Rev. Fac. Agron. (UCV) 39(3): 134-143. 2013.

Evaluación de la multicolinealidad en modelos de regresión lineal múltiple con presencia de valores atípicos

Danny A. Villegas[1*], Marisela Ascanio[2] y Margarita Cobo[2]

[1]Universidad Nacional Experimental de los Llanos Ezequiel Zamora. Programa de Ciencias del Agro y del Mar. Apdo 3350. Guanare, Portuguesa. Venezuela.
[2]Universidad Central de Venezuela, Facultad de Agronomía, Venezuela. Apdo. 4579. Maracay 2101, Aragua. Venezuela

RESUMEN

Se evaluó la multicolinealidad en modelos de regresión lineal múltiple con presencia de valores atípicos, para lo cual se utilizó un estudio de simulación con un modelo lineal con tres variables regresoras (X_1, X_2 y X_3) y una respuesta (Y) para cuatro tamaños de muestra (10, 20, 30 y 50), en donde las variables X_1 y X_2 fueron generadas partiendo de tres distribuciones teóricas continuas (uniforme, normal, y exponencial) y X_3 fue establecida como una combinación lineal ($X_3 = X_1 + 2*X_2 + e$), utilizando un modelo contaminado con un porcentaje dado de valores atípicos. Para evaluar la multicolinealidad se utilizó el índice de colinealidad (K). En ese sentido, se evidenció un potencial efecto del tamaño de muestra y del tipo de distribución teórica de los regresores del modelo sobre el índice de condición K, con el subsecuente efecto sobre el grado de multicolinealidad, lo que demuestra como la presencia de valores atípicos en la muestra no afecta la estructura del modelo, sino más bien a los parámetros del mismo. Finalmente, se sugiere un estudio más exhaustivo del error cuadrático medio del estimador de mínimos cuadrados ordinarios del modelo de regresión lineal múltiple en presencia de regresores colineales, toda vez que se observó como el error cuadrático medio tiende a cero para el estimador de la variable colineal, lo que pudiera sugerir el posible uso del error cuadrático medio del estimador de mínimos cuadrados ordinarios como una alternativa para identificar regresores colineales en un modelo lineal.

Palabras clave: Variables, colinealidad, valores atípicos.

Multicollinearity assessment in multiple linear regression models with presence of outliers

ABSTRACT

To asses the multicollinearity in multiple linear regression models with the presence of outlier, a simulation study was used with a linear model with three regressors (X_1 , X_2 and X_3) and a response (Y) to four samples sizes (10, 20, 30, and 50), where the X_1 and X_2 variables were generated based on three continuous theoretical distributions (uniform, normal, and exponential) and X_3 was established as a linear combination ($X_3 = X_1 + 2 * X_2 + e$), using a contaminated model with a given percentage of outliers. To assess multicollinearity, the collinearity index (K) was used. In that sense, it showed a potential effect of sample size and the theoretical distribution type of model regressors on the condition index (K), with the subsequent effect on the multicollinearity degree, demonstrating the outlier presence in the sample does not affect the model structure, but rather to parameters. Finally, there were suggested further study of the mean square error estimator of the ordinary least squares in the multiple linear regression model in presence of collinear regressors, since it was observed as the mean square error approaches zero for collinear variable estimator, which might suggest the possible use of the mean square error of ordinary least squares estimator as an alternative to identify collinear regressors in a linear model.

