Documento de Trabajo

8 9 2 8

ESTIMACION DE MODELOS ECONOMETRICOS -OBSERVACIONES ATIPICAS EN LA MUESTRA-

José Hernández Alonso

FACULTAD DE CIENCIAS ECONOMICAS Y EMPRESARIALES.- UNIVERSIDAD COMPLUTENSE
Campus de Somosaguas. 28023 - MADRID

ESTIMACION DE MODELOS ECONOMETRICOS
-OBSERVACIONES ATIPICAS EN LA MUESTRA-

JOSE HERNANDEZ ALONSO

RESUMEN

Este documento describe las consecuencias que sobre la estimación mínimo-cuadratica de los modelos econométricos tiene la aparición de observaciones atípicas en los datos muestrales. Se señala la importancia de analizar la posible presencia de puntos atípicos,destacandose las medidas y contrastes estadísticos que facilitan determinar cuando un dato muestral no es homogeneo con el resto, todo ello ilustrado con una aplicación empírica. Finalmente,se enumeran los pasos metodológicos, que conviene atender, para conseguir un diseño robusto en la elaboración de un modelo econométrico.

AUTOR

JOSE HERNANDEZ ALONSO es Doctor en Ciéncias Económicas y Empresariales por la Universidad Complutense de Madrid, Diplomado en Estadística e Investigación Operativa por la misma Universidad y Diplomado en Métodos Cuantitativos de Gestión por la Escuela de Organización Industrial de Madrid. Profesor de la Facultad de Ciéncias Económicas y Empresariales de la Universidad Complutense y colaborador en la Escuela Superior de Gestión Comercial y Marketing, también ha desarrollado su actividad docente en el Centro Asociado de Madrid a la U.N.E.D. Ha trabajado durante varios años en temas relacionados con planificación económica y educativa.

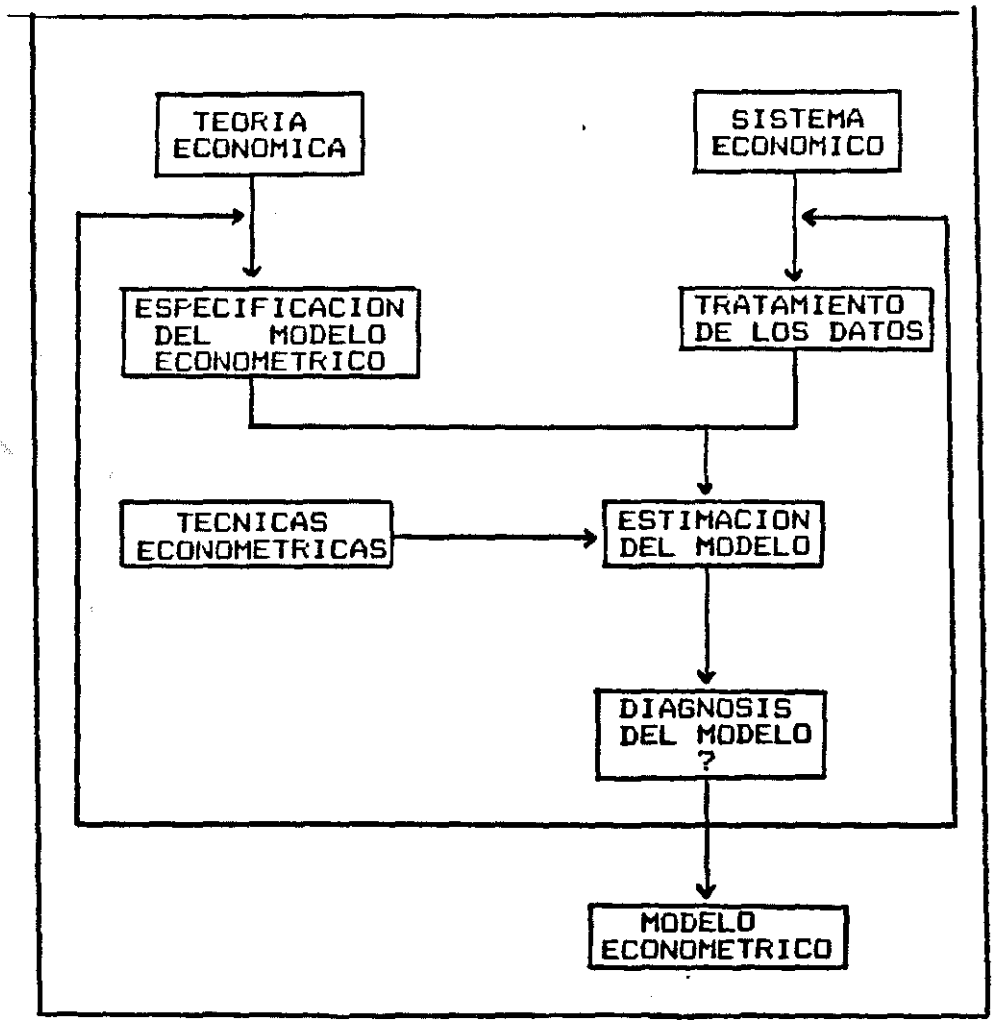
En la elaboración de todo modelo econométrico se presupone, que su especificación es correcta y que una estructura específica ha generado las observaciones en estudio, consistiendo el objetivo principal del trabajo econométrico en la estimación de los valores de los parámetros contenidos en su ecuación, en base a los datos muestrales de las variables endógenas y exógenas que la definen.

Pero en las investigaciones empíricas casi nunca puede partirse de un marco teórico, suficientemente exacto, como para formular el modelo con precisión y garantía de que la especificación es formalmente correcta y que todos los requerimientos precisos para una inferencia válida de los valores de los parametros se cumplen, al menos, a un nivel aproximativo suficiente.

En este sentido, conviene comprobar que la decisión adoptada en cada caso (estimación de los parámetros), no esta basada en un proceso en el que no son confirmadas las condiciones básicas, pues ello nos conduciría a una solución errónea.

