QÜESTIIÓ, Vol 20, 2 pp. 147-166, 1996

ESTABILIDAD DE ALGUNOS CRITERIOS DE SELECCIÓN DE MODELOS

CARMEN GARCÍA OLAVERRI*

Universidad Pública de Navarra

En este artículo se comparan nueve criterios de selección de modelos. Se analiza si el número de modelos influye en la selección. Se estudia la estabilidad y la robustez de la selección. La comparación se lleva a cabo mediante simulación.

Stability of some model selection criteria

Keywords: Model selection criteria. Robustness.

1. INTRODUCCIÓN

La construcción de modelos estadísticos surge de la necesidad de explicar y predecir el comportamiento de fenómenos reales que dependen de distintas variables. Cuando para una misma evidencia muestral existen modelos alternativos surge el problema de la selección. ¿Cuál es el mejor modelo de todas las alternativas formuladas? ¿Tiene sentido seleccionar un modelo en función del uso posterior que se vaya a dar al mismo? ¿Es siempre necesaria una evaluación extramuestral? Para dar respuesta a las anteriores cuestiones se han definido en la literatura estadística distintos criterios de selección de modelos. Algunos de ellos son muy empleados y los paquetes estadísticos incluyen desde hace tiempo información de este tipo (es el caso de los criterios AIC (Akaike), Cp (Mallows), SBIC(Schwarz), cuyo calculo está incluido en el software de uso más común).

*Carmen García Olaverri. Departamento de Estadística e Investigación Operativa. Universidad Pública de Navarra. 31006. Pamplona.

La autora agradece las observaciones y sugerencias del Profesor C. Cuadras y de un evaluador anónimo.

-Article rebut l'octubre de 1994.

-Acceptat el setembre de 1995.

Los distintos métodos de selección de modelos han sido objeto de comparación en la literatura sin que exista una postura unánime sobre cual es la mejor forma de seleccionar el modelo óptimo. Gran parte de la controversia está basada en el hecho de que no todos los criterios han sido definidos con el mismo fin, es decir para todos los autores la idea de "mejor modelo"no es la misma. Sin embargo, parece claro que hay algunas propiedades deseables que todo criterio de selección debiera satisfacer; esto es, existen algunas formas objetivas de comparar los distintos criterios de selección y concluir cuál de ellos es el mejor , al menos en esa parcela. En este sentido, en la literatura estadística se ha comparado el comportamiento de los criterios de selección cuando cambia el tamaño muestral (Geweke y Meese, 1981), la estructura del proceso que generó los datos (Koehler y Murphree, 1988), el grado de colinealidad entre las variables o la distribución del término de error (Mills y Prasad, 1992; García Olaverri y Aznar, 1994).

Una cuestión que apenas se ha abordado ha sido la de analizar el comportamiento de los criterios según sea el número de modelos presentes en la comparación. Unicamente Geweke y Meese (1981) hacen referencia a esta cuestión, no tanto para estudiar como afecta a la selección el número de alternativas sino para poder comparar criterios aplicados a distintos tamaños muestrales.

Desde nuestro punto de vista, es relevante el saber si un determinado criterio proporcionará distintos resultados según sea el número de alternativas que se formulan en el proceso de selección. Es evidente que para ningún criterio es lo mismo seleccionar un modelo entre 2 alternativas que entre 14, pero parece natural que si se hacen conjeturas acerca de modelos muy alejados del verdadero Proceso Generador de Datos (PGD) su presencia o no en el proceso de selección no debiera provocar variaciones en la selección final.

En el presente trabajo presentamos un estudio comparado del comportamiento de distintos criterios de selección cuando se modifica el número de modelos presentes en la comparación siguiendo el siguiente esquema:

Imaginemos que se dispone de cierta evidencia muestral sobre una variable de nuestro interés (Y_t) y desconocemos cuál ha sido el modelo que la ha generado. Supongamos que se hacen distintas conjeturas sobre cuál es el modelo buscado:

$$\begin{aligned}
\mathbf{M}_1 &: \quad Y_t = \beta_1 X_{1t} + u_{1t} \\
\mathbf{M}_2 &: \quad Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + u_{2t} \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\mathbf{M}_k &: \quad Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + u_{kt}
\end{aligned}$$

donde u_{it}, es la perturbación aleatoria de cada modelo ($i = 1, \ldots, k$), que supondremos satisface las condiciones de esfericidad. Las variables X_{it} se supondrán no estocásticas y con ausencia de multicolinealidad exacta.

Supongamos ahora que con esa misma evidencia muestral modificamos el número de modelos presentes en la comparación, ampliándolo e incluyendo en la selección además de $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_K$, otros modelos $\mathbf{M}_{K+1}, \ldots, \mathbf{M}_P$ cada vez más alejados del verdadero PGD. La pregunta que nos hacemos es ¿Cambiarán en algo los criterios su forma de seleccionar por el hecho de haber incluido las nuevas alternativas ? ¿Hay algunos criterios que cambian más que otros?

Para dar respuesta a las anteriores cuestiones se ha realizado un exhaustivo ejercicio de simulación, del que mostramos algunos resultados, llegándose a la conclusión de que para algunos criterios el resultado final del proceso de selección es muy distinto según sea el número de modelos presentes en la comparación. Es decir, son criterios poco robustos pues si la selección se realiza en un ambiente de incertidumbre, los resultados de ésta serán bastante arbitrarios.

