7801

ANALISIS DE REGRESION: ALGUNAS CONSIDERACIONES UTILES PARA EL TRABAJO EMPIRICO

Vicent Poveda y Ricardo Sanz

INTRODUCCION

Este trabajo es un poco extraño. Al menos, eso parece a sus autores. Se trata de una expecie de cajón de sastre donde hemos recogido diversos aspectos, bastante independientes entre sí, relacionados con el análisis de regresión. Su posible unidad radica en que todos ellos son fruto de las dificultades que otros colegas, o nosotros mismos, hemos encontrado al trabajar prácticamente en la estimación de relaciones econométricas o intentar interpretar los resultados obtenidos.

Al pasar de lo que dicen los libros de texto y manuales al trabajo real, se plantean con frecuencia dificultades que sólo la experiencia continuada del trabajo empírico ayuda a salvar de un modo satisfactorio. Si esto es así, difícilmente pueden obviarse con nuevos textos y artículos. No obstante, a lo largo de varios años, hemos observado que algunas de estas dificultades se repiten en las personas que comienzan "a hacer regresiones" y surgen preguntas para las que a veces no existe una respuesta directa en los manuales.

Lo que hemos hecho, simplemente, es recopilar algunas de estas preguntas y estructurarlas a lo largo de seis secciones cuya extensión - y quizás su interés - es bastante desigual . Todos los aspectos tratados son muy elementales y se refieren exclusivamente a modelos uniecuacionales; aún dentro de ellos, no se abordan aspectos relativamente más sofisticados, como podrían ser las trampas que tiende con frecuencia la eliminación cuidadosa - o, simplemente, correcta - de la autocorrelación serial, o el viejo truco de "la variable dependiente desfasada", pongamos por caso.

Esto permite delimitar con bastante precisión el subconjunto de hipotéticos lectores para los que estas pági

nas pueden presentar un mínimo de interés. Se dirigen exclusivamente a los colegas - cada vez más numerosos, afortunadamente - que consideran importante un conocimiento cuantitativo de las relaciones que deben existir en nuestro país entre diferentes variables económicas y, como un primer paso, comienzan a estimar regresiones. Se supone que el lector conoce los aspectos más generales de la teoría del análisis de regresión, y es capaz de efectuar sencillas operaciones matriciales.

La primera sección se dedica a los efectos que produce en la estimación de un modelo lineal, las variaciones en las unidades de medida de las variables. En cierto modo, es una preparación a la sección 2, donde se aborda el mismo problema en un modelo lineal en los logaritmos de las variables. En ambos casos, hemos tratado de demostrar todas las afirmaciones que contiene el texto, y que no se encuentran en los manuales que conocemos. En las secciones 3 y 4 se abordan otros problemas relacionados con el mismo modelo logarítmico lineal: en la primera de ellas, la interpretación del error típico de la regresión y en la segunda se trata someramente el significado de algunos coeficientes estimados y los errores que pueden cometerse al utilizar en algunos casos las elasticidades calculadas. En la sección 5 se discute el delicado problema de la interpretación del coeficiente de determinación en modelos sin término constante, sobre el que no se detienen los libros de texto más generalmente conocidos. Por último, en la sección 6 se insiste en un problema que sí está perfectamente tratado en la literatura: los abusos en el uso del coeficiente de correlación como medida del poder predictivo de un modelo.

1.- CAMBIOS EN LA UNIDAD DE MEDIDA DE LAS VARIABLES EN EL MODELO LINEAL GENERAL

Definiremos el modelo con la notación usual:

$$Y = X \hat{b} + e \qquad (1.1)$$

donde Y es el vector Tx1 de observaciones de la variable dependiente, X es la matriz Tx(K+1) de observaciones sobre las variables explicativas, $\hat{B}$ es un vector (K+1)x1 de coeficientes de regresión y e un vector Tx1 de resíduos estimados.

Los cambios en la unidad de medida los expresaremos como el producto de una o varias variables por las constantes k_i, no necesariamente iguales para todas las variables, aunque constantes para toda la muestra (*). Finalmente utilizaremos un asterisco para distinguir las variables transformadas de las correspondientes al modelo (1.1). Matricialmente

$$Y^* = Y\ k$$
$$X^* = X\ Q$$

donde k es el escalar utilizado en la transformación de Y y Q es la matriz diagonal (K+1)x(K+1)

$$Q = \begin{bmatrix} 1 & & 0 \\ & k_1 & \\ & & \ddots \\ 0 & & k_k \end{bmatrix}$$

(*) Ejemplos típicos son el de modificación del año base en un modelo de series temporales en donde interviene un índice de precios o alguna de las variables está medida por índices, con lo que cada una de ellas se multiplicará por un factor k_i, generalmente distinto, o el de modificación en la unidad de medida - pesetas a dólares, miles a millones, etc - en el que todas las variables expresadas en estas unidades quedarán multiplicadas por un mismo factor k.

que contiene los distintos factores de corrección. En general, el primer elemento, que suponemos corresponde a la constante será igual a uno pudiéndolo ser también algunos de los restantes k_i.

Los cambios en las unidades de medida de las variables pueden agruparse en los tres casos que consideraremos a continuación. En cada uno de ellos se analizan los efectos que tienen sobre ciertos parámetros estimados.

1.1.- Las variables explicativas (o algunas de ellas) se multiplican por las constantes k_i.

Consecuencias:

1.1.1.- Los correspondientes coeficientes de regresión quedan divididos por k_i, a excepción del término constante que no varía (hemos supuesto que las observaciones de esta variable no se modifican).

El estimador mínimo cuadrático de $\hat{b}$ que aparece en (1.1) es:

$$\hat{b} = (X'X)^{-1} X'Y \qquad (1.2)$$

Si el modelo se transforma en

$$Y = X^*\hat{b}^* + e^* \qquad (1.3)$$

el nuevo estimador será

$$\hat{b}^* = (X^{*\prime}X^*)^{-1} X^{*\prime} Y = (Q'X'XQ)^{-1}Q'X' X'Y$$

y dado que la inversa de un producto de matrices cuadradas es el producto de sus inversas en sentido contrario

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

$$= Q^{-1}\,\hat{b}$$

y sabiendo que la inversa de una matriz diagonal es otra matriz diagonal cuyos elementos son los recíprocos de la matriz original

$$\begin{bmatrix} \hat{b}^*_0 \\ \hat{b}^*_1 \\ \vdots \\ \hat{b}^*_K \end{bmatrix} = \begin{bmatrix} 1 & & & 0 \\ & 1/k_1 & & \\ & & \ddots & \\ 0 & & & 1/k_K \end{bmatrix} \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_K \end{bmatrix} \qquad (1.4)$$

1.1.2.- Los resíduos no varían. En efecto,

$$\begin{aligned} e &= Y - X\hat{b} \\ e^* &= Y - X^*\hat{b}^* \\ &= Y - XQQ^{-1}\hat{b} \\ &= e \end{aligned} \qquad (1.5)$$

1.1.3.- Obviamente, tampoco variará el error típico de la regresión, se, definido por $\left[e'e/(T-K-1)\right]^{1/2}$.

