Investigaciones Económicas (Segunda época). Vol. XI, n.º 1 (1987) págs. 25-39

CONTENIDO INFORMATIVO Y SELECCION DE MODELOS ECONOMETRICOS

Antonio AZNAR *
Universidad de Zaragoza

Se caracterizan dos estrategias de selección de modelos diferentes, según den o no entrada a una función de pérdida explícita en el proceso de selección. Se concretan los estadísticos a utilizar en ambas estrategias y se indican las posibles contradicciones resultantes de la aplicación de los mismos. Estos resultados se ilustran con un ejemplo en el que se explica la localización del sector farmacéutico.

1. Introducción

El objeto del presente trabajo es el análisis de una serie de puntos relevantes que afectan la selección de un modelo econométrico.

Se comienza en la sección 1 haciendo referencia a dos estrategias de selección diferentes, relacionando la primera con los procedimientos tradicionales y la segunda con otros procedimientos más recientes que incorporan explícitamente una función de pérdida. En la sección 2 se hace referencia a estos últimos procedimientos centrando el análisis en uno de los posibles objetivos: el análisis estructural. En la sección 3 se comparan ambas estrategias de selección poniendo de manifiesto las posibles contradicciones que puedan derivarse. En la sección 4 se lleva a cabo una aplicación de ambos procedimientos a datos de corte transversal. El trabajo termina con unas conclusiones.

2. Modelo verdadero y modelo funcionalmente verdadero

El punto de partida de la especificación de todo modelo econométrico suele ser una serie de observaciones sobre un grupo de variables; se supone un proceso generador de esos datos, en principio, no conocido. Lo que se llama llegar a

* Quisiera agradecer la inestimable ayuda que muchas personas me han prestado a la hora de elaborar este trabajo. Especialmente querría destacar la ayuda recibida de Jesús Mur Lacambra, quien llevó a cabo todo el trabajo del ordenador. María Teresa Aparicio, Ricardo Sanz y Javier Trivez me hicieron sugerencias y corrigieron errores en un primer manuscrito. Finalmente, un evaluador me propuso numerosas modificaciones que han sido en gran parte, incorporadas en la versión final. De todas formas, los errores que quedan sólo se me pueden atribuir a mi.

una especificación «satisfactoria» consiste en determinar qué estadísticos se van a utilizar para considerar que un modelo dado constituye una buena aproximación a ese proceso generador que nos es desconocido.

En este trabajo vamos a distinguir dos formas diferentes de llevar a cabo esa aproximación caracterizando la primera, con lo que llamaremos la búsqueda del modelo verdadero y la segunda con la búsqueda del modelo funcionalmente verdadero.

Lo que llamo búsqueda del modelo verdadero es un enfoque que basa la selección de modelos en la utilización de una serie de estadísticos paramétricos y no paramétricos, cuyo objetivo es resumir lo más eficientemente posible la evidencia de los datos. Se trata de basar la selección de un modelo en esta evidencia sin hacer referencia a ningún otro elemento ajeno a la misma. A modo de ilustración, veamos algunos de estos estadísticos:

No paramétricos:

$$R^2 \quad \text{y} \quad \bar{R}^2$$

Paramétricos: t-Student, F, o sus generalizaciones: Razón de verosimilitud, Wald y Lagrange, todos ellos con un nivel de significación ($\alpha = 0{,}01$ o $\alpha = 0{,}05$) independiente de la información muestral.

El segundo enfoque, que he llamado búsqueda del modelo funcionalmente verdadero, adopta una perspectiva diferente. No se trata de una simple aproximación a un proceso generador de los datos que nos es desconocido, sino de especificar un modelo para cumplir satisfactoriamente un objetivo establecido *a priori*. Para ello, no basta con los datos, sino que es necesario incorporar al proceso de selección de modelos algún tipo de función de utilidad o función de pérdida, en donde se refleje el grado de cumplimiento del objetivo establecido. Sirvan de ilustración las dos citas siguientes.

Kadane y Dickey (1980) escriben lo siguiente:

«Los enfoques que intentan explicar la especificación de un modelo en términos de la poco apropiada pregunta: ¿Es verdad que ...?, tienen en común que todos ellos funcionan sin referencia a la función de utilidad del científico...»

Nosotros exploramos en este trabajo las consecuencias de plantearnos lo que nos parece una pregunta más relevante: ¿Es útil suponer que...?, que nos exige ser más explícitos respecto a la función de utilidad de los investigadores.

