

La estimación de modelos de regresión semilogísticos con errores en las variables explicativas

Juan C. Estévez Núñez

- Universidad de Santiago de Compostela -

1.- Introducción : modelos con errores.

Las inquietudes que han originado la aparición de los modelos con errores en las variables (EVM) se han constituido en motivo de frecuente reencuentro de los investigadores interesados en los aspectos metodológicos de la Econometría con enfoques ya perfilados en los propios inicios de esta disciplina. En este sentido, teniendo en cuenta algunos resultados que O. Reiersøl (1950) atribuye a C. Gini (1921), R. Frisch (1934) demostró que si un modelo de regresión lineal simple recoge con fidelidad la relación entre una variable cuyo comportamiento se desea explicar (Y_t) y una variable explicativa (χ_t), y de esta última sólo se dispone de observaciones (x_t) con cierto tipo de errores aditivos ($x_t = \chi_t + E_t$), el intervalo construido con los estimadores del parámetro β_1 obtenidos de la regresión de Y_t en función de x_t (EMCO directo: b_1^d) y de la regresión de x_t en función de Y_t (EMCO inverso: b_1^i), “acota consistentemente” el verdadero parámetro. Posteriormente O. Reiersøl (1950) demostró que, suponiendo la normalidad de las perturbaciones y de los errores de medida de un EVM de regresión simple, la propia hipótesis de normalidad referida al regresor medido sin errores (χ_t) es el único supuesto distribucional que impide la identificación de los parámetros. Una generalización de este resultado para un EVM de regresión múltiple ha sido proporcionada más recientemente por A. Kapteyn y T. Wansbeek (1983). Además, S. Klepper y S. Garber (1980) han estudiado los sesgos que se producen en la estimación del EVM de regresión múltiple con errores en más de un regresor. Y el propio Klepper en colaboración con E. E. Leamer (1984) ha conseguido demostrar que los estimadores máximo-verosímiles de los parámetros de un EVM de regresión múltiple con errores en más de un regresor están contenidos en la envoltura convexa formada con los estimadores MCO de Y_t respecto a los

regresores medidos con error y de las K estimaciones MCO de cada x_{it} respecto a las demás variables, con la condición de que los K+1 vectores resultantes se hallen situados en el mismo ortante.

2.- Modelos semilogísticos.

Con objeto de apuntar un camino paralelo al surgido a partir de los resultados de Gini, se presentan aquí las particularidades de la estimación de un tipo de modelos de regresión no lineal cuya forma funcional es semejante a la curva logística y que se denominarán en adelante modelos semilogísticos¹ en la medida en que no suponen más que la incorporación de parámetros adicionales a la expresión matemática de la función logística :

$$(1) \quad Y_t = \frac{1}{1 + e^{-b_1 c_t}}$$

Como es bien sabido, dicha curva trata de recoger el incremento en la probabilidad de un suceso ante incrementos de un factor causante (generalmente una combinación lineal de variables). No es el objeto del presente trabajo analizar las particularidades de la estimación de este tipo de funciones, sino utilizar una expresión matemática semejante para relajar la hipótesis de linealidad de un modelo de regresión y analizar el comportamiento de Y_t condicionado a χ_t como trata de recoger el primer gráfico.

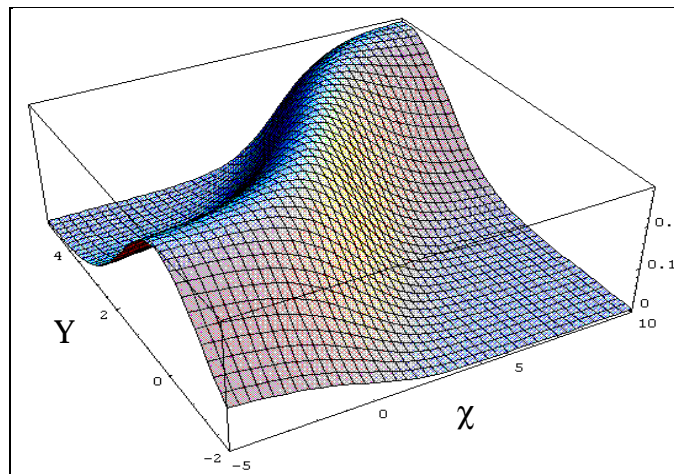


Gráfico 1

¹ También se conocen como modelos de crecimiento en forma de S (S-shaped growth models) tal como se indica en Bewley & Fiebig (1988).