1.1.4.- Las desviaciones típicas de los estimadores de los coeficientes de regresión se modifican en la misma proporción que éstos. Dado que éstas son iguales a los elementos de la diagonal principal de

$$\left[se^2 (X^{*\prime}X^*)^{-1}\right]^{1/2} = \left[se^2\, Q^{-1}(X'X)^{-1}Q'^{-1}\right]^{1/2}$$

es fácil ver, tras operar, que los elementos de la diagonal quedan divididos por k_i^2, lo que, tras extraer la raíz cuadrada, justifica la afirmación anterior.

1.2.- La variable dependiente se multiplica por k

Consecuencias:

1.2.1.- Todos los coeficientes de regresión, incluído el de la constante, quedan multiplicados por k.

$$\begin{aligned} \hat{b}^* &= (X'X)^{-1} X'Y^* \\ &= (X'X)^{-1} X'Yk \\ &= k\,\hat{b} \end{aligned} \qquad (1.6)$$

1.2.2.- Los resíduos quedan multiplicados por k.

$$\begin{aligned} e &= Y - X\hat{b} \\ e^* &= Y^* - X\hat{b}^* \\ &= kY - kX\hat{b} \\ &= ke \end{aligned} \qquad (1.7)$$

1.2.3.- El valor absoluto del error típico de la regresión queda multiplicado por k, pero su valor relativo (con respecto a $\bar{Y}^*$), forma en que se presenta frecuentemente, no se altera

$$se = \sqrt{\frac{e'\,e}{T-K-1}}$$

$$se^* = \sqrt{\frac{ke'\,ek}{T-K-1}}$$

$$= k\,se \qquad (1.8)$$

Y en términos relativos:

$$\frac{se^*}{\bar{Y}^*} = \frac{k\ se}{k\ \bar{Y}}$$

$$= \frac{se}{\bar{Y}} \qquad (1.9)$$

1.2.4.- Finalmente, el error típico de los estimadores de los coeficientes de regresión queda multiplicado por k al ser función lineal del error standard de la regresión.

1.3.- La variable dependiente se multiplica por k y las explicativas - exceptuada la constante - por k_i.

Consecuencias:

1.3.1.- Los coeficientes de regresión varían en una proporción igual al cociente del factor de corrección de la variable dependiente, k, y del correspondiente a la propia variable, k_i (en el caso de la constante, este último es 1).

$$\hat{b}^* = (X^{*\prime}X^*)^{-1}X^{*\prime}Y^*$$
$$= (Q'X'XQ)^{-1}Q'X'Y'k$$
$$= kQ^{-1}(X'X)^{-1}X'Y$$
$$= kQ^{-1}\hat{b}$$

Desarrollando:

$$\begin{bmatrix} \hat{b}^*_0 \\ \hat{b}^*_1 \\ \vdots \\ \hat{b}^*_K \end{bmatrix} = \begin{bmatrix} k & & & 0 \\ & k/k_1 & & \\ & & \ddots & \\ 0 & & & k/k_K \end{bmatrix} \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_K \end{bmatrix} \qquad (1.10)$$

Por ejemplo, en el caso de que la unidad de medida de todas las variables del modelo se modifique en la misma proporción, esto es, $\forall i: k=k_i$, el término constante queda modificado del mismo modo que las variables y los restantes coeficientes de regresión no varían.

1.3.2.- Los resíduos quedam multiplicados por k.

$$\begin{aligned} e^* &= Y^* - X^*\hat{b}^* \\ &= kY - XQkQ^{-1}\hat{b} \\ &= kY - kX\hat{b} \\ &= ke \end{aligned} \qquad (1.11)$$

1.3.3.- El valor absoluto del error típico de la regresión queda multiplicado por k, pero su valor relativo (con respecto a $\bar{Y}^*$) no varía (Véase 1.2.3.)

1.3.4.- Las desviaciones típicas de los coeficientes de regresión se modifican en igual proporción que éstos últimos. En efecto, como vimos en 1.1.1., éstos son iguales al producto del error típico de la regresión, que se multiplica por k, y los elementos de la diagonal principal de la matriz $\left[(X^{*\prime}X^*)^{-1}\right]^{1/2}$ que difieren de los del modelo sin transformar en el factor k_i.

****

En ninguno de los casos estudiados en esta sección se modifican los estadísticos t, el coeficiente de determinación R^2, ni el valor del test de Durbin Watson, D.W. El primero de ellos es igual a

$$t^* = \frac{\hat{b}^*_i}{se_{b^*_i}}$$

donde $se_{b_i^*}$ es el estimador de la desviación típica del coeficiente de regresión y hemos visto que el numerador y denominador de esta expresión, cuando varían, lo hacen en igual proporción. Por la misma razón, no se alteran los estadísticos R^{*2} y DW^*, definidos como

$$R^{*2} = 1 - \frac{e^{*\prime}e^*}{y^{*\prime}y^*}$$

donde $y^* = Y^* - \bar{Y}^*$,

y $$DW^* = \frac{\sum_t (e_t^* - e_{t-1}^*)^2}{\sum_t e_t^{*2}}$$

2.- CAMBIOS EN LA UNIDAD DE LAS VARIABLES EN EL MODELO LOGARITMICO LINEAL

El modelo original que consideramos en esta sección es

$$Y = e^{b_0} X_1^{b_1} \ldots\ldots\ldots X_K^{b_K} \varepsilon \qquad (2.1)$$

donde $X_1 \ldots\ldots X_K$ son los vectores de observaciones de las variables explicativas 1 a K, y ε es un término aleatorio de error. A efectos de estimación, el mismo se suele linealizar aplicando una transformación logarítmica que lo convierte en

$$\ln Y = b_0 + b_1 \ln X_1 + \ldots + b_K \ln X_K + \ln\varepsilon \qquad (2.1)'$$

que, para simplificar la notación, se presentará en lo sucesivo como

$$Y = \imath\hat{b}_0 + X_1\hat{b}_1 + X_2\hat{b}_2 + \ldots + X_K\hat{b}_K + e$$

$$= \imath\hat{b}_0 + X\hat{b} + e \qquad (2.2)$$

donde Y y X_i expresan directamente los vectores de logaritmos de las observaciones sobre las variables dependiente y explicativas, $\hat{b}_0$ es el coeficiente de regresión del término constante, $\imath$ es el vector unitario Tx1, X es la matriz TxK de logaritmos de las variables explicativas excluída la constante, $\hat{b}$ es el vector Kx1 de coeficientes de regresión de las X y e es el vector Tx1 de términos de error calculados.

