Considerando la estimación máximo verosímil de estos parámetros, tendríamos el vector de estimaciones: $\hat{\mu}, \hat{\beta}, \hat{\sigma}_\mu^2, \hat{\Sigma}_\beta$, siendo éstas dos últimas las estimaciones de las matrices de varianzas-covarianzas de estos estimadores.

Ahora bien, supongamos que la observación i -ésima se ha detectado que puede ser un valor atípico y se trata como un "missing value" obteniéndose así los estimadores máximo verosímiles en esta situación: $\hat{\mu}_{(i)}, \hat{\beta}_{(i)}$, siendo los EMV de los parámetros del modelo cuando se considera que la observación i -ésima es un "missing value", $\hat{\sigma}^2$ es un estimador insesgado de la varianza σ^2 y $\hat{\sigma}_{(i)}^2$ representa el mismo estimador cuando la observación i -ésima se considera "missing value".

Teniendo en cuenta estas estimaciones, Peña (1987) deriva tres medidas de influencia:

A - Influencia en el nivel global:

$$D_L(i) = \frac{(\hat{\mu} - \hat{\mu}_{(i)})^2}{\hat{\sigma}_\mu^2} \quad (2.3.1)$$

B - Influencia en los parámetros estructurales:

$$D_s(i) = (\hat{\beta} - \hat{\beta}_{(i)})' \hat{\Sigma}_{\beta} (\hat{\beta} - \hat{\beta}_{(i)}) \quad (2.3.2)$$

medida en la cual se pueden separar los efectos de los componentes del vector de parámetros β .

C - Influencia en la escala:

$$D_v(i) = \frac{(\hat{\sigma}^2 - \hat{\sigma}_{(i)}^2)}{\hat{\sigma}_{(i)}^2} \quad (2.3.3)$$

Mediante estas tres medidas de influencia se puede calcular como será el efecto del valor atípico considerado en la observación i -ésima sobre los estimadores del modelo.

Las medidas de influencia expuestas en esta sección son las más novedosas en estos temas y tienen la ventaja de no ser exclusivas para una situación muy determinada. Se pueden calcular para modelos de muy diversa índole, sin necesidad de considerar una situación especial.

Tradicionalmente, se derivaron otras medidas de influencia bajo la forma de curvas de influencia que expondremos a continuación.

2.4 - CURVAS DE INFLUENCIA

A partir de las siguientes preguntas ¿cómo afecta la contaminación a la precisión (y tal vez el sesgo) de un estimador? ¿la contaminación, influye sobre un estimador dado? ¿el efecto sobre un estimador es proporcional al número de contaminantes presentes?; suponiendo que sólo hay un contaminante, ¿cómo se relaciona su efecto con su magnitud? ¿cuál es el peor efecto que puede tener un contaminante? ¿es un efecto limitado o no?, se derivaron las curvas de influencia para determinar en qué grado los atípicos o contaminantes afectan a los estadísticos y estimadores para una muestra dada.

Un contaminante o un atípico, en una muestra de tamaño n si es de gran magnitud, generará un cambio importante en la media hacia alguno de sus límites; pero, dos contaminantes, en una muestra de tamaño impar, $n=2m-1$, puede al menos cambiar la mediana muestral $x_{(m)}$ a $x_{(m-1)}$ o $x_{(m+1)}$. Aspectos como estos proporcionan herramientas poderosas que dan lugar a la curva ó función de influencia. Para describirlas, Barnett y Lewis (1984) se basan en los trabajos de Hampel (1968,1971) quien las analiza desde un punto de vista general sin centrarse en los casos particulares del modelo lineal general o de las series temporales.

Dadas n "buenas observaciones" (que llamaremos a partir de ahora modelo básico) $x_1, x_2, \dots, x_n$ y un estimador: $T(x_1, x_2, \dots, x_n)$, queremos en principio examinar el efecto sobre T de sustituir una de las n observaciones por un contaminante. Denominaremos a la observación contaminante x_c y supondremos que tiene un valor fijo, ξ . De esta forma, se analizará el efecto sobre el estimador T del contaminante, de manera que T en el modelo alternativo se basa en una muestra de tamaño $n+1$.

Con objeto de aclarar los términos del problema, analizaremos dos ejemplos. Supongamos que $T(x_1, x_2, \dots, x_n) = \bar{x}$ es la media muestral $\bar{x}$. Denotamos $T(x_1, x_2, \dots, x_n, \xi) = \bar{x}_c$ para la media $\bar{x}_c$ en la muestra contaminada mayor. De esta forma, el efecto de añadir ξ supone cambiar el valor del estimador en una cantidad:

$$\bar{x}_c - \bar{x} = \frac{n\bar{x} + \xi}{n+1} - \bar{x} = \frac{\xi - \bar{x}}{n+1} \quad (2.4.1)$$

Esto es proporcional a $1/(n+1)$, esto es a la cantidad de contaminación en la muestra; el efecto estandarizado para la cantidad de contaminación es:

$$(n+1)(\bar{x}_c - \bar{x}) = \xi - \bar{x} \quad (2.4.2)$$

Esta cantidad excederá de cualquier límite para ξ suficientemente grande, de manera que es una función lineal del valor del contaminante.

Por otro lado, si T es la varianza muestral S^2 para una distribución con media y varianza desconocidas, tendremos para la muestra ampliada:

$$nS_c^2 = (n-1)S^2 + n\bar{x}^2 + \xi^2 - (n\bar{x} + \xi)^2 / (n+1) \quad (2.4.3)$$

proporcionando:

$$S_c^2 - S^2 = [(\xi - \bar{x})^2 / (n+1)] - (S^2 / n) \quad (2.4.4)$$

El efecto estandarizado será entonces:

$$(n+1)(S_c^2 - S^2) = (\xi - \bar{x})^2 - \frac{(n+1)}{n} S^2 \quad (2.4.5)$$

El efecto, de nuevo, excede cualquier límite para ξ suficientemente grande, pero esta vez es una función cuadrática de ξ . Los efectos por unidad de contaminación (2.4.2) y (2.4.5) se llaman **funciones de influencia en muestras finitas** (o alguna veces, curvas de influencia en muestras finitas). Una **función de influencia en muestra finita** depende del argumento, ξ , del estimador, T , y en general de la distribución de base, F ; también depende explícitamente del tamaño muestral, n . Esto se denota: $IC_{T,F;n}(\xi)$.

El equivalente asintótico se define entonces:

$$\lim_{n \rightarrow \infty} IC_{T,F;n}(\xi) = IC_{T,F}(\xi) \quad (2.4.6)$$

Esta es la función de influencia asintótica, o simplemente, la función de influencia, $IC_{T,F}(\xi)$, la cual se revela como una herramienta útil.

$$IC_{s^2,F}(\xi) = (\xi - \mu)^2 - \sigma^2 \quad (2.4.7)$$

Esta ecuación convierte la misma información que la versión finita en (2.4.5) con respecto a la influencia cuadrática no limitada por sus contaminantes en S^2 . En esta derivación, cada estimador $T(x_1, \dots, x_n)$ (para $\bar{x} = S^2$) en la parte derecha de (2.4.5) ha sido

reemplazado por $\lim_{n \rightarrow \infty} T(x_1, \dots, x_n) (\mu, \sigma^2)$. Esta forma límite depende solamente de F y la denotaremos $T(F)$.

Definición formal: Decimos que la función de influencia de un estimador $T(x_1, \dots, x_n)$ en la distribución básica F es:

$$IC_{T,F}(\xi) = \lim_{\lambda \rightarrow 0} \{ [T((1-\lambda)F + \lambda G) - T(F)] / \lambda \} \quad (2.4.8)$$

donde G es la distribución atómica:

$$P[x = \xi] = 1$$

Podemos también escribir (2.4.8) como:

$$IC_{T,F}(\xi) = \left. \frac{\partial}{\partial \lambda} \{ T[(1-\lambda)F + \lambda G] \} \right|_{\lambda=0} \quad (2.4.9)$$

Analizaremos a continuación ciertos ejemplos de la curva de influencia asintótica.

Media muestral:

$$IC_{\bar{x},F}(\xi) = \lim_{\lambda \rightarrow \infty} \{ [(1-\lambda)\mu + \lambda\xi - \mu] / \lambda \} = \xi - \mu$$

que podría haberse obtenido haciendo $n \rightarrow \infty$ en (2.4.2).

Varianza muestral:

$$IC_{s^2,F}(\xi) = \lim_{\lambda \rightarrow \infty} \{ [(1-\lambda)(\mu^2 + \sigma^2) + \lambda\xi^2 - (1-\lambda)\mu + \lambda\xi^2 - \sigma^2] / \lambda \} = (\xi - \mu)^2 - \sigma^2$$

Mediana muestral

No es practicable calcular la función de influencia de la mediana muestral $\tilde{x}$ en una muestra finita puesto que el cambio en la mediana muestral en el movimiento de un número de observaciones par o impar ó viceversa no está definido. En la base asintótica, sin embargo, el cálculo es directo. $T(F)$ es ahora la mediana poblacional m , definida por:

$$\int_{-\infty}^m dF = \frac{1}{2}$$

$T((1-\lambda)F+\lambda G)$, la mediana de la distribución mixta, es igual a $m+\Delta$, donde Δ es un número positivo o negativo dependiendo de si ξ es mayor o menor que m . Asumimos que F es continua, con densidad f . Para $\Delta < 0$, tenemos:

$$\frac{1}{2} = (1-\lambda)F(m+\Delta) + \lambda = (1-\lambda)\left[\frac{1}{2} + \Delta f(m)\right] + \lambda = \frac{1}{2} + \frac{1}{2}\lambda + \Delta f(m)$$

generando $\Delta = -\lambda/[2f(m)]$. De forma similar para $\Delta > 0$, $\Delta = +\lambda/[2f(m)]$; entonces la función de influencia, $\lim_{\lambda \rightarrow 0} (\Delta/\lambda)$ es:

$$IC_{\bar{x},F}(\xi) = \frac{\text{sgn}(\xi - m)}{2f(m)}$$

La influencia de un contaminante sobre la mediana muestral está así delimitada, siendo ésta una diferencia esencial con la media muestral.

A partir de las curvas o funciones de influencia se han derivado numerosos estadísticos para detectar la presencia de valores atípicos o de contaminantes, evaluando su importancia en la muestra. Estos estadísticos o métodos de detección de atípicos serán analizados en secciones siguientes.

2.5 - CLASES DE ATÍPICOS EN EL MODELO DE REGRESIÓN LINEAL

En el caso del MRL, se considera que un valor atípico afecta únicamente al período de tiempo en el que se produce aunque si la variable independiente presenta diferencias para lograr la estacionariedad en media, el atípico puede ocasionar variaciones bruscas en los períodos afectados por el orden de diferenciación de la variable. Se considera pues que la observación atípica no se propagará en el tiempo, como ocurre el caso de la metodología de series temporales.

En este sentido, en el modelo lineal, podemos encontrar dos tipos de situaciones: un cambio brusco en el nivel de la serie en un único período, volviendo posteriormente a su nivel anterior (hablaríamos en este caso de un **atípico tipo impulso**) y una segunda situación en la que se produce un cambio en el nivel de la serie, nivel en el que se mantiene en el resto de la muestra (hablándose en este caso de un **atípico tipo escalón**).

Estas dos situaciones descritas son muy comunes en las variables de tipo económico, en las que las intervenciones por parte de las autoridades económicas son bastante frecuentes. Sin embargo, estas intervenciones pueden ser puntuales provocando la aparición de valores atípicos tipo impulso (como la disminución del precio de la gasolina en un mes en el que se espera que la inflación en el resto de los componentes del índice de precios sea elevada). También pueden ser intervenciones con una duración prolongada como es el caso de una devaluación de la moneda que genera un aumento en el nivel del tipo de cambio, manteniéndose en un nivel superior al anterior.

Para corregir la influencia de estos dos tipos de atípicos, se construyen dos variables ficticias, dependiendo del caso que se analice. Por un lado, una variable tipo impulso, que denominaremos I_t^τ , definida mediante la expresión:

$$I_t^\tau = \begin{cases} 0 & t \neq \tau \\ 1 & t = \tau \end{cases}$$

La variable ficticia para la corrección de un atípico tipo escalón, denominada E_t^τ , se define mediante la expresión:

$$E_t^\tau = \begin{cases} 0 & 0 \leq t < \tau \\ 1 & \tau \leq t \leq T \end{cases}$$

Nótese que $I_t^\tau = (1 - B)E_t^\tau$, siendo B el operador retardos.

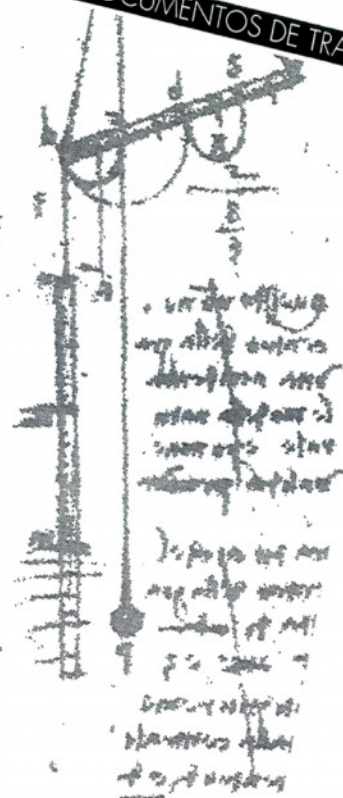
Los datos de la muestra se han obtenido de la siguiente manera: se ha considerado una muestra aleatoria simple de la población, y se han calculado los estadísticos de la muestra. Los resultados se han presentado en la siguiente forma:

Los datos de la muestra se han obtenido de la siguiente manera: se ha considerado una muestra aleatoria simple de la población, y se han calculado los estadísticos de la muestra. Los resultados se han presentado en la siguiente forma:

Los datos de la muestra se han obtenido de la siguiente manera: se ha considerado una muestra aleatoria simple de la población, y se han calculado los estadísticos de la muestra. Los resultados se han presentado en la siguiente forma:

3.- EFECTOS DE LOS VALORES ATÍPICOS SOBRE LOS ESTADÍSTICOS

Los datos de la muestra se han obtenido de la siguiente manera: se ha considerado una muestra aleatoria simple de la población, y se han calculado los estadísticos de la muestra. Los resultados se han presentado en la siguiente forma:



Los datos de la muestra se han obtenido de la siguiente manera: se ha considerado una muestra aleatoria simple de la población, y se han calculado los estadísticos de la muestra. Los resultados se han presentado en la siguiente forma:

Los datos de la muestra se han obtenido de la siguiente manera: se ha considerado una muestra aleatoria simple de la población, y se han calculado los estadísticos de la muestra. Los resultados se han presentado en la siguiente forma:

Los datos de la muestra se han obtenido de la siguiente manera: se ha considerado una muestra aleatoria simple de la población, y se han calculado los estadísticos de la muestra. Los resultados se han presentado en la siguiente forma:

3.- EFECTOS DE LOS VALORES ATÍPICOS SOBRE LOS ESTADÍSTICOS

En el análisis del modelo de regresión, la falta de modelización de un valor atípico se puede enmarcar dentro del problema de errores de especificación. Si consideramos el modelo verdadero como aquel que tiene en consideración el tratamiento de un valor atípico y el modelo erróneo aquel que omite la modelización del atípico, se pueden comparar las estimaciones y sus estadísticos asociados en ambos modelos bajo el enfoque de los errores de especificación.

Para lograr este objetivo, se consideran los dos tipos de atípicos que se presentan en la estimación del modelo de regresión tradicional: un atípico impulso, es decir, aquel que se produce en un momento del período de estimación y el atípico escalón, aquel que produce un cambio de nivel en la variable endógena, nivel en el que se mantendrá hasta el final de la muestra. Estos dos atípicos se introducirán por separado, diferenciando así los efectos que se producen en las estimaciones según se trate de un atípico de tipo impulso o de tipo escalón. Las definiciones de las variables ficticias asociadas a estas dos situaciones serán:

- Para un atípico tipo impulso: $I_t^* = \begin{cases} 0 & t \neq \tau \\ 1 & t = \tau \end{cases}$
- Para un atípico tipo escalón: $I_t^* = \begin{cases} 0 & 0 \leq t < \tau \\ 1 & \tau \leq t \leq T \end{cases}$

Con objeto de simplificar la exposición consideraremos un modelo con una única variable explicativa y en desviaciones con respecto a la media. Esta simplificación no alterará los resultados siendo inmediata, si bien más complicada, su generalización a un modelo con k variables explicativas. Es importante también resaltar que se analizarán por separado las consecuencias de la omisión de la modelización de los valores atípicos, siendo también inmediata la generalización a la presencia de múltiples valores atípicos y de distinto tipo.

Debido a que englobamos el problema dentro de los errores de especificación, trataremos con dos modelos, uno considerado como erróneo:

$$y_t = \beta_1 x_{1t} + v_t \text{ donde el investigador supone } v_t \sim N(0, \sigma_v^2) \quad (3.1)$$

y asumimos que $\sigma_u^2 = \sigma_v^2$. El segundo, será el modelo verdadero, donde se considera la modelización del atípico:

$$y_t = \beta_1 x_{1t} + \delta I_t^\tau + u_t \quad u_t \sim N(0, \sigma_u^2) \quad (3.2)$$

Pasamos a continuación a calcular las estimaciones de los parámetros en cada modelo.

Modelo verdadero

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\delta} \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^T x_{1t}^2 & \sum_{t=1}^T I_t^\tau x_{1t} \\ \sum_{t=1}^T I_t^\tau x_{1t} & \sum_{t=1}^T I_t^{\tau^2} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^T x_{1t} y_t \\ \sum_{t=1}^T I_t^\tau y_t \end{bmatrix} \quad (3.3)$$

A partir de este sistema de ecuaciones, se pueden hallar inmediatamente los estimadores del modelo verdadero que serán:

$$\begin{aligned} \hat{\beta}_1 &= \beta_1 + \frac{\sum_{t=1}^T I_t^{\tau^2} \sum_{t=1}^T x_{1t} u_t - \sum_{t=1}^T x_{1t} I_t^\tau \sum_{t=1}^T I_t^\tau u_t}{\sum_{t=1}^T x_{1t}^2 \sum_{t=1}^T I_t^{\tau^2} - \left(\sum_{t=1}^T x_{1t} I_t^\tau \right)^2} \\ \hat{\delta} &= \delta + \frac{\sum_{t=1}^T x_{1t}^2 \sum_{t=1}^T I_t^\tau u_t - \sum_{t=1}^T x_{1t} I_t^\tau \sum_{t=1}^T x_{1t} u_t}{\sum_{t=1}^T x_{1t}^2 \sum_{t=1}^T I_t^{\tau^2} - \left(\sum_{t=1}^T x_{1t} I_t^\tau \right)^2} \end{aligned} \quad (3.4)$$

Si $x' E[u] = 0$, entonces, estos dos estimadores serán insesgados, $E[\hat{\beta}_1] = \beta_1$; $E[\hat{\delta}] = \delta$, condición que se cumplirá sea cual sea el tipo de atípico que se considere, ya sea impulso o escalón.

Por lo que respecta a las varianzas de ambos estimadores éstas vendrán definidas a partir de la expresión:

$$\begin{bmatrix} \text{Var}(\hat{\beta}_1) \\ \text{Var}(\hat{\delta}) \end{bmatrix} = \sigma_u^2 \begin{bmatrix} \sum_{t=1}^T x_{1t}^2 & \sum_{t=1}^T x_{1t} I_t \\ \sum_{t=1}^T x_{1t} I_t & \sum_{t=1}^T I_t^2 \end{bmatrix}^{-1} \quad (3.5)$$

obteniéndose fácilmente las expresiones para las varianzas:

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \frac{\sigma_u^2 \sum_{t=1}^T I_t^2}{\sum_{t=1}^T x_{1t}^2 \sum_{t=1}^T I_t^2 - \left(\sum_{t=1}^T x_{1t} I_t \right)^2} \\ \text{Var}(\hat{\delta}) &= \frac{\sigma_u^2 \sum_{t=1}^T x_{1t}^2}{\sum_{t=1}^T x_{1t}^2 \sum_{t=1}^T I_t^2 - \left(\sum_{t=1}^T x_{1t} I_t \right)^2} \end{aligned} \quad (3.6)$$

Tanto los estimadores como las varianzas tendrán una expresión ligeramente distinta según se trate de un atípico tipo impulso o de un atípico tipo escalón. Con el objeto de apreciar las diferencias entre los estimadores y las varianzas en cada caso, presentamos ambas expresiones para cada caso.

A- Atípico tipo impulso

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{t=1}^T x_{1t} u_t - x_{1t} u_t}{\sum_{t=1}^T x_{1t}^2 - (x_{1t})^2} \quad (3.7)$$

siendo su varianza

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_u^2}{\sum_{t=1}^T x_{1t}^2 - (x_{1t})^2} \quad (3.8)$$

$$\hat{\delta} = \delta + \frac{u_{\tau} \sum_{t=1}^T x_{1t}^2 - x_{1\tau} \sum_{t=1}^T x_{1t} u_t}{\sum_{t=1}^T x_{1t}^2 - (x_{1\tau})^2} \quad (3.9)$$

$$\text{Var}(\hat{\delta}) = \frac{\sigma_u^2 \sum_{t=1}^T x_{1t}^2}{\sum_{t=1}^T x_{1t}^2 - (x_{1\tau})^2} \quad (3.10)$$

donde $x_{1\tau}$ es el valor de la variable x_{1t} en el momento τ y u_{τ} es el valor de la perturbación en el momento τ .

B- Atípico tipo escalón

$$\hat{\beta}_1 = \beta_1 + \frac{(T-\tau+1) \sum_{t=1}^T x_{1t} u_t - \sum_{t=\tau}^T x_{1t} \sum_{t=\tau}^T u_t}{(T-\tau+1) \sum_{t=1}^T x_{1t}^2 - \left[\sum_{t=\tau}^T x_{1t} \right]^2} \quad (3.11)$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_u^2 (T-\tau+1)}{(T-\tau+1) \sum_{t=1}^T x_{1t}^2 - \left[\sum_{t=\tau}^T x_{1t} \right]^2} \quad (3.12)$$

$$\hat{\delta} = \delta + \frac{\sum_{t=1}^T x_{1t}^2 \sum_{t=\tau}^T u_t - \sum_{t=\tau}^T x_{1t} \sum_{t=1}^T x_{1t} u_t}{(T-\tau+1) \sum_{t=1}^T x_{1t}^2 - \left[\sum_{t=\tau}^T x_{1t} \right]^2} \quad (3.13)$$

$$\text{Var}(\hat{\delta}) = \frac{\sigma_u^2 \sum_{t=1}^T x_{1t}^2}{(T-\tau+1) \sum_{t=1}^T x_{1t}^2 - \left[\sum_{t=\tau}^T x_{1t} \right]^2} \quad (3.14)$$

donde T es el tamaño muestral, $(\tau-1)$ representa el momento muestral justamente anterior a la aparición del atípico

Modelo erróneo

El estimador de β en este caso vendrá dado por la expresión:

$$\tilde{\beta}_1 = \frac{\sum_{t=1}^T x_{1t} y_t}{\sum_{t=1}^T x_{1t}^2} = \{y_t = x_1 \beta_1 + I_t^\tau \delta + u_t\} = \beta_1 + \delta \frac{\sum_{t=1}^T x_{1t} I_t^\tau}{\sum_{t=1}^T x_{1t}^2} + \frac{\sum_{t=1}^T x_{1t} u_t}{\sum_{t=1}^T x_{1t}^2} \quad (3.15)$$

donde se aprecia claramente que es un estimador sesgado, teniendo el sesgo la expresión:

$$B = \delta \frac{\sum_{t=1}^T x_{1t} I_t^\tau}{\sum_{t=1}^T x_{1t}^2} \quad (3.16)$$

Esta expresión depende del parámetro asociado a la intervención en el modelo verdadero y de la relación existente entre la intervención y la variable exógena incluida en el modelo.

Por su parte, la varianza de este estimador será:

$$\text{Var}(\tilde{\beta}_1) = \frac{\sigma_v^2}{\sum_{t=1}^T x_{1t}^2} \quad (3.17)$$

La expresión del sesgo variará en función del tipo de atípico que se considere, mientras que la expresión de la varianza no se verá afectada por el atípico.

A- Atípico tipo impulso:

La expresión del estimador para un atípico tipo impulso será:

$$\tilde{\beta}_1 = \beta_1 + \delta \frac{x_{1\tau}}{\sum_{t=1}^T x_{1t}^2} + \frac{\sum_{t=1}^T x_{1t} u_t}{\sum_{t=1}^T x_{1t}^2} \quad (3.18)$$

viniendo el sesgo en este caso determinado por el parámetro asociado a la variable ficticia de la intervención, por el valor de la variable exógena en el momento de la intervención y por la suma de cuadrados de esta misma variable para todo el período muestral.

B- Atípico tipo escalón

El estimador en este caso será:

$$\tilde{\beta}_1 = \beta_1 + \delta \frac{\sum_{t=\tau}^T x_{1t}}{\sum_{t=1}^T x_{1t}^2} + \frac{\sum_{t=1}^T x_{1t} u_t}{\sum_{t=1}^T x_{1t}^2} \quad (3.19)$$

siendo en este caso el sesgo:

$$B = \delta \frac{\sum_{t=\tau}^T x_{1t}}{\sum_{t=1}^T x_{1t}^2} \quad (3.20)$$

Como puede observarse nos encontramos con dos estimadores, uno insesgado y otro sesgado, y cada uno con varianza distinta, por lo que para decidir cuál de los dos estimadores es mejor tendremos que comparar las varianzas.

Atendiendo a cada uno de los casos, se puede observar que la varianza del estimador del modelo verdadero es mayor que la del estimador del modelo erróneo.

En el caso de un atípico tipo impulso, a partir de la expresión (3.8), la varianza del estimador en el modelo verdadero puede escribirse:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_u^2}{\sum_{t=1}^T x_{1t}^2 \left[1 - \frac{(x_{1T})^2}{\sum_{t=1}^T x_{1t}^2} \right]} = \frac{\sigma_u^2 / \sum_{t=1}^T x_{1t}^2}{1 - \frac{x_{1T}^2}{\sum_{t=1}^T x_{1t}^2}} = \frac{\sigma_u^2 / \sum_{t=1}^T x_{1t}^2}{1 - a} \quad (3.21)$$

de manera que, si comparamos las varianzas de ambos estimadores, todo dependerá del factor a expresado por:

$$a = \frac{(x_{1T})^2}{\sum_{t=1}^T x_{1t}^2} \quad (3.22)$$

estando siempre comprendido entre 0 y 1, de manera que se pueden obtener las siguientes situaciones:

$$\text{- si } a = 0 \Rightarrow \text{Var}(\hat{\beta}_1) = \frac{\sigma_u^2}{\sum_{t=1}^T x_{1t}^2} = \text{Var}(\tilde{\beta}_1)$$

$$\text{- si } a = 1 \Rightarrow \text{Var}(\hat{\beta}_1) \text{ estaría indeterminada.}$$

$$\text{- si } 0 < a < 1 \Rightarrow \text{Var}(\hat{\beta}_1) > \text{Var}(\tilde{\beta}_1)$$

De esta forma, tendremos:

- Modelo verdadero: estimador insesgado y varianza mayor en un gran número de situaciones.

- Modelo erróneo: estimador sesgado y varianza menor en la mayor parte de los casos.

Por este motivo, tendremos que recurrir al análisis de los errores cuadráticos medios para determinar cuál de los estimadores es preferible.

Todo lo realizado para el caso del típico tipo impulso se verifica también para el otro tipo de atípico.

En el caso de un atípico tipo escalón, a partir de la expresión (3.12), la varianza quedará:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_u^2 (T - \tau + 1)}{(T - \tau + 1) \sum_{t=1}^T x_{1t}^2 - \left(\sum_{t=\tau}^T x_{1t} \right)^2} = \frac{\sigma_u^2 \cdot \frac{1}{\sum_{t=1}^T x_{1t}^2}}{1 - \bar{x}_{1\tau} \frac{\sum_{t=\tau}^T x_{1t}}{\sum_{t=1}^T x_{1t}}} = \frac{\text{Var}(\tilde{\beta}_1)}{1 - \bar{x}_{1\tau} \frac{\sum_{t=\tau}^T x_{1t}}{\sum_{t=1}^T x_{1t}}} \quad (3.23)$$

donde $\bar{x}_{1\tau}$ representa a la media de la variable exógena en el período comprendido entre τ y T .

A partir de la desigualdad de Jensen,

$$\frac{1}{T} \sum_{t=1}^T x_{1t}^2 \geq \left(\frac{1}{T} \sum_{t=1}^T x_{1t} \right)^2 \Rightarrow \sum_{t=1}^T x_{1t}^2 \geq \frac{1}{T} \left(\sum_{t=1}^T x_{1t} \right)^2$$

Por otro lado:

$$\begin{aligned} \sum_{t=1}^T x_{1t}^2 &\geq \sum_{t=\tau}^T x_{1t}^2 \{ \text{aplicando desigualdad de Jensen} \} \geq \frac{1}{(T - \tau + 1)} \left(\sum_{t=\tau}^T x_{1t} \right)^2 = \bar{x}_{1\tau} \sum_{t=\tau}^T x_{1t} \\ \Rightarrow \sum_{t=1}^T x_{1t}^2 &\geq \bar{x}_{1\tau} \sum_{t=\tau}^T x_{1t} \Rightarrow \bar{x}_{1\tau} \frac{\sum_{t=\tau}^T x_{1t}}{\sum_{t=1}^T x_{1t}} \leq 1 \end{aligned}$$

A partir de esto se aprecia claramente que $\text{Var}(\hat{\beta}_1) > \text{Var}(\tilde{\beta}_1)$. De esta forma, nos encontramos de nuevo con un estimador insesgado pero con varianza mayor y un estimador sesgado con menor varianza, lo que nos lleva al análisis de los errores cuadráticos medios para decidir cuál de los estimadores es el más apropiado.

Análisis del cociente de Errores Cuadráticos Medios

El error cuadrático medio (ECM), definido mediante la expresión:

$$\text{ECM} = (\text{sesgo})^2 + \text{Varianza} \quad (3.24)$$

nos permite comparar dos estimadores con el objeto de decidir cuál de los dos tiene las mejores propiedades.

1.- Atípico tipo impulso: este cociente tendrá la siguiente expresión:

$$\frac{\text{ECM}(\tilde{\beta}_1)}{\text{ECM}(\hat{\beta}_1)} = \frac{\delta^2 \frac{[x_{1t}]^2}{\left[\sum_{t=1}^T x_{1t}^2\right]^2} + \frac{\sigma_u^2}{\sum_{t=1}^T x_{1t}^2}}{\frac{\sum_{t=1}^T x_{1t}^2 \left[1 - \frac{[x_{1t}]^2}{\sum_{t=1}^T x_{1t}^2}\right]}{\sigma_u^2}} \quad (3.25)$$

donde si operamos se llega, después de unas no demasiado complicadas operaciones algebraicas, a la siguiente expresión para el cociente de errores cuadráticos medios:

$$\frac{\text{ECM}(\tilde{\beta}_1)}{\text{ECM}(\hat{\beta}_1)} = a \left[\frac{\delta^2}{\text{Var}(\hat{\delta})} - 1 \right] + 1; \quad 0 \leq a \leq 1 \quad (3.26)$$

donde a viene dado por la ecuación (3.22) y $\text{Var}(\hat{\delta})$ es la varianza del estimador del parámetro asociado a la intervención estimador en el modelo verdadero. De esta forma, para

decidir acerca de ambos estimadores, tendremos que analizar los posibles valores que puede tomar:

$$\frac{\delta^2}{\text{Var}(\hat{\delta})} \quad (3.27)$$

Casos:

$$(a) \text{ Si } \frac{\delta^2}{\text{Var}(\hat{\delta})} = 0 \Rightarrow \frac{\text{ECM}(\tilde{\beta}_1)}{\text{ECM}(\hat{\beta}_1)} < 1 \Rightarrow \text{ECM}(\tilde{\beta}_1) \leq \text{ECM}(\hat{\beta}_1)$$

Si se produce este caso, $\delta^2 = 0$, $E(\tilde{\beta}_1) = \beta_1$, de manera que el modelo erróneo sería válido no produciéndose el error de especificación.

$$(b) \text{ Si } \frac{\delta^2}{\text{Var}(\hat{\delta})} = 1 \Rightarrow \frac{\text{ECM}(\tilde{\beta}_1)}{\text{ECM}(\hat{\beta}_1)} = 1 \Rightarrow \text{ECM}(\tilde{\beta}_1) = \text{ECM}(\hat{\beta}_1) \text{ lo que implicaría que}$$

la modelización del valor atípico en el modelo verdadero no mejoraría el ECM de la estimación.

$$(c) \text{ Si } \frac{\delta^2}{\text{Var}(\hat{\delta})} > 1 \Rightarrow \frac{\text{ECM}(\tilde{\beta}_1)}{\text{ECM}(\hat{\beta}_1)} > 1 \Rightarrow \text{ECM}(\tilde{\beta}_1) > \text{ECM}(\hat{\beta}_1) \text{ lo que supondría que}$$

la estimación de β en el modelo verdadero sería estadísticamente más idóneo.