Este análisis se pretende relacionar con la información que nos proporcionarían el regresando y las variables medidas con errores (teniendo en cuenta la acepción más amplia de estas, y que incluye la posibilidad de utilizar variables proxy).

Obviamente la expresión (1) presenta ciertas restricciones que resulta poco costoso intentar subsanar incorporando parámetros adicionales:

- a) Y_t varía entre cero y uno;
- b) la función tiene un punto de inflexión en ($Y_t = 0.5, \chi_t = 0$);
- c) es siempre creciente y su derivada es simétrica respecto al eje de ordenadas.

A fin de evitar estos aspectos de la misma, se propone en este trabajo la inclusión de un parámetro β_2 de forma que

$$(2) \quad Y_t = \frac{1}{b_2 + e^{-b_1 \chi_t}}$$

por lo que se cumplirá que :

- a) el máximo valor de Y_t será ($1/\beta_2$);
- b) cuando $\chi_t = 0$ entonces $Y_t = 1/(1+\beta_2)$;
- c) en el punto de inflexión $\chi_t = (-\ln \beta_2)/\beta_1$ con $Y_t = 1/(2\beta_2)$.
- d) la función presenta ahora una derivada simétrica respecto a la recta paralela al eje de ordenadas que pasa por el punto de inflexión

3.- Diseño de una simulación.

Como resulta evidente, los procesos de maximización o minimización de una función objetivo en el marco de modelos no lineales provoca una dificultad añadida respecto a los lineales por cuanto, salvo muy contados casos, no se pueden obtener fórmulas explícitas para los estimadores que se pretende utilizar, y a menudo ni siquiera resultan identificables. El tipo de modelos que nos ocupa no es una excepción y por ello se ha diseñado un experimento de simulación de tipo Monte Carlo para estudiar el comportamiento de dichos estimadores. En concreto, se ha tomado una muestra de 100 observaciones del regresor medido sin errores, y se ha simulado el valor esperado de Y_t con $\beta_2 = 0.04$ y $\beta_1 = 0.6$. Además se han generado 100 muestras ($\xi_1, \xi_2, \dots, \xi_{100}$) pseudo-aleatorias de 100 observaciones de variables que representan las posibles perturbaciones del modelo y otras 100 representando los errores de medición (E_1, E_2, \dots, E_{100}). Cada una de estas 200 muestras tienen media nula y su

varianza es unitaria. El promedio de los valores absolutos de sus coeficientes de correlación simple tomadas dos a dos es de 0,08 y la distribución de frecuencias acumulada de cada una de ellas es semejante a la distribución acumulada de una variable normal.²

Simulando así un sencillo modelo de regresión semilogístico con un único regresor y perturbaciones con distribución normal, se ha calculado mediante un método de estimación no lineal con restricciones en los parámetros, el valor de los estimadores MCO de Y_t en función del regresor medido con errores x_t introduciendo un coeficiente que modula la varianza de las perturbaciones y otro que lo hace con la varianza de los errores de medida.

$$Y_t = E[Y_t / \chi_t] + k \xi_t ; \quad x_t = \chi_t + m E_t ; \quad t=1,2, \dots, 100$$

El coeficiente k indica la raíz cuadrada del porcentaje que se desee que represente la varianza de las perturbaciones con respecto a la variabilidad máxima de la $E(Y_t / \chi_t)$. Esta última viene dada por la expresión : $V_{max}(E[Y_t / \chi_t]) = \frac{1}{4b_2^2}$.

