COMPUTACIÓN Y MATEMÁTICA APLICADA
APLLIED COMPUTATION AND MATHEMATICS

ARTÍCULO ORIGINAL

La Multicolinealidad en modelos de Regresión Lineal Múltiple

The Multicollinearity in Multiple Lineal Regression Models

Juan del Valle Moreno[1] y C. Walkiria Guerra Bustillo[2]

RESUMEN. En este trabajo se aborda la problemática de la Multicolinealidad entre las variables regresoras en el Modelo de Regresión Lineal Múltiple. Diversas son los ambientes en las ciencias agrícolas donde esta dificultad puede presentarse. En este trabajo para crear una situación de multicolinealidad en los datos, necesarios para el estudio, se generaron variables explicativas de modo que entre dos de ellas existiera cierto grado de dependencia es decir "casi combinación lineal". Creadas estas condiciones se establece el análisis de la multicolinealidad transitando por: síntoma, diagnóstico y tratamiento. Para la sintomatología se analizan las correlaciones por pares de variables regresoras, la prueba F parcial y total sobre los coeficientes de regresión, el error estándar de cada estimador y coeficiente de determinación, entre otros aspectos. Para el diagnóstico se usó la diagonalización de la matriz de correlaciones y el examen de los últimos valores propios que brinda una información precisa. Para el tratamiento se aborda la Regresión Ridge y la Regresión sobre Componentes Principales, las cuales resultan efectivas para describir con exactitud y precisión los estimadores en el Modelo de Regresión Lineal Múltiple.

Palabras clave: multicolinealidad, Regresión Lineal Múltiple, Regresión Ridge, Regresión sobre Componentes Principales.

ABSTRACT. This work is about the Multicollinearity problem between the regressive variables in a Multiple Lineal Regression Model. There are many ambiences in the Agro sciences in which this problem can be found. So, in this work and to create, in the data which is really necessary for this study, a multicollinearity situation, explicative variables were generated in a way that at least two of them had a certain degree of dependency, that is to say ¨an almost lineal combination. Once these conditions are created, an analysis of the Multicollinearity is established taking into consideration *the symptom, the diagnose and the treatment.* For *the symptom*, the correlation between regressive variable pairs, the F partial and total test on regressive coefficients, the standard error in each estimator and the determination of the coefficient, among others, were deeply analyzed. For *the diagnose*, a diagonalization of the correlation matrix was used as well the study of the last own values that gives us a precise information. For *the treatment*, the Ridge Regression and the Regression of Principal Components were used which resulted very efficient to accurately describe the estimators in the Multiple Lineal Regression Model.

Keywords: Multicollinearity, Multiple Lineal Regression Model, Ridge Regression, Regression of Principal Components.

INTRODUCCIÓN

El tratamiento de la Multicolinealidad ha sido abordado ampliamente en la literatura estadística desde que Hoerl y Kennard (1970) introdujeron el estimador de Regresión Ridge, de modo que en la actualidad es ya un tópico tratado por muchos autores en distintas publicaciones tales como Gunts y Mason (1980), Draper y Smith (1981), Calero (1987), Bowerman (1990), Helsel (1992), Peña (1994), Gujarati (1995), Chacín (1998), entre otros. Para resolver el problema de la presencia de multicolinealidad entre las variables independientes en el modelo lineal de regresión, $y_i = ß_0 + ß_1x_{1i} + ... + ß_px_{pi} + \varepsilon_i$, i =1, 2,...,n. se han planteado distintos métodos.

En este trabajo se lleva a cabo el estudio de dos métodos para detectar y tratar de paliar su efecto, uno de ellos es la Regresión Ridge, el cual ha resuelto el problema de encontrar mejores estimadores de los parámetros del modelo y el otro es la Regresión sobre Componentes Principales, técnica multivariada de amplio uso en la actualidad.

Realizada una amplia revisión bibliográfica sobre esta temática se pueden citar algunas aplicaciones en investigaciones agrícolas:

Recibido 14/03/11, aprobado 20/07/12, trabajo 60/12, artículo original.

[1] M. Sc., Profesor, Universidad Agraria de La Habana, Facultad de Ciencias Técnicas, departamento de matemática, San José de las Lajas, Mayabeque, Cuba, E✉: juan@isch.edu.cu

[2] Dr. C. Prof. Titular, Centro Universitario Municipal Güines, Mayabeque, Cuba.

