ESTADISTICA ESPAÑOLA
núm. 116, 1988, págs. 45 a 53

Predicciones de variables endógenas en problemas ligados a la existencia de umbrales.

por
JOSE LUIS ROJO GARCIA
Profesor Titular de Estadística y Econometría
de la Facultad de CC.EE. y EE. de Valladolid.
e
ISABEL MARTIN ROJO
Profesora Agregada de I.B. del I.B. Ferrari de Valladolid

RESUMEN

Se construye en este artículo un predictor de la variable endógena en un modelo lineal general, cuando el riesgo se mide en función de los errores en la predicción de la superación o no de un umbral. El predictor construído es de Bayes para cierta distribución a priori, y se demuestra además, que es consistente.

Palabras clave: Econometría, funciones de decisión, modelos de respuesta cualitativa.

Clasificación AMS-1980: Principal 62H12. Secundaria 62P20.

1. INTRODUCCION

En la modelización econométrica se plantea a veces la existencia de umbrales, es decir, valores de ciertas variables cuya superación provoca determinados efectos sobre otra u otras. Es conocido, por ejemplo, el problema de elección dinámica, en el que un suceso ocurre si cierta variable aleatoria, $Y(i,t)$, diferente en el tiempo (t=1,...,T) y según los individuos (i=1,...,I), rebasa cierto umbral. Una técnica ya clásica de resolución de

este problema, que puede verse en Heckman (1981), consiste en definir una variable ficticia, $d(i,t)$ que valdrá uno o cero según que Y rebase o no un cierto umbral y plantear el problema con las técnicas de los modelos de elección discreta. Por ejemplo, una persona se clasifica como pobre si su renta, $E(i,t)$ no rebasa un valor C (en este caso, $Y(i,t) = C - E(i,t)$) o el problema del "salario de reserva" (reservation wage) en el que un individuo se incorpora a la fuerza de trabajo si el salario que ofrece el mercado rebasa el salario de reserva ($Y(i,t)$ sería la diferencia entre ambos, y el umbral sería cero).

En situaciones de este tipo, el investigador necesita predecir, no ya los valores de una cierta variable endógena del modelo, sino si dicha variable rebasará o no el umbral prefijado.

Dentro de la literatura econométrica existen, esencialmente, dos modos de abordar este problema; hablando sucintamente, los siguientes:

a) La estimación de los parámetros del modelo (por MCO, MCG, etc.) para, a continuación, contrastar con criterios probabilísticos la hipótesis de superación del umbral.

b) La construcción de una variable dicotómica cuyos valores se definan, para el periodo muestral, según si la variable endógena rebasa o no el umbral correspondiente, y el estudio de un modelo que sustituya la variable endógena primitiva por dicha variable dicotómica.

La primera de estas opciones utiliza estimadores que, en general, son eficientes de acuerdo con criterios que, desgraciadamente, no son sensibles al problema de los umbrales planteado, sino que tienen en cuenta únicamente la "distancia" entre el estimador (la estimación) y el valor del parámetro. La segunda provoca una pérdida de información muestral en el paso de la variable original a la dicotómica que puede ser importante en ciertas situaciones.

En este trabajo, proponemos la obtención de un estimador que sea óptimo (de Bayes) con respecto a una función de pérdida sensible a la existencia de un umbral; en concreto, que mida el error que cometemos al estimar la probabilidad de que la variable endógena rebase un umbral prefijado.

Puesto que los errores que se estudian son errores en la predicción, y no en la estimación, la función de pérdida construída no es sensible a la segunda y sí a la primera, por lo que se obtiene toda una familia de estimadores que proporcionan una misma predicción; en otras palabras, se obtiene una predicción única, pero no una estimación única, salvo en situaciones muy particulares (uniparamétricas).

2. PLANTEAMIENTO DEL PROBLEMA

Supongamos, pues, que estamos frente a un modelo

$$y_t = \sum_{i=1}^{k} \beta_i x_{ti} + \varepsilon_t , \quad t=1,...,T$$

en las hipótesis clásicas del modelo lineal general,

(*i*) $X = (x_{ti})$, $i=1,...,k$, $t=1,...,T$, es de rango k, y los x_{ti} son no aleatorios.

