EFECTOS DE LA COLINEALIDAD EN EL MODELADO DE REGRESIÓN Y SU SOLUCIÓN

Dr. Jorge L. García A. [1], Ing. Hernando Chagolla G.[1], Dr. Salvador Noriega M.[2]

Resumen

La regresión lineal es una de las técnicas más empleadas cuando se busca determinar una variable dependiente en función de una o varias variables independientes; sin embargo, tradicionalmente se emplea la técnica de mínimos cuadrados ordinarios, la cual enfrenta problemas cuando las variables independientes presentan multicolinealidad; por lo cual en este artículo se describe el problema de la colinealidad y sus efectos en los modelos generados, se discuten las principales técnicas de diagnostico y se presentan los procedimientos más empleados para manejarla o eliminarla.

Palabras clave: colinealidad, regresión ridge, mínimos cuadrados ordinarios.

Abstract

The linear regression is one of the most used techniques for determinate the relation between a dependent variable and one or several independent variables; nevertheless, traditionally the least square technique is used, which faces problems when the independent variables present multicollinearity; that's why in this article we describe the problem of the collinearity between the independent variables and it's main effects in the regression model generated, the main techniques for diagnose it are discussed and appears the procedures for handle and sometimes to eliminate it.

Keywords: collinearity, ridge regression, ordinary least square.

1. Introducción

Frecuentemente existe la necesidad de explicar una variable o conjunto de variables en función de otras.[12] Cuando una variable es explicada por otras, se dice que existe una relación entre ellas; la primera se denomina variable dependiente (VD) y las segundas, variables independientes (VI). Uno de los métodos más comunes para encontrar los parámetros de las VI que explique la VD es la técnica de mínimos cuadrados ordinarios (MCO); sin embargo, uno de los principales supuestos en el modelado de regresión es que las VI no poseen ningún tipo de dependencia lineal entre ellas. Cuando una VI posee alta correlación con otra ú otras ó puede ser explicada como una combinación lineal de algunas de ellas, se dice que el conjunto de datos presentan el fenómeno denominado multicolinealidad, según Wang (1996); sin embargo, Kaciranlar y Sakallioglu (2001) aseguran que no existe una definición totalmente aceptada sobre

[1] Departamento de Ingeniería Industrial, Instituto Tecnológico de Querétaro.
Av. Tecnológico S/N Esquina con M. Escobedo, Col. Centro. CP.76000, Querétaro, Qro, México. Tel. (+52) 442 2163597 Fax (+52) 442 2169931. jlgarcia@itcj.edu.mx

[2] Instituto de Ingeniería y Tecnología, Universidad Autónoma de Cd. Juárez. Henry Dunant 4016, Zona Pronaf, Cd. Juárez, Chihuahua, México. C.P. 32310. Tel:(+52) 656 688-2100. snoriega@uacj.mx

este fenómeno, aunque el enfoque general que proporciona es semejante al anteriormente definido.

Según Akdeniz (2001), cuando se emplean los MCO en la estimación de los parámetros de regresión y existe el problema de multicolinealidad en las VI, se pueden observar problemas de inestabilidad de los mismos, signos incorrectos en los parámetros y frecuentemente elevados errores estándar, lo que conduce a generar modelos con muy poco poder explicativo o de difícil interpretación.

Para resolver el problema anterior se han propuesto varias técnicas que incluyen la detección y diagnostico del fenómeno de la multicolinealidad y su solución, sin que exista un procedimiento objetivo o generalmente aceptado, aunque bajo evaluaciones mediante simulación, unas técnicas son más eficientes que otras. Por ejemplo, Hoerl y Kennard (1970) han propuesto una metodología denominada ridge regresión (RR) donde se sacrifica sesgo de los parámetros por una reducción de error estándar de los parámetros estimados, Liu (1993) y Kaciranlar et al. (1999) han propuesto nuevos estimadores sesgados que mejoran al RR y otros han realizado simulaciones sobre la superioridad de algunas técnicas sobre otras en la estimación de parámetros que son estimados en presencia de colinealidad en las VI, Wichern y Churchill (1978), Delaney y Chatterjee (1986) y Krishnamurthi y Rangaswamy (1987) y Jahufer y Wijekoon (artículo aceptado para su publicación).