Key words: Variables, collinearity, outlier

---

*Autor de correspondencia: Danny Villegas

E-mail: danny_villegas1@yahoo.com

Recibido: 07 octubre, 2013 Aceptado: 18 noviembre, 2013

INTRODUCCIÓN

Los modelos lineales se usan frecuentemente en el análisis de datos en diversas áreas del conocimiento; por ejemplo, en el área agronómica, en la que por muchos años el estudio de las relaciones entre los cultivos y los factores ambientales ha estado dominado por el empirismo (Grenón, 1992). En ese sentido, la predicción del comportamiento de variables como rendimiento, frente a distintas condiciones de suelo, clima, manejo o variedad está también sujeta a interacciones complejas, que desde el punto de vista estadístico presumen la existencia de multicolinealidad y a la presencia de observaciones que tengan una gran influencia sobre los resultados del ajuste de un modelo lineal, denominados valores atípicos. El efecto de estos valores atípicos puede observarse en el estimador de mínimos cuadrados ordinarios, en el vector de predicciones y en la matriz de dispersión. De esta manera, Mandell (1982) señala que una de las dificultades en el uso de estos estimadores es la presencia de este problema. En ese orden, los efectos adversos de la multicolinealidad han sido ampliamente estudiados. Hoerl y Kennard (1970) demostraron que ésta puede afectar el cuadrado de la distancia entre el estimador de mínimos cuadrados y el parámetro estimado. Por esta razón, si se quiere utilizar un modelo lineal general para predecir en presencia de este fenómeno, algunas predicciones pueden ser precisas, mientras que otras pueden verse drásticamente influenciadas (Wang, 1996). Por otra parte, Chacín (1998) señala que la presencia de multicolinealidad tiene efecto potencial sobre los estimadores mínimos cuadrados de los coeficientes de regresión, además que la estrecha multicolinealidad resulta en elevadas varianzas y covarianzas de estos estimadores. En ese sentido, existen alternativas para el diagnóstico de multicolinealidad entre las que destacan el VIF y el índice de condición (k). En ese orden, Chatterjee y Price (1991) sugieren que un VIF mayor que 1 idica una desviación de la ortogonalidad de algunas variables regresoras y una tendencia a la multicolinealidad.

Sin embargo, los modelos de simulación agronómica son herramientas que integran información, y que permiten analizar y cuantificar las relaciones existentes entre los factores mencionados y sus efectos como componentes del sistema, permitiendo evaluar diferentes cultivos, o analizar un factor manteniendo los otros constantes; por ejemplo, la variación del rendimiento por efecto del clima sin modificar el manejo, el genotipo y el suelo. Así pues, numerosos modelos han sido desarrollados y cada uno de ellos muestra debilidades y fortalezas, entre ellos los modelos de simulación de la familia CERES, los cuales simulan el comportamiento de cereales. Es por ello necesario evaluarlos y validarlos en relación a las suposiciones que se plantean en el contexto estadístico, así como en los ambientes en donde se utilizarán. En tal sentido, se evaluó la multicolinealidad en modelos lineales con presencia de valores atípicos.

MATERIALES Y MÉTODOS

Se efectuaron 1 000 simulaciones de modelos lineales con tres variables independientes y una respuesta. Es de resaltar que este tipo de modelo, comúnmente es empleado en el área agronómica, donde la respuesta se puede corresponder con el rendimiento de un cultivo y las variables regresoras con factores, tales como fertilización y manejo, entre otros. Así mismo, se consideraron cuatro tamaños de muestra (10, 20, 30 y 50), mediante el modelo siguiente:

$$Y = \beta o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e,$$

donde $Y = \beta o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e,$

De igual forma, las variables independientes o regresores del modelo (X_1 y X_2) fueron generadas partiendo de tres distribuciones teóricas continuas (uniforme, normal, y exponencial) y X_3 fue establecida como una combinación lineal del tipo $X_3 = X_1 + 2*X_2 + e$, donde el componente aleatorio de esta combinación fue generado de una distribución normal N (0;2). En ese sentido, esta combinación lineal persigue presentar un modelo que acuse multicolinealidad, lo que para fines prácticos pudiera simular una situación de interacciones complejas entre los factores que intervienen en la producción agrícola. Así mismo, los errores aleatorios del modelo se generaron de una distribución normal N (0;1), utilizando un modelo contaminado con un porcentaje dado de valores atípicos, en este caso el modelo:

$$f(x) = (1- \in)\phi(x) + \in \phi(x/k)$$

donde $\phi(x)$ es la función de distribución normal estándar y k es igual a 4. Los valores de contaminación $\in$ que se utilizaron fueron 0,05; 0,10; 0,15 y 0,20, lo que equivale a contaminar la distribución N (0;1) con los porcentajes de valores atípicos de 5, 10, 15 y 20%, generados de una distribución N (0;4). De la misma manera, un modelo similar es considerado por Pulido y Torres (1997) al comparar tres métodos de regresión. En tal sentido, la presencia de estos valores atípicos en la muestra es una situación que con frecuencia se presenta en modelos agronómicos, fundamentalmente aquellos que consideran factores climáticos, dada la ocurrencia de eventos extremos asociados a las precipitaciones, específicamente, lluvias extremas. La simulación antes señalada se realizó partiendo de algoritmos diseñados en el entorno del software libre R-3.0.0.0.

