QÜESTIIÓ, vol. 22, 1, p. 3-37, 1998

EL MODELO LINEAL SIN TÉRMINO INDEPENDIENTE Y EL COEFICIENTE DE DETERMINACIÓN. UN ESTUDIO MONTE CARLO

RAFAELA DIOS PALOMARES*
Universidad de Córdoba

En el presente trabajo se analiza y compara mediante un experimento Monte Carlo el comportamiento de cinco expresiones para el Coeficiente de Determinación cuando el modelo lineal se especifica sin término independiente. Se ensayan distintos valores del parámetro poblacional P^2, que mide la proporción de varianza explicada por el modelo, introduciendo también la multicolinealidad como factor de variación en el diseño. Se confirma el coeficiente propuesto por Heijmans y Neudecker (1987) y el de Barten (1987), como idóneos para medir la bondad del modelo.

The linear model without a constant term and the coefficient of determination. A Monte Carlo study.

Palabras clave: Coeficiente de determinación, bondad de ajuste, modelo lineal sin término independiente, método Monte Carlo.

Clasificación AMS: (MSC): 62J20

* Rafaela Dios Palomares. Dpto. de Estadística e Investigación Operativa. Escuela Técnica Superior de Ingenieros Agrónomos y de Montes de la Universidad de Córdoba.

– Recibido en mayo de 1996.
– Aceptado en octubre de 1997.

1. INTRODUCCIÓN

La econometría empírica tiene como objetivo fundamental llegar a la estimación de un modelo econométrico que represente el comportamiento conjunto de las variables económicas objeto de estudio. Dicha estimación debe ser contrastada tanto para verificar el acierto en la previa especificación del modelo, como para admitir el cumplimiento de las hipótesis supuestas al mismo. Posteriormente se utiliza el modelo con fines predictivos y/o para analizar los parámetros estructurales.

Es de vital importancia, por tanto, lo que se denomina análisis de la bondad del modelo, siendo éste un tema que se puede analizar desde varios puntos de vista. Dependiendo de que los fines del estudio sean predictivos o estructurales, se enfoca de distinta forma el criterio «bondad». En el primer caso, será primordial la medida de la capacidad predictiva, siendo de escasa importancia la identificación de cada parámetro estructural.[1] En el segundo caso, sin embargo, un modelo será tanto mejor cuanto más precisa sea la estimación de dichos parámetros. La problemática inherente a todo trabajo econométrico tiene su base en el hecho del desconocimiento de lo que se denomina el Proceso Generador de los Datos (P.G.D.). Por tanto, en general, el modelo estimado será tanto más «bueno» cuando mejor capte dicho proceso.

Esta circunstancia da lugar a que el analista tenga que decidir la utilización de un modelo estimado, en base a la aplicación de ciertos criterios de bondad que, por supuesto, llevan implícita cierta probabilidad de error en la decisión. Es también muy frecuente, debido al desconocimiento del P.G.D., que se plantee la elección entre varios modelos que se podrían admitir como buenos, porque representen posibles especificaciones alternativas aceptables.

La gran importancia del tema de la selección y aceptación del mejor modelo ha dado lugar al desarrollo de numerosas investigaciones, cuyos resultados se plasman en el establecimiento de distintos criterios de bondad.

El criterio más usado como estadístico de bondad de ajuste es el denominado Coeficiente de Determinación Múltiple, sobre el que se pueden encontrar abundantes trabajos. En Kendall (1960), quedó patente el carácter de «evergreen»[2] de dicho coeficiente, que se sigue manteniendo como tema recurrente en Estadística y Econometría.

Un indicador relacionado con el coeficiente de determinación es el Cp de Mallows, que puede verse en los trabajos de Gorman y Toman (1966), y Malow (1973).

[1] En presencia de multicolinealidad exacta, la estimación de funciones de parámetros estructurales no afecta ni impide la predicción.

[2] Es un coeficiente cuyo caracter relativo lo convierte en tema siempre abierto al debate.

Para evaluar la capacidad predictiva, Theil (1961) planteó el estadístico U cuya descomposición detecta el origen de la discrepancia entre las series observada y predicha.

Otros criterios desarrollados posteriormente buscan encontrar un equilibrio entre la simplicidad del modelo y la bondad de ajuste. La base de los mismos está en minimizar la suma de cuadrados del error de predicción o minimizar los valores de la función de verosimilitud; planteados como una función del Error Cuadrático Medio y del número de parámetros, cabe señalar los siguientes: Criterio de Información de Akaike (AIC) (Akaike,1974), Error de Predicción Finito (FPE) (Akaike,1970), HQ (Hannan and Quinn, 1979), Schawarz (1978), Shibata (1981).

La aplicación de los criterios comentados, que se pueden ver en Judje *et al.* (1985) y Maddala (1988), tiene más sentido en el ámbito de decidir ante varios modelos que para determinar la bondad de un modelo en términos absolutos. Con este objetivo, se han seguido desarrollando métodos de selección, tanto para modelos anidados como no anidados, donde se penaliza la complejidad en pro de la «parsimonia».

Las últimas metodologías econométricas, sobre todo para series de tiempo, estudian lo que se denomina «encompassing» de modelos econométricos (Ver Mizon (1984) o Mizon and Richard (1986)), para determinar si un modelo «incluye» o «domina» a otro, tanto desde el punto de vista de la varianza del error, como de la predicción. En esta línea tienen relevancia el test J propuesto por Davidson and Mackinnon (1981) y los introducidos por Chong and Hendry (1986) y desarrollados por Ericsson (1989) y Clements and Hendry (1991).

Sin embargo, llama la atención la lentitud con que la mayoría de los programas econométricos existentes en el mercado están incorporando estos criterios, que por otra parte considero de gran utilidad en el desarrollo del trabajo en Econometría Empírica.

Hay que reconocer asimismo que, a pesar de que se sigue investigando en análisis de bondad y selección de modelos, el protagonista, y siempre presente en todo trabajo econométrico que incluya la estimación y contrastación del modelo, es el Coeficiente de Determinación en su versión original y Ajustado.

Ambos coeficientes son de fácil interpretación desde el punto de vista del grado de explicación del modelo, al estar definidos entre 0 y 1. Además, el Coeficiente de Determinación Ajustado incorpora una corrección en función del número de variables explicativas y el tamaño de la muestra, que permite establecer comparaciones entre distintos modelos. Koerts y Abrahamse (1969) obtuvieron la función de distribución de dicho coeficiente y, posteriormente, Smith (1973) y Ebberler (1975) completaron los resultados obtenidos por los primeros autores, en el ámbito de establecer proba-

bilidades desde el punto de vista comparativo entre los coeficientes calculados para varios modelos alternativos.