Los criterios de selección de modelos que vamos a comparar son los siguientes: $\overline{R}^2$ (Theil, 1961), Cp (Mallows, 1964), AIC (Akaike, 1973), CAT (Parzen, 1974), BIC (Sawa, 1978), SBIC (Schwarz, 1978), PC (Amemiya,1980), BEC (Geweke y Meese, 1981) y PEC (Aznar y García, 1993). Si la selección se realiza entre m^* modelos anidados: $\mathbf{M}_{m^*} \supset \cdots \supset \mathbf{M}_2 \supset \mathbf{M}_1$, para todos los criterios se trata de elegir aquel valor de $m (m = 1, \ldots, m^*)$, (es decir el modelo con m variables), que minimice las expresiones siguientes:

Tabla 1

Criterio	Elegir $m (m = 1, \ldots, m^*)$ que minimice:
$\overline{R}^2$:	$\hat{\sigma}_m^2 \cdot \dfrac{T}{T-m}$
Cp:	$\hat{\sigma}_m^2 + \dfrac{2m}{T-m^*} \cdot \hat{\sigma}_{m^*}^2$
AIC:	$\ln \hat{\sigma}_m^2 + \dfrac{2m}{T}$
BIC:	$\ln(\hat{\sigma}_m^2) + 2\left(\dfrac{m+2}{T-m^*}\right)\left(\dfrac{\hat{\sigma}_{m^*}^2}{\hat{\sigma}_m^2}\right) - \dfrac{2T}{(T-m^*)^2}\left(\dfrac{\hat{\sigma}_{m^*}^2}{\hat{\sigma}_m^2}\right)^2$
PC:	$\hat{\sigma}_m^2\left(\dfrac{T+m}{T-m}\right)$
SBIC:	$\ln(\hat{\sigma}_m^2) + m\dfrac{\ln T}{T}$
CAT:	$\sum_{j=1}^{m} \dfrac{(T-j)}{T^2} \cdot \hat{\sigma}_j^{-2} - \dfrac{(T-m)}{T} \cdot \hat{\sigma}_m^{-2}$
BEC:	$\hat{\sigma}_m^2 + m\hat{\sigma}_{m^*}^2 \cdot \dfrac{\ln T}{T-m^*}$
PEC:	$\left[\dfrac{1}{T_1}\sum_{p=1}^{T_1} \hat{\mathrm{Var}}(\hat{y}_{mp})\right] \cdot \left[\dfrac{1}{T_1}\sum_{p=1}^{T_1} e_{mp}^{*2}\right]$

Donde $\mathbf{M}_m$ y $\mathbf{M}_{m^*}$ indican modelos con m y m^* variables, siendo $\mathbf{M}_{m^*}$ el modelo más amplio de todos los presentes en la comparación; $\hat{\sigma}^2_m, \hat{\sigma}^2_{m^*}$ son los estimadores máximo verosímiles de la varianza del término de error en $\mathbf{M}_m$ y $\mathbf{M}_{m^*}$ respectivamente; T es el tamaño muestral.

En la expresión del criterio PEC, T_1 indica el número de observaciones que han quedado corroboradas (es decir, que pertenecen al intervalo de confianza de su correspondiente predicción); el segundo factor indica el error cuadrático medio de predicción que se comete al predecir, mediante el modelo de m variables, las T_1 últimas observaciones muestrales.

En las anteriores definiciones se observan grandes diferencias y, como consecuencia de ello en ocasiones se seleccionan distintos modelos según sea el criterio considerado. Como ya ha quedado indicado, estas diferencias son debidas a los distintos objetivos con que han sido definidos los criterios de selección. Por ejemplo los criterios SBIC y BEC han sido diseñados con el objetivo de ser consistentes (esto es, tender a seleccionar siempre el PGD a medida que el tamaño muestral aumenta), los criterios PC, Cp tienen como objetivo elegir aquel modelo que minimice el error cuadrático medio de predicción, el criterio AIC persigue seleccionar el modelo más próximo al PGD (en términos de la medida de Kullback - Leibler). Un estudio exhaustivo sobre los distintos objetivos en los procedimientos de selección puede verse en Aznar (1989).

En el problema que nos ocupa, entendemos que la influencia que pueda tener en un criterio el número de modelos en la comparación, va a depender en gran medida, de la tendencia que presente el criterio hacia la selección de modelos distintos al PGD. Es decir, si un criterio es consistente, en el sentido de que nunca tiende a elegir modelos distintos al PGD, el hecho de que en la comparación haya pocos o muchos modelos no afectará al resultado final. Por el contrario, si un criterio muestra tendencia, por ejemplo, hacia modelos más amplios que el PGD el hecho de que pueda elegir entre varias alternativas de este tipo modificará los resultados.

2. ANÁLISIS DE LA TENDENCIA HACIA MODELOS DISTINTOS AL PGD

En este apartado vamos a analizar cuál es la tendencia de los distintos criterios hacia modelos distintos al PGD y en base a los resultados formularemos hipótesis sobre el comportamiento de los criterios cuando cambian los términos de la comparación.

Por similitud en cuanto a la forma funcional estudiaremos separadamente los criterios «clásicos» y el criterio PEC. Consideremos aquellos en primer lugar. En

todos ellos se trata de elegir aquel modelo con m variables que minimice una expresión que siempre depende de $\hat{\sigma}_m^2$, estimador máximo verosímil de la varianza del término de error del modelo considerado.

Supongamos que los datos se han generado con el modelo de k variables $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}$; con $\boldsymbol{u} \sim N(0, \sigma_k^2 I)$ donde $\boldsymbol{y}$, y $\boldsymbol{u}$ son vectores $T \cdot 1$, β es un vector $k \cdot 1$ y $\boldsymbol{X}$ es una matriz $T \cdot k$.

Si consideramos otro modelo lineal de m variables: $\boldsymbol{y} = \boldsymbol{Z}\gamma + \boldsymbol{v}$; con $\boldsymbol{v} \sim N(0, \sigma_m^2 I)$ donde $\boldsymbol{y}$, y $\boldsymbol{v}$ son vectores $T \cdot 1$, γ es un vector $m \cdot 1$ y $\boldsymbol{X}$ es una matriz $T \cdot m$ y para este modelo calculamos $\hat{\sigma}_m^2 = \dfrac{\hat{v}'\hat{v}}{T}$ se tiene: $\hat{v}'\hat{v} = (\boldsymbol{y} - \boldsymbol{Z}\hat{\gamma}) = \boldsymbol{y}'\boldsymbol{M}_z\boldsymbol{y}$; con $\boldsymbol{M}_z = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ matriz idempotente; sustituyendo $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}$ se tiene:

$$E(\hat{\sigma}_m^2) = \sigma_k^2 \frac{(T-m)}{T} + \frac{\beta'\boldsymbol{X}'\boldsymbol{M}_z\boldsymbol{X}\beta}{T}$$

