Económica, La Plata, Vol. XXXIV, Nro. 1, Enero-Junio 1988

ACERCA DEL PROBLEMA DE LA MULTICOLINEALIDAD EN LA ESTIMACION DEL MODELO LINEAL *

JUANA J. BRUFMAN**

I. El Problema

Sea el modelo de regresión lineal uniecuacional:

$$Y_t = b_0 + b_1 X_{t1} + b_2 X_{t2} + \ldots + b_K X_{tk} + \mu_t \; ; t = 1, 2, \ldots T \quad (1)$$

ó, en notación matricial:

$$y = X\,b + \mu \quad (2)$$

donde:

y: es el vector de T observaciones sobre la variable a explicar
X: es la matriz de T observaciones sobre K+1 variables explicativas ($X_{t0} = 1$ para todo t).
b: es el vector de K+1 parámetros a estimar.
μ: es el vector de T componentes aleatorias.

El problema de la multicolinealidad se presenta cuando las variables explicativas están vinculadas funcionalmente, ó bien, altamente correlacionadas.

(*) Agradezco a la Lic. Hildegart Ahumada y al Dr. Alfredo M. Navarro, por los valiosos comentarios efectuados, que contribuyen, sin duda, a una mejor presentación y comprensión del tema.

(**) Profesora titular de Estadística, Asociada de econometría, Facultad de Ciencias Económicas, U.B.A.

Si en la ecuación (1) dos o más variables X_k son linealmente dependientes, resultará que la matriz (X'X) será de orden K+1 y de rango menor que K+1. Por lo tanto, el determinante $|X'X| = 0$ y no existe la matriz inversa $(X'X)^{-1}$ que se utiliza para la estimación mínimo-cuadrática de b. Es éste un caso extremo, que difícilmente se presenta en los estudios empíricos; por ejemplo, al postular un modelo con variables "dummy" y término constante, si se introducen tantas variables "dummy" como alternativas presenta el atributo, se produce lo que se ha dado en llamar "la trampa de las variables dummy", que es simplemente un caso de extrema multicolinealidad.

En efecto; supóngase el modelo $Y_i = b_0 + b_1 X_{i1} + \mu_i$, siendo X_1: Ingreso e Y: Gasto en indumentaria, y en el que se desea medir además, la incidencia del atributo "sexo" sobre Y. Si se introducen las variables D_1 y D_2 con la siguiente asignación de valores:

	D_1	D_2
varón	1	0
mujer	0	1

el modelo resulta:

$$Y_i = b_0 + b_1 X_{i1} + c_1 D_{i1} + c_2 D_{i2} + \mu_i$$

y la matriz X:

$$X = \begin{vmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 1 & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & 0 & 1 \\ . & . & 0 & 1 \\ . & . & . & . \\ 1 & X_{n1} & 0 & 1 \end{vmatrix}$$

verificándose que la primer columna es combinación lineal de las columnas 3 y 4. Este caso de extrema multicolinealidad se corrige mo-

dificando la especificación del modelo: se deben utilizar tantas variables "dummy" como modalidades presenta el atributo, **menos una**. El modelo correcto sería:

$$Y_i = b_0 + b_1 X_{i1} + c_1 D_{i1} + \mu_i$$

siendo $D_{i1} = 1$, si la i-ésima observación corresponde a individuo varón, e igual a 0, si dicha observación pertenece a individuo mujer.

El extremo opuesto es el de variables explicativas linealmente independientes (ortogonales). Entonces, los coeficientes b_1, b_2, . . . b_K de la ecuación (1) pueden estimarse indistintamente, a partir del modelo de regresión múltiple propuesto, ó separadamente, efectuando regresiones independientes con cada una de las variables explicativas. Obsérvese que esta situación ideal permitiría medir la contribución de cada variable X_k en el comportamiento de Y, permaneciendo constante el resto de las variables explicativas del modelo.

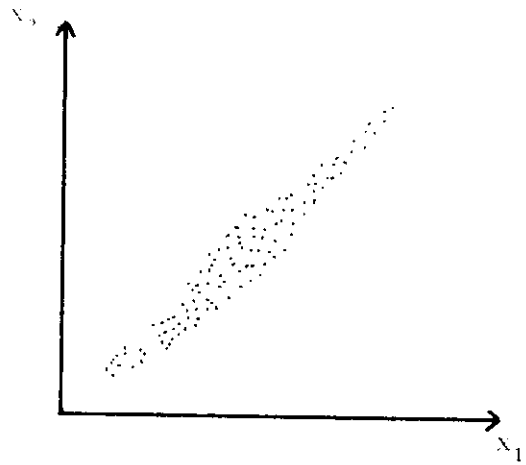
Analizaremos ahora el caso frecuente en las aplicaciones econométricas, en el que las variables explicativas están altamente correlacionadas, presentándose entonces el problema de la multicolinealidad. Por lo expuesto anteriormente, debe quedar claro que la multicolinealidad es una cuestión de "grado". Las variables en Economía no son ni linealmente dependientes (en forma exacta), ni tampoco linealmente independientes. Es decir, no debe pensarse en la multicolinealidad como un problema que está ó no presente; siempre debe esperarse cierto "grado" de multicolinealidad. Lo importante es poder precisar sus "alcances", a los efectos de una correcta apreciación de los resultados obtenidos en el proceso de estimación, y posterior utilización del modelo con fines de predicción.