- Ingestión voluntaria de hierba ensilada, concentrada y peso vivo en ganado de carne (Rokk y Gill, 1990).
- Ritmo de producción de lana en cabras Merino con el tiempos de retención de la digesta (Smuts *et al.,* 1995).
- Condiciones corporales en cerdos (Charette *et al.,* 1996).
- Determinación de grasa añadida a la grasa pura de la leche en muestras de diferentes países de Europa (Lipp, 1996).
- Condiciones físico-corporales en período seco y hasta los 120 días de lactancia en vacas Holstein (Domecq *et al.,* 1997).
- Modelos Estadísticos-Matemáticos en el análisis de la curva de lactancia y factores que la afectan en el genotipo Siboney de Cuba (Fernández, 2004).

En estos trabajos investigativos se incluyen los análisis de Regresión Lineal Múltiple, Regresión Ridge y Regresión sobre Componentes Principales, así como Regresión Logística, Modelos de Calibración, Regresiones no Lineal y Modelos Lineales con efectos aleatorios y mixtos, entre otras técnicas estadísticas.

Se pudo constatar que existen pocos materiales dedicados a este aspecto en el ámbito agrario, por lo que se hace necesario que los profesionales dedicados al trabajo de la Matemática Aplicada en este sector, aborden esta temática con la óptica necesaria para satisfacer la expectativa que esto tiene en nuestro contexto.

MÉTODOS

Para hacer el estudio se estimaron los parámetros considerando el modelo:

$y_i = ß_0 + ß_1x_{1\,i} + ... + ß_px_{p\,i} + \varepsilon_i$, i = 1, 2,..., 20, p = 1, 2,..., 4, donde: $\beta_0=5$, $\beta_1=\beta_2=\beta_3=\beta_4=1$ en cuatro variantes de datos (A, B, C, D), según se resume en la Tabla 1.

TABLA 1. En cuatro variantes de datos

x1	x2	x3	x4
x2+x3+e	N(8; 1,4)	N(20; 2,2)	N(10; 1,7)
Y=5+x1+x2+x3+x4+e			

En este trabajo se analiza la efectividad de la Regresión Ridge y Regresión sobre Componentes Principales en la precisión y exactitud de los estimadores de los parámetros de la regresión lineal múltiple, con respecto a lo obtenido mediante el método de los mínimos cuadrados ordinarios.

El principal problema en la Regresión Ridge es encontrar aquel valor de k que compense el sesgo y la reducción de varianza. Un método usado fue la "traza ridge" (Zárate, 1985; Carvalho *et al.,* 1999), este es un gráfico que brinda una información valiosa de lo que ocurre al incrementarse k, a partir de valores próximo a cero se observa un cambio brusco, después se estabiliza, precisamente se debe seleccionar k en el momento de estabilidad. Durante este proceso el VIF (sigla en inglés del Factor de inflación de varianza) decrece, al principio rápidamente y luego de modo más gradual. La estimación de la varianza de los estimadores de los parámetros del modelo aumenta suavemente cuando se incrementa k (Rook y Gill, 1990).

En la Regresión sobre Componentes Principales, mediante el cálculo de los vectores y valores propios se crean nuevas variables ortogonales (Y_1, Y_2, Y_3 y Y_4), que son combinación lineal de las variables regresoras originales para cada variante, A, B, C y D.

Estas variables ortogonales se utilizan para obtener los estimadores de los parámetros del nuevo modelo usando MCO. Estos coeficientes son difíciles de interpretar por ser combinación de las variables originales, para solucionar esta dificultad se regresa a las variables originales. Es de destacar que siguiendo este método no queda eliminada ninguna de las variables estructurales del modelo, aunque sea necesario algún método de selección de variables, pues estos actúan sobre las componentes (Morzuch y Ruark, 1991).

Se usó el paquete estadístico Statgraphics Plus versión 4.1 (1999) y las posibilidades que tiene el MathCAD versión 5.0 sobre Windows (1996) asistente profesional en Matemática muy útil en trabajo con matrices.

Al aplicar a la Regresión el método de los Mínimo Cuadrado Ordinario, se obtuvieron:

- Los estimadores de los parámetros del modelo.
- Los errores estándar de los estimadores.
- La prueba t sobre los parámetros del modelo.
- La prueba F sobre el modelo (ANOVA).
- Coeficiente de determinación.
- Cuadrado medio del error, y el error estándar de la estimación.