$$E(\hat{\sigma}_k^2) = \sigma_k^2 \frac{(T-K)}{T}$$

Si $m < k$, omisión de variables relevantes, siempre se cumple $E(\hat{\sigma}_m^2) > E(\hat{\sigma}_k^2)$ pues $\beta'\boldsymbol{X}'\boldsymbol{M}_z\boldsymbol{X}\beta$ es definida positiva, además $(T-m) > (T-k)$. Por lo tanto la presencia del estadístico $\hat{\sigma}_m^2$ penaliza fundamentalmente la selección de modelos anidados en el PGD; además, a medida que aumenta el tamaño muestral el resto de sumandos o factores que aparecen en la definición de los criterios son irrelevantes cuando se está comparando un modelo anidado en el PGD. Salvo en casos muy excepcionales como por ejemplo que el PGD tenga muchas variables o se disponga de muy reducido tamaño muestral, los denominados criterios clásicos apenas seleccionarán modelos anidados en el PGD. Por lo tanto podemos concluir que, para estos métodos, el número de modelos anidados en el PGD presentes en la comparación es, en general, irrelevante para la selección final.

Por el contrario si $m > k$, inclusión de variables irrelevantes, $E(\hat{\sigma}_m^2) = \sigma_k^2 \dfrac{(T-m)}{T}$ pues el sumando $\beta'\boldsymbol{X}'\boldsymbol{M}_z\boldsymbol{X}\beta$ se anula (si $m > k$ significa que la matriz de datos $\boldsymbol{Z}$ puede expresarse como $(\boldsymbol{X}, \boldsymbol{Z}^*)$ y reescribiendo $\boldsymbol{M}_z$ se obtiene fácilmente el resultado).

En este caso: $E(\hat{\sigma}_m^2) < E(\hat{\sigma}_k^2)$ lo que no significa necesariamente que se vaya a elegir el modelo amplio, pues en primer lugar $\hat{\sigma}_m^2$ aparece en muchos criterios corregido por el factor $(T-m)$ y en segundo lugar en la definición de los criterios aparecen otras expresiones que precisamente penalizan la selección de modelos muy amplios. No obstante, existe un «trade -off» entre esa penalización y el hecho de que el valor de $\hat{\sigma}_m^2$ pueda ser menor, ello implica que la probabilidad de seleccionar modelos más amplios que el PGD no se anula para los criterios clásicos. Unicamente para los criterios SBIC y BEC se ha comprobado en la literatura que asintóticamente esa probabilidad tiende a cero (Geweke y Meese, 1981).

Para muestras finitas, podemos concluir que al existir una tendencia no despreciable hacia modelos amplios, existe también la posibilidad de que la selección cambie en función del número de alternativas propuestas.

Veamos algunas comparaciones analíticas sobre la robustez que, en este aspecto, presentan algunos de los criterios. Si comparamos por ejemplo los criterios SBIC y AIC que presentan una forma funcional parecida se observa que el primero penaliza más que el segundo la inclusión de modelos amplios, con lo cual cabe esperar que SBIC presente un comportamiento más estable.

En efecto, supongamos que habiéndose generado los datos con el modelo de k variables se plantea la selección entre dos modelos con m_2 y m_1 variables con $m_2 > m_1 \geq k$.

$$\text{SBIC}(m_2) = \ln(\hat{\sigma}^2_{m_2}) + m_2 \frac{\ln T}{T}$$

$$\text{SBIC}(m_1) = \ln(\hat{\sigma}^2_{m_1}) + m_1 \frac{\ln T}{T}$$

Llamando $\nabla(\text{SBIC}) = \text{SBIC}(m_2) - \text{SBIC}(m_1)$, podemos expresar que el modelo con m_1 variables resultará elegido siempre que $\nabla(\text{SBIC}) \geq 0$. Es decir si

$$\nabla(\text{SBIC}) = \ln(\hat{\sigma}^2_{m_2}) - \ln(\hat{\sigma}^2_{m_1}) + (m_2 - m_1)\frac{\ln T}{T} \geq 0$$

Análogamente de la definición del criterio AIC, se deduce que según este criterio el modelo con m_1 variables resultará elegido siempre que

$$\nabla(\text{AIC}) = \ln(\hat{\sigma}^2_{m_2}) - \ln(\hat{\sigma}^2_{m_1}) + (m_2 - m_1)\frac{2}{T} \geq 0.$$

Es decir, para tamaños muestrales por encima de $T = 8$ el criterio SBIC penaliza más que el AIC la inclusión de variables irrelevantes.

Procediendo de modo análogo para los criterios BEC y C_p se tiene:

$$\nabla(\text{BEC}) = \hat{\sigma}^2_{m_2} - \hat{\sigma}^2_{m_1} + (m_2 - m_1)\hat{\sigma}^2_{m^*}\frac{\ln T}{T - m^*}$$

$$\nabla C_p = \hat{\sigma}^2_{m_2} - \hat{\sigma}^2_{m_1} + (m_2 - m_1)\hat{\sigma}^2_{m^*}\frac{2}{T - m^*}$$

Al igual que en la comparación anterior, se concluye que BEC penaliza más que C_p la inclusión de modelos amplios para $T > 8$.

Por último, si comparamos las condiciones correspondientes a PC y $\overline{R}^2$ se tiene:

$$\nabla(\text{PC}) = \hat{\sigma}^2_{m_2} - \hat{\sigma}^2_{m_1} + \left(\frac{T+m_2}{T-m_2}\right) - \left(\frac{T+m_1}{T-m_1}\right) \geq 0$$

$$\nabla(\overline{R}^2) = \hat{\sigma}^2_{m_2} - \hat{\sigma}^2_{m_1} + \left(\frac{T}{T-m_2}\right) - \left(\frac{T}{T-m_1}\right) \geq 0$$

siempre que se satisfaga la condición relativa a $\overline{R}^2$ se satisfará la de PC, por tanto éste seleccionará menos veces modelos muy amplios.