Hay que considerar, sin embargo, que el Coeficiente de Determinación, tal como se define para el modelo lineal general, no cumple las mismas propiedades cuando se aplica al modelo lineal sin término independiente, ya que puede tomar valores inferiores a cero. Es por esta circunstancia que diversos investigadores han aconsejado Coeficientes de Determinación alternativos, cuya idoneidad en el análisis de la bondad del modelo no ha sido extensivamente analizada en pequeñas muestras.

En adición, está claro que se siguen realizando muchos trabajos en Econometría aplicada, sin prestar la debida atención al hecho de que el modelo sin término independiente requiere un tratamiento específico, en lo que al cálculo y la interpretación del Coeficiente de Determinación se refiere, y esta negligencia puede provocar errores importantes en la toma de decisiones.

Por este motivo, he considerado de gran interés estudiar la repercusión que puede tener el uso de las distintas formas propuestas para el Coeficiente de Determinación en el modelo lineal sin término independiente. Para ello se ha diseñado un experimento Monte Carlo que permite calcular las distribuciones empíricas de dichos coeficientes y establecer medidas que nos lleven a definir su comportamiento en relación con el análisis de la bondad del modelo.

2. ASPECTOS TEÓRICOS

La medida de bondad de ajuste más usada en el trabajo aplicado es el Coeficiente de Determinación Múltiple, denominado R^2. Se especifica el modelo uniecuacional

$$y = X\gamma + u \tag{1}$$

donde X contiene una columna de unos y las observaciones de $k-1$ variables explicativas con

$$E(u'u) = \sigma^2 I_T$$ [3]

Para este modelo, el estimador minimocuadrático será $\hat{\gamma} = (X'X)^{-1}X'y$, de modo que podemos escribir

$$y = X\hat{\gamma} + \hat{u} = \hat{y} + \hat{u} \tag{2}$$

[3] Cuando $E(u'u) = \sigma^2 W$, con $W \neq I_T$, o cuando estimamos un sistema de ecuaciones aparentemente no relacionadas, la definición de un R^2 o un estadístico apropiado de bondad de ajuste no es obvia (Judje *et al*, 1985)

Se define el vector $\mathbf{1} = (1,1,1,\ldots,1)'$, la media muestral de y, $\bar{y} = \dfrac{\Sigma y}{T}$, y la matriz $A = I - (1/T)\,\mathbf{1}\,\mathbf{1}' = I - \mathbf{1}\,(\mathbf{1}'\mathbf{1})^{-1}\,\mathbf{1}'$, que multiplicada por cualquier matriz transforma sus columnas en valores desviados con respecto a la media de esa columna.

De ese modo, $Ay = (y - \mathbf{1}\bar{y})$ y

$$y'Ay = \hat{y}'A\hat{y} + \hat{u}'\hat{u} + 2\bar{y}\mathbf{1}'\hat{u} \tag{3}$$

donde se ha aplicado $X'\hat{u} = X'(y - X\hat{\gamma}) = 0$. Si además, el modelo tiene un término independiente, se cumple $\mathbf{1}'\hat{u} = 0$, es decir, que la suma de los residuos se anula y la expresión (3) se puede expresar como

$$y'Ay = \hat{y}'A\hat{y} + \hat{u}'\hat{u}. \tag{4}$$

El Coeficiente de Determinación se define mediante la expresión

$$R^2 = 1 - \frac{\hat{u}'\hat{u}}{y'Ay} \tag{5}$$

y debido a la descomposición en (4), está siempre comprendido entre 0 y uno, y se interpreta como el tanto por uno de la variación de y explicada por el modelo.

Además, R^2 es igual al cuadrado del coeficiente de correlación de Pearson entre la variable endógena, y, y la estimada, $\hat{y}$. Se entiende, por tanto, como una medida de la capacidad predictiva del modelo, y podemos decir que la expresión (5), en el modelo lineal general, coincide con la siguiente

$$R^2 = \frac{(y'A\hat{y})^2}{(y'Ay)(\hat{y}'A\hat{y})} \tag{6}$$

Tiene especial interés la relación entre el R^2 y la prueba F de significación global del modelo, donde se contrasta la hipótesis nula de que todos los parámetros, a excepción del término independiente, son cero.

Dicha relación es la siguiente:

$$F = \frac{R^2(T-k)}{(1-R^2)(k-1)} \tag{7}$$

La aceptación de la hipótesis nula implica que ninguna variable explicativa influye sobre y.

Debido a que la adición de una variable explicativa siempre aumenta el valor de R^2, se usa con frecuencia un coeficiente alternativo que recibe el nombre de Coeficiente

de Determinación Ajustado, cuya expresión es la siguiente:

$$\overline{R}^2 = 1 - \frac{(\hat{u}'\hat{u})/(T-k)}{(y'Ay)/(k-1)} \tag{8}$$

que disminuye al añadir una variable irrelevante al modelo.

Llegados a este punto, vamos a analizar el Coeficiente de Determinación como medida de bondad de ajuste desde un punto de vista teórico y desde la perspectiva poblacional.

Descomponiendo la matriz X del modo:

$$X = (\mathbf{1}, Z), \tag{9}$$

el vector de parámetros queda igualmente descompuesto como:

$$\gamma = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{10}$$

donde α es el término independiente y β el subvector que contiene los $k-1$ parámetros restantes. Así, el vector de estimadores quedará descompuesto de forma que el estimador de β será:

$$b = (Z'AZ)^{-1} Z'Ay. \tag{11}$$

Si se parte de que $(X'X)$ es una matriz de rango completo, también lo será el producto matricial $(Z'AZ)$, y por tanto lo es $M = (1/T)Z'AZ$.

Se deduce fácilmente la existencia del límite en probabilidad de M, que denominaremos M_L, y expresamos como $M_L = \text{plim } M = \lim M$ cuando $T \to \infty$. Igualmente se demuestra que es no singular.

En el modelo con término independiente se puede aplicar la expresión $A\hat{u} = \hat{u}$ para expresar la descomposición (4) del modo:

$$y'Ay = b'Z'AZb + 2\hat{u}'AZb + \hat{u}'A\hat{u} = b'Z'AZb + \hat{u}'\hat{u} \tag{12}$$

Si además se define $m = \dfrac{y'Ay}{T}$, se puede expresar (5) como:

$$R^2 = \frac{b'Z'AZb}{y'Ay} \tag{13}$$

y como

$$R^2 = \frac{b'Mb}{m}. \tag{14}$$

Si $h = \dfrac{\hat{u}'\hat{u}}{T}$, también

$$R^2 = 1 - \frac{h}{m} \tag{15}$$

Analizando la descomposición (13), se puede considerar el término $(b'Z'AZb)$ como la parte explicada de la variación muestral de la variable endógena y el Coeficiente de Determinación como la fracción de la misma que explica el modelo.

