Tema 1. FORMA FUNCIONAL
Econometría II. 3º ºLADE

1 Introducción

Uno de los supuestos en los que se basan los resultados conocidos hasta el momento es que la relación entre la variable endógena y las variables explicativas es una relación lineal. Este supuesto, sin embargo, puede ser ciertamente restrictivo en ocasiones y, por tanto, distante de la realidad. Baste pensar, por ejemplo, que la gran mayoría de las formulaciones teóricas de modelos económicos no presuponen ningún particular de relación entre las variables, sino que lo hacen de una forma mucho más general; del tipo:

$$y_t = f\left(x_t, \beta\right) + u_t \tag{1}$$

Obviamente, no existe ninguna obligación de formular el modelo en términos lineales. No obstante, existen ventajas evidentes que han propiciado que la formulación lineal sea la base de la Econometría. Entre ellos destacamos que, en términos, generales la hipótesis de linealidad parece ser una hipótesis razonable o, cuando esto no ocurre, la hipótesis de linealidad puede ofrecer un primer resultado útil a la hora de con las no lineales.

En este tema vamos a tratar, precisamente, de conocer cómo se puede trabajar cuando los modelos no son lineales. Para ello debemos comenzar por distinguir diversas formas de no linealidad.

La primera de ellas es muy sencilla de resolver. Por ejemplo, pensemos en el siguiente modelo:

$$y_t = \beta_1 + \beta_2 x_t^h + u_t \tag{2}$$

Como es fácil de comprobar, siempre que el valor del parámetro h sea mayor que la unidad, entonces la relación entre las variables del modelo no será lineal. Esto, en principio, nos impide hacer uso de los métodos de estimación que nos son conocidos, mínimos cuadrados o máximo verosimilitud. Sin embargo, una sencilla operación algebraíca nos permite linealizar el modelo de la siguiente manera:

$$y_t = \beta_1 + \beta_2 z_t + u_t \tag{3}$$

siendo $z_t = x_t^h$.

En general podemos decir que si la no linealidad afecta a las variables del modelo, éste es siempre susceptible de ser linealizado a través de unas sencillas operaciones. No podemos decir lo mismo en aquellos modelos en los que aparece un problema de no linealidad en parámetros. Por ejemplo, si queremos estudiar la función de producción de una empresa, es posible que pensemos en una formulación del tipo Cobb-Douglas. En ese caso, el modelo teórico sería el siguiente:

$$y = A\, L^{\beta_2} K^{\beta_3} \qquad \beta_2,\, \beta_3 > 0 \tag{4}$$

donde y representa el output de un producto. mientras que L y K son, respectivamente, el input trabajo y el input capital. Los parámetros del modelo son A, β_2 y β_3. Como vemos, las variables están medidas en niveles, sin embargo, los parámetros que miden los efectos de las variables explicativas sobre la variable endógena no se presentan de forma lineal. En principio este modelo no se puede estimar a través de los métodos conocidos. Sin embargo, tal y como ocurrió en el caso anterior, una simple operación algebráica nos va a permitir expresar este modelo en los términos mínimos cuadrado ordinarios. Asi, si tomamos logaritmos neperianos obtenemos:

$$\ln\, y \,=\, \beta_1 + \beta_2 \ln\, L \,+ \beta_3\, \ln\, K \tag{5}$$

Donde el nuevo modelo es lineal en parámetros pero no en variables. Ahora bien, ya conocemos que este tipo de no linealidades son fácilmente resolubles. En este caso, una simple transformación de variables nos conduce al siguiente modelo:

$$y^* = \beta_1 + \beta_2\, L^* \,+ \beta_3\; K^* \tag{6}$$

donde $y^* = \ln\, y\,, L^* = \ln\, L$ y $L^* = \ln\, L$.

La estimación de este modelo es sencilla. La única diferencia con respecto al modelo lineal general es la interpretación económica de los parámetros. En el modelo (6) β_2 y β_3 son propensiones marginales entre las variables transformadas. Ahora bien, la interpretación correcta debe hacerse con respecto a las variables orignales, no con respecto a las transformadas. Por tanto, observando el modelo (4) vemos que A es la constante técnica y que los parámetros β_2 y β_3 son las elasticidades de los input trabajo y capital, respectivamente. Para comprobarlo, debemos tener en cuenta que:

$$e_{Ly} = \frac{\frac{\delta y}{\delta L}}{\frac{y}{L}} = \frac{\delta y}{\delta L} \times \frac{L}{y} = \beta_2 A L^{(\beta_2 - 1)} K^{\beta_3} \times \frac{L}{A L^{\beta_2} K^{\beta_3}} = \beta_2 \qquad (7)$$

La elasticidad del capital se obtiene de forma análoga. En consecuencia, los parámetros β_2 y β_3 se deben interpretar como los incrementos porcentuales que experimenta el output ante un incremento porcentual en, respectivamente, el input trabajo y el input capital.

Hasta este punto, todos los modelos que hemos analizado eran intrinsicamente lineales, en el sentido que era posible su linealización. Sin embargo, no siempre vamos a poder linealizar el modelo que queremos estudiar. Un ejemplo de ello lo tenemos en la función de producción CES (constant elasticity of substitution). Esta función se define de la siguiente manera:

$$\ln y = \beta_1 + \beta_4 \ln \left[\beta_2 L^{\beta_3} + (1 - \beta_2) K^{\beta_3} \right] \qquad (8)$$

En esta ocasión resulta imposible encontrar una formula que nos permita linealizar el modelo anterior. En consecuencia es necesario encontrar métodos de estimación alternativos a los que conocemos hasta el momento.

