EFECTOS DE ESPECIFICAR UN MODELO INCORRECTO PARA REGRESIÓN LOGÍSTICA, CON DOS VARIABLES INDEPENDIENTES CORRELACIONADAS

EFFECTS OF SPECIFYING AN INCORRECT MODEL FOR LOGISTIC REGRESSION, WITH TWO INDEPENDENT CORRELATED VARIABLES

Rigoberto **Sifuentes-Amaya**, Gustavo **Ramírez-Valverde***

Estadística. Campus Montecillo. Colegio de Postgraduados. 56230. Montecillo, Estado de México. (rsifuentes@colpos.mx), (gramirez@colpos.mx)

Resumen

El análisis de regresión logística se utiliza para estudiar la asociación entre una variable respuesta binaria con un conjunto de variables independientes. Cuando hay correlación alta entre dos variables independientes, se presentan varianzas grandes en los estimadores de los parámetros. Sin embargo, si el modelo lineal está mal especificado las varianzas pueden disminuir al aumentar la correlación entre las variables independientes. En este trabajo se evaluó mediante un estudio de simulación el efecto de la correlación entre las variables independientes en el modelo de regresión logística cuando el modelo tiene una especificación incorrecta. Al omitir una variable relevante el sesgo de $\hat{\beta}$ aumentó y no desapareció al aumentar n. Se detectó un efecto de la correlación en la potencia de la prueba de hipótesis sobre el parámetro estimado y el tamaño de la prueba de hipótesis estuvo cerca del nominal. Al incluir una variable irrelevante no hubo efecto en el sesgo, el error cuadrado medio mostró evidencia de consistencia y la potencia de la prueba de hipótesis disminuyó cuando aumentó la correlación entre las variables independientes.

Palabras clave: correlación, especificación, omisión e inclusión de variables, regresión logística.

Abstract

Analysis of logistic regression is used to study the association between a binary response variable and a set of independent variables. When correlation is high between two independent variables, variances of the parameter estimators are large. However, if the linear model is poorly specified, the variances can decrease when the correlation between the independent variables increases. In this paper, using a simulation study, we evaluated the effect of the correlation between independent variables on the logistic regression model when the model has an incorrect specification. When a relevant variable was omitted, the $\hat{\beta}$ bias increased and did not disappear even when n was augmented. An effect of the correlation was detected in the power of the hypothesis test on the estimated parameter, and the test size of the hypothesis was close to the nominal size. When an irrelevant variable was included, there was no effect on the bias, the mean square error showed evidence of consistency and power of the hypothesis test diminished when the correlation between the independent variables increased.

Key words: correlation, specification, variable omission and inclusion, logistic regression.

Introducción

La regresión logística es un modelo lineal generalizado frecuentemente usado para medir la asociación entre una variable respuesta binaria y una o más variables independientes. El modelo supone que la probabilidad de que Y=1 (π_i) depende de p variables independientes X_1, X_2, ..., X_p. El modelo de regresión logístico está dado por:

Introduction

Logistic regression is a generalized linear model frequently used to measure the association between a binary response variable and one or more independent variables. The model assumes that the probability that Y=1 (π_i) depends on p independent variables X_1, X_2, …, X_p. The logistic regression model is given by:

$$\log it(\pi_i) = \log[\pi_i / (1-\pi_i)] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}$$

* Autor responsable ❖ Author for correspondence.
Recibido: Enero, 2009. Aprobado: Noviembre, 2009.
Publicado como ARTÍCULO en Agrociencia 44: 197-207. 2010.

$$\log it(\pi_i) = \log[\pi_i / (1-\pi_i)] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}$$

El modelo está correctamente especificado cuando las p variables contribuyen a explicar la variabilidad de la variable respuesta, y cualquier otra variable independiente no incluida en el modelo no lo hace; esto es, las p variables son importantes y no hay otra que lo sea.

En regresión lineal Walls y Weeks (1969) advierten que agregar una variable al modelo nunca mejora la precisión de los estimadores de mínimos cuadrados, pero se remueve un posible sesgo y esto ocurre sin importar si la variable es importante o no. Rao (1971) muestra que omitir una variable relevante causa en los estimadores de mínimos cuadrados: 1) sesgo; 2) disminución de sus varianzas; 3) disminución del error cuadrado cuando el valor de su parámetro es menor que la desviación estándar de su estimador. Rosenberg y Levy (1972) determinaron las condiciones con las cuales el estimador de mínimos cuadrados en un modelo incorrectamente especificado por omitir una variable importante, es más eficiente en términos de error cuadrado medio. Según Hocking (1976), omitir una variable importante produce sesgo en los estimadores, excepto cuando las variables omitidas son ortogonales a las incluidas o cuando el coeficiente verdadero de las variables omitidas es cero (las variables omitidas no son relevantes). En relación a la inclusión de variables irrelevantes al modelo de regresión, Rao (1971) muestra que no generan sesgo en los estimadores, pero generan un incremento en sus varianzas y por consiguiente en sus errores cuadrados medios.

Para el modelo de regresión logística, Gail *et al.* (1984) mostraron que omitir variables independientes relevantes produce sesgo en los coeficientes de las variables independientes incluidas, aunque las variables excluidas eran independientes de las incluidas. Neuhaus y Jewell (1993) reportan resultados semejantes cuando hay una correlación de 0.5 entre la variable independiente omitida y la incluida.