En principio, si la variable endógena contiene un valor atípico y éste viene recogido por δ_1 , la estimación de δ en el modelo verdadero sería significativa. En este sentido, si sustituimos $\frac{\delta^2}{\text{Var}(\hat{\delta})}$ por :

$$\frac{\hat{\delta}^2}{\text{Var}(\hat{\delta})} \quad (3.28)$$

esta expresión es equivalente al cuadrado del estadístico t para contrastar la hipótesis de significatividad individual del parámetro δ . De esta forma, el cociente de los errores cuadráticos medios vendría representado por:

$$\frac{ECM(\tilde{\beta}_1)}{ECM(\hat{\beta}_1)} = a[t^2 - 1] + 1 \quad 0 \leq a \leq 1 \quad (3.29)$$

donde en este caso todo el análisis vendría determinado por dicho estadístico. Debido a que la inclusión de variables ficticias para recoger comportamientos atípicos no captados por las variables exógenas únicamente se realiza si realmente dicho valor es atípico¹, se puede suponer fácilmente que dicho estadístico será mayor que 1, lo que implicaría que siempre se cumpliría que:

$$ECM(\tilde{\beta}_1) > ECM(\hat{\beta}_1) \quad (3.30)$$

Esto nos puede llevar a concluir que, si consideramos (3.29) como válida, siempre será mejor modelizar el valor atípico, obteniéndose estimadores con errores cuadráticos medios menores que para los estimadores en los que no se modelizan los valores atípicos.

2. ATÍPICO TIPO ESCALÓN

En este caso, el cociente de los errores cuadráticos medios vendrá dado por la expresión:

$$\frac{ECM(\tilde{\beta}_1)}{ECM(\hat{\beta}_1)} = \frac{\delta^2 \frac{\left[\sum_{t=1}^T x_{1t} \right]^2}{\sum_{t=1}^T x_{1t}^2} + \frac{\sigma_u^2}{\sum_{t=1}^T x_{1t}^2}}{\frac{\sigma_u^2}{\sum_{t=1}^T x_{1t}^2} \frac{1}{1 - \bar{x}_{1t} a}} = \frac{\delta^2 a^2 + \sigma_u^2 / \sum_{t=1}^T x_{1t}^2}{\frac{\sigma_u^2}{\sum_{t=1}^T x_{1t}^2} \frac{1}{1 - \bar{x}_{1t} a}} \quad (3.31)$$

¹ Es decir, si realmente en dicha fecha se produce algún hecho anómalo que provoque la aparición de ese atípico.

expresión que, después de unas modificaciones algebraicas, y considerando el valor de la $\text{Var}(\hat{\delta})$ de la expresión (3.14) quedará:

$$\frac{\text{ECM}(\tilde{\beta}_1)}{\text{ECM}(\hat{\beta}_1)} = 1 + a \cdot \bar{x}_t \left[\left(\frac{\delta^2}{\text{Var}(\hat{\delta})} - 1 \right) \right] \quad (3.32)$$

viniendo en este caso a definido por la expresión:

$$a = \frac{\sum_{t=\tau}^T x_{1t}}{\sum_{t=1}^T x_{1t}^2} \quad (3.33)$$

De esta forma, el cociente de errores cuadráticos medios dependerá de diversas variables y de sus posibles combinaciones:

Los posibles casos que se pueden dar en esta situación son:

1.- Si $a=0$, $\text{ECM}(\tilde{\beta}_1) = \text{ECM}(\hat{\beta}_1)$, lo que implica que no existe error de especificación.

2.- Si $a > 0 \Rightarrow \sum_{t=\tau}^T x_{1t} > 0$, nos encontraremos con las siguientes situaciones:

$$a) \frac{\delta^2}{\text{Var}(\hat{\delta})} > 1 \Rightarrow a \cdot \bar{x}_t \left[\frac{\delta^2}{\text{Var}(\hat{\delta})} - 1 \right] > 0 \Rightarrow \text{ECM}(\tilde{\beta}_1) > \text{ECM}(\hat{\beta}_1)$$

$$b) \frac{\delta^2}{\text{Var}(\hat{\delta})} < 1 \Rightarrow a \cdot \bar{x}_t \left[\frac{\delta^2}{\text{Var}(\hat{\delta})} - 1 \right] < 0 \Rightarrow \text{ECM}(\tilde{\beta}_1) < \text{ECM}(\hat{\beta}_1)$$

$$c) \frac{\delta^2}{\text{Var}(\hat{\delta})} = 1 \Rightarrow a \cdot \bar{x}_t \left[\frac{\delta^2}{\text{Var}(\hat{\delta})} - 1 \right] = 0 \Rightarrow \text{ECM}(\tilde{\beta}_1) = \text{ECM}(\hat{\beta}_1)$$

3.- Si $a < 0 \Rightarrow \sum_{t=\tau}^T x_{1t} < 0 \Rightarrow \bar{x}_\tau < 0$, los casos que se pueden plantear serán:

$$a) \frac{\delta^2}{\text{Var}(\hat{\delta})} > 1 \Rightarrow a \cdot \bar{x}_\tau \left[\frac{\delta^2}{\text{Var}(\hat{\delta})} - 1 \right] > 0 \Rightarrow \text{ECM}(\tilde{\beta}_1) > \text{ECM}(\hat{\beta}_1)$$

$$b) \frac{\delta^2}{\text{Var}(\hat{\delta})} < 1 \Rightarrow a \cdot \bar{x}_\tau \left[\frac{\delta^2}{\text{Var}(\hat{\delta})} - 1 \right] < 0 \Rightarrow \text{ECM}(\tilde{\beta}_1) < \text{ECM}(\hat{\beta}_1)$$

$$c) \frac{\delta^2}{\text{Var}(\hat{\delta})} = 1 \Rightarrow a \cdot \bar{x}_\tau \left[\frac{\delta^2}{\text{Var}(\hat{\delta})} - 1 \right] = 0 \Rightarrow \text{ECM}(\tilde{\beta}_1) = \text{ECM}(\hat{\beta}_1)$$

siendo exactamente los mismos que en el caso 2.

Todo, realmente, salvo en el caso en que $a=0$, dependerá de $\frac{\delta^2}{\text{Var}(\hat{\delta})}$. de manera que si asumimos de nuevo que:

$$\frac{\delta^2}{\text{Var}(\hat{\delta})} \equiv \frac{\hat{\delta}^2}{\text{Var}(\hat{\delta})} = t^2$$

podremos basar el análisis en la inclusión o no de una variable ficticia para captar el comportamiento de un valor atípico. Así, podremos escribir:

$$\frac{\text{ECM}(\tilde{\beta}_1)}{\text{ECM}(\hat{\beta}_1)} = a \cdot \bar{x}_\tau [t^2 - 1] + 1$$

Si el parámetro $\hat{\delta}$ es estadísticamente significativo:

$$t^2 > 1 \Rightarrow \frac{\text{ECM}(\tilde{\beta}_1)}{\text{ECM}(\hat{\beta}_1)} > 1 \Rightarrow \text{ECM}(\tilde{\beta}_1) > \text{ECM}(\hat{\beta}_1)$$

Análisis de la varianza residual

El análisis de los residuos también debe verse afectado por la falta de modelización de un valor atípico ya que todo lo que no es explicado de la variable endógena por las variables exógenas en un modelo de regresión de la variable endógena, se encontrará en los residuos. Por este motivo es conveniente analizar los residuos en ambos modelos.

1) Residuos en el modelo verdadero

Podemos definir los residuos del modelo verdadero mediante la expresión:

$$\hat{u}_t = y_t - \hat{y}_t = (\beta_1 - \hat{\beta}_1)x_{1t} + (\delta - \hat{\delta})I_t^\tau + u_t \quad (3.34)$$

siendo la suma residual en este caso:

$$\sum_{t=1}^T \hat{u}_t^2 = (\beta_1 - \hat{\beta}_1)^2 \sum_{t=1}^T x_{1t}^2 + 2(\beta_1 - \hat{\beta}_1)(\delta - \hat{\delta}) \sum_{t=1}^T x_{1t} I_t^\tau + 2(\beta_1 - \hat{\beta}_1) \sum_{t=1}^T x_{1t} u_t + (\delta - \hat{\delta})^2 \sum_{t=1}^T I_t^{\tau^2} + 2(\delta - \hat{\delta}) \sum_{t=1}^T I_t^\tau u_t + \sum_{t=1}^T u_t^2 \quad (3.35)$$

La expresión (3.34) se verá modificada en cada caso dependiendo del tipo de atípico que se esté analizando, de manera que para un **atípico tipo impulso**, tendremos que la suma residual vendrá dada por:

$$\sum_{t=1}^T \hat{u}_t^2 = (\beta_1 - \hat{\beta}_1)^2 \sum_{t=1}^T x_{1t}^2 + 2(\beta_1 - \hat{\beta}_1)(\delta - \hat{\delta})x_{1\tau} + 2(\beta_1 - \hat{\beta}_1) \sum_{t=1}^T x_{1t} u_t + (\delta - \hat{\delta})^2 + 2(\delta - \hat{\delta})u_\tau + \sum_{t=1}^T u_t^2 \quad (3.36)$$

mientras que dicha expresión para el caso en el que el **atípico** es del **tipo escalón** será:

$$\sum_{t=1}^T \hat{u}_t^2 = (\beta_1 - \hat{\beta}_1)^2 \sum_{t=1}^T x_{1t}^2 + 2(\beta_1 - \hat{\beta}_1)(\delta - \hat{\delta}) \sum_{t=\tau}^T x_{1t} + 2(\beta_1 - \hat{\beta}_1) \sum_{t=\tau}^T x_{1t} u_t + (\delta - \hat{\delta})^2 (T - \tau + 1) + 2(\delta - \hat{\delta}) \sum_{t=\tau}^T u_t + \sum_{t=1}^T u_t^2 \quad (3.37)$$

2) Residuos en el modelo erróneo

Por su parte el modelo erróneo tendrá unos residuos que vendrán dados por la expresión:

$$\tilde{u}_t = y_t - \tilde{y}_t = (\beta_1 - \tilde{\beta}_1)x_{1t} + \delta I_t^\tau + u_t \quad (3.38)$$

y la suma residual tendrá la expresión siguiente:

$$\sum_{t=1}^T \tilde{u}_t^2 = (\beta_1 - \tilde{\beta}_1)^2 \sum_{t=1}^T x_{1t}^2 + 2(\beta_1 - \tilde{\beta}_1)\delta \sum_{t=1}^T x_{1t} I_t^\tau + 2(\beta_1 - \tilde{\beta}_1) \sum_{t=1}^T x_{1t} u_t + \delta^2 \sum_{t=1}^T I_t^{\tau^2} + 2\delta \sum_{t=1}^T I_t^\tau u_t + \sum_{t=1}^T u_t^2 \quad (3.39)$$

donde de nuevo dicha expresión se verá modificada en función del tipo de atípico que se considere. Para un **atípico de tipo impulso**, tendremos que la suma residual del modelo erróneo será:

$$\sum_{t=1}^T \tilde{u}_t^2 = (\beta_1 - \tilde{\beta}_1)^2 \sum_{t=1}^T x_{1t}^2 + 2(\beta_1 - \tilde{\beta}_1)\delta x_{1\tau} + 2(\beta_1 - \tilde{\beta}_1) \sum_{t=1}^T x_{1t} u_t + \delta^2 + 2\delta u_\tau + \sum_{t=1}^T u_t^2 \quad (3.40)$$

mientras que para el caso de un **atípico tipo escalón**, dicha expresión será:

$$\sum_{t=1}^T \tilde{u}_t^2 = (\beta_1 - \tilde{\beta}_1)^2 \sum_{t=1}^T x_{1t}^2 + 2(\beta_1 - \tilde{\beta}_1)\delta \sum_{t=\tau}^T x_{1t} + 2(\beta_1 - \tilde{\beta}_1) \sum_{t=1}^T x_{1t} u_t + \delta^2(T - \tau + 1) + 2\delta \sum_{t=\tau}^T u_t + \sum_{t=1}^T u_t^2 \quad (3.41)$$

Debido a que las expresiones anteriores no nos dicen mucho acerca de la comparación de las sumas residuales para cada uno de los casos, sería más útil analizar la esperanza de dichas sumas residuales, obteniéndose así una expresión para el sesgo de la estimación de la varianza residual generado como consecuencia de la omisión de la modelización de un valor atípico en un modelo de regresión.

La esperanza de la suma residual para el modelo verdadero puede escribirse:

$$E\left[\sum_{t=1}^T \hat{u}_t^2\right] = \text{Var}(\hat{\beta}_1) \sum_{t=1}^T x_{1t}^2 + 2 \text{cov}(\hat{\beta}, \hat{\delta}) \sum_{t=1}^T x_{1t} I_t^\tau + \text{Var}(\hat{\delta}) \sum_{t=1}^T I_t^{\tau^2} + 2E\left[(\beta_1 - \hat{\beta}_1) \sum_{t=1}^T x_{1t} u_t\right] + 2E\left[(\delta - \hat{\delta}) \sum_{t=1}^T I_t^\tau u_t\right] + Tc \quad (3.42)$$

expresión que variará en función del tipo de atípico que se considere.

Además, con independencia del tipo de atípico que se deba modelizar, tendremos que:

$$E\left[(\hat{\beta}_1 - \beta_1) \sum_{t=1}^T x_{1t} u_t\right] = \sigma_u^2$$

$$E\left[(\hat{\delta} - \delta) \sum_{t=1}^T I_t^* u_t\right] = \sigma_u^2$$

- **Si el atípico es del tipo impulso**, considerando los valores de las varianzas para los parámetros $\hat{\beta}_1$ y $\hat{\delta}$, obtenidas en las expresiones (3.8) y (3.10) y considerando que la covarianza entre ambos parámetros es:

$$\text{Cov}(\hat{\beta}_1, \hat{\delta}) = - \frac{\sigma_u^2 x_{1\tau}}{\sum_{t=1}^T x_{1t}^2 \left[1 - \frac{(x_{1\tau})^2}{\sum_{t=1}^T x_{1t}^2} \right]} = - \frac{\sigma_u^2 x_{1\tau}}{\sum_{t=1}^T x_{1t}^2 (1-a)} \quad (3.43)$$

y expresando las varianzas de los estimadores en el modelo verdadero mediante las relaciones siguientes:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_u^2}{\sum_{t=1}^T x_{1t}^2 (1-a)}$$

$$\text{Var}(\hat{\delta}) = \frac{\sigma_u^2}{(1-a)}$$

se demuestra entonces fácilmente que la esperanza de la suma residual es:

$$E\left[\sum_{t=1}^T \hat{u}_t^2\right] = (T-2)\sigma_u^2 \quad (3.44)$$

de manera que se puede definir un estimador insesgado de la varianza residual mediante la expresión:

$$\hat{\sigma}_u^2 = \frac{\sum_{t=1}^T \hat{u}_t^2}{(T-2)} \quad (3.45)$$

Sin embargo, la esperanza de la suma residual para el modelo erróneo puede escribirse de la siguiente forma:

$$E\left[\sum_{t=1}^T \tilde{u}_t^2\right] = E\left[(\beta_1 - \tilde{\beta}_1)^2\right] \sum_{t=1}^T x_{1t}^2 + 2E\left[(\beta_1 - \tilde{\beta}_1)\right] \delta \sum_{t=1}^T I_t^r x_{1t} + \delta^2 \sum_{t=1}^T I_t^{r^2} + 2E\left[(\beta_1 - \tilde{\beta}_1) \sum_{t=1}^T x_{1t} u_t\right] + 2\delta \sum_{t=1}^T I_t^r u_t + E\left[\sum_{t=1}^T u_t^2\right] \quad (3.46)$$

que para un atípico tipo impulso y considerando que:

$$E\left[(\tilde{\beta}_1 - \beta_1)^2 \sum_{t=1}^T x_{1t}^2\right] = \delta^2 a + \sigma_u^2$$

$$E[(\tilde{\beta}_1 - \beta_1) \delta x_{1t}] = \delta^2 a$$

$$E\left[(\tilde{\beta}_1 - \beta_1) \sum_{t=1}^T x_{1t} u_t\right] = \sigma_u^2$$

entonces vendrá recogida por la siguiente expresión:

$$E\left[\sum_{t=1}^T \tilde{u}_t^2\right] = (T-1)\sigma_u^2 + \delta^2 (1-a) \quad (3.47)$$

lo que claramente proporciona un estimador sesgado de la varianza residual cuando el modelo que se utiliza no incluye la modelización de un atípico tipo impulso.

Si consideramos la estimación MCO de la varianza residual, tendremos:

$$\tilde{\sigma}_u^2 = \frac{\sum_{t=1}^T \tilde{u}_t^2}{T-1} \Rightarrow E[\tilde{\sigma}_u^2] = \sigma_u^2 + \frac{\delta^2(1-a)}{T-1} \quad (3.48)$$

siendo pues el sesgo de la estimación de la varianza residual:

$$\text{sesgo}(\tilde{\sigma}_u^2) = \frac{\delta^2(1-a)}{(T-1)} \quad (3.49)$$

sesgo que depende del parámetro asociado a la variable ficticia en el modelo verdadero, δ , y del parámetro a definido en (3.22).

- **Si el atípico es del tipo escalón**, la covarianza entre los estimadores $\hat{\beta}_1$ y $\hat{\delta}$ vendrá dada por la expresión:

$$\text{Cova}(\hat{\beta}_1, \hat{\delta}) = -\frac{\sigma_u^2 \sum_{t=\tau}^T x_{1t}}{(T-\tau+1) \sum_{t=1}^T x_{1t}^2 - \left(\sum_{t=\tau}^T x_{1t} \right)^2} = -\frac{\sigma_u^2 \sum_{t=\tau}^T x_{1t}}{(T-\tau+1) \sum_{t=1}^T x_{1t}^2 (1 - \bar{x}_{1\tau} a)} \quad (3.50)$$

siendo las expresiones para las varianzas las consideradas en (3.12) y (3.14), considerando además que:

$$\begin{aligned} E[(\beta_1 - \hat{\beta}_1)] &= -\sigma_u^2 \\ E[(\delta - \hat{\delta})] &= -\sigma_u^2 \end{aligned}$$

la expresión para la esperanza de la suma residual en el modelo verdadero será:

$$E\left[\sum_{t=1}^T \hat{u}_t^2\right] = \frac{\sigma_u^2(T-\tau+1) \sum_{t=1}^T x_{1t}^2 + \sigma_u^2(T-\tau+1) \sum_{t=1}^T x_{1t}^2 - 2\sigma_u^2 \left(\sum_{t=\tau}^T x_{1t} \right)^2}{(T-\tau+1) \sum_{t=1}^T x_{1t}^2 - \left(\sum_{t=\tau}^T x_{1t} \right)^2} - 4\sigma_u^2 + T\sigma_u^2 \quad (3.51)$$

De nuevo operando llegamos a la expresión de la esperanza de la suma residual en el caso de un atípico tipo escalón para el modelo verdadero, considerando siempre la definición del parámetro a de (3.22)

$$E\left[\sum_{t=1}^T \hat{u}_t^2\right] = (T-2)\sigma_u^2 \quad (3.52)$$

de manera que un estimador insesgado de la varianza residual para el modelo verdadero cuando el atípico considerado es del tipo escalón será:

$$\hat{\sigma}_u^2 = \frac{\sum_{t=1}^T \hat{u}_t^2}{(T-2)} \quad (3.53)$$

Por otro lado, considerando, para el atípico tipo escalón que:

$$E\left[(\tilde{\beta}_1 - \beta_1)^2 \sum_{t=1}^T x_{1t}^2\right] = \delta^2 a \sum_{t=\tau}^T x_{1t} + \sigma_u^2$$

$$E\left[(\beta_1 - \tilde{\beta}_1) \delta \sum_{t=\tau}^T x_{1t}\right] = -\delta^2 a \sum_{t=\tau}^T x_{1t} \quad y$$

$$E\left[(\beta_1 - \tilde{\beta}_1) \sum_{t=1}^T x_{1t} u_t\right] = -\sigma_u^2$$

la expresión para la esperanza de la suma residual en el modelo erróneo en el caso de la omisión de un atípico del tipo escalón será:

$$E\left[\sum_{t=1}^T \tilde{u}_t^2\right] = \delta^2 a \sum_{t=\tau}^T x_{1t} + \sigma_u^2 - 2\delta^2 a \sum_{t=\tau}^T x_{1t} - 2\sigma_u^2 + \delta^2 (T - \tau + 1) + T\sigma_u^2 \quad (3.54)$$

expresión en la cual realizando ciertas operaciones nos quedará:

$$E\left[\sum_{t=1}^T \tilde{u}_t^2\right] = (T-1)\sigma_u^2 + \delta^2 (T-\tau+1)[1-a.\bar{x}_\tau] \quad (3.55)$$

donde a viene definida en este caso por la expresión (3.33). De esta forma, el estimador MCO de la varianza residual será:

$$\hat{\sigma}_u^2 = \frac{\sum_{t=1}^T \tilde{u}_t^2}{(T-1)} \quad (3.56)$$

que será un estimador sesgado de la varianza residual, con expresión para el sesgo dada por:

$$\text{sesgo}(\hat{\sigma}_u^2) = \frac{\delta^2 (T-\tau+1)(1-a.\bar{x}_\tau)}{(T-1)} \quad (3.57)$$

donde de nuevo depende del parámetro asociado a la variable ficticia en el modelo verdadero y del parámetro a .

Conclusiones de los efectos de los valores atípicos sobre los estadísticos del MLG

El error de especificación cometido al ignorar la modelización de un valor atípico en un modelo de regresión tiene como consecuencias:

- 1.- Los estimadores obtenidos serán estimadores sesgados, viniendo el sesgo determinado por la relación entre las variables exógenas incluidas en el modelo y las variables ficticias que recogen el comportamiento de los valores atípicos.
- 2.- Al considerar el cociente de los errores cuadráticos medios de los estimadores en un modelo verdadero (con la modelización de los valores atípicos) y en un modelo erróneo (sin la modelización de los valores atípicos), se puede decir que en la mayor

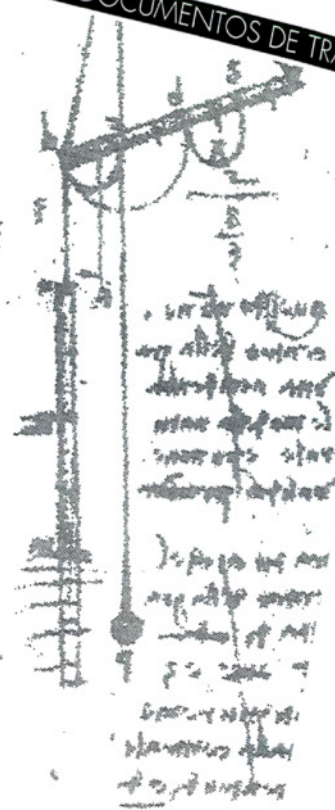
Los tres tipos de
 errores que se cometen al
 analizar los datos de un
 experimento son: errores
 de observación, errores
 de medición y errores
 de interpretación. Los
 errores de observación
 son los que se cometen
 al registrar los datos
 del experimento. Los
 errores de medición
 son los que se cometen
 al medir las variables
 del experimento. Los
 errores de interpretación
 son los que se cometen
 al interpretar los
 resultados del experimento.

Los errores de observación
 pueden ser evitados
 tomando algunas
 precauciones al registrar
 los datos. Los errores
 de medición pueden
 ser evitados tomando
 algunas precauciones
 al medir las variables.
 Los errores de interpretación
 pueden ser evitados
 tomando algunas
 precauciones al interpretar
 los resultados.

Los errores de observación
 pueden ser evitados
 tomando algunas
 precauciones al registrar
 los datos. Los errores
 de medición pueden
 ser evitados tomando
 algunas precauciones
 al medir las variables.
 Los errores de interpretación
 pueden ser evitados
 tomando algunas
 precauciones al interpretar
 los resultados.

4.- DETECCIÓN DE LOS VALORES ATÍPICOS

Los valores atípicos son
 aquellos valores que
 se desvían mucho de
 los demás valores de
 una muestra. Los
 valores atípicos pueden
 ser detectados
 utilizando algunas
 técnicas estadísticas.
 Una de las técnicas
 más comunes para
 detectar valores atípicos
 es el uso del coeficiente
 de variación. El
 coeficiente de variación
 es una medida de la
 dispersión de los
 datos. Si el coeficiente
 de variación es muy
 alto, esto indica que
 hay muchos valores
 atípicos en la muestra.



Los valores atípicos
 pueden ser detectados
 utilizando algunas
 técnicas estadísticas.
 Una de las técnicas
 más comunes para
 detectar valores atípicos
 es el uso del coeficiente
 de variación. El
 coeficiente de variación
 es una medida de la
 dispersión de los
 datos. Si el coeficiente
 de variación es muy
 alto, esto indica que
 hay muchos valores
 atípicos en la muestra.

4.- DETECCIÓN DE VALORES ATÍPICOS

Los trabajos realizados sobre la detección de valores atípicos, tanto en series temporales como en el modelo lineal general, centran especialmente su atención en la importancia de encontrar contrastes o métodos de detección que eviten los dos principales problemas que se presentan: el problema del enmascaramiento de atípicos, que aparece cuando los efectos de un valor atípico enmascaran los de otros y el problema de la inundación que surge al considerar una observación como atípico cuando realmente no lo es.

Los contrastes derivados a lo largo de las investigaciones han puesto de manifiesto la dificultad de evitar estos dos problemas, no habiéndose conseguido en todos los casos. Por ello, en esta sección expondremos los principales métodos propuestos, centrándonos en los desarrollos más recientes.

Tal y como se puso de manifiesto en las secciones anteriores, el problema de los valores atípicos puede provocar efectos de magnitud importante por lo que la detección de los valores atípicos se convierte así en una etapa indispensable en todo análisis de un modelo de regresión.

En el modelo de regresión, al plantearse la explicación de una variable, Y , en función de un conjunto de variables explicativas, incluidas en la matriz X , nos podremos encontrar con atípicos tanto en Y como X . En este sentido, se han definido los **puntos influyentes** ("leverage points") **en dirección y** y **en dirección x** , debido a que puntos elevados en cualquiera de las variables incluidas en el modelo pueden provocar la aparición de valores atípicos que distorsionan claramente la estimación y los estadísticos asociados a ésta.

Se han desarrollado diferentes métodos de detección de atípicos que podremos englobar en las siguientes categorías:

- 1.- Métodos basados en los residuos.
- 2.- Métodos de eliminación de una o múltiples filas

3.- Métodos basados en la distancia de Mahalanobis

4.- Métodos gráficos

Dentro de cada una de las categorías anteriores, se han propuesto diferentes métodos de detección de atípicos. Aquí simplemente nos referiremos a los métodos más habituales y a los de desarrollo más reciente, no pretendiendo por tanto una enumeración exhaustiva de todos los contrastes derivados.

Asimismo, la clasificación anterior de los contrastes es una de las muchas que se hubiesen podido obtener. Obedece únicamente al juicio de la autora por razones de simplicidad expositiva. Además, los contrastes que se presentan podrían haber sido incluidos en otra categoría diferente debido a que muchos de ellos presentan características comunes.

4.1.- MÉTODOS BASADOS EN LOS RESIDUOS

En el modelo de regresión:

$$Y = X\beta + u$$

donde $Y = (y_1, y_2, \dots, y_T)'$, X es la matriz de regresores de dimensión $T \times k$ y $u = (u_1, u_2, \dots, u_T)'$ es un vector de errores aleatorios con $E[u] = 0$ y $V(u) = \sigma_u^2 I_T$, tendremos los siguientes resultados básicos.

Definimos la matriz H ("hat matrix") de la siguiente forma:

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

siendo la matriz de proyección que transforma el vector Y en su estimador mínimo cuadrático: $\hat{Y} = HY$, donde H es una matriz simétrica e idempotente y $\text{tr}(H) = k$. Gracias a estas propiedades podemos escribir:

$$h_{ii} = \sum_{j=1}^T h_{ij}^2 \quad \forall i$$

donde h_{ii} es el i -ésimo elemento de la diagonal principal de la matriz H .

Si definimos los residuos mínimo cuadráticos $\hat{u} = Y - \hat{Y}$, podemos expresar las covarianzas de $\hat{Y}$ y $\hat{u}$ de la siguiente forma:

$$\begin{aligned}\text{Cov}(\hat{Y}) &= \sigma^2 H \\ \text{Cov}(\hat{u}) &= \sigma^2 (I - H)\end{aligned}$$

La matriz de proyección H se ha utilizado en numerosas ocasiones para determinar la influencia de determinados puntos de observaciones. En concreto, se analizan los elementos de la diagonal principal de la matriz H , considerando como influyentes aquellos elementos en los que h_{ii} sean mayores que $2k/T$ (en otras ocasiones, $3k/T$). Sin embargo, tal y como ponen de manifiesto Rousseeuw y Leroy (1987), este método no tiene en consideración las posibles observaciones influyentes en Y debido a la forma de construir la matriz H que sólo considera las variables en X .

Por este motivo, se desarrollan los métodos basados en los residuos, que son:

- Residuos estandarizados, definidos por:

$$s_i = \frac{e_i}{S} \quad (4.1.1)$$

$$\text{siendo } S^2 = \frac{1}{T-k} \sum_{j=1}^T e_j^2$$

- Residuos estudentizados definidos por:

$$t_i = \frac{e_i}{S\sqrt{1-h_{ii}}} \quad (4.1.2)$$

el cual, según Belsley, Kuh y Welsch (1980), en un gran número de situaciones prácticas, se distribuye como una t con $T-k-1$ grados de libertad, pudiéndose interpretar como un medio para medir la significatividad de un residuo estudentizado.

De esta forma, los residuos estudentizados proveen un buen camino para examinar la información en los residuos tanto porque tienen la misma varianza como por su fácil relación, en numerosas ocasiones, con la distribución t . Valores elevados de (4.1.1) y (4.1.2) en valor

absoluto determinarían la presencia de puntos influyentes que deberían ser analizados como posibles atípicos. Según Rousseeuw y Leroy (1987), un residuo estandarizado o estudentizado superior a 2.5 (considerado como punto de referencia debido a la gaussianidad supuesta en numerosos casos prácticos) podría establecerse como un valor atípico a analizar.

Los residuos estandarizados y en especial los residuos estudentizados han sido los métodos más ampliamente utilizados en las investigaciones para detectar valores atípicos. Sin embargo, en numerosas situaciones, nos encontramos con que puntos influyentes tienen un residuo estudentizado pequeño. En problemas en los que hay múltiples atípicos en los datos de las variables X, ni los residuos estandarizados ni los estudentizados detectan los atípicos debido al desplazamiento que se produce en el ajuste mínimo cuadrático. De esta forma, se produce el efecto de enmascaramiento de atípicos. (Véase Rousseeuw y Leroy (1987) para gran número de ejemplos).

Rousseeuw (1985) propone otro método para detectar múltiples atípicos basado en los residuos que denomina **diagnóstico resistente (RD_i)**, que busca identificar todos los puntos que son atípicos. La idea básica de este diagnóstico está estrechamente relacionada con la definición de la regresión de mediana de cuadrados mínima (LMS), consistente en:

$$\text{Minimizar}_{\beta} \text{med}_i \hat{u}_i^2 \quad (4.1.3)$$

para cada subconjunto $J = \{i_1, i_2, \dots, i_k\}$ de $(1, 2, \dots, T)$ conteniendo k subíndices. Se obtienen así $\hat{\beta}_J$ y los residuos correspondientes $\hat{u}_i(\hat{\beta}_J)$, a partir de los cuales se construyen los residuos estandarizados mediante la expresión:

$$\frac{\hat{u}_i(\hat{\beta}_J)}{\text{med}_{j=1, \dots, T} |\hat{u}_j(\hat{\beta}_J)|}$$

A partir de esta expresión, el caso i -ésimo se define como atípico mediante la expresión:

$$P_i = \max_j \frac{|\hat{u}_i(\hat{\beta}_J)|}{\text{med}_{j=1, \dots, T} |\hat{u}_j(\hat{\beta}_J)|} \quad (4.1.4)$$

De esta forma, las observaciones con P_i elevado se considerarán aberrantes o al menos sospechosas. El problema surge en ¿qué se puede considerar elevado?. Rousseeuw y Leroy (1987) consideran que el punto i -ésimo se identifica como atípico si su P_i es mayor que un valor crítico C_n , el cual podrá derivarse a partir de la distribución de P_i bajo condiciones de normalidad. Desgraciadamente, es muy difícil derivar C_n que depende también de las variables X . Por ello, propone una versión estandarizada de P_i , definiendo el **diagnóstico resistente**:

$$RD_i = \frac{P_i}{\text{med } P_j}_{j=1, \dots, T} \quad (4.1.5)$$

para cada observación i . Esta expresión únicamente se puede aplicar cuando la mediana de todos los P_j es diferente de 0, hecho que únicamente ocurre cuando el ajuste es exacto, es decir, cuando más de la mitad de las observaciones cae en el mismo hiperplano, hecho que se descubre realizando una estimación robusta, como la LMS. Su experiencia les lleva a asegurar que RD_i se puede considerar elevado si su valor excede de 2.50.