En lo que que sigue vamos a continuar la discusión en dos vertientes. La primera estará encaminada a estudiar algunos modelos que no son lineales en los parámetros, pero que se pueden linealizar fácilmente. La segunda estudiará qué hacer cuando el modelo es intrinsecamente no lineal, esto es, no es linealizable.

2 MODELOS LINEALIZABLES

En la presente sección vamos a presentar diversos modelos que no son lineales en sus parámetros, pero que pueden ser linealizados. No obstante, la razón de presentarlos aquí no es la presencia de no linealidad, sino su utilidad para estudiar diversos fenómenos económicos.

2.1 Ley de Zipf

Supongamos que queremos estudiar la evolución del tamaño de las aglomeraciones urbanas de una determinada region, como Aragon, o de un pais, supongamos España. En esta literatura de economía urbana juega un papel fundamental la denominada ley de Zipf. Esta es una ley de tipo empírico, sin

un transfondo teórico claro, aunque recientemente se le ha intentado dotar del mismo (ver, por ejemplo, Gabaix, 1999). De acuerdo con esta ley, la evolución de las ciudades se puede aproximar mediante la siguiente relación:

$$R_i X_i^a = A \tag{9}$$

donde R_i es el rango de la ciudad i-ésima y X_i es el tamaño, medido por el numero de habitantes, por ejemplo, de la aglomeración i-ésima.. A y a son sendos parámetros, Cuando $a = 1$, se dice que se cumple la ley de Zipf, en cuyo caso la relación entre R y X representa una hypérbola rectangular. De acuerdo con esto, si suponemos que la aglomeracion con mayor numero de habitantes tiene 100 habitantes, el cumplimiento de la ley de Zipf exige que la segunda algomeracion tenga 50 habitantes (100/2), la tercera 33 habitantes (100/3), la cuarta 25 (100/4), etc. El parametro a se puede interpretar como un indice de metropolizacion en el sentido de que una tendencia decreciente de este valor indica papeles relativamente mas importantes para las mayores ciudades y, por tanto, una mayor concentracion. Por contra, una tendencia creciente indica una mayor dispersion de la poblacion fuera de las aglomeraciones urbanas y, por tanto, una distribucion mas homogenea entre areas urbanas y rurales. En la literatura relacionada con economia urbana se demuestra que esta relacion se cumple en USA, pero la evidencia es menor en Europa.

Si estamos interesados en comprobar si la ley de Zipf se cumple para un conjunto de aglomeraciones, deberíamos estimar el valor del parámetro a y comprobar si este es igual a la unidad. Sin embargo, topamos con la no linealidad del modelo (9). Esta es fácilmente resoluble sin más que aplicar logaritmos neperianos a ambos lados de la igualdad, lo que nos lleva a la siguiente ecuación:

$$\begin{aligned} \ln\,(R_i X_i^a) &= \ln\,(R_i) + \ln\,(X_i^a) = \ln\,(R_i) + a \ln\,(X_i) = \ln A \Rightarrow \\ &\Rightarrow \ln\,(r_i) = \ln A - a \ln\,(x_i) \\ &\Rightarrow r_i = \ln A - a\, x_i \end{aligned}$$

Entonces, podemos estimar el siguiene modelo:

$$r_i = \beta_1 + \beta_2\, x_i + u_i$$

si el parámetro β_2 es igual a -1, entonces se puede admitir que se cumple la ley de Zipf.

Lanaspa et al. (2003) han analizado el cumplimiento de la ley de ZIpf en España. Estos autores obtienen que para el año 1900 la estimacion del parametro a es igual a 1.44, mientras que para 1999 es igual a 1.10, lo que señala un claro proceso de creacion de areas urbanas mas grandes en detrimento de la rurales.

2.2 Función de aprendizaje

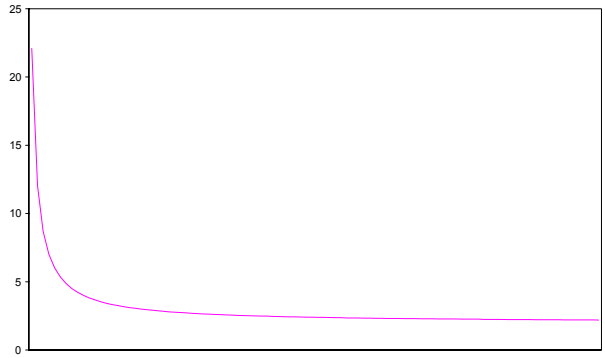
Supongamos que estamos interesados en estudiar como evoluciona la cadena de montaje de un determinado bien, por ejemplo, vagones de tren. Un aspecto que podría ser interesante para la empresa sería conocer cuál es el número de horas que se tarda por unidad construida. Esto es lo que se conoce como función de aprendizaje. En este caso la forma funcional lineal no nos sirve de ayuda. La razón es que el proceso de construcción de las piezas funciona de tal forma que al inicio de la actividad el número de horas por pieza construida es mayor que una vez que la cadena de trabajo funciona a pleno ritmo. Las razones de ser de esta relación son fácilmente entendibles. Imaginemos que estamos en una cadena de montaje, en la que tenemos que poner un asiento dentro de un coche. Cuando nos viene el primer coche no sabemos muy bien como ponermos dentro del coche, donde estan los tornillos, cómo se enroscan éstos, etc. Todo ello supone un retraso en la cadena. Sin embargo, con el tiempo estos pequeños parones desaparecen y al encargado realiza el trabajo en el menor tiempo posible. Como vemos, la relación entre coste de tiempo por unidad y número de piezas trabajadas es no lineal.