II. Causas de Multicolinealidad.

● Tendencia de la mayoría de las variables económicas a evolucionar conjuntamente (al unísono), respondiendo a un factor común que las afecta a todas ellas. Por ejemplo, en épocas de prosperidad económica, se observa el crecimiento del Consumo, Ingreso, Inversión, precios y disminuye la tasa de desempleo; en épocas de depresión, se produce movimientos en contrario de las mismas variables. Es decir, existe un patrón intrínseco de comportamiento de las variables econó-

micas, que origina problemas de multicolinealidad cuando se utilizan dos ó más de ellas como explicativas. En estos casos suele decirse que la multicolinealidad es un problema **poblacional.**

● Cuando se utilizan datos muestrales (de sección ó corte transversal), el diseño de la encuesta puede resultar "pobre" en información. Supóngase que deba explicarse el consumo familiar de cierto bien, en función de X_1: nivel de ingreso; y X_2: índice de riqueza. Es sabido que, en términos generales, las familias de mayores ingresos muestran un mayor índice de riqueza. Graficando la información de estas dos variables, nos encontraremos con una nube de puntos en torno a una función lineal. Esto significa que ambas variables suministran prácticamente la misma información a los efectos de explicar el consumo del bien. Un diseño apropiado deberá incluir, para **cada** tramo de ingreso, datos referentes a familias de distinto nivel de riqueza, con lo que se logra enriquecer la información. En este caso, la multicolinealidad es un problema **muestral.**



● Inclusión en el modelo de una variable explicativa con gran cantidad de desfasajes. Si se pretende explicar una variable Y_t en función de X_t, X_{t-1}, . . ., X_{t-s}, es obvio que, por tratarse de una misma variable referida al momento t, t-1, t-2, . . . etc., habremos de encontrar problemas de multicolinealidad, originados, en este caso, en la **especificación** del modelo, máxime cuando las variables están altamente autocorrelacionadas.

III. Consecuencias de la Multicolinealidad

● Imposibilidad de medir separadamente la incidencia de cada variable explicativa sobre la variable a explicar. Recordemos el significado de los parámetros de un modelo de regresión, como el (1): un coeficiente b_k indica la variación promedio de Y, ante un incremento unitario de X_k, **suponiendo** el resto de las variables **constantes** (condición ceteris-paribus). Pero, si por efectos de la multilinealidad, algunas variables explicativas se mueven simultaneamente, ya no es posible tal interpretación; en este caso los coeficientes reflejan efectos combinados. Sin pérdida de generalidad, consideremos la siguiente relación lineal exacta entre X_1, X_2 y X_3:

$$X_{t1} = \lambda_2 X_{t2} + \lambda_3 X_{t3} \qquad (3)$$

Reemplazando (3) en el modelo (1) resulta:

$$Y_t = b_0 + b_1(\lambda_2 X_{t2} + \lambda_3 X_{t3}) + b_2 X_{t2} + b_3 X_{t3} + \ldots + \mu_t \qquad (4)$$

y reordenando términos:

$$Y_t = b_0 + (b_1\lambda_2 + b_2) X_{t2} + (b_1\lambda_3 + b_3) X_{t3} + \ldots + \mu_t \qquad (5)$$

Resulta claro que no podemos estimar separadamente los coeficientes b_1, b_2 .. . . b_k, (la matriz X'X sería singular), aunque sí las siguientes combinaciones:

$$b_0 ; (b_1\lambda_2 + b_2) ; (b_1\lambda_3 + b_3) ; b_4 ; \ldots ; b_K \qquad (6)$$

● Falta de precisión de las estimaciones mínimo-cuadráticas, por resultar con altas varianzas y covarianzas.