Premultiplicando (2.2) por el factor $\frac{1}{T}\imath'$, donde T es un escalar, se obtiene, teniendo en cuenta que la media de e en la regresión con término constante es igual a cero:

$$\bar{Y} = \hat{b}_0 + \bar{X}\hat{b} \qquad (2.3)$$

donde $\bar{Y}$ es un escalar que representa la media de las observaciones sobre la variable dependiente, y $\bar{X}$ un vector 1xK formado por las medias de las K variables explicativas. Premultiplicando (2.3) por ι y sustrayéndolo de (2.2), obtenemos,

$$y = x\hat{b} + e \qquad (2.4)$$

donde las letras minúsculas representan desviaciones de cada variable con respecto a su media. Claramente, el modelo expresado de esta forma carecerá de constante, siendo y un vector Tx1 y x una matriz TxK. Al igual que hicimos en la sección anterior usaremos los asteriscos para expresar valores transformados de las variables o parámetros asociados a las nuevas relaciones. Así,

$$Y^* = \iota \ln k + Y$$
$$X_i^* = \ln k_i + X_i$$
etc,

donde k y k_i son los factores de corrección aplicados al modelo original (2.1).

Los cambios en la unidad de medida de las variables los agruparemos en los tres mismos casos de la sección precedente.

2.1.- Las variables explicativas, excluída la constante, se multiplican por una cantidad k_i (las variables del modelo en logaritmos se incrementan en $\ln k_i$).

Consecuencias:

2.1.1.- Los correspondientes coeficientes de regresión no varían, pero el término constante sí.

El estimador $\hat{b}$ del modelo (2.4) es

$$\hat{b} = (x'x)^{-1}x'y \tag{2.5}$$

y depués de transformarse las variables

$$\hat{b}^* = (x^{*\prime}x^*)^{-1}\, x^{*\prime}y \tag{2.6}$$

Para que $\hat{b}^*$ sea igual a $\hat{b}$ basta que x^* sea igual a x. Y puede verse fácilmente que esta última igualdad se satisface para cualquiera de las variables x_i en el momento t.

$$\begin{aligned} x^*_{it} &= X^*_{it} - \bar{X}^*_{it} \\ &= (X_{it}+\ln k_i) - \frac{1}{T} \sum_t (X_{it}+\ln k_i) \\ &= X_{it}+\ln k_i - \frac{1}{T} \sum_t X_{it} - \frac{1}{T} T \ln k_i \\ &= X_{it} - \bar{X}_i \\ &= x_{it} \end{aligned} \tag{2.7}$$

con lo que queda demostrada la igualdad de $\hat{b}$ y $\hat{b}^*$.

El término constante, sin embargo, sí variará. Del modelo (2.3) se deduce, después de transformar las variables,

$$\begin{aligned} \hat{b}^* &= \bar{Y} - \bar{X}^*\hat{b}^* \\ &= \bar{Y} - (\bar{X} + q')\hat{b} \end{aligned}$$

donde $q=[\ln k_i]$ es el vector formado con los logaritmos de los factores de transformación k_i

$$\begin{aligned} &= \bar{Y} - \bar{X}\hat{b} - q'\hat{b} \\ &= \hat{b}_0 - q'\hat{b} \end{aligned}$$

es decir,

$$\hat{b}_0^* = \hat{b}_0 - \sum_i^K \ln k_i \, \hat{b}_i \qquad (2.8)$$

o lo que es lo mismo, el término constante queda modificado en una cantidad igual a la suma ponderada de los coeficientes de regresión de las variables en que se ha modificado la unidad de medida, siendo las ponderaciones los logaritmos del factor de modificación de cada una de ellas. Obsérvese que si no se cambia la unidad de medida de una variable, el correspondiente valor de k_i sería la unidad, y su logaritmo cero, por lo que su coeficiente de regresión no figuraría en el sumatorio de (2.8).

2.1.2.- Los resíduos no se modifican

Por (2.2),

$$\begin{aligned} e &= Y - \imath\hat{b}_0 - X\hat{b} \\ e^* &= Y - \imath\hat{b}_0^* - X^*\hat{b}^* \\ &= Y - \imath(\hat{b}_0 - q'\hat{b}) - (X + \imath q')\hat{b} \\ &= Y - \imath\hat{b}_0 + \imath q'\hat{b} - X\hat{b} - \imath q'\hat{b} \\ &= e \end{aligned} \qquad (2.9)$$

2.1.3.- Como consecuencia de 2.1.2. el error típico de la regresión no variará en valor absoluto. Sin embargo, su expresión en términos relativos (con respecto a $\bar{Y}$) carece de significación clara, como se pondrá de manifiesto en el epígrafe 2.2.3.

2.2.- La variable dependiente se multiplica por una constante (la variable en el modelo en logaritmos se incrementa en lnk)

Consecuencias:

2.2.1.- Todos los coeficientes de regresión permanecen invariables excepto el término constante.

Partiendo del estimador $\hat{b}$ del modelo (2.4) (con Y transformada) en el que ha desaparecido el término constante:

$$\begin{aligned}\hat{b}^* &= (x'x)^{-1}x'\ y^* \\ &= (x'x)^{-1}x'(y + \imath\ln k) \\ &= (x'x)^{-1}x'y + (x'x)^{-1}x'\ \imath\ln k \\ &= \hat{b} + (x'x)^{-1}x'\imath\ln k\end{aligned}$$

Para que $\hat{b}^*=\hat{b}$ es necesario que el segundo término sea nulo. Para ello es suficiente que el producto

$$x'\ \imath\ln k$$

tenga todos sus elementos nulos. El elemento típico del vector es

$$\begin{aligned}\sum_t x_{it}\ln k &= \sum_t (x_{it} - \frac{1}{T}\sum_t x_{it})\ln k \\ &= \sum_t x_{it}\ln k - \sum_t \frac{1}{T}\ t\ x_{it}\ln k \\ &= 0\end{aligned}$$

Por consiguiente

$$\hat{b}^* = \hat{b}. \qquad (2.10)$$

Pero el término constante sí variará:

$$\begin{aligned}\hat{b}_0^* &= \bar{Y}^* - \bar{X}\ \hat{b}^* \\ &= \bar{Y} + \ln k - \bar{X}\ \hat{b} \\ &= \hat{b}_0 + \ln k \qquad (2.11)\end{aligned}$$

2.2.2.- Los resíduos no varían

Por (2.2),

$$e = Y - \imath\hat{b}_0 - X\hat{b}$$

$$\begin{aligned} e^* &= Y^* - \imath\hat{b}_0^* - X\hat{b}^* \\ &= Y + \imath lnk - \imath\hat{b}_0 - \imath lnk - X\hat{b} \\ &= e \end{aligned} \qquad (2.12)$$

2.2.3.- Puesto que los resíduos no varían, el error típico de la regresión tampoco se verá afectado por las transformaciones de Y. En cambio, es evidente que la presentación en términos relativos no tiene ninguna significación en una regresión sobre logaritmos de las variables, ya que la supuesta medida del error dependerá de la unidad de medida, necesariamente arbitraria, elegida para expresar las observaciones de Y.