Para modelizarla, se utiliza lo que se conoce como modelo inverso. Este se representa de la siguiente manera:

$$y_i = \beta_1 + \beta_2 \frac{1}{x_i} \tag{10}$$

En el caso de la función de aprendizaje, y_i sería el coste de tiempo por unidad producida, mientras que x_i recogería el número de piezas producidas.

Gráfico 1.1. Función de aprendiza



Es evidente que el modelo (10) no es lineal, pero se puede linealizar simplemente sin más que realizar una transformación del siguiente tipo

$$y_i = \beta_1 + \beta_2 x_i^* + u_i \tag{11}$$

donde $x_i^* = \frac{1}{x_i}$.

La interpretación del modelo es la siguiente. Cuando $x_i \to \infty$, entonces $y_i \to \beta_1$.Si además, tenemos que $\beta_2 > 0$, entonces, β_1 representa el valor mínimo que puede adoptar la variable y_i

El modelo anterior no solamente se puede aplicar al caso de la función de aprendizaje. Es también válido para toda aquella relación entre dos (o más) variables que presente bien un suelo (valor mínimo que no puede sobrepasar) o bien un valor techo (valor máximo). En este último caso, el parámetro β_2 debería ser negativo.

3 Modelos no lineales. Estimación por métodos no lineales.

We consider the following general nonlinear model*:*

$$\begin{aligned} y_t &= f(x_t;\beta) + \varepsilon_t \quad t = 1,\ldots,T \\ \varepsilon_t &\sim N\left(0,\sigma^2 I\right) \\ x_t &: k \times 1 \text{ vector of exogeneous variables} \end{aligned}$$

β : $n \times 1$ vector of parameters

$f(\cdot)$: some function satisfying some regularity conditions

We assume that the function is such that all the parameters are asymptotically identified (see Davidson and MacKinnon, ch. 5.2 for a discussion). Since the ε_t's are normal, we can write

$$\log L\left(\beta, \sigma^2\right) = -\frac{T}{2}\log 2\pi - \frac{T}{2}\log \sigma^2 - \frac{1}{2\sigma^2}S(\beta)$$

where

$$S(\beta) = \Sigma_{t=1}^{T}\left\{y_t - f(x_t;\beta)\right\}^2 = \Sigma_{t=1}^{T}\varepsilon_t^2.$$

Hence, the MLE is the same as the nonlinear least squares estimator.

Define the following vector of first derivatives :

$$z_t = \frac{\partial f(x_t;\beta)}{\partial \beta} = -\frac{\partial \varepsilon_t}{\partial \beta}.$$

Then,

$$\begin{aligned}
\frac{\partial \log L(\beta, \sigma^2)}{\partial \beta} &= \sigma^{-2}\Sigma_{t=1}^{T} z_t\left(y_t - f(x_t;\beta)\right) \\
\frac{\partial^2 \log L(\beta, \sigma^2)}{\partial \beta \partial \beta'} &= \sigma^{-2}\Sigma_{t=1}^{T}\left[\frac{\partial z_t}{\partial \beta}\varepsilon_t - z_t z_t'\right]
\end{aligned}$$

and

$$I(\beta) = E\left[\frac{\partial^2 \log L}{\partial \beta \partial \beta'}\right] = \sigma^{-2}\Sigma_{t=1}^{T} z_t z_t'$$

since

$$E\left[\frac{\partial z_t}{\partial \beta}\varepsilon_t\right] = 0,$$

given that the x's are fixed.

The off-block diagonal elements of the information matrix are 0 since

$$E\left[-\frac{\partial^2 \log L}{\partial \beta \partial \sigma^2}\right] = E\left[\sigma^{-4}\Sigma_{t=1}^{T} z_t \varepsilon_t\right] = 0$$

by the assumption that the elements of x_t are fixed. Therefore, under regularity conditions on $f(\cdot)$, we have (using the MLE asymptotic result)

$$\sqrt{T}\left(\hat{\beta} - \beta\right) \to^d N\left(0, \sigma^2 \lim_{T\to\infty}\left(T^{-1}\Sigma_1^T z_t z_t'\right)^{-1}\right)$$

and, as in the linear case,

$$\sqrt{T}\left(\hat{\sigma}^2 - \sigma^2\right) \rightarrow^d N\left(0, 2\sigma^4\right).$$

Remark 1 *As in the linear case, the nonlinear least squares estimator will have the same asymptotic distribution if the normality assumption on the errors is not imposed.*

Remark 2 *The asymptotics of the nonlinear case looks much like that of the linear case. Most results carry over after suitable modifications. What is different (and more difficult) are the precise regularity conditions upon which such results are valid.*