Dhrymes et al. (1975), por su parte, escriben lo siguiente:

«En este contexto, la validación de un modelo se convierte en un proceso que depende del problema planteado, variando de caso a caso conforme lo hace el uso establecido para el modelo objeto de estudio. Así, un modelo particular puede ser validado para una aplicación concreta, pero no para otra. En cada caso, el proceso de validación está orientado a responder la pregunta: ¿Está este modelo cumpliendo la tarea propuesta?»

Creo que estas dos referencias sirven para situar perfectamente la nueva alternativa que he llamado: búsqueda del modelo funcionalmente verdadero. Se trata de especificar un modelo que sea útil y satisfaga las demandas que se fijen *a priori*.

¿Qué demandas se pueden hacer de un modelo econométrico? La literatura econométrica —ver Zellner y Palm (1974) e Intriligator (1978)— ha hecho referencia a tres fundamentales:

— Predicción.
— Análisis estructural.
— Control o evaluación de políticas.

Si se está interesado en especificar un modelo que cumpla, con satisfacción, alguno de los tres objetivos indicados habrá que pensar en incorporar, en el proceso mismo de especificación del modelo, algún tipo de función de utilidad o función de pérdida en la que se refleje el grado de cumplimiento, por cada modelo, del objetivo perseguido. Lo que pretendo mostrar en este trabajo es que si se sigue un proceso de selección de modelos en el cual se incorpora una función de pérdida, el modelo resultante puede ser muy diferente al que se obtendría siguiendo el proceso relacionado, con lo que hemos llamado búsqueda del modelo verdadero. Para ello, en la sección siguiente voy a describir lo que sería un proceso tipo de selección de modelos en el que se incorpora una función de pérdida. Me centraré en el análisis estructural entendiendo este como una aproximación cuantitativa a ciertos coeficientes, elasticidades o propensiones que nos dan cuenta de las interrelaciones entre las variables del sistema; este tipo de análisis es el que parece tener un mayor interés en los estudios con datos de corte transversal aunque los resultados a los que se llega son fácilmente generalizables a los otros dos objetivos enunciados. Ver, para ello, Aznar (1984 y 1986-a) y Aparicio (1985).

3. Selección del modelo funcionalmente verdadero: Análisis estructural

En el marco del modelo lineal general, suponer los dos modelos siguientes:

$$\begin{aligned} M_1&: \; y = \beta_{1i}x_i + X_1\beta_1 + u_1 \\ M_2&: \; y = \beta_{2i}x_i + X_2\beta_2 + u_2 \end{aligned} \qquad [1]$$

en donde:

y: vector de n observaciones de la variable dependiente.
x_i: vector de n observaciones de la variable i-ésima.
X_1: matriz $n \times k_1$ de observaciones de las variables incluidas en M_1.
X_2: matriz $n \times k_2$ de observaciones de las variables incluidas en M_2.
β_{1i}, β_{2i}: parámetros de la variable x_i en M_1 y M_2, respectivamente.

β_1, β_2: vectores de k_1 y k_2 parámetros, respectivamente.
u_1, u_2: vectores de perturbaciones aleatorias. Supongamos que son ruido blanco. Es decir:

$$Eu_j = 0 \quad \text{var}\,(u_j) = \sigma_j^2 I \quad j = 1, 2$$

Suponemos, además, que siguen una distribución normal.

Vemos que ambos modelos incluyen la misma variable, x_i. El objetivo de análisis estructural, hemos dicho anteriormente, es proporcionar información cuantitativa acerca del parámetro de la variable x_i. Se trata entonces de elegir aquel modelo que mayor evidencia proporcione sobre este parámetro.

Formulando así el problema, la cuestión que queda pendiente es: ¿Qué significa cumplir de la forma más satisfactoria el objetivo establecido *a priori* de aproximarnos al parámetro de la variable x_i?

Suponer que $\hat{\beta}_{1i}$ y $\hat{\beta}_{2i}$ son los estimadores *MCO* de β_{1i} y β_{2i}, respectivamente, utilizando M_1 y M_2. Ambos, en principio, con los supuestos enunciados, son estimadores insesgados. Podíamos responder a la pregunta anterior diciendo que el modelo que cumple el objetivo de la forma más satisfactoria es aquel que minimiza:

$$|\hat{\beta}_{ji} - \beta_{ji}|, \quad j = 1, 2 \qquad [2]$$

Pero $\hat{\beta}_{ji}$ son variables aleatorias y β_{ji} son constantes, por lo que la comparación hay que establecerla en términos probabilísticos y no como aparece en [2].