En el modelo lineal generalizado Neuhaus (1998) encuentra que omitir una variable independiente relevante puede inducir pérdida en la eficiencia de los estimadores cuando la variable excluida es independiente de la variable incluida. Neuhaus (1998) reporta, en el modelo lineal generalizado, una pérdida de la potencia de las pruebas de hipótesis de las variables

The model is correctly specified when the p variables contribute to explaining the variability of the response variable, and any other independent variable not included in the model does not; that is the p variables are important and there are no others that are.

Walls and Weeks (1969) warn that, in linear regression, adding a variable to a model never improves the precision of the minimum square estimators, but possible bias is removed; this occurs regardless of whether the variable is important or not. Rao (1971) shows that omitting a relevant variable causes, in the minimum square estimators: 1) bias; 2) reduced variances; 3) reduction of the square error when the value of its parameter is less than the standard deviation of its estimator. Rosenberg and Levy (1972) determined the conditions under which the estimator of minimum squares in a model incorrectly specified by omission of an important variable is more efficient in terms of the mean square error. According to Hocking (1976), omitting an important variable produces bias in the estimators, except when the omitted variables are orthogonal to those included, or when the true coefficient of the omitted variables is zero (the omitted variables are not relevant). Regarding the inclusion of irrelevant variables in the regression model, Rao (1971) shows that they do not generate bias in the estimators, but they do generate an increase in their variances and, consequently, in their mean square errors.

For the logistic regression model, Gail *et al.* (1984) showed that omitting relevant independent variables produced bias in the coefficients of the included independent variables, even when the excluded variables were independent of those included. Neuhaus and Jewell (1993) report similar results when there is a correlation of 0.5 between the omitted and the included independent variable.

In the generalized linear model Neuhaus (1998) found that omitting a relevant independent variable can induce loss of estimator efficiency when the excluded variable is independent of the included variable. Neuhaus (1998) reports that in the generalized linear model, power is lost in the hypothesis tests of variables included in the model when relevant variables, independent of the included variables, are omitted. According to Begg and Lagakos (1990), Score's test of logistic models conserves the nominal test size, although relevant random variables are omitted.

incluidas al modelo al omitir variables relevantes e independientes de las variables incluidas. Según Begg y Lagakos (1990), la prueba de Score de los modelos logísticos conserva el tamaño de prueba nominal, aunque se omitan variables aleatorias relevantes.

En regresión lineal un modelo correctamente especificado produce estimadores insesgados de los parámetros. Sin embargo, la correlación entre las variables independientes puede causar un aumento de la varianza de los estimadores de los parámetros (Wittink, 1988; Lehmann *et al*., 1997), aunque Mela y Praveen (2002) mencionan que en ciertas circunstancias la varianza de los estimadores de los parámetros puede disminuir.

Un modelo incorrectamente especificado es el caso donde el sesgo puede ser problemático y el problema puede aumentar si hay correlación alta entre las variables independientes (Mela y Praveen, 2002). Además, Clarke (2009) menciona que ante el riesgo de sesgo por omitir variables se podría suponer una mejora al incluir un número grande de variables de control pero en ambos casos, regresión lineal o modelo lineal generalizado, el aumento de variables de control puede aumentar o disminuir el sesgo y es difícil saber cual es el caso en una situación particular.

En el presente trabajo mediante un estudio de simulación, se analizó el efecto de una incorrecta especificación del modelo de regresión logística en presencia de diferentes grados de correlación en un modelo con dos variables independientes.

Materiales y Métodos

Estudio de simulación

El estudio de simulación se realizó con dos escenarios: uno para mostrar el desempeño del estimador de máxima verosimilitud y su prueba de hipótesis al omitir una variable relevante al modelo; otro para mostrar el desempeño del estimador de máxima verosimilitud y su prueba de hipótesis al incluir una variable irrelevante.

Primer escenario: Omisión de una variable relevante

El modelo generador de los valores de la variable respuesta Y_i fue:

$$\log\left[\pi_i / (1-\pi_i)\right] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \qquad (1)$$

In linear regression, a correctly specified model produces unbiased estimators of the parameters. The correlation between independent variables, however, can cause an increase in the variance of the parameter estimators (Wittink, 1988; Lehmann *et al*., 1997), although Mela and Praveen (2002) mention that under certain circumstances the variance in the parameter estimators can decrease.

An incorrectly specified model is the case in which bias can be problematic and the problem can grow in the presence of high correlation between the independent variables (Mela and Praveen, 2002). Moreover, Clarke (2009) mentions that because there is a risk of bias when omitting variables, it might be assumed that including a large number of control variables can improve the situation, but that in both cases, linear regression or generalized linear model, adding more control variables could increase or decrease the bias, and it is difficult to know which is the case in a given situation.

In this work, through a simulation study, the effect of incorrect specification of a logistic regression model was analyzed in the presence of different degrees of correlation in a model with two independent variables.

Materials and Methods

Simulation study

The simulation study was conducted with two different scenarios: one to show the performance of the estimator of maximum likelihood and its hypothesis test when a relevant variable is omitted from the model, and the other to show the performance of the estimator of maximum likelihood and its hypothesis test when an irrelevant variable is included.