Remark 3 *The finite sample properties may be very different. Asymptotically, the slope of $f(\cdot)$ approximates the true function arbitrarily well in the neighborhood of the true value. In finite samples, the approximation will depend on the function.*

3.1 Numerical optimization.

We consider the problem of minimizing a criterion function $f(\psi)$ with ψ an n vector of parameters (if the function is to be maximized we simply take its negative). As a matter of notation, we let

$$\begin{aligned} \underset{n\times 1}{g(\psi)} &= \frac{\partial f(\psi)}{\partial \psi} \text{ vector of first derivatives} \\ \underset{n\times n}{G(\psi)} &= \frac{\partial f(\psi)}{\partial \psi \partial \psi'} \text{ matrix of second derivatives (Hessian).} \end{aligned}$$

A sufficient condition for a local minimum is that

$$g(\hat{\psi}) = 0 \text{ and } G(\hat{\psi}) \text{ is positive definite.}$$

This only ensures a local minimum, however. It is very difficult to be certain whether a global minimum has been found. No algorithm can ensure a global minimum with certainty. The best thing to do is to repeat the optimization with different starting values.

3.1.1 Numerical evaluation of the derivatives.

Most numerical optimization techniques involve the calculation of the first and possibly second derivatives. The first derivatives are approximated numerically by

$$g_j(\hat{\psi}) = \left\{ f(\hat{\psi} + h_j e_j) - f(\hat{\psi}) \right\} / h_j, \quad j = 1, \ldots, n,$$

where

$$e_j = \begin{pmatrix} 0 & 0 \ldots & \underset{\substack{\downarrow \\ \text{j}^{th} \text{ element}}}{1} & \ldots 0 & 0 \end{pmatrix}$$

is the unit vector with a one in position j and h_j is called the *step length.* We choose h_j small enough for the difference estimate to be close to the true derivative, but not so small as to be dominated by rounding errors. Using the same notation the second derivatives are calculated as

$$G_{ji}(\hat{\psi}) = \left\{ g_i(\hat{\psi} - h_i e_i) - g_j(\hat{\psi}) \right\} / h_i, \quad i, j = 1, \ldots, n.$$

3.1.2 Newton-Raphston method.

Take a Taylor series expansion of $f(\psi)$ around $\hat{\psi}$, the minimum, then we have:

$$f(\psi) \approx f(\hat{\psi}) + (\psi - \hat{\psi})' g(\hat{\psi}) + \frac{1}{2}(\psi - \hat{\psi})' G(\hat{\psi})(\psi - \hat{\psi}).$$

Differentiate with respect to ψ

$$g(\psi) \approx g(\hat{\psi}) + G(\hat{\psi})(\psi - \hat{\psi}).$$

But since $\hat{\psi}$ is the minimum $g(\hat{\psi}) = 0$ and

$$\hat{\psi} \approx \psi - G^{-1}(\hat{\psi}) g(\psi).$$

This suggests the following iterative scheme. Let $\tilde{\psi}$ be some initial estimate and ψ^* the revised estimate; then,

$$\psi^* = \tilde{\psi} - G^{-1}(\hat{\psi}) g(\tilde{\psi}).$$

This involves evaluating $G^{-1}(\hat{\psi})$ at $\hat{\psi}$ the minimum which is unknown. The solution is to evaluate $G(\cdot)$ at the current estimate $\tilde{\psi}$. Then,

$$\psi^* = \tilde{\psi} - G^{-1}(\tilde{\psi}) g(\tilde{\psi}).$$

The estimate are obtained by iterating until convergence.

3.1.3 Convergence criteria.

The convergence may be defined in a number of ways, for example:
1) $f(\psi^*)$ close to $f(\tilde{\psi})$,
2) ψ^* close to $\tilde{\psi}$,
3) $g(\psi^*)$ close to $g(\tilde{\psi})$,
where "close" is a small number.

Remark 4 *Achieving convergence is quite difficult sometimes for the following reasons: a) problem of local versus global maximum, b) in practice, the objective function is often flat so that convergence may be slow.*

Potential problems:

There are also issues related to two potential problems:

a) $G(\tilde{\psi})$ may not be positive definite,
b) the difference $\psi^* - \tilde{\psi}$ obtained may be too large (overshooting).

Solutions*:*

For problem (a), we consider the iterations :

$$\psi^* = \tilde{\psi} - \left[G(\tilde{\psi}) + \mu I\right]^{-1} g(\tilde{\psi})$$

where μ is a positive scalar that varies as the iteration proceeds and that ensures a positive definite matrix. Some methods of choosing μ have been proposed by Goldfeld, Quandt and Trottier (1966).

For the problem (b) consider the iteration :

$$\psi^* = \tilde{\psi} - \lambda G^{-1}(\tilde{\psi}) g(\tilde{\psi}). \tag{12}$$

We choose λ by a line search either to accelerate convergence or avoid overshooting (method proposed by Fletcher-Powell).