Con los supuestos indicados, es bien conocido que:

$$\begin{aligned} \hat{\beta}_{1i} &\sim N(\beta_{1i}, \sigma_1^2 Q_{11}) \\ \hat{\beta}_{2i} &\sim N(\beta_{2i}, \sigma_2^2 Q_{22}) \end{aligned} \qquad [3]$$

en donde Q_{11} y Q_{22} son los elementos correspondientes de las siguientes matrices:

$$\left[\begin{array}{cc|cc} x_i' & x_i & x_i' & X_1 \\ \hline X_1' & x_i & X_1' & X_1 \end{array}\right]^{-1} \qquad \left[\begin{array}{cc|cc} x_i' & x_i & x_i' & X_2 \\ \hline X_2' & x_i & X_2' & X_2 \end{array}\right]^{-1}$$

Se ve fácilmente que (Theil (1971)):

$$\begin{aligned} Q_{11} &= \frac{1}{(1 - R_{1i}^2)x_i'x_i} \\ Q_{22} &= \frac{1}{(1 - R_{2i}^2)x_i'x_i} \end{aligned} \qquad [4]$$

en donde R_{ji}^2 es el coeficiente de determinación en la regresión de x_i sobre las variables incluidas en el modelo M_j, $j = 1, 2$.

Sea ahora ε una masa de probabilidad dada y sean $N_1(\varepsilon)$ y $N_2(\varepsilon)$ dos valores definidos como:

$$Pr\{|\hat{\beta}_{1i} - \beta_{1i}| > N_1(\varepsilon)\} = \varepsilon$$

y

$$Pr\{|\hat{\beta}_{2i} - \beta_{2i}| > N_2(\varepsilon)\} = \varepsilon$$

Entonces decimos que el modelo M_1 nos informa más satisfactoriamente sobre el coeficiente de x_i que el modelo M_2 si se cumple que:

$$N_1(\varepsilon) < N_2(\varepsilon) \qquad [5]$$

y esto es equivalente a:

$$\text{var } (\hat{\beta}_{1i}) < \text{var } (\hat{\beta}_{2i})$$

o bien:

$$\frac{\sigma_1^2}{(1 - R_{1i}^2)} < \frac{\sigma_2^2}{(1 - R_{2i}^2)}$$

como σ_1^2 y σ_2^2 no son conocidos, se sustituyen por sus estimadores *MCO*, resultando un proceso de selección en el que nos quedamos con el modelo M_1 frente al modelo M_2 si:

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} < \frac{1 - R_{1i}^2}{1 - R_{2i}^2} \qquad [6]$$

Hemos perfilado un método de selección de modelos que depende del objetivo que el investigador persigue. En todo el proceso seguido hay una función de pérdida implícita que podemos escribir ahora como:

$$(\hat{\beta}_{ji} - \beta_{ji})^2, \quad j = 1, 2$$

y, en términos de esta función de pérdida y de la correspondiente función de riesgo, se obtiene el criterio de selección escrito en [6].

Podríamos precisar más y determinar no sólo cómo elegir un modelo entre varios disponibles, sino si un modelo es o no útil en términos absolutos para el fin propuesto. ¿Qué significa que un modelo es útil para el fin propuesto? En el marco estocástico en el que nos movemos y para el objetivo del análisis estructural establecido, entiendo que un modelo es útil si la diferencia $|\hat{\beta}_{ji} - \beta_{ji}|$ es «pequeña» en un porcentaje «elevado» de todos los posibles resultados. Sólo resta concretar lo que se entiende por «pequeña» y «elevado». En principio, el sentido de estos términos lo debe fijar el usuario del modelo, es decir, la persona

que va a valorar la utilidad del modelo. Así, puede decir que lo de «pequeño» y «elevado» lo concreta en dos indicadores, d_0 y ε_0 y concluye diciendo que un modelo cualquiera le es útil sólo si:

$$Pr\{|\hat{\beta}_{ji} - \beta_{ji}| > d_0\} \leqslant \varepsilon_0 \qquad [7]$$

es decir, si le garantiza que el error cometido será menor que d_0 con una probabilidad igual a $1 - \varepsilon_0$. El que sea o no posible alcanzar estas dos cotas dependerá de la información muestral disponible y, como consecuencia, del tamaño de la varianza del estimador.

En el proceso de selección habría que pensar en otra etapa en la que se compararan la estimación concreta realizada y el verdadero valor del parámetro. Pero esto, que en la predicción es factible, pues el valor de la variable que no es conocido en un momento dado sí que lo es en otro posterior, en el caso del análisis estructural no lo es porque el valor verdadero del parámetro nunca se puede conocer.