Los resultados producto de la simulación permitieron realizar comparaciones de cada uno de los escenarios propuestos con base en los siguientes criterios:

- Índice de condición (K) establecido por Belsley *et al.* (1980), el cual sugiere tres grados de multicolinealidad: leve (K<10), moderada (10<K<30) y severa (K≥30). Este índice está dado por la siguiente expresión, $K = \sqrt{\frac{\lambda max}{\lambda min}}$ el cual mide la sensibilidad de las estimaciones mínimo cuadrático ante pequeños cambios en los datos.
- Cantidades promedios de valores atípicos detectados usando el método de los residuales estandarizados, el cual según Birkes y Dodge (1993), sugieren que un residual estandarizado mayor a 2,5 en valor absoluto puede catalogarse como valor atípico.
- Error cuadrático medio (ECM) de los estimadores de los coeficientes del modelo (EMCO), los cuales son aquellos estimadores de los parámetros del modelo poblacional que minimizan la varianza media insesgada cuando los errores tienen varianzas infinitas. Bajo la suposición de normalidad en los errores, estos estimadores son de máxima verosimilitud.
- Promedios de los coeficientes de determinación del modelo (R^2).

Algoritmos en el entorno de programación del software libre R-3.0.0.0:

Índice de condición (K):

```
n = n_o
po = p
X_1 < -rdist(n, ,)
X_2 < -rdist(n, ,)
e_1 < -rnorm(n, ,)
X_3 < -X_1 + 2*X_2 + e_1
e_3 < -(1-po)*(rnorm(n, ,)) + po*(rnorm(n, ,))
Y < -X_1 + X_2 + X_3 + e_3
g < -lm(Y~X_1 + X_2 + X_3)
X < -as.matrix(cbind(1, X_1, X_2, X_3)[,-7])
XtX < - solve(t(X) %*% X)
e_2 < -eigen(t(X)%*%X)
e2$val
sqrt(e2$val[1]/e2$val)
```

Error cuadrático medio de los EMCO del modelo de regresión:

```
n = n_o
po = p
X_1 < -rdist(n, ,)
X_2 < -rdist(n, ,)
e_1 < -rnorm(n, ,)
X_3 < -X_1 + 2*X_2 + e_1
e_3 < -(1-po)*(rnorm(n, ,)) + po*(rnorm(n, ,))
Y < -X_1 + X_2 + X_3 + e_3
g < -lm(Y~X_1 + X_2 + X_3)
X < -as.matrix(cbind(1, X_1, X_2, X_3)[,-7])
solve(t(X)%*% X,t(X)%*% Y)
```

Ajuste del modelo:

```
n = n_o
po = p
X_1 < -rdist(n, ,)
X_2 < -rdist(n, ,)
e_1 < -rnorm(n, ,)
X_3 < -X_1 + 2*X_2 + e_1
e_3 < -(1-po)*(rnorm(n, ,)) + po*(rnorm(n, ,))
Y < -X_1 + X_2 + X_3 + e_3
g < -lm(Y~X_1 + X_2 + X_3)
X < -as.matrix(cbind(1, X_1, X_2, X_3)[,-7])
1-sum(g$res*g$res)/sum((Y-mean (Y))*(Y-mean (Y)))
```

RESULTADOS Y DISCUSIÓN

En el Cuadro 1 se muestran los resultados del diagnóstico de multicolinealidad en un modelo de regresión lineal múltiple contaminado con proporciones de valores atípicos; es decir, cantidades de valores atípicos incorporados a la muestra de datos, considerando una distribución uniforme de los regresores (X_1 y X_2). Se observa que el índice de condición K decrece conforme incrementa el tamaño de la muestra. Para muestras mayores o iguales a 30, este índice se ubicó en el rango que sugiere un grado de multicolinealidad leve

Cuadro 1. Diagnóstico de multicolinealidad en un modelo de regresión lineal múltiple contaminado con proporciones de valores atípicos considerando una distribución uniforme de los regresores (X_1 y X_2).