Si se enfoca desde el punto de vista poblacional, y siguiendo a Cramer (1984) y a Barten (1987), se puede deducir el valor al que converge en probabilidad la variable aleatoria R^2, cuando $T \to \infty$.

Se puede demostrar que:

$$\text{plim } b'Mb = \beta' M_L \beta \tag{16}$$

$$\text{plim } m = \text{plim } b'Mb + \text{plim } h = \beta' M_L \beta + \sigma^2 \tag{17}$$

Por tanto, aplicando la definición (15), el límite en probabilidad del Coeficiente de Determinación será:

$$\text{plim } R^2 = \frac{\beta' M_L \beta}{\beta' M_L \beta + \sigma^2} = P^2 \tag{18}$$

Teniendo en cuenta que los regresores no son estocásticos y que son controlados por el investigador, M_L se puede considerar como una característica poblacional. Por tanto, P^2 puede ser enfocado como la fracción de la varianza poblacional de la variable endógena que es explicada por la variación de las variables explicativas.

Analizando a continuación las particularidades de los desarrollos anteriores en el caso de que el modelo no tenga término independiente, se especifica (1) con la restricción $\alpha = 0$, quedando del modo:

$$y = Z\beta + u \tag{19}$$

El estimador minimocuadrático de β, será óptimo, y se expresa como:

$$b = (Z'Z)^{-1} Z'y \tag{20}$$

y el vector de residuos correspondiente a (20) y (21), será:

$$\hat{u} = y - Zb \tag{21}$$

En este caso, sin embargo, hay que tener en cuenta que aunque se cumple que $Z'\hat{u} = 0$, no se cumple siempre que la suma de los residuos se anule $(\mathbf{1}'\hat{u} = 0)$, aunque se demuestre fácilmente que plim $\overline{u} = 0$, siendo $\overline{u}$ la media aritmética de $\hat{u}$.

Si se analiza además la descomposición (13), se ve que la variación de y alrededor de su media será:

$$y'Ay = b'Z'AZb + \hat{u}'\hat{u} - T(2\overline{u}\overline{z}'b + \overline{u}^2) \tag{22}$$

y por tanto,

$$m = b'Mb + (1/T)\,\hat{u}'\hat{u} - (2\overline{u}\overline{z}'b + \overline{u}^2), \tag{23}$$

con lo que no es difícil verificar que:

$$\text{plim } m = \beta' M_L \beta + \sigma^2, \tag{24}$$

es la misma expresión deducida en (18), para el modelo con término independiente.

Queda patente de este modo que se puede admitir el valor de P^2 como el límite al que debería tender cualquier medida de bondad de ajuste que se plantea. Basado en estos resultados, Barten (1987) propone las siguientes condiciones como deseables a cumplir por el Coeficiente de Determinación en todos los casos:

$$\begin{aligned} &a) \quad \text{plim } R^2 = P^2, \\ &b) \quad 0 \leqslant R^2 \leqslant 1, \\ &c) \quad \text{Si } b = 0,\ R^2 = 0, \\ &d) \quad \text{Si } \hat{u} = 0,\ R^2 = 1. \end{aligned} \tag{25}$$

Ya que el objetivo propuesto es analizar la bondad de ajuste en el modelo lineal sin término independiente, el primer paso será estudiar el comportamiento del R^2 definido en (5) y que, por supuesto, cumple las condiciones (26) en el modelo completo.

En primer lugar, cabe decir que no coincide con el cuadrado del coeficiente de correlación entre la y y la $\hat{y}$ estimada, ni se relaciona con la prueba F como se vio en (7).

En segundo lugar, tampoco se obtiene el mismo resultado con (5) y con (15), ya que no se cumple la expresión (13).

En cuanto a las condiciones (26), hay que decir que sólo se cumplen *a* y *d*, ya que si $b = 0$, no necesariamente $R^2 = 0$, pudiendo tomar valores negativos, de modo que está definido entre $-\infty$ y 1.

Por este motivo, queda claro que no es el coeficiente idóneo para medir la bondad de ajuste en el caso que nos ocupa, sobre todo en términos comparativos, ya que se pierde además el sentido de ser el tanto por uno de la varianza explicada por el modelo.

3. ALTERNATIVAS PROPUESTAS

Debido a la particularidad expuesta del modelo sin término independiente, se han propuesto diversas alternativas al Coeficiente de Determinación definido en (5), que pasamos a comentar.

Judje *et al.* (1985) ofrecen dos formas para R^2 que consideran razonables. La primera es la medida de la variación de y alrededor de 0 y no con respecto a su media, aludiendo a que si es razonable especificar un modelo sin constante, también lo es esta medida. Además, hay que tener en cuenta el hecho de que si bien, como hemos visto, no se cumple la expresión (13), sí se cumple en cambio

$$y'y = \hat{y}'\hat{y} + \hat{u}'\hat{u}, \tag{26}$$

por lo que R^2 puede ser definido como:

$$R^2 = 1 - \frac{\hat{u}'\hat{u}}{y'y} \tag{27}$$

Esta definición tiene, además, correspondencia con la prueba F de significación global del modelo sin término independiente, ya que el estadístico se calcula mediante:

$$F = \frac{\hat{y}'\hat{y}/k}{\hat{u}'\hat{u}/(T-k)}$$

y la relación sería:

$$F = \frac{R^2}{(1-R^2)} \frac{(T-k)}{k} \tag{28}$$

La segunda opción aconsejada por dichos autores es el cuadrado del coeficiente de correlación entre la variable y observada y la $\hat{y}$ estimada, que ya introdujimos en la expresión (6), y que en este caso no coincide con (5) ni tampoco con (28). Se presenta como una buena medida de la capacidad predictiva del modelo.

Al analizar la bondad de ajuste del modelo, Drymes (1984) plantea como «necesario» calcular el Coeficiente de Determinación usando la formula (28) y propone asímismo el cálculo del Coeficiente de Determinación Ajustado como:

$$\overline{R}^2 = 1 - \frac{\hat{u}'\hat{u}/(T-k)}{y'y/T} \tag{29}$$

En dicho análisis llama la atención sobre la precaución de no trabajar por error con datos desviados con respecto a la media, ya que los resultados no coinciden con los

de los datos sin desviar (como en el modelo completo) y dicho tratamiento carece de sentido.

Por otro lado, en Greene (1993) se advierte también del problema inherente al cálculo del Coeficiente de Determinación (5) y se comenta que la alternativa

$$R^2 = \frac{b' Z' y}{y' A y}$$

resulta igualmente problemática ya que puede tomar valores superiores a 1.