Remark 5 *If $\tilde{\psi}$ is not a minimum, it can be shown that there exists a λ such that $f(\psi^*) < f(\tilde{\psi})$ using (12). So, using such a method avoids overshooting and permits faster convergence.*

3.1.4 Maximization of a likelihood function.

When maximizing the log-likelihood function, we may write the Newton-Raphson iteration as :

$$\psi^* = \tilde{\psi} - \left[D^2 \log L(\tilde{\psi})\right]^{-1} D \log L(\tilde{\psi})$$

where

$$\begin{aligned} D \log L(\tilde{\psi}) &= \frac{\partial \log L(\psi)}{\partial \psi}|_{\psi=\tilde{\psi}} \\ D^2 \log L(\tilde{\psi}) &= \frac{\partial^2 \log L(\psi)}{\partial \psi \partial \psi'}|_{\psi=\tilde{\psi}}. \end{aligned}$$

The method of scoring. Instead of the matrix of second derivatives, we use its expectation

$$E\left[D^2 \log L(\tilde{\psi})\right] = -I(\tilde{\psi}),$$

the information matrix evaluated at the current estimate $\tilde{\psi}$.

Drawback*:*

- The information matrix is only an approximation to the Hessian, hence its use may imply slower convergence.

Advantages:

- It is often easier to compute, so this method may be quicker and may compensate for a slower convergence rate.
- The information matrix is positive definite by construction, so it avoids some of the difficulties of Newton-Raphston. But a variable step length λ should still be used.

3.1.5 Gauss-Newton method.

This method is specifically designed for the calculation of the nonlinear least-squares estimates; i.e., when we have the minimization of a sum-of-squares function of the form :

$$f(\psi) = S(\psi) = \Sigma_{t=1}^{T} \varepsilon(\psi)_t^2 .$$

The first derivatives are:

$$g(\psi) = \frac{\partial f(\psi)}{\partial \psi} = 2\Sigma_{t=1}^{T} \frac{\partial \varepsilon_t}{\partial \psi} \varepsilon_t.$$

and the Hessian is:

$$G(\psi) = \frac{\partial^2 f(\psi)}{\partial \psi \partial \psi'} = 2\Sigma_{t=1}^{T} \left\{ \overbrace{\frac{\partial \varepsilon_t}{\partial \psi} \frac{\partial \varepsilon_t}{\partial \psi'}}^{1} + \overbrace{\frac{\partial^2 \varepsilon_t}{\partial \psi \partial \psi'} \varepsilon_t}^{2} \right\}.$$

The difference now is that we usually have the second term small relative to the first one, so we can neglect it. Therefore, the Gauss-Newton iteration scheme is

$$\psi^* = \tilde{\psi} - \left[\Sigma_{t=1}^{T} \frac{\partial \varepsilon_t}{\partial \psi} \frac{\partial \varepsilon_t}{\partial \psi'} \right]^{-1} \Sigma_{t=1}^{T} \frac{\partial \varepsilon_t}{\partial \psi} \varepsilon_t(\tilde{\psi})$$

where the derivatives are evaluated at the current estimate.

Remark 6 *This method involves only the vector of the first derivatives, not the second derivatives.*

Remark 7 *This method can be interpreted as an OLS regression. Let* $z_t = -\frac{\partial \varepsilon_t}{\partial \psi}$. *Then,*

$$\psi^* = \tilde{\psi} + \left(\Sigma_{t=1}^{T} z_t z_t' \right)^{-1} \Sigma_{t=1}^{T} z_t \varepsilon_t.$$

So, Gauss-Newton is a series of OLS regressions. At each step, z_t *and* ε_t *are evaluated at the point* $\tilde{\psi}$; *and* $\tilde{\psi}$ *is then updated by means of a regression of* ε_t *on* z_t.

Remark 8 *Similar modifications can be incorporated as in the Newton-Raphson case; for example using a Marquart quadratic hill climbing method*

$$\psi^* = \tilde{\psi} + \left(\Sigma_{t=1}^{T} z_t z_t' + \mu I \right)^{-1} \Sigma_{t=1}^{T} z_t \varepsilon_t$$

for a suitably chosen μ.

The Berndt, Hall, Hall and Hausman method This method attempts to retain the advantage of Gauss-Newton in the more general minimization problem by not using the matrix of second derivatives. The suggestion is to approximate the Hessian by

$$\Sigma_{t=1}^{T} \frac{\partial \log L\left(y_t;\psi\right)}{\partial \psi} \frac{\partial \log L\left(y_t;\psi\right)}{\partial \psi'}.$$

This is suggested by the equality (see ch. 8).

$$E\left[\frac{\partial \log L\left(\psi\right)}{\partial \psi} \frac{\partial \log L\left(\psi\right)}{\partial \psi'}\right] = -E\left[\frac{\partial^2 \log L\left(\psi\right)}{\partial \psi \partial \psi'}\right].$$

However, the approximation in finite samples may be inadequate.

4 TRANSFORMADAS DE BOX-COX

Hasta el momento, hemos considerado diversas formas no lineales alternativas. Cada una de ellas las hemos asociado con el estudio de una variable concreta y, en consecuencia, hemos asociado su uso a una serie de casos particulares. Sin embargo, no siempre es posible realizar estas disposiciones, y el investigador, a priori, puede dudar entre diversas formas funcionales alternativas. Por ello, parece conveniente disponer de un armazón teórico que nos permita determinar cuál es la forma funcional que mejor se ajusta a mis datos. Dicho de otro modo, endogeneizar la determinación de la forma funcional. Este es el objetivo de las transformadas de Box-Cox.