4. Búsqueda del modelo verdadero *versus* búsqueda del modelo funcionalmente verdadero

El objetivo de la presente sección es demostrar que los procedimientos de selección de modelos diseñados para lo que hemos llamado búsqueda del modelo verdadero, pueden llegar a resultados muy diferentes a los que se llegaría utilizando el procedimiento descrito en la sección anterior para la búsqueda del modelo funcionalmente verdadero. En concreto, voy a hacer referencia a dos procedimientos de selección, muy utilizados en la práctica y los voy a comparar con el criterio descrito en [6]. Las dos estrategias son:

a) Elige aquel modelo que maximiza el $\bar{R}^2$.
b) Elige aquel modelo para el que los valores del estadístico *t*-Student indican rechazar la hipótesis nula de que los coeficientes son cero supuesto un nivel de significación.

Comparemos ahora estas dos estrategias con la formulada en la sección anterior.

a) Maximización del $\bar{R}^2$

El coeficiente $\bar{R}^2$ para el modelo M_j es:

$$\bar{R}_j^2 = 1 - \frac{\hat{\sigma}_j^2}{\sum (y - \bar{y})^2/n - 1}$$

La maximización de $\bar{R}_j^2$ equivale a la minimización de $\hat{\sigma}_j^2$. Entre dos modelos dados, M_1 y M_2, seleccionamos M_1 si:

$$\hat{\sigma}_1^2 < \hat{\sigma}_2^2$$

o bien, si:

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} < 1 \qquad [8]$$

Comparando [8] con [6] vemos que sólo coinciden cuando, en esta última expresión numerador y denominador coinciden. Es decir, cuando el grado de colinieal idad que afecta a la variable x_i es el mismo en ambos modelos. Un caso extremo es cuando dicha variable es ortogonal al resto de las variables en ambos modelos, en cuyo caso el término de la derecha de [6] es 1. Por lo tanto, podemos decir que: sólo excepcionalmente ambos procedimientos coincidirán y llegarán a la misma conclusión.

b) VALORES DEL ESTADÍSTICO t

Vamos a restringir ahora el análisis al caso de modelos anidados. Vamos a suponer que M_1 está anidado en M_2, de forma que las variables incluidas en M_1 es un subconjunto de las variables incluidas en M_2, así que, en éste, aparecen variables que no aparecen en M_1.

Vamos a ver, primeramente, que el criterio escrito en [6] es equivalente a la aplicación del test-F, para contrastar la hipótesis nula de que los coeficientes de las variables que aparecen en M_2, pero no en M_1, son conjuntamente cero, con un nivel de significación que depende de la información muestral. Para ello, multiplicamos a [6] por $n - k_1 - 1$ y lo dividimos por $n - k_2 - 1$ y restamos la unidad a cada término resultando:

$$\frac{(n - k_1 - 1)\,\hat{\sigma}_1^2 - (n - k_2 - 1)\hat{\sigma}_2^2}{\hat{\sigma}_2^2} <$$

$$< \frac{(n - k_1 - 1)(1 - R_{1i}^2) - (n - k_2 - 1)(1 - R_{2i}^2)}{1 - R_{2i}^2}$$

Se ve cómo el término de la izquierda de la desigualdad es el contraste de la F para la hipótesis nula indicada y que el punto crítico que aparece a la derecha depende de la información muestral, lejos, en general, del que correspondería a un nivel de significación constante. Por lo tanto, también en este caso, sólo excepcionalmente coincidirán los resultados derivados de la aplicación de los dos procedimientos indicados.

¿Cuáles son las conclusiones de todo este análisis? La conclusión más importante es que si se desea especificar un modelo para cumplir un objetivo determinado, en este caso el análisis estructural, la referencia a ese objetivo hay que introducirla en el proceso de selección de modelos mediante la correspondiente función de pérdida. Como consecuencia de esto, si la selección del modelo se realiza siguiendo los procedimientos tradicionales que no prestan atención explícita a ninguna función de pérdida, hay una alta probabilidad de realizar una selección inapropiada.

La generalización del contenido de esta sección, al caso de un vector de coeficientes, puede verse en Aparicio-Aznar (1985).

