ESTADISTICA ESPAÑOLA
Núm. 90, 1981, págs. 21 a 28

En torno al cambio de medida de las variables que intervienen en la regresión lineal

por JOAN BARO LLINAS
Escuela Universitaria de Estudios Empresariales
Universidad de Barcelona

RESUMEN

Las transformaciones lineales en las variables que intervienen en los modelos uniecuacionales se justifican en la medida que tengan algún significado económico o por el simple hecho de facilitar los cálculos.

Se analiza el efecto de estos cambios de medida en la especificación de los modelos, en la estimación de los parámetros, en las pruebas de significabilidad y en los contrastes de hipótesis.

Palabras clave: Cambio de medida, transformaciones lineales, coeficientes de nueva estructura, sensibilidad de los contrastes.

I. INTRODUCCION

Es muy frecuente, ya sea para facilitar los cálculos operativos o para mejor aproximarnos a la realidad que queramos modelizar, que los valores de todas o de algunas de las variables que entran en juego en las relaciones funcionales, sean transformaciones de unos datos originales; buenos ejemplos de ello lo constituyen: las magnitudes deflactadas, los cambios de origen en el tiempo, el empleo de otra unidad monetaria, la utilización de magnitudes per cápita, el manejo de proporciones, pudiendo así continuar un largo etcétera de supuestos que justifican la importancia de este planteamiento.

El empleo más claro, sobre todo con la frecuencia que se emplea, es la presentación de los valores de las variables centrados respecto a la media, simplificando de tal modo las operaciones a efectuar; así, recogiendo uno de los ejemplos citados, podemos centrar una serie de datos en determinado año (t), que sera para nosotros el origen del tiempo, siendo la nueva variable:

$$T = t - \bar{t}$$

donde t es la serie primitiva de años, $\bar{t}$ puede ser la media aritmética y T la nueva sucesión de años.

O incluso para un número par de años podría plantearse

$$T = 2(t - \bar{t})$$

empleando así sólo valores enteros para la nueva variable.

Otro caso válido que no cabe encontrar entre los que se han mencionado, podría resultar de un cambio de sueldos en una empresa después de un convenio colectivo, que mejora lineal y proporcionalmente las remuneraciones antiguas. Podría escribirse

$$W' = k + q \cdot W$$

con q, tasa proporcional de aumento sobre los sueldos anteriores W, para determinar con k incremento lineal de nuevas percepciones W'.

Como ya se ha dicho, son muchos más los ejemplos a tener en cuenta, pero en cualquier caso vamos a ceñirnos a:

1) Funciones deterministas, evitando así relaciones estocásticas en la transformación (por ejemplo, $W' = k + qW$ + componente aleatoria), puesto que estos aspectos se encuadrarían mejor en otro contexto: el de los modelos multiecuacionales que ya están suficientemente desarrollados, al menos en la simplicidad que nos proponemos.

2) Y también a cambios lineales que al fin y al cabo son los que casi exclusivamente tienen sentido (¿qué interpretación cabría dar a $T = 2(t - \bar{t})^2$?).

Ciñiéndonos, pues, a la idea primitiva veamos hasta qué punto y, en su caso, de qué modo pueden alterarse los parámetros más característicos del modelo lineal.

II. EN TORNO AL MODELO

Sea en su forma original la relación

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

y sea la nueva ecuación con variables transformadas y, por su puesto, con nuevos parámetros

$$Y_i^* = \alpha^* + \beta_1^* X_{1i} + \beta_2^* X_{2i} + \ldots + \beta_k^* X_{ki}^* + \varepsilon_i^*$$

indicándonos las variables con asterisco (Z^*), cambios lineales en las originales (Z), siendo Z cualquier exógena o la misma endógena de modo que con

$$Z^* = \delta + \gamma Z$$

podrán establecerse las igualdades

$$\overline{Z}^* = \delta + \gamma \overline{Z}$$

$$\text{Var } (Z^*) = \gamma^2 \text{ var } (Z)$$

$$\text{DS } (Z^*) = \text{DS } (Z)$$

Siendo, además, igual el grado de dependencia lineal existente entre dos variables originales que la que presentan las dos nuevas variables resultantes de sendas combinaciones lineales de aquéllas; ello es evidente si tenemos en cuenta que el coeficiente de correlación de Pearson es insensible ante cambios de medida, al menos en valor absoluto.