First scenario: Omission of a relevant variable

The model generator of the values of the response variable Y_i was:

$$\log\left[\pi_i / (1-\pi_i)\right] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \qquad (1)$$

There are two relevant independent variables X_1 and X_2. The properties in the inference associated to the variable X_1 when the relevant variable X_2 is omitted were studied in three situations:

Hay dos variables independientes relevantes X_1 y X_2. Las propiedades en la inferencia asociada a la variable X_1 al omitir la variable relevante X_2; se estudiaron en tres situaciones:

1) El modelo fue correctamente especificado, la estimación y la prueba de hipótesis se hizo usando correctamente el modelo (1).
2) El modelo fue incorrectamente especificado; la estimación y la prueba de hipótesis se hizo usando incorrectamente el modelo:

$$\log\left[\pi_i/(1-\pi_i)\right]=\beta_0+\beta_1X_{1i} \qquad (2)$$

Esto es, se omitió la variable X_2 que es una variable importante para predecir Y_i.

3) Para estudiar el tamaño de la prueba de hipótesis H_0:β_1=0 *vs* H_s:$\beta_1\neq0$ en un modelo incorrectamente especificado donde la variable importante X_2 fue omitida y el valor de β_1=0 (Ho es cierta). El modelo generador de los datos fue $\log\left[\pi_i/(1-\pi_i)\right]=\beta_0+\beta_2X_{2i}$ y se probó H_0:β_1=0 en el modelo $\log\left[\pi_i/(1-\pi_i)\right]=\beta_0+\beta_1X_{1i}$.

Segundo escenario:
Inclusión de una variable irrelevante

El modelo generador de los valores de la variable respuesta Y_i fue:

$$\log\left[\pi_i/(1-\pi_i)\right]=\beta_0+\beta_1X_{1i}$$

Sólo está la variable independiente relevante X_1. Las propiedades en la inferencia asociada a la variable X_1 al incluir la variable irrelevante X_2 se estudiaron en dos situaciones:

1) El modelo fue correctamente especificado, la estimación y la prueba de hipótesis se hizo usando correctamente el modelo (2):

$$\log\left[\pi_i/(1-\pi_i)\right]=\beta_0+\beta_1X_{1i}$$

2) El modelo fue incorrectamente especificado; la estimación y la prueba de hipótesis se hizo usando incorrectamente el modelo (1):

$$\log\left[\pi_i/(1-\pi_i)\right]=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}$$

Esto es, se incluyó la variable X_2, que es irrelevante para predecir Y_i. En el Cuadro 1 se muestra un resumen de los dos escenarios estudiados.

1) The model was correctly specified; the estimation and the hypothesis test were done using the correctly model (1).
2) The model was specified incorrectly; the estimation and the hypothesis test were done using incorrectly the model:

$$\log\left[\pi_i/(1-\pi_i)\right]=\beta_0+\beta_1X_{1i} \qquad (2)$$

That is, the variable X_2, an important variable in predicting Y_i, was omitted.

3) To study the test size of hypothesis H_0: β_1=0 *vs* H_s: $\beta_1\neq0$ in an incorrectly specified model where the important variable X_2 was omitted and the value of β_1=0 (H_0 is true), the data generator model was $\log\left[\pi_i/(1-\pi_i)\right]=\beta_0+\beta_2X_{2i}$, and H_0: β_1=0 was tested in the model $\log\left[\pi_i/(1-\pi_i)\right]=\beta_0+\beta_1X_{1i}$.

Second scenario:
Inclusion of an irrelevant variable

The model generator of the values of the response variable Y_i was:

$$\log\left[\pi_i/(1-\pi_i)\right]=\beta_0+\beta_1X_{1i}$$

We have only the relevant independent variable X_1. The properties of the inference associated to the variable X_1 when the irrelevant variable X_2 was included, were studied in two situations:

1) The model was correctly specified; the estimation and the hypothesis test were done using correctly the model (2) :

$$\log\left[\pi_i/(1-\pi_i)\right]=\beta_0+\beta_1X_{1i}$$

2) The model was incorrectly specified; the estimation and the hypothesis test were done using incorrectly the model (1):

$$\log\left[\pi_i/(1-\pi_i)\right]=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}$$

That is, the variable X_2, which is irrelevant in the prediction of Y_i, was included. A summary of the two scenarios studied is shown in Table 1.

The variable X_1 was simulated as a uniform variable with a mean of zero and a variance of one. The second independent variable X_2 was generated using an auxiliary variable X_3 and the variable X_1. The auxiliary variable X_3 was generated independently of X_1 as a uniform variable with a mean of zero and variance of one. Finally, X_2 was obtained with the equation $X_2=\alpha_0+\alpha_1X_1+\alpha_2X_2$, where α_0=1, α_0=.5, and for the value α_2, we began with α_2=30, which was reduced by 0.1 until obtaining a value of α_2

La variable X_1 se simuló como una variable uniforme con media cero y varianza uno; la segunda variable independiente X_2 se generó usando una variable auxiliar X_3 y la variable X_1; la variable auxiliar X_3 se generó independiente de X_1 como una variable uniforme con media cero y varianza uno; finalmente, se obtuvo X_2 mediante la ecuación $X_2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$, donde α_0=1, α_0=.5 y para el valor de α_2 se inició con α_2=30 y se disminuyó en 0.1 hasta obtener un valor de α_2 que obtenga las correlaciones deseadas entre los valores X_1 y X_2.

Se estudiaron tres correlaciones en cada uno de los escenarios generados: a) la correlación nula (≈.05); b) correlación moderada (≈.5); c) correlación severa (≈.99). Una vez logradas las tres distintas correlaciones, las variables X_1 y X_2 se mantuvieron fijas para todas las situaciones estudiadas. Este proceso se repitió para cada tamaño de muestra estudiado.