(*ii*) $\varepsilon_t \sim NIID(0,\sigma^2)$.

Supongamos asímismo que σ es conocida.

Sea $X^o = (x_1^o,..., x_k^o)$ el vector de valores de las variables exógenas para el periodo de predicción, $\beta = (\beta_1,..., \beta_k)'$ y $u = X^o\beta + \varepsilon^o$, (obviamente, u es función de β, pero no lo pondremos de manifiesto en la notación). Supondremos que $(\varepsilon_1,..., \varepsilon_T, \varepsilon^o)'$ verifica las hipótesis clásicas de las perturbaciones en un modelo lineal general.

Sea a un umbral para la variable endógena; nuestro problema consiste en obtener una estimación de β que sea óptima en el sentido de que nos permita valorar, de forma adecuada, la probabilidad de que u rebase dicho umbral.

Siguiendo las ideas que ya expusimos en Rojo (1987) para problemas generales de decisión, vamos a definir la función de riesgo para una elección, $\tilde{\beta}$ de β,

$$F(\tilde{\beta}; y, X) = \int_{R^k} [p(u \geq a \mid \beta) - p(u \geq a \mid \tilde{\beta})]^2 \; p(\beta \mid y, X) \; d\beta ,$$

donde

$$p(u \geq a \mid \beta) = \int_a^{\infty} (\sigma \sqrt{(2\pi)})^{-1} \; exp \; \{ - (2\sigma^2)^{-1} (u - X^o\beta)^2 \} \; du ,$$

lo mismo para $\tilde{\beta}$ y $p(\beta \mid y, X)$ es una densidad a posteriori para β condicionada por los datos que, suponiendo una distribución a priori para β difusa o no informativa, neutral en el sentido de Jeffreys (Box y Tiao (1973), págs. 41 y sigs.) es normal, de media $\hat{\beta}$ y matriz de covarianzas $\sigma^2(X'X)^{-1}$ (ver, por ejemplo, Zellner (1971)), pudiéndose escribir

$$p(\beta \mid y, X) = (2\pi\sigma^2)^{-k/2} \mid X'X \mid^{1/2} exp \{ - (2\sigma^2)^{-1} (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \}$$

siendo $\hat{\beta}$ el estimador MCO, $\hat{\beta} = (X'X)^{-1} X'y$.

El estimador de Bayes, que es el estimador buscado, se obtendrá minimizando F con respecto a $\tilde{\beta}$, que es lo que vamos a hacer a continuación.

3. ESTIMACION

Puede comprobarse sin dificultad, aunque los cálculos son algo tediosos, que son posibles las derivaciones bajo el signo integral y que, en consecuencia,

$$\partial F/\partial \tilde{\beta}_i = \int_{R^k} -2\,[p(u \geq a \mid \beta) - p(u \geq a \mid \tilde{\beta})]\,\frac{\partial\, p(u \geq a \mid \tilde{\beta})}{\partial\, \tilde{\beta}_i}\, p(\beta \mid y, X)\, d\beta$$

$$= -2\,\frac{\partial\, p(u \geq a \mid \tilde{\beta})}{\partial\, \tilde{\beta}_i} \int_{R^k} [p(u \geq a \mid \beta) - p(u \geq a \mid \tilde{\beta})]\; p(\beta \mid y, X)\, d\beta$$

Se comprueba que la derivada parcial que está en el último miembro de esta cadena de igualdades es diferente de cero salvo que x_i^o sea igual a cero (Obviamente, en este caso $p(u \geq a \mid \tilde{\beta})$ no depende de $\tilde{\beta}_i$). Supondremos que x_i^o es diferente de cero para todo i. En consecuencia, para resolver la ecuación $\partial F/\partial \tilde{\beta}_i = 0$, basta con resolver

$$\int_{R^k} p(u \geq a \mid \beta)\, p(\beta \mid y, X)\, d\beta = p(u \geq a \mid \tilde{\beta}). \qquad [1]$$

Calculemos el primer miembro de [1].