La necesaria diagnosis del modelo requiere analizar,en base a los propios resultados de la estimación, la validez general de las condiciones teóricas iniciales, de acuerdo con la aplicación de determinados contrastes de hipótesis usuales en todo proceso de modelización econométrica, lo que implica,según los casos, la reformulación del modelo o la reestimación del mismo por métodos alternativos que sean más satisfactorios.

Gráfico 1.- Proceso de modelización econométrica.



Estas limitaciones de la práctica econométrica se producen, en gran parte,de los habituales errores en la materia prima de los modelos construidos,las estadísticas económicas.La calidad de los datos es lo que implica que,cuando se da el salto de la teoría a la práctica,haya que convencerse de que todos los modelos y todas las hipótesis en que se fundamenta un determinado metodo de estimación son 'falsos'. 'No hay modelos correctos; hay modelos útiles', como señalan los profesores RAYMOND y URIEL (1). Los modelos deben ser utilizados no creídos.

El problema de la calidad de las estadísticas económicas, que deriva a que no puedan aceptarse, sin ninguna objeción las cifras que se publican,sin someterlas a un proceso de contraste o crítica razonable, supone también, la normal preocupación que debe representar este hecho para la Econometría Aplicada, debiendo incluirse en los contrastes para análisis de validez general del modelo, test que permitan encontrar,sí los hay,fallos en la información numérica de las variables.

Entre éstos vamos a centrarnos en aquellos que tratan de la detección de valores anómalos o atípicos, observaciones no homogéneas con el resto de datos de la muestra, 'outliers' en terminología anglosajona.

Su estudio y tratamiento, se aborda en muy pocos manuales econométricos, a pesar de que, como demostraremos en estas páginas, pueden tener una incidencia importante en la resolución de cualquier modelo.

---

(1) J.L RAYMOND / E. URIEL (1987): Investigación econométrica aplicada.

CONSECUENCIAS SOBRE ESTIMACION MINIMO-CUADRATICA.

La estimación del modelo lineal:

$$Y = X \beta + u$$

donde, Y es un vector de T*1 observaciones de la variable endógena, X es una matriz de T*K valores de las variables exógenas, de rango K, ß es un vector de K parámetros y u es un vector T*1 de perturbaciones aleatorias, con $E(u)=0$ y $E(uu')= \sigma^2 I$ y distribucion normal, conduce, aplicando el criterio de estimación de los mínimos-cuadrados, a la ecuación:

$$X'X = b\ X'Y$$

en donde, bajo el supuesto de que la matriz X'X es invertible, se obtiene como solución para el vector de estimadores:

$$b = (X'X)^{-1}\ X'Y$$

Gráficamente, el ajuste mínimo-cuadrático significa obtener la ecuación matemática $\hat{Y} = Xb$, tal que discurriendo por entre la nube de puntos que representa los T datos muestrales, resume estos dándonos una escala de transformación de los valores X en Y.

La solución obtenida es claramente dependiente de los valores contenidos en la matriz X y en el vector Y,y por lo tanto, la validez de los datos condiciona la estimación de los parámetros del modelo.

En concreto, si suponemos la presencia de puntos atípicos que pueden mostrarse por un valor anómalo de Y para un punto del vector X correspondiente o un valor atípico para X este segundo, acompañado o no de una respuesta atípica en Y, los mismos incidiran en el resultado final de la estimación según mínimos-cuadrados.

Además, condicionado por este cambio,los residuos de la regresión también se ven afectados, en mayor o menor cuantía, lo cual conduce inevitablemente a que el conjunto de medidas evaluatorias de la ecuación estimada, varianza residual, varianza de los estimadores, coeficiente de determinación, ..., también se vean alterados variando, por tanto, todo el conjunto de resultados de la regresión y la diagnosis final sobre el modelo que se estima.

En efecto,ilustrando el proceso descrito con un ejemplo numérico en base a datos muestrales de dos variables, con objeto de que podamos representar gráficamente los resultados, se ha pasado a su resolución m-c en base a los datos originales de las variables (cuadro 1.a), a esos mismos datos perturbando un un valor de la variable endógena,dato atípico en Y(cuadro 1.b) o cometiendo un error en la variable exógena X(cuadro 1.c). Podemos así, analizar los efectos que la presencia de un punto atípico,supuestamente generado por un error en la manipulación de los datos, bien de Y o de X,tiene sobre la estimación de la ecuación.

En las tres salidas de resultados obtenidas,tanto los valores de la estimación como las medidas evaluatorias, varian de una regresión a otra,cambiando el sentido de la interpretación

que de ellas debe realizarse. Si, para la primera regresion, la estimación conseguida puede considerarse aceptable en atención a la representatividad general alcanzada en la ecuación ajustada (R^2= 0.96) y a la fiabilidad conseguida en la estimación de los parámetros, en especial en el término de pendiente de la recta (CD_{β_2} = 15%), en la segunda, con dato atípico en Y y en la tercera,con dato atípico en X, estos resultados claramente empeoran.

Cuando la regresión la realizamos en base a los valores del cuadro 1.b,se debe rechazar la estimación pues el contraste de la hipótesis $H_o:\beta_2 = 0$ acepta esta, al nivel de significacion del 5%, indicándonos la no dependencia lineal entre variables endógena y exógena. Mientras que si analizamos resultados del cuadro 1.c, esta dependencia se acepta pero,frente a la primera regresión, tiene la desventaja, de que no nos permite asegurar que se haya cuantificado adecuadamente los parámetros de la estructura que relaciona Y con X (CD_{β_2} = 61%).

En la representación gráfica de los tres ajustes,se observa el sentido del cambio que los puntos atípicos determinan sobre la ecuación de regresión, línea de trazo continuo por la línea punteada. Las observaciones anómalas atraen la recta estimada hacia sí, tanto más, cuanto más se alejen éstas del centro de gravedad medio del resto de observaciones empleadas en la estimación.