En el presente trabajo se ha realizado el análisis para los siguientes valores de k :

$$0, \frac{\sqrt{0.1}}{2b_2}, \dots, \frac{\sqrt{0.5}}{2b_2}$$

Puesto que la variable explicativa puede tomar valores del intervalo $[-\infty, \infty]$ la decisión de cual es la varianza máxima de los errores de medición digna de ser estudiada resulta más complicada y subjetiva (si cabe) que en el caso anterior. A este respecto, teniendo en cuenta que la derivada de $E(Y_t / \chi_t)$ respecto a χ_t es siempre positiva con dos puntos de inflexión (uno en el tramo creciente y otro en el tramo decreciente), que la distancia que separa dichos puntos depende exclusivamente del parámetro β_1 , y que esa distancia nos indica en cierta medida la velocidad con que crece la función semilogística, se propone establecer la varianza máxima de los errores de medición como un porcentaje del cuadrado de la distancia que separa los mencionados puntos de inflexión :

$$Distancia_{c_{inf(1^{a}derivada)}} = \frac{\ln(2 + \sqrt{3}) - \ln(2 - \sqrt{3})}{b_1} = \frac{2,63391579}{b_1}$$

² Calculando el estadístico de Kolmogorov-Smirnov resulta una probabilidad superior a 0.999 de que no se pueda rechazar la hipótesis de que la muestra proceda de una variable con distribución $N(0,1)$. De esta forma se ha pretendido generar muestras de variables aleatorias independientes con distribución normal.

Por lo tanto m expresa la raíz cuadrada del porcentaje que se desee que la varianza de los errores de medición represente respecto al cuadrado de la susodicha distancia. Se han tenido en cuenta los siguientes valores de m :

$$0, Dist * \sqrt{0.1}, \dots, Dist * \sqrt{0.5}$$

El gráfico de la derecha representa una de las 10.000 muestras de 100 observaciones simuladas con un porcentaje para definir k y m del 20%, además de la función $E[Y_t / \chi_t]$ con $\beta_2 = 0.04$ y $\beta_1 = 0.6$.

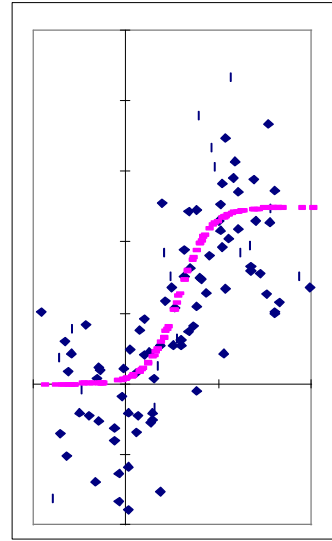


Gráfico 2

4.- Resultados.

Además de los resultados obtenidos que demuestran la necesidad de disponer de información sobre todos los tramos relevantes de la función³ con objeto de analizar comportamientos sistemáticos de los estimadores, también se han calculado los valores promedio de los estimadores MCO de los parámetros de las regresiones directas que se presentan a continuación. Así, en la tabla 1 se puede observar que considerando las 100 muestras simuladas del regresando generadas con $k = \frac{\sqrt{0.4}}{2b_2}$ en combinación con las 100 muestras simuladas del regresor medido con errores generadas con $m = Dist * \sqrt{0.3}$ el promedio de las estimaciones MCO del parámetro β_2 resulta 0.04159, información que se ofrece en las tablas 2 y 3 sobre b_1 (0.617) y la suma de cuadrados de los errores .

Tabla 1

b2		% Varianza Perturbación vs Varianza máxima de E[Y]						
		0	0,1	0,2	0,3	0,4	0,5	0,6
%	0	0,04000	0,03999	0,03999	0,03998	0,03998	0,03998	0,03999
Varza Error	0,1	0,04050	0,04049	0,04049	0,04048	0,04048	0,04048	0,04048
vs	0,2	0,04104	0,04104	0,04102	0,04102	0,04102	0,04102	0,04102
Distancia ²	0,3	0,04159	0,04159	0,04158	0,04158	0,04159	0,04159	0,04160
	0,4	0,04217	0,04217	0,04217	0,04218	0,04219	0,04220	0,04221
	0,5	0,04275	0,04276	0,04277	0,04279	0,04280	0,04281	0,04282

³ Que no son presentados aquí por considerarse triviales.