$$\int_{R^k} p(u \geq a \mid \beta)\, p(\beta \mid y, X)\, d\beta =$$

$$= \int_{R^k} \int_a^\infty (\sigma \sqrt{(2\,\pi)})^{-1}\, exp\{-(2\,\sigma^2)^{-1}\,(u - X^o\,\beta)^2\}\; du \cdot$$

$$\cdot\, (2\,\pi\,\sigma^2)^{-k/2}\, |X'X|^{1/2}\, exp\{-(2\,\sigma^2)^{-1}\,(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\}\; d\beta =$$

$$= \int_a^\infty (\sigma \sqrt{(2\,\pi)})^{-k-1}\, |X'X|^{1/2} [\int_{R^k} exp\{(-2\,\sigma^2)^{-1} [(u - X_o\,\beta)^2 +$$

$$+ (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})]\}\, d\beta]\; du =$$

$$= \int_a^\infty (\sigma \sqrt{(2\,\pi)})^{-k-1}\, |X'X|^{1/2}\, exp\{SSE / 2\,\sigma^2\} \cdot$$

$$\cdot \int_{R^k} exp\{-(2\,\sigma^2)^{-1}\,(w - Z\beta)'(w - Z\beta)\}\, d\beta\, du \qquad [2]$$

donde $SSE = (y - X\hat{\beta})'(y - X\hat{\beta})$, $w = (y', u)'$ y $Z = (X', X^{o\prime})'$.

Si llamamos $\hat{\beta}^o$ al estimador MCO de β en el modelo ampliado es decir $\hat{\beta}^o = (Z'Z)^{-1} Z'w$, se comprueba con facilidad que

$$(w - Z\beta)'(w - Z\beta) = SSE^o + (\beta - \hat{\beta}^o)'Z'Z(\beta - \hat{\beta}^o),$$

donde

$$SSE^o = (w - Z\hat{\beta}^o)'(w - Z\hat{\beta}^o) .$$

En consecuencia, [2] vale

$$\int_a^\infty (\sigma \sqrt{(2\pi)})^{-k-1} \mid X'X \mid^{1/2} exp \{ - (2\sigma^2)^{-1} (SSE^o - SSE) \} \cdot$$

$$\cdot \int_{R^k} exp \{ - (2\sigma^2)^{-1} (\beta - \hat{\beta}^o)' Z' Z (\beta - \hat{\beta}^o) \} \, d\beta du =$$

$$= \int_a^\infty \mid X'X \mid^{1/2} \mid Z'Z \mid^{-1/2} (\sigma \sqrt{(2\pi)})^{-1} exp \{ - (2\sigma^2)^{-1} (SSE^o - SSE) \} \, du \cdot \qquad [3]$$

Calculemos SSE^o - SSE.

$$SSE^o = (y',u)(y',u)' - (y',u)(X',X^{o\prime}) [(X',X^{o\prime})(X',X^{o\prime})']^{-1} (X',X^{o\prime})(y',u)' ,$$

y

$$SSE = y'y - y'X(X'X)^{-1} X'y .$$

Por tanto,

$$SSE^o - SSE = y'X(X'X)^{-1} X'y + u^2 -$$

$$- (y'X + uX^o)(X'X + X^{o\prime}X^o)^{-1} (X'y + X^{o\prime}u) =$$

$$= u^2 [1 - X^o(X'X + X^{o\prime}X^o)^{-1} X^{o\prime}] -$$

$$- 2u [X^o(X'X + X^{o\prime}X^o)^{-1} X'y] +$$

$$+ y'X(X'X)^{-1} X'y - y'X(X'X + X^{o\prime}X^o)^{-1} X'y . \qquad [4]$$

La expresión [4] es el cuadrado de un polinomio de primer grado en u. En efecto, si escribimos [4] en la forma eu^2-$2fu$+g, entonces,

$$e \cdot g = [1 - X^o(X'X + X^{o\prime}X^o)^{-1} X^{o\prime}] \cdot$$

$$\cdot y'X(X'X)^{-1} - (X'X + X^{o\prime}X^o)^{-1}) X'y \qquad [5.a]$$

y

$$f^2 = y'X(X'X + X^{o\prime}X^o)^{-1} X^{o\prime}X^o(X'X + X^{o\prime}X^o)^{-1} X'y . \qquad [5.b]$$