En efecto

$$\frac{se}{\bar{Y}^*} = \frac{se}{\bar{Y} + lnk} \neq \frac{se}{\bar{Y}} \qquad (2.13)$$

Volveremos sobre otros aspectos de este problema en la sección 3.

2.3.- Todas las variables del modelo-dependiente e independientes- se multiplican por constantes no necesariamente iguales (las variables del modelo en logaritmos se incrementan en k_i)

Consecuencias:

2.3.1.- Todos los coeficientes de regresión permanecen invariables excepto el del término constante

$$\begin{aligned} \hat{b}^* &= (x^{*\prime}x^*)^{-1}x^{*\prime}y^* \\ &= (x'x)^{-1}x'y, \quad \text{por (2.7)} \\ &= \hat{b} \end{aligned} \qquad (2.14)$$

Por el contrario, el término constante sí varía

$$\hat{b}_0 = \bar{Y} - \bar{X}\,\hat{b}$$

$$\hat{b}^*_0 = \bar{Y}^* - \bar{X}^*\hat{b}^*$$

$$= \bar{Y} + \ln k - (\bar{X} + q')\hat{b}$$

donde q es el vector de los logaritmos de los factores de transformación definido en 2.1.1.

$$= \bar{Y} - \bar{X}\hat{b} + \ln k - q'\hat{b}$$

$$= \hat{b}_0 + \ln k - q'\hat{b}$$

es decir,

$$\hat{b}^*_0 = \hat{b}_0 + \ln k - \sum_{i=1}^{K} \hat{b}_i \ln k_i \qquad (2.15)$$

2.3.2. Los resíduos no varían

$$e^* = Y^* - \imath\hat{b}^*_0 - X^*\hat{b}^*$$

$$= Y + \imath\ln k - \imath\hat{b}_0 - \imath\ln k + \imath q'\hat{b} - (X + \imath q')\hat{b}$$

$$= Y - \imath\hat{b}_0 - X\hat{b}$$

$$= e \qquad (2.16)$$

2.3.3.- En cuanto al error típico de la regresión se aplica totalmente lo dicho en el epígrafe 2.2.3..

****** ******

Es fácil ver que en ninguno de los casos tratados en esta sección, con una sola excepción, se modifican los erro-

Resumen de las modificaciones en algunos estimadores

	modelo lineal			modelo logarítmico lineal		
	$X_i^* = k_i X_i$	$Y^* = kY$	$Y^* = kY$ $X_i^* = k_i X_i$	$\ln X_i^* = \ln(X_i k_i)$	$\ln Y^* = \ln(kY)$	$\ln Y^* = \ln(kY)$ $\ln X_i^* = \ln(X_i k_i)$
$\hat{b}_0^*$	$\hat{b}_0$	$k\,\hat{b}_0$	$k\,\hat{b}_0$	$\hat{b}_0 - \Sigma \hat{b}_i \ln k_i$	$\hat{b}_0 + \ln k$	$\hat{b}_0 + \ln k - \Sigma \hat{b}_i \ln k_i$
$\hat{b}_i^*$	$\hat{b}_i / k_i$	$k\,\hat{b}_i$	$(k/k_i)\hat{b}_i$	$\hat{b}_i$	$\hat{b}_i$	$\hat{b}_i$
e^*	e	$k\,e$	$k\,e$	e	e	e
$se^*_{b_0}$	se_{b_0}	$k\,se_{b_0}$	$k\,se_{b_0}$	-- (a)	-- (a)	-- (a)
$se^*_{b_i}$	se_{b_i}/ki	$k\,se_{b_i}$	$(k/k_i)\,se_{b_i}$	se_{b_i}	se_{b_i}	se_{b_i}
se^*	se	$k\,se$	$k\,se$	se	se	se

Las variables y parámetros con asteriscos corresponden al modelo transformado

$\hat{b}_0$ = coeficiente de regresión del término constante.

$\hat{b}_i$ = coeficientes de regresión de las variables explicativas.

e = residuo calculado en la regresión.

se_{b_0} = desviación típica de b_0.

se_{b_i} = desviación típica de b_i.

se = desviación típica (error típico) de la regresión.

No se incluyen los valores de R^2, DW y estadísticos t por no variar en ninguno de los supuestos contemplados (excepto para se_{b_0} en el modelo logaritmo lineal. Véase la nota (a)).

(a) Las modificaciones son irrelevantes desde el punto de vista de la significación estadística de esta variable, de obligada inclusión. Ver texto.

res típicos de los coeficientes de regresión, estadísticos t, R^2 y DW ,por depender estos estadísticos del término de error estimado, varianzas y covarianzas de las variables y los propios coeficientes de regresión, que como hemos visto no varían. La excepción la constituye el estadístico t del coeficiente del término constante, variable de uso obligado en las regresiones a logaritmos por recogerse en ella todas las modificaciones en la escala elegida para las restantes variables del modelo (ver Sección 4).

3.- EL SIGNIFICADO DEL ERROR TIPICO EN LAS REGRESIONES EN LOGARITMOS

En lo que respecta al trabajo empírico, la implicación más importante que se desprende de la sección 2 es que el error típico de la regresión sobre logaritmos de las variables pierde toda su significación cuando se le presenta en la forma habitual de porcentaje sobre la media de las observaciones de la variable a explicar. A pesar de ello, en la presentación de resultados incluída en numerosos trabajos, se encuentra con cierta frecuencia dicho parámetro. En realidad, al "investigador" deseoso de mostrar un error relativo bajo en sus estimaciones, le basta con escoger una unidad de medida baja para su variable dependiente - millones de pesetas en lugar de miles de millones o miles en lugar de millones - para reducir artificialmente el aparente error medio cometido en su estimación.

Con todo ello, esta dificultad pierde buena parte de su importancia cuando se comprende que, en general, la información relevante no es el error cometido al explicar la evolución del logaritmo de una variable sino el cometido al explicar la evolución de la propia variable. El modelo realmente propuesto es el recogido en (2.1) y es claro que el recurso a los

logaritmos no es más que una solución de facilidad para poder abordar la estimación con las ventajas obvias de los métodos lineales. Una vez obtenidos los coeficientes de regresión, las medidas globales sobre la calidad del ajuste hay que juzgarlas desde el modelo original (2.1).

De un modo más concreto, el coeficiente de determinación R^2 obtenido en la regresión sobre logaritmos no contiene, en general, una información válida sobre el "porcentaje explicado de la varianza total a explicar" del modelo que nos interesa, ya que está calculado sobre la varianza del logaritmo de la variable, y no sobre la varianza de la propia variable. Y la transformación logarítmica modifica dicha varianza.