5. Aplicación empírica

Para llevar a cabo la aplicación empírica he utilizado la información contenida en Mur (1986) en un trabajo en el que se intenta explicar la localización del sector farmacéutico en el Estado español. Se dispone de diecisiete observaciones, una para cada una de las comunidades autónomas. La variable dependiente es el crecimiento de la producción del sector farmacéutico entre los años 1978 y 1982 y las variables explicativas se refieren al año 1982.

Con base en la teoría del ajuste de perfiles de un sector y de las regiones, a la que se hace amplia referencia en el trabajo citado de Mur, se hace una primera especificación de veinticuatro variables agrupadas en cinco perfiles: perfil de atracción, perfil clásico, perfil ambiental, perfil socio-cultural y perfil político. Las fuentes estadísticas de donde se obtuvieron los datos de todas las variables estudiadas en cada uno de los perfiles, pueden verse en el Apéndice 2.

En el perfil de atracción se incluyen todas aquellas variables que se refieren a sectores que mantienen una relación, por la oferta o la demanda, con el sector farmacéutico. Como indicadores de oferta, se han incluido nueve variables que recogen las producciones de nueve sectores químicos relacionados con dicho sector. Como indicadores de demanda, se han utilizado seis variables, que van desde el número de médicos colegiados, la población total y el número de trabajadores afiliados al régimen general de la Seguridad Social.

El perfil clásico se refiere a la disponibilidad de mano de obra, capital, recursos energéticos así como la abundancia y características de terrenos y agua. Se tomaron seis variables como indicadores de estos factores.

El perfil socio-cultural se refiere a las variables relacionadas con la infraestructura social; se tomaron dos variables, una indicativa del grado de urbanización y otra el número de profesores universitarios como indicador del nivel de investigación.

Por último, en lo que se refiere al perfil político, se tomó una variable definida como el número de funcionarios superiores y directivos de empresa.

Teniendo en cuenta el reducido número de observaciones disponibles, $n = 17$, lo primero que había que hacer era reducir el número de variables incurriendo en la menor pérdida posible de la información contenida en los datos. Para ello, se hizo uso de técnicas multivariantes, análisis de componentes principales y análisis discriminante, para determinar las dimensiones de los datos y determinar cómo reducir el número de variables sin perder apenas información.

Se aplicó un análisis de componentes diferente a las variables incluidas en cada uno de los perfiles, obteniéndose, para cada uno de ellos, lo que puede considerarse la variable representativa del mismo, que era aquella variable más rela-

cionada con el componente de mayor varianza, es decir, el primero. A continuación se formaron dos grupos con las observaciones, incluyendo en el primero las comunidades con un valor del coeficiente de especialización en el sector farmacéutico superior a uno y, en el otro, las comunidades a las que corresponde un valor inferior a uno; tomando diferentes grupos de variables y mediante el análisis discriminante, se determinaron aquellas que daban cuenta de un mayor poder discriminatorio entre los dos grupos definidos en la forma indicada.

Como resultados de este análisis se obtuvieron cuatro variables que daban cuenta de la mayor parte de la información contenida en los datos. Las variables fueron:

- x_1: Profesores universitarios
- x_2: Producción de bienes y servicios para la venta del sector químico
- x_3: Total de camas hospitalarias
- x_4: Funcionarios superiores y directores de empresa

Llegados a este punto, se pensó en especificar un modelo que permitiera conocer el efecto de la variable que reflejaba el nivel de investigación de cada comunidad autónoma, x_1, sobre la producción del sector farmacéutico. Para ello, se consideraron todos aquellos modelos en los que aparecía dicha variable.

El primer paso que se dió es aportar alguna evidencia de que la perturbación aleatoria cumplía las hipótesis mencionadas anteriormente. Se utilizó un contraste de autocorrelación espacial basado en la aplicación del estadístico de Lagrange para contrastar la hipótesis nula de ausencia de correlación espacial entre las comunidades contiguas de primer orden. La derivación de este contraste puede verse en Aznar (1986.b)[1]. El contraste puede escribirse de la siguiente forma:

$$SCLM = \frac{\hat{u}' P^* \hat{u}}{\hat{u}' \hat{u}} \times \frac{n^2}{Q}$$

en donde $\hat{u}$ es el vector de residuos *MCO*, P^* es una matriz de ceros y unos que depende de las relaciones de vecindad entre las comunidades, n es el número de observaciones y Q es la suma de vecindades de orden uno que afectan a cada una de las unidades. Se distribuye como una χ^2 con un grado de libertad.