En Ramanathan (1992 y 1993) se comenta que aunque algunos paquetes informáticos calculan el R^2 (28), hay que tener en cuenta que dicho coeficiente no es comparable con el de (5), al no tener el mismo denominador. En ambos textos se propone como coeficiente ajustado la expresión:

(30) $$\overline{R}^2 = 1 - \frac{\widehat{\text{var}}(\hat{u})}{\widehat{\text{var}}(y)}$$

Ramanathan (1993), también considera buena medida el coeficiente de determinación que relaciona la y observada y estimada, definido en (6).

Para hacer comparaciones entre modelos con y sin término independiente, Ramanathan (1992), recomienda usar (5), a pesar de que puede tomar valores negativos. Llama la atención también sobre el hecho de que no todos los paquetes econométricos calculan el Coeficiente de Determinación mediante la misma fórmula, por lo que conviene conocer dicha fórmula en cada caso, con el fín de poder interpretar los resultados.

Heijmans and Neudecker (1987) estudian las propiedades asintóticas del coeficiente que hemos definido en (6) y lo plantean como idóneo en todos los casos, demostrando que su límite en probabilidad es P^2, y que en el modelo completo este coeficiente coincide con el de (5).

Por último, Barten (1987) dedica su trabajo al tema que nos ocupa, y analiza el comportamiento de algunas fórmulas propuestas para el Coeficiente de Determinación en el modelo sin término independiente.

Con respecto al coeficiente (28), que es uno de los más aconsejados en la bibliografía y calculado en varios paquetes econométricos, hay que decir que falla en la condición *a*, ya que su límite en probabilidad resulta superior a P^2.

Barten (1987) concluye que el que ofrece mejores resultados es el siguiente:

(31) $$R^2 = \frac{(\hat{y}' A \hat{y})}{(\hat{y}' A \hat{y}) + \hat{u}' \hat{u}}$$

demostrando que cumple todas las condiciones (26), que hemos planteado como deseables para cualquier medida de bondad del modelo.

4. TRATAMIENTO EN PAQUETES ECONOMÉTRICOS

Se comenta a continuación el tratamiento que se hace sobre la bondad del modelo, en los paquetes econométricos más usados en la actualidad y que he tenido ocasión de revisar.

Llama la atención el hecho de que el paquete TSP 4.2 A (1991) para micro, así como el (TSP 7.03 (1990) y la nueva versión para Windows denominada EVIEWS (1994), no efectúan ningún cálculo específico para el modelo sin término independiente, aportando la misma salida de resultados y ofreciendo por tanto sólo los coeficientes (5), y el (7).

Sin embargo, el programa TSP 4.1 (1987) aplica para todos los casos el coeficiente definido en (6), que como vimos, mide el cuadrado de la correlación entre la variable y y la estimada por el modelo.

Otros paquetes como el PCGIVE 6.01 (1988) y el BMDP, en todas sus versiones, incluso la versión 7.0 (1993) para micro y el BMDP New System 1.1 (1994) para Windows, distinguen el caso en que no se estima el término independiente y aplican la fórmula (28).

El STATGRAPHICS 6.0 (1994) y el STATISTICA for Windows (1993) calculan, además de (28), el Coeficiente de Determinación Ajustado que vimos en (30).

En cambio, el RATS 4.0 (1992) y el SHAZAM 6.2 (1990) dan una información completa al aportar no sólo el coeficiente (28) sino tambien el (5) y el (7).

Con respecto a los demás criterios de bondad comentados en la introducción, cabe resaltar que son el PCGIVE, RATS y SHAZAM, los más completos en proporcionar dichas medidas.

5. DISEÑO Y METODOLOGÍA

El objetivo de la investigación que se ha realizado es, como ya se comentó en la introducción, el análisis individual y comparativo del comportamiento del Coeficiente de Determinación calculado, según las definiciones más aconsejadas para el estudio de la bondad del modelo lineal sin término independiente. Teniendo en cuenta que

el coeficiente (5) no es idóneo, resulta de gran interés estudiar cuales de las fórmulas propuestas tendrán propiedades óptimas en pequeñas muestras.

Como se vio en el apartado 2, una de las condiciones (26) obliga a que el límite en probabilidad de R^2 coincida con el valor P^2, que se puede considerar la porción de la varianza poblacional explicada por el modelo. En base a esto, la distribución muestral del Coeficiente de Determinación debe reflejar del mejor modo posible dicho valor P^2.

El planteamiento se ha basado en el establecimiento previo del Proceso Generador de Datos y la comprobación posterior del acoplamiento encontrado entre la distribución muestral de cada estadístico y el valor poblacional (P^2) con que se ha generado la muestra. Para ello se ha diseñado un experimento de simulación Monte Carlo (Hendry, 1984), cuyas particularidades se comentan a continuación.

El modelo a estudiar se ha especificado con dos variables explicativas y del modo:

$$y_t = \beta_1 z_{1t} + \beta_2 z_{2t} + u_t, \qquad \text{para } t = 1, \ldots, T \qquad \text{y } u \approx \text{N}\left(0_T, \sigma^2 I_T\right).$$

El tamaño de la muestra se ha fijado en $T = 15$, ya que se pretende analizar el fenómeno para pequeñas muestras. Los valores teóricos asignados a los parámetros[4] han sido por simplicidad $\beta_1 = \beta_2 = 10$.

En dicho diseño se han considerado dos factores de variación que han sido P^2 y la correlación entre z_1 y z_2. Ante la posibilidad de que el comportamiento de los estadísticos calculados no fueran indiferentes al valor de P^2 con que generamos la muestra, hemos realizado el experimento para distintos valores de dicho parámetro. Por otro lado, hemos investigado la incidencia de la presencia de multicolinealidad en el modelo, por lo que se ha introducido una segunda fuente de variación concretada en la correlación muestral entre z_1 y z_2.

Las variables explicativas son por hipótesis fijas en el muestreo, y por las características del método su magnitud no influye en los resultados. Por tanto, la construcción de la matriz X ha estado en función de la correlación que debía de haber entre dichas variables. El procedimiento que se ha seguido es partir de variables ortogonales para correlación cero y crear z_2 en función de z_1 para conseguir una determinada correlación. Un posterior ajuste de valores se realizó para que la correlación muestral fuera exactamente la requerida.

En nuestro diseño, los factores de variación son parámetros acotados entre 0 y 1. Como quiera que nuestro objetivo era tomar decisiones generales, el procedimiento adecuado sería plantear un diseño con factores aleatorios.

[4] Se han ensayado otros valores para los parámetros estructurales, pero no se recogen en el trabajo porque los resultados, como cabía esperar, son practicamente iguales.

Tras unos primeros ensayos de sensibilidad, y ante la evidencia de que los resultados no eran sensibles a valores con diferencias inferiores a una décima, se ensayaron todos los valores entre 0 y 1, con intervalo de una décima.