Posteriormente se aplicó la Regresión Ridge y la Regresión sobre Componentes Principales. Una vez obtenidas todos los resultados con los diferentes procedimientos, se analizó para cada una de las variantes de datos, el comportamiento de las variables en cuanto a ser colineal con otra variable, para lo cual se usaron todas las herramientas, planteadas por los diferentes autores, para atenuar o resolver la multicolinealidad.

RESULTADOS Y DISCUSIÓN

En las Tablas 2 y 3 aparecen la media aritmética, desviación típica y coeficiente de variación de los estimadores de la Regresión Lineal Múltiple, haciendo uso de los MCO, en cada variante A, B, C y D.

TABLA 2. Media, desviación típica y coeficiente de variación de los estimadores

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Media aritmética	4,19	0,96	1,04	1,06	1,04
Desv. típica	1,26	0,72	0,75	0,70	0,06
C.V.	30,09%	74,79%	72,08%	66,19%	5,72%

TABLA 3. Media aritmética y desviación típica de los estimadores sin multicolinealidad

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Media aritmética	5,08	0,995	0,99	0,98	1,00
Desv. típica	0,19	0,0052	0,0056	0,0053	0,0058

En la Tabla 2 se resume, la media aritmética en cada uno de los estimadores MCO, la desviación típica así como el coeficiente de variación. También se resume en la Tabla 3 estos mismos estadígrafos para un juego de datos que tiene la característica de no presentar problemas de multicolinealidad.

Como se conoce los estimadores MCO poseen la propiedad de ser insesgados, esto es avalado por lo obtenido en las muestras representadas por las variantes de datos A, B, C y D. Los estimadores en promedio están próximos a β_0=5, β_1=β_2=β_3=β_4=1 (según construcción del modelo), este resultado es importante por cuanto confirma lo planteado teóricamente. Por otra parte también se presenta la desviación típica y el coeficiente de variación, esto permite ver el comportamiento en cuanto a la variabilidad de los estimadores $\hat{\beta}$ de los parámetros del modelo.

En la Tabla 2 se observa que la Desviación Típica en los estimadores de β_1, β_2, β_3, β_4, sobrepasa un valor aceptable (debe estar próximo a los que se sugieren en la Tabla 3 ó también el que se encuentra en la Tabla 2 (β_4). El estimador de β_4 corresponde a la única variable que se excluyó de la colinealidad, por lo que entre otras cosas cumple con el propósito de ser fuente de comparación, aquí es útil para ello, observe que la Desviación Típica de $\hat{\beta}_4$ es de 0,06. Se agrega el Coeficiente de Variación (C.V.) en todos los casos, este estadígrafo cumple con dar una información de la variabilidad en el conjunto de datos en forma relativa, muy útil para hacer comparaciones (Ostle, 1981).

En la Tabla 4 se recogen los resultados de media aritmética, desviación típica y coeficiente de variación.

TABLA 4. Información resumen de todas las variantes

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
			Regresión Múltiple (MCO)		
Media	4,19	0,96	1,04	1,06	1,04
Desviación	1,26	0,717	0,75	0,703	0,159
CV	30,09%	74,79%	72,08%	66,19%	5,72%
			Regresión Ridge		
Media	4,58	1,02	0,97	0,98	1,04
Desviación	1,152	0,122	0,16	0,12	0,06
CV	25,5%	11,85%	16,94%	12,3%	5,94%
			Regresión Componentes Principales		
Media	5,03	1,33	0,665	0,68	1,05
Desviación	1,12	0,039	0,06	0,03	0,06
CV	22,3%	2,98%	9,7%	5,11%	5,9%

Como se observa en la Tabla 4, los estimadores que se obtienen por la Regresión Múltiple en las variables que están afectadas por la colinealidad, en promedio tienden a estar aproximadamente sobre 1 en caso de los términos de las variables, y aproximadamente 5 en el caso del término independiente, esto confirma el hecho conocido de que los estimadores Mínimo Cuadrado Ordinarios son insesgados, es decir el valor esperado es precisamente el valor del parámetro que están estimando, en este caso se conoce ese parámetro pues se parte del modelo conocido. Por otra parte se observa que precisamente esos estimadores están alejados del parámetro en muestras individuales en los juegos de datos, esto se debe al hecho discutido hasta ahora de que los estimadores en presencia de multicolinealidad se vuelven inestables haciendo que sus varianzas sean grandes (Aksu y Gunter, 1994; Gómez, 1995, Gunter y Aksu, 1997).