En cuanto al error típico de la regresión no todo está perdido, ya que el mismo puede interpretarse como una medida aproximada del error relativo medio - es decir, expresado como porcentaje de la variable en niveles - que obtendríamos si calculásemos el antilogaritmo de los valores estimados y los comparásemos con los valores realmente observados. Si llamamos e_t^N al error así obtenido y e_t^L al resíduo estimado en la regresión en logaritmos, tenemos que:

$$e_t^L = -\left|\ln \hat{Y}_t - \ln Y_t\right|$$

$$= -\ln \frac{\hat{Y}_t}{Y_t}$$

$$= -\ln \left|1 - \frac{Y_t - \hat{Y}_t}{Y_t}\right|$$

que para valores pequeños (*) del cociente incluído dentro del paréntesis puede aproximarse por

(*) Una medida de lo que significa pequeño en este contexto nos lo da el hecho de que a un valor de 0,10 del cociente- es decir, de la tasa de error - corresponde un valor del logaritmode 0,105.

$$\frac{Y_t - \hat{Y}_t}{Y_t} = \frac{e_t^N}{Y_t} \qquad (3.1)$$

Es decir, el error absoluto cometido en cada período en la regresión en logarmitmos aproxima el error relativo sobre los valores originales de la variable en el período correspondiente. En consecuencia, el error típico de la regresión en logaritmos tiene el mismo significado que el obtenido al estimar una ecuación en niveles si expresamos éste último como porcentaje de la media de las observaciones de la variable dependiente.

En efecto, en la ecuación en logaritmos el error típico es:

$$se^L = \sqrt{\frac{\sum_t e_t^{L2}}{T - K - 1}} = \sqrt{\frac{1}{T - K - 1} \sum_t \left(\frac{e_t^N}{Y_t}\right)^2} \qquad (3.2)$$

y en la ecuación en niveles el error dividido por la media de la variable dependiente será

$$\frac{se}{\bar{Y}} = \frac{1}{Y} \sqrt{\frac{\sum_t e_t^2}{T - K - 1}} = \sqrt{\frac{1}{T-K-1} \sum_t \left(\frac{e_t}{\bar{Y}}\right)^2} \qquad (3.3)$$

expresiones que sólo se diferencian en el divisor del término que aparece entre paréntesis que es igual al valor correspondiente de la variable dependiente, ya expresada en niveles en el primer caso y a la media de todos las observaciones de esta variable en el segundo. Nótese, sin embargo, que esta comparación tiene sentido únicamente a efectos de ilustrar el signifi

cado del error típico de ambos tipos de modelos. En general, si se estima una misma ecuación en niveles y logaritmos

$$e_t \neq e_t^N$$

por responder ambos a especificaciones distintas, no siendo (3.2) y (3.3) los estadísticos más adecuados para la elección de la mejor especificación.

No obstante, existe al menos un tipo de modelos en logaritmos en el que la comparación directa del error típico con los valores medios de la variable dependiente tal como se hace en (3.3) es relevante. Son aquellos en los que la variable dependiente es igual al logaritmo del cociente - o a la diferencia de logaritmos - de una variable en niveles y la misma con un desfase. Es obvio que en este caso los cambios en la unidad de medida no afectarán el valor numérico de la variable dependiente y dado que el valor de ésta es aproximadamente igual a la tasa de variación de la variable en niveles, los errores típicos de la regresión recobran el significado que tenínan en el modelo original.

4.- LOS COEFICIENTES DE REGRESION EN LOS MODELOS EN LOGARITMOS. SU INTERPRETACION Y USO CON FINES PREDICTIVOS.

Un modelo estimado con relativa frecuencia sobre series cronológicas - junto con la versión lineal en las variables - es el siguiente:

$$Y_t = e^{b_0} e^{b_1 T} Y_{1t}^{b_2} X_{2t}^{b_3} \ldots\ldots \qquad (4.1)$$

donde T trata de captar una tendencia, y las variables X_{1t}, etc representan las observaciones de las restantes variables explicativas.

Corriendo el riesgo de insistir nuevamente en conceptos ya sabidos, hay que recalcar el papel crucial que juega la constante, e^{b_0}, en este tipo de modelos. Como vimos en la sección 2, es un factor de escala que hace compatibles las unidades de medida de todas las variables que intervienen en él. Es obvio que, en ausencia de este término, dado el carácter multiplicador del modelo (4.1), al incrementar una variable en la proporción k_i, el lado derecho de la ecuación aumentaría en la proporción $k_i^{\beta_i}$. Como las unidades de medida son necesariamente arbitarias - y además las distintas variables de un modelo representan frecuentemente magnitudes heterogéneas no comparables entre sí - la ausencia de la constante haría que los resultados obtenidos por cada investigador dependieran de la unidad de medida elegida. En consecuencia, la presencia de este término es necesaria, cualquiera que sea el valor de su desviación típica y, consiguientemente, del estadístico t, y no es posible dar una interpretación económica al valor obtenido en la estimación.

Es sabido que tomando logaritmos el modelo queda linealizado en la forma:

$$\ln Y_t = b_0 + b_1 t + \ln X_{1T} + \ldots + e_t \qquad (4.2)$$

El estimador b_1 es simplemente

$$\frac{d \ln Y_t}{d T} = \hat{b}_1$$

esto es, la cantidad en que aumenta el logaritmo de Y_t por unidad de tiempo o, lo que es lo mismo, una aproximación generalmente válida (*), de la tasa de variación de Y_t por unidad de tiempo. A $\hat{b}_1$ se le llama frecuentemente la "tasa de crecimiento autónomo" de la variable Y_t y, claramente, recoge la influencia de las variables no recogidas explícitamente en el modelo (**).

Por el contrario, el coeficiente $\hat{b}_2$ - y análogamente $\hat{b}_3, \ldots \hat{b}_K$ - es

$$\frac{d \ln Y_t}{d \ln X_{1t}} = \frac{d Y_t/Y_t}{d X_{1t}/X_{1t}} = \hat{b}_2$$

esto es, la relación entre las tasas de crecimiento instantáneas de Y_y y X_{1t}, o sea, la elasticidad de Y_t con respecto a X_{1t}.

Como en el caso del coeficiente $\hat{b}_1$, la medida de la elasticidad así obtenida aproxima peor la relación entre las

---

(*) La aproximación es más exacta cuanto menor es la tasa de variación, y el error tiende a cero cuando la variación de Y tiende a cero.

(**) De un modo más preciso, habría que decir: la influencia de las variables no recogidas explícitamente en el modelo, dadas las variables incluídas. Es obvio que la estimación de la influencia de la tendencia dependerá de la especificación del modelo - como sucede con cualquier otra variable que no sea ortogonal con el resto de variables incluídas - por lo que hay que ser particularmente cuidadosos al pronunciarse sobre "las tasas de variación autónoma" de una variable o deducir conclusiones apoyadas en su estimación. En recientes trabajos sobre la evolución del paro en España, por ejemplo, no se dedica suficiente importancia a esta elemental salvedad y se aventuran conclusiones más que dudosas.

tasas de variación de las variables cuanto mayor sea la tasa de variación de la variable en cuestión, en este caso X_{1t}. Este hecho puede plantear problemas en el trabajo empírico cuando $\hat{b}_2$ se utiliza para hacer predicciones de la variable dependiente fuera de la muestra. El razonamiento que suele seguirse en estos casos es: si la variable X_{1t} aumenta en x por ciento, Y_t aumentará en $x\hat{b}_2$ por ciento. Implícitamente, se está utilizando el modelo

$$\dot{Y}_t = \hat{b}_1 + \hat{b}_2 \dot{X}_{1t} + \ldots\ldots \qquad (4.3)$$

donde el punto sobre una variable indica que la misma se expresa en tasas de variación. Este modelo es una aproximación de (4.1)-(4.2), que difiere más de aquel cuanto mayor sea $\dot{X}_{1t}$.