n	Valor atípico	Aciertos multicolinealidad			Índice de condición $\overline{K}$
		Leve (K ≤ 10)	Moderada (10<K<30)	Severa (K≥30)	
	%	---------- % ----------			
10	0	0	35,60	64,40	35,27
	5	0	33,70	66,30	35,92
	10	0	35,50	64,50	35,41
	15	0	36,00	64,00	35,39
	20	0	36,30	63,70	35,47
20	0	0	53,00	47,00	30,59
	5	0	49,10	50,90	30,94
	10	0	49,40	50,60	30,84
	15	0	50,70	49,30	30,71
	20	0	51,10	48,90	30,64
30	0	0	60,80	39,20	29,39
	5	0	58,60	41,40	29,54
	10	0	59,20	40,80	29,39
	15	0	57,70	42,30	29,45
	20	0	58,50	41,50	29,54
50	0	0	66,20	33,80	28,91
	5	0	67,60	32,40	28,80
	10	0	67,30	32,70	28,80
	15	0	68,30	31,70	28,50
	20	0	67,70	32,30	28,70

(10<K<30), independientemente de la proporción de valores atípicos en la muestra, mientras que para muestras menores a 30, el mismo sugiere un grado de multicolinealidad severa (K≥30), lo que evidencia la consistencia del mismo. De igual forma, estos resultados muestran como el porcentaje de casos donde se detecta multicolinealidad severa disminuye conforme aumenta el tamaño de muestra, incrementándose a su vez el porcentaje de aciertos de multicolinealidad moderada. No obstante, estos resultados no sugieren un potencial efecto de la presencia de valores atípicos sobre el fenómeno de la multicolinealidad, toda vez que a medida que se incrementa la proporción de valores atípicos en el modelo, el índice de condición K se mantiene estable en cada uno de los tamaños de muestra considerados.

En el Cuadro 2 se muestran los resultados del diagnóstico de multicolinealidad en un modelo de regresión lineal múltiple contaminado con proporciones de valores atípicos, considerando una distribución normal de los regresores (X_1 y X_2). Se observa que el índice de condición K reporta valores elevados (143,08 a 186,42) para cada uno de los tamaños de muestra considerados, independientemente de la proporción de valores atípicos presentes en la muestra, lo que indica un grado severo de multicolinealidad. No obstante, pese a que este índice decrece conforme incrementa el tamaño de la muestra, estos resultados no sugieren un cambio significativo en el grado de multicolinealidad, en comparación a los obtenidos en el caso anterior, cuando se utilizó una distribución uniforme para los regresores

Cuadro 2. Diagnóstico de multicolinealidad en un modelo de regresión lineal múltiple contaminado con proporciones de valores atípicos considerando una distribución normal de los regresores (X_1 y X_2).

n	Valor atípico	Aciertos multicolinealidad			Índice de condición $\overline{K}$
		Leve ($K \leq 10$)	Moderada ($10<K<30$)	Severa ($K \geq 30$)	
	%	---------- % ----------			
10	0	0	0	100	183,11
	5	0	0	100	181,87
	10	0	0	100	186,42
	15	0	0	100	181,40
	20	0	0	100	184,53
20	0	0	0	100	155,66
	5	0	0	100	155,23
	10	0	0	100	154,72
	15	0	0	100	156,39
	20	0	0	100	154,87
30	0	0	0	100	149,15
	5	0	0	100	148,16
	10	0	0	100	148,94
	15	0	0	100	148,64
	20	0	0	100	148,50
50	0	0	0	100	143,60
	5	0	0	100	143,50
	10	0	0	100	144,02
	15	0	0	100	143,08
	20	0	0	100	143,42

(X_1 y X_2). De igual forma, estos resultados evidencian como el porcentaje de aciertos de multicolinealidad severa se mantiene estable conforme aumentan el tamaño de muestra y la proporción de valores atípicos, mostrando un comportamiento distinto al ocurrido cuando se utilizó la distribución uniforme.

En el Cuadro 3 se muestran los resultados del diagnóstico de multicolinealidad en un modelo de regresión lineal múltiple contaminado con proporciones de valores atípicos, considerando una distribución exponencial de los regresores (X_1 y X_2). Se observa que el índice de condición K reporta valores relativamente bajos (10,57 a 15,49) para cada uno de los tamaños de muestra considerado, independientemente de la proporción de valores atípicos presentes en la muestra, lo que indica en líneas generales un grado moderado de multicolinealidad. No obstante, pese a que el promedio de este índice decrece conforme se incrementa el tamaño de la muestra, estos resultados no sugieren un cambio significativo en el grado de multicolinealidad, en comparación a los obtenidos cuando se utilizó una distribución uniforme para los regresores (X_1 y X_2). Sin embargo, estos resultados evidencian como el porcentaje de aciertos de multicolinealidad moderada decrece conforme aumentan el tamaño de muestra, en favor de un aumento en el porcentaje de aciertos de multicolinealidad leve, independientemente de la proporción de valores atípicos, sugiriendo un comportamiento distinto al ocurrido cuando se utilizó la distribución uniforme y la normal. Más aún, para muestras menores a 30 se detecta en menores proporciones casos de multicolinealidad