Ahora bien, si A, B y C son matrices de las mismas dimensiones, y existen las inversas de A, C y (A + B),

$$A^{-1} - C^{-1} = A^{-1} (C - A) C^{-1},$$

y por tanto,

$$A^{-1} - (A + B)^{-1} = A^{-1} B(A + B)^{-1}, \qquad [6.a]$$

y

$$(A + B)^{-1} = A^{-1} - A^{-1} B(A + B)^{-1} . \qquad [6.b]$$

Aplicando [6.a] y [6.b] a [5.a] y [5.b] respectivamente, obtendremos

$$e \cdot g = [1 - X^o(X'X + X^{o\prime}X^o)^{-1} X^{o\prime}] \cdot$$

$$\cdot y'X [(X'X)^{-1} X^{o\prime}X^o(X'X + X^{o\prime}X^o)^{-1}] X'y ,$$

y

$$f^2 = y'X [(X'X)^{-1} - (X'X)^{-1} X^{o\prime}X^o(X'X + X^{o\prime}X^o)^{-1}] \cdot$$
$$\cdot X^{o\prime}X^o(X'X + X^{o\prime}X^o)^{-1} X'y =$$

$$= y'X(X'X)^{-1} X^{o\prime}X^o(X'X + X^{o\prime}X^o)^{-1} X'y -$$

$$- y'X(X'X)^{-1} X^{o\prime}X^o(X'X + X^{o\prime}X^o)^{-1} X^{o\prime}X^o(X'X + X^{o\prime}X^o)^{-1} X'y .$$

En consecuencia, $f^2 = e \cdot g$, como puede comprobarse fácilmente sin más que tener en cuenta que las matrices $X^o(X'X + X^{o\prime}X^o)^{-1} X^{o\prime}$, $y'X(X'X)^{-1} X^{o\prime}$ y $X^o(X'X + X^{o\prime}X^o)^{-1} X'y$ conmutan por ser matrices cuadradas de orden uno. [4] es, pues, el cuadrado de un polinomio de primer grado en u.

Podemos escribir, por tanto,

$$SSE^o - SSE = e(u - f/e)^2 .$$

Si hacemos entonces en [3] el cambio de variable (*)

$$v = (u - f/e) \sqrt{e}$$

(*) El número $e = 1 - X^o(X'X + X^{o\prime}X^o)^{-1} X^{o\prime}$ es positivo, ya que

$$e = 1 - X^o(X'X)^{-1} X^{o\prime} + X^o(X'X)^{-1}X^{o\prime}X^o(X'X + X^{o\prime}X^o)^{-1} X^{o\prime} =$$
$$= 1 - X^o(X'X)^{-1} X^{o\prime}e ,$$

luego

$$e = 1 / (1 + X^o(X'X)^{-1} X^{o\prime}) > 0$$

obtenemos

$$| X'X |^{1/2} | Z'Z |^{-1/2} (\sigma \sqrt{(2\pi)}) e^{-1/2} \int_{(a - f/e)\sqrt{e}}^{\infty} exp \{ - (2\sigma^2)^{-1} v^2 \} dv =$$

$$= | X'X |^{1/2} | Z'Z |^{-1/2} e^{-1/2} p[N (0,1) \geq e^{1/2} (a - f/e) / \sigma] .$$

En consecuencia, la ecuación [1] se convierte en

$$p[N (0,1) \geq (a - X^o \tilde{\beta}) / \sigma] =$$

$$[| X'X | / (e | Z'Z |)]^{1/2} p[N (0,1) \geq (a - f/e) \sqrt{e} / \sigma] \qquad [5]$$

que permite obtener $X^o\tilde{\beta}$ utilizando los datos y las tablas de la N(0,1).

Finalmente, se comprueba sin dificultad que la predicción obtenida minimiza realmente la función de riesgo, es decir que la derivada segunda de F con respecto a $X^o\tilde{\beta}$ es positiva.

Obviamente, la predicción obtenida es óptima en el sentido de que minimiza la función de riesgo para la distribución a priori neutral elegida. Demostraremos a continuación que converge en probabilidad hacia $X^o\beta$.

Proposición

Si $\lim_{T \to \infty} (X'X)/T = \underline{Q}$ es no singular, la predicción obtenida converge en probabilidad a $X^o\beta$.

Demostración.