De las comparaciones establecidas anteriormente y de algunas equivalencias probadas en la literatura estadística (AIC $\approx$ PC $\approx$ CAT, SBIC $\approx$ BEC asintóticamente) podemos pensar en tres grandes categorías para los criterios clásicos. Por un lado estarían los criterios SBIC y BEC que apenas muestran tendencia a la sobreparametrización o subparametrización y por ello estimamos que no se verán apenas afectados por el número de modelos en la comparación. Por otro lado está el criterio $\overline{R}^2$ que penaliza muy poco la selección de modelos con muchas variables y por tanto, tendrá un comportamiento arbitrario dependiendo del número de modelos. En una situación intermedia estarían el resto de criterios.

En cuanto al criterio PEC su comportamiento es bien distinto del resto de modelos considerados. La definición del estadístico tiene dos partes una que mide la varianza del predictor y otra los errores de predicción cometidos.

Si la selección se plantea entre dos modelos anidados $\mathbf{M}_m \supset \mathbf{M}_k$ la varianza del predictor para el modelo más restringido $\mathbf{M}_k$ es siempre menor que la del modelo amplio $\mathbf{M}_m$, independientemente del proceso que haya generado los datos.

Si utilizamos el modelo amplio para predecir definimos el predictor como:

$$\hat{y}_{mp} = \boldsymbol{z}'_p\hat{\gamma} \quad \text{con} \quad \hat{\gamma} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

donde $\boldsymbol{z}'_p$ es un vector fila con m elementos cuyo correspondiente valor de y queremos predecir.

Si la predicción se hace con el modelo restringido el predictor será:

$$\hat{y}_{kp} = \boldsymbol{x}'_p\hat{\beta} \quad \text{con} \quad \hat{\beta} = (\boldsymbol{X}' - \boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

donde $\boldsymbol{x}'_p$ es un vector fila con k elementos Si el proceso generador de datos ha sido el modelo restringido se tiene que:

$$\text{Var}(\hat{y}_{kp}) = \sigma^2_k\boldsymbol{x}'_p(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_p \; < \; \text{Var}(\hat{y}_{mp}) = \sigma^2_k\boldsymbol{z}'_p(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_p$$

Si por el contrario los datos se hubieran generado con el modelo amplio:

$$\text{Var}(\hat{y}_{kp}) = \sigma^2_m\boldsymbol{x}'_p(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_p \; < \; \text{Var}(\hat{y}_{mp}) = \sigma^2_m\boldsymbol{z}'_p(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{z}_p$$

Obsérvese que al ser $m > k$, $Z = (X, Z^*)$ y $z_p' = (x_p', z_p^{*\prime})$ pudiendo escribirse:

$$z_p'(Z'Z)^{-1}z_p = x_p'(X'X)^{-1}x_p + C, \quad \text{con } C \text{ escalar positivo}$$

Por lo tanto, el primer factor en la definición del criterio PEC propicia el que se seleccionen modelos anidados en el PGD, penalizando únicamente la inclusión de variables irrelevantes.

En cuanto al segundo factor, denotando $e_{mp} = y_p - \hat{y}_{mp}$; $e_{kp} = y_p - \hat{y}_{kp}$ a los errores que se cometen al predecir con el modelo amplio y el restringido respectivamente se tiene que: si los datos fueron generados con el modelo restringido $\mathrm{E}(e_{mp}) = \mathrm{E}(e_{kp}) = 0$; mientras que si fueron generados con el amplio $\mathrm{E}(e_{kp}) \neq \mathrm{E}(e_{mp}) = 0$. Por esta razón este segundo factor penaliza los modelos restringidos en favor de los más amplios. No obstante el hecho de exigir previamente la fase de corroboración (acotación para los errores de predicción de datos muestrales) elimina cualquier modelo muy alejado del PGD.

En síntesis, cabe esperar que al realizar distintas comparaciones, incluyendo modelos cada vez más amplios que contienen al verdadero PGD, los criterios SBIC, BEC y PEC apenas modificarán la selección, los criterios AIC, PC, Cp, BIC y CAT lo harán de forma moderada y el criterio $\overline{R}^2$ mostrará gran variabilidad en la selección. Veamos a través de un ejercicio de simulación como se confirman algunos de estos resultados.

3. RESULTADOS DEL EJERCICIO DE SIMULACIÓN

El ejercicio de simulación se ha realizado mediante la elaboración de un programa en lenguaje FORTRAN utilizando la librería estadística IMSL, la descripción sobre la forma de generar las observaciones, el procedimiento de estimación y los criterios de selección empleados puede verse en (Garcia Olaverri, Aznar ,1994).

Para efectuar la selección se comparan m^* modelos anidados, entre los que siempre se encuentra uno con el mismo número de regresores que el PGD. Nuestro objetivo es analizar si los criterios seleccionan de distinta forma cuando se modifica m^*.

El ejercicio de simulación se realiza dentro del siguiente esquema:

En primer lugar se generan observaciones para las variables X_t, según el modelo AR(2):

$$X_t = 1.6X_{t-1} - 0.64X_{t-2} + \xi_t \quad \text{con} \quad \xi_t \sim \mathrm{N}(0,1).$$

En segundo lugar, se obtienen las observaciones de la variable Y_t a través del PGD.

Presentamos los resultados correspondientes a dos posibles PGD:

$$Y_t = X_t + u_t$$

$$Y_t = 1.25X_t + 1X_{t-1} + 0.8X_{t-2} + u_t \quad \text{con} \quad u_t \sim \text{N}(0;0.5)$$

en ambos modelos.[1]

Por último, se estiman distintos modelos para el conjunto de observaciones y se van aplicando los criterios de selección de modelos. Para los dos PGD anteriores realizamos distintos procesos de selección en los que se va modificando el número de modelos presentes en la comparación. Mostramos a continuación algunos resultados:

Por ejemplo, generando los datos con el modelo $Y_t = X_t + u_t$, utilizando una muestra de tamaño $T = 100$, $\text{var}(u_t) = 0.5$ y realizando 100 iteraciones, se obtuvieron los resultados que se muestran en la Tabla 2.