Para no diversificar demasiado los resultados, se presentarán los que producen diferencias más notables. Éstos son 0.2, 0.5, y 0.9 para P^2, y 0.4 y 0.8 para r.

El número de muestras generadas en cada caso ha sido de 1000.

El procedimiento metodológico que se ha planteado para el tratamiento de cada muestra ha sido el siguiente:

1) *Generar una muestra de tamaño 15 con un P^2 y un $r(z_1, z_2)$ conocidos:*

   Si en (19), se supone $M_L = M$,

$$P^2 = \frac{\beta' M \beta}{\beta' M \beta + \sigma^2},$$

   Si se establece el valor teórico de β, y unos valores para las variables z, esto nos determina el valor $\beta' M \beta$, que se llamará K, y se puede escribir:

(32) $$P^2 = \frac{K}{K + \sigma^2},$$

   y por tanto deducir

(33) $$\sigma^2 = K \frac{(1 - P^2)}{P^2}$$

   La expresión (33) nos indica que variando la relación entre la varianza teórica explicada K y la magnitud de la varianza del error (σ^2), se puede conseguir un valor de P^2 para el Proceso Generador de Datos. Enfocado en sentido inverso, (34) establece que dado un valor de K, el P.G.D. con un P^2 deseado, se simula generando la variable de error u con una varianza cuyo valor cumpla la expresión (34).

2) *Estimar el modelo para esa muestra por mínimos cuadrados ordinarios.*

3) *Calcular los siguientes estadísticos:*

   a) El Coeficiente de Determinación que se definió en (5) y que goza de excelentes propiedades en el modelo con término independiente

$$R_1^2 = 1 - \frac{\hat{u}' \hat{u}}{y' A y}$$

b) El Coeficiente de Determinación que aconsejan con más frecuencia en la bibliografía y que definimos en (28)

$$R_2^2 = 1 - \frac{\hat{u}'\hat{u}}{y'y}$$

c) El cuadrado de la correlación entre la y observada y la y estimada que vimos en (6)

$$R_3^2 = \frac{(y'A\hat{y})^2}{(y'Ay)(\hat{y}'A\hat{y})}$$

d) El Coeficiente de Determinación aconsejado por Barten (1987), que vimos en (32)

$$R_4^2 = \frac{(\hat{y}'A\hat{y})}{(\hat{y}'A\hat{y}) + \hat{u}'\hat{u}}$$

e) Por último, teniendo en cuenta que algunos autores calculan el Coeficiente de Determinación en el modelo completo, dividiendo la Suma de Cuadrados Explicada por la Suma de Cuadrados Total debido a que coincide con (5)[5], se ha creído de interés valorar la importancia del error que se comete al aplicar esta fórmula en el caso que nos ocupa. Por este motivo, se ha calculado también el estadístico:

$$R_5^2 = \frac{\hat{y}'\hat{y} - T\bar{y}^2}{y'Ay}$$

La información aportada por los estadísticos calculados para las 1000 muestras nos han permitido establecer las siguientes medidas de comportamiento para cada Coeficiente de Determinación (R_i^2):

a) Medidas Descriptivas: Valores máximos y mínimos, media, varianza e histograma de frecuencias.

b) Medidas de Proximidad a P^2 creadas al efecto:

b1) *Error Cuadrático medio con respecto a P^2:*

$$ECP_i = \frac{\Sigma(R_i^2 - P^2)^2}{1000}$$

b2) *Probalilidad de que el Coeficiente R_i^2 sea mayor que P^2:*

$$PA_i = \frac{fma}{1000},$$

siendo *fma* la frecuencia absoluta de valores de R_i^2 que han resultado superiores a P^2.

[5] Peña (1989) propone dicha formula y no la (5) para el cálculo del Coeficiente de Determinación.

Esta probabilidad (empírica) indica el tanto por uno de veces que cabe esperar que el valor calculado de R_i^2 sea superior a P^2, y por tanto se incurra en sobrevalorar la adecuación de la muestra al modelo teórico.

b3) *Sesgo de R_i^2 con respecto a P^2:*

$$SE_i = \frac{\Sigma R_i^2}{1000} - P^2$$

Todo el trabajo se ha desarrollado con el paquete econométrico EVIEWS, para el que se han realizado los programas necesarios, gracias a la posibilidad añadida de cálculo matricial que ha incorporado esta nueva versión para Windows de μTSP.

6. RESULTADOS

El carácter de la investigación, así como las dos fuentes de variación introducidas para analizar los cinco coeficientes estudiados, proporcionan una gran complejidad de resultados que resulta difícil resumir para una eficiente interpretación.

En dicha interpretación, vamos a estudiar en primer lugar el comportamiento de los cinco coeficientes individualmente, para lo cual presentaremos sus medidas descriptivas, sus histogramas y sus medidas de proximidad a P^2. En segundo lugar, estableceremos las oportunas comparaciones entre los cinco coeficientes, con el fin de establecer criterios que nos permitan deducir cuales de ellos resultan con mayor capacidad para captar el valor de P^2.

La presentación de resultados se hará en forma gráfica en su mayoría, con el fin de ayudar a la comprensión de los mismos y con el ánimo de que en el análisis comparativo resultará más enriquecedor e ilustrativo.

Los valores descriptivos de los cinco coeficientes se presentan en las tablas n° 1 al 6, de modo que cada una de ellas se refiere a un caso distinto cuyas características de P^2 y r aparecen el la cabecera correspondiente.

Tablas. *Valores descriptivos de los coeficientes R_i^2*

Tabla n° 1	$P^2 = 0.2$		$r = 0.4$		
	R_1^2	R_2^2	R_3^2	R_4^2	R_5^2
Media	0.250487	0.999485	0.262428	0.255417	0.256866
Mediana	0.234359	0.999509	0.244715	0.236102	0.234215
Desv. Típica	0.176111	0.000207	0.166722	0.164881	0.166513

Tabla n° 2		$P^2 = 0.5$		$r = 0.4$	
	R_1^2	R_2^2	R_3^2	R_4^2	R_5^2
Media	0.527197	0.999869	0.532573	0.533201	0.537237
Mediana	0.542468	0.999875	0.544730	0.542801	0.544409
Desv. Típica	0.167218	5.09E-05	0.159083	0.150426	0.155234

Tabla n° 3		$P^2 = 0.9$		$r = 0.4$	
	R_1^2	R_2^2	R_3^2	R_4^2	R_5^2
Media	0.910702	0.999985	0.911282	0.910932	0.912448
Mediana	0.915039	0.999986	0.915781	0.914716	0.915551
Desv. Típica	0.035041	5.73E-06	0.034804	0.034387	0.052371