El objetivo de este artículo es presentar los principales efectos que tiene la multicolinealidad en la estimación de parámetros de regresión lineal y como puede ésta ser detectada o diagnosticada en las VI, así como los principales procedimientos adoptados para manejarla o eliminarla.

El artículo está organizado de la siguiente manera; después de esta introducción, en la segunda sección se discuten las principales consecuencias de la colinealidad en la regresión lineal y su impacto en la eficiencia de los modelos generados, en la sección tres se discuten las principales técnicas de detección y diagnóstico que se reportan en la literatura, en la cuarta se analizan las técnicas de corrección o manejo empleadas y finalmente, en la quinta sección se discuten los resultados.

2. Principales Efectos de la Colinealidad en Modelos de Regresión

Cuando se sospecha de la presencia de multicolinealidad en las VI, este fenómeno debe ser investigado antes de generar un modelo de regresión, ya que puede generar errores en los pronósticos y dificultar la interpretación de la importancia de cada una de las VI en el modelo. Según Wang y Akabay (1994), las principales consecuencias de las altas colinealidades entre las VI son las siguientes:

- En un modelo de dos variables, el error estándar de los coeficientes estimados es muy grande; esto es debido a que al coeficiente de variación tiene un factor de la forma $1/(1-r^2)$, donde r es el coeficiente de regresión entre las dos VI y su valor está en el intervalo [-1,1]. Este índice es comúnmente denominado factor de inflación de la varianza (FIV). Cuando r=0 no existe colinealidad, las VI son ortogonales y su FIV es igual a 1. A medida que el valor absoluto de r se incrementa en valor absoluto, es decir, existe una correlación negativa o positiva entre las variables, el FIV también se incrementa, ya que el denominador tiende a cero a medida que r tiende a uno (correlación perfecta). Algunos autores recomiendan que los FIV sean menores a 10, de lo contrario se concluye que existe multicolinealidad.
- Los coeficientes estimados pueden ser insignificantes o de signo contrario al esperado y consecuentemente son muy sensibles a cambios en los datos muestrales. Esto es debido a la colinealidad de las VI, entonces los errores estándar serán grandes y consecuentemente el estadístico de prueba *t* será pequeño. Los coeficientes estimados con error estándar muy grande serán inestables; además, una adición de nuevas observaciones o puntos muestrales provoca grandes cambios en los valores de los parámetros estimados y algunas veces en el signo.
- Cuando existe colinealidad en las VI es difícil estimar adecuadamente la importancia de

éstas en el modelo generado, especialmente cuando existe signo contrario al esperado en uno de los coeficientes estimados. Por ejemplo, se espera que a mayor calidad de un producto terminado, la demanda se incremente si el precio se mantiene constante; sin embargo, puede encontrarse mediante un modelo de regresión lineal empleado como pronostico, que a mayor calidad del producto la demanda disminuya, lo cual es ilógico.

- La colinealidad de las VI puede sugerir al usuario de los modelos generados que excluyan importantes variables en éstos. Sin embargo, este proceso puede generar modelos menos objetivos o que no representa la realidad, dado que estadísticamente no son suficientes.

Es importante señalar que la colinealidad de las VI no es la única fuente de inestabilidad y grandes errores estándar en los coeficientes estimados; cuando otros supuestos del modelado de regresión se han violado, esos errores estándar serán grandes también y los parámetros eran inestables.

3. Principales Técnicas de Detección

La literatura provee muchas técnicas para manejar y diagnosticar la presencia de la colinealidad, las cuales comprenden desde reglas de eliminación de variables al cálculo de índices complejos. Algunos de los más ampliamente usados son el análisis de la matriz de correlaciones de todas las VI, otros se basan en el análisis de la eigenestructura de los datos de la matriz **X**, incluyendo factores de inflación de la varianza, traza de $(\mathbf{X}'\mathbf{X})^{-1}$ y el número de condición; los cuales se discuten a continuación.