Supontamos el modelo lineal con sólo dos regresores, expresando en unidades centradas:

$$y'_t = b_1 x_{t1} + b_2 x_{t2} + \mu'_t$$

Si se tiene presente que el coeficiente de correlación lineal entre X_1 y X_2 es:

$$r_{12} = \frac{\Sigma x_{t1}\, x_{t2}}{(\Sigma x_{t1}^2 \;\; \Sigma x_{t2}^2)^{1/2}}$$

las varianzas y covarianzas de los estimadores mínimo-cuadráticos serán:

$$\text{var}(\hat{b}_1) = \frac{\sigma_\mu^2}{(1 - r_{12}^2)\,\Sigma x_{t1}^2} \quad ; \quad \text{var}(b_2) = \frac{\sigma_\mu^2}{(1 - r_{12}^2)\;\Sigma x_{t2}^2} \quad ;$$

$$\text{cov}\,(\hat{b}_1\, \hat{b}_2) = \frac{-\sigma_\mu^2\; r_{12}}{(1 - r_{12}^2)\,(\Sigma x_{t1}^2 \;\; \Sigma x_{t2}^2)^{1/2}}$$

Como puede observarse, las varianzas y covarianza son:

a) Directamente proporcionales al σ_μ^2.

b) Inversamente proporcionales a la variabilidad muestral de los regresores en torno a sus respectivas medias:

$$\Sigma x_{t1}^2 \;\; ; \;\; \Sigma x_{t2}^2 \;\; ; \;\; (\Sigma x_{t1}^2 \;\; \Sigma x_{t2}^2)^{1/2}.$$

c) Seriamente afectadas por la correlación muestral de las variables explicativas: r_{12}. Si $|r_{12}| \to 1$, las varianzas y covarianza pueden resultar excesivamente grandes, $(\to \infty)$ por lo que las estimaciones son de escasa precisión. Respecto a la covarianza, obsérvese que este resultado confirma lo indicado en el párrafo anterior: si la $\text{cov}(\hat{b}_1\, \hat{b}_2)$ es muy alta, no pueden estimarse separadamente, con suficiente precisión, los coeficientes b_1 y b_2, por falta de variación independiente de X_1 y X_2.

Si el modelo contiene más de dos regresores, se requiere un análisis más detallado, a fin de indagar sobre las consecuencias de la multicolinealidad en las varianzas y covarianzas de los estimadores. Ya no basta con los coeficientes de correlación simple r_{ij} entre el par de variables X_i, X_j, puesto que pueden presentarse relaciones cuasi lineales que involucran más de 2 variables explicativas.

Consideremos el modelo (1), expresado en unidades centradas y estadarizadas. La matriz $X = [x_1\; x_2 \; \ldots \; x_k]$ está formada por vectores-columna normalizados, es decir, de longitud unitaria, y la matriz $(X'X)$ resulta ser la matriz de correlación $R = r_{ij}$, cuyos elementos

representan los coeficientes de correlación lineal **simple** entre cada par de variables X_i, X_j.

$$R = \{r_{ij}\} = \begin{vmatrix} 1 & r_{12} & r_{13} & \cdot & \cdot & \cdot & r_{1K} \\ r_{12} & 1 & r_{23} & \cdot & \cdot & \cdot & r_{2K} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ r_{1K} & r_{2K} & r_{3K} & \cdot & \cdot & \cdot & 1 \end{vmatrix}$$

La matriz de varianzas-covarianzas de los coeficientes $b_1^{*\,1}$ del modelo es:

$$\Sigma \hat{b}^* \hat{b}^* = \sigma_\mu^2 \; R^{-1} = \sigma_\mu^2 \; \{r^{ij}\} = \sigma_\mu^2 \begin{vmatrix} r^{11} & r^{12} & \cdot & \cdot & \cdot & r^{1K} \\ r^{12} & r^{22} & \cdot & \cdot & \cdot & r^{2K} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ r^{1K} & r^{2K} & \cdot & \cdot & \cdot & r^{KK} \end{vmatrix}$$

De modo que:

$$\text{var}\, \hat{b}_1^* = \sigma_\mu^2 \, r^{ii} \; ; \; \text{cov} \hat{b}_1^* \, \hat{b}_j^* = \sigma_\mu^2 \, r^{ij} \tag{7}$$

Se prueba que[2]:

$$r^{ii} = \frac{|R_{ii}|}{|R|} = \frac{1}{1 - R_i^2} \tag{8}$$

$$r^{ij} = \frac{|R_{ij}|}{|R|} = - \, r_{ij.} \, (r^{ii} \, r^{jj})^{1/2} \tag{9}$$

donde: $|R_{ij}|$ es el cofactor de r_{ij} ; $R_i^2 = R^2_{X_i . X_1 X_2 \cdots X_K}$ es el coeficiente

(1) El uso del asterisco obedece al cambio de escala en las variables del modelo. La relación entre los coeficientes es: $b_i = b_i^* \, \sigma_Y / \sigma_{X_i}$. Este cambio de escala no afecta el análisis que sigue, ya que solamente indagamos sobre los efectos de la multicolinealidad.

(2) Véase, por ejemplo, E. Malinvaud: Métodos Estadísticos de la Econometría, Ariel, Barcelona, 1967.

de determinación lineal entre x_i y las restantes variables explicativas del modelo; y r_{ij} es el coeficiente de correlación lineal **parcial** entre x_i y x_j, de orden K-2, es decir, quitada la influencia de las restantes K-2 variables explicativas.