Tabla 2

b1		% Varianza Perturbación vs Varianza máxima de E[Y]						
		0	0,1	0,2	0,3	0,4	0,5	0,6
%	0	0,600	0,601	0,602	0,603	0,604	0,605	0,607
Varza Error	0,1	0,601	0,602	0,603	0,605	0,606	0,608	0,610
vs	0,2	0,603	0,605	0,606	0,608	0,611	0,614	0,617
Distancia ²	0,3	0,605	0,607	0,610	0,613	0,617	0,622	0,629
	0,4	0,609	0,612	0,615	0,620	0,638	0,639	0,647
	0,5	0,614	0,618	0,624	0,630	0,684	0,681	0,691

Tabla 3

SCE		% Varianza Perturbación vs Varianza máxima de E[Y]						
		0	0,1	0,2	0,3	0,4	0,5	0,6
%	0	0,0	1516,7	3033,3	4549,7	6066,1	7582,4	9098,5
Varza Error	0,1	556,7	2072,3	3587,7	5103,0	6618,1	8133,1	9647,9
vs	0,2	1074,9	2586,9	4103,6	5617,7	7131,6	8645,3	10159,0
Distancia ²	0,3	1562,3	3075,5	4588,6	6101,3	7614,0	9126,2	10638,3
	0,4	2021,4	3533,3	5044,9	6556,2	8067,2	9578,1	11088,7
	0,5	2452,2	3962,6	5472,8	6982,9	8492,3	10001,7	11510,7

De la primera tabla parece poder deducirse que el estimador b_2 de mínimos cuadrados de esta función semilogística con el regresor medido con errores, presentan un sesgo sistemático a medida que aumenta la varianza de los errores, y un sesgo menos regular y perceptible a medida que aumenta la varianza de las perturbaciones. Este comportamiento se repite para el estimador b_1 . Los gráficos 3, 4 y 5 recogen la información de las tablas 1, 2 y 3, mostrando la existencia de una relación aproximadamente lineal entre el estimador b_2 y los porcentajes de perturbación y de error (situación que se repite para la suma de cuadrados de los errores), y una relación no lineal entre b_1 y dichos porcentajes.