Cuadro 3. Diagnóstico de multicolinealidad en un modelo de regresión lineal múltiple contaminado con proporciones de valores atípicos considerando una distribución exponencial de los regresores (X_1 y X_2).

n	Valor atípico	Aciertos multicolinealidad			Índice de condición $\overline{K}$
		Leve ($K \leq 10$)	Moderada ($10<K<30$)	Severa ($K \geq 30$)	
	%	---------- % ----------			
10	0	13,40	83,70	2,90	15,49
	5	16,00	81,50	2,50	15,01
	10	16,70	81,00	2,30	14,93
	15	14,60	82,70	2,70	15,32
	20	14,10	83,80	2,10	15,11
20	0	25,70	74,20	0,10	11,97
	5	24,70	75,20	0,10	12,12
	10	24,50	75,50	0,00	12,22
	15	26,60	73,40	0,00	12,14
	20	25,10	74,80	0,10	12,07
30	0	30,10	69,90	0	11,27
	5	32,50	67,50	0	11,28
	10	31,20	68,80	0	11,22
	15	30,10	69,90	0	11,23
	20	35,50	64,50	0	11,13
50	0	39,20	60,80	0	10,57
	5	33,40	66,60	0	10,75
	10	38,30	61,70	0	10,64
	15	37,20	62,80	0	10,65
	20	35,40	64,60	0	10,73

severa (0,1 a 2,9%), mientras que para muestras mayores o iguales a 30 solo se detectan casos de multicolinealidad leve (30,1 a 39,2%) y moderada (60,8 a 69,9%). No obstante, al igual que en los casos anteriores, cuando se utilizó la distribución uniforme y la normal para los regresores (X_1 y X_2), estos resultados no sugieren un potencial efecto de la presencia de valores atípicos sobre el fenómeno de la multicolinealidad, toda vez que a medida que se incrementa la proporción de valores atípicos en el modelo, el índice de condición K se mantiene estable en cada uno de los tamaños de muestra considerados.

Cuadro 4. Efecto de multicolinealidad en un modelo de regresión lineal múltiple contaminado con proporciones de valores atípicos considerando una distribución uniforme de los regresores (X_1 y X_2).

n	Valor atípico	Error cuadrático medio de estimadores MCO				Ajuste R^2
		b_o	b_1	b_2	b_3	
	%					
10	0	1,35	0,58	0,62	0,12	0,9450
	5	1,21	0,57	0,56	0,11	0,9842
	10	1,28	0,58	0,60	0,11	0,9836
	15	1,40	0,59	0,62	0,12	0,9814
	20	1,44	0,65	0,66	0,12	0,9764
20	0	0,78	0,35	0,35	0,06	0,9349
	5	0,78	0,35	0,35	0,06	0,9385
	10	0,72	0,33	0,35	0,06	0,9382
	15	0,79	0,36	0,37	0,07	0,9310
	20	0,87	0,39	0,43	0,08	0,9194
30	0	0,59	0,28	0,27	0,05	0,9335
	5	0,59	0,26	0,27	0,05	0,9357
	10	0,60	0,27	0,27	0,04	0,9354
	15	0,65	0,28	0,29	0,05	0,9284
	20	0,66	0,30	0,31	0,06	0,9159
50	0	0,45	0,20	0,22	0,04	0,9318
	5	0,45	0,21	0,20	0,03	0,9360
	10	0,46	0,19	0,21	0,04	0,9344
	15	0,47	0,21	0,22	0,04	0,9277
	20	0,53	0,23	0,24	0,04	0,9135

En el Cuadro 4 se muestran algunos criterios de evaluación de la multicolinealidad en un modelo de regresión lineal múltiple contaminado con proporciones de valores atípicos, considerando una distribución uniforme de los regresores (X_1 y X_2). Se observa que el error cuadrático medio de los estimadores del modelo disminuye conforme se incrementa el tamaño de muestra, lo que evidencia la consistencia del estimador de mínimos cuadrados ordinarios del modelo de regresión lineal múltiple. En ese sentido, estos resultados coinciden con lo expuesto por Montgomery y Peck (1992), quienes demostraron que el error cuadrático medio es inversamente proporcional al tamaño de la muestra y al número de regresores del modelo. No obstante, en