De las expresiones (7), (8) y (9) surge:

a) La varianza de un coeficiente $\hat{b}_i^*$ se verá incrementada cuando sea R_i^2

b) La covarianza entre $\hat{b}_i^*$ y $\hat{b}_j^*$ será mayor, cuando el coeficiente de correlación parcial entre x_i y x_j sea alto (próximo a la unidad en valor absoluto), pudiendo ser positiva o negativa según el sentido de la relación entre dichas variables.

En síntesis, cuando el modelo tiene más de dos variables explicativas, deben analizarse no sólo los coeficientes r_{ij}, sino los r_{ij} y la correlación entre cada una de las variables y el resto de los regresores, es decir, la estructura de la matriz R.

● Inestabilidad de las estimaciones: El agregado ó quita de pocas observaciones, modifica sensiblemente los resultados, no sólo en magnitud, sino también en signo.

● Como consecuencia de las altas varianzas, las estimaciones resultan no significativas: los valores empíricos del test "t" son en general bajos, e inducen a aceptar con frecuencia la hipótesis nula de "no significación" de las variables involucradas; en otros términos, los tests "t" pierden potencia.

IV. Diagnóstico de la Multicolinealidad

Citaremos a continuación algunas técnicas de diagnóstico comunmente utilizadas, efectuando paralelamente las críticas pertinentes.

● Si coeficientes asociados a variables importantes del modelo resultan no significativos (t bajos), y al mismo tiempo se ha obtenido

un coeficiente de determinación alto, ello suele tomarse como síntoma de multicolinealidad, y más aún, la multicolinealidad se utiliza como justificación de tales resultados. Al respecto, cabe recordar que las varianzas altas de los coeficientes (y por lo tanto t bajos) pueden deberse a un σ_μ^2 excesivamente grande, como así también poca variabilidad de las variables explicativas.

Como consecuencia de un uso indiscriminado de este criterio, el econometrista tiende a omitir variables que, desde el punto de vista teórico, son importantes en el modelo, pero estadísticamente resultaron "no significativas"; debe tenerse presente que en esa forma se incurre en "errores de especificación", originando sesgos y hasta inconsistencia en las estimaciones de las restantes coeficientes del modelo.

• Análisis de la matriz de correlación R y de su inversa R^{-1}. La observación de los coeficientes r_{ij} suele utilizarse como criterio para detectar la multicolinealidad: si algunos de ellos son mayores de 0.8 ó 0.9, se dice que hay problemas de multicolinealidad. Sin embargo, a menos que el modelo contenga sólo dos variables explicativas, tal criterio resulta restrictivo e induce a conclusiones erróneas.

Debe recordarse que el coeficiente de correlación simple r_{ij} mide la fuerza de la asociación lineal entre el **par** de variables X_i, X_j, ignorando las restantes variables. Cuando el modelo contiene más de dos variables explicativas, pueden existir relaciones lineales que involucren 3 ó más variables, que no pueden detectarse a partir de los coeficientes de correlación simple. Tan es así, que en muchos estudios empíricos se han obtenido matrices R con coeficientes r_{ij} bajos, (0.5; 0.4; etc.), y sin embargo se detectaron problemas serios de multicolinealidad.

El valor del determinante de R, es usado como criterio para detectar multicolinealidad. Es sabido que $|R|$ está comprendido entre 0 y 1: Si las variables del modelo son linealmente independiente, todo $r_{ij} = 0$, y $|R| = 1$; si existe por lo menos una dependencia lineal exacta entre los regresores del modelo, entonces $|R| = 0$. Por lo tanto, cuanto más próximo a 0 se halle $|R|$, cabría esperar problemas de multicolinealidad.

Si bien este criterio preanuncia el problema, debe tenerse presente que valores bajos de $|R|$, pueden originarse en muchos tipos de multicolinealidad: es decir, el criterio no permite identificar las variables involucradas en la ó las relaciones lineales existentes.

Los elementos de la matriz R^{-1} proporcionan también información general sobre la multicolinealidad. Recuérdese la relación (8);

$$r^{ii} = \frac{1}{1-R_{i.}^2} \qquad R_{i.}^2 = 1 - \frac{1}{r^{ii}} \tag{10}$$

Valores altos de R_i^2 revelan que X_i está fuertemente asociada a las restantes variables. Ahora bien; si existen **varias** dependencias cuasi-lineales, las variables explicativas de estas regresiones auxiliares, estarán a su vez, afectadas por multicolinealidad, y entonces será imposible detectar la naturaleza de las relaciones entre variables explicativas, a partir de los coeficientes de regresión calculados. Es decir, la debilidad de este criterio radica en la imposibilidad de distinguir entre **varias** relaciones lineales coexistentes.