Los tamaños de muestra usados en cada situación simulada fueron 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500 y 1000 observaciones.

La generación de la variable respuesta Y_i se realizó utilizando el método de la transformada inversa (Ross, 1999), con los pasos siguientes:

1) Primero se generaron los valores de π_i (i=1, 2,…, n) para cada una de las n observaciones de la muestra. Para el primer escenario estos valores se obtuvieron con la ecuación:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})} \qquad (i=1, 2, ..., n)$$

Para el segundo escenario estos valores se obtuvieron con la ecuación:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{1i})}{1 + \exp(\beta_0 + \beta_1 X_{1i})} \qquad (i=1, 2, ..., n)$$

El vector de parámetros β se seleccionó como el vector propio asociado al valor propio mayor de $X^T X$, que es el mejor de los escenarios teóricos para estimar β (Lee y Silvapulle, 1988; Duffy y Santner, 1989).

2) Se generó una variable auxiliar Z_i (i=1,2,…,n)donde Z_i tiene distribución uniforme en (0,1).
3) Se generó el valor de Y_i (i=1,2,…,n) con la siguiente regla de decisión:

$$Y_i \begin{cases} 1 & Si\ \pi_i > Z_i \\ 0 & Si\ \pi_i \leq Z_i \end{cases}$$

Los aspectos de la inferencia en el modelo de regresión logística evaluadas en la simulación siempre fueron sobre β_1 y consideraron:

Cuadro 1. Resumen de los escenarios utilizados en el estudio de simulación.

Table 1. Summary of the scenarios used in the simulation study.

Modelo generador de los datos	Modelo analizado	Situación analizada
Primer escenario $\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}$	$\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}$	Modelo correcto
	$\beta_0+\beta_1 X_{1i}$	Modelo incorrectamente especificado (omite una variable importante)
Segundo escenario $\beta_0+\beta_1 X_{1i}$	$\beta_0+\beta_1 X_{1i}$	Modelo correcto
	$\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}$	Modelo incorrectamente especificado (incluye una variable irrelevante)

that resulted in the desired correlations between the values of X_1 and X_2.

Three correlations were studied in each of the generated scenarios: a) null correlation (≈.05); b) moderate correlation (≈.5); c) severe correlation (≈.99). Once the three different correlations were achieved, the variables X_1 and X_2 were maintained fixed for all of the situations studied. This process was repeated for each sample size studied.

The sample sizes used in each simulated situation were 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, and 1000 observations.

The response variable Y_i was generated using the inverse transform method (Ross, 1999), proceeding with the following steps:

1) First, values of π_i (i=1,2, … , n) were generated for each of the *n* observation of the sample. For the first scenario, these values were obtained with the equation:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})} \qquad (i=1, 2, ..., n)$$

For the second scenario, these values were obtained with the equation:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{1i})}{1 + \exp(\beta_0 + \beta_1 X_{1i})} \qquad (i=1, 2, ..., n)$$

The vector of parameters β was selected as the proper vector associated to the highest proper value of $X^T X$, which is the best of the theoretical scenarios for estimating β (Lee and Silvapulle, 1988; Duffy and Santner, 1989).

1) Sesgo del estimador. El sesgo del estimador de β_1 dado por E $\left(\hat{\beta}_1\right) - \beta_1$, se estimó como:

$$\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\beta}_{1i} - \beta_1\right)$$

donde $\hat{\beta}_{1i}$ es el estimador de máxima verosimilitud de β_1 obtenido en la i-ésima simulación.

2) El error cuadrado medio (ECM). El ECM de β_1, dado por $E\left[\left(\hat{\beta}_{1i} - \beta_1\right)^2\right]$, se estimó como:

$$\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\beta}_{1i} - \beta_1\right)^2$$

3) La potencia y tamaño de la prueba. Se estudió la función potencia de la prueba de hipótesis estadística H_0:β_1=0 *vs* H_a:$\beta_1 \neq 0$; la función de potencia está dada por:

$$P(\beta_1) = \begin{cases} \alpha(\beta_1) & \text{Para valores de } \beta_2 \text{ bajo } H_0 \\ 1-\beta(\beta_1) & \text{Para valores de } \beta_2 \text{ bajo } H_a \end{cases}$$

donde $\alpha(\beta_1)$= Pr (Error tipo I) = Pr (Rechazar H_0 cuando H_0 es verdadero), y $\beta(\beta_1)$ = Pr (Error tipo II) = Pr (No rechazar H_0 cuando H_0 es falso)

El valor $\alpha(\beta_1)$ para valores de $\beta_1$bajo H_0 representa el tamaño de la prueba y el valor $1-\beta(\beta_1)$ para valores de β_1 bajo H_a representa la potencia.

El valor de $\alpha(\beta_1)$ se estimó como la proporción de rechazos cuando en la simulación β_1=0.

El valor de $1-\beta(\beta_1)$ se estimó como la proporción de rechazos cuando en la simulación $\beta_1 \neq 0$.

El nivel de significancia en todas las pruebas de hipótesis fue α=0.05.

En los dos escenarios propuestos se realizaron 1000 simulaciones con el paquete R versión 2.7.1.