Los resultados obtenidos para todos los posibles modelos en los que aparece la variable x_1 pueden verse en el cuadro 1. Los valores que toma el estadístico *SCLM* indican ausencia de autocorrelación en todos los casos. Si se siguiera una estrategia de selección de modelos según la práctica tradicional, parece haber poca duda de que el modelo seleccionado sería aquel en el que aparecen las cuatro variables ya que le corresponde el mayor valor del coeficiente $\bar{R}^2$ y los valores de la t para todas las variables indican que los coeficientes son diferentes

[1] Véase Apéndice 1.

CUADRO 1
Resultados de la aplicación empírica

Modelo	$\bar{R}^2$	SCLM	Varianza estimada del estimador del coeficiente de x_1	Valores de las estimaciones de los coeficientes de las variables y del estadístico t-student				
				c	x_1	x_2	x_3	x_4
x_1, x_2, x_3, x_4	0,9551	0,00205	0,4359	9.800,8 (2,04)	3,1127 (4,71)	0,2387 (12,80)	−0,9461 (−4,43)	−7,7943 (−1,95)
x_1, x_2, x_3	0,9453	0,0217	0,5222	771,65 (0,55)	2,9493 (4,08)	0,2333 (11,47)	−0,9492 (−4,03)	
x_1, x_2, x_4	0,8107	0,2672	0,3634	6.289,3 (0,8538)	0,7375 (1,22)	0,1938 (7,93)		−7,9286 (−1,27)
x_1, x_3, x_4	0,3926	1,2239	5,7801	−4.363,7 (−0,25)	1,8904 (0,78)		0,5376 (0,81)	−0,2902 (−0,02)
x_1, x_2	0,8858	0,6647	0,3601	−2.907,9 (−1,91)	0,5632 (0,93)	0,1882 (7,66)		
x_1, x_3	0,4360	1,4190	5,3056	−4.695,2 (−1,11)	1,8853 (0,81)		0,5362 (0,84)	
x_1, x_4	0,4071	1,1866	1,2559	−3.680,4 (−0,21)	3,6196 (3,22)			0,9555 (0,06)
x_1	0,4429	1,3783	13,2576	−2.570,8 (−0,76)	3,6517 (3,72)			

de cero, suponiendo un nivel de significación del 5 por 100 (la variable x_4 requiere un nivel ligeramente superior para ser significativa). Si cambiamos de estrategia y nos quedamos con el modelo en el que la varianza del estimador del parámetro de x_1 es menor, entonces el modelo elegido sería aquel en el que aparecen las variable x_1 y x_2. Sea M_1 el modelo que incluye las cuatro variables y M_2 el que incluye las dos variables x_1 y x_2. Puede verse, directamente, que M_1 garantiza que la diferencia entre las estimaciones del parámetro de x_1 y su valor verdadero será menor que 1,4322 el 95 por 100 de las veces; el modelo M_2, con el mismo porcentaje, garantiza que la diferencia será menor que 1,284. ¿Sirven para algo estos modelos? Pues depende de las exigencias que el usuario establezca. Supongamos que el usuario dice que el modelo sólo le resulta útil si le garantiza que la desviación de las estimaciones proporcionadas por dicho modelo respecto al valor verdadero del parámetro no supera el 0,5 el 95 por 100 de las veces. Entonces está claro que ninguno de los modelos que aparecen en el cuadro 1 le sirven para nada. Si las exigencias fueran otras, uno o los dos modelos, podrían serle útiles.

A la vista de los valores que toman las estimaciones del parámetro de la variable x_1, tal como puede verse en el cuadro 1, cabría pensar en que es difícilmente aceptable el que todos los modelos sean insesgados. Esto habría que resolverlo en una etapa posterior, de corroboración, en la que se comparaban las estimaciones con el verdadero valor del parámetro. Pero el problema es que éste nunca va a conocerse. Esto me hace pensar en la necesidad de diseñar un proceso especial de acercamiento a estos parámetros, pues, de otro modo, las afirmaciones que se hacen a la hora de hacer un análisis estructural no son controlables *a posteriori* y, como consecuencia, son de dudosa utilidad. Cuando se hace una predicción para una variable cuyo valor no es conocido, cuando éste se conoce se pueden comparar la predicción y el verdadero valor y así comprobar si lo que afirmaba el modelo tenía o no sentido. Este control *a posteriori* de las conclusiones derivadas de todo modelo es una etapa necesaria en todo proceso de selección de modelos.