Según Rousseeuw y Leroy (1987), por construcción, el diagnóstico RD_i no es susceptible del efecto de enmascaramiento y sugieren que el RD_i debe interpretarse conjuntamente con los residuos estandarizados de un ajuste muy robusto, pudiendo concluirse que todos los puntos con un residuo estandarizado elevado también poseen un RD_i grande, no siendo cierto el inverso. Puede suceder que una o más observaciones con $RD_i > 2.50$ tengan residuos estandarizados de un ajuste robusto reducidos. Tales observaciones son "puntos de influencia inofensivos", que no influyen en el ajuste.

4.2.- MÉTODOS BASADOS EN LA ELIMINACIÓN DE UNA O MÚLTIPLES FILAS

Estos métodos buscan valorar la influencia de una ó múltiples observaciones en los estadísticos derivados a partir del modelo de regresión.

Se fundamentan en la idea de determinar si la eliminación de una ó múltiples filas afecta de forma sustancial a los estadísticos. Si esto ocurre, podremos determinar una ó múltiples observaciones influyentes que deberán ser analizadas como posibles atípicos.

En la primera parte de esta sección, analizaremos la influencia de la eliminación de una única fila a partir de los estadísticos propuestos por Cook (1977) y Belsley, Kuh y Welsch (1980).

4.2.1.- ELIMINACIÓN DE UNA ÚNICA FILA (EN X Y EN Y)

Cook (1977) busca detectar observaciones influyentes en el modelo de regresión lineal. Para ello, basándose en la definición de la elipsoide de confianza al $(1-\alpha)\%$, la cual si consideramos normalidad para el vector desconocido β , viene dada por el conjunto de todos los vectores β^* que satisfacen:

$$\frac{(\beta^* - \hat{\beta})' X' X (\beta^* - \hat{\beta})}{kS^2} \leq F_{(k, T-k)}(1-\alpha)$$

siendo $\hat{\beta} = (X'X)^{-1}X'Y$, $S^2 = \frac{\hat{u}'\hat{u}}{T-k}$ y $F_{(k, T-k)}(1-\alpha)$ el valor de la distribución F con k y $T-k$ grados de libertad al $1-\alpha$, donde k es el número de parámetros del modelo.

Para determinar el grado de influencia que el i -ésimo dato tiene en el estimador $\hat{\beta}$, el camino más natural es calcular el estimador mínimo cuadrático (EMCO) de β eliminando dicha observación, estimador que denotaremos $\hat{\beta}_{(-i)}$. De esta forma, Cook (1977) considera que la distancia entre ambos estimadores puede entenderse como una buena manera de medir esa influencia. Define así una medida de la naturaleza crítica de cada punto mediante la expresión:

$$D_i \equiv \frac{(\hat{\beta}_{(-i)} - \hat{\beta})' X' X (\hat{\beta}_{(-i)} - \hat{\beta})}{kS^2} \quad i = 1, 2, \dots, T \quad (4.2.1.1)$$

estadístico que provee una medida de la distancia entre $\hat{\beta}_{(-i)}$ y $\hat{\beta}$ en términos de niveles de significación descriptivos. Si suponemos que $D_i \sim F_{(k, T-k)}(1-\alpha)$, un dato se considera influyente si $\hat{\beta}_{(-i)}$ estuviese fuera de la región de confianza del 10% (Cook (1977)). Cook y Weisberg (1982) y Montgomery y Peck (1982) consideran que un valor de D_i alrededor de 1.0 sería generalmente considerado como grande.

Con el objeto de evitar realizar (T-1) regresiones eliminando cada vez una observación, Cook (1977) expresa D_i de una forma más sencilla y a la vez clarificadora. Para ello, considera ciertas derivaciones. Se demuestra fácilmente que:

$$\hat{\beta} - \hat{\beta}_{(-i)} = (X'_{(-i)} X_{(-i)})^{-1} x_i [y_i - x_i' \hat{\beta}] \quad (4.2.1.2)$$

donde $X_{(-i)}$ es la matriz X con la i -ésima fila (x_i') eliminada, y_i es la i -ésima observación del vector Y . Si definimos $h_{ii} = x_i' (X' X)^{-1} x_i$ y asumimos $v_i < 1$ (lo cual se produce cuando $X_{(-i)}$ es de rango completo) entonces:

$$(X'_{(-i)} X_{(-i)})^{-1} = (X' X)^{-1} + (X' X)^{-1} x_i x_i' (X' X)^{-1} / (1 - h_{ii}) \quad (4.2.1.3)$$

de lo que se deduce que:

$$\frac{(X' X)^{-1} x_i}{1 - h_{ii}} = (X'_{(-i)} X_{(-i)})^{-1} x_i \quad (4.2.1.4)$$

Sustituyendo (4.2.1.4) en (4.2.1.2):

$$\hat{\beta} - \hat{\beta}_{(-i)} = \frac{(X' X)^{-1} x_i}{1 - h_{ii}} [y_i - x_i' \hat{\beta}] \quad (4.2.1.5)$$

a partir de la cual se desprende inmediatamente que:

$$D_i = \left[\frac{y_i - x_i' \hat{\beta}}{S \sqrt{1 - h_{ii}}} \right]^2 \frac{h_{ii}}{k(1 - h_{ii})} \quad (4.2.1.6)$$

De esta forma, D_i depende de tres cantidades relevantes, relacionadas todas ellas con el conjunto de datos completo: el número de parámetros, k , el residuo estudentizado i -ésimo:

$$t_i = \left[\frac{y_i - x_i' \hat{\beta}}{S \sqrt{1 - h_{ii}}} \right]$$

y el ratio entre la varianza del i -ésimo valor predicho:

$$V(\hat{y}) = x_i' (X'X)^{-1} x_i \sigma^2 = h_{ii} \sigma^2$$

y la varianza del i-ésimo residuo:

$$V(\hat{u}_i) = \sigma^2(1 - h_{ii})$$

De esta forma, se puede expresar D_i más simplemente:

$$D_i = \frac{t_i^2}{k} \cdot \frac{V(\hat{y}_i)}{V(\hat{u}_i)} \quad (4.2.1.7)$$

Claramente, t_i^2 es una medida del grado con el que la observación i-ésima puede ser considerada como un atípico a partir del modelo asumido. El cociente de varianzas mide la sensibilidad relativa del estimador $\hat{\beta}$ a atípicos potenciales en cada punto. Un valor elevado de este cociente indica que el punto asociado tiene una fuerte ponderación en la determinación de los $\hat{\beta}$. Ambas medidas, combinadas en (4.2.1.7) proporcionan una medida del impacto total que una única observación tiene sobre la solución mínimo cuadrática.

El problema de esta medida surge cuando $\rho(X_{(-i)}) < k$, de forma que $\hat{\beta}_{(-i)}$ no está esencialmente definido.

A partir de esta medida, Belsley, Kuh y Welsch (1980) desarrollan un conjunto de contrastes con el objeto de valorar los efectos que la eliminación de una fila tiene sobre los estadísticos más usuales en el análisis de regresión.

El efecto de eliminar una fila sobre el vector de coeficientes estimados se analiza en Belsley, Kuh y Welsch (1980) mediante:

$$DFBETA_i \equiv \hat{\beta} - \hat{\beta}_{(-i)} = \frac{(X'X)^{-1} x_i' \hat{u}_i}{1 - h_{ii}} \quad (4.2.1.8)$$

Si el interés se centra en el efecto sobre un único elemento del vector de parámetros estimados, el j-ésimo, se valora en relación a la varianza de $\hat{\beta}_j, \sigma^2 (X'X)^{-1}_{jj}$. Si definimos $C = (X'X)^{-1} X'$, entonces:

$$\hat{\beta} - \hat{\beta}_{(-i)j} = \frac{c_{ji} \hat{u}_i}{1 - h_{ii}} \quad (4.2.1.9)$$

Puesto que: $\sum_{i=1}^T (X'X)^{-1} x_i' x_i (X'X)^{-1} = (X'X)^{-1}$

Mosteller y Tukey (1977) demuestran que:

$$\text{Var}(\hat{\beta}_j) = \sigma^2 \sum_{k=1}^T c_{jk}^2 \quad (4.2.1.10)$$

De esta forma, una medida del cambio que se produce en el parámetro $\hat{\beta}_j$ al eliminar la i -ésima fila se define mediante la expresión:

$$\text{DFBETAS}_{ij} \equiv \frac{\hat{\beta}_j - \hat{\beta}_{(-i)j}}{S_{(-i)} \sqrt{(X'X)^{-1}_{jj}}} = \frac{c_{ji}}{\sqrt{\sum_{k=1}^T c_{jk}^2}} \cdot \frac{\hat{u}_i}{S_{(-i)} (1 - h_{ii})} \quad (4.2.1.11)$$

donde Belsley et al (1980), con el objeto de hacer el denominador estocásticamente independiente del numerador en el caso de gaussianidad, consideran en vez de S^2 (el estimador usual de σ^2), $S_{(-i)}^2$ definido como:

$$S_{(-i)}^2 = \frac{1}{T - k - 1} \sum_{k \neq i} [y_k - x_k \hat{\beta}_{(-i)}]^2$$

En el caso general, si $|\text{DFBETAS}_{ij}|$ es elevado, se considerará que la observación i -ésima es influyente en la determinación del parámetro $\hat{\beta}_j$. Tanto Belsley et al (1980) como Rousseeuw y Leroy (1987) consideran que si $|\text{DFBETAS}_{ij}| > 2/\sqrt{T}$, la observación i -ésima se considerará influyente debiendo ser analizada.

Bajo la hipótesis de normalidad, Belsley et al (1980) consideran un contraste útil para analizar el efecto sobre el estadístico t , valorándose así la sensibilidad del output de regresión ante la eliminación de una fila. En este sentido, definen:

$$\text{DFTSTAT}_{ij} \equiv \frac{\hat{\beta}_j}{S\sqrt{(\mathbf{X}'\mathbf{X})_{jj}^{-1}}} - \frac{\hat{\beta}_{(-i)j}}{S_{(-i)}\sqrt{[\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}]_{jj}^{-1}}} \quad (4.2.1.12)$$

estadístico que ayudará a determinar si la eliminación de una fila afectará de forma importante a las conclusiones del contraste de hipótesis.

El efecto que la eliminación de una fila tiene sobre los valores ajustados se mide a través del contraste:

$$\text{DFFITS}_i = \hat{y}_i - \hat{y}_{(-i)i} \equiv \left[\frac{h_{ii}}{1 - h_{ii}} \right]^{1/2} \cdot \frac{\hat{u}_i}{S_{(-i)}\sqrt{1 - h_{ii}}} \quad (4.2.1.13)$$

expresión muy similar a D_i en (4.1.b.1.6) donde se emplea S en vez de $S_{(-i)}$. Según Rousseeuw y Leroy (1987) y Belsley et al (1980), una observación que cumpla: $\text{DFFITS}_i > 2(k/T)^{1/2}$ debería ser analizada como influyente.

Con objeto de valorar el efecto de la eliminación de una fila sobre los residuos, Belsley et al (1980) definen los residuos estudentizados de la siguiente forma:

$$\text{RSTUDENT} = \hat{u}_i^* \equiv \frac{\hat{u}_i}{S_{(-i)}\sqrt{1 - h_{ii}}} \quad (4.2.1.14)$$

diferenciándose de la expresión anterior en (4.1.2) en el uso de $S_{(-i)}$ en vez de S , presentando los mismos problemas anteriormente comentados.

Finalmente, Belsley et al (1980) analizan los efectos de la eliminación de una fila sobre la matriz de covarianzas de los estimadores, proponiendo dos formas de analizarlos.

En primer lugar, proponen realizar el cociente de los determinantes:

$$\det[(\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)})^{-1}] / \det[(\mathbf{X}'\mathbf{X})^{-1}]$$

sugiriendo que, al diferir ambas matrices únicamente en la inclusión y exclusión de una sola fila, un valor de este ratio cercano a la unidad indicará que ambas matrices son muy cercanas, reflejando que la matriz de covarianzas es insensible a la eliminación de la fila i -ésima.

En segundo lugar, y con objeto de considerar además los efectos de la eliminación de una fila sobre la variable y , mediante la estimación de su varianza, proponen el siguiente estadístico:

$$\begin{aligned} \text{COVRATIO} &\equiv \frac{\det\{S_{(-i)}^2[(X'_{(-i)}X_{(-i)})^{-1}]\}}{\det\{S^2(X'X)^{-1}\}} \\ &= \frac{S_{(-i)}^{2k}}{S^{2k}} \cdot \left[\frac{\det[(X'_{(-i)}X_{(-i)})^{-1}]}{\det[(X'X)^{-1}]} \right] \end{aligned} \quad (4.2.1.15)$$

el cual al considerar que:

$$\begin{aligned} * \det[X'_{(-i)}X_{(-i)}] &= (1 - h_{ii}) \det(X'X) \\ *(T - k - 1)S_{(-i)}^2 &= (T - k)S^2 - \frac{i^2}{1 - h_{ii}} \\ * \frac{i}{S_{(-i)}} &= \frac{i}{\sqrt{1 - h_{ii}}} \end{aligned}$$

puede expresarse:

$$\text{COVRATIO} \equiv \frac{1}{\left[\frac{(T - k - 1)}{T - k} + \frac{i^2}{T - k} \right]^k \cdot (1 - h_{ii})} \quad (4.2.1.16)$$

Como herramienta de diagnóstico, el interés se centrará en aquellas observaciones de (4.2.1.16) cuyos valores no sean cercanos a la unidad, siendo estas observaciones las que probablemente sean más influyentes y necesiten una investigación más profunda. Belsley et al (1980) analizan los casos extremos que se pueden producir, deduciendo que deberán analizarse los puntos con $|\text{COVRATIO} - 1|$ cercanos o mayores que $3k/T$.

Por último, definen un estadístico para analizar el efecto de la eliminación de una fila sobre la varianza de $\hat{y}_i$. Partiendo de:

$$\text{Var}(\hat{y}_i) = S^2 h_{ii} \quad \text{y}$$

$$\text{Var}(\hat{y}_{(-i)i}) = \text{Var}(x_i \hat{\beta}_{(-i)}) = S_{(-i)}^2 \left[\frac{h_{ii}}{1 - h_{ii}} \right]$$

forman el cociente:

$$\text{FVARATIO} = \frac{S_{(-i)}^2}{S^2(1 - h_{ii})}$$

ratio que exhibe los mismos esquemas de comportamiento que el COVRATIO.

Rousseeuw y Leroy (1987) analizan estos estadísticos para medir la influencia de una fila. Ponen de manifiesto que estas medidas se definen en función de los residuos MCO y de los elementos de la diagonal de las matrices H y C, conteniendo así dos tipos de información: los residuos MCO, e_i , que son medidas del mal ajuste y las matrices H y C que reflejan la distancia o alejamiento de x_i en el espacio de las variables explicativas. Sin embargo, su interpretación no es tan sencilla cuando los datos contienen más de un caso atípico. Estos estadísticos son susceptibles del efecto enmascaramiento puesto que dos o más atípicos pueden actuar conjuntamente para reforzar o cancelar la influencia de los otros. El resultado final es que estos estadísticos de un único caso no tienen éxito en la identificación de atípicos, fracasando a menudo en revelar el impacto de pequeños grupos de casos porque la influencia de un punto podría estar enmascarada por otro.

4.2.2.- ELIMINACIÓN DE MÚLTIPLES FILAS

Ante la posibilidad del enmascaramiento de atípicos, se plantea la necesidad de encontrar técnicas que sean capaces de detectar los efectos potencialmente influyentes de un grupo de observaciones.

En este sentido, Cook y Weisberg (1982) generalizan la expresión (4.2.1.1) para medir el efecto de más de un caso, obteniendo:

$$D_i(I) \equiv \frac{(\hat{\beta}_{(-I)} - \hat{\beta})' X' X (\hat{\beta}_{(-I)} - \hat{\beta})}{kS^2} \quad (4.2.2.1)$$

donde I recoge los índices correspondientes al grupo de casos potencialmente influyentes y $\hat{\beta}_{(-I)}$ es el estimador MCO eliminando las filas indexadas por I .

El problema fundamental de esta aproximación es la selección de los casos que se deben incluir en I . Según Belsley et al (1980) se pueden utilizar las técnicas expuestas anteriormente para seleccionar los casos a incluir en I . Sin embargo, tal y como exponen Rousseeuw y Leroy (1987), se puede producir el caso de que existan puntos conjuntamente influyentes pero que lo sean de forma separada, de manera que no consideran válidos los diagnósticos de un único caso para detectar el subconjunto a incluir en I . Esto lleva a que se tengan que calcular C_T^m combinaciones donde $m=1,2,\dots,T/2$, lo que dificulta extremadamente el cálculo.

Belsley et al (1980), al considerar que sus técnicas de eliminación de una fila son válidas, estiman que el subconjunto a incluir en I se puede detectar fácilmente y generalizan los estadísticos anteriores al caso de eliminación de múltiples filas.

La generalización del efecto sobre la estimación de un parámetro de la eliminación de m filas es:

$$\frac{|\hat{\beta}_j - \hat{\beta}_{(-I_m)j}|}{\text{Escala}} \quad (4.2.2.2)$$

para $j=1,2,\dots,k$ y $m=2,3,\dots$ donde escala indica una medida apropiada del error estándar.

Si el interés se centra en los valores ajustados, la medida se transforma en:

$$\frac{|x_r [\hat{\beta} - \hat{\beta}_{(-I_m)}]|}{\text{Escala}} \quad (4.2.2.3)$$

para $r=1,\dots,T$. Estas medidas también pueden expresarse:

$$\frac{[\hat{\beta}_j - \hat{\beta}_{(-I_m)j}][\hat{\beta}_j - \hat{\beta}_{(-I_m)j}]}{\text{Escala}} \quad (4.2.2.4)$$

$$\frac{[\hat{\beta} - \hat{\beta}_{(-I_m)}]' X' X [\hat{\beta} - \hat{\beta}_{(-I_m)}]}{\text{Escala}} \quad (4.2.2.5)$$

Los efectos sobre el ajuste pueden analizarse mediante la expresión:

$$\text{MDFIT} = [\hat{\beta} - \hat{\beta}_{(-I_m)}]' X'_{(-I_m)} X_{(-I_m)} [\hat{\beta} - \hat{\beta}_{(-I_m)}] \quad (4.2.2.6)$$

lo que también se puede expresar (Bingham (1977)):

$$I_m' X_{I_m} [X'_{(-I_m)} X_{(-I_m)}]^{-1} X'_{I_m} I_m \quad (4.2.2.7)$$

donde $\hat{u}$ es el vector columna de los residuos MCO e I_m como subíndice denota la matriz o el vector con filas cuyos índices están contenidos en I_m .

La generalización del cociente de varianzas para la eliminación de m filas es:

$$\text{COVRATIO}(I_m) \equiv \frac{\det\{S^2_{(-I_m)} (X'_{(-I_m)} X_{(-I_m)})^{-1}\}}{\det\{S^2 (X' X)^{-1}\}} \quad (4.2.2.8)$$

Tal y como señalan Rousseeuw y Leroy (1987) las técnicas de eliminación de múltiples filas no son tan populares como las de eliminación de una única fila debido a los problemas de cálculo. Estos vienen motivados por la dificultad de encontrar aquellas filas que forman el subconjunto I_m . Se han derivado medidas para aliviar este problema pero en ningún caso se garantiza la identificación de los casos "malos" presentes en el conjunto de datos.

4.3.- MÉTODOS BASADOS EN LA DISTANCIA DE MAHALANOBIS

La aproximación clásica para la detección de valores atípicos se realiza mediante la distancia de Mahalanobis al cuadrado, definida mediante la expresión:

$$MD_i^2(x_i, X) = (x_i - T(X))C^{-1}(X)(x_i - T(X)) \quad (4.3.1)$$

calculada para cada observación, siendo $T(X)$ la media aritmética y $C(X)$ el estimador clásico de la covarianza:

$$C(X) = \frac{1}{T} \sum_{i=1}^T (x_i - \bar{x})(x_i - \bar{x})' \quad (4.3.2)$$

Los puntos para los cuales $MD_i^2(x_i, X)$ es elevada se eliminan de la muestra y se procede de nuevo con la muestra "limpiada". Esta aproximación produce buenos resultados si únicamente hay un atípico, pero sufre, en los casos de múltiples atípicos, del efecto enmascaramiento. Los estadísticos como la media y la matriz de covarianzas son muy sensibles a la presencia de atípicos ya que son estadísticos no robustos.

Por estas razones, los métodos de detección de atípicos basados en el cuadrado de la distancia de Mahalanobis se han centrado en la obtención de estimaciones robustas de $T(X)$ y $C(X)$ para evitar los problemas anteriormente comentados.

Rousseeuw (1985) y Rousseeuw y Leroy (1987) se basan en dos técnicas de estimación robusta: Elipse de Volumen Mínimo (MVE) y Determinante de Covarianza Mínimo (MCD) que, para los parámetros a estimar en MD_i^2 se definen de la siguiente forma:

a.- El MVE para $T(X)$ se define como el centro de la elipse de volumen mínimo cubriendo al menos h puntos de X , donde $h = \lceil (T + k + 1) / 2 \rceil$ ²

El estimador de la matriz de covarianzas $C(X)$ correspondiente viene dado por la misma elipse, multiplicada por un factor adecuado para conseguir la consistencia con los datos normales multivariantes.

b.- El MCD, estimador de determinante de covarianza mínimo, para $T(X)$ se define como la media de los h puntos de X , para los cuales el determinante de la matriz de covarianzas es mínimo.

² La notación $\lceil . \rceil$ significa la parte entera de . .

Rousseeuw y van Zameren (1989) proponen reemplazar la media y la covarianza clásicas en MD_i^2 por sus MCD ó MVE robustos, resultando así distancias robustas que permiten identificar los puntos de influencia en los datos de regresión.

Rousseeuw y Leroy (1987) dicen que, en el análisis de regresión, las técnicas anteriores se aplicarían sobre la matriz de variables explicativas X con el objeto de encontrar los puntos de influencia. Sin embargo, conocer estos puntos no es suficiente ya que se tienen que tener también en cuenta la respuesta en y_i con el objeto de determinar si (y_i, x_i) es un atípico en el modelo de regresión (es decir, si (y_i, x_i) se desvía del esquema lineal de la mayoría de los datos). De esta forma, la cuestión de si x_i es un "buen" o "mal" punto de influencia únicamente puede averiguarse por medio de una técnica de regresión robusta como el LMS que comentaremos en la sección dedicada a la corrección de los valores atípicos.

Netkov y Neytchev (1991) proponen un procedimiento, basado en la técnica LMS y MVE, para distinguir entre "buenos" y "malos" puntos de influencia.

El procedimiento, para el modelo:

$$Y = X\beta + u$$

donde Y es un vector $(Tx1)$, X es una matriz (Txk) y u es $(Tx1)$ se compone de las siguientes etapas:

- 1.- Hacer la regresión LMS, es decir encontrar el estimador de mínima mediana de residuos al cuadrado:

$$\underset{\hat{\beta}}{\text{minimizar}} (\hat{u}^2)_{p,T} \quad (4.3.3)$$

donde $p = [T/2] + [(k+1)/2]$, $\hat{\beta}$ son los coeficientes de la regresión, $(\hat{u}^2)_{i,T}$, $i=1, \dots, T$, son los residuos al cuadrado ordenados de menor a mayor. Calcular los residuos estandarizados $\hat{u}_i / S$ donde S es un estimador de la desviación típica final y $\hat{u}_i$ son

los residuos basados en la regresión LMS. Identificar entonces la i -ésima observación como atípico LMS si $\left| \hat{u}_i / S \right| \geq 2.5$ (Rousseeuw y Leroy (1987)).

2.- Aplicar la técnica MVE sólo a las variables explicativas. Calcular la distancia robusta (RD_i)³ correspondiente para cada observación. Calcular el valor crítico $(\chi^2_{k,0.975})^{1/2}$.

3.- Dibujar los residuos estandarizados de la regresión LMS contra las RD_i calculadas en la segunda etapa. Los puntos de este gráfico caerán en las siguientes categorías:

- Observaciones regulares, con reducida RD_i $\left[RD_i < (\chi^2_{k,0.975})^{1/2} \right]$ y residuos estandarizados pequeños $\left[\hat{u}_i / S \leq 2.5 \right]$.
- Atípicos verticales (VO), con RD_i pequeña y $\hat{u}_i / S$ grande.
- "Buenos" puntos de influencia (GLP), con RD_i elevada y $\hat{u}_i / S$ reducidos.
- "Malos" puntos de influencia (BLP), con RD_i y $\hat{u}_i / S$ elevados.

De esta forma, se detectan los atípicos en ambas direcciones (Y y X), no estando sujeto este procedimiento a los efectos de enmascaramiento ni de inundación de atípicos.

Rousseeuw y van Zomeren (1990) ilustran gráficamente este procedimiento ya que ponen de manifiesto que la mejor forma de decidir sobre los "buenos" y "malos" puntos de influencia es mediante el análisis gráfico de los residuos robustos, $\hat{u}_i / S$, versus las distancias robustas RD_i .

En un primer gráfico de los valores de la variable endógena versus la variable exógena se pueden distinguir qué tipo de atípicos nos podemos encontrar. Así en la figura 1, las observaciones (a) serán observaciones regulares, la observación (b) se considera un atípico

³ Es la distancia de Mahalanobis de (4.1.c.1) sustituyendo $T(X)$ y $C(X)$ por las estimaciones MVE robustas.

vertical, el punto (c) es un buen punto de influencia y finalmente (d) se denomina mal punto de influencia.

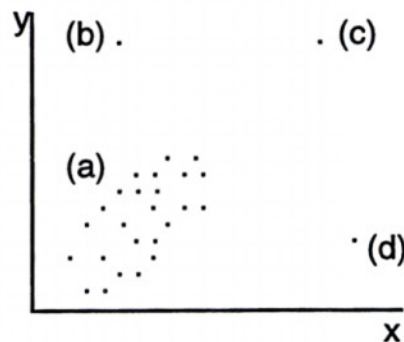


Figura 1: Determinación gráfica de los puntos de influencia

Sin embargo, el gráfico más ilustrativo es el que recoge los residuos robustos estandarizados contra las distancias robustas:

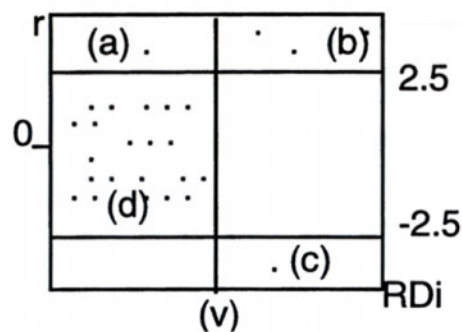


Figura 2: Residuos estandarizados versus distancias robustas

A partir de la figura 2, en la que se dibujan las distancias robustas frente a los residuos estandarizados, se pueden clasificar las observaciones en cuatro categorías. Antes de pasar a definir las clases de observaciones, hay que señalar que la recta (v) recoge el valor de $\sqrt{\chi^2_{k,0.975}}$. Las cuatro categorías en que se pueden dividir las observaciones se han denotado en la figura 2 y son las siguientes:

(a) Este punto se considera como **atípico vertical** ya que tiene una RD_i de valor reducido y un residuo estandarizado elevado.

(b) Las observaciones incluidas en este cuadrado se denominan **malos puntos de influencia** ya que presentan una distancia robusta elevada y un $\hat{u}/S$ elevada.

(c) El punto (c) presenta una distancia robusta grande y sin embargo un residuo estandarizado reducido por lo que se denomina **buen punto de influencia**.

(d) Los puntos incluidos en el cuadrante (d) se consideran observaciones regulares ya que presentan tanto la distancia robusta como los residuos estandarizados reducidos.

Rousseeuw y van Zomeren (1990) destacan que un único diagnóstico no es suficiente para clasificar las observaciones en estas cuatro categorías y además revelan que este método es mucho más eficiente en la detección de los valores atípicos y puntos de influencia que el que otros autores han propuesto, consistente en dibujar los residuos mínimo cuadráticos estandarizados frente a la distancia de Mahalanobis. Neykov y Neytchev (1991) recomiendan la utilización del método de Rousseeuw y van Zomeren (1990) en todo análisis de regresión.

Por su parte, Hadi (1992) considera que la estimación MVE de los parámetros de la distancia de Mahalanobis propuesta por Rousseeuw (1985), aunque genera estimaciones de $T(X)$ y $C(X)$ robustas, plantea un coste de cálculo demasiado elevado. Por este motivo, se centra en la obtención de estimadores robustos para la media y la matriz de covarianzas de la distancia de Mahalanobis con un menor coste operativo.

El procedimiento se centra en la matriz de variables explicativas X , de orden $T \times k$, siendo las etapas a seguir:

- Etapla 0: Ordenación inicial.

Reordenar las T observaciones en forma ascendente de acuerdo con una distancia robusta elegida de forma adecuada. Por ejemplo:

$$D_i(C_R, S_R) = \sqrt{(x_i - C_R)' S_R^{-1} (x_i - C_R)} \quad i = 1, \dots, T \quad (4.3.4)$$

siendo C_R y S_R estimadores robustos de localización y de la matriz de varianzas-covarianzas. Se dividen entonces las observaciones en dos subconjuntos iniciales:

1.- Subconjunto básico: que contiene las $k+1$ primeras observaciones.

2.- Subconjunto no básico, conteniendo las T-k-1 observaciones restantes.

- Etapa 1(a): Conjunto básico de rango completo.

Si el conjunto básico es de rango completo, calcular:

$$\sqrt{(x_i - C_b)' S_b^{-1} (x_i - C_b)} \quad i = 1, \dots, T \quad (4.3.5)$$

donde C_b y S_b son la media y la matriz de covarianzas del subconjunto básico.

- Etapa 1(b): Conjunto básico no de rango completo.

Si el conjunto básico no es de rango completo, calcular primero los autovalores de S_b , $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k = 0$ y la matriz de autovectores correspondiente, V_b . Entonces, calcular las distancias:

$$\sqrt{(x_i - C_b)' V_b W_b V_b' (x_i - C_b)} \quad i = 1, \dots, T \quad (4.3.6)$$

donde W_b es una matriz diagonal cuyo j-ésimo elemento es:

$$w_j = \frac{1}{\max \{\lambda_j, \lambda_s\}} \quad j = 1, \dots, T \quad (4.3.7)$$

donde λ_s es el autovalor, distinto de 0, más pequeño de S_b

- Etapa 2: Incrementar el tamaño del subconjunto básico.

Reordenar las observaciones en orden ascendente de acuerdo con las expresiones (4.3.5) ó (4.3.6), según sea el caso. Sea r el número de observaciones en el subconjunto básico actual. Dividir las observaciones en dos subconjuntos - uno, conteniendo las r+1 primeras observaciones y - otro, conteniendo las T-r-1 observaciones restantes.

- Etapa 3: Criterio de parada.

Repetir los pasos 1-2 hasta encontrar un determinado criterio de parada y calcular entonces las distancias robustas:

$$D_i(C_b, S_b) = \sqrt{(x_i - C_b)' (c_b S_b)^{-1} (x_i - C_b)} \quad i = 1, \dots, T \quad (4.3.8)$$

donde $c_b = C_{Tr} m_j / \chi_{k,0.50}^2$ es un factor de corrección para obtener la consistencia cuando los datos provienen de una distribución normal multivariante y m_j es el $100(h/T)$ th percentil de los T valores de (4.3.8). Los estudios de simulación sugieren que $C_{Tr} = \{1 + r/(T-k)\}^2$, siendo r el número final de observaciones en el subconjunto básico final.