Resultados y Discusión

Primer escenario

El modelo generador de los datos es:

$$\log\left[\pi_i / (1-\pi_i)\right] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

2) An auxiliary variable Z_i (i=1, 2, …, n) was generated, where Z_i has uniform distribution at (0,1)

3) The value of Y_i (i=1, 2, …, n) was generated with the following decision rule:

$$Y_i \begin{cases} 1 & Si\ \pi_i > Z_i \\ 0 & Si\ \pi_i \leq Z_i \end{cases}$$

The inference aspects in the logistic regression model evaluated in the simulation were always above β_1 and considered:

1) Estimator bias. The bias of the β_1 estimator given by E $\left(\hat{\beta}_1\right) - \beta_1$ was estimated as:

$$\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\beta}_{1i} - \beta_1\right)$$

where $\hat{\beta}_{1i}$ is the estimator of maximum likelihood of β_1 obtained in the i^{th} simulation.

2) The mean square error (MSE). The MSE of β_1, given by $E\left[\left(\hat{\beta}_{1i} - \beta_1\right)^2\right]$, was estimated as:

$$\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\beta}_{1i} - \beta_1\right)^2$$

3) Power and test size. The power function of the statistical hypothesis test H_0:β_1=0 vs H_a: $\beta_1 \neq 0$ was studied; the power function is given by:

$$P(\beta_1) = \begin{cases} \alpha(\beta_1) & \text{For } \beta_2 \text{ values under } H_0 \\ 1-\beta(\beta_1) & \text{For } \beta_2 \text{ values under } H_a \end{cases}$$

Where $\alpha(\beta_1)$=Pr (type I error) = Pr (Reject H_0 when H_0 is true), and $\beta(\beta_1)$ = Pr (type II error) = Pr (Do not reject H_0 when H_0 is false).

The value $\alpha(\beta_1)$ for values of β_1 under H_0 represents the test size and the value $1-\beta(\beta_1)$ for values of β_1 under H_a represents the power.

The value$\alpha(\beta_1)$ was estimated as the proportion of rejections when β_1=0 in the simulation.

The value $1-\beta(\beta_1)$ was estimated as the proportion of rejections when $\beta_1 \neq 0$ in the simulation.

The level of significance in all of the hypothesis tests was α=0.05.

In the two proposed scenarios 1000 simulations were conducted with the R package software, version 2.7.1.

Sesgo al omitir la variable relevante X_2

En el modelo correctamente especificado (Figura 1A) hay evidencia de que el sesgo tiende a cero al aumentar *n*; alrededor de n=300 el sesgo prácticamente desaparece.

En el modelo incorrectamente especificado (Figura 1B) el sesgo de $\hat{\beta}_1$ al omitir X_2 se mantiene casi constante para cualquier tamaño de muestra, dando evidencias de que el sesgo no desaparece al aumentar n. Estos resultados coinciden con los de Gail *et al*. (1984) y Neuhaus y Jewell (1993).

Hay un efecto de la correlación en el sesgo, en general a mayor correlación mayor sesgo. El efecto es más marcado en el modelo incorrectamente especificado (Figura 1B).

Error cuadrado medio al omitir la variable relevante X_2

En la Figura 2 se observa un efecto fuerte de la correlación. El ECM es marcadamente más alto en los casos con correlación alta que cuando la correlación es nula o moderada.

El modelo incorrectamente especificado (Figura 2B) mostró ECM menor que el modelo correctamente especificado (Figura 2A) en tamaños de muestra

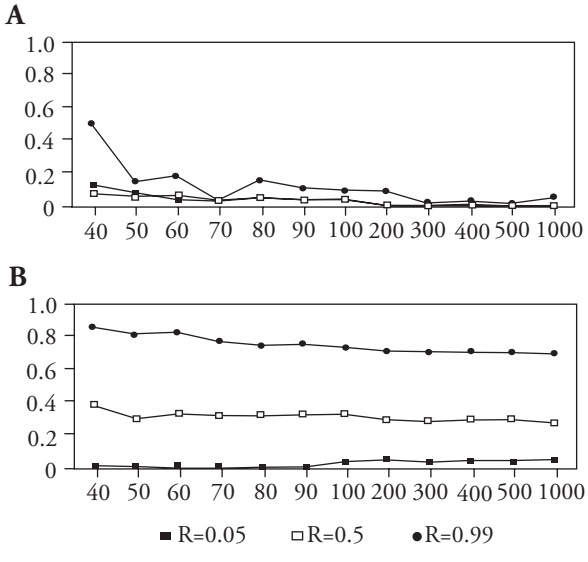

Figura 1. Sesgo de $\hat{\beta}_1$: A) con el modelo correctamente especificado; B) al omitir la variable relevante X_2.

Figure 1. Bias of $\hat{\beta}_1$: A) with the correctly specified model; B) when the relevant variable X_2 is omitted.

Results and Discussion

First scenario

The data generator model is:

$$\log[\pi_i / (1-\pi_i)] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

Bias when relevant variable X_2 is omitted

In the correctly specified model (Figure 1A) there is evidence that the bias tends toward zero when *n* increases; around *n*=300 the bias practically disappears.

In the incorrectly specified model (Figure 1B) the bias of $\hat{\beta}_1$ when X_2 is omitted remains almost constant for any sample size, giving evidence that the bias does not disappear when *n* increases. These results coincide with those of Gail *et al*. (1984) and Neuhaus and Jewell (1993).

There is a correlation effect on the bias; in general, the higher the correlation, the greater the bias. The effect is more marked in the incorrectly specified model (Figure 1B).