Como puede observarse, independientemente del objetivo con que fueran construídos los criterios, algunos de ellos muestran la propiedad de que apenas se ven afectados por el número de modelos presentes en la comparación (siempre que entre ellos haya un modelo con las mismas variables que el PGD), en esta situación están los criterios SBIC, BEC y PEC. En el extremo opuesto aparecen criterios como el $\overline{R}^2$ que varía de forma notable la selección cuando se incluyen más modelos en la comparación, aun cuando estén realmente alejados del PGD. En una situación intermedia se encuentran los criterios BIC, AIC, PC, C_p y CAT que modifican bastante la selección cuando se pasa de 2 a 5 modelos, pero apenas cambian cuando se formulan alternativas muy alejadas del PGD (comparación entre 10 o más modelos).

El ejercicio se ha realizado para distintos modelos, tamaños muestrales y número de alternativas en la comparación. Con el fin de mostrar lo más relevante de estos resultados presentamos en la Tabla 3 información sobre el número de veces que, en 100 iteraciones, se eligió el PGD según los distintos criterios, variando el número de modelos, m^*, presentes en la comparación, así como el tamaño muestral. La fila correspondiente a $m^* = 2$ indica el porcentaje de veces que se eligió el modelo con un

[1] Para generar las X_t se han probado otros modelos AR(2): $X_t = \rho_1 X_{t-1} + \rho_2 X_{t-2} + \xi_t$; modificando ρ_1, ρ_2 y Var (ξ_t). En general, cuanto mayor es la varianza muestral de las X_t aparece una tendencia más clara a seleccionar el PGD, sin embargo el grado de cumplimiento de la propiedad que nos ocupa (estabilidad de los criterios) se mantiene, es decir si un criterio es muy inestable cuando cambia el número de modelos en la comparación, seguirá siéndolo independientemente de los valores que tomen ρ_1, ρ_2 y Var(ξ_t); por tanto las tablas que mostramos si bien se refieren a una forma concreta de generar las variables X_t, son representativas de un conjunto más amplio de posiblte situaciones.

regresor (PGD) cuando se comparó únicamente con otro modelo (de dos regresores). La fila correspondiente a $m^* = 5$ indica el porcentaje de veces que se eligió el PGD cuando se comparó con otros cuatro modelos, cada uno de ellos incluyendo una variable más que el anterior. Omitimos por tanto, el detalle de cuantas veces se seleccionaron modelos distintos al PGD, pero la tendencia es, en todos los casos muy parecida a la mostrada en la Tabla 2.

Por otras simulaciones realizadas en trabajos anteriores es sabido que de todos los criterios considerados, sólo el PEC muestra alguna tendencia a la subparametrización, pues como ha quedado indicado, al depender la selección de la varianza del predictor, penaliza claramente la selección de modelos con muchas variables. En este primer ejercicio como no es posible comparar el PGD con modelos anidados en él, el criterio PEC muestra unos resultados óptimos pero su comportamiento general no es así. De hecho, como se verá en el siguiente ejemplo, cuando intervienen en la comparación modelos anidados en el PGD, el criterio PEC tiende a seleccionarlos de forma no despreciable.

Al final de cada columna se indica entre paréntesis la desviación típica de la variable «número de veces que, en 100 iteraciones, se eligió el PGD».

Como es natural una desviación típica pequeña nos informa de que la selección apenas se ve afectada por el hecho de modificar el número de modelos alternativos. Por el contrario, una desviación típica grande muestra la poca robustez del criterio ante cambios en el número de modelos a comparar.

Es interesante observar cómo los distintos criterios mejoran (en términos del número de veces que se elige el PGD) cuando aumenta el tamaño muestral, pero no debiera obviarse el hecho de que criterios tan empleados como AIC seleccionan un 24% de las veces modelos distintos al PGD cuando la selección se plantea entre 5 alternativas, incluso para un tamaño muestral de 400. Para este mismo tamaño muestral el número de veces que no se selecciona el PGD se reduce al 16% si sólo se comparan dos modelos.

Como es de esperar para tamaños muestrales pequeños las fluctuaciones en la selección son más evidentes, así para $T = 60$ se observa que la diferencia entre $\overline{R}^2$ y C_p o CAT es de un 19% cuando la selección es entre dos modelos pero alcanza más de un 40% cuando se comparan 16 especificaciones alternativas.

Los criterios SBIC, BEC y PEC se muestran claramente como los más robustos, la selección mejora a medida que aumenta el tamaño muestral, pero apenas se modifica si para un mismo tamaño muestral se formulan más alternativas.

Tabla 2

Porcentaje de veces que se elige un modelo con m variables (Habiéndose generado los datos con $m = 1$; PGD: $Y_t = X_t + u_t$)

$T = 100$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 1$	66	81	78	78	78	78	96	96	100
$m = 2$	34	19	22	22	22	22	4	4	0

$T = 100$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 1$	43	74	70	70	73	70	96	96	100
$m = 2$	17	13	15	15	13	15	4	4	0
$m = 3$	10	7	7	7	7	7	0	0	0
$m = 4$	10	2	3	3	2	3	0	0	0
$m = 5$	20	4	5	5	5	5	0	0	0

$T = 100$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 1$	28	71	66	66	68	66	96	96	100
$m = 2$	10	14	14	14	13	14	4	4	0
$m = 3$	5	8	6	6	7	6	0	0	0
$m = 4$	5	1	3	3	2	3	0	0	0
$m = 5$	5	2	3	3	3	3	0	0	0
$m = 6$	2	1	1	1	0	1	0	0	0
$m = 7$	5	0	0	0	0	0	0	0	0
$m = 8$	14	1	3	3	3	3	0	0	0
$m = 9$	15	0	1	1	1	2	0	0	0
$m = 10$	11	2	3	3	3	2	0	0	0

$T = 100$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 1$	20	72	65	65	68	65	96	96	100
$m = 2$	7	15	14	14	12	14	4	4	0
$m = 3$	4	8	6	6	6	6	0	0	0
$m = 4$	5	0	3	3	2	3	0	0	0
$m = 5$	3	0	3	3	3	3	0	0	0
$m = 6$	2	1	1	1	1	1	0	0	0
$m = 7$	3	0	0	0	0	0	0	0	0
$m = 8$	8	0	3	3	3	3	0	0	0
$m = 9$	9	0	1	1	1	1	0	0	0
$m = 10$	3	2	2	2	2	2	0	0	0
$m = 11$	6	1	1	1	1	1	0	0	0
$m = 12$	4	1	0	0	1	1	0	0	0
$m = 13$	5	0	0	0	0	0	0	0	0
$m = 14$	5	0	0	0	0	0	0	0	0
$m = 15$	8	0	1	1	0	0	0	0	0
$m = 16$	8	0	0	0	0	0	0	0	0