Por otro lado, al margen de todo esto, fácilmente se intuye que la endógena nueva no pierde su carácter aleatorio y que, por supuesto, las ahora exógenas siguen siendo regresores fijos en la medida que lo fuesen los primitivos y de modo que la posible otorgonalidad, o quizá multicolinealidad, no puede haberse alterado por los simples cambios lineales. En definitiva, el modelo mantiene en relación a las variables todas sus características, queda, pues, ahora configurado el modo:

$$(\delta_0 + \gamma_0 Y_i) = \alpha_0^* + \beta_1^*(\delta_1 + \gamma_1 X_{1i}) + \beta_2^*(\delta_2 + \gamma_2 X_{2i}) + \ldots + \beta_k^*(\delta_k + \gamma_k X_{ki}) + \varepsilon_i^*$$

que para buscar semejanza con el modelo primitivo se transforma en

$$Y_i = \frac{\alpha_0^* + \beta_1^*\delta_1 + \beta_2^*\delta_2 + \ldots + \beta_k^*\delta_k - \delta_0}{\gamma_0} + \frac{\beta_1^*\gamma_1}{\gamma_0} X_{1i} + \ldots$$

$$\ldots + \frac{\beta_2^*\gamma_2}{\gamma_0} X_{2i} + \frac{\beta_k^*\gamma_k}{\gamma_0} X_{ki} + \frac{\varepsilon_i^*}{\gamma_0}$$

de modo que entre los coeficientes habrá de existir la relación:

$$\beta_k^* = \frac{\gamma_0}{\gamma_k}\beta_k \qquad \text{con } k = 1, 2, \ldots, k$$

$$\alpha_0^* = (\delta_0 + \gamma_0 \cdot \alpha) - \gamma_0 \sum_{k=1}^{k} \frac{\delta_k}{\gamma_k}\beta_k$$

y es por lo que la constancia de los coeficientes de la nueva estructura se mantendrá para todos los elementos muestrales en la medida que aceptemos esta hipótesis para la estructura original y que hayamos fijado los parametros del cambio.

Por fin, las perturbaciones no habrán alterado sus características iniciales, así como las premisas habituales para:

$$\varepsilon_i = \frac{\varepsilon_i^*}{\gamma_0}$$

puede afirmarse que también

$$E(\varepsilon_i^*) = 0 \qquad \text{el valor medio de la perturbación es 0}$$

$$E(\varepsilon_i^* \varepsilon_j^*) = \begin{cases} \gamma_0^2 \cdot \sigma_i^2 & \forall i = j \quad \text{homocedasticidad} \\ 0 & \forall i \neq j \quad \text{independencia} \end{cases}$$

$$\varepsilon_i^* = N(0,\ \gamma_0^2 \cdot \sigma_i^2) \qquad \text{normalidad}$$

Todo ello, por supuesto, aceptando el conjunto de hipótesis válido para la perturbación del modelo inicial.