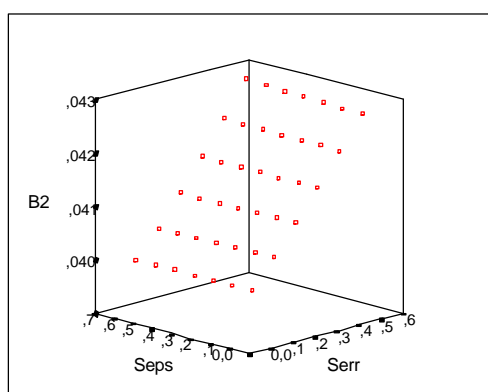


Gráfico 3

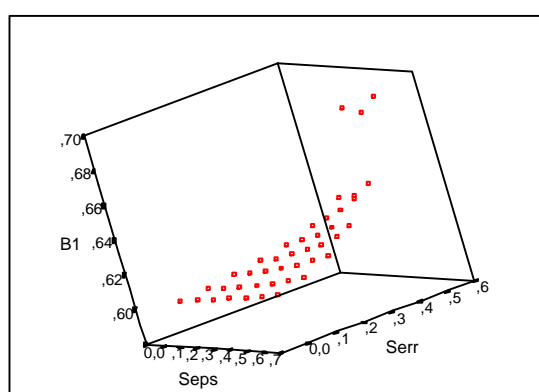


Gráfico 4

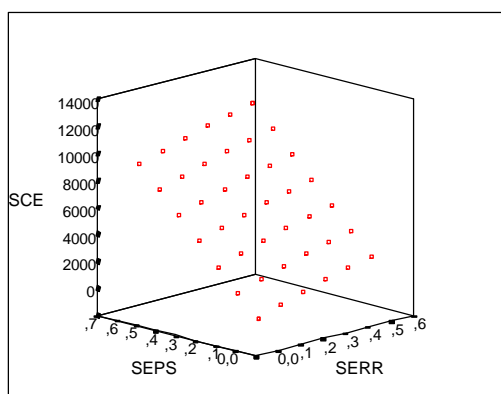


Gráfico 5

De manera parecida en las tablas 4, 5 y 6 se ofrecen los resultados obtenidos con el mismo procedimiento que los anteriores pero teniendo en cuenta únicamente 25 muestras simuladas del regresor medido con errores, y se pueden observar pautas muy similares a las observadas en las tablas 1, 2 y 3

por cuanto los sesgos se siguen manteniendo, y los porcentajes de varianza de los errores tienen una influencia mayor en la generación de dichos sesgos que los porcentajes de varianza de las perturbaciones.

Tabla 4

b2		% Varianza Perturbación vs Varianza máxima de E[Y]						
		0	0,1	0,2	0,3	0,4	0,5	0,6
%	0	0,04000	0,04020	0,04027	0,04031	0,04034	0,04037	0,04038
Varza Error	0,1	0,04050	0,04067	0,04073	0,04076	0,04078	0,04080	0,04081
vs	0,2	0,04104	0,04125	0,04124	0,04127	0,04130	0,04131	0,04133
Distancia ²	0,3	0,04159	0,04174	0,04179	0,04182	0,04186	0,04189	0,04191
	0,4	0,04217	0,04232	0,04238	0,04243	0,04246	0,04248	0,04250
	0,5	0,04275	0,04291	0,04297	0,04302	0,04305	0,04308	0,04310

Tabla 5

b1		% Varianza Perturbación vs Varianza máxima de E[Y]						
		0	0,1	0,2	0,3	0,4	0,5	0,6
%	0	0,600	0,605	0,607	0,610	0,612	0,614	0,615
Varza Error	0,1	0,601	0,605	0,608	0,610	0,613	0,616	0,618
vs	0,2	0,603	0,608	0,611	0,614	0,618	0,622	0,626
Distancia ²	0,3	0,605	0,611	0,615	0,619	0,625	0,634	0,643
	0,4	0,609	0,615	0,621	0,627	0,684	0,671	0,679
	0,5	0,614	0,622	0,629	0,639	0,834	0,803	0,818

Tabla 6

SCE		% Varianza Perturbación vs Varianza máxima de E[Y]						
		0	0,1	0,2	0,3	0,4	0,5	0,6
%	0	0,0	1549,5	3098,8	4648,0	6197,1	7746,0	9294,8
Varza Error	0,1	556,7	2095,7	3640,0	5185,1	6730,5	8276,0	9821,5
vs	0,2	1074,9	2594,2	4143,7	5684,7	7226,4	8768,4	10310,9
Distancia ²	0,3	1562,3	3082,7	4617,7	6154,9	7693,5	9231,9	10770,6
	0,4	2021,4	3533,7	5064,5	6597,9	8132,1	9667,4	11203,5
	0,5	2452,2	3957,1	5483,8	7014,0	8544,7	10077,5	11610,6

A continuación se presentan gráficamente los histogramas de las estimaciones mínimo cuadráticas obtenidas en el proceso de simulación. En los gráficos 6 y 7 se

representa las correspondientes a la estimación de las 100 muestras del regresando respecto al regresor medido sin errores, con $k = \frac{\sqrt{0.4}}{2 b_2}$.