Los errores en que se puede incurrir quedan ilustrados en el ejemplo siguiente. Supongamos, por razones de simplicidad, que el modelo es

$$Y_t = 0{,}01\, X_t^2 + e_t \qquad (4.4)$$

y en el último período de la muestra utilizada en la regresión la variable X_t toma el valor $X_T = 100$ y el resíduos es nulo. Se desea efectuar una predicción de Y_{T+1}, dado un valor de X_{T+1}.

El razonamiento habitual, lleva a utilizar el modelo

$$\dot{Y}_t = 2\,\dot{X}_t \qquad (4.5)$$

En el cuadro siguiente se incluyen las predicciones sobre Y_{T+1} bajo dos supuestos distintos en cuanto a la evolución de X_{T+1} y utilizando en ambos casos el modelo (4.4) y la aproximación (4.5).

Predicciones de Y_{T+1} condicionalmente a X_{T+1}

período	X_t	$\dot{X}_t$	Y_t	
T	100	--	100	
			modelo (4.4)	modelo (4.5)
T+1	105	5 %	110,2	110,0
T+1	200	100 %	400,0	300,0

Vemos que si la tasa de variación de X_t es del 5 por ciento, la diferencia existente entre las dos predicciones de Y_t es relativamente débil, pero cuando la tasa de variación de X_t es del 100 por cien, el error cometido al utilizar el modelo (4.5) es muy importante.

Como hemos visto en el ejemplo anterior, la sustitución del modelo (4.2) por el (4.3) supuesto se haya utilizado el primero para la estimación de los coeficientes de regresión, introducirá errores en los resultados obtenidos, errores que aunque normalmente son lo suficientemente pequeños como para justificar esta sustitución, pueden ser graves en el caso de variaciones importantes de las variables explicativas. En todo caso, sería conveniente indicar en los trabajos publicados el método de cálculo usado para evitar posibles confusiones al analista que desea contrastar los resultados.

5.- R^2 EN REGRESIONES SIN TERMINO CONSTANTE

El origen de esta sección se encuentra en la sorpresa sufrida recientemente por un economista al recibir los resultados de la estimación de una regresión y encontrar que el ordenador le presentaba un valor negativo del coeficiente de determinación, R^2.

Lo que sigue es una exposición detallada que trata de justificar semejante atrevimiento por parte del ordenador. Su contenido, obviamente, sólo interesará a los colegas que tiendan a pensar que la culpa debe ser del programa utilizado, porque " como todo el mundo sabe, R^2 está comprendido entre cero y la unidad", como decía nuestro amigo.

Comenzaremos partiendo de un modelo de regresión simple del que deduciremos ciertas propiedades generalmente conocidas y que serán de gran utilidad posteriormente. El modelo, en su forma más sencilla, y sin aplicarle restricción alguna, lo escribiremos en la forma usual:

$$Y_t = b_0 + b_1 X_t + \varepsilon_t \qquad (t=1,....,T) \qquad (5.1)$$

Al minimizar la suma de los cuadrados de los resíduos con respecto a $\hat{b}_0$ y $\hat{b}_1$, se obtienen las conocidas ecuaciones normales:

$$\sum_t Y_t = T\hat{b}_0 + \hat{b}_1 \sum_t X_t \qquad (5.2)$$

$$\sum_t X_t Y_t = \hat{b}_0 \sum_t X_t + \hat{b}_1 \sum_t X_t^2 \qquad (5.3)$$

de las que se deducen una serie de propiedades que se satisfarán en cualquier muestra.

En primer lugar (*):

$$e_t = \Sigma(Y_t - \hat{Y}_t)$$
$$= \Sigma(Y_t - \hat{b}_0 - \hat{b}_1 X_t)$$
$$= \Sigma Y_t - T\hat{b}_0 - \hat{b}_1 \Sigma X_t$$
$$= 0 , \quad \text{por (5.2)} \tag{5.4}$$

Y también:

$$\Sigma X_t e_t = \Sigma X_t(Y_t - \hat{b}_0 - \hat{b}_1 X_t)$$
$$= \Sigma X_t Y_t - \hat{b}_0 \Sigma X_t - \hat{b}_1 \Sigma X_t^2$$
$$= 0, \text{ por (5.3)} \tag{5.5}$$

Además, la suma de los cuadrados de los residuos puede expresarse como:

$$\Sigma e_t^2 = \Sigma(Y_t - \hat{Y}_t)^2$$
$$= \Sigma Y_t^2 + \Sigma \hat{Y}_t^2 - 2\Sigma Y_t \hat{Y}_t , \tag{5.6}$$

donde el último término puede transformarse en:

$$2\Sigma Y_t \hat{Y}_t = 2\Sigma(\hat{Y}_t + e_t)\hat{Y}_t$$
$$= 2\Sigma \hat{Y}_t^2 + 2\Sigma \hat{Y}_t e_t$$
$$= 2\Sigma \hat{Y}_t^2 \tag{5.7}$$

puesto que

$$\Sigma \hat{Y}_t e_t = \Sigma(\hat{b}_0 + \hat{b}_1 X_t)e_t$$
$$= \hat{b}_0 \Sigma e_t + \hat{b}_1 \Sigma X_t e_t$$
$$= 0, \text{ por (5.4) y (5.5)}$$

(*) Puesto que todos los sumatorios que siguen es evidente que son sobre el tiempo,t, omitimos el índice para simplificar la notación.

Introduciendo (5.7) en (5.6) se llega a un importante resultado:

$$\Sigma Y_t^2 = \Sigma \hat{Y}_t^2 + e_t^2 \qquad (5.8)$$

Por otra parte, la varianza de una serie puede expresarse como

$$\Sigma (Y_t - \bar{Y})^2 = \Sigma Y_t^2 - \frac{1}{T} (\Sigma Y_t)^2$$

y

$$\Sigma (\hat{Y}_t - \bar{\hat{Y}})^2 = \Sigma \hat{Y}_t^2 - \frac{1}{T} (\Sigma \hat{Y}_t)^2$$

lo que nos será de gran utilidad para transformar (5.8). En efecto, de la primera ecuación normal, (5.2), se deduce que $\Sigma e_t = 0$, por lo que $\Sigma Y_t = \Sigma \hat{Y}_t$. Sustrayendo $1/T(\Sigma Y_t)^2$ de los dos miembros de (5.8):

$$\Sigma Y_t^2 - \frac{1}{T} (\Sigma Y_t)^2 = \Sigma \hat{Y}_t^2 - \frac{1}{T} (\Sigma Y_t)^2 + \Sigma e_t^2$$

que expresaremos de la forma:

$$\Sigma y_t^2 = \Sigma \hat{y}_t^2 + \Sigma e_t^2 \qquad (5.9)$$

donde las letras minúsculas expresan desviaciones con respecto a la media. La expresión (5.9) es la conocida descomposición de la varianza total "a explicar", en dos partes: la varianza "explicada" y la varianza residual.