Tabla n° 4		$P^2 = 0.2$		$r = 0.8$	
	R_1^2	R_2^2	R_3^2	R_4^2	R_5^2
Media	0.246804	0.999963	0.301187	0.283515	0.299356
Mediana	0.276058	0.999965	0.286882	0.264450	0.264482
Desv. Típica	0.219090	1.45E-05	0.178472	0.097233	0.137430

Tabla n° 5		$P^2 = 0.5$		$r = 0.8$	
	R_1^2	R_2^2	R_3^2	R_4^2	R_5^2
Media	0.539314	0.999991	0.576692	0.571616	0.612587
Mediana	0.578774	0.999992	0.590899	0.567129	0.571265
Desv. Típica	0.191573	3.51E-06	0.150818	0.096894	0.227150

Tabla n° 6		$P^2 = 0.9$		$r = 0.8$	
	R_1^2	R_2^2	R_3^2	R_4^2	R_5^2
Media	0.910573	0.999999	0.917666	0.913282	0.938364
Mediana	0.916493	0.999999	0.922364	0.917313	0.917330
Desv. Típica	0.036991	4.00E-07	0.033539	0.030635	0.154538

Los histogramas de frecuencias aparecen en los gráficos n° del 1 al 5, donde la presentación se ha hecho en otra dimensión, de modo que cada gráfico incluye las salidas de un coeficiente para los seis casos presentados. Pretendemos así facilitar la labor comparativa de los resultados.

Gráfico 1. Histogramas de Frecuencias de R_1^2.

Gráfico 2. Histogramas de Frecuencias de R_2^2.

Gráfico 3. Histogramas de Frecuencias de R_3^2.

Gráfico 4. Histogramas de Frecuencias de R_4^2.

Gráfico 5. Histogramas de Frecuencias de R_5^2.

Si se observan los histogramas de frecuencias[6], se puede apreciar que las distribuciones empíricas calculadas para R_1^2, R_3^2, R_4^2, y R_5^2 se agrupan alrededor del valor correspondiente de P^2 en todos los casos. En cambio, llama la atención el hecho de que la de R_2^2 se distribuye invariantemente entre 0.9 y 1, para todos los valores de P^2 generados.

En cuanto a las medidas de posición, se ve que tanto la media muestral como la mediana son superiores a P^2. En concordancia con lo comentado para las distribuciones, la media y la mediana del coeficiente R_2^2 resultan muy alejadas de P^2, cuando este parámetro toma los valores 0.2 y 0.5. En todos los demás casos, y para el resto de los coeficientes, ambas medidas de posición se encuentran relativamente próximas a P^2.

En cuanto al ámbito del espacio muestral, cabe señalar que los de R_2^2, R_3^2, y R_4^2 están comprendidos entre 0 y 1, mientras que el de R_1^2 incluye valores negativos con un mínimo de -0.6, y el de R_5^2 supera a 1, alcanzando un valor máximo de 2.15.

La desviación típica disminuye en general al aumentar P^2, advirtiéndose en los histogramas menos dispersión para valores más altos de P^2.

Los gráficos del n° 6 al 8 representan la posible incidencia de los parámetros P^2 y r, que se han introducido en el diseño como fuentes de variación en las medidas de proximidad calculadas para los coeficientes.

A la vista del gráfico n° 6, se puede ver que, como regla general, el valor de ECP_i disminuye en gran manera al aumentar P^2. Esta circunstancia confirma el comportamiento observado para la dispersión de las distribuciones, ya que al estar menos dispersas también es menor la magnitud de ECP_i para los casos en que la media está próxima a P^2.

La influencia de r en ECP, sin embargo, parece menos importante y, además, se detectan tendencias distintas, ya que ECP_1, ECP_3 y ECP_5 aumentan con r, y ECP_4 tiende a disminuir, mientras que ECP_2 se mantiene constante.

Según se puede ver en el gráfico n° 7, las medidas PA_1 y PA_3 aumentan claramente con P^2, mientras que PA_4 y PA_5 no presentan uniformidad ya que adoptan distintas tendencias según sea el valor de r. En cuanto a PA_2, hay que decir que su valor se mantiene en 1 para todos los casos, debido a que se encuentra toda la distribución a la derecha del valor P^2. En cuanto a la incidencia de la multicolinealidad, vemos un aumento generalizado de PA_i provocado por este problema, salvo en el caso ya comentado de PA_2.

[6]Para ilustrar los resultados hemos sombreado el área que está a la derecha de P^2, que mediremos porteriormente como PA.

Gráfico 6. Representaciones de *ECP* frente a P^2 y r.

Gráfico 7. Representaciones de *PA* frente a P^2 y r.

Gráfico 8. Representaciones de *SE* frente a P^2 y r.

La tercera medida de proximidad, recogida en el gráfico n° 8, presenta una gran uniformidad en su tendencia a disminuir cuando aumenta el valor de P^2. Por el contrario, un incremento en la correlación de las variables explicativas redunda en un aumento de SE_i, salvo para SE_1 y SE_2 en los que no se aprecia incidencia alguna.

Si se aborda la perspectiva comparativa entre los cinco coeficientes calculados, se pueden comentar los siguientes resultados:

Gráfico 9. Representaciones de ECP_1, ECP_3, ECP_4 y ECP_5.

En el gráfico 9 se han representado los valores calculados de la medida ECP_i, para todos los coeficientes a excepción de ECP_2, que se presenta en el gráfico 10 debido a que la diferencia de magnitud con el resto desvirtuaría la gráfica conjunta. Se puede observar que el coeficiente que menor error cuadrático presenta para todos los casos es R_4^2, y los mayores se deben a R_1^2 con valores bajos de P^2 y a R_5^2 con valores altos de dicho parámetro.

Gráfico 10. Representación de ECP_2.

Los gráficos 11 y 12 se ocupan de representaciones semejantes pero con respecto a la medida PA_i. Se puede obsevar aquí que la multicolinealidad provoca que PA_4 aumente considerablemente, ofreciendo el peor resultado en cuanto a este criterio el coeficiente R_4^2 para valores de $r = 0.8$. En estos casos las medidas PA_3 y PA_5 resultan ser las más bajas. Sin embargo, para los casos en que $r = 0.4$, son inferiores las de PA_4.

Gráfico 11. Representaciones de PA_1, PA_3, PA_4 y PA_5.

La representación comparativa de las medidas SE_i, se presentan en los gráficos 13 y 14, donde se puede apreciar un mejor comportamiento generalizado desde este punto de vista para R_1^2, siendo el peor en todos los casos el de R_4^2, al obtenerse mayores valores del sesgo calculado.

Gráfico 12. Representación de PA_2.

Gráfico 13. Representaciones de SE_1, SE_3, SE_4 y SE_5.

Gráfico 14. Representación de SE_2.

7. CONCLUSIONES

Se extraen a continuación las conclusiones más relevantes en base a los resultados obtenidos.