Cuadro 5. Efecto de multicolinealidad en un modelo de regresión lineal múltiple contaminado con proporciones de valores atípicos considerando una distribución normal de los regresores (X_1 y X_2).

n	Valor atípico	Error cuadrático medio de estimadores MCO				Ajuste R^2
		b_o	b_1	b_2	b_3	
	%					
10	0	1,88	0,37	0,28	0,11	0,9823
	5	1,82	0,35	0,26	0,10	0,9842
	10	1,83	0,35	0,28	0,11	0,9836
	15	1,94	0,36	0,28	0,11	0,9814
	20	2,12	0,40	0,31	0,12	0,9764
20	0	1,10	0,21	0,16	0,06	0,9811
	5	1,04	0,21	0,16	0,06	0,9825
	10	1,08	0,21	0,16	0,06	0,9812
	15	1,18	0,22	016	0,06	0,9796
	20	1,24	0,24	0,18	0,07	0,9761
30	0	0,83	0,16	0,13	0,05	0,9810
	5	0,85	0,16	0,13	0,05	0,9820
	10	0,83	0,16	0,12	0,04	0,9809
	15	0,84	0,17	0,13	0,05	0,9789
	20	0,96	0,18	0,15	0,05	0,9749
50	0	0,67	0,12	0,10	0,03	0,9804
	5	0,64	0,12	0,09	0,03	0,9816
	10	0,64	0,12	0,09	0,03	0,9807
	15	0,66	0,12	0,09	0,04	0,9789
	20	0,72	0,13	0,11	0,04	0,9749

cada tamaño de muestra considerado se observa como el error cuadrático medio del estimador aumenta en la medida en que se incrementa la proporción de valores atípicos, lo que evidencia el efecto de los valores atípicos sobre el estimador de mínimos cuadrados ordinarios del modelo de regresión lineal múltiple. En ese orden, esto verifica lo señalado por Montgomery y Peck (1992), quienes mostraron que el error cuadrático medio es directamente proporcional a la la suma de cuadrados del error, la cual se incrementa con la presencia de valores atípicos, siempre y cuando estos sean valores extremadamente altos. De la misma manera, se observa

Cuadro 6. Efecto de multicolinealidad en un modelo de regresión lineal múltiple contaminado con proporciones de valores atípicos considerando una distribución exponencial de los regresores (X_1 y X_2).

n	Valor atípico	Error cuadrático medio de estimadores MCO				Ajuste R^2
		b_o	b_1	b_2	b_3	
	%					
10	0	0,55	0,41	0,82	0,11	0,9482
	5	0,55	0,40	0,81	0,10	0,9501
	10	0,56	0,40	0,81	0,11	0,9483
	15	0,59	0,45	0,86	0,11	0,9451
	20	0,63	0,47	0,91	0,13	0,9351
20	0	0,36	0,23	0,47	0,06	0,9400
	5	0,34	0,22	0,46	0,06	0,9440
	10	0,36	0,23	0,46	0,06	0,9431
	15	0,35	0,23	0,47	0,06	0,9362
	20	0,39	0,25	0,50	0,07	0,9277
30	0	0,27	0,18	0,34	0,05	0,9404
	5	0,26	0,17	0,34	0,05	0,9440
	10	0,26	0,16	0,34	0,05	0,9410
	15	0,27	0,18	0,36	0,05	0,9365
	20	0,31	0,20	0,39	0,05	0,9239
50	0	0,20	0,13	0,25	0,03	0,9397
	5	0,20	0,13	0,25	0,03	0,9427
	10	0,19	0,12	0,25	0,03	0,9419
	15	0,25	0,13	0,26	0,03	0,9343
	20	0,22	0,14	0,29	0,04	0,9246

que el ajuste del modelo (R^2) disminuye en la medida que se incrementa el tamaño de muestra y la proporción de valores atípicos, lo que sugiere un potencial efecto de los valores atípicos sobre la bondad de ajuste del modelo, toda vez que se ha evidenciado en la literatura como los mismos afectan el ajuste del modelo de regresión lineal. En ese orden, los Cuadros 5 y 6 muestran un comportamiento similar para las distribuciones normal y exponencial, respectivamente, de los regresores (X_1 y X_2). Por otro lado, estos tres escenarios considerados (distribución uniforme, normal y exponencial para los regresores) muestran un resultado muy particular con