La atracción que todo punto atípico produce, es debida a la particular forma de la función que el método de estimación de los mínimos-cuadrados trata de minimizar, una suma de cuadra-

Cuadro 1.a.- Resultados para datos originales.

*** DATOS ***

t	Y_t	X_{2t}
1	4.96	4
2	5.68	5
3	6.24	6
4	6.82	7
5	6.95	8
6	8.04	9
7	8.51	10
8	8.33	11
9	8.64	12
10	8.94	13
11	9.96	14

MATRICES DATOS

** (X'X) **

11	99
99	1001

** (X'Y) **

83.07
797.45

** (Y'Y) **

650.841901

** $(X'X)^{-1}$ **

.827272727	-.0818181818
-.0818181818	9.09090909E-03

1.MODELO DESCRIPTIVO

* ECUACION MINIMO-CUADRATICA *

$\hat{Y}_t$=+3.4756364 +.452909089 X_{2t}

S_e^2=.0862210881 R^2=.959662415

2.ESTRUCTURA ESTIMADA

* ESTIMACION PUNTUAL *

b_1=3.4756364

b_2=.452909089

S_u^2=.10538133

** VAR(b) **

.0871791002	-8.62210881E-03
-8.62210881E-03	9.5801209E-04

* ESTIMACION POR INTERVALOS *

I_{B_1}=2.80775585 ; 4.14351694 CD_{B_1}=19.2160649

I_{B_2}=.382896185 ; .522921994 CD_{B_2}=15.4584896

* DEPENDENCIA LINEAL CONJUNTA F_0 =214.116977

Cuadro 1.b.- Resultados con valor anómalo en Y.

*** DATOS ***

t	Y_t	X_{2t}
1	4.96	4
2	5.68	5
3	6.24	6
4	6.82	7
5	6.95	8
6	18.4	9
7	8.51	10
8	8.33	11
9	8.64	12
10	8.94	13
11	9.96	14

MATRICES DATOS

** (X'X) **

11	99
99	1001

** (X'Y) **

93.43
890.69

** (Y'Y) **

924.760301

** $(X'X)^{-1}$ **

.827272727	-.0818181818
-.0818181818	9.09090909E-03

1.MODELO DESCRIPTIVO

* ECUACION MINIMO-CUADRATICA *

$\hat{Y}_t$=+4.41745457 +.452909091 X_{2t}

S_e^2=9.87599297 R^2=.171981382

2.ESTRUCTURA ESTIMADA

* ESTIMACION PUNTUAL *

b_1=4.41745457

b_2=.452909091

S_u^2=12.0706581

** VAR(b) **

9.98572623	-.987599297
-.987599297	.109733255

* ESTIMACION POR INTERVALOS *

I_{B_1}=-2.73051062 ; 11.5654197 CD_{B_1}=161.811855

I_{B_2}=-.29640126 ; 1.20221944 CD_{B_2}=165.443875

* DEPENDENCIA LINEAL CONJUNTA F_0 =1.86932081

Cuadro 1.c.- Resultados con valor anómalo en X.

*** DATOS ***

t	Y_t	X_{2t}
1	4.96	4
2	5.68	5
3	6.24	6
4	6.82	7
5	6.95	8
6	8.04	9
7	8.51	10
8	8.33	11
9	8.64	22
10	8.94	13
11	9.96	14

MATRICES DATOS

** (X'X) **

11	109
109	1341

** (X'Y) **

83.07
883.85

** (Y'Y) **

650.841901

** $(X'X)^{-1}$ **

.467247387	-.0379790941
-.0379790941	3.83275262E-03

1.MODELO DESCRIPTIVO

* ECUACION MINIMO-CUADRATICA *

$\hat{Y}_t$=+5.24641812 +.232655051 X_{2t}

S_e^2=.853616389 R^2=.60064499

2.ESTRUCTURA ESTIMADA

* ESTIMACION PUNTUAL *

b_1=5.24641812

b_2=.232655051

S_u^2=1.04330892

** VAR(b) **

.487483367	-.0396239277
-.0396239277	3.99874499E-03

* ESTIMACION POR INTERVALOS *

I_{B_1}=3.66708952 ; 6.82574673 CD_{B_1}=30.1029878

I_{B_2}=.0896160543 ; .375694048 CD_{B_2}=61.4811482

* DEPENDENCIA LINEAL CONJUNTA F_0 =13.5363393

Cuadro 1.d.- Resultados con valor anómalo en X incrementado.

*** DATOS ***

t	Y_t	X_{2t}
1	4.96	4
2	5.68	5
3	6.24	6
4	6.82	7
5	6.95	8
6	8.04	9
7	8.51	10
8	8.33	11
9	8.64	112
10	8.94	13
11	9.96	14