La lógica de este procedimiento radica en que al ordenar las observaciones en la etapa 0, el subconjunto básico estará libre de atípicos, de manera que las estimaciones del parámetro de localización y de la matriz de covarianzas serán robustas no viéndose afectadas por los atípicos. De esta forma, la distancia que se calcula en la etapa 1, utilizando (4.3.5) ó (4.3.6), será robusta. Después se incrementa en una observación el subconjunto, realizando las mismas operaciones hasta encontrar un criterio de parada que hay que determinar. Se han propuesto dos criterios de parada:

- a) parar cuando $\text{Min}\{D_i(C_b; S_b); i \in \text{conjunto no básico}\} \geq C_\alpha$
- b) parar cuando el subconjunto básico contenga h observaciones.

El valor crítico de C_α se define tal que:

$$\Pr[\text{min}\{D_i(C_b; S_b); i \in \text{conjunto no básico}\} \geq C_\alpha | X \text{ no contenga atípicos}] = 1 - \alpha$$

El problema reside en que C_α depende de la distribución de $D_i(C_b; S_b)$ que no es conocida. Por ello, Hadi adopta el criterio (b), definiendo h como la parte entera de $[T+k+1]/2$. Alcanzado el criterio de parada, se utilizan las observaciones contenidas en el subconjunto básico para estimar los parámetros de localización y de escala, usándose para calcular la distancia (4.3.8), declarándose como valores atípicos las observaciones superiores a esta distancia. Debido al desconocimiento de la distribución de $D_i(C_b; S_b)$, no se podrá hallar

un límite de manera que Hadi sugiere la realización de un gráfico de los valores de $D_i(C_b; S_b)$ donde se presentarán claramente los valores atípicos como aquellas observaciones que se distancian claramente del resto.

Hadi aprecia ventajas claras de este procedimiento con respecto al de la distancia robusta propuesto por Rousseeuw, destacando las siguientes:

- 1.- Menos costoso en tiempo de cálculo que la RD_i de Rousseeuw.
- 2.- Mayor eficiencia y precisión en el cálculo de los parámetros de localización y de escala: la RD_i utiliza $p+1$ observaciones mientras que la $D_i(C_b; S_b)$ usa $h=(T+k+1)/2$ observaciones.
- 3.- S_b es más estable que la calculada en el procedimiento de Rousseeuw (1985).
- 4.- Mientras que en la RD_i puede ocurrir que la matriz de covarianzas sea 0, en el procedimiento sugerido por Hadi se propone la expresión (4.3.6) para evitar estas situaciones.

Tras dos años de perfeccionamiento de este procedimiento, Hadi (1994) modifica el contraste anterior, presentando como ventajas adicionales que es uniformemente más potente y que se encuentra mucho menos afectado por los problemas de enmascaramiento e inundación de atípicos, permitiendo además un mayor control sobre el tamaño del contraste.

Hadi (1994) redefine las etapas de su contraste:

- Etapla 0: Ordenación inicial: Sin cambios respecto al contraste anterior.
- Etapla 1: Si el subconjunto básico es de rango de columnas completo, calcular $D_i^2(C_b; S_b)$ donde C_b y S_b son la media y la matriz de covarianzas del subconjunto básico. Si éste no es de rango completo, incrementarlo en tantas observaciones como se requieran hasta conseguir que sea de rango completo.
- Etapla 2: Incrementar el subconjunto básico. Sin cambios con respecto al contraste anterior.

- Etapa 3: Repetir las etapas 1 y 2 hasta que el subconjunto básico contenga

$$h = \left\lceil \frac{T+k+1}{2} \right\rceil \text{ observaciones. Entonces ir a la etapa 4.}$$

- Etapa 4: Sea r el número de observaciones en el subconjunto básico actual. Define entonces el factor de corrección C_{Tk} mediante la expresión:

$$C_{Tk} = \left[1 + \frac{k+1}{T-k} + \frac{1}{T-h-k} \right]^2 = \left[1 + \frac{2}{T-1-3k} + \frac{k+1}{T-k} \right]^2$$

el cual sólo depende de T y k . A partir de esto:

(a) Calcular $D_i^2(C_b; C_{Tk}S_b)$, y ordenarlas de forma ascendente. Sea $D_{(r+1)}^2$ el estadístico de orden r -ésimo de las $D_i^2(C_b; C_{Tk}S_b)$. [Si S_b es singular, utilizar la ecuación (4.3.6) del contraste anterior].

(b) Si $D_{(r+1)}^2 \geq \chi_{k,\alpha/T}^2$, entonces el proceso termina y se declara que todas las observaciones que satisfacen $D_i^2(C_b; C_{Tk}S_b) \geq \chi_{k,\alpha/T}^2$ son atípicas. Si $D_{(r+1)}^2 < \chi_{k,\alpha/T}^2$ ir al paso 5.

- Etapa 5: Dividir las observaciones en dos subconjuntos: un subconjunto básico que incluya las $r+1$ primeras observaciones y un subconjunto no básico conteniendo las $T-r-1$ observaciones restantes. Si $T=r+1$, se para el proceso y se declara que no hay atípicos. En otro caso, volver al paso 4.

Este método de detección de atípicos en la matriz de datos X únicamente puede utilizarse cuando el tamaño muestral es elevado, $T > 3k+1$. Esto puede parecer una desventaja, pero presenta una lógica bastante grande. En un conjunto de datos en el que el tamaño muestral sea reducido, los atípicos se confundirán con las observaciones "buenas", no pudiéndose entonces hablar de atipicidad pues serán realmente típicas. No se podría apreciar claramente cuáles son las observaciones que se desvían del conjunto de datos que podríamos denominar como atípicos.

El problema de este contraste puede entonces residir en la consideración por parte de Hadi de un tamaño muestral, a nuestro entender, tan reducido. Según Hadi, es menos costoso en tiempo que el método de estimación de la distancia robusta propuesto por Rousseeuw. Puede que este menor coste sea debido al reducido tamaño muestral que considera y que al trabajar con tamaños muestrales más amplios, este procedimiento pueda llegar a convertirse en tedioso.

4.4.- MÉTODOS GRÁFICOS

Se han propuesto una gran variedad de métodos gráficos para detectar la presencia de valores atípicos.

Daniel (1959) sugiere los "half-normal plots" en los cuales se consideran experimentos factoriales de la forma 2^p , esto es p factores de dos niveles. Considera los valores absolutos ordenados de los efectos totales. Si la distribución del error es normal, los efectos totales absolutos se comportarán como desviaciones de la distribución "half-normal" con varianza común. Esto es, su función de densidad de probabilidad tiene la forma:

$$f(x) = \sqrt{(2\pi\sigma^2)} \cdot \exp\left[-x^2 / 2\sigma^2\right] \quad (x > 0)$$

Dibujando los valores ordenados se encontrarán observaciones que caen muy cercanas a una línea recta a través del origen. Desviaciones de la linealidad, indicarán efectos reales (valores grandes fuera de la línea recta) ó violaciones de las hipótesis básicas del modelo. En particular, la presencia de un único valor atípico inflará los valores absolutos de los efectos totales, estando no dirigido hacia el origen pero hacia un valor similar al sesgo de contaminación reflejado por el atípico. Con este método, nos volveremos conscientes de que existe un atípico pero no determinaremos cual es la observación individual que lo genera.

Con más de dos atípicos, este gráfico no revela de forma clara su presencia, a menos que el sesgo sea del mismo signo y de magnitud similar. Asimismo, las violaciones de las

hipótesis básicas del modelo pueden afectar a la no linealidad de estos gráficos de forma que sea muy difícil distinguir la manifestación de un valor atípico.

Una presentación gráfica simple se obtiene dibujando los valores ordenados de los residuos en papel de probabilidad normal. Con estructura de error normal, esperaríamos una relación de línea recta. No linealidad sería indicativa de asimetría de la distribución del error, manifestándose los atípicos en alejamientos marcados en los finales del dibujo. Este método, cuya principal ventaja es que considera valores extremos de los residuos y no valores absolutos, presenta el problema del camuflaje de atípicos debido a la técnica de estimación. Por ello, Gnanaderikon y Kettenning (1972) sugieren la utilización de residuos modificados, basándose en los valores ajustados robustamente de los atípicos mejor que la estimación mínimo cuadrática para la muestra completa. Sugieren que los dibujos de probabilidad de tales residuos modificados ordenados pueden ser más informativos sobre los atípicos. Estos gráficos de probabilidad son de gran utilidad cuando sólo hay un valor atípico. Cuando hay dos ó más atípicos, Gentleman y Wilk (1975a, 75b) demuestran que puede haber confusión debido a la compensación de los efectos, teniendo en estos casos, escaso valor los gráficos de probabilidad.

Atkinson (1981, 82a, 82b) analiza métodos gráficos para detectar atípicos de los modelos lineales. Para la respuesta de atípicos, recomienda los "half-normal plots" de los residuos estimados adecuadamente modificados. En vez de usar los residuos estudentizados, sugiere utilizar:

$$\hat{u}_i^* = \frac{y_i - \mathbf{x}_i' \hat{\beta}_{(-i)}}{S_{(-i)} \left[1 + \mathbf{x}_i' (\mathbf{X}_{(-i)}' \mathbf{X}_{(-i)})^{-1} \mathbf{x}_i \right]^{1/2}} \quad (4.4.1)$$

Según Atkinson, estos residuos reflejan más adecuadamente los atípicos, sugiriendo un "half-normal plot" de los $|\hat{u}_i^*|$ para detectar respuestas de los atípicos.

En los casos de regresión simple, una forma adecuada para detectar los atípicos son los "scatter plots" en los cuales se representan los valores (y_i, x_i) , analizándose de esta forma la estructura de los datos, detectándose fácilmente los puntos de influencia en ambas

direcciones. En el caso de la regresión múltiple, este método se extendería a los "partial regression leverage plots".

Otro método gráfico también bastante ilustrativo es el histograma de los residuos. Bajo la hipótesis de normalidad, un histograma de frecuencias presentando colas muy gruesas y asimetría, podría ser indicativo de la presencia de atípicos. El problema es el método de estimación que puede llevar a generar unos residuos que presenten valores muy elevados debido a la escasa robustez del ajuste. Rousseeuw y Leroy (1987) proponen utilizar los residuos calculados a partir de un método de estimación muy robusto, que incluyen información muy potente para detectar todos los atípicos presentes en los datos. Un gráfico de estos residuos estandarizados contra el valor estimado de y_i proporciona una descripción cuantitativa de esta información, permitiendo al analista detectar los puntos "malos" de forma simple.

En esta sección se han analizado múltiples contrastes derivados para detectar la presencia de valores atípicos. Como se comentó anteriormente, se han incluido los que se han considerado más importantes. Existen otros contrastes no recogidos en este trabajo, algunos de los cuales se pueden encontrar en Atkinson y Weisberg (1991), Portnoy (1991), Davies y Gather (1993) y Barnett y Lewis (1984).

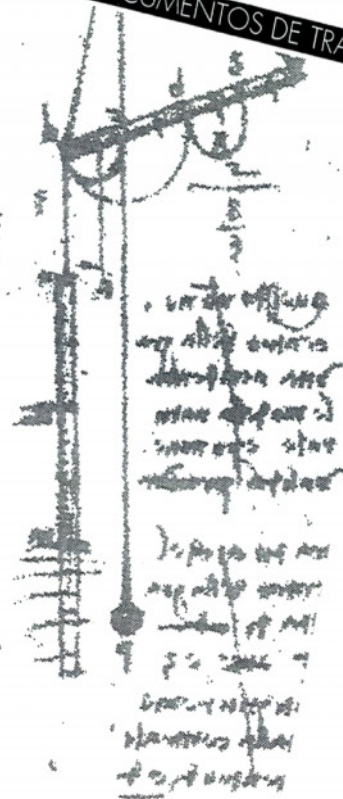
Los tres ejes son
 ortogonales entre sí y se
 sitúan en el centro de la
 pieza que se estudia. Los ejes
 están etiquetados como X, Y y Z.
 Los ejes X y Y están en el
 plano de la pieza y el eje Z
 es perpendicular a ese plano.
 Los ejes X y Y están
 paralelos a las caras de la
 pieza y el eje Z es
 perpendicular a esas caras.
 Los ejes X y Y están
 paralelos a las aristas de la
 pieza y el eje Z es
 perpendicular a esas aristas.
 Los ejes X y Y están
 paralelos a las diagonales de la
 pieza y el eje Z es
 perpendicular a esas diagonales.

Los ejes X, Y y Z son
 ortogonales entre sí y se
 sitúan en el centro de la
 pieza que se estudia. Los ejes
 están etiquetados como X, Y y Z.
 Los ejes X y Y están en el
 plano de la pieza y el eje Z
 es perpendicular a ese plano.
 Los ejes X y Y están
 paralelos a las caras de la
 pieza y el eje Z es
 perpendicular a esas caras.
 Los ejes X y Y están
 paralelos a las aristas de la
 pieza y el eje Z es
 perpendicular a esas aristas.
 Los ejes X y Y están
 paralelos a las diagonales de la
 pieza y el eje Z es
 perpendicular a esas diagonales.

Los ejes X, Y y Z son
 ortogonales entre sí y se
 sitúan en el centro de la
 pieza que se estudia. Los ejes
 están etiquetados como X, Y y Z.
 Los ejes X y Y están en el
 plano de la pieza y el eje Z
 es perpendicular a ese plano.
 Los ejes X y Y están
 paralelos a las caras de la
 pieza y el eje Z es
 perpendicular a esas caras.
 Los ejes X y Y están
 paralelos a las aristas de la
 pieza y el eje Z es
 perpendicular a esas aristas.
 Los ejes X y Y están
 paralelos a las diagonales de la
 pieza y el eje Z es
 perpendicular a esas diagonales.

5.- MODELIZACIÓN Y CORRECCIÓN DE VALORES ATÍPICOS

Los ejes X, Y y Z son
 ortogonales entre sí y se
 sitúan en el centro de la
 pieza que se estudia. Los ejes
 están etiquetados como X, Y y Z.
 Los ejes X y Y están en el
 plano de la pieza y el eje Z
 es perpendicular a ese plano.
 Los ejes X y Y están
 paralelos a las caras de la
 pieza y el eje Z es
 perpendicular a esas caras.
 Los ejes X y Y están
 paralelos a las aristas de la
 pieza y el eje Z es
 perpendicular a esas aristas.
 Los ejes X y Y están
 paralelos a las diagonales de la
 pieza y el eje Z es
 perpendicular a esas diagonales.



Los ejes X, Y y Z son
 ortogonales entre sí y se
 sitúan en el centro de la
 pieza que se estudia. Los ejes
 están etiquetados como X, Y y Z.
 Los ejes X y Y están en el
 plano de la pieza y el eje Z
 es perpendicular a ese plano.
 Los ejes X y Y están
 paralelos a las caras de la
 pieza y el eje Z es
 perpendicular a esas caras.
 Los ejes X y Y están
 paralelos a las aristas de la
 pieza y el eje Z es
 perpendicular a esas aristas.
 Los ejes X y Y están
 paralelos a las diagonales de la
 pieza y el eje Z es
 perpendicular a esas diagonales.

5.- MODELIZACIÓN Y CORRECCIÓN DE LOS VALORES ATÍPICOS

Una vez que se han detectado los valores atípicos que pueden presentarse en los modelos econométricos habituales analizados en este trabajo, el paso siguiente será modelizar, corregir ó eliminar éstos con el fin de minimizar al máximo las nefastas consecuencias que estas observaciones tienen sobre los estadísticos y estimaciones tal y como se ha analizado en secciones anteriores.

De nuevo, los métodos de corrección que han sido sugeridos por diferentes autores son distintos dependiendo de qué tipo de modelo econométrico estemos analizando. Por este motivo, analizaremos los métodos de corrección y/ó modelización de los valores atípicos para cada uno de los esquemas econométricos considerados en este trabajo.

En esta sección presentaremos los métodos que han sido propuestos para tratar el problema de los valores atípicos en el modelo de regresión.

Expondremos dos tipos de soluciones que pensamos son las más acertadas, aunque existen otros métodos, no tan ampliamente utilizados, como son la eliminación de las observaciones atípicas, la sustitución de éstas por la media de las observaciones anterior y posterior, la sustitución por una extrapolación, etc.. (ver Barnett y Lewis (1984)), que aquí no comentaremos.

En la investigación realizada sobre los métodos de corrección de los valores atípicos, se han encontrado dos que han recibido la mayor atención. Por un lado, se encuentran los métodos de estimación robusta que, además de presentar una elevada utilidad en la etapa de detección de los valores atípicos, son capaces de generar estimaciones robustas de los parámetros ante la presencia de valores atípicos, lo que supone que en numerosas situaciones no será necesario corregir estos valores debido a que los parámetros estimados del modelo no se verán negativamente afectados por estas observaciones anómalas.

Por otro lado, se encuentra el método que podríamos denominar "clásico" de corrección de los valores atípicos, consistente en la inclusión en el proceso de estimación de variables

ficticias para eliminar los efectos nocivos de los atípicos, obteniéndose así una estimación del valor del efecto que un atípico tiene sobre la variable analizada.

Pasamos ahora a exponer ambos tipos de soluciones.

5.1.- ESTIMACIÓN ROBUSTA DE LOS PARÁMETROS DEL MODELO

El desarrollo tecnológico, proporcionando grandes avances en la informática con ordenadores cada vez más rápidos y la posibilidad de desarrollar programas ó aplicaciones mucho más potentes y eficientes, ha resucitado el interés sobre los métodos de estimación robusta de los parámetros de los modelos econométricos.

Como ya se ha comentado, las series de datos de variables económicas presentan numerosos valores atípicos que pueden provocar importantes efectos negativos sobre los estimadores y los estadísticos, conseguidos mediante la estimación mínimo cuadrática, método que se ve muy afectado por estas observaciones anómalas.

Antes de exponer los diferentes métodos de estimación robusta más ampliamente utilizados, creemos que es importante clarificar ciertos conceptos muy directamente relacionados con estos procedimientos de estimación.

El principal objetivo de los procedimientos de estimación robusta es la obtención de estimaciones de los coeficientes que se vean afectados lo menos posible por las observaciones anómalas ó los valores atípicos. En este sentido, existe un concepto fundamental en torno al cual giran todos los métodos de estimación robusta: el "breakdown point" (BP a partir de ahora). La definición dada por Donoho y Huber (1983) establece que representa a "la fracción más pequeña de contaminación que puede causar que un estimador tome valores aberrantes arbitrariamente grandes". En otras palabras, se entiende por BP al número mínimo de observaciones atípicas que pueden estar presentes en una muestra de forma que los estimadores basados en estas observaciones se vean claramente afectados, variando substancialmente los resultados si estos valores anormales no estuviesen presentes. En este sentido, se buscan pues estimadores con un BP elevado, significando que la muestra puede

contar con un número importante de observaciones atípicas de manera que los estimadores no se verán afectados por éstas.

El problema fundamental que se plantea es que los estimadores robustos con un BP elevado carecen en numerosas situaciones de otra de las propiedades elementales de todo estimador, la eficiencia, de manera que las investigaciones más recientes en torno a este problema se centran en encontrar métodos de estimación robustos con elevado BP y eficientes. De todas formas, tal y como exponen Rousseeuw y Leroy (1987), el BP es sólo uno de los muchos criterios de robustez exigibles a los estimadores, siendo el BP elevado únicamente una condición necesaria para que el método de estimación pueda ser considerado como aceptable, debiendo cumplir este método con otras cualidades típicas en el análisis de regresión.

Los principales objetivos de la estimación robusta fueron resumidos por Staudte y Sheather (1990) quienes establecen que las estimaciones robustas de los parámetros β del modelo de regresión deberían satisfacer los siguientes puntos:

- 1) Consistencia, normalidad asintótica y elevada eficiencia de los estimadores cuando se produzcan violaciones importantes de las hipótesis del modelo.
- 2) Métodos válidos para construir intervalos de confianza de los parámetros desconocidos y para contrastar hipótesis acerca de ellos.
- 3) Débil sensibilidad relativa de la propiedad 1) y de los resultados de 2) ante pequeñas variaciones del modelo.
- 4) Simplicidad teórica.
- 5) Cálculo sencillo.

El principal problema de la estimación mínimo cuadrática de los parámetros del modelo de regresión:

$$Y = X\beta + u$$

es que, a pesar de tener una eficiencia elevada cuando los errores u se distribuyen de forma normal, las observaciones con residuo elevado tienen una contribución mayor a los residuos al cuadrado, afectando de forma importante a su resolución. Es decir, la derivación de la estimación mínimo cuadrática se basa en la minimización de los residuos al cuadrado, de manera que un residuo elevado generará un aumento considerable de la suma residual, obteniéndose de esta forma unas estimaciones de los parámetros que la minimizan distintas a las que se hubiesen conseguido sin la presencia de ese residuo elevado. En consecuencia, la estimación mínimo cuadrática no es robusta y tiene un BP muy reducido debido a un simple atípico ó punto influyente que puede ocasionar importantes variaciones en los estimadores.

Con el objeto de evitar los problemas ocasionados al minimizar los residuos al cuadrado se han propuesto distintos estimadores alternativos, todos ellos basados en los residuos. Pasamos ahora a exponer los principales métodos de estimación propuestos. La idea principal que subyace en todos estos métodos es que los residuos elevados, pudiendo ser atípicos, no deben tener una contribución elevada de manera que los estimadores no se vean demasiado alterados.

5.1.1.- ESTIMADOR DE DESVIACIONES ABSOLUTAS MÍNIMAS Ó ESTIMADOR L_1

En el modelo de regresión lineal, el estimador L_1 se obtiene encontrando el vector de estimaciones $\hat{\beta}$ tal que

$$\sum_{i=1}^T |\hat{u}_i| = \sum_{i=1}^T |Y_i - \hat{\beta}X_i| \quad [5.1.1]$$

sea mínima.

La idea es sustituir los residuos al cuadrado, fuentes de perturbación de los estimadores, por su valor absoluto, provocando así que la contribución de un residuo elevado sea menor en la función objetivo. De esta forma, se genera una ganancia considerable en la robustez de los estimadores. Sin embargo, este estimador es robusto únicamente con respecto a los atípicos que se pueden encontrar en la variable endógena, pero es muy vulnerable a los puntos de influencia en dirección X .

Por otro lado, Wilson (1978) analizó la eficiencia de este estimador concluyendo que ésta disminuye a medida que aumenta el tamaño muestral. Es importante destacar también que este estimador no tiene una formulación explícita y que puede verse afectado por dos problemas: el estimador resultante puede no ser único, es decir, puede existir otro estimador que se ajuste igualmente a estos datos y la estimación encontrada puede ajustarse también a otro conjunto de datos.

5.1.2.- ESTIMADOR M DE HUBER (1973)

Se trata del primer estimador que busca ponderar los residuos con el objeto de que los atípicos no tengan una elevada contribución a la función objetivo a minimizar. Huber (1973) propone reemplazar los residuos al cuadrado por otra función de los residuos y escoger el estimador $\hat{\beta}$ tal que:

$$\sum_{i=1}^T \rho(\hat{u}_i) = \sum_{i=1}^T \rho(Y_i - \hat{\beta}X_i) \quad [5.1.2]$$

sea mínima, donde ρ es una función de los residuos convexa, simétrica y con un mínimo en 0. Derivando [5.1.2] e igualando a 0, se obtiene:

$$\sum_{i=1}^T \psi_i(\hat{u}_i)X_i = \sum_{i=1}^T \psi(Y_i - X_i\hat{\beta}_\psi)X_i = 0 \quad [5.1.3]$$

con $\psi = \rho'$ (derivada de ρ). Se obtendrían así los estimadores $\hat{\beta}_\psi$, llamados estimadores M ó estimadores de Huber que minimizan la expresión [5.1.2].

En su idea inicial, Huber desarrolló los estimadores M para situaciones en las que una parte de los errores o perturbaciones provenían de una distribución diferente de la normal, es decir para trabajar en situaciones en las que la distribución del error no es enteramente normal pero se encuentra "cercana" a la distribución normal.

Debido a su falta de equivarianza con respecto a una modificación en Y_i , se deben estandarizar los residuos por medio de algún estimador de la varianza, de manera que [5.1.3] se transformaría en:

$$\sum_{i=1}^T \psi(\hat{u}_i / \hat{\sigma}) X_i = 0 \quad [5.1.4]$$

donde $\hat{\sigma}$ se estima conjuntamente.

Tal y como ponen de manifiesto Staudte y Sheather (1990), si se elige de forma adecuada la función ψ , los estimadores M pueden alcanzar una elevada eficiencia en relación con los estimadores LS y pueden ser muy robustos contra los residuos muy elevados. Sin embargo, no son tan robustos contra los atípicos en X_i , de manera que presentan un BP bastante reducido, de $1/T$.

5.1.3.- ESTIMADORES M GENERALIZADOS Ó ESTIMADORES GM

Se construyen con el objeto de limitar la influencia de los puntos de las variables incluidas en la matriz X que sean atípicos. La idea consiste en cambiar el dominio de la función ψ , incluyéndose además de los residuos, los puntos de X , solucionándose con respecto a $\hat{\beta}$ la expresión:

$$\sum_{i=1}^T \eta(X_i, \hat{u}_i) X_i = \sum_{i=1}^T \eta(X_i, Y_i - X_i' \hat{\beta}_\eta) X_i = 0 \quad [5.1.5]$$

siendo ésta la aproximación sugerida por Hampel et al (1986) y Hettmansperger (1984).

Por su parte, Mallows (1975) propone reemplazar [5.1.4] por:

$$\sum_{i=1}^T \omega(X_i) \psi(\hat{u}_i / \hat{\sigma}) X_i = 0 \quad [5.1.6]$$

mientras que Schweppe (Hill (1977)) sugiere usar:

$$\sum_{i=1}^T \omega(X_i) \psi(\hat{u}_i / (\omega(X_i) \hat{\sigma})) X_i = 0 \quad [5.1.7]$$

Todos los $\hat{\beta}$ resultantes de estas tres últimas expresiones se denotarían como estimadores M generalizados ó estimadores GM y según Rousseeuw y Leroy (1987) tienen un BP bastante más elevado que los estimadores M. El punto débil de estos estimadores es su reducida eficiencia en presencia de buenos puntos de influencia.

Los residuos resultantes de la estimación M, estimación GM de Mallows y de Schweppe se analizan en McKean et al (1991), quienes estudian la utilidad de los residuos de estas estimaciones robustas para detectar las inadecuaciones en el modelo. Concluyen que los métodos de estimación robusta con elevado BP pueden conducir a situaciones erróneas si se utilizan sus residuos para verificar la adecuación del modelo, debiendo ser muy precavidos en este punto.

5.1.4.- ESTIMADORES MM - YOHAI (1988)

Los estimadores MM, propuestos por Yohai (1988), presentan el máximo BP posible (50%) y además consiguen una elevada eficiencia, dos de las principales características que se desean en todo estimador.

Se definen en tres etapas:

- 1.- Primero se calcula un estimador para el vector de parámetros β con un BP elevado ($\hat{\beta}^*$) sin necesidad de que sea eficiente. En este sentido, se puede calcular cualquier estimador de los que se presentan en esta sección.
- 2.- Sobre los residuos obtenidos con el estimador de elevado BP $\left(\hat{\beta}^* \right)$, se calcula un estimador M de la escala (S_T) que tenga el máximo BP posible (50%)
- 3.- Finalmente , el estimador MM ($\hat{\beta}$) se define como la solución de:

$$\sum_{i=1}^T \psi\left(\frac{\hat{\beta}}{S_T}\right) X_i = 0 \quad [5.1.8]$$

que satisface $S(\hat{\beta}) \leq S(\hat{\beta}^*)$ donde

$$S(\hat{\beta}) = \sum_{i=1}^T \rho\left(\frac{\hat{\beta}}{S_T}\right)$$

La función ρ debe satisfacer las siguientes propiedades:

A.- Debe ser simétrica, continuamente diferenciable y $\rho(0) = 0$.

B.- Existe $c > 0$ tal que ρ es estrictamente creciente en $[0; c]$ y constante en $[c; \infty)$.

Yohai (1988) demostró que los estimadores MM heredan el BP del 50% de la primera etapa y poseen la propiedad del ajuste exacto. Además son altamente eficientes cuando los errores tienen una distribución normal y comparan muy favorablemente con los estimadores M Generalizados.

Yohai, Stahel y Zamar (1990) proponen aumentar el estimador MM incluyendo la estimación de la covarianza asintótica con el objeto de permitir la inferencia aproximada. Según ellos, con este estimador MM consiguen la mayor parte de los objetivos que todo procedimiento de estimación robusta debe perseguir:

- a) Tener un BP elevado
- b) Tener una elevada eficiencia cuando el mínimo cuadrático lo es.
- c) Ser adecuado para la inferencia en el sentido de que los resultados asintóticos deberían estar disponibles.
- d) Mostrar un comportamiento del sesgo satisfactorio, es decir, que el sesgo máximo causado por la contaminación arbitraria sea reducido.

5.1.5.- ESTIMADOR DE MÍNIMOS CUADRADOS PONDERADOS

Cuando se sospecha que puede haber atípicos que distorsionan las estimaciones mínimo cuadráticas se pueden usar los mínimos cuadrados ponderados (WLS) consistentes en encontrar $\hat{\beta}_w$ que minimiza:

$$S(\hat{\beta}) = \sum_{i=1}^T \omega_i \hat{u}_i^2 \quad [5.1.9]$$

de manera que se otorgaría una menor ponderación a los residuos atípicos.

El primer paso consistiría en estimar las ponderaciones o escogerlas a partir de funciones de los datos preespecificadas. Después se obtienen los WLS de β y σ^2 y se usan los residuos resultantes de este modelo ajustado para obtener nuevas ponderaciones y repetir el proceso. Este procedimiento iterativo es especialmente efectivo en la estimación robusta.

Una ventaja de este problema es que se puede resolver aplicando la metodología mínimo cuadrática:

$$\text{Si } E[u] = 0, \text{cov}(u) = \sigma^2 W^{-1} \quad [5.1.10]$$

siendo W una matriz $T \times T$ simétrica y definida positiva, el problema de la estimación de β se puede transformar en uno en el cual los errores satisfacen las hipótesis ideales:

$$W = W^{1/2} \cdot W^{1/2}; Y_w = W^{1/2} Y; X_w = W^{1/2} X; u_w = W^{1/2} u.$$

Con estas nuevas variables, transformamos el modelo: $Y_w = X_w \beta + u_w$, donde la perturbación satisface las hipótesis de manera que:

$$\hat{\beta}_w = (X_w' X_w)^{-1} X_w' Y_w = (X' W X)^{-1} X' W Y$$

$$\begin{aligned} E[\hat{\beta}_w] &= \beta; \quad \text{cov}(\hat{\beta}_w) = \sigma^2 (X' W X)^{-1} \\ \text{con} \quad s_w^2 &= \frac{w' w}{n - k - 1}; \quad w = Y_w - X_w \hat{\beta}_w \end{aligned}$$

Cuando W es una matriz diagonal, $\hat{\beta}_w$ es el estimador de mínimos cuadrados ordinarios.

Staudte y Sheather (1990) demuestran que el $\hat{\beta}_w$ converge en probabilidad a β cuando se considera la estructura de regresión.

El mayor problema de este método es que es altamente ineficiente cuando se aplica sobre datos no contaminados y se requiere una mayor investigación sobre la eficiencia de este método en relación con el método mínimo cuadrático.

5.1.6.- ESTIMADOR LMS (ROUSSEEUW (1985))

Rousseeuw, para construir el estimador de mediana de cuadrados mínima (LMS), se basa en la constatación de los estimadores de mínimos cuadrados ordinarios para el modelo de regresión lineal no son robustos, viéndose muy afectados por los valores atípicos que puedan existir. Además, busca un estimador con BP elevado, siendo el mayor posible (el 50%). Por otro lado, considera que la mediana no se verá afectada por los valores atípicos, de manera que define el estimador de mediana de cuadrados mínima de la siguiente forma.

Para el modelo lineal general, $Y=X\beta+u$, donde en principio se supone que se cumplen las hipótesis ideales del modelo, se trataría de encontrar el vector de parámetros β tal que:

$$\text{mediana}_i^2 = \text{mediana}_i(Y - X\hat{\beta})^2 \quad [5.1.11]$$

sea mínima.

Rousseeuw pone de manifiesto que el estimador resultante posee bastantes ventajas: siempre tiene solución, aunque no se obtiene una formulación explícita, no se ve afectado por los cambios en la escala y si el número de variables explicativas es mayor que 1, el BP del método LMS es $([T/2]-k+2)/T$ (donde $[.]$ se refiere a la parte entera de $.$). Además, si $k>1$ y existe un β tal que al menos $T-[T/2]+k-1$ de las observaciones satisfacen exactamente $Y = X_i\hat{\beta}$ y se encuentran en posición general, entonces la solución LMS de [5.1.11] iguala a $\hat{\beta}$ sean cuales sean las otras observaciones.