Tabla 3

Porcentaje de veces que se elige un modelo con las mismas variables que el PGD en funcion de m^* (numero de modelos anidados presentes en la comparacion)

PGD: $Y_t = X_t + u_t$

$T = 60$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m^* = 2$	58	78	77	77	77	77	93	93	98
$m^* = 5$	34	65	61	61	62	62	93	93	98
$m^* = 10$	24	63	57	58	60	60	93	92	96
$m^* = 16$	18	66	56	57	60	60	93	91	96
	(15.25)	(5.87)	(8.43)	(8.07)	(7.12)	(7.12)	(0.0)	(0.83)	(1)

$T = 200$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m^* = 2$	74	84	83	84	84	84	97	97	98
$m^* = 5$	38	74	73	73	73	73	97	97	98
$m^* = 10$	23	72	68	68	69	68	97	97	98
$m^* = 16$	18	71	66	66	68	67	97	97	97
	(21.91)	(5.16)	(6.57)	(6.98)	(6.34)	(6.75)	(0.0)	(0.0)	(0.43)

$T = 400$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m^* = 2$	72	85	84	84	85	85	98	98	100
$m^* = 5$	45	76	76	76	76	76	98	98	100
$m^* = 10$	33	73	73	73	73	73	98	98	98
$m^* = 16$	24	73	73	73	73	73	98	98	97
	(18.06)	(4.92)	(4.5)	(4.5)	(4.92)	(4.92)	(0.0)	(0.0)	(1.29)

Si observamos las fluctuaciones de la selección a través de las desviaciones típicas (escritas al final de cada tabla) se observa con nitidez la robustez de los distintos criterios. Insistimos en la cautela con que deben ser interpretados los resultados correspondientes al criterio PEC cuando, como en este caso no hay modelos anidados en el PGD.

Las Tablas 4 y 5 nos muestran los resultados de la simulación cuando la selección se plantea conjuntamente para modelos más amplios y más restringidos que el PGD. Consideremos el caso de que el PGD sea $Y_t = 1.25X_t + 1X_{t-1} + 0.8X_{t-2} + u_t$. El resto de variables (varianza, semilla, etc) son idénticas a las empleadas en el anterior ejercicio.

Tabla 4

Porcentaje de veces que se elige un modelo con m variables (Habiéndose generado los datos con m = 3;

$$Y_t = 1.25X_T + X_{t-1} + 0.8X_{t-2} + U_t)$$

$T = 100$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 2$	0	0	0	0	0	0	0	0	11
$m = 3$	100	100	100	100	100	100	100	100	87

$T = 100$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 3$	73	89	89	89	89	89	98	98	99
$m = 4$	27	11	11	11	11	11	2	2	1

$T = 100$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 1$	0	0	0	0	0	0	0	0	0
$m = 2$	0	0	0	0	0	0	0	0	11
$m = 3$	57	78	78	78	78	78	97	97	88
$m = 4$	16	8	8	8	8	8	2	2	1
$m = 5$	27	14	14	14	14	14	1	1	0

$T = 100$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 1$	0	0	0	0	0	0	0	0	0
$m = 2$	0	0	0	0	0	0	0	0	11
$m = 3$	31	73	68	68	69	68	97	97	89
$m = 4$	8	6	7	7	7	7	2	2	0
$m = 5$	6	10	9	9	9	9	1	1	0
$m = 6$	3	3	4	4	4	4	0	0	0
$m = 7$	7	1	3	3	3	3	0	0	0
$m = 8$	15	4	4	4	4	5	0	0	0
$m = 9$	18	1	2	2	2	2	0	0	0
$m = 10$	12	2	3	3	2	2	0	0	0

$T = 100$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 1$	0	0	0	0	0	0	0	0	0
$m = 2$	0	0	0	0	0	0	0	0	11
$m = 3$	23	74	68	68	69	68	97	97	89
$m = 4$	7	6	7	7	7	7	2	2	0
$m = 5$	3	8	8	8	9	8	1	1	0
$m = 6$	3	3	4	4	3	4	0	0	0
$m = 7$	4	1	3	3	2	3	0	0	0
$m = 8$	8	3	4	4	5	5	0	0	0
$m = 9$	11	1	2	2	1	1	0	0	0
$m = 10$	4	0	2	2	2	2	0	0	0
$m = 11$	6	1	1	1	1	1	0	0	0
$m = 12$	4	1	0	0	1	1	0	0	0
$m = 13$	5	0	0	0	0	0	0	0	0
$m = 14$	5	0	0	0	0	0	0	0	0
$m = 15$	8	0	1	1	0	0	0	0	0
$m = 16$	9	0	0	0	0	0	0	0	0

Tabla 5

Porcentaje de veces que se elige un modelo con las mismas variables que el PGD en funcion de m^*

PGD: $Y_t = 1.25X_t + X_{t-1} + 0.8X_{t-2} + u_t$

$T = 60$	$\overline{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m^* = 2$ (2 vs.3)	100	100	100	100	100	100	100	100	82
$m^* = 2$ (3 vs.4)	72	82	79	79	80	80	92	94	96
$m^* = 5$	62	77	74	74	75	75	89	90	83
$m^* = 10$	31	73	61	61	64	63	89	88	83
$m^* = 16$	26	74	56	57	62	60	89	88	83
	(27.31)	(9.91)	(15.45)	(15.22)	(13.66)	(14.26)	(4.26)	(4.56)	(5.31)

$T = 200$	$\overline{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m^* = 2$ (2 vs.3)	100	100	100	100	100	100	100	100	93
$m^* = 2$ (3 vs.4)	72	82	82	82	82	82	96	96	98
$m^* = 5$	57	72	70	70	71	70	96	96	90
$m^* = 10$	29	66	64	64	65	64	96	96	91
$m^* = 16$	20	63	61	61	62	61	96	96	90
	(29.04)	(13.38)	(14.29)	(14.29)	(13.81)	(14.29)	(1.6)	(1.6)	(3)