En efecto; con las notaciones anteriores,

$$f = X^o(Z'Z)^{-1} X'y = X^o((Z'Z)/T)^{-1} ((X'X)/T)\beta + X^o((Z'Z)/T)^{-1} X'\varepsilon/T .$$

Ahora bien,

$$\lim_{T \to \infty} ((Z'Z)/T) = \lim_{T \to \infty} ((X'X)/T) + \lim_{T \to \infty} ((X^{o\prime}X^o)/T) = \underline{Q},$$

y $\operatorname{plim}_{T \to \infty} X'\varepsilon/T = 0$.

NOTA: Obsérvese que la función de pérdida no es sensible al estimador de β, sino que únicamente lo es a la predicción. La condición [5] determina tan sólo $X^o\tilde{\beta}$, determinación que es compatible con un continuo de vectores $\tilde{\beta}$.

Luego

$$\underset{T\to\infty}{plim}\ f = X^o\underline{Q}^{-1}\underline{Q}\beta + X^o\underline{Q}^{-1}\cdot 0 = X^o\beta\ .$$

Además,

$$\underset{T\to\infty}{lim}\ e = 1 - \underset{T\to\infty}{lim}\ (1/T)\ X^o\ (\ (Z'Z)/T)^{-1}\ X^{o\prime} = 1\ ,$$

y, puesto que el determinante es una función continua de sus argumentos,

$$\underset{T\to\infty}{lim}\ [\ |X'X|\ /\ (e\ |Z'Z|)\ ]^{1/2} = 1\ .$$

Como $p[\ N\ (0,1)\ \geq x]$ es continua en x, tomando límites probabilísticos en [5], obtenemos

$p[\ N\ (0,1)\ \geq (a - plim\ X^o\tilde{\beta})\ /\ \sigma\ ] = p[\ N\ (0,1)\ \geq (a - X^o\beta)\ /\ \sigma]$,

y por tanto $plim\ X^o\tilde{\beta} = X^o\beta$.

5. CONCLUSIONES

En este trabajo, hemos obtenido una predicción de la variable endógena en un modelo lineal general, que goza de propiedades de convergencia en probabilidad a la predicción (que es una variable aleatoria) y optimalidad, entendida esta última en el sentido del criterio descrito en el apartado 2, es decir, como predicción a partir de un estimador de B es para una distribución a priori neutral en el sentido de Jeffreys. Obviamente, y aunque es intención de los autores valorar más detenidamente esta cuestión en un trabajo posterior, presenta ciertas desventajas con respecto a las predicciones que se obtienen a partir del estimador MCO: su dependencia del umbral, y una mayor complejidad en los cálculos. Consideramos, sin embargo, que posee una innegable ventaja, que es su sensibilidad al problema de los umbrales, si éste es el que nos preocupa. Como decíamos, nos planteamos realizar un estudio comparado de ambos estimadores, así como (y ésta es una extensión obvia), lo que ocurre cuando se supone que la varianza es desconocida, como es habitual.

BIBLIOGRAFIA

BOX, G. E. P. y TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Massachusetts. Addison-Wesley Pub.Co.

HECKMAN, J. J. (1981). *Statistical Models for Discrete Panel Data.* Structural Analysis of Discrete Data with Econometric Applications, Ed. Manski, F. y McFadden, D. Págs. 114 a 178. Cambridge, Massachusetts. The MIT Press.

ROJO GARCIA, J. L. (1987). *Estimación bayesiana con funciones de pérdida ligadas a la cola de una distribución.* Homenaje al Profesor Gonzálo Arnáiz Vellando. Ed. I.N.E. Madrid.

ZELLNER, A. (1971). *An introduction to Bayesian Inference in Econometrics.* New York. John Wiley & Sons Inc.

SUMMARY

PREDICTIONS OF ENDOGENOUS VARIALES FOR PROBLEMS RELATED TO THRESHOLDS.

In this paper we suggests a predictor for the endogenous variable of a general linear model, when the risk function depends on prediction errors in relation to surpassing a threshold. This predictor is a Bayes'one for a certain a priori distribution; moreover, it's consistent.

Key words: Econometrics, Decision functions, Qualitative Response models.

AMS 1980. Subject Classification: Primary 62H12 - Secondary 62P20.