Su fundamento es permitir que sean los propios datos quienes dicten qué forma funcional es la más adecuada. Para ello, a partir de una transformación sobre alguna o todas de las variables de una especificación, crean una familia de funciones, todas ellas anidadas entre sí. La transformación que usan Box y Cox es la siguiente:

$$z^{(\lambda)} = \begin{cases} \frac{z^{\lambda}-1}{\lambda} & \text{si } \lambda \neq 0 \\ \ln z & \text{si } \lambda = 0 \end{cases}$$

Así cuando, por ejemplo, $\lambda = 1$, tenemos que $z^{(\lambda)} = z - 1$. Si se cumple que $\lambda = -1$, tenemos que $z^{(\lambda)} = -\frac{1}{z} + 1$. Estos dos resultados son directos. No lo es tanto comprobar que cuando $\lambda = 0$, $z^{(\lambda)} = \ln z$. Para ello debemos tener en cuenta que:

$$z^{(0)} = \frac{z^0 - 1}{0} = \frac{1-1}{0} = \frac{0}{0}$$

Por lo tanto llegamos a una indeterminacion. Para resolverla aplicamos la regla de L'Hôpital, de forma que se cumple que:

$$\lim_{\lambda \to 0} z^{(\lambda)} = \lim_{\lambda \to 0} \frac{\frac{d\left(z^\lambda - 1\right)}{d\lambda}}{\frac{d\lambda}{d\lambda}} = \lim_{\lambda \to 0} \frac{z^\lambda \ln z}{1} = \ln z$$

La gran utilidad que presentan estas transformadas Box-Cox es que anidan los modelos estudiados con anterioridad. Para ver cómo se anidan estas formas funcionales, supongamos que disponemos de dos variables, x e y, que se transforman de acuerdo a los parámetros λ_x y λ_y, respectivamente. Esta relación genérica la podemos expresar de la siguiente manera:

$$y^{(\lambda_y)} = \beta_1 + \beta_2 x^{(\lambda_x)} + u \tag{13}$$

En nuestro caso concreto, al imponer la restricción $\lambda_x = \lambda_y = 1$, en realidad no estamos sino considerando el modelo lineal simple. Para verlo, basta con desarrollar el modelo anterior, resultando que:

$$\begin{aligned} \frac{y^1 - 1}{1} &= y - 1 = \beta_1 + \beta_2 \frac{x^1 - 1}{1} + u = \beta_1 - \beta_2 + \beta_2 x + u \Rightarrow \\ y &= \beta_1 - \beta_2 + 1 + \beta_2 x + u = \beta_1^* + \beta_2 x + u \end{aligned}$$

Del mismo modo, si tenemos que $\lambda_x = \lambda_y = 0$, la ecuación (13) se convierte en:

$$\ln y = \beta_1 + \beta_2 \ln x + u$$

esto es, en el modelo doble logaritmico. El resto de posibles modelos se obtiene de forma similar.

El modelo ((13) se puede extender al caso del modelo lineal general de k parametros de la siguiente manera:

$$y^{(\lambda_1)} = \beta_1 + \beta_2 x_2^{(\lambda_2)} + ... + \beta_k x_k^{(\lambda_k)} + u \tag{14}$$

Si conocemos a priori el valor de los parametros de las transformadas Box-Cox, el proceso para estimar los parametros del modelo lineal general sera el de, en primer lugar, imponer estos valores de los parámetros $\lambda_i,\ i = 1, 2, ..., k$. Una vez incluida esta información, entonces se estiman los parámetros de posición y el de dispersión para una forma funcional dada. En el caso de que tengamos información a priori sobre el valor de los parámetros de las transformadas Box-Cox esta es una buena estrategia. Sin embargo, es posible que no dispongamos de información sobre estos parámetros o, simplemente, que estemos interesados en endogeneizar estos parámetros, de forma que sean los datos los que nos indiquen cuál es la forma funcional más adecuada para la muestra disponible. En este caso, parece aconsejable estimar conjuntamente todos los parámetros del modelo. El procedimiento a seguir para ello se estudia en el siguiente apartado.

4.1 Estimación de las transformadas Box-Cox

El punto de partida de esta sección es el modelo (14), en el que asumimos que la perturbación cumple todas las hipotesis basicas. El objetivo es estimar los parámetros de los que depende el modelo: $\beta_i,\ \lambda_i,\ i = 1, 2, ..., k$ y σ^2. La aplicación de los métodos de estimación basados en la hipótesis de linealidad no es posible, dada la evidente no linealidad del modelo. En consecuencia, debemos acudir al uso de métodos no lineales de estimación. El métodoo que se emplea con mayor frecuencia en este caso es el basado en la estimación maximo verosímil no lineal. Suponiendo normalidad en las perturbaciones, la función de densidad conjunta del vector de perturbaciones viene dada por:

$$f(u) = \left(2\pi\sigma^2\right)^{-\frac{T}{2}} \exp\left(-\frac{u'u}{2\,\sigma^2}\right)$$