Los elementos no diagonales de R^{-1}, r^{ij}, están ligados a los coeficientes de correlación parcial $r_{ij.}$ según la relación (9):

$$r_{ij.} = - r^{ij} (r^{ii} r^{jj})^{1/2}$$

Valores altos de $r_{ij.}$ revelan que las variables X_i y X_j, depuradas del efecto de las restantes K-2 explicativas, presentan un grado importante de asociación lineal.

Belsley, Kuh y Welsch (1980), hacen notar, sin embargo, que **todos** los coeficientes de correlación parcial tienden a ± 1, cuando hay problemas de multicolinealidad. Así, $r_{ij.}$ puede estar próximo a 1, sin que las variables X_i y X_j estén involucradas en una relación lineal; en otras palabras, estos coeficientes carecerían de poder discriminatorio.

● Test de Farrar y Glauber.

En 1967, D.E. Farrar y R.R. Glauber idearon un procedimiento para detectar problemas de multicolinealidad, basado en la información muestral contenida en la matriz R y su inversa R^{-1}. El mismo consiste en la realización de tres tests consecutivos, según la siguiente secuencia:

1) Para determinar la **presencia** de multicolinealidad, se calcula |R| y su transformación logarítmica:

$$-[T - 1 - 1/6\,(2K + 5)] \ln |R|$$

que admite una distribución Ji-cuadrado con ϑ grados de libertad, siendo $\vartheta = {}^{1}/_{2}\,K\,(K-1)$.
Si el valor empírico del estadístico es mayor que el valor crítico, según el nivel de significación prefijado, diremos que hay un problema serio de multicolinealidad.

2) Para **localizar** el problema de la multicolinealidad, es decir detectar cuál ó cuales son las variables afectadas, se calcula el estadístico:

$$w_i = (r^{ii} - 1)\,\frac{T - K}{K - 1}$$

que se distribuye como una $F_{K-1\ ;\ T-K}$. (Recuérdese la relación entre r^{ii} y R_i^2.). En síntesis, este es un test para contrastar la significatividad de las regresiones auxiliares. Si el valor empírico de w_i es superior al valor crítico de la distribución F, al nivel de significación fijado, diremos que la variable X_i está altamente correlacionada con las restantes variables explicativas del modelo.

3) Finalmente, a los efectos de descubrir el "patrón" de multicolinealidad, se utiliza el estadístico

$$r_{ij.}\ \sqrt{T-K}\ /\ (1 - r_{ij.}^2)^{-1/2}$$

con distribución t de Student con T-K grados de libertad, realizándose un test de dos colas en la forma habitual, y al nivel de significación fijado de antemano: valores empíricos pertenecientes a la zona crítica de la distribución indicarían que las variables X_i y X_j, libres del efecto de los restantes regresores, se hallan altamente colineadas.

El procedimiento de Farrar y Glauber ha merecido las siguientes consideraciones:

a) No es un test, en el sentido riguroso con que se utiliza este término en la teoría estadística; todo test requiere previamente la especificación de una hipótesis a contrastar. En este sentido, la

condición de Gauss-Markov:

rango (X) = K + 1

no es una hipótesis contrastable, pues si no se cumple, la estimación mínimo-cuadrática clásica se torna inválida. Por otra parte, no debe olvidarse, como se mencionó al principio de este trabajo que, en la mayoría de los casos, la multicolinealidad es un problema "muestral", originado por un diseño pobre de información; por lo que no es pertinente la realización de un test para inferir respecto a la población.

b) Los autores parten del supuesto de que la matriz X está integrada por una muestra de T observaciones independientes (las filas de la matriz), generadas por una distribución normal K-dimensional, siendo las columnas de X ortogonales. Todo ello permite la utilización de las distribuciones Ji-cuadrado, F y t, en los sucesivos tests. Este supuesto es discutible, por cuanto:
   i. exigiría que los regresores fuesen estocásticos. Recuérdese que en el modelo lineal estándar, las variables X_i se suponen fijas, en sucesivas muestras; (variables matemáticas).
   ii. es difícil, cuando se maneja información temporal, que las observaciones resulten independientes.
   iii. requiere una distribución K-dimensional para los regresores del modelo, que difícilmente se cumple, dado el carácter no-experimental de la información económica.

c) El determinante $|R|$ y los coeficientes $R_{i.}^2$ y $r_{ij.}$ utilizados en los tests de la Ji-cuadrado, F y t, respectivamente, no permiten diagnosticar la presencia de **varias** dependencias lineales **coexistentes** entre los regresores del modelo.

V. Nuevas Técnicas de Diagnóstico

● Exámen de las Raíces y Vectores Característicos de la matriz (X'X). Esta técnica de diagnóstico ha sido utilizada por Kendall (1957) y Silvey (1969). Constituye un avance respecto a las anteriores por cuanto no sólo detecta la presencia de multicolinealidad, sino que permite identificar las variables involucradas en una relación lineal.