MATRICES DATOS

** (X'X) **

11	199
199	13401

** (X'Y) **

83.07
1661.45

** (Y'Y) **

650.841901

** $(X'X)^{-1}$ **

.124302013	-1.84588399E-03
-1.84588399E-03	1.02031331E-04

1.MODELO DESCRIPTIVO

* ECUACION MINIMO-CUADRATICA *

$\hat{Y}_t$=+7.2589975 +.016186068 X_{2t}

S_e^2=1.90405776 R^2=.109207582

2.ESTRUCTURA ESTIMADA

* ESTIMACION PUNTUAL *

b_1=7.2589975

b_2=.016186068

S_u^2=2.32718171

** VAR(b) **

.289273371	-4.29560486E-03
-4.29560486E-03	2.37445495E-04

* ESTIMACION POR INTERVALOS *

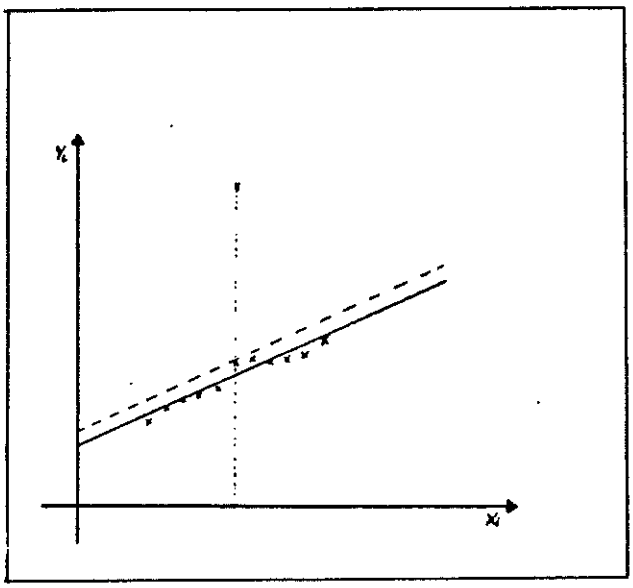
I_{B_1}=6.04240025 ; 8.47559475 CD_{B_1}=16.7598521

I_{B_2}=-.0186696929 ; .0510418289 CD_{B_2}=215.344214

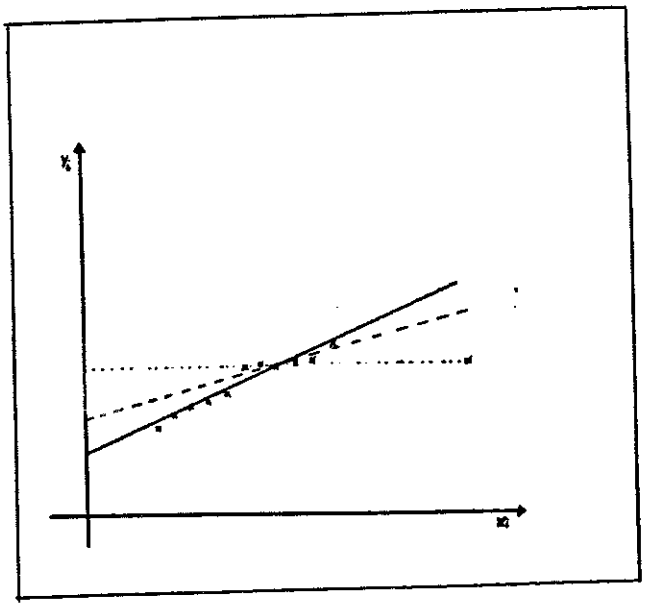
* DEPENDENCIA LINEAL CONJUNTA F_0 =1.10336395

Gráfico 2.- Resultados de estimación con datos anómalos

a) Anomalía en Y



b) Anomalía en X



dos de residuos, que hace que aquellas observaciones con residuo grande, ponderadas según una función cuadrática, arrastren la ecuación ajustada hacia esas observaciones.

Dependiendo de la gravedad con que se nos presente el problema del punto atípico podría determinar, casi en exclusiva, los valores de los coeficientes de la ecuación que se ajusta, como ocurre, por ejemplo, en la regresión que se muestra en el cuadro 1.d, en la que hemos supuesto un error en la variable X de mucha mayor cuantía que el inicialmente considerado en la estimación recogida en el cuadro 1.c.

El método de los mínimos-cuadrados se muestra así poco eficaz, distorsionando en gran medida los resultados a alcanzar, dando lugar a estimadores que son sesgados y no eficientes.

Además, la propia hipótesis de normalidad de las perturbaciones aleatorias del modelo, necesaria para la definición válida de los contrastes y test estadísticos, de cuya aplicación dependemos para evaluar el modelo, pueden verse alterados como consecuencia de la presencia de puntos atípicos e invalidarnos esta parte de los resultados econométricos.

En conclusión, que la presencia de punto/s atípico/s malogra el conjunto de resultados que se obtienen por el método de estimación de los mínimos-cuadrados, dependiendo la validez de la solución de la gravedad propia del problema.

ANALISIS DE PUNTOS INFLUYENTES EN LA REGRESION.

Los incovenientes apuntados sobre mínimos-cuadrados suponen que este método de estimación no debe ser empleado, si existen puntos atípicos que influyen fuertemente en la regresión. La solución de la ecuación deducida de la presencia de los valores atípicos, independientemente del resto de valores no anómalos, prejuzga conseguir una estimación,cuyas propiedades son debidas únicamente a esos puntos y no podemos tener confianza en el modelo construido, que aunque pudiera ser válido,no confirma sus propiedades por el total de información disponible en la muestra.

Es preciso analizar los datos en que se basa la estimación, si deseamos conocer la magnitud del problema con que nos estamos enfrentando, para tomar las medidas correctoras correspondientes.

A tal efecto,se han propuesto tradicionalmente, la utilización de gráficos de residuos frente a la variable endógena (o su estimación), o bien frente a cada una de las variables exógenas de la ecuación.

La utilización de estos gráficos, convenientes en cualquier aplicación econométrica empírica, puesto que su empleo permite conocer aproximativamente la presencia de determinados problemas especiales como no linealidad de la relación, autocorrelación o heteroscedasticidad, puede también ser válido para detectar puntos atípicos asociados a un Y o a una X particular.

No podemos,en cambio,detectar valores atípicos multivariantes, caracterizados por tener varias coordenadas X alejadas de sus valores medios, y por ello vamos a describir medidas cuantitativas de influencia que permitan conocer la presencia de puntos atípicos bajo circunstancias generales.

Cualquier medida de la influencia real de un valor anómalo, debemos establecerla en función de la alteración que su inclusión produce en el vector de estimadores m-c, o bien, alternativamente, por la modificación que establece sobre la ecuación estimada.

Supuesto el modelo lineal general resuelto para una muestra de tamaño T,tal y como lo hemos planteado en estas páginas,vamos a pasar a resolverlo para un tamaño de muestra T-1,tras la supresión en el vector Y, de valores de la variable endógena, del dato y en T, y en la matriz X el vector x_T que contiene los K datos de las variables exógenas, en dicho instante del tiempo.