$T = 400$	$\overline{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m^* = 2$ (2 vs.3)	100	100	100	100	100	100	100	100	95
$m^* = 2$ (3 vs.4)	68	85	84	84	84	84	100	100	98
$m^* = 5$	50	76	74	74	74	74	98	98	96
$m^* = 10$	35	70	69	69	70	69	98	98	96
$m^* = 16$	24	70	69	69	70	69	98	98	96
	(26.76)	(11.32)	(11.75)	(11.75)	(11.41)	(11.75)	(0.98)	(0.98)	(0.98)

Cuando indicamos $m^* = 2$ (2 vs. 3) nos referimos a la comparación entre dos modelos lineales anidados de 2 y 3 variables respectivamente, cuando indicamos $m^* = 2$ (3 vs. 4) nos referimos a la comparación entre dos modelos anidados de 3 y 4 variables. Como puede observarse, en la primera fila de cada tabla, excepto el criterio PEC que muestra cierta tendencia a la subparametrización, el resto de métodos seleccionan sin excepción el PGD. La situación cambia claramente cuando la selección se establece entre el PGD y un modelo más amplio que él (fila 2 de cada tabla).

Las filas restantes corresponden a la comparación de m^* modelos lineales anidados de $1, 2, 3, \ldots, m^*$ variables respectivamente. Al igual que en ejemplo anterior, mientras algunos criterios se ven muy poco afectados por el número de modelos en la comparación, otros cambian significativamente el resultado de la selección. Así, los criterios AIC, PC, C_p, CAT, son poco robustos en este sentido y el $\overline{R}^2$ nada robusto. Además para tamaños muestrales pequeños ($T = 60$), criterios como BIC, CAT o C_p pueden parecer semejantes si la selección se hace entre dos modelos, pero en un hipotético ambiente de más incertidumbre donde la selección se estableciera entre 10

ó más modelos, el criterio BIC es superior a los otros dos. Claramente por encima de todos ellos se sitúan los criterios, SBIC, BEC y PEC.

Hay que tener en cuenta que el hecho de realizar la selección entre modelos con muchas variables, puede plantear problemas de grados de libertad, que podrían eliminarse aumentando el tamaño muestral; queremos señalar, no obstante, que aun estando ambas cuestiones (tamaño muestral y tamaño del modelo) íntimamente relacionadas hay que separar los dos problemas.

Si observamos la selección desde un punto de vista verificacionista en el que exclusivamente estuviéramos interesados en conocer cuántas veces se elige cada modelo, es evidente que los criterios PEC, SBIC y BEC mantienen fija la selección, mientras que otros van disminuyendo el número de veces que se elige el PGD a medida que aumenta el número de modelos en la comparación.(obsérvense por ejemplo, las columnas correspondientes al criterio $\overline{R}^2$).

Entendemos que esta robustez relativa al número de modelos que entran en la comparación es una buena propiedad de los criterios PEC, SBIC y BEC frente a los otros métodos de selección.

BIBLIOGRAFÍA

[1] **Akaike, H.** (1969). «Fitting Autoregressive Models for Prediction». *Annals of Institute of Statistical Mathematics*, bf 21, 243–247.

[2] **Akaike, H.** (1974). «A New Look at the Statistical Model Identification». *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

[3] **Amemiya, T.** (1980). «Selection of Regressors». *International Economic Review*, **21**, 331–354.

[4] **Aznar, A.** (1989). *Econometric Model Selection: A New Approach*. Dordrecht: Kluwer Academic Publishers.

[5] **Chow, G.C.** (1981). «A Comparison of the Information and Posterior Probability Criteria for Model Selection». *Journal of Econometrics*, **16**, 21–33.

[6] **Geweke, J.** y **Meese, R.** (1981). «Estimating Regression Models of Finite but Unknown Order». *International Economic Review*, **22**, 55–70.

[7] **García Olaverri, C.** y **Aznar, A.** (1994). «Estudio comparado de la robustez de distintos criterios de selección de modelos econométricos ante cambios en la varianza». *Estadística Española*, **36, 136**, 287–318.

[8] **García Olaverri, C.** (1993). *El Criterio ACOR: nuevos desarrollos teóricos y resultados de un estudio de Monte-Carlo.* Tesis Doctoral. Universidad de Zaragoza.

[9] **Koehler, A. B.** y **Murphree, B. S.** (1988). «A Comparison of the Akaike and Schwarz Criteria for selecting Model Order». *Applied Statistics*, **37**, 187–195.

[10] **Mallows, C.L.** (1973). «Some Comments on C_p». *Technometrics*, **15**, 661–676.

[11] **Mills, J.A.** y **Prasad, K.** (1992). «A Comparison of Model Selection Criteria». *Econometric Reviews*, **11**, 201–223.

[12] **Parzen, E.** (1974). «Some Recent Advances in Time Series Analysis». *IEEE Transactions on Automatic Control*, **AC-19**, 723–730.

[13] **Sawa, T.** (1978). «Information Criteria for Discriminating among Alternative Regression Models». *Econometrica*, **46**, 1273–1282.

[14] **Schwarz, G.** (1978). «Estimating the dimension of a Model». *Annals of Statistics*, **6**, 461–464.

[15] **Theil, H.** (1961). *Economic Forecasts and Policy*. Amsterdam: North Holland.

ENGLISH SUMMARY:

STABILITY OF SOME MODEL SELECTION CRITERIA

Carmen García Olaverri

In this paper we compare the behavior of nine model selection criteria. The aim of the study is to analyze if the number of models in the comparison has any influence on the results of the selection procedure. Are the results of selection the same if we compare 2 or 10 models? Are all the selection criteria similar behavior in this aspect? To answer these questions we conduct a Monte - Carlo simulation study.

Let us suppose that we have sample information relative to the variables $Y_t, X_{1t}, X_{2t}, \ldots, X_{mt}$.