relación al error cuadrático medio del estimador b_3, el cual mide precisamente el efecto de la variable X_3, la cual está expresada como una combinación lineal de X_1 y X_2. Dicho estimador (b_3) muestra un error cuadrático medio cercano a cero, además de ser el menor en comparación al error cuadrático medio de los otros tres estimadores (b_0, b_1 y b_2). Con los resultados antes mencionados, es importante destacar lo señalado por Canavos (1988), quien sugiere que la multicolinealidad no impide tener un buen ajuste ni evita que la respuesta sea, en forma adecuada, predicha dentro del intervalo de las observaciones. Lo que sucede es que ésta afecta en forma severa las estimaciones de los mínimos cuadrados, ya que bajo los efectos de multicolinealidad éstas tienden a ser menos precisas para los efectos individuales de las variables de predicción; es decir, cuando dos o más variables de predicción son colineales los coeficientes de regresión no miden los efectos individuales sobre la respuesta, sino que reflejan un efecto parcial sobre la misma, sujeto a todo lo que ocurra con las demás variables de predicción en la ecuación de regresión.

CONCLUSIONES

Se evidenció un efecto potencial del tamaño de muestra, así como del tipo de distribución teórica de los regresores del modelo sobre el índice de condición K, con el subsecuente efecto sobre el grado de multicolinealidad. En ese orden, se verificó el efecto de los valores atípicos sobre la estabilidad de los estimadores de mínimos cuadrados ordinarios y la bondad de ajuste del modelo de regresión lineal múltiple. No obstante, los resultados obtenidos en esta investigación no sugirieron un efecto potencial de la presencia de los valores atípicos sobre el grado de multicolinealidad presente en el modelo. Este aspecto demuestra que la presencia de valores atípicos en la muestra no afecta la estructura del modelo, sino más bien a los parámetros del mismo. Sin embargo, se recomienda considerar un intervalo consecutivo de valores para la proporción de valores atípicos que verifique estos resultados. Finalmente, estos resultados sugieren un estudio más exhaustivo del error cuadrático medio del estimador de mínimos cuadrados ordinarios del modelo de regresión lineal múltiple en presencia de regresores colineales, toda vez que se observó como el error cuadrático medio tiende a cero para el estimador de la variable colineal, lo que pudiera sugerir el potencial uso del error cuadrático medio del estimador de mínimos cuadrados ordinarios como una alternativa para identificar regresores colineales en un modelo lineal. De igual forma, se recomienda considerar modelos agronómicos, específicamente los de la familia CERES para ser evaluados a la luz de la multicolinealidad y la presencia de valores atípicos.

REFERENCIAS BIBLIOGRÁFICAS

Belsley, D.; E. Kuth; R. Welsh. 1980. Regression diagnostics. Identifying influential data and sources of collinearity. New York, John Wiley & Sons. New York, EUA. 1183 p.

Birkes, D.; Dodge, Y. 1993. Alternative methods of regression. John Wiley & Sons. New York, EUA. 228 p.

Canavos, G. 1988. Probabilidad y Estadística. Aplicaciones y Métodos. Mc Graw Hill. Ciudad de México, México. 520 p.

Chacín, F. 1998. Análisis de Regresión y Superficie de Respuesta. Facultad de Agronomía, Universidad Central de Venezuela. Maracay, Venezuela. 274 p.

Chatterjee, S.; B. Price. 1991. Regression Analysis by Example. 2da ed. Wiley. New York, EUA. 278 p.

Grenón, D. 1992. Utilidad de los modelos de simulación en la formación en ingeniería agronómica. Facultad de Agronomía y Veterinaria. Universidad Nacional del Litoral. Santa Fe, Argentina. 22 p.

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

Mandell, J. 1982. Use of the singular value decomposition in regression analysis. Am. Statis. 36: 15-24.

Montgomery, D.; E. Peck. 1992. Introduction to Linear Regression Analysis. 2da ed. John Wiley & Sons. New York, EUA. 129 p.

Pulido, E.; J. Torres. 1997. Análisis comparativo de tres métodos de regresión: mínimos cuadrados, no paramétrica basada en rangos y mínima desviación absoluta usando métodos de simulación. Tesis de Grado. Universidad Nacional de Colombia. Bogotá, Colombia.

Wang, G. 1996. How to handle multicollinearity in regression modelling. J. Business Forecas. 15: 23-27.