Cálculo de los coeficientes de correlación. En un modelo con solamente dos VI, se puede estimar su coeficiente de correlación para determinar el grado de colinealidad. En algunos casos la construcción de una matriz de correlación y la representación gráfica es de gran utilidad. Mason y Perreault (1991) recomiendan que sea eliminada una de las variables que tenga un coeficiente de correlación mayor a 0.8 con otras. Para conocer esas correlaciones, generalmente se construye una *matriz de correlaciones* como la que se indica en la Tabla 1, donde las variables se colocan en filas y columnas y sus intercepciones deben representar el

coeficiente de regresión lineal que obtienen. En este caso se presenta una matriz de correlaciones para un conjunto de datos en que se tienen dos VI y una VD, en este caso la variable X_1 tiene alta correlación con la variable X_2 (0.824215); por lo que de acuerdo a lo propuesto por Mason y Perreault (1991), una de las variables se puede eliminar. Obsérvese que los valores de la diagonal es un uno.

	X_1	**X_2**	**Y**
X_1	1.000000	**0.824215**	0.964615
X_2	**0.824215**	1.000000	0.891670
Y	0.964615	0.891670	1.000000

Tabla 1. Matriz de Correlaciones

Asimismo, es de gran utilidad la construcción de una matriz con los diagramas de dispersión de los datos. En la Figura 1 se ilustra el caso de un diagrama de dispersión para un conjunto de veinte observaciones, donde además se ha agregado una línea de ajuste obtenida por mínimos cuadrados y se puede observar la lejanía o cercanía a de los puntos a dicha línea. Debe mencionarse que este tipo de matrices de correlación y de dispersión son fácilmente generados por programas tradicionales como MINITAB, SPSS, NCSS y STATISTICA.

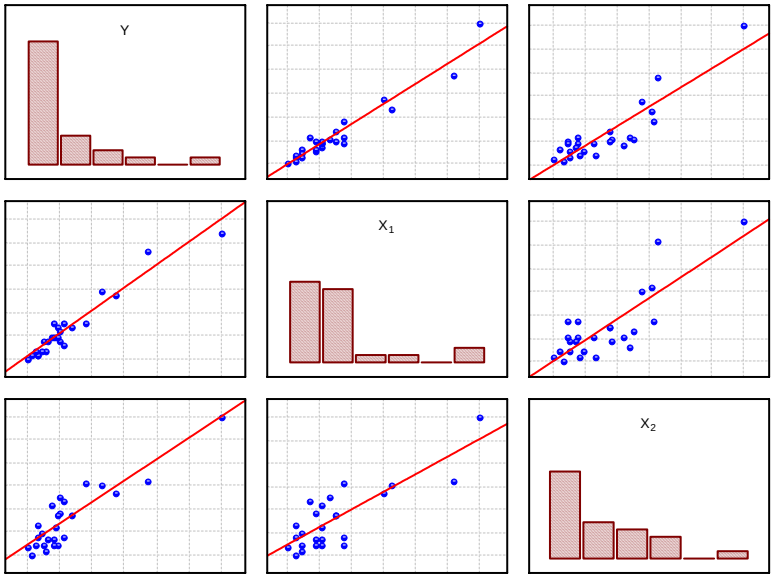

Figura 1. Matriz de Dispersión

Inspección de las R^2 y estadístico F. Cuando los valores de R^2 y el estadístico F son grandes, esto indica una fuerte relación entre las VI analizadas. Además, si algunos de los coeficientes son insignificantes (valores pequeños o muy grandes) y los valores de R^2 y F son grandes, esto es un indicativo de que algunas VI poseen alta correlación y se puede sospechar de la multicolinealidad.