La nueva estimación del modelo ($b_{(T)}$), nos supone resolver la ecuación:

$$(X'_{(T)}X_{(T)})\ b_{(T)} = X'_{(T)}Y_{(T)} \qquad e.1$$

en donde:

$X_{(T)}$, es la submatriz de dimensión T-1,K de valores de las variables exógenas,obtenida al suprimir el vector fila x_T correspondiente a la observación T.

$Y_{(T)}$, es el vector columna de dimensión T-1 resultante de eliminar la observación y_T.

es decir:

$$Y = \begin{bmatrix} Y_{(T)} \\ y_T \end{bmatrix} \qquad X = \begin{bmatrix} X_{(T)} \\ x_T \end{bmatrix}$$

ello nos permite expresar e.1, por:

$$(X'X - x_T'x_T)\, b_{(T)} = X'Y - x_T'y_T$$

y restando de la ecuación general:

$$X'X = b\, X'Y$$

nos conduce a:

$$X'X\,(b-b_{(T)}) = x_T'\,(y_T - x_T b_{(T)})$$

o lo que es lo mismo:

$$b-b_{(T)} = (X'X)^{-1}\, x_T'(y_T-\hat{y}_{(T)}) \qquad \text{e.2}$$

relación que nos permite analizar la diferencia entre la estimación con T y T-1 datos, como función de la discrepancia entre valor observado y_T y el previsto por la ecuación estimada con T-1 observaciones ($e_{(T)} = y_T-\hat{y}_{(T)}$), ponderado por una cantidad que depende de la columna T asociada a esa observación en la matriz $(X'X)^{-1}\, X'$.

En esta ecuación, premultiplicando por el vector fila x_T, nos surge la relación entre predicciones del punto T, a priori ($\hat{y}_{(T)}$) y a posteriori ($\hat{y}_T$), esta última, obtenida incluido el el propio dato T en la estimación.

$$\hat{y}_T - \hat{y}_{(T)} = x_T (X'X)^{-1} x_T' (y_T - \hat{y}_{(T)})$$

relación que vamos a escribir como:

$$\hat{y}_T - \hat{y}_{(T)} = v_{TT} \ (y_T - \hat{y}_{(T)}) \qquad \text{e.3}$$

en donde:

$$v_{TT} = x_T \ (X'X)^{-1} \ x_T'$$

es el término T de la diagonal de la matriz:

$$V = X \ (X'X)^{-1} X'$$

Las predicciones coinciden, haciéndolo también los estimadores, bien cuando el residuo $e_{(T)}$ sea cero o cuando lo sea el elemento v_{TT}. El punto T se comporta como los T-1 previos y no será un valor atípico.

Por el contrario, si el residuo obtenido es grande o lo es el término v_{TT} la diferencia de previsión de comportamiento de y estará indicándonos la presencia de un punto atípico.

El termino v_{TT} en el que x_T es una fila contenida en la ma-

triz X de datos cumple:

$$v_{TT} = \frac{1}{T} (1 + (x_T - \bar{x})\ S^{-1} (x_T - \bar{x})')$$

donde, S es la matriz de varianzas y covarianzas entre las variables exógenas del modelo. Por tanto, v_{TT} puede interpretarse como una medida de la distancia del punto x_T al centro de las medias muestrales $\bar{x}$, siendo mayor cuanto más alejado esté el punto de su media. En concreto, puede demostrarse que su valor esta acotado:

$$1/T \leq v_{TT} \leq 1$$

mostrando un valor próximo a la cota inferior 1/T, que sucederá cuando $x_T = \bar{x}$, que el punto muestral T no es atípico en cuanto a las variables x, si bien puede serlo respecto a y si y_T está alejado de $\hat{y}_{(T)}$, lo que puede producirse si el punto es atípico en y.

Alternativamente, un valor próximo a 1, representa un punto atípico en el vector x_T que incidirá sobre los resultados de la estimación en función del comportamiento concreto que tengamos en y.

Las ecuaciones e.2 y e.3 que nos permiten conocer si el punto T es atípico o no frente al resto de valores de la muestra, suponen la ejecución de dos regresiones independientes que contemplen las mismas en función del total de datos muestrales y de un tamaño de muestra T-1.

Desconocido a priori si un punto t es o no atípico, este es-

tudio debe extenderse al conjunto de puntos T de la muestra, lo que nos supone ejecutar T+1 regresiones independientes.

El hecho de tener que emplear en las ecuaciones e.2 y e.3 el residuo a priori $e_{(T)}$ para t=1,2...T, es el que condiciona este tipo de actuación. Por ello resulta interesante analizar la relación entre residuos a priori y a posteriori, que puede permitirnos llevar a cabo el análisis de puntos atípicos según una sola regresión sobre el total de puntos muestrales.

Basta tener en cuenta, que en general:

$$y_T - \hat{y}_{(T)} = y_T - \hat{y}_T \;/\; 1-v_{TT} \qquad e.4$$

o, lo que es lo mismo:

$$e_{(T)} = e_T \;/\; 1-v_{TT} \qquad e.5$$

pudiendo entonces escribirse la ecuacion e.3, por:

$$\hat{y}_T - \hat{y}_{(T)} = (y_T - \hat{y}_T)\; v_{TT} \;/(1-v_{TT}) \qquad e.6$$

y, a su vez, la ecuacion e.2 que marca la variación producida en el vector de estimadores, como:

$$b - b_{(t)} = (X'X)^{-1} x_t' \;(y_t - \hat{y}_{(t)}) \;/\; (1-v_{tt}) \qquad e.7$$

Estas dos últimas ecuaciones, nos permiten llevar a cabo el análisis de puntos atípicos en cualquier muestra, mediante la ejecución de una única regresión sobre el total de datos, que

nos facilita el conocimiento de los residuos a posteriori, y de la obtención de los elementos v_{tt} de la diagonal de la matriz V. Las distancias entre b y $b_{(t)}$ o $\hat{y}_t$ e $\hat{y}_{(t)}$ marcarán su presencia.