III. EN TORNO A LA ESTIMACION

El modelo con variables transformadas puede presentarse en desviaciones del modo

$$y^* = \beta^* \cdot x^* + \varepsilon^*$$

$$\text{con } y^* = \begin{Vmatrix} y_i^* - \bar{y}^* \\ \vdots \end{Vmatrix} = \begin{Vmatrix} \gamma_0(y_i - \bar{y}) \\ \vdots \end{Vmatrix} = \gamma_0 \cdot y$$

$$x^* = \begin{Vmatrix} X_{1i}^* - \bar{X}_1^* & X_{2i}^* - \bar{X}_2^* & \vdots & X_{ki}^* - \bar{X}_k^* \\ \vdots & \vdots & \vdots & \vdots \end{Vmatrix} =$$

$$= \begin{Vmatrix} \gamma_1(X_{1i} - \bar{X}_1) & \gamma_2(X_{2i} - \bar{X}_2) & \vdots & \gamma_k(X_{ki} - \bar{X}_k) \\ \vdots & \vdots & \vdots & \vdots \end{Vmatrix} = x \cdot \gamma$$

$$\text{con } \gamma = \gamma' = \begin{Vmatrix} \gamma_1 & 0 & \dots & 0 \\ 0 & \gamma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \gamma_k \end{Vmatrix}$$

$$\varepsilon^* = \begin{Vmatrix} \gamma_0 \cdot \varepsilon_i \\ \vdots \end{Vmatrix} = \gamma_0 \cdot \varepsilon \qquad \beta^* = \begin{Vmatrix} \beta_1^* \\ \beta_2^* \\ \vdots \\ \beta_k^* \end{Vmatrix}$$

Nótese que al trabajar en desviaciones se ha excluido el término independiente, por lo que prescindiendo de él la estimación por MCO nos conducirá a

$$\begin{Vmatrix} \hat{\beta}_1^* \\ \hat{\beta}_2^* \\ \vdots \\ \hat{\beta}_k^* \end{Vmatrix} = \hat{\beta}^* = (x^{*\prime}x^*)^{-1}x^{*\prime}y^* = \gamma_0 \cdot \gamma^{-1}(x'x)^{-1}x'y = \gamma_0 \cdot \gamma^{-1}\hat{\beta} =$$

$$= \begin{Vmatrix} \frac{\gamma_0}{\gamma_1}\hat{\beta}_1 \\ \frac{\gamma_0}{\gamma_2}\hat{\beta}_2 \\ \vdots \\ \frac{\gamma_0}{\gamma_k}\hat{\beta}_k \end{Vmatrix}$$

mientras que para el término correspondiente tendremos

$$\hat{\alpha}^* = \overline{Y}^* - \sum_{r=1}^{k} \hat{\beta}_r^* \overline{X}_r^* = (\delta_0 + \gamma_0\hat{\alpha}) - \gamma_0 \sum_{r=1}^{k} \frac{\delta_r}{\gamma_r}\beta_r$$

relaciones tanto para los coeficientes de regresión como para el parametro independiente que son las que cabría esperar de acuerdo con las obtenidas a nivel poblacional.

La matriz de varianzas y covarianzas de los parámetros funcionales será ahora

$$\mathrm{Var}(\hat{\beta}^*) = \sigma_2^2(x^{*\prime}x^*)^{-1} = \gamma_0^2 \cdot \sigma_2^2(\gamma'x'x\gamma)^{-1} = \gamma_0^2\gamma^{-1}\,\mathrm{Var}(\hat{\beta})\gamma^{-1} =$$

$$= \begin{Vmatrix} \frac{\gamma_0^2}{\gamma_1^2}\mathrm{Var}(\hat{\beta}_1) & \frac{\gamma_0^2}{\gamma_1\gamma_2}\mathrm{Cov}(\hat{\beta}_1\hat{\beta}_2) & \ldots & \frac{\gamma_0^2}{\gamma_1\gamma_k}\mathrm{Cov}(\hat{\beta}_1\beta_k) \\ & \frac{\gamma_0^2}{\gamma^2}\mathrm{Var}(\hat{\beta}_2) & \ldots & \frac{\gamma_0^2}{\gamma_2\gamma_k}\mathrm{Cov}(\hat{\beta}_2\beta_k) \\ & & & \ldots\ldots\ldots\ldots \\ & & & \frac{\gamma_0^2}{\gamma_2^2}\mathrm{Var}(\hat{\beta}_2) \end{Vmatrix}$$

Resultados lógicos a la vista de la relación existente entre los parámetros.