La varianza de cada uno de los parámetros estimados puede ser obtenida por la ecuación (1), donde R^2_k es el coeficiente de determinación de la variable X_k como VD sobre las demás VI y el factor $(1- R^2_k)$ es conocido como factor de inflación de la varianza (FIV). Así, a medida que R^2_k incrementa su valor, el FIV tienda a cero y por estar éste en el denominador, la varianza del parámetro estimado se incrementa, dado que σ^2 se mantiene constante.

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum_i (X_{ki} - \bar{X})^2 (1 - R^2{}_k)} \qquad (1)$$

Por su parte Marquardt (1970) sugirió un valor máximo admisible para el FIV de 10 y para valores superiores a este límite, se considera que existen problemas de colinealidad. En la actualidad existe software que considera ese valor como límite, después de lo cual recomienda estimaciones sesgadas de los parámetros.

Por su parte, Willan y Watts (1978) han proveído una extensión de la interpretación que se tiene sobre los FIV, los cuales son los elementos de la diagonal de la matriz $(\mathbf{X'X})^{-1}$ y han desarrollado un análisis del determinante de la matriz $\mathbf{X'X}$. Especialmente, ellos han interpretado $FIV^{1/2}$ como una medida de perdida que tiene cada una de las variables debida a la multicolinealidad que pueda tener con otras VI. Esta medida tiene la ventaja de proporcionar información para cualquier variable en particular y es también más fácil de interpretar que el número de condición η propuesto por Belsley et al. (1980) y que se discute en párrafos posteriores.

Finalmente, Willan y Watts (1978) proponen que la raíz cuadrada del determinante de las VI sea usado como una medida de la eficiencia general del modelo de regresión generado. Este índice es interpretado como un radio de confidencia generado por el modelo generado en relación al generado sobre un diseño ortogonal hipotético. El valor de $|X'X|^{1/2}$ toma valores en el intervalo de [0, 1], esto es, si el índice es pequeño, significa que el modelo tiene poca eficiencia, ya que en un diseño ortogonal el valor sería 1.

Análisis del eigensistema. Por su parte Belsley et al. (1980) propuso un índice denominado número de condición (η), el cual está basado en la descomposición de valores singulares de la matriz de datos **X**, mismo que es definido como una relación entre el máximo eigenvalor y el mínimo, tal como se indica en 2. Algunos autores consideran que un $\eta < 5$ puede ser ignorado, para valores de $5<\eta<10$ existe una colinealidad débil, para valores

10<η<30 se califica como moderada, para 30<η<100 se considera fuerte y para η>100 se considera muy fuerte. Además, algunos programas computacionales que permiten la regresión ridge, como NCSS, consideran conveniente otro análisis diferente al de MCO cuando η>100.

$$\eta = \frac{\lambda_{\max}}{\lambda_{\min}} \qquad (2)$$

Otras técnicas de diagnostico incluyen el análisis del determinante de (**X´X**), el cual en presencia de colinealidad tiene un valor pequeño y valores elevados en los elementos de la diagonal de la inversa de la matriz de correlación simple; dado que σ^2 es constante, se puede observar en 3 que la varianza esta directamente relacionada con los elementos de la diagonal de $(\mathbf{X´X})^{-1}$.

$$\mathbf{Var\ (\beta) = \sigma^2 (X´X)^{-1}} \qquad (3)$$

4. Técnicas de corrección o manejo de la colinealidad

Muchos investigadores se han planteado técnicas y algoritmos parar corregir la colinealidad en los datos; sin embargo, algunos procedimientos funcionan en un modelo, mientras que en otros no. Wang (1996) propone las siguientes reglas para tratar la colinealidad de los datos.

Transformación de las variables por diferenciación. En algunos casos, la diferenciación consecutiva de cada variable del conjunto de datos puede reducir el impacto de la multicolinealidad. Así por ejemplo, la variable dependiente puede ser recalculada como $y_t = Ln(y_t) - Ln(y_{t-1})$ y también para cada una de las VI en la matriz **X**, $x_t = Ln(x_t) - Ln(x_{t-1})$.