Las dos relaciones propuestas, que definen distancias absolutas entre estimadores y/o previsión, tienen la desventaja de medir estas diferencias en base a la magnitud de los datos de las variables que se están manejando en la muestra, siendo conveniente que sean sustituidas por otras de caracter relativo, adimensionales, independientes de las unidades en que se formalizan los datos muestrales.

Así, si suponemos que un punto t es atípico y conlleva como error absoluto de previsión el descrito en la ecuacion e.6, el error relativo que estamos cometiendo al incluirlo en la estimación, vendrá dado por la razón:

$$\hat{y}_t - \hat{y}_{(t)} \; / \hat{y}_{(t)} = ((y_t - \hat{y}_t) / \hat{y}_{(t)}) * (v_{tt}/(1-v_{tt})) \qquad e.8$$

en donde, según la ecuación e.4:

$$\hat{y}_{(t)} = y_t - ((y_t - \hat{y}_t)/(1-v_{tt}))$$

es decir:

$$\hat{y}_{(t)} = (\hat{y}_t - v_{tt} y_t)/(1-v_{tt})$$

que sustituido en e.7 determina:

$$\hat{y}_t - \hat{y}_{(t)} \; / \hat{y}_{(t)} = v_{tt}(y_t - \hat{y}_t)/(\hat{y}_t - v_{tt} y_t) \qquad e.9$$

relación que facilita efectuar el cálculo en base a los resultados propios de la regresión única sobre el total de datos de la muestra.

A efectos de interpretación conviene que se escriba en términos absolutos y en tantos por cien, como forma mas expresiva de medir esta distancia de previsión, pareciendo recomendable que, en cualquier regresión, su magnitud no supere una cota máxima del 10%, valores superiores serán representativos de presencia de puntos atípicos.

Aplicando lo descrito en este apartado al ejemplo analizado anteriormente, presentamos en los cuadros 2.a, 2.b y 2.c, los residuos de la regresión, las ponderaciones v_{tt} , las distancias absolutas de previsión de cada punto t y las correspondientes relativas.

Así, para el modelo resuelto con los datos iniciales no se presentan divergencias de previsión importantes, para la regresión sobre el total de puntos muestrales y la ejecutada sin el correspondiente punto t, sin embargo, en las dos estimaciones con dato atípico en Y o en X, las distancias obtenidas marcan la presencia del error de manipulación cometido en los datos, con valores del orden del 13% y 23%, respectivamente.

Adicionalmente, en estos cuadros de resultados, puede apreciarse como se incrementan todas las distancias de predicción, derivadas, exclusivamente, de la incorporación de un único dato atípico en la muestra.

Cuadro 2.a.- Resultados para datos originales.

t	e_t	v_{tt}	$\hat{y}_t - \hat{y}_{(t)}$	$\hat{y}_t - \hat{y}_{(t)} / \hat{y}_{(t)}$
1	-.327272752	.318181818	-.152727284	2.80748683
2	-.0601818413	.236363637	-.0186277128	.323464643
3	.0469090715	.172727273	9.79420175E-03	.158397731
4	.17399998	.127272727	.0253749971	.383271926
5	-.148909109	.1	-.0165454565	.232528455
6	.488181802	.090909091	.0488181803	.65064881
7	.505272713	.1	.0561414126	.706306926
8	-.127636373	.127272727	-.0186136378	.219597555
9	-.270545464	.172727273	-.0564875145	.629946546
10	-.423454553	.236363636	-.131069266	1.38047225
11	.143636357	.318181818	.0670302999	.687537266

Cuadro 2.b.- Resultados con valor anómalo en Y.

t	e_t	v_{tt}	$\hat{y}_t - \hat{y}_{(t)}$	$\hat{y}_t - \hat{y}_{(t)} / \hat{y}_{(t)}$
1	-1.26909093	.318181818	-.592242434	8.68220922
2	-1.00200002	.236363637	-.310142864	4.43559104
3	-.894909108	.172727273	-.186849155	2.5519711
4	-.767818201	.127272727	-.111973488	1.45424048
5	-1.09072729	.1	-.121191922	1.48484589
6	9.90636362	.090909091	.990636364	13.2032035
7	-.436545473	.1	-.0485050526	.53924158
8	-1.06945456	.127272727	-.155962124	1.63218548
9	-1.21236365	.172727273	-.253130873	2.50488358
10	-1.36527274	.236363636	-.42258442	3.93913168
11	-.798181832	.318181818	-.372484855	3.34647389

Cuadro 2.c.- Resultados con valor anómalo en X.

t	e_t	v_{tt}	$\hat{y}_t - \hat{y}_{(t)}$	$\hat{y}_t - \hat{y}_{(t)} / \hat{y}_{(t)}$
1	-1.21703833	.224738676	-.352804369	5.40295358
2	-.729693377	.183275261	-.163745186	2.49101265
3	-.402348427	.149477352	-.0707117883	1.05334655
4	-.055003481	.123344948	-7.73896356E-03	.112440116
5	-.15765853	.104878049	-.0184722529	.25921855
6	.699686417	.0940766551	.0726597435	.999768909
7	.937031366	.0909407665	.0937390518	1.25332497
8	.52437632	.0954703833	.0553463451	.714120833
9	-1.72482924	.651219512	-3.22048537	23.705637
10	.669066217	.127526132	.0977948224	1.19653933
11	1.45641116	.155052265	.267258956	3.24487921

CONTRASTES PARA DETECCION DE PUNTOS ATIPICOS.

¿ Cúando un valor concreto expresivo de la diferencia en el vector de estimadores o en las predicciones será significativo y, por tanto, característico de presencia en la muestra de un punto atípico ?.