La estimación de la varianza del término de perturbación habrá variado en proporción al cuadrado del parametro funcional en el cambio lineal de la endógena, ya que

$$\hat{\sigma}_t^{2*} = \frac{\hat{\varepsilon}^{*\prime}\hat{\varepsilon}^*}{N - (k+1)} = \frac{y^{*\prime}y^* - \hat{\beta}^{*\prime}x^{*\prime}y^*}{N - (k+1)} = \frac{\gamma_0^2 y'y - \gamma_0\hat{\beta}'\gamma^{-1}\gamma x'y + \gamma_0 \cdot y}{N - (k+1)} =$$

$$= \gamma_0^2 \frac{y'y - \hat{\beta}'x'y}{N - (k+1)} = \gamma_0^2 \cdot \hat{\sigma}_t^2$$

y por fin, el coeficiente de determinación no se habrá alterado

$$R^{*2} = \frac{\text{Var}(\hat{y}^*)}{\text{Var}(y^*)} = \frac{\hat{\beta}^{*\prime}x^{*\prime}y^*}{y^{*\prime}y^*} = \frac{\gamma_0^2\hat{\beta}'\gamma^{-1}\gamma x'y}{\gamma_0^2 * y'y} = \frac{\hat{\beta}'x'y}{y'y} = R^2$$

y por su puesto, tampoco el corregido de grados de libertad $\bar{R}^{*2} = \bar{R}^2$.

IV. EN TORNO A LAS PRUEBAS DE SIGNIFICABILIDAD

Se intuye fácilmente que todos los estadísticos empleados en los contrastes de la calidad del modelo y de sus variables exógenas no se habrán alterado; esto es, que las conclusiones derivadas del modelo inicial seguirán siendo válidas para el modelo con transformaciones lineales en las variables.

Como ya he dicho, el resultado es el lógico si tenemos en cuenta que las variables han sido modificadas mediante relaciones de tipo determinista que en ningún modo pueden hacer variar su capacidad explicativa ni, por supuesto, la del modelo, veamos

$$t^* = \left| \frac{\hat{\beta}_r^*}{DS(\hat{\beta}_r^*)} \right| = \left| \frac{\hat{\beta}_r}{DS(\hat{\beta}_r)} \right| = t$$

$$F^* = \frac{R^{*2}(N - K - 1)}{(1 - R^{*2}) \cdot K} = \frac{R^2(N - K - 1)}{(1 - R^2) \cdot K} = F$$

Prueba esta última que puede cómodamente ser entendida, no sólo para el modelo como un todo, sino para grupos de variables por separado y para el estudio de la capacidad explicativa de éstas o de la relevancia de la información perdida por el grupo de variables omitidas.

En definitiva, una vez más, pues, tanto los contrastes de hipótesis nulas para los coeficientes como el análisis de la varianza conjuntamente o por separado, no verán modificadas sus conclusiones de un modelo respecto a otro.

V. EN TORNO A LOS CONTRASTES DE LAS HIPOTESIS

Los distintos tests aplicables para corroborrar los presupuestos básicos de partida en el modelo de regresión o, en su caso, detectar su incumplimiento, no modifican el estadístico a emplear, puesto que como ya indicábamos no se ven alteradas sus conclusiones en relación al modelo primitivo, por lo menos en los contrastes que ahora probaremos; así, entre los más conocidos cabe plantearse:

— El test de Farrar y Glauber, para la multicolinealidad en el que la matriz de coeficientes de correlación entre las exógenas (Rx) sería la única expresión sospechosa de cambio, al haber alterado las variables independientes del modelo, no obstante, como ya he indicado anteriormente, el coeficiente de correlación simple es insensible ante cambios de medida en las variables; o cuando menos, su valor absoluto, que es lo que importa, no se altera. Es por ello por lo que fácilmente se comprueba que el determinante de aquella matriz no se ve modificado

$$|R^*x| = |Rx| = \begin{vmatrix} 1 & r_{12} & r_{13} & \dots & r_{1k} \\ & 1 & r_{23} & \dots & r_{2k} \\ & & 1 & \dots & r_{3k} \\ & & & \dots\dots & \\ & & & & 1 \end{vmatrix}$$

en consecuencia, tampoco se habrá alterado el estadístico χ^2, válido para el contraste.