- Quedan confirmados empíricamente los siguientes puntos:
  - Los coeficientes R_1^2, R_3^2 y R_5^2 no ofrecen resultados iguales cuando se aplican al modelo sin término independiente.

  - El coeficiente R_1^2 toma valores inferiores a 0.
  - El coeficiente R_5^2 toma valores superiores a 1.

- El problema que presentan R_1^2 y R_5^2 de no estar definidos entre 0 y 1, hace que se pierda la interpretación como tanto por uno de varianza explicada y por tanto se concluye que no son los coeficientes idóneos para su uso en el modelo sin término independiente.

- Las distribuciones de los coeficientes no son indiferentes al valor teórico P^2, de modo que queda patente la influencia de este parámetro así como la de la multicolinealidad introducida en el modelo, según hemos podido apreciar en los resultados.

- El hecho de que se hayan obtenido todos los valores de PA_i mayores que 0.5, indica que existe un mayor peligro de sobrevalorar la adecuación del modelo que de subestimarla.

- Si se tiene en cuenta cómo se ha diseñado el P.G.D., se puede decir que las muestras generadas con un valor bajo de P^2 son aquellas para las que la varianza del error era grande en relación a la varianza teórica explicada que se denominó K. Esto ha provocado, en muestras de tamaño 15, el hecho de que la variable endógena explicada no estuviera próxima a la observada. Sin embargo, el modelo que se ha especificado en el experimento era exactamente igual al teórico, de forma que la hipótesis nula de anulación de los parámetros siempre debería de ser rechazada, puesto que nunca se cumplía.

  En relación con esto, se extrae una conclusión importante del comportamiento de los coeficientes calculados. Los coeficientes R_1^2, R_3^2, R_4^2, y R_5^2, captan perfectamente el valor de P^2, es decir, que detectan la falta de adecuación de la variable endógena observada con la explicada y, por tanto, podemos decir que serían buenos para tomar decisiones de cara a la predicción. Sin embargo, el coeficiente R_2^2 es indiferente al valor de P^2, ya que ofrece siempre buenos resultados indicando fielmente que el modelo se ha generado con parámetros distintos de 0. Esto se explica perfectamente dada su relación con la prueba F de significación global. Presenta, en cambio, el problema de no detectar la relación entre la varianza del error y la explicada. Nos podría llevar, por tanto, a conclusiones demasiado optimistas en pequeñas muestras, siendo mayor el riesgo de sobrevaloración para valores pequeños de P^2.

- En consecuencia, de todo lo anterior pensamos que los mejores coeficientes para analizar la bondad del modelo son R_3^2 y R_4^2, teniendo en cuenta además que la información aportada por R_2^2 ya la tenemos con la prueba F.

- En cuanto a discriminar entre R_3^2 y R_4^2 concluimos que, en lo que respecta a la investigación efectuada, no hemos encontrado razones suficientes para decidir cual se comporta mejor dentro de los límites del experimento planteado, que supone el inicio de un proyecto mucho más ambicioso. Concretamente, estamos preparando

un análisis en el contexto de plantear un error de especificación y estudiar la relación de los coeficientes con otras medidas de bondad del modelo.

8. AGRADECIMIENTOS

Quiero mostrar mi agradecimiento a D. José Angel Roldán Casas por su colaboración en la revisión de paquetes econométricos y en la realización de las gráficas del presente trabajo.

Agradezco igualmente las sugerencias de tres evaluadores anónimos.

REFERENCIAS

[1] **Akaike, H.** (1970). «Statistical Predictor Identification». *Annals of Institute. Stat. Math.*, **22**, 203-217.

[2] **Akaike, H.** (1974). «A new look at Statistical Model Identification». *IEEE Trans. Auto. Control*, **19**, 716-723.

[3] **Barten, A.P.** (1987). «The coefficient of determination for regression without a constant term». *The Practice of Econometrics*. Ed. R.D.H. Heijman y H. Neudecker, 181-189.

[4] **Chong, Y.Y.** y **Hendry, D.F.** (1986). «Econometric evaluat ion of linear macro-economic models». *Review of Economic Studies*, **53**, 671-690.

[5] **Clements, M.P.** y **Hendry, D.F.** (1991). «On the invalidity of mean square error forecast comparisons in economics». *Institute of Economics and Statistics and Nuffield College*, Oxford University, Oxford (mimeo).

[6] **Cramer, J.S.** (1984). «Sample size and R^2», *Discussion paper AE N2/84 of Faculty of Actuarial Science and Econometrics of the University of Amsterdam*.

[7] **Davidson, R.** y **Mackinnon, J.G.** (1981). «Several tests for model especification in the presence of alternative hypotheses». *Econometrica*, **49**, 781-793.

[8] **Dhrymes Phoebus J.** (1984). *Econometría*. Ed. AC. Madrid.

[9] **Ebbeler, D.H.** (1975). «On the Probability of Correct Model Selection Using the Maximum R^2 Choice Criterion». *International Economic Review*, **16**, 516-521.

[10] **Ericsson, N.R.** (1989). «Mean square error and forecast encompassing». *Discussion Paper*, International Finance Division, Board of Governors of the Federal Reserve System.

[11] **Gorman, J.W.** y **Toman, R.J.** (1966). «Selection of Variables for Fitting Equations to Data». *Techometrics*, **8**, 27–51.

[12] **Greene, W.H.** (1993). *Econometric Analysis*. Ed. Macmillan. New York.

[13] **Hannan, E.J.** y **Quinn, B.** (1979). «The Determination of the Order of an Autoregression». *J. Royal Stat. Society*, **series B, 41**, 190-195.

[14] **Heijmans, R.D.H.** y **Neudecker, H.** (1987). «The coefficient of determination revisited». *The Practice of Econometrics*, Ed. R.D.H. Heijmans y H. Neudecker, 191-204.

[15] **Hendry, D.F.** (1984). «Monte Carlo Experimentation in Econometrics» en Griliches, Z. and Intriligator, M.D. (eds) *Handbook of Econometric*, Volume II. Elsevier Science Publishers BV. North-Holland.

[16] **Judge, G.G., Griffiths, W.E., Hill, R.C. Lütkepohl, H.** y **Lee, T.C.** (1985). *The Theory and Practice of Econometrics*, Ed. Wiley, New York.

[17] **Kendall, M.G.** (1960). «The evergreen correlation coefficient, in: Contributions to probabilty and statistics». *Essays in Honor of Harold Hotelling*. Ed. I. Olkin a.o. Stanford : Stanford University Press, 274-277.

[18] **Koerts, J.** y **Abrahamse, A.P.J.** (1970). «The correlation coefficient in the general linear model». *European Economic Review*, **1**, 401-427.