Suponiendo, por cuestiones de comodidad en la exposición, que todas las variables están transformadas de forma idéntica, lo que supone que $\lambda_i = \lambda_1, \forall i = 1, 2, ..., k$, la función de densidad conjunta para el vector de observaciones de la endógena viene dado por:

$$f(y) = \left(2\pi\sigma^2\right)^{-\frac{T}{2}} \exp\left\{-\frac{\left[Y^{(\lambda)} - X^{(\lambda)}\beta\right]' \left[Y^{(\lambda)} - X^{(\lambda)}\beta\right]}{2\,\sigma^2}\right\}\ abs\left(\left|\frac{\partial u'}{\partial y}\right|\right) \tag{15}$$

donde $abs\left(\left|\frac{\partial u'}{\partial y}\right|\right)$ es el valor absoluto del determinante del Jacobiano de la transformación de u sobre y. En el caso general, tenemos que:

$$\frac{\partial u'}{\partial y} = \begin{bmatrix} \frac{\partial u_1}{\partial y_1} & \frac{\partial u_2}{\partial y_1} & \cdots & \frac{\partial u_T}{\partial y_1} \\ \frac{\partial u_1}{\partial y_2} & \frac{\partial u_2}{\partial y_2} & \cdots & \frac{\partial u_T}{\partial y_2} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial u_1}{\partial y_T} & \frac{\partial u_2}{\partial y_T} & \cdots & \frac{\partial u_1}{\partial y_T} \end{bmatrix} = \begin{bmatrix} y_1^{\lambda_1 - 1} & 0 & \cdots & 0 \\ 0 & y_2^{\lambda_1 - 1} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & y_T^{\lambda_1 - 1} \end{bmatrix}$$

por lo que resulta directo probar que:

$$\left|\frac{\partial u'}{\partial y}\right| = \begin{vmatrix} y_1^{\lambda_1 - 1} & 0 & \cdots & 0 \\ 0 & y_2^{\lambda_1 - 1} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & y_T^{\lambda_1 - 1} \end{vmatrix} = \prod_{i=1}^{T} y_i^{\lambda_1 - 1} \tag{16}$$

Si incorporamos el valor de (16) en (15) y tomamos logaritmos neperianos, comprobamos que la funcion log-verosímil es igual a:

$$\begin{aligned} \ell\left(\beta, \sigma^2, \lambda / \, y, x\right) &= -\frac{T}{2} \ln 2\pi - \frac{T}{2} \ln \, \sigma^2 \\ &\quad -\frac{1}{2\,\sigma^2} \, \left[Y^{(\lambda)} - X^{(\lambda)}\beta\right]' \, \left[Y^{(\lambda)} - X^{(\lambda)}\beta\right] \qquad (17) \\ &\quad + (\lambda_1 - 1) \sum_{i=1}^{T} \ln \, y_i \qquad (18) \end{aligned}$$

Diferenciando esta función con respecto al vector de parametros de posicion y al parametro de dispersión, e igualando las respectivas derivadas a 0, obtenemos:

$$\tilde{\beta}(\lambda) = \left[X^{(\lambda)\prime} X^{(\lambda)}\right]^{-1} \, X^{(\lambda)\prime} Y^{(\lambda)} \tag{19}$$

$$\tilde{\sigma}^2(\lambda) = \frac{\left[Y^{(\lambda)} - X^{(\lambda)}\tilde{\beta}(\lambda)\right]' \left[Y^{(\lambda)} - X^{(\lambda)}\tilde{\beta}(\lambda)\right]}{T} \tag{20}$$

Como vemos, asumiendo conocidos los parametros λ, los estimadores del resto de los parametros del modelo son equivalentes a los ya conocidos. Si, por el contrario, el valor de λ es desconocido, el proceso de estimación es bietápico. Primero debemos encontrar el valor de λ que maximiza la función

de verosimilitud. Una vez determinado este, se procede de forma habitual para hallar $\tilde{\beta}(\lambda)$ y $\tilde{\sigma}^2(\lambda)$. Para determinar λ, debemos concentrar la función de verosimilitud, sustituyendo (19) y (20) por, respectivamente, β y σ^2 en (**??**). La funcion log-verosimil concentrada es igual a:

$$\ell^* (\lambda / y, x) = -\frac{T}{2} \ln 2\pi - \frac{T}{2} \ln \tilde{\sigma}^2(\lambda) - \frac{T}{2} + (\lambda_1 - 1) \sum_{i=1}^{T} \ln y_i \qquad (21)$$

Su maximización se suele hacer mediante algoritmos de optimización muy sencillos. Uno de ellos, por ejemplo, sería obtener el mayor valor de la función log-verosímil para un conjunto de valores del parámetro λ_1: Nos quedaríamos con aquel valor que maximizara la función log-verosimil o, lo que es igual, con el que minimizara el valor de la suma residual del modelo.

Esta primera posibilidad es factible en la medida que dispongamos de un ordenador con capacidad suficiente. Sin embargo, resulta un poco engorrosa al tener que comparar continuamente los valores y no saber cuál es la aproximación que debemos utilizar. Por ello resulta conveniente utilizar una forma alternativa para obtener la estimación de los parámetros del modelo. Para ello, podemos calcular la media geométrica de los valores del vector y:

$$\bar{y}_G = (y_1 \times y_2 \times \cdots \times y_T)^{\frac{1}{T}}$$

Una vez calculada esta media geométrica, dividimos cada una de las observaciones por dicha media geométrica. Es obvio que ahora la nueva función log-verosimil es igual a:

$$L^* (\lambda / \tilde{y}, \tilde{x}) \sim -\frac{T}{2} \ln \tilde{\sigma}^2(\lambda)$$

por cuanto $\sum_{i=1}^{T} \ln \tilde{y}_i = 0$. Entonces, la función log-verosímil es función exclusiva, al margen de una constante, de $\tilde{\sigma}^2(\lambda)$. A partir de aquí, es directo deducir que el valor óptimo del parámetro λ es aquél que minimiza la suma residual de la regresión entre $\tilde{y}$ sobre $\tilde{x}$, donde estas variables estan expresadas como cociente de los valores originales con respecto a sus respectivas medias geometricas.