Rousseeuw también reconoce que su método de estimación de los parámetros presenta desventajas como es la escasa eficiencia de los estimadores cuando los errores del modelo de regresión son realmente normales. Para conseguir mayor eficiencia, Rousseeuw propone utilizar los LMS como valores iniciales de un **estimador M de una sola etapa**, denominados OSM cuya expresión sería:

$$\hat{\beta}_{OSM} = \hat{\beta}_{LMS} + (X'X)^{-1}X' \left[\psi(\hat{u}_1/\sigma^*), \dots, \psi(\hat{u}_T/\sigma^*) \right]' \cdot \frac{\sigma^*}{B(\psi, \phi)} \quad [5.1.12]$$

$$\text{donde } B(\psi, \phi) = \int \psi'(v) d\phi(v); \hat{u}_i^* = Y_i - X_i \hat{\beta}_{LMS}.$$

Otra alternativa para mejorar la eficiencia sería utilizar los **mínimos cuadrados reponderados**, consistentes en que para cada observación (Y_i, X_i) , se asigna una ponderación ω_i que es una función de los residuos LMS estandarizados, $\hat{u}_i/\sigma^*$ en valor absoluto. Se pueden usar diferentes funciones:

1.- Utilizar la función de ponderación:

$$\omega_i = \begin{cases} 1 & \text{si } |\hat{u}_i/\sigma^*| \leq c_1 \\ 0 & \text{en el resto de los casos} \end{cases} \quad [5.1.13]$$

2.- Introducir una parte lineal que suavice la transición de la ponderación 1 a la ponderación 0:

$$\omega_i = \begin{cases} 1 & \text{si } |\hat{u}_i/\sigma^*| \leq c_2 \\ (c_3 - |\hat{u}_i/\sigma^*|)/(c_3 - c_2) & \text{si } c_2 \leq |\hat{u}_i/\sigma^*| \leq c_3 \\ 0 & \text{en el resto de los casos} \end{cases} \quad [5.1.14]$$

donde c_1, c_2 y c_3 son constantes predefinidas.

Una vez seleccionada la función de ponderación, se reemplaza cada observación $(Y_i; X_i)$ por $(\omega_i^{1/2} Y_i; \omega_i^{1/2} X_i)$, pudiéndose aplicarse ya los métodos de mínimos

cuadrados sobre estas observaciones reponderadas, consiguiéndose así un estimador con mayor eficiencia que el LMS original.

5.1.7.- ESTIMADOR LTS - (ROUSSEEUW (1983))

Rousseeuw introduce otro estimador robusto en el que se deben ordenar los residuos al cuadrado de menor a mayor, obteniéndose:

$$(\hat{u}^2)_{1:T} \leq (\hat{u}^2)_{2:T} \leq \dots \leq (\hat{u}^2)_{T:T}$$

Considerando únicamente los h primeros términos de esta ordenación, Rousseeuw (1983) define el estimador de residuos al cuadrado ordenados mínimo (LTS) como el estimador $\hat{\beta}$ que minimiza:

$$\sum_{i=1}^h (\hat{u}^2)_{i:T} = \sum_{i=1}^h (Y_i - X_i \hat{\beta})_{i:T}^2 \quad [5.1.15]$$

Si se considera $h=[T/2]+1$, el estimador LTS tiene el mismo BP que el estimador de mínima mediana de cuadrados. Además, para $h=[T/2]+[(k+1)/2]$, el LTS alcanza el máximo "breakdown point" posible, es decir, $([(T-k)/2]+1)/T$.

Junto a la propiedad de que alcanza el máximo BP, también presenta las mismas ventajas que las citadas para el estimador LMS, siendo además más eficiente que éste debido a una convergencia más rápida.

La principal desventaja de este método es que requiere ordenar los residuos de menor a mayor y además no se obtiene una formulación explícita de este estimador.

5.1.8.- ESTIMADORES S - (ROUSSEEUW Y YOHAI (1984))

Rousseeuw y Yohai (1984) derivan otra clase de estimadores robustos, con un BP elevado, invariantes y con una convergencia elevada, que se basan en la minimización de la dispersión de los residuos. Se trata de encontrar el estimador $\hat{\beta}_s$ tal que:

$$S(\hat{u}_1(\hat{\beta}), \dots, \hat{u}_T(\hat{\beta})) \quad [5.1.16]$$

es mínimo, con estimador final de la escala dado por la expresión:

$$\hat{\sigma} = S(\hat{u}_1(\hat{\beta}), \dots, \hat{u}_T(\hat{\beta})) \quad [5.1.17]$$

La dispersión $S(\hat{u}_1(\hat{\beta}), \dots, \hat{u}_T(\hat{\beta}))$ se define como:

$$\frac{1}{T} \sum_{i=1}^T \rho\left(\frac{\hat{u}_i}{s}\right) = k \quad [5.1.18]$$

donde k a menudo se iguala a $E_{\Phi}[\rho]$, siendo Φ la distribución normal estándar.

5.1.9.- ESTIMADOR DE MEDIANA REPETIDA - (SIEGEL (1982))

Concluimos esta sección dedicada a los estimadores robustos con el primer estimador con BP máximo que fue derivado por Siegel (1982). Se basa en todos los subconjuntos de k puntos que se pueden formar, que denominaremos $i_1, i_2, \dots, i_k$. Cualquiera de las k observaciones $(y_{i_1}, x_{i_1}), \dots, (y_{i_k}, x_{i_k})$ determinan un único vector de parámetros denotado por $\beta_j(i_1, \dots, i_k)$. El estimador de mediana repetida se define entonces como:

$$\hat{\beta}_j = \text{med}_{i_1} \left(\dots \left(\text{med}_{i_{k-1}} \left(\text{med}_{i_k} \beta_j(i_1, \dots, i_k) \right) \right) \dots \right) \quad [5.1.19]$$

Entre las ventajas de este estimador se encuentra el hecho de que se puede calcular explícitamente y que, al calcularse secuencialmente las medianas en vez de una mediana para el total, este estimador alcanza un BP del 50%, es decir, el máximo posible.

Los problemas que se plantean son que tanto la eficiencia asintótica como la función de influencia de este estimador son desconocidas, problemas que unidos al hecho de que se puede tener un número elevado de subconjuntos (tantos como variables explicativas se incluyan en X), hacen que no sea de los estimadores robustos más utilizados.

Como conclusión a las soluciones propuestas en esta sección dedicada a los métodos de estimación robusta, cabe destacar que todos estos métodos presentan problemas.

Destaca asimismo el hecho de que ninguno de los métodos busca considerar la información que puede aportar un valor atípico a la hora de explicar la variable dependiente, sino que simplemente buscan estimar los parámetros del modelo de forma que éstos no se vean afectados por los valores atípicos. Pasaremos ahora a exponer el método "clásico" de modelización de valores atípicos, método en el cual si que se tiene en consideración la información suministrada por el atípico.

5.2.- MÉTODO "CLÁSICO" - VARIABLES FICTICIAS

Como ya se comentó anteriormente, los métodos de estimación robustos han empezado a ser más utilizados recientemente gracias al desarrollo de los ordenadores que son más potentes y al desarrollo de programas o aplicaciones que se están adecuando a los ordenadores. Sin embargo, estos métodos no son recientes; lo que si es más reciente, es la preocupación por este tipo de estimación.

Por este motivo, los métodos de modelización o eliminación de los valores atípicos que más ampliamente se han utilizado son, por un lado, los propuestos por algunos autores consistente en la eliminación ó sustitución de estas observaciones (véase Barnett y Lewis (1984)), métodos en los cuales no se tenía en consideración la información que podía suministrar esta observación anómala. Hay que recordar que, a nuestro entender, en las series económicas, los valores atípicos vienen causados por algún hecho que provoca que la variable, en ese momento, se comporte de forma diferente al resto de los puntos, de manera que puede estar incluyendo información que creemos es necesario incluir en el proceso de estimación.

En segundo lugar, el método más utilizado en las investigaciones económicas ha sido la modelización de estos valores atípicos mediante la introducción de variables ficticias en el

proceso de estimación⁴. De esta forma, se recoge la información que proporciona ese valor atípico debido a que la variable ficticia, ya sea tipo impulso ya sea tipo escalón, tiene asociado un coeficiente que se estima conjuntamente con el resto de los parámetros del modelo.

Debido a que es ampliamente conocido por la mayor parte de los investigadores, nos centramos exclusivamente en realizar una serie de comentarios sobre él, en vez de exponer las consideraciones teóricas que subyacen en el método "clásico".

La idea que subyace en este método pensamos que se ajusta de forma adecuada al objetivo que se persigue mediante la estimación de los parámetros del modelo de regresión. En éste, se busca establecer una relación, generalmente lineal, entre la variable dependiente, Y , y las variables explicativas, incluidas en la matriz X , de manera que se consiga una función de ajuste ($\hat{Y}$) que sea lo más próxima posible a la variable a explicar (Y), es decir, se intenta establecer una relación de la variable Y con las variables incluidas en X de tal forma que se explique la mayor proporción posible de Y . Los residuos representarán así lo que dejamos de explicar de manera que un residuo muy elevado, en valor absoluto, representará el punto de Y que el modelo ha sido incapaz de explicar, pudiendo denominarse atípico. Debido a que las variables incluidas en X no pueden explicar este valor, la idea de incluir una variable ficticia para recogerlo se encuentra en perfecta consonancia con el objetivo de la regresión. Asimismo, ayudará a mejorar la varianza residual, haciéndola más reducida, de manera que disminuirán los intervalos de predicción, otro de los objetivos que se persiguen con el modelo de regresión. De esta forma, la inclusión de una variable ficticia, además de contribuir a explicar a la variable endógena, contribuye a obtener unas predicciones con intervalos más pequeños, siendo así más fiable.

Sin embargo, pensamos que, en numerosas situaciones, la modelización de los valores atípicos no se ha hecho correctamente ó mejor dicho, no se detectaron de forma adecuada y, en consecuencia, la inclusión de las variables ficticias no se hizo de forma adecuada. Tal vez por falta de medios ó por complejidad, los valores atípicos se suelen identificar por medio de los residuos de un ajuste mínimo cuadrático que, como se ha comentado en secciones anteriores, no es demasiado robusto, de manera que la identificación de los valores atípicos

⁴ Este método es el que denominamos "clásico" debido a que es el más ampliamente utilizado.

puede no ser todo lo fiable que se desearía. Además, en la mayor parte de las ocasiones, esta identificación se basa en la hipótesis de normalidad de los residuos lo que ocurre en pocas ocasiones. Debe recordarse que los atípicos distorsionan de forma importante los coeficientes de asimetría y curtosis, de manera que la supuesta normalidad de los residuos no se suele verificar. Por todo ello, pensamos que la identificación de los valores atípicos debería realizarse sobre unos residuos basados en estimaciones robustas, tal y como se ha expuesto en secciones anteriores, combinando el método con la modelización de los atípicos gracias a la introducción de variables ficticias. Es decir, el método consistiría, primero en identificar los valores atípicos a partir de los residuos obtenidos mediante un ajuste robusto, y en una segunda etapa, se realizaría la estimación de los parámetros del modelo elegido, incluyendo en la matriz X las variables ficticias oportunas para recoger el comportamiento de los atípicos anteriormente identificados. De esta forma, se obtendrían los efectos que los distintos atípicos tuvieron sobre la variable objeto del estudio.

La combinación de ambos métodos nos permite también conocer qué tipo de atípicos se presentan debido a que los métodos de estimación robusta, cuando los utilizamos para la detección de valores atípicos, nos permiten clasificar éstos (véase sección 4.1), conociendo, de esta forma, además del tipo, la influencia que presenta en el modelo de regresión.

[illegible]

[Faint, illegible handwritten text]

[Faint handwritten notes or bleed-through from another page.]

6.- EJEMPLO DE LOS DISTINTOS ATÍPICOS EN EL MODELO DE REGRESIÓN

DE LOS DISTINTO
 EN EL MODELO DE R

6.- EJEMPLOS DE LOS DISTINTOS ATÍPICOS EN EL MODELO DE REGRESIÓN

De cara a ilustrar todo lo anteriormente expuesto, presentamos en esta sección diferentes ejemplos de cómo se verán afectados los estadísticos cuando tanto la variable endógena como alguna de las exógenas presenta valores atípicos. En este sentido se ha realizado un ejercicio de simulación que pasamos a detallar.

Aunque en el campo económico habitualmente los atípicos se presentan tanto en la variable endógena como en las exógenas y por lo general se pueden presentar múltiples atípicos, de diferentes tipos, en unas y en otras, por sencillez expositiva, aquí distinguiremos entre atípicos en la variable endógena y atípicos en las variables exógenas. Para ello, partimos de un modelo general simulado sobre el cual se irán introduciendo las diferentes modificaciones para ilustrar los efectos de los atípicos considerados. El modelo inicial se construyó a partir de la simulación de variables $N(0,1)$, tanto para la variable endógena (Y) como para las 4 variables exógenas introducidas en el modelo (X_1, X_2, X_3 y X_4), todas de 300 observaciones. El modelo inicialmente simulado fue el siguiente:

$$Y = 45.235 + 2.87X_1 - 10.876X_2 + 9.8754X_3 + 74.234X_4 + u_t$$

donde u es una variable aleatoria $N(0,1)$.

A partir de este modelo inicial, se fueron introduciendo los diferentes atípicos, para la variable endógena y las variables explicativas.

* Atípicos para la variable endógena

Atípico tipo impulso: se introdujo un atípico tipo impulso en la observación 100, con un efecto de magnitud importante (100), obteniéndose $YI1$

Atípico tipo escalón: se introdujo en la observación 150, también con un efecto de magnitud 100 y de duración hasta el final de la muestra. La variable endógena con el atípico tipo escalón recibe el nombre de $YI2$.

* Atípicos para las variables explicativas

Atípico tipo impulso: se introduce en la variable X_2 , con efecto, magnitud y momento de ocurrencia exactamente iguales a los descritos en el caso del atípico tipo impulso para la variable endógena. La variable con dicho atípico recibe el nombre de X2I1

Atípico tipo escalón: Al igual que en el caso anterior, este tipo de atípico se introduce sobre la variable X_2 con los mismos esquemas seguidos para la variable endógena. Esta nueva variable recibe el nombre de X2I2.

De esta forma, tendremos:

- a) Atípico tipo impulso en Y
- b) Atípico tipo escalón en Y
- c) Atípico tipo impulso en X_2
- d) Atípico tipo escalón en X_2

dando paso así a la estimación de diferentes modelos a partir de los cuales consideraremos los principales efectos de los atípicos, los contrastes para la detección de los atípicos y las diversas formas de tratamiento que se han propuesto para disminuir o aminorar los efectos de los atípicos sobre las estimaciones

6.1.- EFECTOS DE LOS VALORES ATÍPICOS SOBRE LOS ESTIMADORES Y ESTADÍSTICOS

En esta sección, analizaremos los efectos de los diferentes valores atípicos en los estimadores, los estadísticos etc.. comparando los resultados con los obtenidos en la estimación del modelo inicial simulado. En la tabla 1, se presentan los resultados de las estimaciones para los diferentes modelos considerados, siendo éstos los siguientes:

Modelo 1: Modelo inicial simulado:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u_t$$

Modelo 2: Modelo con atípico tipo impulso en Y

$$YI1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u_t$$

Modelo 3: Modelo con atípico tipo escalón en Y

$$Y_{I2} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u_t$$

Modelo 4: Modelo con atípico tipo impulso en X_2 :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 I1 + \beta_3 X_3 + \beta_4 X_4 + u_t$$

Modelo 5: Modelo con atípico tipo escalón en X_2 :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 I2 + \beta_3 X_3 + \beta_4 X_4 + u_t$$

Dividiremos los efectos en función de dónde se encuentre el valor atípico, ya sea en la variable endógena ó sobre la variable exógena. Los resultados de las estimaciones, que se comentarán a continuación, se encuentran en la tabla siguiente. (Entre paréntesis figuran los errores estándar)

Efectos de los atípicos sobre los estimadores y estadísticos					
Parámetros y estadísticos	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
$\hat{\beta}_0$ (ES)	45.255 (0.0559)	45.5997 (0.3403)	95.7117 (2.9055)	46.4360 (0.6303)	46.3074 (0.9113)
$\hat{\beta}_1$ (ES)	29135 (0.0578)	3.0804 (0.3517)	7.3282 (3.0030)	4.2961 (0.6494)	4.2784 (0.6616)
$\hat{\beta}_2$ (ES)	-10.791 (0.0549)	-10.8018 (0.3341)	-12.2480 (2.8524)	-0.3233 (0.1082)	0.0010 (0.0129)
$\hat{\beta}_3$ (ES)	9.9530 (0.0617)	9.3735 (0.3754)	7.2394 (3.2049)	10.5855 (0.7001)	10.8005 (0.7077)
$\hat{\beta}_4$ (ES)	74.2468 (0.0544)	73.9214 (0.3309)	77.3788 (2.8254)	73.3224 (0.6141)	73.3995 (0.6340)
R^2	0.9998	0.9942	0.7214	0.9801	0.9795
$\hat{\sigma}$	0.9632	5.8578	50.0133	10.8945	11.0581
LBQ(12)	9.3540	0.9100	3148.6	18.027	17.479
LBQ(24)	17.937	1.1919	5647.9	27.232	27.941
LBQ(36)	23.998	2.1906	7544.5	44.403	43.290

6.1.1.- EFECTO DE LOS ATÍPICOS EN LA VARIABLE ENDÓGENA SOBRE LOS ESTIMADORES Y LOS ESTADÍSTICOS

* Modelo 2 - Atípico tipo impulso

Los principales comentarios que se pueden extraer de los resultados presentados en la tabla 1 son los siguientes:

- No se aprecian importantes cambios en las estimaciones de los coeficientes con respecto al modelo 1, no siendo muy diferentes a los obtenidos en la estimación del modelo inicial. Se producen sesgos reducidos de distinto signo.
- Los errores estándar aumentan considerablemente con respecto al modelo 1, debido al incremento que se produce en el error estándar de la regresión ($\hat{\sigma}$).
- El coeficiente de determinación disminuye ligeramente con respecto al obtenido en el modelo inicial, pudiéndose decir que se consigue un ajuste bastante satisfactorio: se explica el 99.42% de la variable endógena.
- La mayor variación observada con respecto al modelo inicial se produce en los estadísticos de Ljung-Box, que disminuyen significativamente. Esto corrobora el resultado de Chan (1995) quien demuestra que, cuando se produce un atípico tipo impulso de magnitud elevada (que tiende a ∞), las autocorrelaciones tienden a 0, llevando a identificar un proceso de ruido blanco de la serie analizada, en este caso, los residuos.

*** Modelo 3 - Atípico tipo escalón**

- Se aprecian cambios muy importantes tanto en las estimaciones de los coeficientes del modelo como en los estadísticos. En algunos casos los coeficientes duplican en valor al realmente simulado (coeficientes $\hat{\beta}_0$ y $\hat{\beta}_1$), mientras que en el resto se aprecian variaciones en los parámetros en torno a los 3 puntos. Los sesgos de los coeficientes estimados también son de diverso signo, aunque de magnitud bastante elevada.
- Los errores estándar de los coeficientes aumentan considerablemente con respecto a los obtenidos en el modelo inicial y también con respecto a los obtenidos en el modelo 2, es decir, con atípico tipo impulso en la variable endógena.
- Se aprecia una importante reducción en el coeficiente de determinación, así como un considerable incremento en la desviación estándar del modelo, aumento que

provoca la elevación de los errores estándar de las estimaciones que pierden precisión.

- Los estadísticos de Ljung-Box también aumentan significativamente, indicando claramente que los residuos no se pueden considerar ruido blanco.

A partir de los resultados obtenidos en la estimación de los modelos 2 y 3 queda claro que los efectos del atípico tipo escalón en la variable endógena son mucho más fuertes que los ocasionados por el atípico tipo impulso, produciéndose alteraciones de los resultados mucho más importantes que las provocadas por el atípico tipo impulso.

6.1.2.- EFECTO DE LOS ATÍPICOS EN LA VARIABLE EXPLICATIVA SOBRE LAS ESTIMACIONES Y LOS ESTADÍSTICOS.

*** Modelo 4 - Atípico tipo impulso en X_2**

- El mayor efecto observado en las estimaciones de los coeficientes del modelo se produce para el parámetro $\hat{\beta}_2$, es decir, el coeficiente asociado a la variable X_2 que incluye el atípico tipo impulso. Este parámetro, aunque mantiene el signo de la simulación, disminuye notablemente de valor con respecto al modelo inicial simulado.
- El resto de las estimaciones también se ven afectadas, si bien sus parámetros cambian relativamente poco. Como en los casos anteriores, todos los coeficientes presentan sesgos con respecto al modelo original, siendo éstos positivos para casi todos los parámetros, excepto para el parámetro $\hat{\beta}_4$, cuyo valor disminuye con respecto al modelo inicial.
- Se aprecian también importantes incrementos en los errores estándar de las estimaciones, causadas por el notable incremento en la desviación estándar de la estimación con respecto a la del modelo inicial.
- Se produce una ligera disminución en el coeficiente de determinación, aunque se podría aceptar perfectamente que el modelo alcanza un nivel de explicación de la variable endógena muy elevado (98%)

- Se produce un notable incremento en los estadísticos de Ljung-Box, que en todos los casos casi duplican a los obtenidos en el modelo inicial. Aún así, no podría rechazarse la hipótesis de que los residuos son ruido blanco.

*** Modelo 5 - Atípico tipo escalón en X_2**

Los resultados de las estimaciones y los efectos sobre los principales estadísticos son muy similares a los obtenidos y comentados en el modelo anterior, apareciendo unos coeficientes estimados de magnitud prácticamente igual a los del modelo 4, pudiéndose pues extender los comentarios anteriormente realizados a este caso.

Únicamente apreciamos una disminución, en valor absoluto, del coeficiente asociado a X_2 , variable afectada por el atípico tipo escalón, que pierde toda su significatividad estadística.

Comentar finalmente que se observa una ligera disminución en el coeficiente de determinación, consiguiéndose explicar un 97.95% de la variable endógena. Además, se produce un leve incremento, con respecto al modelo 4, del error estándar de la regresión y los estadísticos de Ljung-Box nos permiten aceptar la hipótesis de que los residuos son ruido blanco.

A partir de los resultados comentados para la inclusión de atípicos en las variables exógenas, se puede concluir que tanto el tipo impulso como el tipo escalón tienen efectos muy similares, no pudiendo decidir cuál de ellos es más perjudicial de cara a la regresión. Destacar, sin embargo, que el atípico tipo escalón incluido en la variable X_2 provoca la pérdida de significatividad estadística del coeficiente asociado, mientras que el atípico tipo impulso en X_2 , aunque provoca una importante disminución, en valor absoluto, del valor del estimador con respecto al modelo inicial, este coeficiente sigue siendo significativo.

6.2.- DETECCIÓN DE LOS VALORES ATÍPICOS

Sobre los modelos 2, 3, 4 y 5, se efectuarán algunos de los contrastes de detección analizados en la sección "Detección de valores atípicos" con el objeto de verificar si dichos métodos concluyen o detectan siempre los mismos atípicos y sobre todo si son capaces de indicar cuál de las observaciones atípicas introducidas en las variables puede presentar problemas. Es decir, se trata de verificar si la detección de atípicos mediante los distintos

contrastes concluyen todos en la misma dirección.

Los contrastes efectuados en esta parte del trabajo fueron los siguientes:

- a.- Gráficos de la variable endógena con atípicos versus las variables explicativas.
- b.- Gráficos de los residuos con los límites $\pm 2\sigma$.
- c.- Histograma de los residuos.
- d.- Contraste de los residuos estandarizados comparándolos con los límites ± 2.5 .
- e.- Contraste de los residuos estudentizados comparándolos con los límites ± 2.5 .
- f.- Residuos estandarizados versus las distancias de Mahalanobis robustas de las variables explicativas.
- g.- Métodos de eliminación de filas.

Hay que destacar que los contrastes a y f, es decir aquellos que intentan relacionar los atípicos con las variables explicativas para los modelos 2 y 3 sólo se utilizan a modo ilustrativo ya que los atípicos incluidos en dichos modelos se deben a la variable endógena, es decir, la simulación de los dos atípicos considerados (impulso y escalón) se efectuó directamente sobre la variable endógena. Sin embargo, como veremos posteriormente, estos contrastes se muestran muy ilustrativos para los modelos 4 y 5 en los que el atípico se encuentra en la variable exógena. Finalmente decir que todos los contrastes se efectúan sobre los modelos sin la consideración de valores atípicos con el objeto de verificar si los métodos de detección nos indican la presencia de los mismos.

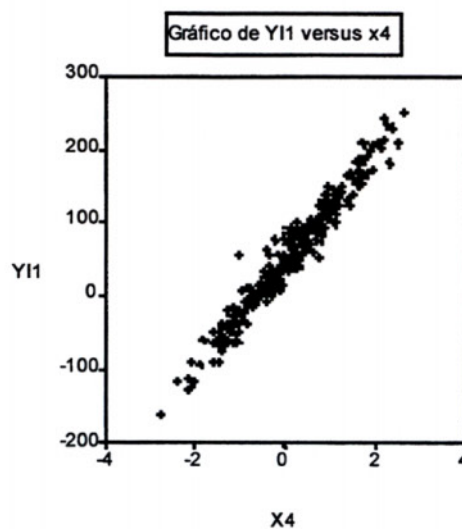
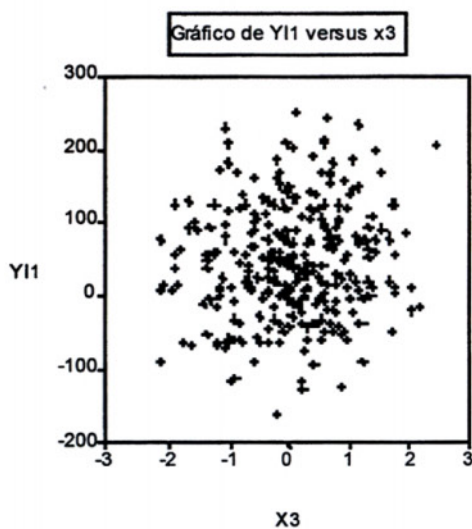
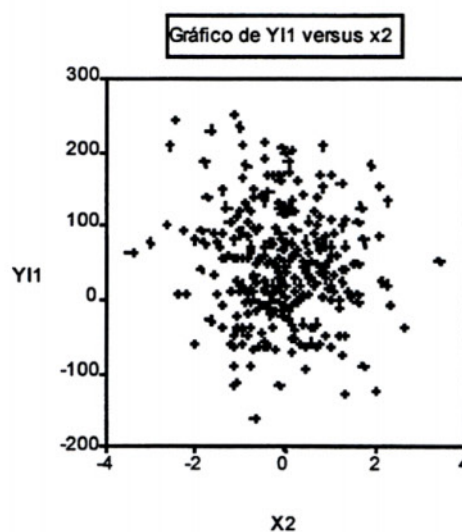
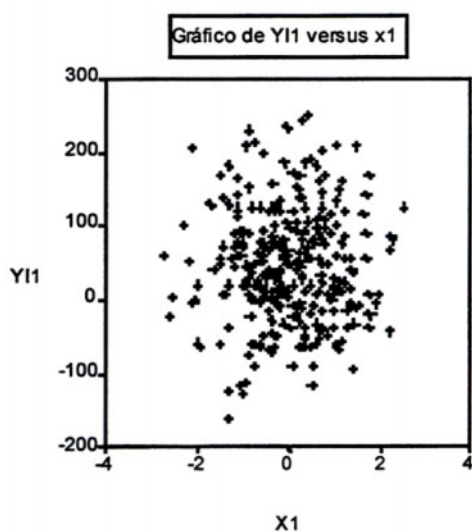
6.2.1.- DETECCIÓN DE LOS VALORES ATÍPICOS SOBRE LA VARIABLE ENDÓGENA

6.2.1.1 DETECCIÓN DE ATÍPICOS SOBRE EL MODELO 2 - ATÍPICO IMPULSO EN Y

6.2.1.1.A.- GRÁFICOS DE LA ENDÓGENA VERSUS LAS VARIABLES EXPLICATIVAS

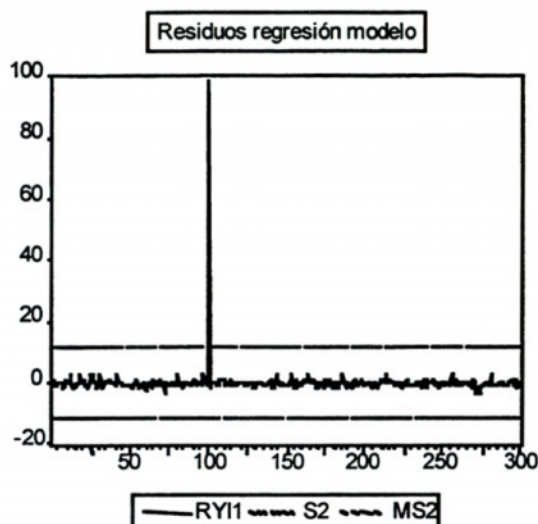
Tal y como se comentó anteriormente, el atípico incluido sobre la variable endógena no venía causado por ninguna de las variables explicativas, de manera que en estos gráficos no se puede apreciar ni distinguir el atípico tipo impulso incluido en Y. Únicamente, destacar el gráfico de Y_{11} versus X_4 que presenta un esquema diferente al del resto tal y como se puede

apreciar en los siguientes gráficos.



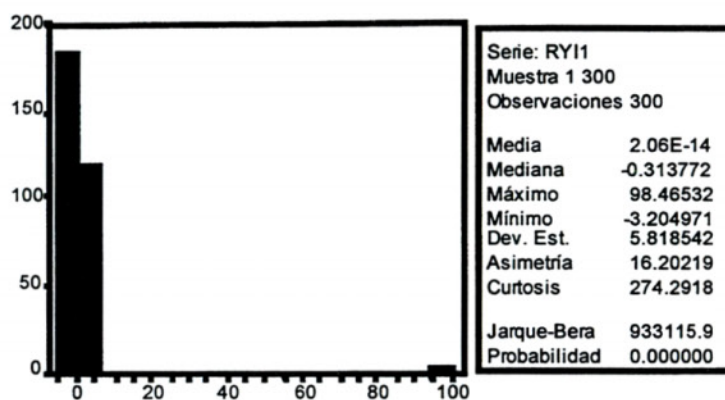
6.2.1.1.B.- GRÁFICO DE LOS RESIDUOS CON LAS BANDAS $\pm 2\sigma$

Se aprecia claramente el atípico en la observación 100, siendo este último el único valor que sobrepasa los límites $\pm 2\sigma$ siendo el resto de los residuos muy reducidos, tal y como se puede apreciar en el gráfico siguiente.



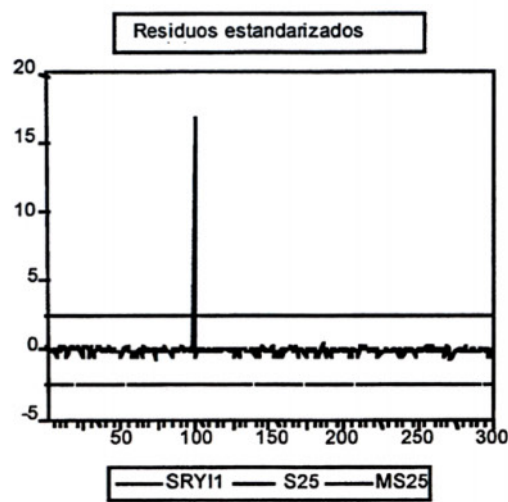
6.2.1.1.C.- HISTOGRAMA DE LOS RESIDUOS

El efecto del atípico en el histograma de los residuos es “contundente”, provocando la concentración de 199 valores residuales en el intervalo $(-5,5)$ y un único valor en el intervalo $(95,100)$. De esta forma la distribución es claramente leptocúrtica, muy asimétrica a la derecha, presentando así una fuerte desviación de la normalidad.



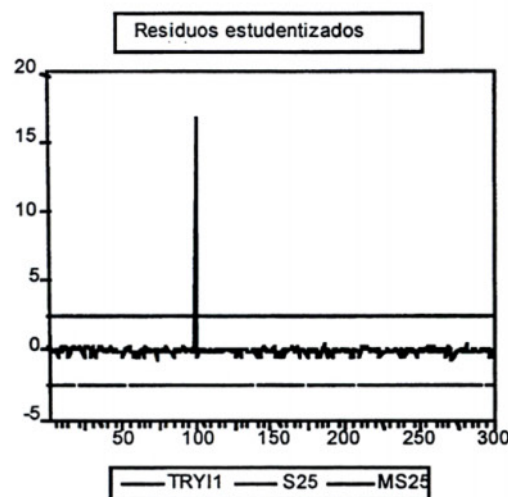
6.2.1.1.D.- RESIDUOS ESTANDARIZADOS

El gráfico de los residuos estandarizados, así como sus valores, no arrojan duda alguna. Únicamente el residuo estandarizado correspondiente a la observación 100 presenta un valor muy superior a los límites ± 2.5 establecidos como los valores críticos que permiten detectar la presencia de un posible atípico. El residuo estandarizado para la observación 100 arroja un valor de 16.80911.