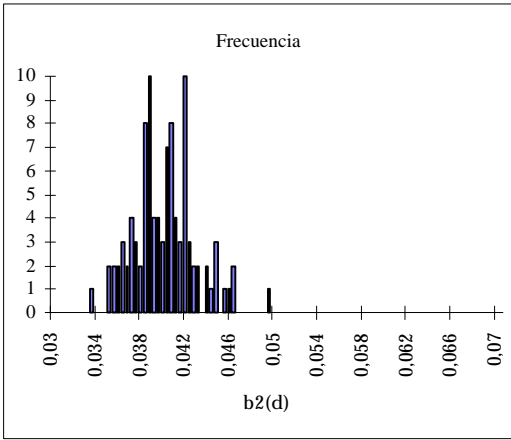


Gráfico 6

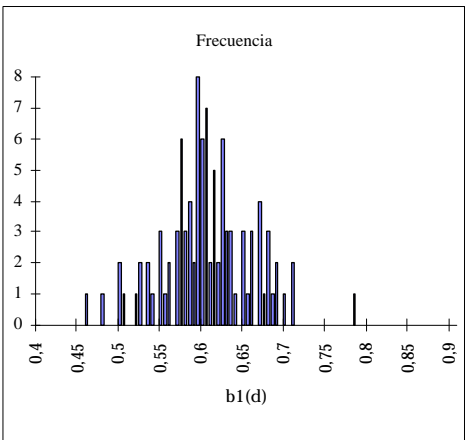


Gráfico 7

Por último, los gráficos 8 y 9 muestran los histogramas de las regresiones semilogísticas de esas 100 muestras del regresando respecto a las 100 muestras simuladas del regresor medido con errores suponiendo $m = Dist * \sqrt{0.3}$.