Sí un punto t es atípico frente al resto de valores muestrales su comportamiento, desde el punto de vista de la estructura que relaciona y_t con x_t, no será plenamente explicable por la estructura estimada en base al resto de datos muestrales.

La previsión del valor y_t realizada con la ecuación ajustada a T-1 datos, no compatibilizará entre los valores deducidos de su intervalo de confianza, a éste que estamos tratando de predecir.

Así, podemos plantear un contraste basado en la predicción a priori del punto t, dado por la expresión:

$$\frac{\hat{y}_{(t)} - y_t}{S_{e_{(t)}}\sqrt{1 + x_t (X'_{(t)} X_{(t)})^{-1} x'_t}} \qquad e.10$$

que como se demuestra en los manuales econométricos, dentro del capítulo específico dedicado al modelo lineal, se comporta según una distribución 't' de Student con T-K-1 grados de libertad, que nos permite, determinado un nivel de significación ε, aceptar o rechazar que y_t pertenece a la estructura estimada en la ecuación o en nuestro caso, que el dato y_t correspondiente

al vector x_t es o no un valor atípico.

La ecuación e.10 podemos también escribirla como:

$$\frac{\hat{y}_{(t)} - y_t}{s_{e_{(t)}}\sqrt{1 \,/\, (1-v_{tt})}} \qquad \text{e.11}$$

si tenemos en cuenta la relación entre la inversa de $X'X$ y de $X'_{(t)}X_{(t)}$

$$(X'_{(t)}X_{(t)})^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x'_t x_t (X'X)^{-1}}{1 - v_{tt}}$$

que implica que:

$$x_t\ (X'_{(t)}X_{(t)})^{-1}\ x'_t = v_t / (1-v_{tt})$$

Operativamente, las ecuaciones e.10 y e.11 tienen el inconveniente de basarse en la varianza de los residuos obtenidos sin incluir el punto t, lo que nos supone en la aplicación del contraste, según se ha explicitado en puntos anteriores,ejecutar un total de T+1 regresiones independientes con las consiguientes complicaciones de cálculo.

Los T-1 residuos de la regresión sin el punto t (2), pueden ser conocidos sin necesidad de ajuste específico sobre los correspondientes puntos muestrales, basándonos en los residuos de la regresión ajustada al total de datos T de la muestra.

---

(2) Denominaremos a estos residuos por:

$$e^1_{(t)},\ e^2_{(t)} \ldots,\ e^{t-1}_{(t)},\ e^{t+1}_{(t)} \ldots,\ e^T_{(t)}$$

En efecto, si en la ecuación e.7 premultiplicamos por el vector x_j, obtenemos:

$$x_j b - x_j b_{(t)} = x_j'(X'X)^{-1}x_t' \frac{(y_t-\hat{y}_t)}{1-v_{tt}}$$

siendo $x_j b$ la predicción del punto y_j con la ecuación de regresión sobre el total de datos y $x_j b_{(t)}$ la llevada a cabo con la ecuación sin el punto t ($\hat{y}^j_{(t)}$), es decir:

$$y_j-\hat{y}^j_{(t)} = x_j'(X'X)^{-1}x_t' \frac{(y_t-\hat{y}_t)}{1-v_{tt}}$$

ó lo que es lo mismo:

$$(y_j-\hat{y}^j_{(t)}) = (y_j-\hat{y}_j) + x_j'(X'X)^{-1}x_t' \frac{(y_t-\hat{y}_t)}{1-v_{tt}}$$

ó también (3):

$$e^j_{(t)} = e_j + x_j'(X'X)^{-1}x_t' \frac{e_t}{1-v_{tt}} \qquad \text{e.12}$$

relación que nos permite calcular la varianza residual:

$$S^2_{e(t)} = \sum_{j\neq t}^{T} e_t^{j\,2} \;/\; (T-K-1)$$

necesaria en la resolución del contraste descrito.

Alternativamente al contraste 't', que supone aplicar procedimientos de detección de cambio estructural, puede emplear-

---

(3) Si j=t nos surje la relación ya conocida:

$$e^t_{(t)} = e_{(t)} = e_t/(1-v_{tt})$$

se también, el método propuesto por G.CHOW(4),un contraste 'F' obtenido por comparación de la suma de cuadrados de residuos de la regresión total y de la regresión sin el dato t,según la expresión:

$$F^{1}_{T-K-1} = \frac{\sum_{t=1}^{T} e^{2}_{t} - \sum_{j\neq t}^{T} e^{j\ 2}_{(t)}}{\sum_{j\neq t}^{T} e^{j\ 2}_{(t)} \ / \ (T-K-1)}$$

en la que los residuos de la regresión sin el punto t los calculamos de acuerdo con la ecuación e.12.

De los contrastes mencionados,en el cuadro 3, se contiene la aplicación del estadístico 't' al ejemplo que venimos desarrollando en este documento. En el mismo,se muestra la clara atipicidad de los correspondientes puntos muestrales que han sido manipulados en las regresiones b y c.

Cuadro 3.- Resultados del contraste 't' para valores atipicos

t	Regresion a	Regresion b	Regresion c
1	-1.26019716	-.421687277	-1.42954236
2	-.200517614	-.313059401	-.772583503
3	.149997514	-.268198627	-.406839931
4	.551116291	-.223733941	-.0542341625
5	-.461909815	-.31391468	-.154041168
6	1.74812522	35.4735968	.698950335
7	1.84763149	-.124994286	.957736522
8	-.400768303	-.312548059	.517363057
9	-.90724153	-.36470991	-8.9079349
10	-1.62246701	-.428814098	.680005076
11	.513467251	-.263452078	1.70859283

---

(4) CHOW G: Text of equality between sets of coeffients in two linear regression.

TRATAMIENTO DE DATOS ATIPICOS. CONCLUSIONES.