Los estimadores en la Regresión Ridge se comportan en cuanto al insesgamiento en estos juegos de datos mejor que los de Mínimo cuadrados Ordinarios, pero la cualidad a resaltar en ellos es la precisión con que estiman a los parámetros del modelo, siendo realmente por teoría estimadores sesgados (Leiby y Adams, 1991; Lynch *et al.,* 1993; Parternak *et al.,* 1997).

En cuanto a los estimadores componentes principales, la cualidad a destacar en estos es que se han obtenidos en variables o regresoras incorrelacionadas por lo que la inflación de varianza para cada uno de ellos es exactamente uno, es decir la varianza no está inflada por lo que se convierten en mejores estimadores, de hecho son los más atractivos, eso se deja ver pues en la mayoría de los trabajos revisados en la bibliográfica resuelven el problema de la colinealidad recurriendo a los Componentes Principales (Ovalles y Collins, 1988; Aparicio, 1992; Lipp, 1996; Genachte *et al.,* 1996, Plotto y Anita, 1997), entre otros. La gran dificultad es que los estimadores no tienen interpretación hasta tanto no se retorne a las variables originales, y esto no está contemplado en muchos programas estadísticos. Una vez que se tiene la salida se procede a aplicar algún método de eliminación de variables, este no elimina las variables originales pues se aplica a las componentes y después se retorna, obteniéndose como resultado final todas las variables originales.

Las técnicas existentes para detectar y tratar la multicolinealidad no siempre logran el éxito teniéndose que recurrir a más de una de éstas, por lo que no se puede sugerir una receta para abordar el problema, pero si se proponen los siguientes pasos para evitar su efecto indeseable:

- Análisis de la matriz de correlaciones entre las variables explicatorias, como medida de diagnóstico de la multicolinealidad
- Resumen o sumario estadístico de ajuste del modelo como son: prueba F parcial y total.
- Error estándar de los estimadores y coeficiente de determinación.
- La prueba T sobre cada uno de los parámetros del modelo.
- Determinación y análisis de los valores y vectores propios.
- Identificación de las variables involucradas en la multicolinealidad, mediante la contribución proporcional de los componentes sobre el VIF.
- Determinación y análisis del Número de Condición (NC), Índice de Condición (IC) y el Factor de Inflación de Varianza (VIF), con el objetivo de precisar la existencia de una multicolinealidad, y posteriormente atenuarla imponiendo restricciones:
- Sobre las variables independientes añadiendo una constante a sus varianzas antes de resolver las ecuaciones normales (Regresión Ridge).
- Sobre las variables independientes ortogonalizándolas (Regresión sobre Componentes Principales).