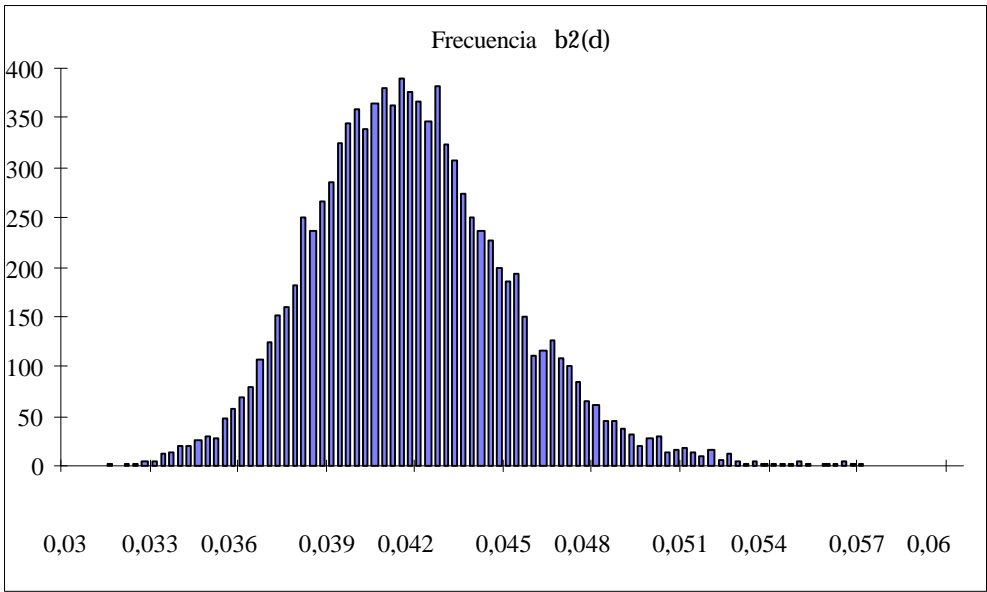


Gráfico 8

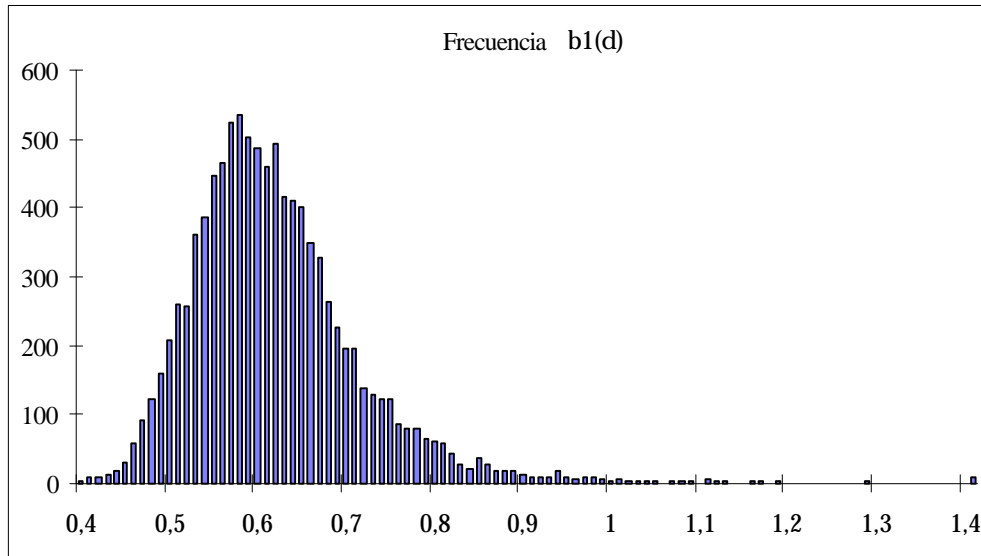


Gráfico 9

5.- Conclusiones.

De los resultados obtenidos, y que se ha tratado de resumir en los cuadros y gráficos anteriores, parece poder deducirse que si se dispone de información suficiente sobre todos los tramos relevantes de la función, las perturbaciones del modelo y los errores de medición tienen distribución Normal (i.i.d.), entonces la estimación de mínimos cuadrados proporciona estimadores sesgados. Es decir, el método de estimación MCO tiende a ajustar curvas cuyo punto de saturación estimado es menor que el verdadero punto de saturación, y cuya velocidad de crecimiento es mayor que la verdadera. Además, a medida que se incrementa la varianza de la perturbación el sesgo es mayor, al igual que sucede cuando se incrementa la varianza de los errores de medición, pero los incrementos en ambos tipos de varianzas provocan incrementos cuasi-lineales en el sesgo del estimador b_1 , e incrementos de tipo exponencial en el sesgo del estimador b_2 .

Teniendo en cuenta que hemos restringido los estimadores, obligándoles a tomar valores positivos, también se puede concluir que su distribución puede considerarse semejante a una normal, a pesar de que el contraste de Kolmogorov-Smirnov ofrezca como resultado el rechazo de la hipótesis de normalidad puesto que $K-S_{b1} = 11.49$ y $K-S_{b2} = 3.4^4$.

⁴ A partir de $K-S = 3.1$ la probabilidad de no rechazar la hipótesis de normalidad es cero.

BIBLIOGRAFIA

- Aigner, D. J.: "MSE Dominance of Least Squares with Errors-of-Observation", *Journal of Econometrics*, 2 (Dicembre-1974), 365-372.
- Bewley, R.; Fiebig, D.G.: "A flexible logistic growth model with applications in telecommunications", *International Journal of Forecasting*, 4, 177-192.
- Frisch, R.: *Statistical Confluence Analysis by Means of Complete Regression Systems*. Publ. n° 5, Oslo: University Institute of Economics, 1934, 192 pp.
- Gini, C.: "Sull'interpolazione de una retta quando i valori della variable indipendente sono affetti da errori accidentali", *Metroeconomica*, 63-82.
- Garber, S.; Klepper, S.: "Extending the Classical Normal Errors-in-Variables Model", *Econometrica*, 48 (Settembre-1980), 1541-1546.
- Kapteyn, A.; Wansbeek, T.: "Identification in the Linear Errors in Variables Model", *Econometrica*, 51 (Novembre-1983), 1847-1849.
- Klepper, S.; Leamer, E. E.: "Consistent Sets of Estimates for Regressions with Errors in All Variables", *Econometrica*, 52 (Enero-1984), 163-183.
- Reiersøl, O.: "Identifiability of a Linear Relation Between Variables which are Subject to Error", *Econometrica*, 18 (1950), 375-389.