El R^2 suele recibir la interpretación de "proporción explicada de la varianza total a explicar", esto es:

$$R^2 = \frac{\Sigma \hat{y}_t^2}{\Sigma y_t^2} \qquad (5.10)$$

o, lo que es lo mismo, por (5.9)

$$R^2 = 1 - \frac{\Sigma e_t^2}{\Sigma y_t^2} \qquad (5.10)'$$

Es bien sabido, que en caso de un ajuste perfecto, $\hat{Y}_t = Y_t$, por lo que $R^2=1$. En el peor de los casos, $\Sigma e_t^2 = \Sigma y_t^2$, la recta ajustada será igual a $\bar{Y}$, y la varianza explicada nula, con lo que $R^2=0$.

Es muy frecuence que los programas de regresión calculen R^2 a partir de esta última fórmula, (5.10)'.

La situación se complica cuando al estimar el modelo (5.1) se impone la restricción $b_0=0$, reduciéndolo a

$$Y_t = b_1^* X_t + \varepsilon_t^* \qquad (5.11)$$

En efecto, veamos el importante papel que jugó el término constante ahora omitido para llegar a las dos definiciones alternativas de R^2. Estas se obtuvieron de la descomposición de la varianza a explicar, (5.9), que a su vez surge de tomar desviaciones con respecto a la media en (5.8). Si nos centramos en esta última expresión, vemos que procede de manipular la suma de los cuadrados de los resíduos desde (5.6); al único resultado precedente que tenemos que recurrir necesariamente en este proceso es a la segunda ecuación normal, (5.5), que nos garantiza que $\Sigma X_t e_t = 0$ (*). Esta propiedad surge de la derivada parcial de Σe_t^2 con respecto a $\hat{b}_1$ por lo que subsiste en el caso de modelo sin término constante. En consecuencia, (5.8) es válido, independientemente de que la regresión contenga, o no una constante.

(*) En el texto anterior se ha utilizado también (5.4) porque el modelo incluía un término constante. Claramente, si $b_0=0$, llegaríamos a (5.7) sin que apareciera el término $\hat{b}_0 \Sigma e_t$.

El paso de (5.8) a (5.9), en cambio, es posible exclusivamente porque $\Sigma e_t=0$, propiedad que se deduce de la primera ecuación normal, que no existe en el caso de un modelo sin término constante. Así, en el modelo (5.11), en general, $\Sigma e_t^* \neq 0$ y $\Sigma Y_t \neq \Sigma \hat{Y}_t$, por lo que:

$$\Sigma y_t^2 \neq \Sigma \hat{y}_t^2 + \Sigma e_t^{*2} \qquad (5.12)$$

La descomposición de la varianza ya no es posible, y R^{*2} no puede recibir la interpretación usual de "proporción explicada". Además, debido a (5.12), las expresiones (5.10) y (5.10)' ya no son equivalente, ni sus valores numéricos estarán comprendidos en el intervalo 0,1. El cálculo de R^2 a partir de (5.10) podrá exceder de la unidad porque la varianza de $\hat{Y}_t^*$ puede ser superior a la varianza de Y_t. Por el contrario, el cálculo a partir de (5.10)' puede conducir a valores negativos, porque la varianza de los resíduos puede exceder a la varianza de Y_t.

Esto puede ilustrarse con un sencillo ejemplo, en el que el modelo (5.11) se ha estimado para los pares de observaciones siguientes: (4,2), (5,6) y (6,10). La fórmula de R^2 incorporada al programa utilizado es (5.10)' que, en este caso, conduce a:

$$R^{*2} = 1 - \frac{\Sigma e_t^{*2}}{\Sigma y_t^2} = -3.077,$$

mientras que, utilizando (5.10), se obtendría:

$$R^{*2} = \frac{\Sigma \hat{y}_t^{*2}}{\Sigma y_t^2} = 8.488$$

Este ejemplo, claramente, es un caso particular y límite. La representación gráfica de los tres puntos observados mostraría que el investigador debía incluir un término constante que tendría que presentar un signo positivo y, en esas condiciones, obtendría un excelente ajuste global. Se trata solamente de mostrar incluso exageradamente, que, en efecto, los límites 0 y 1 de R^2 desaparecen con el término constante.

Un valor de R^2 que se sitúe fuera de estos límites evidencia - y en el ejemplo anterior es obvio - un error de especificación del modelo. En un caso más complejo, incluyendo varias variables explicativas, la representación gráfica del hiperplano a ajustar resulta imposible y la decisión de incluir o no, un término constante puede resultar más delicada. En general, el camino lógico a seguir en el trabajo empírico consistiría en no excluir la constante a priori. Pero es cierto, que frecuentemente, con los resultados en la mano, hay dudas razonables sobre si no sería preferible eliminarla. En estos casos, lo ideal sería poder seguir un criterio similar al que se adopta cuando esta misma duda se plantea - en un modelo con constante - a propósito de una variable explicativa, es decir, comparar los R^2 corregidos por los grados de libertad de las regresiones con y sin dicha variable. El problema, en el caso que nos ocupa, es claro que no puede abordarse de ese modo: en la regresión sin término constante, el ordenador presentará un valor de R^2 - generalmente, calculado a partir de (5.10)', como ya dijimos - pero el investigador sabe que, partiendo de (5.10) puede obtener otro valor distinto y que ninguno de los dos puede interpretarlos como "la proporción explicada de la varianza" ni por consiguiente, compararlo - previa corrección por los grados de libertad - al F^2 de la regresión con constante.

Si bien todo cuanto precede es estrictamente válido, se podría matizar la imposibilidad de la comparación se

ñalando que, en realidad, si el modelo sin término constante está bien especificado, las diferencias numéricas entre las dos estimaciones de R^2 no diferirán mucho entre sí y pueden aproximar razonablemente la idea de "proporción explicada". En cualquier caso, más que confiar excesivamente en el propio modelo, desde un punto de vista teórico, parece más interesante buscar una medida de R^2 que sea unívoca y aplicable a todos los modelos, contengan éstos o no, un término constante.