Incorporación de información priori en el modelo. En este caso se pretende incorporar información o valores que han sido estimados en modelos anteriores en el nuevo modelo, la cual puede ser para cualquiera de los regresores. Por ejemplo, puede saberse por medio de un modelo anterior que el valor de β_1 sea un dos.

Agregar datos adicionales o nuevos en la muestra. Algunas veces el problema de colinealidad puede ser eliminado mediante la obtención de una nueva muestra u obteniendo más información para la ya existente. Este procedimiento tiene el impedimento de que muchas veces la obtención de nueva información no tiene fundamento económico.

Eliminar variables del análisis. Este procedimiento cosiste en eliminar una o más variables correlacionadas del análisis. Para la determinación de las variables que se integrarán en el nuevo modelo, generalmente se emplean técnicas de análisis multivariable, como el análisis factorial, donde en base a los eigenvalores de la matriz **X** se estima el poder de explicación de cada una de las VI. Este enfoque es aceptado por ser reduccionista y simplificar el modelo, sin embargo reduce el rango de **X** y esto lo puede convertir en una técnica que genere un modelo con menor poder explicativo.

Sin embargo, muchos otros autores han propuesto sacrificar ciertas características de los estimadores obtenidos mediante MCO, como es el caso del sesgo. En la figura 2 se observa el caso en que se tiene un parámetro β insesgado pero con un error estándar muy grande, mientras que en la figura 3 se observa el caso de un parámetro β que es sesgado pero que tiene un menor error estándar.

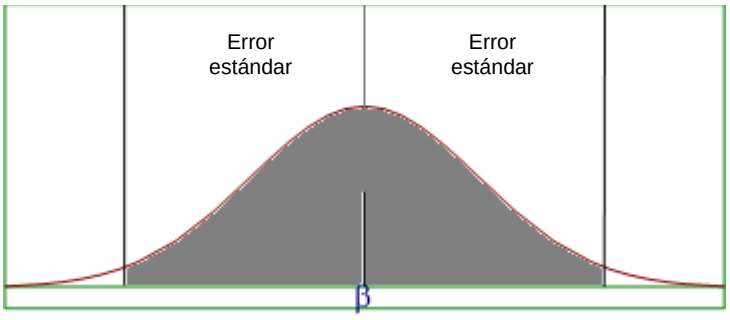

Figura 2. Parámetro β insesgado

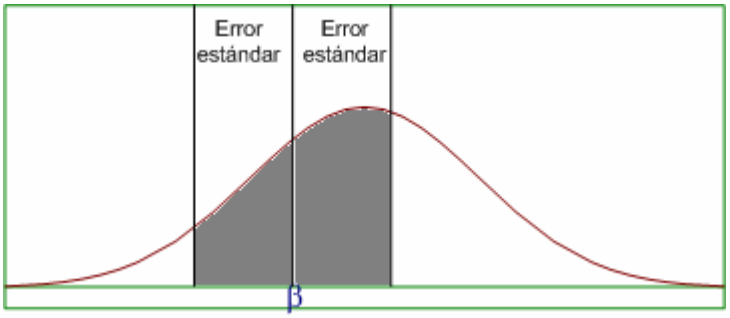

Figura 3. Parámetro β sesgado

La principal técnica empleada para obtener parámetros sesgados fue propuesta por Hoerl y Kennard (1970) y se denominada regresión ridge, donde se agrega un sesgo a los parámetros estimados con la finalidad de reducir el error estándar de éstos, donde agrega una matriz constante **k**, con valores que se encuentran entre 0 y 1. En el caso en que los valores de **k** son 0, entonces RR es igual a MCO. Lógicamente, según el teorema de Gauss-Markov, cuando los parámetros son obtenidos por MCO, éstos son insesgados; entonces a medida que los valores de **k** crecen, el sesgo de los parámetros estimados por RR también crece, así que se pretende minimizar los valores de **k** para minimizar el sesgo y a su vez el error estándar. Los parámetros del modelo pueden ser estimados según (4).