Debemos reconocer que estos métodos de estimación condicionada tienden a estimar por debajo de sus valores, por lo que otros métodos basados en una estimación no lineal conjunta de todos los parámetros nos pueden conducir a mejores resultados.

5 CONTRASTES DE FORMA FUNCIONAL

Una vez que hemos estimado el modelo con todos sus parámetros, incluido el valor de los parámetros de las transformadas Box-Cox λ, parece adecuado disponer de métodos que nos permitan contrastar cuál de la forma funcional es la más adecuada. De esta manera, por ejemplo, si los valores de los parametros de las transformadas están próximos a la unidad, podriamos imponer la restriccion de que todos son igual a la unidad, estimando el modelo lineal. Si esta restricción es cierta, el modelo así estimado contará con mejores propiedades, en el sentido de presentar una menor varianza que el modelo sin restricciones. Por ello, resulta conveniente disponer de estadisticos que nos permitan discriminar entre las diversas formas funcionales.

Una primera aproximación a este tema nos la proporciona el contraste de la razón de verosimilitud. Recordemos que este estadistico se define, en terminos generales, de la siguiente manera:

$$RV \; = \; -2 \; \ln \; \frac{L(H_0)}{L(H_A)} = -2 \left[\ln L(H_0) - \ln L(H_A) \right] \overset{as.}{\sim} \; \chi^2_r$$

donde $L(H_0)$ y $L(H_A)$ son la función de verosimilitud del modelo bajo la hipotesis nula (modelo restringido) y la alternativa (modelo sin restricciones), respectivamente, mientras que r representa el numero de restricciones impuestas.

Véamos su uso en el siguiente ejemplo. Supongamos que para una muestra de 100 observaciones, la estimacion por Box-Cox de un modelo nos arroja el siguiente resultado:

$$\hat{y}_i^{(0.95)} \; = \; 0.7 \; + \; 0.6 \; \; x_i^{(1.04)}, \ell = -100$$

donde ℓ es el logaritmo neperiano de la funcion de verosimilitud. Al mismo tiempo, conocemos los resultados para la estimacion lineal y doble logarítmica:

$$\hat{y}_i = \; 0.71 \; + \; 0.6 \; x_i, \ell = -100.2$$

$$\ln \hat{y}_i = \; -2 \; + \; 0.04 \; \ln \; x_i, \ell = -200$$

Entonces, podemos contrastar cuál de las formas funcionales es la más adecuada a partir del cálculo de la razón de verosimilitud:

$$RV_1 = -2\left[-100.2 - (-100)\right] = 0.04$$

$$RV_2 = -2\left[-200 - (-100)\right] = 200$$

El número de restricciones impuestas en ambos casos es de 2: $\lambda_1 = \lambda_2 = 1$, en el modelo lineal, y $\lambda_1 = \lambda_2 = 0$ en el doble logarítmico. Como el valor de una distribucion χ^2_2 al nivel de significacion del 5% es igual a 5.99, es claro que el estadistico RV_1 nos permite aceptar la hipotesis nula $H_o : \lambda_1 = \lambda_2 = 1$. Por contra, RV_2 nos conduce al rechazo de la hipotesis nula $H_o : \lambda_1 = \lambda_2 = 0$. En consecuencia, la evidencia empírica de este caso nos lleva a considerar la forma funcional lineal como la más adecuada.

Bibliografía

Cuairán, R., M. Sanso y F. Sanz (1991). "Flujos bilaterales de comercio internacional, ecuación de gravedad y forma funcional", *Revista Española de Economía*, 8 (2), 331-348.

Gabaix, X. (1999). "Zipf's Law and the Growth of Cities", *American Economic Review*, 89, 129-132.

Greene, W. H. 1999. Análisis econométrico. 3^a Edición. Prentice-Hall, Madrid. Cap. 8 y 10.

Johnston, J. (1987). Métodos Econométricos. Ed. Vicens Vives, Barcelona. Cap. 3.2 y 3.3.

Judge, G. G., R. C. Hill, W. E. Griffiths, H. Lütkepohl y T. Lee (1988). Introduction to the Theory and Practice of Econometrics. Ed. Wiley. Cap. 12

Lanaspa, L., F. Pueyo and F. Sanz, (2003), "The Evolution of the Spanish Urban Structure during the Twentieth Century", *Urban Sudies*, 40 (3), 567-580

Novales, A. (1993). Econometría. Ed. McGraw-Hill. Madrid. Cap.11 y 12.

Maddala, G.S. (1992). Introduction to Econometrics. Ed. Prentince-Hall. New-Jersey. Cap. 5.6

Sanso, M., R. Cuairán y F. Sanz (1993). "Bilateral Trade Flows, the Gravity Equation, and Functional Form", *The Review of Economic and Statistics*, 75, 266-375.