— La fórmula de Spearman para medir la correlación entre los rangos de cada una de las exógenas y el del valor absoluto del error, si bien puede cambiar de resultado en cuanto al signo, puesto que según sea la ordenación de la variable explicativa cambia su sentido, no altera para nada su valor absoluto, que como he dicho es lo más relevante y, en consecuencia, tampoco se ve modificada la prueba t para la correlación, para la heterocedasticidad en nuestro caso.

— El contraste de Goldfield y Quandt también válido, sólo con grandes muestras para probar la constancia de la varianza del término de perturbación, analogamente seguirá proporcionando las mismas conclusiones, puesto que de la partición efectuada en la muestra, las regresiones resultantes en las dos submuestras extremas de igual tamaño nos proporcionan una relación entre variaciones residuales que no cambiará al manejar cualquiera de los dos bloques de exógenas de que disponemos, así en la medida que el cociente no sea inferior a uno, tendremos

$$\frac{\text{Máx}\,(\Sigma e_{1t}^2,\ \Sigma e_{2t}^2)}{\text{Mín}\,(\Sigma e_{1t}^2,\ \Sigma e_{2t}^2)} = \frac{\text{Máx}\,(\Sigma e^{*2},\ \Sigma e^{*2})}{\text{Mín}\,(\Sigma e_{1t}^{*2},\ \Sigma e_{2t}^{*2})}$$

Esto es, la prueba F para la heterocedasticidad sigue manteniendo su resultado pese al cambio habido en las variables.

— El estadístico de Durbin y Watson, idénticamente a lo razonado para los otros tests, mantiene el resultado en cualquiera de las dos especificaciones alternativas que venimos empleando.

$$dw = \frac{\Sigma(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\Sigma\hat{\varepsilon}_t^2} = \frac{\Sigma(\hat{\varepsilon}_t^* - \hat{\varepsilon}_{t-1}^*)^2}{\Sigma\hat{\varepsilon}_t^{*2}}$$

La independencia o autocorrelación que pueda presentar la perturbación ε_t en el primer modelo sigue estando presente en la perturbación ε_t^* del modelo transformado.

Detengamonos aquí en el estudio de la sensibilidad de los contrastes de hipótesis, aunque sólo sea por no alargarlo inútilmente o por haber razonado simplemente los tests más empleados en la práctica. Razonablemente, la rigidez que manifiestan las pruebas estudiadas no insinúan la posibilidad de que se mantengan para otros contrastes, pese a ser más sofisticados, en cualquier caso una generalización en mis apreciaciones, no puede ser entendidas más que analizando como se ha hecho en estos casos otras décimas para revalidar las hipótesis del modelo general.

No obstante, y a la vista de las pruebas efectuadas, podemos concluir que las pruebas derivadas de la estimación por MCO de la estructura original, son exactamente las mismas que los de la nueva estructura, al mantenerse o rechazarse a la par para aquéllas las hipótesis de cumplimiento general en el modelo.

SUMMARY

Linear Transformations in variables, ocurring in uniequational models, are justified provided that they are somehow economically significant or simplify calculations.

The effect of these changed meausures is analysed in the specification of the model, estimation of paramenters, significability tests and checkings of hipotheses.

Key words: Changed measures, linear transformations, new structure coefficients, sensitivity of checkings.

AMS, 1970, subject clasification: 52J05.