El fundamento del método es el siguiente:

Sea $Q = \{q_i\}$ la matriz de vectores característicos de (X'X), asociados a las raíces características λ_i ; i = 1, 2, . . ., K. Es decir:

$$(X'X)\, q_i = \lambda_i\, q_i$$

$$q'_i q_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad (11)$$

Se verifican entonces las siguientes relaciones:

$$Q'(X'X)Q = \Lambda = \begin{vmatrix} \lambda_1 & & & & \\ & \lambda_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \lambda_K \end{vmatrix} \qquad (12)$$

$$q_i'\,(X'X)q_i = \lambda_i \qquad (13)$$

Veamos ahora cómo la magnitud de una raíz característica se relaciona con el problema de la multicolinealidad.

Supóngase que $\lambda_i = 0$; debe verificarse entonces, por (13):

$$X\, q_i = [x_1\ x_2\ .\ .\ x_K] \begin{vmatrix} q_{1i} \\ q_{2i} \\ \\ q_{Ki} \end{vmatrix} = 0$$

ó sea:

$$q_{1i}x_1 + q_{2i}x_2 + . \ . \ . + q_{Ki}x_K = 0 \qquad (14)$$

Es decir, la raíz $\lambda_i = 0$ y su vector asociado q_i permiten identificar una dependencia lineal exacta entre las variables del modelo: es el caso de extrema multicolinealidad. Si λ_i es pequeña, la (14) se verifica "aproximadamente", por lo que estamos ante una relación cuasi-lineal. Por cada raíz característica pequeña de la matriz (X'X), habremos identificado así, una relación lineal aproximada

Además, de la (12):

$$(X'X)^{-1} = Q \Lambda^{-1} Q' = \sum_{1}^{K} \lambda_i^{-1} q_i q_i' \qquad (15)$$

y la varianza de un coeficiente b_i resulta:

$$\text{var } \hat{b}_i = \sigma_\mu^2 \left| \frac{q_{i1}^2}{\lambda_1} + \frac{q_{i2}^2}{\lambda_2} + \ldots + \frac{q_{iK}^2}{\lambda_K} \right| \qquad (16)$$

Esta expresión permite apreciar la contribución de las raíces y vectores característicos en la precisión de las estimaciones. Se observa que la varianza de un $\hat{b}_i$ depende de σ_μ^2, de **todas** las raíces características y de los elementos i-ésimos de los vectores asociados a cada raíz. En primer lugar, como

$$\text{tr } (X'X) = \sum_{1}^{K} x_i' x_i = \sum_{1}^{K} \lambda_i \qquad (17)$$

y las $\lambda_i > 0$, por ser (X'X) definida positiva, cuanto mayor sea la variabilidad de las variables explicativas, mayores serán las λ_i, y por lo tanto mayor será la precisión del coeficiente $\hat{b}_i$, suponiendo constantes los restantes elementos de la (16). Además, si hacemos tr (X'X) = constante, lo que se logra normalizando los vectores x_i , ($x_i' x_i = 1$), cuando mayor sean las diferencias entre las raíces características, menor precisión tendrá $\hat{b}_i$: si λ_2, por ejemplo, es pequeña en relación a las demás raíces características, el término $\sigma_\mu^2 \, q_{i2}^2/\lambda_2$ puede resultar grande, y por ende, la varianza de $\hat{b}_i$.

Ahora bien; los elementos q_{ij} (j = 1, 2, . . ., K), también inciden en la varianza de $\hat{b}_i$; como Q es ortogonal ($\sum_{1}^{K} q_{ij}^2 = 1$), y siguiendo con el ejemplo anterior, un valor pequeño de q_{i2} puede anular los efectos de λ_2, y no afectar la precisión de $\hat{b}_i$.

Este análisis nos indica que, si bien una raíz característica pequeña, y su vector asociado permiten detectar relaciones lineales entre variables explicativas, **no siempre** se verá afectada la precisión de las estimaciones mínimo-cuadráticas clásicas.

El inconveniente que se señala para esta técnica de diagnóstico, es el no establecer qué se entiende por una raíz "pequeña".

- Descomposición de la matriz X en valores singulares.

Belsley, Kuh y Welsch (1980), proponen un método de diagnóstico similar al anterior, basado en la descomposición de la matriz X en "valores singulares". Los autores aconsejan trabajar con variables normalizadas, pero no centradas; es decir, la matriz X está formada por vectores de longitud unitaria.

Sea X una matriz de orden T.K; la máxima puede descomponerse en la siguiente forma:

$$X = U\,D\,V' \quad ; U_{T,K}\ ,\ D_{K,K}\ ,\ V_{K,K} \tag{18}$$

siendo:

$$U'U = V'V = I_K \tag{19}$$

y D una matriz diagonal con elementos $\mu_j > 0$; j = 1, 2, . . . , K, llamados valores singulares de X. En realidad, los elementos μ_j se relacionan con las raíces características de (X'X), pues se verifica que:

$$\mu_j = \sqrt{\lambda_j} \tag{20}$$

Además, V es la matriz de vectores característicos de (X'X), y U la de vectores característicos de (X X'), asociados a sus K raíces características diferentes de cero.