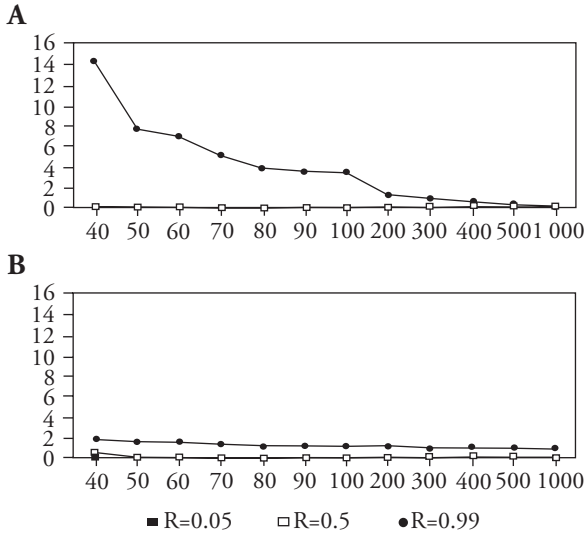

Figura 2. Error cuadrado medio de $\hat{\beta}_1$: A) con el modelo correctamente especificado; B) al omitir la variable relevante X_2.

Figure 2. Mean squared error of $\hat{\beta}_1$: A) with the correctly specified model; B) when the relevant variable X_2 is omitted.

pequeños. Sin embargo, al aumentar el tamaño de muestra disminuye la diferencia en ECM, hasta ocurrir lo contrario alrededor de n=300.

Potencia al omitir la variable relevante X_2

En la Figura 3 se observa que para correlación nula y media el modelo correctamente especificado (Figura 3A) presentó mayores potencias que en el modelo incorrectamente especificado (Figura 3B). Sin embargo, con una correlación alta, la potencia de la prueba tuvo una gran disminución en el modelo correctamente especificado (Figura 3A) pero en el modelo incorrectamente especificado (Figura 3B) la potencia tuvo un gran aumento.

Hay un efecto de correlación en la potencia de la prueba de Wald, en ambos casos. Así, cuando el modelo está correctamente especificado (Figura 3A) la potencia disminuye y el efecto es muy marcado para correlación alta; la potencia es menos de 0.3 cuando n=1000. Para el modelo mal especificado (Figura 3B) que omitió la variable X_2, el efecto de la correlación es positivo; al aumentar la correlación la potencia aumenta drásticamente y en n=50 la potencia es casi 1 en la correlación alta.

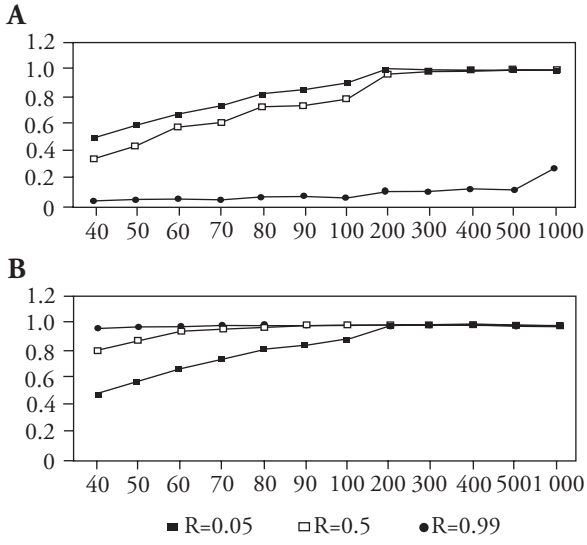

Figura 3. Potencia de la prueba de hipótesis sobre β_1: A) con el modelo correctamente especificado; B) al omitir la variable relevante X_2.

Figure 3. Power of the hypothesis test over β_1: A) with the correctly specified model; B) when the relevant variable X_2 is omitted.

Mean square error when the relevant variable X_2 is omitted

In Figure 2 a strong effect of the correlation can be observed; MSE is markedly higher in the cases in which correlation is high than in those where the correlation is null or moderate.

The incorrectly specified model (Figure 2B) had a lower MSE than the correctly specified model (Figure 2A) in small sample sizes. However, when the sample size increases, the difference in MSE diminishes until the contrary occurs around n=300.

Power when relevant variable X_2 is omitted

In Figure 3 it can be observed that for null and medium correlation, the correctly specified model (Figure 3A) had higher powers than the incorrectly specified model (Figure 3B). However, with a high correlation, the power of the test decreased greatly in the correctly specified model (Figure 3A), while in the incorrectly specified model (Figure 3B) power greatly increased.

Correlation has an effect on the power of the Wald test in both cases. Thus, when the model is correctly specified (Figure 3A), power decreases and the effect is strongly marked for high correlation; power is less than 0.3 when n= 1000. For the poorly specified model (Figure 3B), which omitted the variable X_2, the effect of the correlation is positive; when correlation increases, the power increases drastically and with n=50, power is almost 1 in the high correlation.

Hypothesis test size when the relevant variable X_2 is omitted

The behavior of the size of the test H_0: β_1=0 *vs* $\beta_1 \neq 0$ when a relevant variable is omitted is shown en Figure 4. Note that the values fluctuate around the nominal value of 0.05. The fluctuation can be attributed to the fact that only 1000 simulations were repeated in each simulated situation. The result obtained coincides with that reported by Begg and Lagakos (1990) when the included and the non-included variables are independent.