6. Conclusiones

En este papel se ha intentado demostrar que si, a la hora de especificar un modelo econométrico, se tienen en mente determinadas demandas que habrían de satisfacerse a partir de ese modelo, tales demandas deberían incorporarse en el propio proceso de selección del modelo con la correspondiente función de pérdida. Se ha hecho referencia a un proceso de selección en el que se da cabida a una función de pérdida y se ha demostrado que los resultados a los que se llega con este procedimiento pueden ser sustancialmente diferentes a los que se obtendrían utilizando otro tipo de procedimientos en los que no se incorpora una función de pérdida. Se presenta, finalmente, un ejemplo con el que se ilustran bastante claramente estas conclusiones.

Apéndice I. Contraste de correlación espacial

Suponer una muestra obtenida a partir de una población con función de distribución de probabilidad, $f(x/\theta)$, que depende de un vector de parámetros, θ. Tratamos de contrastar la siguiente hipótesis nula:

$$H_0: \quad h(\theta) = 0 \qquad [A1]$$

Sea $L(x/\theta)$ la función de verosimilitud. El estadístico de Lagrange para contrastar la hipótesis nula escrita en [A1] puede escribirse como:

$$LM = \left[\frac{\partial \log L(x/\theta)}{\partial \theta}\right]_{\theta=\theta_R} B(\tilde{\theta})^{-1} \left[\frac{\partial \log L(x/\theta)}{\partial \theta}\right]_{\theta=\theta_R} \qquad [A2]$$

en donde $B(\tilde{\theta}_R)^{-1}$ es la matriz de información y $\tilde{\theta}_R$ es el estimador máximo-verosimil restringido.

Considerar ahora el siguiente modelo:

$$y = X\beta + u \qquad [A3]$$

en donde y es $n \times 1$, X es $n \times k$ y u es $n \times 1$.

Supongamos que las observaciones corresponden a n regiones tomadas para un período. En este caso, los contraste habituales para detectar autocorrelación, como el Durbin-Watson, no son aplicables ya que estos requieren el carácter serial de las observaciones, que está ausente, por definición, en los datos de corte transversal.

Para presentar una alternativa, supongamos un esquema de correlación del siguiente tipo:

$$u = Pu + \varepsilon \qquad [A4]$$

en donde ε es un vector de variables que son ruido blanco y P es una matriz de parámetros que, en general, puede escribirse:

$$P = \begin{bmatrix} 0 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} \\ \rho_{21} & 0 & \rho_{23} & \cdots & \rho_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \cdots & 0 \end{bmatrix} \qquad [A5]$$

ρ_{ij} indica el efecto de la perturbación aleatoria de la región j sobre la región i. En general, tal como aparece en [A5], la obtención de un contraste sería impracticable, por lo que es necesario asumir hipótesis más restrictivas acerca del proceso de correlación espacial. Así, supondremos que todos los ρ_{ij} son

iguales y que sólo aparecen influencias entre regiones vecinas. Por ejemplo, si sólo se tratara de tres regiones del siguiente tipo:

la matriz P puede escribirse como:

$$P = \begin{bmatrix} 0 & \rho & 0 \\ \rho & 0 & \rho \\ 0 & \rho & 0 \end{bmatrix}$$

Podemos transformar el modelo escrito en [A3], obteniendo:

$$y^* = X^*\beta + \varepsilon \qquad [A6]$$

en donde:

$$y^* = (I - P)y \quad ; \quad X^* = (X - PX) \quad ; \quad \varepsilon = (I - P)u$$

A partir de [A6] la derivación de los términos que aparecen en [A2] es directa, obteniéndose así la expresión concreta que el contraste toma en el texto, teniendo en cuenta que P^* es la matriz P sustituyendo a ρ por 1. Los detalles pueden verse en la cita, Aznar (1986b), que aparece en el texto.

Apéndice II. Información estadística

Los valores de la variable dependiente y de las cuatro variables explicativas que se mantuvieron son las siguientes:

Regiones	Y^*	X_1^{**}	X_2^*	X_3^{**}	X_4^{**}
Andalucía	238	5.673	63.387	28.898	18.100
Aragón	1.573	1.800	13.674	7.790	6.700
Asturias	339	1.085	7.294	6.514	3.300
Baleares	34	249	533	3.865	2.600
Canarias	0	1.315	2.493	7.566	4.100
Cantabria	54	926	21.492	3.180	2.300
Castilla-León	5.431	3.009	26.599	16.020	5.800
Castilla-La Mancha	3.740	0	32.991	6.817	5.900
Cataluña	53.381	6.155	266.895	31.195	34.200
Valencia	752	2.600	35.714	15.067	22.300
Extremadura	132	687	1.113	4.268	2.200
Galicia	670	1.481	13.385	11.551	12.200
Madrid	27.123	9.903	96.684	27.499	32.300

Regiones	Y^*	X_1^{**}	X_2^*	X_3^{**}	X_4^{**}
Murcia	603	794	12.063	4.133	2.800
Navarra	1.270	1.123	6.474	3.782	3.100
País Vasco	1.201	1.620	58.564	12.518	13.500
La Rioja	56	0	1.521	2.068	1.100

* En millones.
** En unidades.