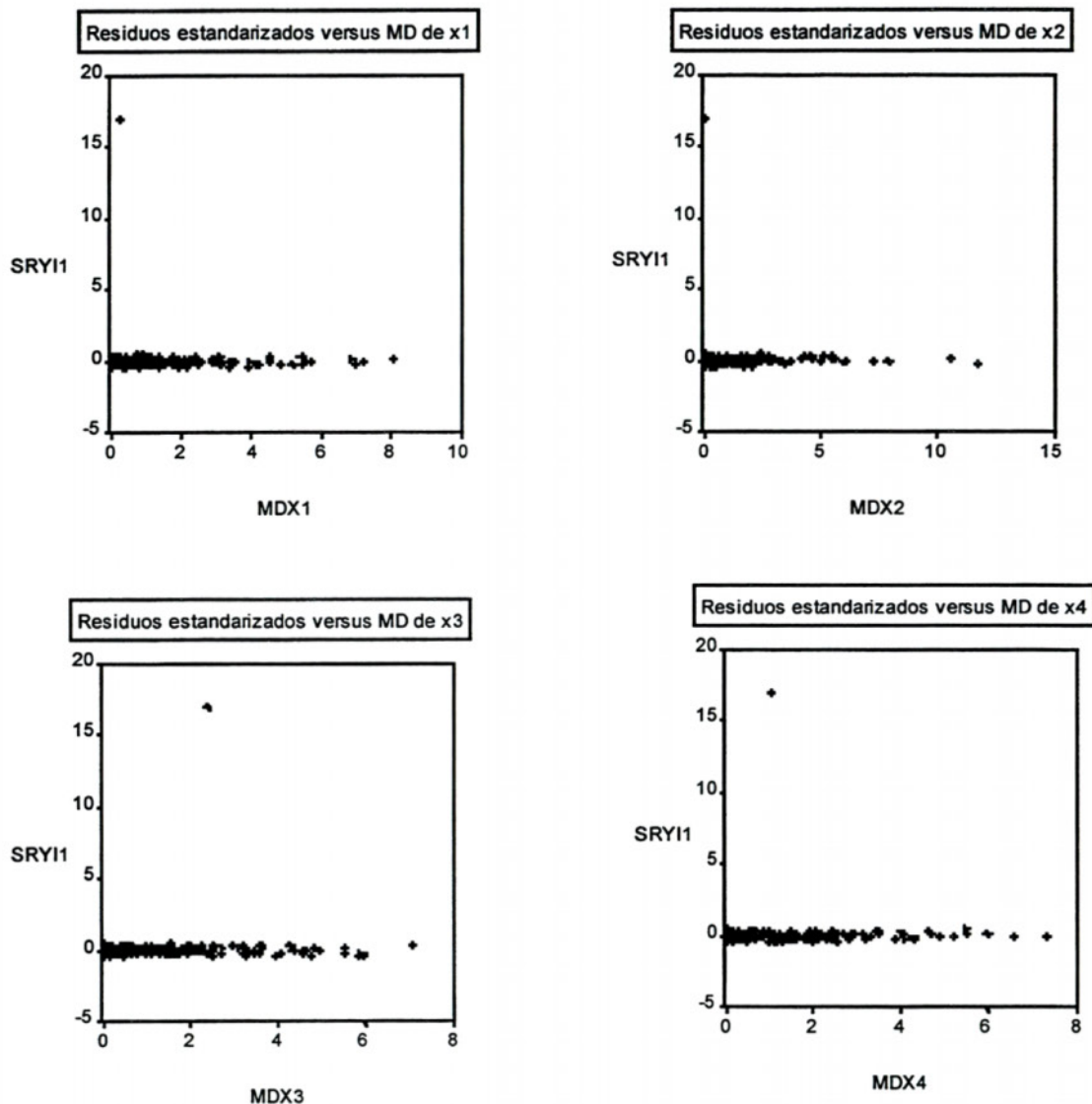
6.2.1.1.E.- RESIDUOS ESTUDENTIZADOS

Se extraen las mismas conclusiones que para el contraste anterior, siendo en este caso el valor para el residuo estudentizado correspondiente a la observación 100 de 16.91163, sin apreciarse ningún otro residuo que supere los límites ± 2.5 .



6.2.1.1.F.- RESIDUOS ESTANDARIZADOS VERSUS DISTANCIAS DE MAHALANOBIS ROBUSTAS DE LAS VARIABLES EXPLICATIVAS

Aunque el atípico tipo impulso en este caso no venía causado por ninguna de las variables explicativas decidimos realizar este contraste a modo ilustrativo, si bien a partir de estos gráficos se pueden extraer conclusiones en este sentido.



En los cuatro gráficos de los residuos estandarizados versus las distancias de Mahalanobis robustas de cada una de las variables explicativas se aprecia que un único punto se aleja del conjunto de las observaciones que se mantienen todas alrededor de 0. Debido a que en todos los casos, es decir para todas las MD, este hecho se repite, se evidencia que este atípico no viene causado por ninguna de las variables explicativas encontrándose pues en la variable endógena. Este sería pues un buen método para descubrir en qué dirección se produce el

atípico, es decir, si se encuentra en la variable dependiente o en alguna de las variables explicativas.

6.2.1.1.G.- MÉTODOS DE ELIMINACIÓN DE FILAS

Con el objeto de comprobar la influencia de la observación 100, donde se sitúa el atípico tipo impulso introducido en la variable endógena, se calculan los estadísticos de eliminación de fila comentados anteriormente.

Es importante destacar que estos estadísticos son válidos únicamente para detectar la presencia de atípicos tipo impulso. En el caso del atípico tipo escalón incluido en la variable endógena YI2, estos contrastes se pueden realizar pero nos obliga a eliminar las 150 últimas observaciones, lo que pone ciertas dudas sobre las conclusiones a cerca de los estadísticos.

Centrándonos ya en los resultados para la eliminación de la observación 100, podemos destacar los siguientes puntos:

- a) Todos los estadísticos coinciden en señalar la observación 100 como muy influyente.
- b) El contraste D_i , dedicado a medir la naturaleza crítica del punto, arroja un valor de 0.9097, siendo prácticamente igual a 1.0, valor que tanto Cook y Weisberg (1982) como Montgomery y Peck (1982) coinciden en señalar como un valor considerado como elevado. Por lo tanto, la naturaleza crítica de la observación 100 es muy importante.
- c) El efecto sobre las estimaciones de los parámetros, medido a través del estadístico $DFBETAS_{ij}$, también es bastante importante, aceptándose para los parámetros $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_3$ y $\hat{\beta}_4$ que la observación 100 es muy influyente sobre las estimaciones de los parámetros. Destaca el hecho de que esta influencia es más leve en el caso del parámetro $\hat{\beta}_2$ para el cual el estadístico no señala que el punto 100 sea muy influyente, tal y como se puede apreciar en la siguiente tabla en la que se recogen los resultados para este contraste.

Efecto de la observación 100 sobre la estimación de los parámetros

Parámetro	DFBETAS _{ij}
$\hat{\beta}_0$	1.478702 *
$\hat{\beta}_1$	0.704754 *
$\hat{\beta}_2$	-0.046662
$\hat{\beta}_3$	-2.369011 *
$\hat{\beta}_4$	-1.416701 *

donde * significa valores del estadístico que superan, en valor absoluto a $2/\sqrt{T}=0.11547$.

d) Los efectos de la eliminación de la observación 100 sobre los estadísticos t son también considerables. Aunque la eliminación de dicha fila no afecta a las conclusiones que se pueden extraer de dicho contraste (en todos los casos se verifica la significatividad estadística de los coeficientes estimados) si se aprecian cambios importantes en los valores de los estadísticos, como consecuencia de las alteraciones que se producen en los errores estándar de los estimadores.

De esta forma, los estadísticos DFTSTAT_{ij} presentan valores muy elevados, destacando principalmente el obtenido para el parámetro $\hat{\beta}_4$ presentando una diferencia entre los estadísticos superior a 1000 puntos.

Efecto de la observación 100 sobre los estadísticos t

Parámetro	DFTSTAT _{ij}
$\hat{\beta}_0$	-671.9597
$\hat{\beta}_1$	-41.5138
$\hat{\beta}_2$	163.7724
$\hat{\beta}_3$	-135.3469
$\hat{\beta}_4$	-1136.4885

e) El estadístico DFFITS_i, calculado para medir la influencia de la observación eliminada sobre los valores ajustados, arroja también un valor elevado ($DFFITS_{100} = 11.33854 > 2\left(k/T\right)^{1/2} = 0.258199$) que nos conduce a concluir, también en este caso, que la observación 100 es muy influyente sobre los valores

ajustados.

f) El efecto de la observación 100 sobre los residuos, medido a través del estadístico RSTUDENT es también muy importante, obteniéndose para la observación 100, $\hat{u}_{100}^* = 102.518$, valor extremadamente elevado, lo que pone de manifiesto la importancia de esta observación sobre los residuos estudentizados.

g) La matriz de varianzas-covarianzas de los estimadores también se ve afectada significativamente por la eliminación de la fila 100. El estadístico que mide esta influencia, COVRATIO, presenta un valor muy reducido, de manera que al realizar el contraste $|\text{COVRATIO} - 1|$ se obtiene un valor unitario muy superior a $\frac{3k}{T} = 0.05$, pudiéndose concluir que el efecto sobre la matriz de varianzas-covarianzas de la fila 100 es muy importante.

h) Ocurre exactamente lo mismo en la medición del efecto de la observación sobre la varianza de los valores ajustados mediante el estadístico FVARATIO, concluyéndose también que la observación es muy influyente.

Finalmente, quisiera comentar brevemente cuales son los principales efectos de la eliminación de la observación 100 sobre el resto de los estadísticos que no se tienen en cuenta a la hora de construir estos contrastes y que habitualmente se analizan en econometría. Para su mejor interpretación, se resumen en la siguiente tabla:

Efectos de la eliminación de la observación 100 sobre otros estadísticos		
Estadísticos	Con todas las observaciones	Sin la observación 100
$\bar{R}^2$	0.9941	0.9998
$\hat{\sigma}$	5.857857	0.964835
SSR	10122.77	273.6863
AIC	3.552094	-0.055016
S(p,q)	3.613824	0.006864
LBQ(12)	0.9100	9.3320
LBQ(24)	1.1919	17.889
LBQ(36)	2.1906	23.927

Los efectos más importantes de la eliminación de la observación 100 se aprecian de forma más acusada en ciertos estadísticos que en otros destacando, por un lado, las importantes reducciones que se observan en la desviación estándar ($\hat{\sigma}$), la suma de residuos al cuadrado (SSR), los valores para los criterios de selección de modelos de Akaike y Schwarz, pudiéndose concluir que, en este caso, la eliminación de la observación 100 tiene también importantes efectos reductores sobre estos estadísticos. Por otro lado, destacar el incremento, aunque leve, del coeficiente de determinación corregido ($\bar{R}^2$) al eliminar la observación 100. Si bien, con todas las observaciones, este coeficiente era muy elevado, la eliminación de la observación 100 produce un ajuste casi perfecto, pudiendo decirse que prácticamente se explica el 100% de la variable endógena.

También destaca el importante incremento que se observa en los estadísticos de Ljung-Box, confirmándose el resultado conocido de que un atípico impulso de magnitud elevada tiene importantes efectos en las funciones de autocorrelación, efectos que se traducen en valores de las autocorrelaciones prácticamente nulas, llevando a la identificación de procesos de ruido blanco. En este caso, aunque se incrementan las autocorrelaciones de los residuos y los estadísticos de Ljung-Box, se sigue aceptando la hipótesis de que los residuos no tienen ninguna estructura y son ruido blanco.

6.2.1.2.- DETECCIÓN DE ATÍPICOS SOBRE EL MODELO 3

- Atípico escalón en Y

Debido a las especiales características que supone la inclusión de un atípico tipo escalón, y en especial un escalón como el incluido en este caso, desde la observación 150 hasta el final de la muestra, se han realizado ciertos contrastes de eliminación de filas. El objetivo fundamental de estos contrastes es analizar como se verán afectados los principales estadísticos de la regresión cuando sólo consideramos la mitad de la muestra.

Aunque se ha hecho alusión a las especiales características de este atípico, no se puede decir que este no se produzca nunca. De hecho, el efecto de la devaluación de una moneda de gran magnitud genera un esquema muy similar en los tipos de cambio, por ejemplo,

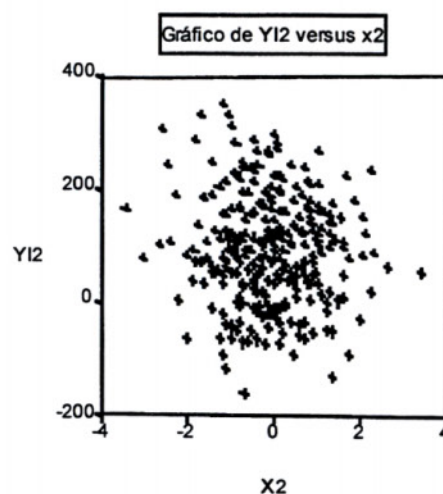
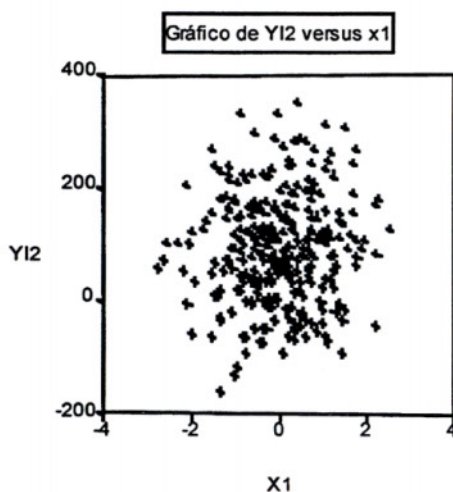
poniéndose pues de manifiesto la dificultad de su detección si no se conocen las causas del atípico.

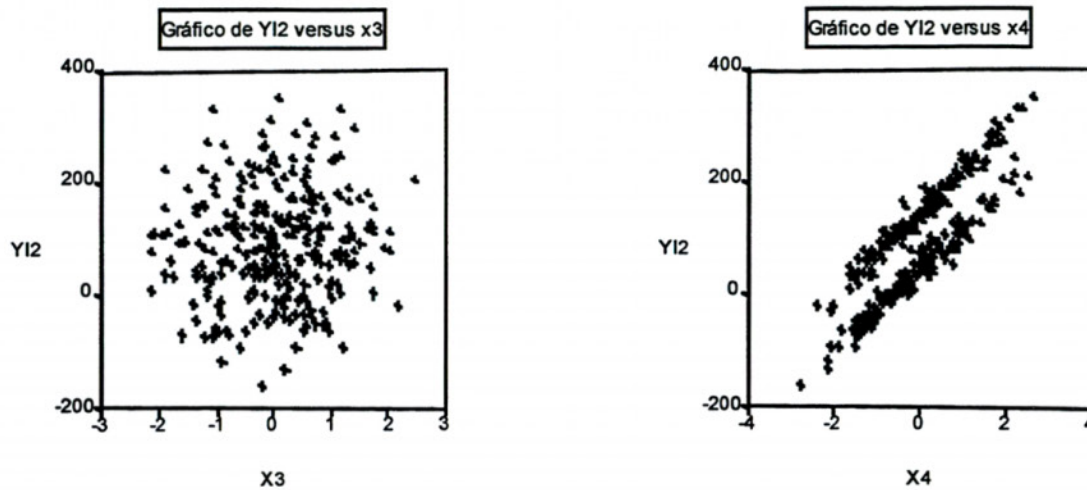
Después de esta breve introducción, pasamos a comentar los resultados de los contrastes de atípicos, siguiendo el mismo esquema que en el caso del atípico tipo impulso.

6.2.1.2.A.- GRÁFICOS DE LA ENDÓGENA VERSUS LAS VARIABLES EXPLICATIVAS

Al haberse incluido el atípico tipo escalón directamente sobre la variable endógena, este contraste de atípicos aporta poco a la detección del atípico ya que este método se encamina a decidir si el atípico se encuentra en alguna de las variables explicativas.

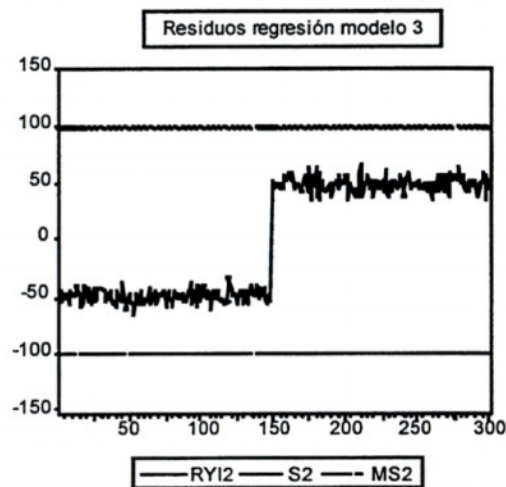
De hecho los gráficos que se presentan no nos permiten obtener ninguna conclusión. El único comentario que cabría realizar sobre éstos se refiere al gráfico de la variable endógena versus la variable x_4 , en el que se aprecia la misma tendencia observada que en el caso del atípico tipo impulso, aunque en esta ocasión aparece claramente dividida en dos tendencias, como consecuencia del cambio en el nivel de la variable endógena tras la inclusión del atípico tipo escalón en la observación 150.



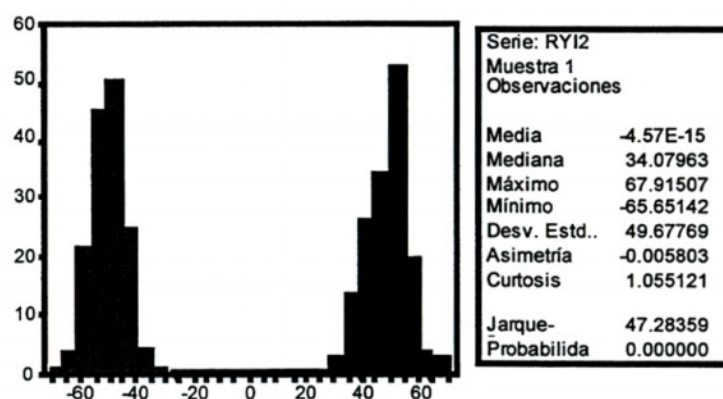


6.2.1.2.B.- GRÁFICO DE LOS RESIDUOS CON LAS BANDAS $\pm\sigma$

En este caso, no se aprecia ningún residuo que sobrepase los límites $\pm 2\hat{\sigma}$ debido a la elevada varianza residual que se obtiene en este modelo al no considerarse el tratamiento del atípico. Sin embargo, si se aprecian unos residuos que presentan dos niveles claramente distintos, indicando la presencia del atípico tipo escalón en la observación 150.



6.2.1.2.C.- HISTOGRAMA DE LOS RESIDUOS

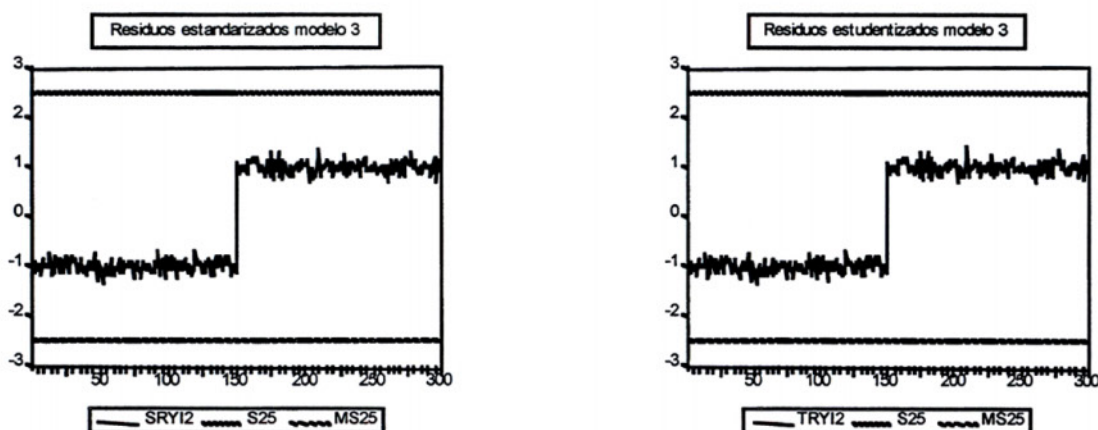


El efecto del atípico tipo escalón en el histograma de los residuos es muy fuerte, presentando un esquema muy peculiar. Aparecen dos distribuciones de frecuencias, concentrados en dos medias distintas, que en valor absoluto podrían considerarse casi iguales. Además, cada distribución de frecuencias presenta un esquema muy semejante a la distribución normal aunque, globalmente, no podría considerarse que los residuos sigan una distribución normal. De hecho, el estadístico Jarque-Bera arroja un valor de 47.28359, lo que supone un rechazo contundente de normalidad.

Un hecho importante que corroboraría los comentarios efectuados acerca de las dos medias es que la media total para todas las observaciones es prácticamente nula, presentándose unos valores máximo y mínimo muy distantes entre sí debido a las dos distribuciones apreciadas.

6.2.1.2.D.- RESIDUOS ESTANDARIZADOS:

El gráfico de los residuos estandarizados comparándolos con los valores “críticos” ± 2.5 es muy similar al realizado en el contraste b). Ningún valor sobrepasa los límites críticos pero claramente se observa un cambio de nivel en los residuos a partir de la observación 150.



6.2.1.2.E.- RESIDUOS ESTUDENTIZADOS

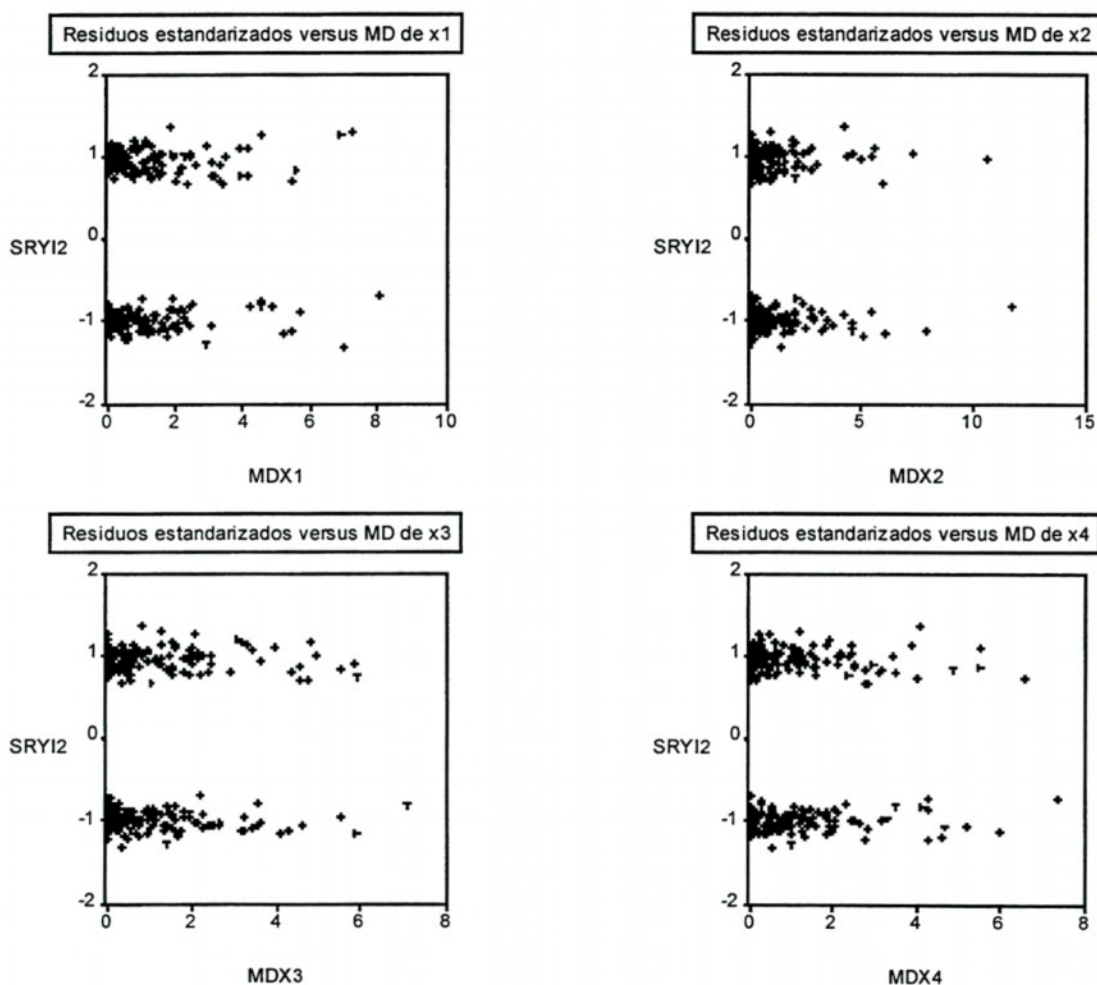
En el gráfico anterior, donde se presentan los residuos estudentizados con los límites ± 2.5 , se observa exactamente el mismo esquema que en el contraste anterior por lo que se pueden extrapolar los comentarios anteriores a este contraste.

6.2.1.2.F.- RESIDUOS ESTANDARIZADOS VERSUS DISTANCIAS DE MAHALANOBIS ROBUSTAS DE LAS VARIABLES EXPLICATIVAS.

De nuevo realizamos este contraste con un único propósito ilustrativo debido a que el atípico tipo escalón no venía causado por ninguna de las variables explicativas del modelo. Sin embargo, los gráficos que se obtienen nos permiten realizar ciertos comentarios interesantes.

Todos los gráficos presentan un esquema muy parecido, observándose una concentración de los puntos en torno a dos valores: -1 y 1, hecho que coincide con las dos distribuciones de frecuencias observadas en el histograma de los residuos.

Además, ningún punto sobrepasa los valores críticos para los residuos estandarizados (± 2.5) ni se observa ningún punto que se encuentre en las categorías reseñadas en el texto que pudiera indicarnos la presencia de algún atípico.



6.2.1.2.G.- ELIMINACIÓN DE FILAS

En este caso, conociendo que la variable endógena contiene un atípico tipo escalón en la observación 150, se procedió a calcular estos estadísticos eliminando de la observación 150 a la 300 con el objeto de verificar cuáles serían los principales efectos de esta eliminación sobre los principales estadísticos. No se calculan exactamente los mismos que para el caso del atípico tipo impulso, aunque las medidas que aquí presentamos creemos que son suficientemente ilustrativas como para extraer conclusiones importantes. Pasamos a continuación a analizar estos resultados.

a) Naturaleza crítica de las observaciones eliminadas: El elevado valor que se obtiene para esta medida de influencia ($D_i = 18904.007$) indica que las observaciones eliminadas tienen un efecto muy importante en el conjunto de la regresión, hecho que también se verifica en las medidas de eliminación de filas que se detallan a continuación.

b) El efecto de la eliminación de las filas sobre los parámetros se analiza a través de dos estadísticos. Por un lado, el estadístico $DFBETA_i$ y por otro el estadístico $DFBETAS_{ij}$, estando recogidos los resultados para ambos en la siguiente tabla.

Efecto de las observaciones 150-300 sobre los parámetros

Parámetros	$DFBETA_i$	$DFBETAS_{ij}$
$\hat{\beta}_0$	50.405	623.2746
$\hat{\beta}_1$	4.3850	50.9786
$\hat{\beta}_2$	-1.4766	-19.0026
$\hat{\beta}_3$	-2.7554	-30.3323
$\hat{\beta}_4$	3.1749	40.4439

En los resultados anteriores, se aprecia, para ambos estadísticos, que el parámetro más afectado por la eliminación de filas es el asociado al término constante, que presenta una reducción en el valor de la estimación considerable. En el resto de los estadísticos los resultados no son tan fuertes, apreciándose cambios menos

importantes. Sin embargo, conviene señalar que los resultados para el estadístico $DFBETAS_{ij}$ son mayores que los obtenidos para el otro contraste.

c) Los efectos sobre los estadísticos t de la eliminación de las filas 150 a 300 se miden a través del estadístico $DFTSTAT_{ij}$, que en este caso arrojó los siguientes resultados:

Efectos de la eliminación de las filas 150-300 sobre los estadísticos t

Parámetros	$DFTSTAT_{ij}$
$\hat{\beta}_0$	-527.6445
$\hat{\beta}_1$	-31.7972
$\hat{\beta}_2$	134.3238
$\hat{\beta}_3$	-107.9125
$\hat{\beta}_4$	-918.0712

En este caso se aprecian importantes cambios en los estadísticos t , destacando principalmente las variaciones observadas para los parámetros $\hat{\beta}_0$ y $\hat{\beta}_4$, que presentan cambios muy importantes con respecto a los estadísticos obtenidos al estimar con todas las observaciones. Destacar que en todos los casos, los estadísticos t multiplican sus valores cuando se eliminan las observaciones, de manera que las filas presentan una influencia muy elevada.

d) El estadístico COVRATIO, encargado de medir el efecto de las observaciones eliminadas sobre la matriz de varianzas-covarianzas de los estimadores, arroja un valor muy reducido, de manera que el contraste $|\text{COVRATIO} - 1| = 1$, lo que pone de manifiesto que, el efecto de las observaciones 150 a 300 es muy importante pudiendo decirse que las observaciones eliminadas tienen un efecto significativo sobre las covarianzas.

La detección de atípicos llevada a cabo en esta parte del trabajo nos permite establecer dos conclusiones principales:

1) El atípico tipo impulso ha sido detectado por todos los métodos utilizados, no

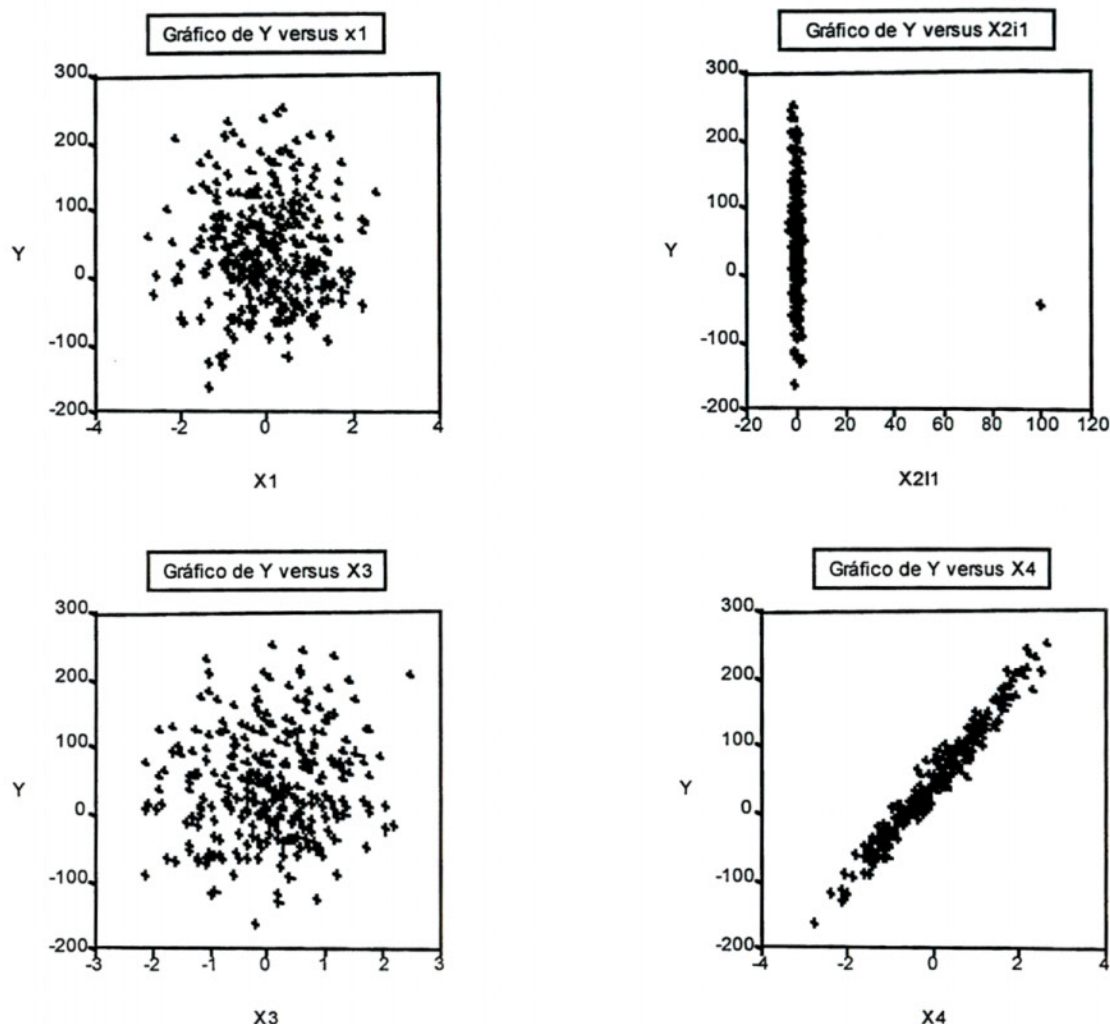
presentándose ninguna duda debido a que todos los contrastes concluyen en la misma dirección. Los métodos de eliminación de filas indican todos que la observación 100 es un punto muy influyente.