Esto es factible a partir de (5.8) que ya vimos es válido en todos los casos. Un R^2 definido como

$$R^2 = \frac{\Sigma \hat{Y}_t^2}{\Sigma Y_t^2} = 1 - \frac{\Sigma e_t^2}{\Sigma Y_t^2} \qquad (5.13)$$

y calculado a partir de los dos modelos alternativos permite decidir sobre la conveniencia de incluir, o no, una constante. La justificación de ésto se apoya en que dicha comparación es consistente con un test F sobre la hipótesis de que el término constante sea nulo, que puede efectuarse a partir del estadístico (*):

$$F_{1,T-K-1} = \frac{(T-K-1)(\Sigma e_t^{*2} - \Sigma e_t^2)}{\Sigma e_t^2}$$

El precio a pagar al utilizar (5.13) es que R^2 pierde toda su significación como "proporción explicada de la varianza a explicar", que sin duda presenta un gran interés desde el punto de vista de la estimación y predicción y que lógicamente interesa conservar. (5.13) puede tomarse simple-

(*) En general, en los manuales de econometría no se encuentra discusión alguna sobre el problema objeto de esta sección, y la deducción habitual de R^2 parece no estar sujeta a ningún tipo de limitación. Una excepción es el libro de Aigner, Basic Econometrics, Prentice-Hall, donde el lector interesado puede profundizar la relación entre el test F y el estimador (5.13) de R^2 en un epígrafe dedicado a este problema en el Capítulo 3.

mente, como una medida de "la bondad del ajuste"(*) perfectamente válida para tratar el problema que nos ocupa, y no como una alternativa superior al R^2 habitual.

La conclusión es que no existe una medida de R^2 que simultáneamente sea aplicable a todo tipo de modelos y estrictamente interpretable como "proporción explicada". En general, si la constante se elimina correctamente, el R^2 calculado puede aproximar de un modo razonable el concepto habitual que tiene en las regresiones con constante. El tratamiento correcto del problema de la eventual eliminación del término constante debe plantearse a partir de (5.13) y no del valor de R^2 ofrecido por el ordenador.

6.- EL COEFICIENTE DE CORRELACION Y LA CALIDAD DE LAS PREDICCIONES

El coeficiente de correlación lineal entre dos series es una de las medidas más utilizadas para juzgar sobre la "similitud" existente entre allas. Utilizamos deliberadamente un término tan poco preciso como "similitud" porque suponemos es la idea que tenía en mente el autor de un trabajo en el que se recurría a este parámetro para ofrecer un juicio global sobre el valor predictivo de un modelo, que había utilizado para predecir valores de la variable dependiente fuera de la muestra utilizada en la estimación.

El único objeto de esta sección es recordar una vez más que una alta correlación lineal entre dos series no significa necesariamente que una sea una buena aproximación de la otra, por lo que no puede utilizarse como una medida de la calidad de una predicción.

(*) Quizás algún lector piense que otra forma válidad de medir la bondad del ajuste podría consistir en calcular el coeficiente de correlación entre las series calculada y observada. Si este lector existe, quizás la sección 6 de este trabajo pueda interesarle.

En el cuadro adjunto se incluyen dos posibles predicciones alternativas, X^p_1 y X^p_2, sobre los valores de la variable X en tres períodos futuros, de los que ya se conocen las observaciones "reales" de dicha variable, que supondremos son 4, 5 y 6.

VALORES OBSERVADOS (X) Y PREDICCIONES (X^p_1, X^p_2) DE UNA VARIABLE

X	X^p_1	error %	X^p_2	error %
4	4.102	2.55	1.442	-63.95
5	4.837	-3.26	3.605	-27.90
6	6.061	1.02	7.209	20.15

Si en lugar de presentar los resultados de este modo, se resumen con el coeficiente de correlación calculado entre X y X^p_1, o entre X y X^p_2, podría pretenderse que ambas predicciones aproximan del mismo modo la evolución observada de la variable X, puesto que en ambos casos dicho coeficiente toma el valor de 0.9897.

Este elevado valor sólo indica que, en efecto, entre los dos pares de variables existe una fuerte relación del tipo

$$X_t = b_0 + b_1 X^p_{1t} \quad , \quad o$$

$$X_t = b_0 + b_1 X^p_{2t}$$

mientras que una buena predicción requeriría que $b_0=0$ y $b_1=1$.

Como medida global de la bondad de una predicción, Theil propuso utilizar su coeficiente de desigualdad, U, calculado como:

$$U = \frac{\sqrt{\frac{1}{T}\sum_t (X_t - X_t^p)^2}}{\sqrt{\frac{1}{T}\sum_t X_t^2} + \sqrt{\frac{1}{T}\sum_t X_t^{p2}}}$$

cuyos valores estarán comprendidos entre cero y la unidad. En caso de predicción exacta, $X_t = X_t^p$ paratodo t, por lo que U=0, mientras que U=1 indica el límite de predicciones equivocadas.

El lector interesado en su utilización, puede remitirse directamente a unas páginas de sencilla lectura del propio trabajo de Theil (*) donde se presentan diversas propiedades de dicho coeficiente.

(*) Véase, H. Theil, Economic Forecast and Policy, North-Holland, 1965, pp. 31 y siguientes.

DOCUMENTOS DE TRABAJO

7801 Vicent Poveda y Ricardo Sanz: Análisis de regresión: algunas consideraciones útiles para el trabajo empírico.

7802 Julio Rodríguez López: El PIB trimestral de España, 1958-1975. Avance de cifras y comentarios.

7803 Antoni Espasa: El paro registrado no agrícola 1964-1976: un ejercicio de análisis estadístico univariante de series económicas.

7804 Pedro Martínez Méndez y Raimundo Poveda Anadón: Propuestas para una reforma del sistema financiero.

7805 Gonzalo Gil: Política monetaria y sistema financiero. Respuestas al cuestionario de la CEE sobre el sistema financiero español.

7806 Ricardo Sanz: Modelización del índice de producción industrial y su relación con el consumo de energía eléctrica.

7807 Luis Angel Rojo y Gonzalo Gil: España y la CEE. Aspectos monetarios y financieros.

7901 Antoni Espasa: Modelos Arima univariantes, con análisis de intervención para las series de agregados monetarios (saldos medios mensuales) M_3 y M_2

7902 Ricardo Sanz: Comportamiento del público ante el efectivo.

7903 Nicolás Sánchez-Albornoz: Los precios del vino en España, 1861-1890. Volumen I: Crítica de la fuente.

7904 Nicolás Sánchez-Albornoz: Los precios del vino en España, 1861-1890. Volumen II: Series provinciales.

7905 Antoni Espasa: Un modelo diario para la serie de depósitos en la banca: primeros resultados y estimación de los efectos de las huelgas de febrero de 1979.

7906 Agustín Maravall: Sobre la identificación de series temporales multivariantes.

7907 Pedro Martínez Méndez: Los tipos de interés del mercado interbancario.

***Estas publicaciones —que, por su carácter especializado, son de tirada reducida— se distribuyen gratuitamente a las personas o entidades interesadas que las soliciten por correo.*