[19] **Maddala, G.S.** (1988). *Introduction to Econometrics*. Ed. Macmillan. Londres.

[20] **Mallows, C.L.** (1973). «Some Comments on C_p». *Technometrics*, **15**, 661–675.

[21] **Mizon, G.E.** (1984). «The encompassing approach in econometrics», en Hendry, D. F. y Wallis, K.F. (eds), *Econometrics and Quantitative Economics*. Ed. Basil Blackwell, Oxford.

[22] **Mizon, G.E.** y **Richard, J.F.** (1986). «The encompassing principle and its application to testing non-nested hypothesis». *Econometrica*, **54**, 657-678.

[23] **Peña Sanchez de Rivera, D.** (1989). *Estadística Modelos y Métodos 2. Modelos lineales y series temporales*. 2ª Edición revisada. Ed. Alianza, Madrid.

[24] **Ramanathan, R.** (1992). *Introductory Econometrics with Applications*. Ed. Harcourt Brace Jovanovich, Fort Worth, Texas.

[25] **Ramanathan, R.** (1993). *Statistical Methods in Econometrics*. Ed. Academic Press, San Diego, California.

[26] **Schmidt, T.** (1973). «Calculating the Power of the Minimum Standard Error Choice Criterion». *International Economic Review*, **XIV**, 253-255.

[27] **Schwarz, G.** (1978). «Estimating the Dimension of a Model». *Annals of Stat.*, **6**, 461–464.

[28] **Shibata, R.** (1981). «An Optimal Selection of Regression Variables». *Biometrika*, **68**, 45–54.

[29] **Theil, H.** (1971). *Principles of Econometrics*. Ed. North-Holland, Amsterdam.

Referencias de paquetes econométricos

TSP. 4.2 A. TSP International. 1991.

MICRO TSP. 7.03. David M. Lilien. 1990.

ECONOMETRIC VIEWS. 1.1 C . David M. Lilien. 1994.

PC GIVE . 6.01. David F. Hendry. 1988.

BMDP. 7.0. BMDP Statistical Software. 1993.

BMDP NS. 1.1. BMDP Statistical Software. 1994.

STATGRAPHICS. 6.0 . Statistical Graphics Sistem. 1994.

PC RATS. 4.00. Tomas A. Doan. 1992.

SHAZAM. 6.2. K. J. White. 1990.

ESTATISTICA . 4.1. Statsoft. 1993.

ENGLISH SUMMARY

THE LINEAR MODEL WITHOUT A CONSTANT TERM AND THE COEFFICIENT OF DETERMINATION. A MONTE CARLO STUDY

RAFAELA DIOS PALOMARES*

Universidad de Córdoba

In the present work a Monte Carlo experiment has been done in order to compare the empirical performance of five differents expressions for the coefficient of determination in the linear model without intercept. All of these have been calculated introducing a knwon value of P2, which may be considered as the fraction of the population variance of the dependent variable explained by the variation of the regressors. A given level of multicolineality has been also introduced between the regressors. The results reveal that the best coefficients for the measure of goodness of fit in this case are R_3^2 *(Heijmans y Neudecker, 1987) y* R_4^2 *(Barten, 1987).*

Keywords: Coefficient of determination, goodness of fit, lineal model without an intercept, Monte Carlo method.

AMS Classification: (MSC): 62J20

* Rafaela Dios Palomares. Dpto. de Estadística e Investigación Operativa. Escuela Técnica Superior de Ingenieros Agrónomos y de Montes de la Universidad de Córdoba.

– Received May 1996.
– Accepted October 1997.

The commonly used expressions for the coefficient of determination are somewhat faulty as a maesure of goodness of fit in regressions without a constant term. They can give negative values for R^2 or values larger than unity.

The regression model considered here is writen as

$$y = Z\beta + u$$

where the matrix Z does not have a column with all elements equal to one. The best linear unbiased and consistent estimator of β is

$$b = \left(Z'Z\right)^{-1} Z'y \qquad \text{and the vector of residuals is} \qquad \hat{u} = y - Zb.$$

Let $\mathbf{1}$, A and M be the vector $\mathbf{1} = (1,1,1,\ldots,1)'$ and the matrices $A = I - (1/T)\,\mathbf{1}\mathbf{1}' = I - \mathbf{1}(\mathbf{1}\mathbf{1})^{-1}\mathbf{1}'$ and $M = (1/T)Z'AZ$, respectively.

In order to establish the parent counterpart of R^2 we follow Cramer (1984) and use the value to which the random variable

$$R^2 = \frac{b'Z'AZb}{y'Ay}$$

converges in probabilily for $T \to \infty$.

Assuming that $M_L = \operatorname{plim} M = \lim M$ for $T \to \infty$ we have the following expression

$$\operatorname{plim} R^2 = \frac{B'M_L\beta}{\beta'M_L\beta + \sigma^2} = P^2$$

So P^2 may be considered as a fraction of the population variance of the dependent variable explained by the variations of the regressors.

Barten (1987) have proposed four good properties for the coefficient of de determination

a) $\operatorname{plim} R^2 = P^2$

b) $0 \leqslant R^2 \leqslant 1$,

c) Si $b = 0$, $R^2 = 0$,

d) Si $\hat{u} = 0$, $R^2 = 1$.

These properties are the ones which one would like to see satisfied also by the coeficient of determination for regressions without intercept.

The econometric literature, as far as consulted by the author, presents the following five expressions for the coefficient of determination in this case.

$$R_1^2 = 1 - \frac{\hat{u}'\hat{u}}{y'Ay}, \quad R_2^2 = 1 - \frac{\hat{u}'\hat{u}}{y'y}, \quad R_3^2 = \frac{(y'A\hat{y})^2}{(y'Ay)(\hat{y}'A\hat{y})}$$

$$R_4^2 = \frac{(\hat{y}'A\hat{y})}{(\hat{y}'A\hat{y}) + \hat{u}'\hat{u}}, \quad \text{and} \quad R_5^2 = \frac{\hat{y}'\hat{y} - T\bar{y}^2}{y'Ay}.$$

One can easily verify that only R_3^2 and R_4^2 meet properties *a)* through *d)*.

In order to study the empirical performance of all of these statistics Monte Carlo simulations have been used where the DGP considered was

$$y_t = \beta_1 z_{1t} + B_2 z_{2t} + u_t, \qquad \text{para} \quad t = 1, \ldots, T$$

y $u \approx N\left(0_T, \sigma^2 I_T\right)$.

All simulations were carried for different values for P^2 and the multicolineality level which lie in $[0,1]$ using $T = 15$ and 1000 replications.

The results reveal that the best coefficients for the measure of goodness of fit in this case are R_3^2 (Heijmans y Neudecker, 1987) y R_4^2 (Barten, 1987).