2) El atípico tipo escalón no ha sido detectado por ninguno de los contrastes efectuados, lo que puede poner en entredicho la validez de estos métodos cuando el atípico que se presenta es tipo escalón. Destacar además que las medidas de influencia, a través de la eliminación de filas, sugieren que dichas observaciones presentan un efecto importante.

6.2.2.- DETECCIÓN DE ATÍPICOS SOBRE LAS VARIABLES EXPLICATIVAS

6.2.2.1.- DETECCIÓN DE ATÍPICOS SOBRE MODELO 4 - ATÍPICO IMPULSO EN X2

6.2.2.1.A.- GRÁFICO DE LA ENDÓGENA VERSUS LAS EXPLICATIVAS

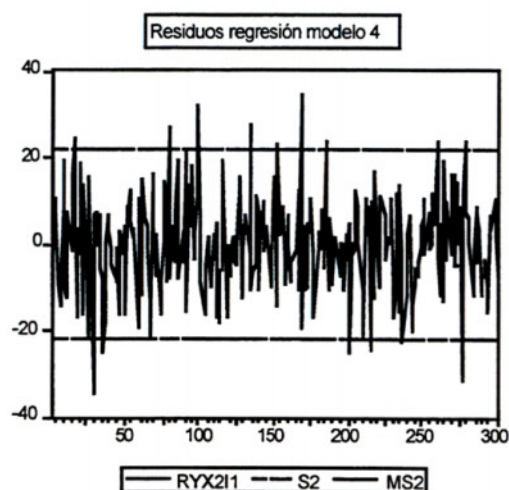


En este caso, resulta muy ilustrativo el gráfico de la variable endógena versus la explicativa X2I1, que es variable que presenta el atípico tipo impulso. El resto de los gráficos presentan un esquema muy similar al observado en el caso del atípico tipo impulso en Y.

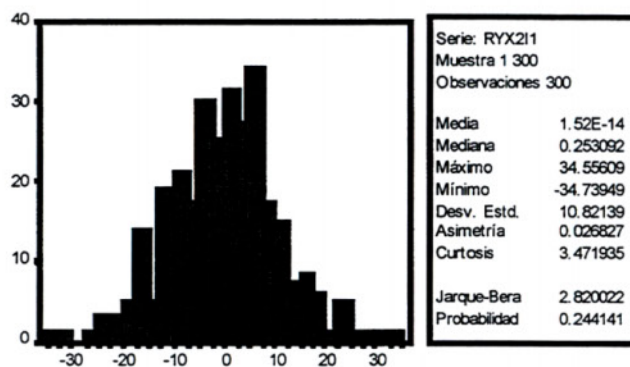
En el gráfico de Y versus X2I1, se aprecia claramente un único punto distanciado del resto de las observaciones, en dirección de X2 lo que pone de manifiesto que la variable explicativa presenta un atípico. Se trataría pues de un punto influyente en dirección X2.

6.2.2.1.B.- GRÁFICO DE LOS RESIDUOS CON LAS BANDAS

En este gráfico, se pretende descubrir aquellas observaciones que sobrepasan las bandas $\pm\hat{\sigma}$. Destacar que numerosos valores de los residuos sobrepasan estas bandas ($\pm\hat{\sigma} = 21.789$) destacando cuatro residuos que superan los 30 puntos. Una de estas observaciones se sitúa en la observación 100, pudiendo venir causado este elevado valor residual por el atípico tipo impulso incluido en X2.



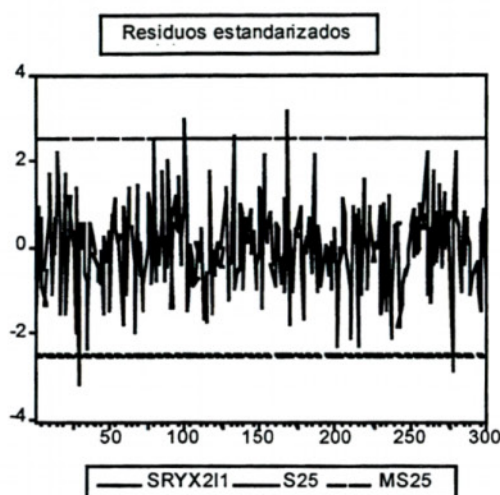
6.2.2.1.C.- HISTOGRAMA DE LOS RESIDUOS



Los numerosos residuos sobrepasando los límites $\pm\hat{\sigma} = 21.789$ también se aprecian en el histograma de los residuos que presenta siete observaciones en las colas que superan ± 25 . Estos valores se corresponden con las observaciones realizadas en el gráficos anterior.

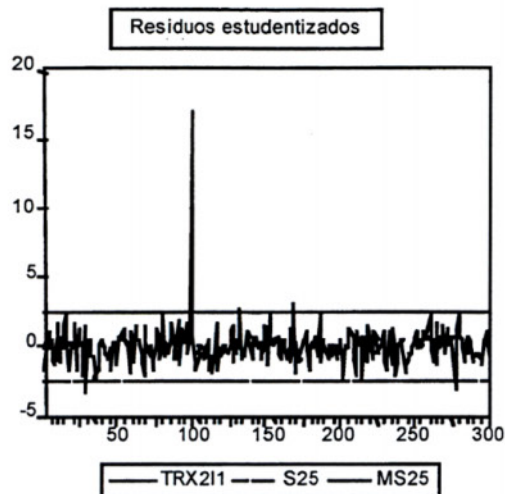
6.2.2.1.D.- RESIDUOS ESTANDARIZADOS

El residuo estandarizado correspondiente a la observación 100 presenta un valor superior a ± 2.5 , valor crítico, lo que nos puede estar indicando que debería ser un punto a investigar. Sin embargo, no es el valor más elevado de los residuos estandarizados; existen otros valores superiores que deberían investigarse, tal y como se puede apreciar en el siguiente gráfico.



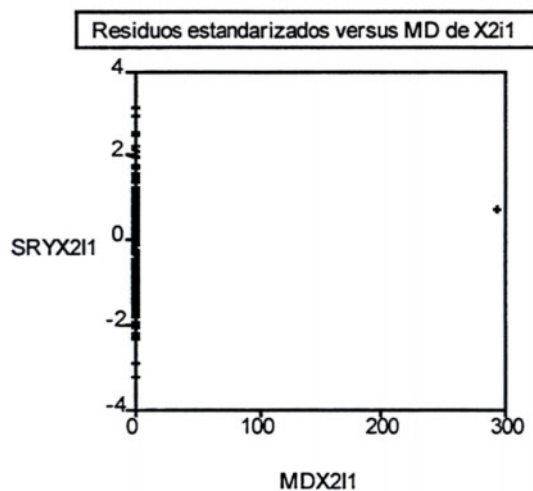
6.2.2.1.E.- RESIDUOS ESTUDENTIZADOS

A diferencia de lo que ocurre con los residuos estandarizados, se aprecia claramente el residuo correspondiente a la observación 100 que supera claramente los límites ± 2.5 . Se aprecian también otros valores elevados que sobrepasan estos límites, pero en ningún caso presentan un t_i^* tan elevado. La razón de esto es la consideración de la matriz H , que incluye la información de las variables explicativas, y en consecuencia, de $X2I1$, hecho que no ocurre con los residuos estandarizados. En este caso, el residuo estudentizado presenta un valor de 17.108, claramente superior a ± 2.5 , siendo el residuo de la regresión igual a 32.10268 y $\sqrt{1 - h_{ii}} = 0.17272$.



6.2.2.1.F.- RESIDUOS ESTANDARIZADOS VERSUS DISTANCIAS DE MAHALANOBIS ROBUSTAS DE LAS VARIABLES EXPLICATIVAS

En este caso, únicamente presentamos el gráfico correspondiente a los residuos estandarizados versus la distancia de Mahalanobis robusta de la variable X2I1. El resto de los gráficos no aportan nada debido a que el atípico se encuentra en esta variable explicativa.



En el gráfico anterior, claramente se aprecia un punto distanciado del resto de las observaciones que se sitúan en torno al valor 0 para la MD_i^2 de X2I1. Este punto se corresponde con la observación 100, con un residuo estandarizado de 2.95 y una MD_i^2 de X2I1 de 290.8229.

6.2.2.1.G.- MEDIDAS DE INFLUENCIA - ELIMINACIÓN DE FILAS

Debido a que algunos de los contrastes anteriores nos señalaban la aparición de un atípico en la observación 100 y conociendo nosotros de antemano que el atípico se incluía en la observación 100 de la variable X2, se procedió al cálculo de las principales medidas de influencia de la observación 100. Pasamos a comentar brevemente los resultados obtenidos para cada medida.

El estadístico D_i para medir la naturaleza crítica del punto eliminado arroja un valor de 1903.73 lo que pone de manifiesto que la observación 100 es claramente influyente.

Por lo que respecta a los efectos sobre las estimaciones de los parámetros, el estadístico $DFBETA_i \equiv \hat{\beta} - \hat{\beta}_{(-100)}$ arroja los siguientes resultados:

Parámetro	DFBETA _i
$\hat{\beta}_0$	1.28078
$\hat{\beta}_1$	1.382647
$\hat{\beta}_2$	10.465749
$\hat{\beta}_3$	0.632326
$\hat{\beta}_4$	-0.92448

donde claramente se observa que el parámetro más afectado por la eliminación de la fila 100 es el asociado a la variable X2I1, variable que contiene el atípico tipo impulso en la observación 100.

Algo semejante se observa en el estadístico $DFBETAS_{ij}$ cuyos resultados fueron:

Parámetro	DFBETAS _{ij}
$\hat{\beta}_0$	22.9459
$\hat{\beta}_1$	24.0413
$\hat{\beta}_2$	1092.546
$\hat{\beta}_3$	10.1991
$\hat{\beta}_4$	-16.9977

El mayor efecto se observa para el coeficiente $\hat{\beta}_2$ que presenta un valor muy superior al del resto de los estadísticos. Sin embargo, para el resto de los parámetros, el efecto de la eliminación de la observación 100 no es despreciable.

Los efectos sobre los estadísticos t también son importantes. Los resultados de $DFTSTAT_{ij}$ fueron:

Parámetro	$DFTSTAT_{ij}$
$\hat{\beta}_0$	-732.2789
$\hat{\beta}_1$	-43.6561
$\hat{\beta}_2$	193.1161
$\hat{\beta}_3$	-145.1566
$\hat{\beta}_4$	-1240.4705

presentando pues valores muy elevados, indicando que los estadísticos t presentan grandes variaciones. Pone de nuevo de manifiesto que la eliminación de la observación 100 tiene efectos muy significativos sobre los estadísticos.

Sin embargo, el efecto sobre los valores ajustados, medido a través de $DFFITS_i$ arroja un valor, para la observación 100 de 0.190276, menor que el valor crítico $2\sqrt{\frac{k}{T}}=0.2582$, indicándonos que en este caso, la eliminación de dicha observación no presenta un gran efecto, pudiendo concluir a través de este estadístico que se muestra poco influyente.

Algo semejante ocurre con el efecto sobre los residuos, medido a través de $RSTUDENT_i$, siendo el valor para la observación 100 de 0.033366, indicando que la observación 100 en este caso es poco influyente.

Las dos últimas medidas de la influencia de la observación 100 calculadas en este apartado ponen de manifiesto la elevada influencia de la observación 100. Se trata de las medidas $COVRATIO$ y $FVARATIO$, definidas ambas en el texto, siendo los valores obtenidos:

$$|COVRATIO - 1| = 0.9999$$

$$|FVARATIO - 1| = 0.737091$$

ambos superiores a $3k/T=0.05$ lo que indica que, para estas dos medidas, la observación 100 aparece como claramente influyente.

Los efectos de la observación 100 sobre el resto de los estadísticos se analizarán a través de los resultados que se presentan en la siguiente tabla:

Efectos de la eliminación de la observación 100 sobre otros estadísticos		
Estadísticos	Con todas las observaciones	Sin la observación 100
$\bar{R}^2$	0.9798	0.9998
$\hat{\sigma}$	10.89451	0.964835
SSR	35013.66	273.6863
AIC	4.793044	-0.055016
S(p,q)	4.854774	0.006864
LBQ(12)	18.027	9.3320
LBQ(24)	27.793	17.889
LBQ(36)	44.403	23.927

Los principales comentarios que se pueden efectuar a partir de los resultados anteriores son los siguientes:

- 1) La eliminación de la observación 100 provoca la consecución de un ajuste prácticamente perfecto, pasando de un 98% de explicación de la endógena considerando todas las observaciones a un 99.98% al eliminar la observación 100.
- 2) Se produce una importante disminución en el error estándar de la regresión, llegando a ser, en el caso de la eliminación de la observación 100, muy similar a la del modelo inicial. Este se ve acompañado de la importante reducción que se observa en la suma residual: una única observación genera una reducción de la suma residual superior al 78%.
- 3) Destacar asimismo la reducción que se observa en los valores del AIC y S(p,q), fruto únicamente de la eliminación de una única observación.
- 4) Observar finalmente la fuerte disminución que se produce en los estadísticos de Ljung-Box, no rechazándose la hipótesis de que los residuos son ruido blanco.

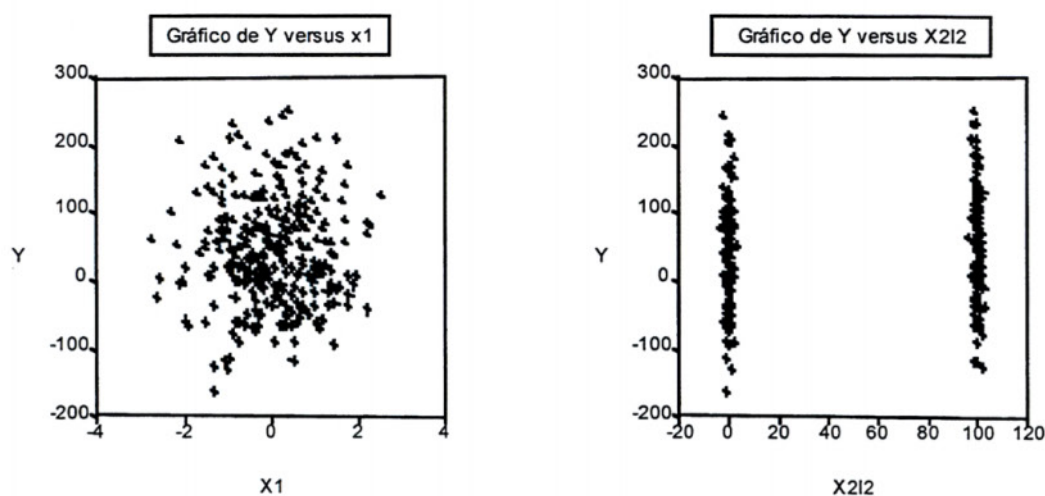
La detección del atípico tipo impulso incluido en la variable X_2 no ha presentado resultados tan concluyentes como en el caso de este mismo atípico en la variable endógena. Así como en este último caso los contrastes de detección no ofrecían duda alguna sobre donde se encontraba el atípico, en este caso, dos contrastes gráficos nos han ayudado claramente en esta dirección. Estamos hablando, por un lado, del gráfico de la variable endógena versus la variable explicativa X_{2I1} y, por otro lado, del gráfico de los residuos estandarizados versus la distancia de Mahalanobis robusta de X_{2I1} , gráficos que, en ambos casos, nos indicaban claramente que el atípico, detectado principalmente a través de los residuos estudentizados, se encontraba en la variable explicativa X_{2I1} .

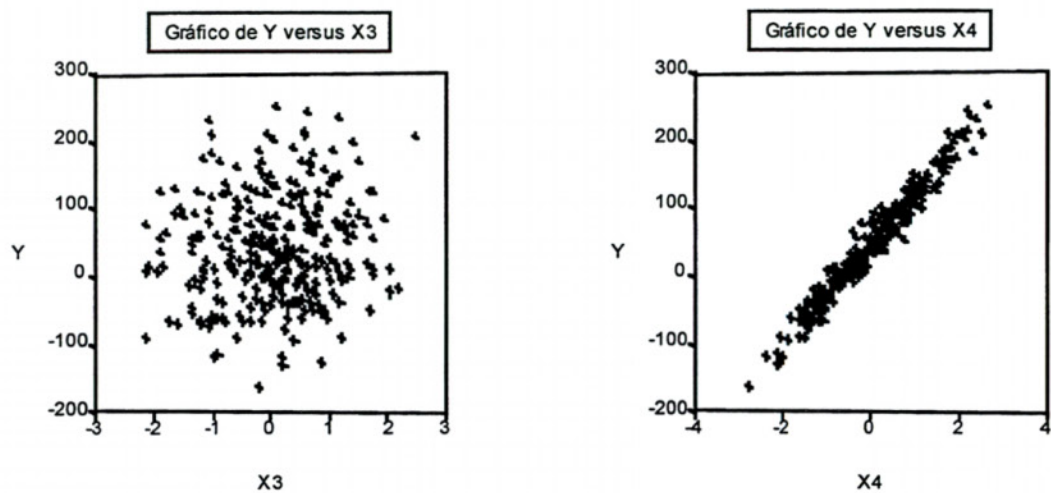
El resto de los contrastes no se mostraron muy adecuados para la detección de este tipo de atípicos en las variables explicativas ya que no nos permitieron, en algunos de los casos decidir si era un atípico y, en otros casos, donde se encontraba el atípico.

Hay que decir, sin embargo, que este hecho es bastante lógico debido a que la mayor parte de estos contrastes se basan en unos residuos que no provienen de una estimación robusta (la estimación mínimo cuadrática no es robusta frente a los atípicos) y pueden verse afectados por el efecto enmascaramiento.

6.2.2.2.- DETECCIÓN DE ATÍPICOS SOBRE MODELO 5 - ATÍPICO ESCALÓN EN X_2

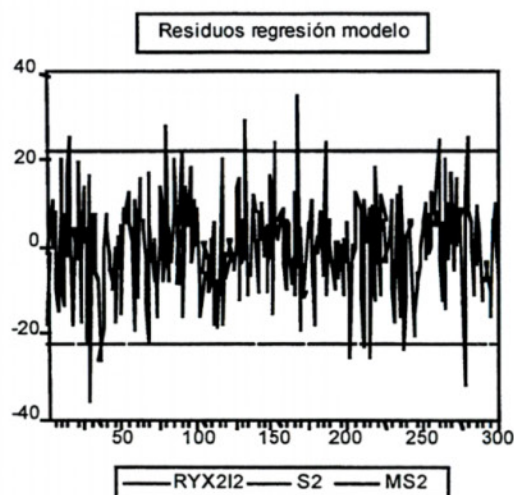
6.2.2.2.A.- GRÁFICO DE LA ENDÓGENA VERSUS LAS VARIABLES EXPLICATIVAS





Tal y como ocurría en el caso anterior, el gráfico de la endógena versus X2I2 es muy ilustrativo debido a que se observan dos esquemas de comportamiento claramente diferenciados. Mientras que para ambos esquemas, los valores de la variable endógena siempre oscilan entre -150 y 250, los de X2I2 presentan dos valores diferentes, dando paso a las dos relaciones observadas. Por un lado, oscilan alrededor de 0 y, en otro lado, claramente diferenciado del anterior, oscilan alrededor de 100. El resto de los gráficos presentan los mismos esquemas que en el caso anterior, luego los comentarios anteriormente realizados son válidos en este caso.

6.2.2.2.B.- GRÁFICO DE LOS RESIDUOS CON LAS BANDAS $\pm\sigma$

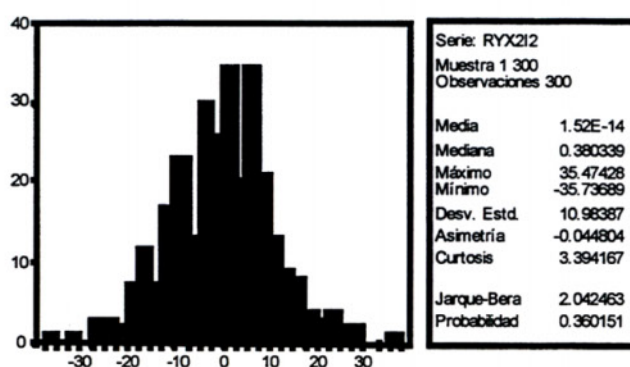


Tal y como se puede apreciar, el gráfico no es nada ilustrativo del atípico tipo escalón introducido en X2 debido a que no se aprecia ningún comportamiento diferente a partir de la observación 150, momento en el que se incluye el escalón en la variable X2. Se observan

valores superiores a las bandas, pero sin seguir un esquema tipo escalón ni coincidiendo con el momento en que éste se produce.

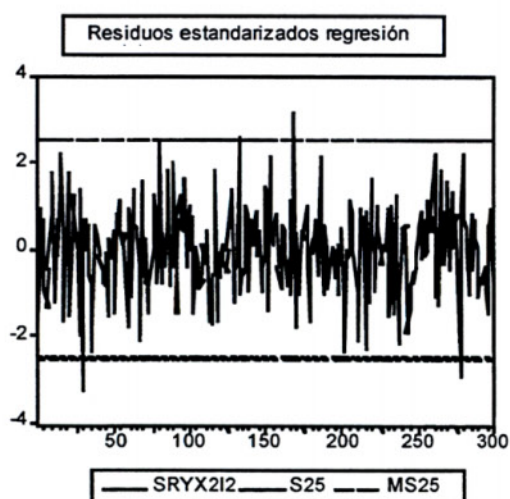
6.2.2.2.C.- HISTOGRAMA DE LOS RESIDUOS

Aparecen siete valores superiores a ± 25 , valores que se corresponden con los que en el gráfico de los residuos anterior se salían de las bandas $\pm\sigma$. Sin embargo, no se ve reflejado el atípico tipo escalón en X2I2.



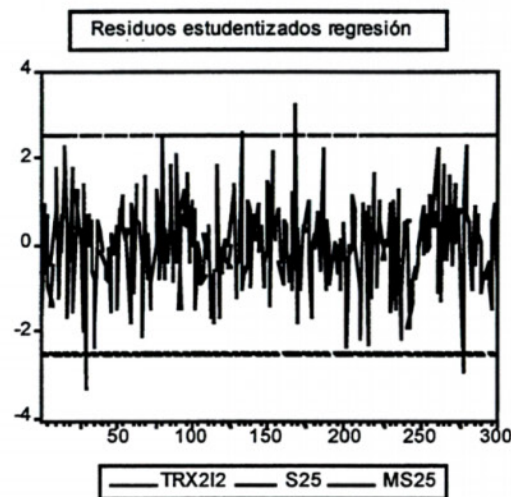
6.2.2.2.D.- RESIDUOS ESTANDARIZADOS

De nuevo, diversos valores superan los límites críticos ± 2.5 , pero ninguno se corresponde con el atípico tipo escalón de la observación 150 en X2. Asimismo, no se aprecia ningún comportamiento distinto a partir de la observación 150.



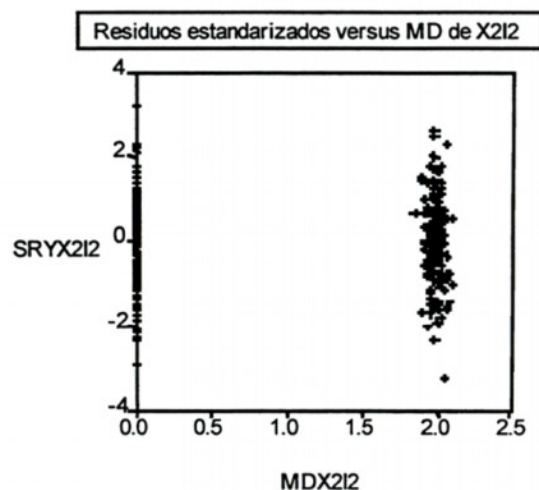
6.2.2.2.E.- RESIDUOS ESTUDENTIZADOS

Los resultados que se presentan en el gráfico siguiente son muy similares a los expuestos para los residuos estandarizados de manera que los comentarios realizados en el caso anterior se pueden hacer extensibles a esta situación.



6.2.2.2.F.- RESIDUOS ESTANDARIZADOS VERSUS DISTANCIA DE MAHALANOBIS ROBUSTA DE X2I2

Aparecen dos esquemas de comportamiento claramente diferenciados. Por un lado, valores que presentan residuos estandarizados comprendidos entre -2.5 y 2.5 y con distancias de Mahalanobis robustas prácticamente nulas. Por otro lado, se observa otro conjunto de valores con residuos estandarizados muy similares a los anteriores pero con valores para la distancia de Mahalanobis superiores, estando centrados éstos en torno al valor 2.0. Este es uno de los métodos que mejor nos indican la presencia de un atípico, en este caso escalón, en la variable explicativa X2I2.



6.2.2.2.G.- MEDIDAS DE INFLUENCIA - ELIMINACIÓN DE FILAS

De nuevo en este caso, al no ser los contrastes demasiado ilustrativos sobre el momento en que se encontraba el valor atípico, pero conociendo de antemano en dónde y en qué momento se producía, procedimos a calcular las principales medidas de influencia eliminando las filas de la 150 a la 300. Los resultados obtenidos para las diferentes medidas se presentan y se comentan a continuación.

1) Naturaleza crítica de los puntos 150-300: EL estadístico en este caso arroja un valor muy elevado ($D_i=284505.98$) poniendo de manifiesto la importancia de dichas observaciones.

2) Efecto sobre los parámetros estimados: Se miden a través de dos estadísticos. Por lo que respecta a $DFBETA_i$ se obtuvieron los siguientes resultados:

Parámetro	$DFBETA_i$
$\hat{\beta}_0$	1.00106
$\hat{\beta}_1$	1.335123
$\hat{\beta}_2$	10.772398
$\hat{\beta}_3$	0.805769
$\hat{\beta}_4$	-0.80444

apreciándose claramente que el mayor efecto se produce para el parámetro $\hat{\beta}_2$, asociado a la variable con el atípico escalón.

Por lo que respecta al segundo estadístico, $DFBETAS_{ij}$, los resultados fueron:

Parámetro	$DFBETAS_{ij}$
$\hat{\beta}_0$	-12.3783
$\hat{\beta}_1$	15.5217
$\hat{\beta}_2$	138.6320
$\hat{\beta}_3$	8.8702
$\hat{\beta}_4$	-10.2474

donde de nuevo, destaca el elevado valor que se obtiene para el parámetro $\hat{\beta}_2$. Esto demuestra que, en cuanto al efecto de las filas 150 a 300 sobre los parámetros, el

coeficiente $\hat{\beta}_2$ es el que más sufre sus consecuencias. Este hecho es lógico debido a que precisamente la variable asociada a ese parámetro es la que contiene el atípico escalón objeto del análisis.

3) Efecto sobre los estadísticos t: Los efectos de la eliminación de las filas 150 a 300 sobre los estadísticos t son muy importantes, a tenor de los resultados que se producen:

Parámetro	DFTSTATS _{ij}
$\hat{\beta}_0$	-509.7709
$\hat{\beta}_1$	-27.77114
$\hat{\beta}_2$	138.6946
$\hat{\beta}_3$	-94.91056
$\hat{\beta}_4$	-827.8469

sin que se pueda diferenciar claramente alguno de ellos debido a que se producen importantes alteraciones entre estos valores. Se puede destacar que el estadístico t asociado a $\hat{\beta}_2$ pasa de no permitir rechazar la hipótesis nula de no significatividad del parámetro a presentar un valor que nos indica que ese coeficiente es claramente significativo.

4) Efecto sobre la matriz de varianzas-covarianzas de los estimadores: De nuevo, en este caso, la medida CORVATIO arroja un valor muy reducido ($2.3108 \cdot 10^{-6}$) que, al realizar el contraste $|\text{COVRATIO} - 1| = 0.9999$ lo que claramente supera al valor crítico $3k/T$. Este resultado nos lleva a concluir que las observaciones eliminadas son muy influyentes.

Destacar que el resto de los contrastes ó medidas de influencia construidos para los casos de los atípicos tipo impulso en este caso no tienen sentido alguno debido a que tratan de medir la influencia de una única observación eliminada, encontrándonos ahora con la eliminación de múltiples filas y además correlativas, es decir, se reduce la muestra a más de la mitad.

Concluiremos esta sección de eliminación de filas comentando brevemente los efectos de dicha eliminación sobre otros estadísticos también representativos del análisis de regresión.

Efectos de la eliminación de las observaciones 150 a 300 sobre otros estadísticos		
Estadísticos	Con todas las observaciones	Sin las observaciones 150 a 300
$\bar{R}^2$	0.9792	0.9998
$\hat{\sigma}$	11.05802	0.969497
SSR	36072.99	135.3491
AIC	4.82285	-0.028975
S(p,q)	4.88458	0.071828
LBQ(12)	17.479	9.9398
LBQ(24)	27.941	20.754
LBQ(36)	43.290	29.250

A partir de los resultados presentados en la tabla anterior, se puede decir que la eliminación de las observaciones 150 a 300 produce cambios importantes en todos los estadísticos, destacando el incremento en el coeficiente de determinación, consiguiéndose un ajuste prácticamente perfecto, la disminución en el error estándar de la regresión así como en la suma residual y la disminución en los estadísticos de Ljung-Box, concluyéndose que los residuos son ruido blanco.

6.3.- TRATAMIENTO DE LOS VALORES ATÍPICOS

Los importantes efectos que los valores atípicos tienen tanto sobre las estimaciones de los coeficientes como en el resto de los estadísticos en el modelo de regresión hacen imprescindible su tratamiento. En esta sección pues se comentarán los resultados obtenidos en la corrección de los valores atípicos incluidos en este trabajo.

Una vez contrastada la presencia de los valores atípicos se procederá a su tratamiento con el objeto de observar cómo cambiarán los resultados de la estimación y de los contrastes cuando se considera la presencia de un atípico.

Aunque los métodos de detección “fracasan” en indicar la presencia del atípico tipo escalón, los resultados de ciertos contrastes si nos indicaban que en la observación 150 ocurría

algo que nos distorsionaba la detección de los atípicos. Por este motivo y conociendo la presencia de los mismos, también consideraremos su tratamiento.

Debido a la diferencia que se plantea a la hora del tratamiento de los valores atípicos en la variable endógena y en las variables exógenas, se expondrán primeramente los resultados para el tratamiento de éstos en la endógena y posteriormente se comentará la corrección de los atípicos en la variable exógena.

6.3.1.- TRATAMIENTO DE ATÍPICOS EN LA VARIABLE ENDÓGENA

Tal y como se comentó en la sección dedicada a la exposición de la corrección de atípicos en el modelo lineal general, existen diversas formas de tratar este problema. La primera sería mediante métodos de estimación robustos, que no eliminan la observación en cuestión sino que buscan minimizar su efecto mediante estimaciones robustas de los parámetros. Las dificultades de cálculo, también comentadas anteriormente, nos conducen a la estimación del modelo de regresión corrigiendo los valores atípicos mediante lo que denominamos el “método clásico”, es decir, mediante la introducción de variables ficticias para captar ese comportamiento anómalo. Pasamos a continuación a exponer los principales resultados.