Tamaño de la prueba de hipótesis al omitir la variable relevante X_2

En la Figura 4 se muestra el comportamiento del tamaño de la prueba H_0:β_1=0 *vs* $\beta_1 \neq 0$, al omitir una variable relevante; se nota que los valores fluctúan cerca del valor nominal de 0.05. Las fluctuaciones se pueden atribuir a que sólo se repitieron 1000 simulaciones en cada situación simulada. El resultado obtenido coincide con el reportado por Begg y Lagakos (1990) cuando la variable incluida y la no incluida son independientes.

Segundo escenario
Sesgo al incluir la variable irrelevante X_2

En la Figura 5 se muestra que para ambos casos: A) sesgo de $\hat{\beta}_1$ con el modelo correctamente especificado y B) sesgo de $\hat{\beta}_1$ al incluir la variable irrelevante X_2, los sesgos tienden a desaparecer al aumentar *n* (consistencia) y son mayores cuando el modelo está incorrectamente especificado por incluir la variable irrelevante X_2.

Error cuadrado medio al incluir la variable irrelevante X_2

En la Figura 6 se observa que con el modelo correctamente especificado (Figura 6A) el error cuadrado medio de $\hat{\beta}_1$ no muestra efecto de la correlación.El ECM de $\hat{\beta}_1$ al incluir la variable irrelevante X_2 (Figura 6b) es mayor sólo en las correlaciones severas ($r_{ij} \approx 0.99$).

Potencia de la prueba de hipótesis al incluir la variable irrelevante X_2

En la Figura 7 se observa un gran efecto de la correlación en la potencia de la prueba sobre $\hat{\beta}_1$ al

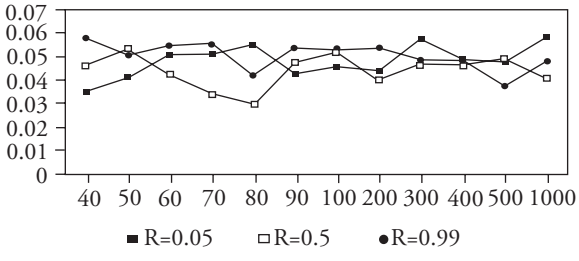

Figura 4. Tamaño de la prueba de hipótesis sobre β_1 al omitir la variable relevante X_2.

Figure 4. Test size of hypothesis over β_1 when the relevant variable X_2 is omitted.

Second scenario
Bias when the irrelevant variable X_2 is included

In Figure 5 it is shown that for both cases: A) $\hat{\beta}_1$ bias with the correctly specified model and B) $\hat{\beta}_1$ bias when the irrelevant variable X_2 is included, the biases tend to disappear when *n* increases (consistency) and are greater when the model in incorrectly specified by including the irrelevant variable X_2.

Mean square error when the irrelevant variable X_2 is included

In Figure 6 it can be observed that with the correctly specified model (Figure 6A) the mean square error of $\hat{\beta}_1$ does not exhibit effect of the correlation. When the irrelevant variable X_2 (Figure 6B) is included, the MSE of $\hat{\beta}_1$ is larger only in severe correlations ($r_{ij} \approx 0.99$).

Power of the hypothesis test when the irrelevant variable X_2 is included

In Figure 7 a large effect of the correlation on the power of the test over $\hat{\beta}_1$ is observed when the irrelevant variable X_2 is included, especially in small sample sizes. This effect of the correlation decreases when *n* increases, mainly in null ($r_{ij} \approx 0.05$) and moderate ($r_{ij} \approx 0.05$) correlations.

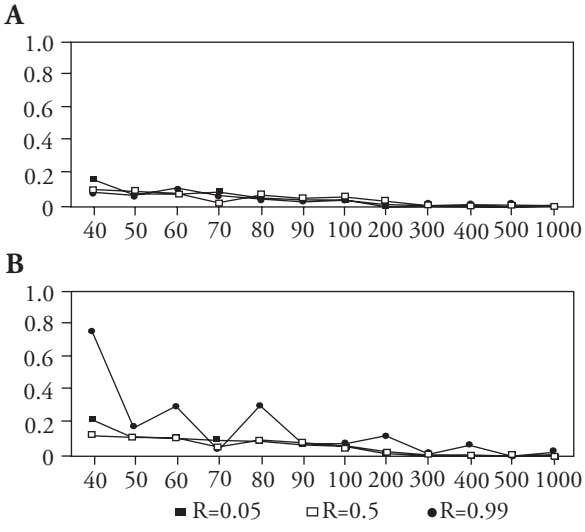

Figura 5. Sesgo de $\hat{\beta}_1$: A) con el modelo correctamente especificado; B) al incluir la variable irrelevante X_2.

Figure 5. Bias of $\hat{\beta}_1$: A) with the correctly specified model; B) when the irrelevant variable X_2 is included.

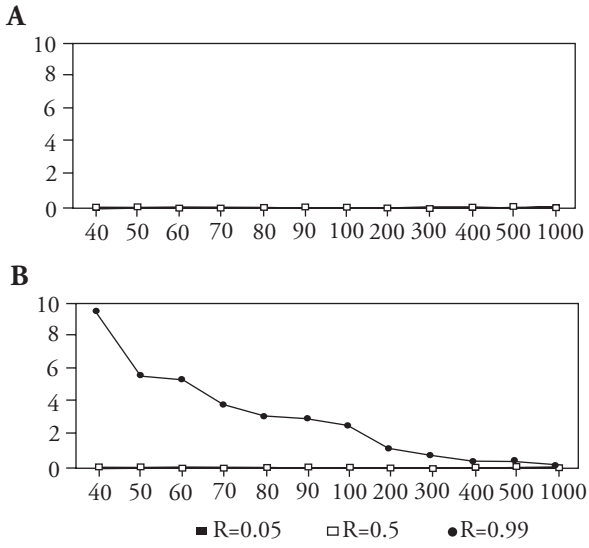

Figura 6. Error cuadrado medio de $\hat{\beta}_1$: A) con el modelo correctamente especificado; B) al incluir la variable irrelevante X_2.