REFERENCIAS BIBLIOGRÁFICAS

1. AKSU, C. & L. GUNTER: *Efficiency of combinations of forecast using inequality restricted least squares,* pp. 47-60, Economical & Financial Modelling, Spring, 1994.
2. APARICIO, R.; F. GUTIÉRREZ & R. MORALES: "Relationship between flavour descriptors and overall grading of analytical panels for virgin olive oil", *J. Sci. Food. Agric.,* 58(4): 555-562, 1992.
3. BOWERMAN, B.L. & T. O'CONNELL: *Linear Statistical Models: an Applied Approach*, 1024pp., PWS-Kent, Boston, 1990.
4. CARVALHO, C.G.P.; R. OLIVEIRA; D. CRUZ & D. CASALI: "Análise de trilha sob multicolinearidade em pimentao", *Pesq. Agropec. Bras.,Brasilia,* 34(4): 603-613, 1999.
5. CHACÍN LUGO, F.: *Análisis de Regresión y Superficie de Respuesta,* 274pp., Universidad Central de Venezuela, Caracas, 1998.
6. CHARETTE, R.; M. BIGRAS-POULIN & P. MARTINEAU: "Body condition evaluation in sows" *Livestock-Production-Science,* 46(2): 107-115, 1996.
7. Draper, N. & H. Smith: *Applied regression analysis,* Second Edition, John Wiley and Sons, Inc. New York, NY, USA, 1981.
8. DOMECQ, J.J.; L. SKIDMORE; W. LLOYD & B. KANEENE: "Relationship between body condition scores and milk yield in a large dairy herd of high yielding Holstein cows", *Journal off Dairy Science,* 80(1): 101-112, 1997.
9. FERNÁNDEZ, L.: *Modelos Estadísticos-Matemáticos en el análisis de la curva de lactancia y factores que la afectan en el genotipo Siboney de Cuba*, 100pp., **Tesis (presentada en opción al grado de Doctor en Ciencias Veterinarias)**, Instituto de Ciencia Animal, La Habana, Cuba, 2004.
10. GENACHTE, G. VAN DE; D. MALLANTS; J. RAMOS; A. DECKERS; J. FEYEN & G. VAN DE GENACHTE: "Estimating infiltration parameters from basic soil properties", *Hidrological Processes,* 10(5): 687-701, 1996.
11. GUJARATI, D. N.: *Econometría,* 2ª edición. ENSPES,Universidad de La Habana, Cuba, 1992.
12. GÓMEZ, S. M.J.: *Contribución al Análisis Multivariante directo del Gradiente mediante estudios combinados de Configuraciones Espaciales*, **Tesis Doctoral,** Universidad de Salamanca,Salamanca, España, 1995.
13. GUNST, F.R. & R. MASON: *Regression analysis and its Applications*, 402pp., A Data Oriented Approach, Marcel Dekker, New York and Basel. USA, 1980.
14. GUNTER, S.I. & C. AKSU: "The usefulness of heuristic N(E)RLS algorithms for combining forecasts", *J. Forecasting,* 16(6): 439-463, 1997.
15. HELSEL, D.R. & M. HIRSCH: *Statistical Methods in Water Resources,* 522pp., Elsevier, New York, USA, 1992.
16. HOERL, A.E. & W. KENNARD: "Ridge Regression; Apllications to nonorthogonal problems", *Technometrics,* 12:55-67, 69-82,1970.
17. LEIBY, J.D. & D. ADAMS: "The returns to agriculture research in Maine: the case of a small northeastern experiment station", *Northeast J. Agric Resour Econ.*,20(1): 1-14, April 1991.
18. LIPP, P.: "Comparison of PLS, PCR and MLR for the quantitative determination of foreign oils and fats in butter fats of several European countries by their triglyceride composition", *Zeitschrift fur Lebensmittel Untersuchung und Forschung,* 202(3): 193-198, 1996.
19. LYNCH, T.B.; A. MAX; E. BURKHART & J. LIU: "Prediction equations for centers of gravity and moments of inertia of loblolly pine stems", *For Sci, Bethesda, Md.: Society of American Foresters*, 39(2): 260-274, May 1993.
20. MORZUCH, B.J. & A. RUARK: Principal Components Regression to Mitigate the Effects of Multicollinearity. *Forest Science*, 37(1): 191-199, 1991.
21. OSTLE, B.: *Estadística Aplicada*, 629pp., Editorial Científico Técnico, La Habana, Cuba, 1981.
22. OVALLES, F. A. & E. COLLINS: "Variability of Northwest Florida Soils by Principal Component Analysis", *Soil. Sci. Soc. Am. J.,* 52: 1430-1435, 1988.
23. PASTERNAK, H. Z., E. SCHMILOVICH & Y. FALLIK: Edan. Ridge Regression for NIR analysis with multicollinearity, In: **3rd International Symposium on Sensors in Horticulture**, Tiberias, Israel, August 17-21. 1997.
24. PEÑA SÁNCHEZ DE RIVERA, D.: *Estadística, modelos y métodos,* Alianza Universidad Textos, Tomo II: Modelos Lineales y Series Temporales, Segunda edición, primera reimpresión, España, 1992.
25. PLOTTO, A.; N. AZARENKO; R. MCDANIEL; W. CROCKETT & P. MATTHEIS: "Eating quality of 'Gala' and 'Fuji' apples from multiple harvests and storage durations", *Hortscience,* 32(5): 903-908, 1997.
26. ROKK, A. J. & M. GILL: "Prediction of the voluntary intake of grass silages by beff cattle, 2. Principal Component and Ridge Regression Analysis", *Anim. Prod.,* 50: 439-454, 1990.
27. SMUTS, M.; H. MEISSNER &; B. CRONJE: "Retention time of digesta in the rumen: its repeatability and relationship with wool production of Merino rams", *J. anim. sci*., 73(1): 206-210, 1995.
28. ZÁRATE DE L., G.P. y B. DÍAZ: "Técnicas de Manejo y Puntos de Influencia y Multicolinealidad en regresión Lineal", *Agrociencia,* 61: 41-58, 1985.