$$\beta = (X'X + kI)^{-1} X'Y \qquad (4)$$

Otra técnica ha sido propuesta por Liu (2003) donde se mejoran los parámetros estimados por RR; este método considera que aún después de admitir un sesgo mediante k en RR, se requiere de un segundo parámetro para disminuir aun más los efectos de la colinealidad, al que denomina *d*. Las fórmulas para obtener el estimador de Liu se listan a continuación en (5), (6) y (7).

$$\beta_{k,d} = (X'X + kI)^{-1} * (X'Y - d\beta) \qquad (5)$$

$$k = \frac{\lambda_1 - 100 * \lambda_p}{99} \tag{6}$$

$$d = \frac{\sum_{i=1}^{p} ((\lambda_i (\sigma^2{}_R - k\alpha^2{}_{Ri})) / (\lambda_i + k)^3}{\sum_{i=1}^{p} ((\lambda_i (\lambda_i \alpha^2{}_{Ri} + \sigma^2{}_R)) / (\lambda_i + k)^4} \tag{7}$$

5. Resultados y Conclusiones

En artículo se han discutido los principales efectos que tiene la multicolinealidad de las variables independientes en la eficiencia de los parámetros estimados mediante mínimos cuadrados ordinarios, se han planteado los principales procedimientos de diagnóstico reportados en la literatura y se han expuesto las principales técnicas para manejarla o corregirla; por lo que se concluye que una vez detectada la multicolinealidad, deben seguirse procedimientos de alternos a mínimos cuadrados, tales como aquellos que reportan parámetros sesgados pero con menor error estándar.

6. Bibliografía

Akdeniz, F. 2001. The examination and analysis of residuals for some biased estimators in linear regression. Communications in Statistics: Theory and Methods, 30: 1171-1183.

Belsley, D. A., Kuth, E. and Welsh, R. E. 1980. Regression diagnostics –identifying influencial data and sources of collinearity-. New York, Jhon Wiley & Sons, Inc.

Delaney, N. J. and Chatterjee, S. 1986. Use of the bootstrap and cross validation in ridge regression. Journal of Business and Economic Statistics, 4: 255-262.

Hoerl, A. E. and Kennard, R. W. 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 12: 55-67.

Jahufer, A. and Wijekoon, P. A Monte Carlo evaluation of new biased estimators in regression model. Article accepted by Indian Journal of Statistics (To be published).

Kaciranlar, S. and Sakallioglu, S. 2001. Combining the LIU estimator and the principal component regression estimator. Communications in Statistics: Theory and Methods, 22: 393-402.

Kaciranlar, S., Sakallioglu, S., Akdeniz, F., Styan, G. P. H. and Werner, H. J. 1999. A new biased estimator in linear regression and a detailed analysis of the widely-analyzed dataset on Prtland

cement. Sankhya, Series B. Indian Journal of Statistics, 61: 443-459.

Kinshnamurthi, L. and Ranganwamy, A. 1987. The equity estimators of marketing research. Marketing Science, 6: 336-357.

Liu, H. 1993. A new class of biased estimate in linear regression. Communications in Statistics: Theory and Methods, 22: 393-402.

Liu, K. 2003. Using Liu-Type estimator to combat collinearity. Communications in Statistics: Theory and Methods. 32(5): 1009-1020.

Marquardt, D. W. 1970. Generalized inverses, ridge regression and linear biased estimation. Technometrics, 12: 591-612.

Mason, C. C. and Perreault, W. 1991. Collinearity, power and interpretation of multiple regression analysis. Journal of Marketing Research. 28: 268-280.

Wang, G. C. S. 1996. How to handle multicollinearity in regression modelling. The Journal of Business Forecasting, Spring.

Wang, G. C. S. and Akabay, C. 1994. Autocorrelation: problems and solution in regression analysis. The Journal of Business and Forecasting Methods and Systems. 13(4): 18-26.

Willan, A. R. and Watts, D. G. (1978). Meaningful multicollinearity measures. Technometrics, 407-412.

Winchern, D. W. and Churchill, G. A. 1978. A comparison of ridge regression estimators. Technometrics, 20: 301-311.