6.3.1.1.- TRATAMIENTO DEL VALOR ATÍPICO TIPO IMPULSO EN LA VARIABLE ENDÓGENA

Con el objeto de captar el comportamiento anómalo en la observación 100 detectado por los diferentes contrastes efectuados en la etapa de detección de los valores atípicos, se construyó la variable ficticia I1, definida de la siguiente forma:

$$I1 = \begin{cases} 1 & t = 100 \\ 0 & t \neq 100 \end{cases}$$

variable que se incluyó posteriormente en el modelo de regresión, proporcionando los siguientes resultados:

Estimaciones de los modelos con tratamiento de valores atípicos en la variable endógena			
Parámetros y estadísticos	Modelo 1	Modelo 6	Modelo 7
$\hat{\beta}_0$ (ES)	45255 (0.0559)	45255 (0.0561)	453089 (0.0799)
$\hat{\beta}_1$ (ES)	29135 (0.0578)	29135 (0.0579)	29182 (0.0580)
$\hat{\beta}_2$ (ES)	-10791 (0.0549)	-10791 (0.059)	-107256 (0.0549)
$\hat{\beta}_3$ (ES)	99530 (0.0617)	99532 (0.0546)	99502 (0.0618)
$\hat{\beta}_4$ (ES)	742468 (0.0544)	742469 (0.0546)	742501 (0.0549)
$\hat{\delta}_1$ (ES)		1000260 (0.9724)	
$\hat{\delta}_2$ (ES)			998937 (0.1121)
$\bar{R}^2$	0.9998	0.9998	0.9999
$\hat{\sigma}$	0.9632	0.9648	0.9634
LBQ(12)	9.3540	9.3624	9.5706
LBQ(24)	17.937	17.946	18.148
LBQ(36)	23.998	24.002	24.560
Errores estándar entre paréntesis			
Modelo 6: corrección atípico tipo impulso; $\hat{\delta}_1$ parámetro asociado a la ficticia I1.			
Modelo 7: corrección atípico tipo escalón; $\hat{\delta}_2$ parámetro asociado a la ficticia I2			

Tal y como se puede apreciar, el parámetro asociado a esta variable arroja una estimación de 100.026, valor que prácticamente coincide con la magnitud del efecto incluido al contaminar la variable endógena con el atípico tipo impulso.

Destacar también que los sesgos encontrados en la estimación del modelo sin el tratamiento de atípicos con respecto al modelo inicial desaparecen, obteniéndose estimaciones de los coeficientes prácticamente iguales a los simulados en el modelo inicial, siendo muy semejantes los errores estándar.

El resto de los estadísticos típicos en el análisis de regresión también vuelven a ser prácticamente iguales a los obtenidos en la estimación del modelo inicial, de manera que la

corrección del atípico tipo impulso mediante la introducción de la variable ficticia genera unos resultados muy satisfactorios y pone de manifiesto la importancia del tratamiento del atípico para evitar sus efectos.

6.3.1.2.- TRATAMIENTO DEL VALOR ATÍPICO TIPO ESCALÓN EN LA VARIABLE ENDÓGENA

En este caso, se trató ó corrigió el atípico tipo escalón mediante una variable ficticia tipo escalón (I2), definida mediante la expresión:

$$I2 = \begin{cases} 0 & 1 \leq t \leq 149 \\ 1 & 150 \leq t \leq 300 \end{cases}$$

Introduciéndola en la regresión se obtienen los resultados del modelo 7, recogidos en la tabla anterior.

A partir de estos resultados podemos decir que de nuevo la variable ficticia recoge perfectamente el escalón que se introdujo en la variable endógena, siendo la estimación del parámetro asociado a dicha variable de 99.8937, recogiendo casi perfectamente la magnitud del atípico incluido en la observación 150.

El resto de los parámetros estimados son también prácticamente idénticos a los del modelo inicial simulado, al igual que el resto de los estadísticos analizados. Se pone de nuevo de manifiesto que la inclusión de la variable ficticia para captar el atípico tipo escalón en la variable endógena es un buen método para evitar los “nefastos” efectos de los atípicos en las estimaciones de los modelos de regresión.

6.3.2.- TRATAMIENTO DE ATÍPICOS EN LA VARIABLE EXÓGENA

La inclusión de un atípico sobre alguna de las variables explicativas en el modelo de regresión plantea problemas y métodos de detección diferentes a los tratados en el caso anterior. De hecho, los métodos de detección de este tipo de atípicos, como ya hemos podido comprobar, no se muestran tan eficaces si el atípico se encuentra en las variables exógenas. Por este motivo, se deben analizar muchas cosas y, en especial, si los efectos de los

valores atípicos sobre las variables exógenas son tan importantes como para requerir un tratamiento especial.

Aunque, como hemos visto, los efectos de los atípicos, tipo impulso y tipo escalón sobre la variable X_2 , son importantes pero de menor intensidad que los correspondientes a la variable endógena, se proponen en este caso diferentes formas de considerarlos, formas que pasamos a detallar.

6.3.2.1.- TRATAMIENTO DEL VALOR ATÍPICO TIPO IMPULSO EN LA VARIABLE EXÓGENA X_2 .

El tratamiento del valor atípico tipo impulso en X_2 se ha realizado siguiendo diferentes esquemas o formas de corrección que fueron los siguientes:

Solución 1: Primero se realizó una regresión de X_{2I1} sobre una constante y la variable ficticia I_1 . Los residuos de esta regresión (RX_{2I1}), que recogen la variable X_2 libre de atípicos, se introducen en el modelo completo, habiéndose eliminado el efecto del atípico en X_2 .

Solución 2: Se realizó una regresión de X_{2I1} sobre la variable ficticia I_1 , obteniéndose los residuos denominados $R2X_{2I1}$, residuos que se introducen de nuevo en el modelo completo.

Solución 3: Sustituir la observación 100 de X_{2I1} (que es la que contiene el atípico) por la media de los valores inmediatamente anterior y posterior, obteniéndose la variable X_{2I1MED} . Esta nueva variable se introduce en el modelo completo.

Los resultados que se obtienen en estas estimaciones se presentan en la siguiente tabla en la que también se recogen las estimaciones del modelo inicial, con el cual se compararán los resultados obtenidos en estas correcciones del valor atípico en X_{2I1} .

Tratamiento del valor atípico tipo impulso en X2I1				
Parámetros y estadísticos	Modelo 1	Modelo solución 1	Modelo solución 2	Modelo solución 3
$\hat{\beta}_0$ (ES)	45.255 (0.0559)	46.3572 (0.05573)	45.2616 (0.0563)	45.2222 (0.0663)
$\hat{\beta}_1$ (ES)	29135 (0.0578)	29147 (0.05789)	29167 (0.0582)	29001 (0.0685)
$\hat{\beta}_2$ (ES)	-10.791 (0.0549)	-10.7909 (0.05499)	-10.7902 (0.0553)	-10.7565 (0.0650)
$\hat{\beta}_3$ (ES)	9.9530 (0.0617)	9.9489 (0.0617)	9.9426 (0.0621)	10.0171 (0.0731)
$\hat{\beta}_4$ (ES)	74.2468 (0.0544)	74.2445 (0.0545)	74.2408 (0.0547)	74.2787 (0.0645)
$\bar{R}^2$	0.9998	0.9998	0.9998	0.9998
$\hat{\sigma}$	0.9632	0.964132	0.969	1.14122
LBQ(12)	9.3540	9.1308	8.8158	10.778
LBQ(24)	17.937	17.680	17.279	17.904
LBQ(36)	23.998	23.910	23.916	27.153

A partir de los resultados presentados en la tabla anterior, se pone de manifiesto que todas las soluciones propuestas eliminan casi en su totalidad los efectos del atípico tipo impulso que se pretende corregir. Lo que si parece claro es que existen ciertos matices que nos pueden llevar a decidir cuál de los métodos de corrección propuestos generan resultados más satisfactorios. Pasaremos pues a comentar los resultados.

Las soluciones 1 y 2 generan resultados muy similares, siendo además muy semejantes a las obtenidas en el modelo inicial: los parámetros son prácticamente iguales, así como el resto de los estadísticos.

Los resultados obtenidos con la solución 3 difieren algo de los del modelo inicial: los parámetros $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$ son inferiores, en valor absoluto, a los obtenidos con el modelo inicial mientras que $\hat{\beta}_3$ y $\hat{\beta}_4$ son superiores. Los errores estándar, en casi todos los casos, son más elevados como consecuencia del incremento experimentado en $\hat{\sigma}$. Esto pone de manifiesto que se estima con menor precisión que en el resto de los casos. Aún así, se puede decir que la solución 3 tampoco es una mala solución.

Como conclusión al tratamiento del valor atípico tipo impulso en la variable explicativa se puede decir que las tres soluciones propuestas generan resultados bastante satisfactorios, pudiendo optarse por cualquiera de ellas en caso de que se presente esta situación.

6.3.2.2.- TRATAMIENTO DEL VALOR ATÍPICO TIPO ESCALÓN EN LA VARIABLE EXÓGENA X2.

El tratamiento del atípico tipo escalón en X2 se realizó utilizando dos posibles soluciones. Por un lado, se corrige el atípico tipo escalón en X2 mediante la realización de una regresión de la variable explicativa sobre una constante y la variable ficticia I2. Los residuos resultantes de esta regresión se introducen en la regresión inicial. Este enfoque es lo que denominamos solución 4. Por otro lado, se realiza una regresión de X2 sobre la variable ficticia tipo escalón para corregir el atípico y a continuación se incluyen los residuos de esta regresión en el modelo inicial. Este caso se denomina solución 5. Estos dos enfoques se comentan a continuación, presentándose los resultados de las dos estimaciones así como la del modelo inicial, para realizar comparaciones, en la siguiente tabla.

Tratamiento del valor atípico tipo escalón en X2I2			
Parámetros y estadísticos	Modelo 1	Modelo solución 4	Modelo solución 5
$\hat{\beta}_0$ (ES)	45.255 (0.0559)	46.3590 (0.0582)	45.9896 (0.0679)
$\hat{\beta}_1$ (ES)	29135 (0.0578)	29456 (0.0604)	29829 (0.0704)
$\hat{\beta}_2$ (ES)	-10.791 (0.0549)	-10.792 (0.0574)	-10.7687 (0.0669)
$\hat{\beta}_3$ (ES)	9.9530 (0.0617)	9.9348 (0.0645)	9.9172 (0.0752)
$\hat{\beta}_4$ (ES)	74.2468 (0.0544)	74.268 (0.0569)	74.2887 (0.0663)
$\bar{R}^2$	0.9998	0.9998	0.9998
$\hat{\sigma}$	0.9632	1.006683	1.173465
LBQ(12)	9.3540	26.882	342.63
LBQ(24)	17.937	52.442	604.19
LBQ(36)	23.998	63.597	772.41

Hay que destacar que, en este caso, no se podía optar por la sustitución del valor atípico por la media de los valores inmediatamente anterior y posterior debido a la naturaleza del atípico considerado.

Con respecto a los resultados presentados en la tabla anterior, se puede decir que, en ambas soluciones, no se produce una buena recuperación de las estimaciones de los parámetros del modelo inicial, tal y como ocurría en las soluciones presentadas para el atípico tipo impulso en X2I1. Los sesgos en las estimaciones son más importantes que en el caso anterior añadiéndose además el hecho de las estimaciones son más imprecisas debido a que los errores estándar de las estimaciones son mayores. Esto se debe al incremento que se ha producido en $\hat{\sigma}$ con respecto al modelo inicial en ambas soluciones. Sin embargo, se puede decir que el error estándar de la regresión de la solución 5 es mayor que el obtenido para la solución 4 de manera que las estimaciones en este último caso son menos imprecisas.

Frente a estos resultados, el coeficiente de determinación en ambos casos nos indica que obtenemos un ajuste casi perfecto aunque los residuos no presentan esquema de ruido blanco. De hecho, los estadísticos de Ljung-Box experimentan un incremento significativo al compararlos con los obtenidos en el modelo inicial, destacando los resultados en la solución 5 cuyos residuos muestran un esquema típico de no estacionariedad, reflejándose en los valores que se obtienen en los estadísticos de Ljung-Box.

Como principal conclusión a estas dos soluciones propuestas para el atípico escalón en X2I2 cabe decirse que la primera de ellas (corrección del atípico escalón en X2 mediante una regresión de X2I2 sobre una constante y la variable ficticia I2) parece ser la menos mala, en el sentido de que, aunque no genera los resultados del modelo inicial, si se obtienen mejores resultados que con la estimación de la solución 5.

6.4.- CONCLUSIONES DE LOS EJEMPLOS

Los ejemplos que se han introducido en este trabajo buscaban ilustrar los problemas que presentan los valores atípicos, ya sean en la variable endógena como en las exógenas, en el modelo de regresión. Para ello, se han incluido los dos tipos de atípicos considerados en la variable endógena y en una de las variables exógenas de un modelo simulado.

Los efectos de los atípicos en los dos tipos de variables son diferentes. Mientras que los atípicos en la variable endógena afectan de forma significativa a todas las estimaciones de los parámetros y al resto de los estadísticos de la regresión, los atípicos en la variable exógena presentan su mayor efecto en el parámetro asociado a esta variable, que en el caso del atípico tipo escalón pierde la significatividad estadística.

Por lo que respecta a los métodos de detección de los valores atípicos se puede decir, a partir de lo expuesto en las páginas anteriores, que los atípicos incluidos en la variable endógena son más fáciles de detectar, concluyendo casi todos los contrastes en la misma dirección. Sin embargo, los atípicos introducidos en la variable exógena son algo más difíciles de detectar debido a que la mayor parte de los contrastes utilizados se basan en los residuos de la regresión mínimo cuadrática y no siempre se traslada de forma clara el atípico a los residuos.

El tratamiento de los valores atípicos se ha realizado siguiendo diferentes enfoques según donde se encontrase el valor atípico. Por lo que respecta a los atípicos en la variable endógena, las soluciones propuestas, es decir la inclusión de una variable ficticia para captar el comportamiento anómalo, genera resultados satisfactorios, obteniéndose tanto unas estimaciones como unos estadísticos muy similares a los del modelo inicial simulado.

Sin embargo, las soluciones propuestas para el tratamiento de los valores atípicos en la variable exógena no generan los mismos resultados. Mientras que las soluciones propuestas para el atípico tipo impulso en X_2 generan unos resultados bastante satisfactorios, recuperándose en todos los casos los parámetros del modelo inicial simulado, no ocurre lo mismo con las soluciones que se proponen para el atípico tipo escalón en la variable exógena. En este caso, las estimaciones de los parámetros presentan sesgos que son más importantes en la solución 5, los errores estándar son superiores, indicando la menor precisión de las estimaciones, la varianza de la regresión se incrementa de forma significativa y los estadísticos de Ljung-Box no nos permiten aceptar la hipótesis de que los residuos son ruido blanco.

A partir de todo lo expuesto en esta sección, se puede decir que si los atípicos se encuentran en la variable endógena, aunque los efectos son más importantes, el tratamiento

de éstos genera unos resultados bastante aceptables. Sin embargo, la presencia de un atípico en alguna de las variables exógenas tiene unos efectos menos perniciosos que en el caso anterior aunque su detección es bastante más complicada y su tratamiento no genera resultados demasiado satisfactorios. En este caso, si sería recomendable más que nunca recurrir a los métodos de estimación robusta presentados en secciones anteriores.

[illegible][illegible]

YES

7.- CONCLUSIONES Y EXTENSIONES

CONCLUSIONES Y EXT

7.- CONCLUSIONES Y EXTENSIONES

A lo largo de este trabajo, se han tratado de recoger, en parte, los últimos desarrollos que se han producido en el campo econométrico sobre el tema de los valores atípicos. Las investigaciones realizadas en este campo han sido muy abundantes por lo que en este trabajo se han recogido únicamente las más novedosas, es decir las investigaciones sobre los valores atípicos más recientes y, por otro lado, las investigaciones que más atención han recibido por parte de los investigadores. Con ello, quiero dejar claro que existen otras tendencias dentro del análisis de los valores atípicos que por distintos motivos no han sido recogidas en esta investigación y que podrían resultar enfoques alternativos, igualmente válidos.

Entrando ya en los principales resultados de esta investigación, debe citarse como hecho fundamental el haber recogido las principales consecuencias, métodos de detección y métodos de corrección de los valores atípicos más ampliamente utilizados en el análisis econométrico.

Si nos centramos en alguno de los puntos desarrollados, se han examinado las consecuencias de la falta de tratamiento de los valores atípicos en el modelo de regresión utilizando la aproximación de las consecuencias de los errores de especificación, habiéndose demostrado que los valores atípicos no tratados provocan estimaciones de los parámetros sesgadas y varianzas residuales superiores a las que se obtendrían si se hubieran tratado los valores atípicos. Estos dos hechos tienen consecuencias sobre estos estadísticos de manera que los resultados de la estimación de los mínimos cuadrados pueden ser totalmente erróneos. Asimismo, y considerando los errores cuadráticos medios de los estimadores con el tratamiento de los valores atípicos y sin él, se deduce que los ECM correspondientes a los parámetros del modelo con errores de especificación, en general, serán superiores a los del modelo correctamente especificado, de manera que los estimadores de este último modelo serán los adecuados. En este sentido, se pone de manifiesto que los valores atípicos pueden tener consecuencias bastante nefastas en la estimación de los parámetros del modelo de regresión.

La detección de los valores atípicos ha recibido también una atención especial a lo largo de este trabajo. Se han expuesto los métodos más utilizados en el análisis del modelo de regresión. Entre los métodos expuestos se han presentado algunos que se basan en la estimación robusta de los parámetros que ahora están recibiendo una especial atención debido al desarrollo que se está produciendo en el campo electrónico siendo los ordenadores actuales más potentes, lo que facilita notablemente la obtención de las estimaciones robustas en las que se basan los métodos de detección de atípicos. También se ha considerado la detección de los valores atípicos en lo que podríamos denominar enfoque "clásico", es decir mediante el análisis de los residuos y mediante la eliminación de las filas correspondientes a la observación sospechosa.

En la sección dedicada a la corrección de los valores atípicos, se han expuesto las tendencias desarrolladas. En el caso del modelo de regresión, se han expuesto los principales estimadores robustos desarrollados presentando casi todos un BP elevado pero escasa eficiencia. Entre estos estimadores destaca el estimador MM derivado por Yohai (1988) que presenta un BP elevado y además tiene una elevada eficiencia, de manera que puede considerarse como uno de los mejores métodos de estimación robusta. Además, se ha expuesto brevemente el método de corrección que denominamos "clásico" consistente en el tratamiento de los valores atípicos mediante la introducción de variables ficticias. Si bien este método ha recibido numerosas críticas, pensamos que puede ayudar a la mejor comprensión de las variables estudiadas debido a que, según pensamos en esta investigación, todo valor atípico en las series económicas debe tener una causa. Por este motivo, creemos que una correcta detección del valor atípico, basada en los métodos de estimación robusta, junto con una corrección adecuada de la observación anómala, acompañados por la investigación de la causa ya sea económica o de otro tipo que provoca la presencia de esta observación anómala, sería el método adecuado para corregir ó por lo menos intentar eliminar los problemas que causa este valor atípico.

En la sección 6 se ha presentado un ejemplo con el objeto de ilustrar la importancia de los valores atípicos. Se ha considerado un modelo simulado en el que se han introducido los dos atípicos que hemos tratado en este trabajo, tanto en la variable endógena como en las

variables exógenas. Para cada uno de los casos, se han analizado los efectos de los valores atípicos sobre las estimaciones y los estadísticos, concluyéndose que las consecuencias de los valores atípicos son más importantes cuando éstos se encuentran sobre la variable endógena. Los métodos de detección de los valores atípicos empleados en estos ejemplos nos llevaron a concluir que la detección del atípico tipo impulso es muy clara, mientras que en el caso del tipo escalón los métodos de detección utilizados no son capaces, en la mayor parte de los casos, de indicar la presencia del atípico tipo escalón. Por lo que respecta a la detección de los atípicos en las variables explicativas, ésta se presenta bastante insatisfactoria con los contrastes utilizados, contrastes no robustos. En estos casos, los métodos de estimación robustos serían más adecuados.

El tratamiento de los valores atípicos en los ejemplos, en casi todos los casos, se ha realizado utilizando el método "clásico" (variables ficticias) dando resultados bastante satisfactorios. Se pone así de manifiesto que es un buen método para el tratamiento de los valores atípicos siempre conociendo la causa de su aparición.

Tal y como se ha puesto de manifiesto en párrafos anteriores, este trabajo tiene claras extensiones a las que nos dedicaremos en investigaciones futuras.

En todo este artículo, se ha hablado del tema de los valores atípicos sin hacer referencia en ningún momento a la estimación bayesiana. Es lógico pensar que, una vez conocido el período de ocurrencia y el tipo de atípico que se presenta, se podría incorporar al proceso de estimación como información a priori, de manera que se obtuviesen estimadores más eficientes. En este sentido, la estimación bayesiana nos permitirá incorporar esta información por medio de funciones de probabilidad a priori generando mejores resultados en el proceso de estimación. Por ello, una línea de investigación futura será el tratamiento de los atípicos desde el punto de vista bayesiano.

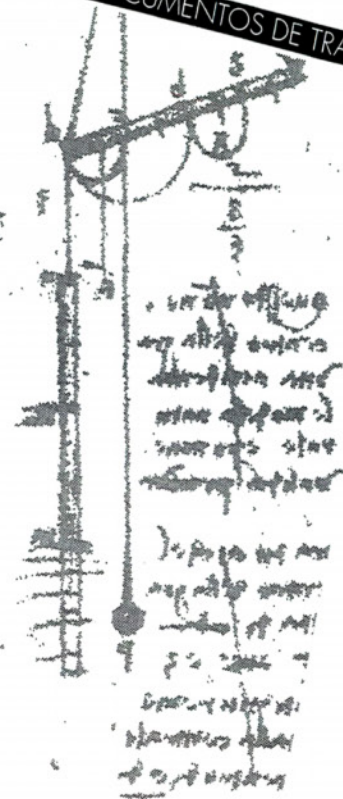
Otra extensión posible a este trabajo se enmarca dentro de los métodos no paramétricos que están siendo ahora muy utilizados. Birkes y Dodge (1993) ponen de manifiesto que pueden ser de gran utilidad en el tratamiento de los valores atípicos.

Los tres ejes de la
 estructura son: el eje de
 la columna, el eje de la
 viga y el eje de la
 losa. El eje de la columna
 es el eje de la estructura
 que soporta la carga. El
 eje de la viga es el eje de
 la estructura que soporta
 la carga. El eje de la losa
 es el eje de la estructura
 que soporta la carga.

La estructura es una
 estructura de hormigón
 armado. La estructura es
 una estructura de hormigón
 armado. La estructura es
 una estructura de hormigón
 armado. La estructura es
 una estructura de hormigón
 armado. La estructura es
 una estructura de hormigón
 armado.

3 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86 88 90 92 94 96 98 100

La estructura es una
 estructura de hormigón
 armado. La estructura es
 una estructura de hormigón
 armado. La estructura es
 una estructura de hormigón
 armado. La estructura es
 una estructura de hormigón
 armado. La estructura es
 una estructura de hormigón
 armado. La estructura es
 una estructura de hormigón
 armado. La estructura es
 una estructura de hormigón
 armado.



La estructura es una
 estructura de hormigón
 armado. La estructura es
 una estructura de hormigón
 armado. La estructura es
 una estructura de hormigón
 armado. La estructura es
 una estructura de hormigón
 armado.

BIBLIOGRAFIA

BIBLIOGRAFÍA

- Atkinson, A.C. (1981): "Two Graphical Displays for Outlying and Influential Observations in Regression". *Biometrika*, 68, pp. 13-20.
- Atkinson, A.C. (1982a): "Regression Diagnostics, Transformations and Constructed Variables (with Discussion)". *Journal of the Royal Statistical Society, Series B*. nº 44, pp. 1-36.
- Atkinson, A.C. (1982b): "Robust and Diagnostic Regression Analysis". *Commun. Statist. Theo. Meth.*, 11, pp. 2559-2571.
- Atkinson, A.C. (1985): *Plots, Transformations and Regression*. New York. Oxford University Press.
- Atkinson, A.C. y Weisberg, S. (1991): "Simulated Annealing for the Detection of Multiple Outliers Using Least Squares and Least Median of Squares Fitting". en *Directions in Robust Statistics and Diagnostics. Part II*. eds. Stahel, W. y Weisberg, S. Springer-Verlag.
- Aznar, A. y Tívez, F.L: (1993a): *Métodos de Predicción en Economía (I)*. Ed Ariel. Barcelona.
- Barnett, V. y Lewis, T. (1984): *Outliers in Statistical Data*. (2nd Ed). John Wiley and Sons. New York.
- Bartlett, M.S. (1946): "On the Theoretical Specification of Sampling Properties of Autocorrelated Time Series". *Journal of the Royal Statistical Society, series B*, pp. 8-27.
- Belsley, D.A., Kuh, E. y Welsch, R.E. (1980): *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley and Sons. New York.
- Birkes, D. y Dodge, Y. (1993): *Alternatives Methods on Regression*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons. New York.
- Chan, W.S. (1995): "Understanding the Effect of Time Series Outlier on Sample Autocorrelation". Próxima publicación en Test.
- Cook, R.D. (1977): "Detection of Influential Observation in Linear Regression". *Technometrics*, 19, pp. 15-18.

- Cook, R.D. y Weisberg, S. (1982): *Residuals and Influence in Regression*, Chapman and Hall, Londres.
- Cook, R.D., Hawkins, H y Weisberg, S. (1992): "Comparison of Model Misspecification Diagnostics using Residuals from Least Mean of Squares and Least Median of Squares Fits". *Journal of the American Statistical Association*, 87, pp. 419-424.
- Daniel, C. (1959): "Use of Half-Normal Plots in Interpreting Factorial Experiments". *Technometrics*, 20, pp. 149-156.
- Davies, L. y Gather, U. (1993): "The Identification of Multiple Outliers". *Journal of the American Statistical Association*, vol. 88, n°. 423. pp. 782-792.
- de Juan, A. (1990): Eficiencia del Mercado Español de Divisas: Análisis Mensual y Diario de los Tipos de Cambio. *Tesis Doctoral*. Diciembre 1990.
- de Juan, A. (1993): "Daily Exchange Rates: Non-Linearities or Misspecification Errors?". *Documento de Trabajo DT 93#12*. Serie Econometría. Departamento de Análisis Económico. Universidad Autónoma de Madrid.
- de Juan, A. (1995): "Importancia de los Valores Atípicos en el Modelo de Regresión: Causas, Consecuencias, Detección y Tratamiento". *Documento de Trabajo DT 95#1*. Serie Econometría. Departamento de Análisis Económico. Universidad Autónoma de Madrid.
- Donoho, D.L. y Huber, P.J. (1983): "The notion of Breakdown Point" en *A Festschrift for Erich Lehmann*. ed. P. Bikel, Doksum, K y Hodges, J.L. Jr; Wadsworth. Statistics/Probability Series.
- Fuller, W.A. (1976): *Introduction to Statistical Time Series*. New York. Wiley and Sons.
- Gentleman, J.F. y Wilk, M.B. (1975a): "Detecting Outliers: II. Supplementing the Direct Analysis of Residuals". *Biometrics*, 31, pp. 387-410.
- Gentleman, J.F. y Wilk, M.B. (1975b): "Detecting Outliers in a Two Way Table. I. Statistical Behaviour of Residuals". *Technometrics*, 17, pp. 1-14.
- Gnanadesikan, R. y Kettenring, J.R. (1972): "Robust Estimates, Residuals and Outlier Detection with Multiresponse Data". *Biometrics*, 28, pp. 81-124.

- Hadi, A.S. (1992): "Identifying Multiple Outliers in Multivariate Data". *Journal of the Royal Statistical Society, Series B*, 54, pp.761-771.
- Hadi, A.S. (1994): "A Modification of a Method for the Detection of Outliers in Multivariate Samples". *Journal of the Royal Statistical Society, Series B*, 56.
- Hampel, F.R. (1974): "The Influence Curve and its Role in Robust Estimation". *Journal of the American Statistical Association*, 69, pp. 383-393.
- Hampel, F.R., Ronchetti, E. M. Rousseeuw, P.J. and Stahel, W.A. (1986): *Robust_Statistics: The Approach Based on Influence Functions*. Wiley and Sons. New York.
- Hettmansperger, T.P. (1984): *Statistical Influence Based on Ranks*. Wiley and Sons. New York.
- Hill, R.W. (1977): Robust Regression When There are Outliers in the Carriers. *Unpublished Ph. D. Dissertation*, Harvard University, Boston, MA:
- Huber, P.J. (1973): "Robust Regression: Asymptotics, Conjectures and Monte Carlo". *Annals of Statistics*, 1. pp. 799-821.
- Justel, A., Peña, D. y Sánchez, M^a. J. (1993): "Grupos de Atípicos en Modelos Económicos". *Cuadernos Económicos de ICE*, n° 55, 1993/3. pp. 285-325.
- Mallows, C.L. (1975): "On Some Topics in Robustness". *Unpublished Memorandum*, Bell Telephone Laboratories, Murray Hill, New York.
- McKean, J.W. y Sheather, S.J. (1991): "Small Sample Properties of Robust Analyses of Linear Models Based on R-Estimates: a survey" en *Directions in Robust_Statistics and Diagnostics, Part II*. eds Stahel, W and Weisberg, S. Springer Verlag, New York.
- McKean, J.W., Sheather, S.J. y Hettmansperger, TP. (1983): "The Use and Interpretation of Residuals Based on Robust Estimation". *Journal of the American_Statistical Association*, vol. 88, n° 424, pp. 1254-1263.
- Montgomery, D.C. y Peck, A.E. (1982): *Introduction to Linear Regression Analysis*. John Wiley and Sons. New York.
- Mosteller, F. y Tukey, J.W. (1977): *Data Analysis and Regression*. Addison-Wesley, Reading, MA.

- Netkov, N.M. y Neytchev, P.N. (1991): "Unmasking Multivariate Outliers and Leverage Points by Means of BMDP3R". en *Directions in Robust Statistics and Diagnostics. Part II.* eds. Stahel, W. y Weisberg, S. Springer-Verlag.
- Portnoy, S. (1991): "Regression Quantile Diagnostics for Multiple Outliers". en *Directions in Robust Statistics and Diagnostics. Part II.* eds. Stahel, W. y Weisberg, S. Springer Verlag.
- Rousseeuw, P.J. (1983): "Multivariate Estimation with High Breakdown Point" en *Mathematical Statistics and Applications*, vol. B. Eds. Grossman, W., Pflug, G. Vincze, I. y Wertz, W. pp. 283-297.
- Rousseeuw, P.J. (1984): "Least Median of Squares Regression". *Journal of the American Statistical Association*, 79, n° 388. pp. 871-880.
- Rousseeuw, P.J. (1985): "A Regression Diagnostic for Multiple Outliers and Leverage Points". *Abstract in IMS Bulletin.*, 14, pp.399.
- Rousseeuw, P.J. and Croux, C. (1993): "Alternatives to the Median Absolute Deviation". *Journal of the American Statistical Association*, vol. 88, n°424, pp. 1273-1283.
- Rousseeuw, P.J. y Basset, G.W. Jr. (1990): "The remedian: A Robust Averaging Method for Large Data Sets". *Journal of the American Statistical Association*, vol. 85, n° 409, pp. 97-104.
- Rousseeuw, P.J. y Leroy, A.M. (1987): *Robust Regression and Outlier Detection*. John Wiley and Sons. New York.
- Rousseeuw, P.J. y van Zomeren, B.C. (1990): "Unmasking Multivariate Outliers and Leverage Points". *Journal of the American Statistical Association*, 85, n°411, pp. 633-639.
- Rousseeuw, P.J. y Yohai, V. (1984): "Robust Regression by means of S-Estimators" en *Robust and NonLinear Time Series Analysis*. eds. Franke, W., Hardle, W. and Martin, R.D., Lecture Notes in Statistics, n° 26, Springer Verlag, New York. pp. 256-272.
- Siegel, A.F. (1982): "Robust Regression using Repeated Medians" *Biometrika*, 69, pp. 242-244.

- Simonoff, J.S. (1991): "General Approaches to Stepwise Identification of Unusual Values in Data Analysis". en *Directions in Robust Statistics and Diagnostics. Part II.* eds. Stahel, W. y Weisberg, S. Springer-Verlag.
- Staudte, R.G. y Sheather, S.J. (1990): *Robust Estimation and Testing*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons. New York
- Wilson, H.C. (1978): "Least Squares versus Minimum Absolute Deviations Estimation in Linear Models". *Decision Science*, pp. 322-335.
- Yohai, V.J. (1988): "High-breakdown Point and High Efficiency Robust Estimates for Regression". *Annals of Statistics*, 15. pp. 642-665.
- Yohai, V.J., Stahel, W.A. and Zamar, R.H. (1991): "A Procedure for Robust Estimation and Inference in Linear Regression". en *Directions in Robust Statistics_and Diagnostics. Part II.* eds. Stahel, W.A. and Weisberg, W. Springer Verlag, New York.
- Yohai, V.J. and Zamar, R.H. (1988): "High Breakdown Estimates of Regression by Means of the Minimization of an Efficient Scale". *Journal of the American Statistical Assotiation*, 83, pp. 406-413.