Respecto al problema de la multicolinealidad, la descomposición propuesta por los autores proporciona la misma información que el sistema de raíces y vectores característicos. Sin embargo, existen razones para preferir la descomposición en valores singulares: al trabajar directamente con la matriz X, en lugar de hacerlo con la (X'X), se obtienen ventajas de cálculo, por la mayor estabilidad numérica con que se efectúa tal descomposición.

Como se explicó anteriormente, en la mayoría de las aplicaciones econométricas, suelen presentarse relaciones lineales "aproximadas" que se detectarán por valores singulares "pequeños". Para responder a la pregunta: qué se entiende por valor singular "pequeño"[2], los autores definen el "índice condicionante", (condition index):

$$\eta_j = \frac{\mu_{m\acute{a}x}}{\mu_j} \quad ; \quad j = 1, 2, \ldots, K \tag{21}$$

Resultará $\eta_j \geqslant 1$ para todo j; la cota inferior corresponde a la máxima μ_j. Un valor singular pequeño **en relación** a $\mu_{máx.}$, determinará un índice condicionante alto. Habrá tantas relaciones lineales aproximadas, como índices elevados. Experimentalmente se ha podido determinar que índices de alrededor de 5 ó 10, corresponden a dependencias lineales débiles; mientras que valores de 30 a 100, revelan relaciones moderadas a fuertes; la ocurrencia simultánea de varios índices altos, permiten diagnosticar la existencia de más de una relación lineal aproximada.

Con el propósito de "medir" hasta qué punto las dependencias lineales afectan la precisión de las estimaciones, los autores utilizan la técnica de "descomposición de la varianza", esbozada por Silvey, en la siguiente forma:

Partiendo de la expresión $\Sigma_{\hat{b}\hat{b}} = \sigma^2_{\mu} (X'X)^{-1}$, aplican la descomposición en valores singulares de la matriz X, obteniéndose:

$$\Sigma_{\hat{b}\hat{b}} = \sigma^2_{\mu} (VD^{-1} V') \tag{22}$$

De donde:

$$\text{var } (\hat{b}_i) = \sigma^2_{\mu} \sum_{1}^{K} \frac{v^2_{ij}}{\mu^2_j} \tag{23}$$

siendo los v_{ij} elementos de la matriz V.

Es decir, la (23) descompone la varianza de $\hat{b}_i$ en una suma de componentes, cada una de las cuales está asociada a **uno y sólo uno de** los valores singulares μ_j.
Indicando con:

$$\theta_{ij} = \frac{v^2_{ij}}{\mu^2_j} \quad ; \quad \theta_i = \sum_{1}^{K} \theta_{ij}$$

se definen los coeficientes

$$\pi_{ij} = \frac{\theta_{ij}}{\theta_i} \quad ; \quad i,j = 1, 2, \ldots, K \tag{24}$$

que representan la proporción de varianza de $\hat{b}_i$, asociada con la compo-

nente j-ésima de su descomposición.

Si se disponen los π_{ij} en el siguiente cuadro de doble entrada:

Valor Singular Asociado	Proporciones de var $\hat{b}_1$	var $\hat{b}_2$	var $\hat{b}_K$
μ_1	π_{11}	π_{12}	π_{1K}
μ_2	π_{21}	π_{22}	π_{2K}
.	.	.	.
.	.	.	
μ_K	π_{K1}	π_{K2}	π_{KK}
	1.00	1.00	1.00

su lectura permite una rápida interpretación en la siguiente forma:
Si se observa una **alta** proporción de la varianza de dos o más coeficientes concentrada en un mismo valor singular **bajo**, ello constituye evidencia que la relación lineal está causando problemas de precisión en los $\hat{b}_i$.

En síntesis, la multicolinealidad "degrada" la calidad de las estimaciones, si se observa:

1) Un valor singular que origine un **alto** índice condicionante, asociado con
2) Altas proporciones en la descomposición de las varianzas de dos o más coeficientes[3]

La cantidad de índices condicionantes altos (mayores de 30), determina el número de relaciones lineales aproximadas; los coeficien-

(3) Ello es así, por cuanto en toda relación lineal quedan involucradas, por lo menos, dos de las variables que componen la matriz X.

tes cuyas varianzas presentan en su descomposición, una alta proporción (mayor de 0.50), concentrada en un determinado índice alto, identifican las variables involucradas en la respectiva relación lineal.