El comportamiento atípico de un dato o de un subconjunto de datos, que desvirtua los resultados de cualquier regresión mínimo-cuadrática, puede ser la consecuencia práctica de un normal error de medición,tal y como nos ha servido en estas páginas para presentar el problema, o también puede surgir, porque el punto o puntos analizados determinan un comportamiento estructural distinto del que representan el conjunto muestral de valores base de la regresión. Incluso puede llegar a pensarse, que estos puntos anómalos esten encubriendo fallos en la definición de la especificación del modelo, bien en la forma funcional de la ecuación, bien en la enumeración de las variables exógenas, presentando, en este último caso, el grave problema econométrico de la variable omitida.

Es por ello,que el necesario tratamiento de puntos atípicos en la estimación mínimo-cuadrática, requiere, como soluciones, seguir un conjunto estructurado de pasos que nos garantizen la no presencia de cada una de las posibles causas de atipicidad, tal y como las hemos enumerado.

Así, en primer lugar,ante la detección de valores anómalos, es conveniente revisar la información estadística utilizando, si es posible, criterios empíricos que permitan garantizar la validez de esos datos.

Esta operación es recomendable de ejecutar incluso antes de la estimación de la ecuación, analizandose cada serie de datos

de las variables que van a introducirse en el modelo, endógena y exógenas, por separado, prestando atención a comportamientos no homogeneos de algunos datos particulares con el resto.

Una vez depuradas las series de datos sí el problema continua, es necesario precisar sí esos datos atípicos son debidos a los errores de especificación que hemos mencionado.Sera conveniente probar nuevas regresiones con forma funcional distinta o con variables exógenas adicionales.

De persistir el problema,la presencia de cambio estructural parece la única causa directa de la presencia de puntos atípicos, implicando que su solución puede venir dada, bien por la estimación de dos estructuras distintas o por la eliminación de esos puntos, si son pocos los datos atípicos,del proceso de estimación.

Adicionalmente,puede pensarse en un cambio de método de estimación, linea seguida dentro de la Econometría por los denominados métodos robustos de estimación. Procedimientos que modifican la ponderación cuadrática del método mínimo-cuadrático para garantizar que unos pocos valores con alto peso no arrastren la ecuación hacia sí (5).

En cualquier caso,vistas las negativas consecuencias que la presencia de anomalias muestrales tiene sobre el método de estimación de los mínimos-cuadrados, resulta imprescindible analizar, previamente a la estimación,la posible incorporación de datos atípicos en la muestra.

---

(5) Métodos de estimación planteados por primera vez por P.J. HUBER: Robust estimation of a location parameter.

La robustez del diseño obtenido y la confianza en una solución correcta, solo estaran garantizadas en el caso de no presencia de puntos atípicos entre los datos muestrales.

Así, para el ejemplo estudiado en este documento, suponiendo que en el caso concreto de atipicidad manifiesta en un dato de X (dato noveno del cuadro 1.c), no nos fuera posible detectar la causa originaria de la anomalía, convendría abordar la estimación del modelo lineal eliminando la observacion atípica.

La escasez de información muestral (solo once datos), no es óbice para que los resultados así calculados (cuadro 4), representen una aproximación válida a los iniciales, cuando no aparecía el problema de punto atípico. Ello indica claramente que es más adecuado perder grados de libertad en la resolución por mínimos-cuadrados que incorporar datos atípicos en la información muestral.

Cuadro 4.- Resultados estimación sobre T-1 puntos.

*** DATOS ***

t	Y_t	X_{2t}
1	4.96	4
2	5.68	5
3	6.24	6
4	6.82	7
5	6.95	8
6	8.04	9
7	8.51	10
8	8.33	11
9	8.94	13
10	9.96	14

MATRICES DATOS

** (X'X) **

10	87
87	857

** (X'Y) **

74.43
693.77

** (Y'Y) **

576.192301

** $(X'X)^{-1}$ **

.856143855	-.0869130868
-.0869130868	9.99000998E-03

1.MODELO DESCRIPTIVO

* ECUACION MINIMO-CUADRATICA *

$\hat{Y}_t$=+3.42509494 +.46182817 X_{2t}

S_u^2=.0859954834 R^2=.961280406

2.ESTRUCTURA ESTIMADA

* ESTIMACION PUNTUAL *

b_1=3.42509494

b_2=.46182817

S_u^2=.107494354

** VAR(b) **

.0920306309	-9.34266615E-03
-9.34266615E-03	1.07386967E-03

* ESTIMACION POR INTERVALOS *

I_{B_1}=2.72553408 ; 4.1246558 CD_{B_1}=20.4245684

I_{B_2}=.386260663 ; .537395676 CD_{B_2}=16.3626889

* DEPENDENCIA LINEAL CONJUNTA F_0 =198.613735

BIBLIOGRAFIA .

BARNETT V /LEWIS T (1978). Outliers in statistical data.Wiley.

BOX G.E.P/ DRAPER N.A (1975): Robust desing. Biometrica, 62.

CHOW G. (1960). Text of equality between sets of coefficients in two linear regression. Econometrica, 28.

COOK R.D (1977). Deteccion of influential observations in linear regression. Technometrics, 19.

COOK R.D (1979). Influential observation in linear regression. J.A.S.A. 74.

DRAPER N.R / JOHN J.A (1981): Influential observations and outliers in regression. Technometrics, 23.

HUBER P.J. (1964). Robust estimation of a location parameter. Annals of Mathematical Statistics, 135.

PEÑA D/RUIZ-CASTILLO J(1982): Métodos robustos de construcción de modelos de regresión. Estadística Española, 97.

PEÑA D (1987). Observaciones influyentes en modelos econométricos. Investigaciones Económicas, XI.

PEÑA D (1987). Modelos y métodos 2.Alianza Universidad Textos.

RAYMOND J.L / URIEL E (1987): Investigación econométrica, -un caso de estudio-. Editorial A.C.