There are some relationships that can be established between the former set of variables, if we don't know what is the «best» specification we can formulate some alternative models:

$$\mathbf{M}_1: \quad Y_t = \beta_1 X_{1t} + u_{1t}$$
$$\mathbf{M}_2: \quad Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + u_{2t}$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\mathbf{M}_{m^*}: \quad Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_m^* X_{mt}^* + u_{mt}^*$$

and then, select the better model by means of a criterion.

The criteria to be compared are the following: (Theil, 1961), Cp (Mallows, 1964), AIC (Akaike, 1973), CAT (Parzen, 1974), BIC (Sawa, 1978), SBIC (Schwarz, 1978), PC (Amemiya, 1980), BEC (Geweke and Meese, 1981), PEC (Aznar y García 1994). To establish the comparison we have adopted the following notation:

The selection consists in choose one of the m^* linear and nested models $\mathbf{M}_{m^*} \supset \cdots \supset \mathbf{M}_2 \supset \mathbf{M}_1$, being:

- m: the number of variables contained in $\mathbf{M}_m$ model
- m^*: the number of variables contained in the biggest model considered $\mathbf{M}_{m^*}$.
- $\hat{\sigma}_m^2, \hat{\sigma}_{m^*}^2$: maximum likelihood estimates of the variance on the error term in $\mathbf{M}_m$ and $\mathbf{M}_{m^*}$ respectively.
- T: sample size.

Using this notation, the expressions of the criteria to be compared are the following:

Choose the m value ($m = 1, \ldots, m^*$) that minimizes:

- $\overline{R}^2$ ***Criterion:***

$$\hat{\sigma}_m^2 \cdot \frac{T}{T-m}$$

- C_p ***Criterion:***

$$\hat{\sigma}_m^2 + \frac{2m}{T-m^*} \cdot \hat{\sigma}_{m^*}^2$$

- ***AIC Criterion:***

$$\ln \hat{\sigma}_m^2 + \frac{2m}{T}$$

- ***BIC Criterion:***

$$\ln(\hat{\sigma}_m^2) + 2\left(\frac{m+2}{T-m^*}\right)\left(\frac{\hat{\sigma}_{m^*}^2}{\hat{\sigma}_m^2}\right) - \frac{2T}{(T-m^*)^2}\left(\frac{\hat{\sigma}_{m^*}^2}{\hat{\sigma}_m^2}\right)^2$$

- ***PC Criterion:***

$$\hat{\sigma}_m^2\left(\frac{T+m}{T-m}\right)$$

- ***SBIC Criterion:***

$$\ln(\hat{\sigma}_m^2) + m\frac{\ln T}{T}$$

- ***CAT Criterion:***

$$\sum_{j=1}^{m} \frac{(T-j)}{T^2} \cdot \hat{\sigma}_j^{-2} - \frac{(T-m)}{T} \cdot \hat{\sigma}_m^{-2}$$

- ***BEC Criterion:***

$$\hat{\sigma}_m^2 + m\,\hat{\sigma}_{m^*}^2 \frac{\ln T}{T-m^*}$$

- ***PEC Criterion:***

$$\left[\frac{1}{T_1}\sum_{p=1}^{T_1} \hat{\mathrm{Var}}(\hat{y}_{mp})\right] \cdot \left[\frac{1}{T_1}\sum_{p=1}^{T_1} e_{mp}^{*2}\right]$$

where the second factor indicates the mean squared prediction error corresponding to the last T_1 observations.

Following the definition of the criteria we can observe that there are some methods that assign a high penalty to the models with many variables. For these critera the number of models in the comparison will be irrelevant.

For example SBIC penalizes the big models more than AIC; BEC more than C_p and PC more than $\overline{R}^2$. So, we can expect a different behavior of the criteria.

SIMULATION RESULTS

A Monte-Carlo experiment is conducted in order to study the behavior of the nine compared criteria. The program was written in FORTRAN language using IMSL library. Robustness of the experiment was tested previously.

Here we have limited our attention to two models as Data Generating Process (DGP), one hundred replications were generated from each of the following models:

$$\begin{aligned} Y_t &= X_t + u_t \\ Y_t &= 1.25X + X_{t-1} + 0.8X_{t-1} + u_t \end{aligned}$$

The generating model for the regressors is $X_t = 1.6X_{t-1} - 0.64X_{t-2} + \xi_t$ where ξ_t is a standard normal variable with mean zero and a variance that can take different values. (Following a experiment from Geweke and Meese). We consider four possible sample sizes 60, 100, 200 and 400.

Once the data are generated we estimate m^* nested models (one of them containing the same variables as the true DGP) and we select the «best» model using the different selection criteria.

The results are summarized in TABLES 2-5. Tables 2 and 3 show the selection when the DGP is $Y_t = X_t + u_t$. Tables 4 and 5 show the results when DGP is the model $Y_t = 1.25X + X_{t-1} + 0.8X_{t-1} + u_t$.

For each criterion we evaluate how many times the DGP is selected and we repeat the selection changing the number of the models in the comparison. Tables 2 and 4 are refered to $T = 100$ and show the number of times that every model on the comparison has been selected. Tables 3 and 5 are relative to sample sizes $T = 60$, 200 and 400, and show the number of times that the DGP has been selected, m^* indicates the number of models being compared.

As can be seen there are some criteria as SBIC, BEC or PEC that rarely change the results of the comparison when the number of models is modified. So we can

conclude that they are robust in this aspect. On the other hand the $\overline{R}^2$ criterion shows a poor behavior. The BIC, AIC, PC, C_p and CAT criteria present an intermediate situation; when the comparison is established between two models they can seem similar, but when the number of models increases the BIC criterion is better than the other ones (see the correponding table to $T = 60$ in Table 5).

If we suppose that we have to select a model in a uncertain environment and we decide to establish the comparison between a few models, it is interesting to know what are the criteria that are «independent» on the number of models in the comparison and which are not. In this framework we conclude that the criteria having this good property are: SBIC, BEC and PEC; being the $\overline{R}^2$ criterion the worse.

Working with the simulation experiment we get another results well known in the literature of model selection criteria. The consistency of SBIC and BEC criteria; the tendency to overfit that does not vanish in AIC, PC, C_p, and CAT criteria being around 30% of selections, even when the sample size increases; and the fact that PEC criterion is more parsimonious than the other criteria.