Las fuentes estadísticas fueron las siguientes:

Nueve variables: Valor de la producción en los siguientes sectores:

— Fabricación de productos químicos básicos.
— Industria del vidrio.
— Transformación del papel y del cartón.
— Transformación de materias plásticas.
— Sacrificio de ganado y preparación de carne.
— Fabricación de grasas y aceites, vegetales y animales.
— Industrias de alcohóles etílicos.
— Producción de bienes y servicios para la venta del sector químico.
— Fabricación de productos químicos destinados a la industria.

Fuente: *Censo Industrial.* Año 1982. INE.

Cinco variables:

— Población total.
— Médicos colegiados.
— Veterinarios colegiados.
— Farmacéuticos colegiados.
— Población en ciudades de más de 50.000 habitantes.

Fuente: *Anuario Estadístico.* Año 1983. INE.

Seis variables:

— Población activa.
— Desempleo.
— Empleo femenino industrial.
— Relación empleo industrial sobre población activa.
— Profesionales y técnicos.
— Funcionarios superiores y directores de empresa.

Fuente: *Encuesta de Población Activa.* Año 1982. INE.

Cuatro variables:

— Profesores universitarios.

Fuente: *Estadística de la Enseñanza en España.* Año 1985. INE.

— Costes salariales, por empleado, del sector químico.

Fuente: *Encuesta Industrial.* Año 1984. INE.

— Camas hospitalarias.

Fuente: *Estadística de Establecimientos Sanitarios.* Año 1982. INE.

— Trabajadores afiliados a la Seguridad Social.

Fuente: *Información Estadística, núm. 12.* Año 1983. Ministerio de Trabajo y Seguridad Social.

Referencias

Aparicio, M. T. (1985): «Selección de Modelos Econométricos: Estudio Comparado de un Nuevo Criterio». Tesis Doctoral. Universidad de Zaragoza.

Aparicio, M. T., y Aznar, A. (1985): «Estimación de un Subconjunto de Coeficientes: Una Aplicación del Criterio de la Varianza Estimada». II Reunión de Econometría. Zaragoza, 1-2 julio.

Aznar, A. (1984): «Buscando el Modelo Econométrico útil». *Estadística Española,* 102.

Aznar, A. (1986a): *Econometric Model Selection: A New Approach.* Universidad de Zaragoza.

Aznar, A. (1986b): «The Application of the Lagrange Multiplier to Test Spatial Correlation». CI-86-2.

Dhrymes, P. J., Howrey, E. P., Hymans, S. H., Kmenta, J., Leamer, E. E., Quandt, R. E., Ramsey, J. B., Shapiro, H. T., y Zarnowitz, V. (1975): «Criteria for Evaluation of Econometric Models», en *The Brookings Model: Perspective and Recent Developments.* Edi. by C. Fromm and L. R. Klein. North-Holland.

Intriligator, M. D. (1978): *Econometric Models, Techniques and Applications.* North-Holla·

Kadane, J. D., y Dickey, J. M. (1980): «Bayesian Decision Theory and the Simpli tion of Models», en *Evaluation of Econometric Models.* Edi.: by J. Kmenta and Ramsey. New York: Academic Press.

Mur Lacambra, J. (1986): «Tratamiento Econométrico de la Localización. La Industria Farmacéutica». Tesina leída en Junio. Universidad de Zaragoza.

Theil, H. (1971): *Principles of Econometrics.* North-Holland.

Zellner, A., y Palm, F. I. (1974): «Time Series Analysis and Simultaneous Equation Econometric Models». *Journal of Econometrics, 2,* 17-54.

Abstract

Two alternative econometric model selectión procedures are charecterised according to whether they consider explicitily or not a loss function in the selectión process. The methods and statistics to be used in both procedures are outlined indicating the posibility of contradictions among them. In order to illustrate these results the estimation and selectión of an econometric model for the spanish pharmaceutical sector is presented.

Recepción del original, octubre, 1986
Versión final, diciembre, 1986