VI. Aplicación

La técnica descripta ha sido implementada con información de las siguientes variables de la Economía Argentina, período 1950-1972:

Y^{A} : Ingreso de los Asalariados

$Y^{No\ A}$: Ingreso de los No Asalariados

AF : Activos Financieros

Dichas variables fueron utilizadas como explicativas en una función Consumo[4]. La matriz (X'X) normalizada resultó:

1.0000000	0.9755803	0.9601641	0.9864861
0.9755803	1.0000000	0.9869128	0.9803713
0.9601641	0.9869128	0.9999999	0.9475847
0.0864861	0.9803713	0.9475847	0.9999998

Los valores singulares μ_j y números índices condicionantes η_j fueron:

$\mu_1 = 1.97955146$ $\eta_1 = 1.0$

$\mu_2 = 0.24810809$ $\eta_2 = 7.97858$

$\mu_3 = 0.13310963$ $\eta_3 = 14.87160$

$\mu_4 = 0.04580227$ $\eta_4 = 43.21951$

(4) H.L. Urbisaia y J.Z. Brufman: Econometría. Problemas y Ejercicios. Ed. Macchi. Buenos Aires 1985. Pág. 46.

De donde surge la existencia de una única relación lineal aproximada, (que corresponde a un $\eta > 30$):

$$-0,9999997 - 0.0002125\ Y^{A} + 0.0001412\ Y^{N\ o\ A} +$$

$$+0.0003913\ AF \simeq 0$$

Los coeficientes de esta ecuación son los elementos del vector característico asociado a μ_4^2.

La descomposición de la varianza resultó:

Valor Singular Asociado	Proporciones de			
	var $\hat{b}_0$	var $\hat{b}_1$	var $\hat{b}_2$	var $\hat{b}_3$
μ_1	0.000936	0.000276	0.000581	0.000458
μ_2	0.034763	0.003370	0.075889	0.034899
μ_3	0.435885	0.057447	0.021622	0.078259
μ_4	0.528417	0.938907	0.901908	0.886384
	1.00	1.00	1.00	1.00

Se aprecia que las tres variables explicativas están involucradas en la relación lineal; la precisión de las estimaciones se halla afectada por la multicolinealidad, por cuanto una alta proporción de sus respectivas varianzas, (> 0.50) está asociada al número condicionante alto (η_4).

BIBLIOGRAFIA

BELSLEY, D.A.; KUH, E. & WELSCH, R.E.: *Regression Diagnostic*. J. Wiley & Sons. New York, 1980.

FARRAR, D.E. & GLAUBER, R.R.: *Multicollinearity in Regression Analysis: The Problem Revisited*. Review of Economics and Statistics. 49, 1967.

FOMBY, T.C.; HILL, R.C. & JOHNSON, S.R.: *Advanced Econometric Methods*. Springer Verlag. New York, 1984.

JUDGE, C.; GRIFFITHS, W.; HILL, R.C.; LUTKEPOHL, H. & LEE, T.C.: *Introduction to the Theory and Practice of Econometrics*. J. Wiley & Sons. New York, 1982.

JUDGE, C.; W. GRIFFITHS; R.C. HILL; H. LUTKEPOHL & T.C. LEE: *The Theory and Practice of Econometrics*. 2º Edición. J. Wiley & Sons. New York, 1985.

KENDALL, M.G.. *A. Course in Multivariate Analysis*. Griffin. London, 1957.

LEAMER, E.E.: *Model Choice and Specification Analysis*. Handbook of Econometrics. Vol. I (Ed. Z. Griliches & M.D. Intriligator). North Holland. Amsterdam, 1983.

MADDALA, G.S.: *Econometría*. Mc Graw Hill. México, 1985.

SILVEY, S.D.: *Multicollinearity and Imprecise Estimation*. Journal of the Royal Statistical Society. Series B. 31, 1969.

ACERCA DELPROBLEMA DE LA MULTICOLINEALIDAD EN LA ESTIMACION DEL MODELO LINEAL

RESUMEN

En este trabajo se sintetizan aspectos relacionados con la presencia de Multicolinealidad en el Modelo de Regresión Lineal.

Se analizan sus causas, sus consecuencias desde el punto de vista de la calidad de las estimaciones mínimo-cuadráticas clásicas, como así también diferentes criterios utilizados para su diagnóstico.

En particular, se detallan las técnicas basadas en el análisis de raíces características, valores singulares y números condicionantes de la matriz de observaciones, con énfasis en su vinculación con la varianza de las estimaciones. Finalmente se ofrece una aplicación, con datos de la economía argentina.

ABOUT THE PROBLEM OF THE MULTICOLLINEARITY IN THE ESTIMATION OF THE LINEAR MODEL

SUMMARY

This paper summarizes various aspects related with the existence of Multicollinearity in the standard linear regression model. The core of the paper focuses on the causes for multicollinearity, its effects on the appraisal of least squares estimates, and diagnostic procedures. Details about these techniques, based on eigenvalues, singular values and condition numbers of the matrix of observations, are provided, stressing its relation with the variance of the estimates.

The last section of the paper presents an application with data of Argentina.