Figure 6. Mean squared error of $\hat{\beta}_1$: A) with the correctly specified model; B) when the irrelevant variable X_2 is included.

incluir la variable irrelevante X_2, principalmente en muestras pequeña. Este efecto de la correlación disminuye al aumentar n en las correlaciones nulas ($r_{ij} \approx 0.05$) y moderadas ($r_{ij} \approx 0.05$) principalmente.

Conclusiones

Primer escenario
Al omitir una variable independiente relevante

El estimador de β_1 en el modelo correctamente especificado es consistente. Cuando el modelo estaba mal especificado el estimador fue sesgado y el sesgo no desapareció al aumentar el tamaño de muestra.

Existe efecto de la correlación en la potencia de la prueba de hipótesis. En el modelo correctamente especificado hay un efecto negativo de la correlación (la potencia disminuye al aumentar la correlación). En el modelo incorrectamente especificado hay un efecto positivo de la correlación (al aumentar la correlación la potencia aumentó).

El valor del tamaño de la prueba en todos los casos estuvo cercano al nominal de α=0.05.

Conclusions

First scenario
A relevant independent variable is omitted

The estimator of β_1 in the correctly specified model was consistent. When the model was poorly specified the estimator was biased and the bias did not disappear when sample size increased.

There was an effect of the correlation on the power of the hypothesis test. In the correctly specified model, a negative effect of the correlation was observed (power decreases when correlation increases). In the incorrectly specified model there was a positive effect of correlation (when correlation increased, power increased).

Value of the test size in all cases was near the nominal of α=0.05.

Second scenario
An irrelevant independent variable is included

Including an irrelevant variable had no effect on bias, except when the correlation was high and larger biases resulted than when the model was correctly specified.

When an irrelevant variable was added, correlation had an effect on the mean square error, increasing as

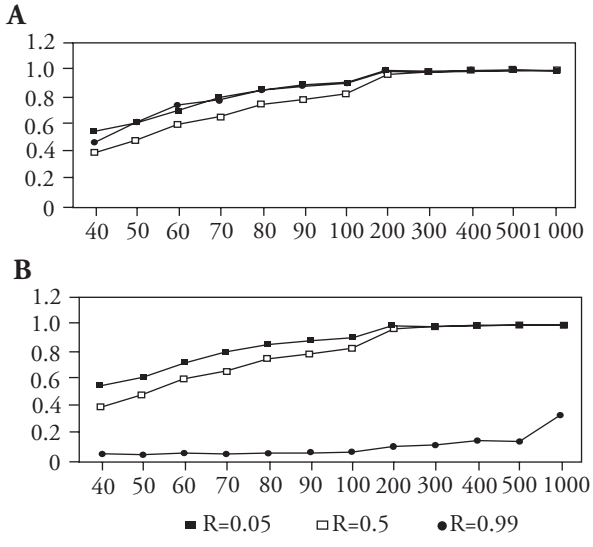

Figura 7. Potencia de $\hat{\beta}_1$: A) con el modelo correctamente especificado; B) al incluir la variable irrelevante X_2.

Figure 7. Power of $\hat{\beta}_1$: A) with the correctly specified model; B) when the irrelevant variable X_2 is included.

Segundo escenario
Se incluye una variable independiente irrelevante

No hay efecto por incluir una variable irrelevante en el sesgo, excepto cuando la correlación es alta donde resultaron sesgos mayores que cuando el modelo estaba correctamente especificado.

Al aumentar una variable irrelevante hay efecto de la correlación en el error cuadrado medio y aumenta conforme incrementa la correlación. El efecto disminuye al aumentar el tamaño de muestra.

Al aumentar una variable irrelevante hay efecto de la correlación en la potencia de la prueba; al aumentar la correlación, disminuye la potencia.

Literatura Citada

Begg, M. D., and S. W. Lagakos. 1990. On the consequences of model misspecification in logistic regression. Environ. Health Perspectives 87: 69-75.

Clarke, K. A. 2009. Return of the phantom menace omitted variable bias in political research. Conflict Manage. Peace Sci. 26(1): 46-66.

Duffy, D. E., and T. J. Santner. 1989. On the small sample properties of norm restricted maximum likelihood estimators for logistic regression models. Comm. Statistics-Theory and Methods 18(3): 959-980.

Gail, M. H., S. Wieand, and S. Piantadosi. 1984. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrics 71: 431-444.

Hocking, R. R. 1976. The analysis and selection of variables in linear regression. Biometrics 32(1):1-49.

the correlation increased. The effect decreased when sample size increased.

When an irrelevant variable was added, correlation had an effect on the test power: when the correlation increased, power decreased.

—End of the English version—

Lee, A. H., and M.J. Silvapulle. 1988. Ridge estimation in logistic regression. Comm. Statistics-Theory and Methods 17(4): 1231-1257.

Lehmann, D. R., S. Gupta, and J. Steckel. 1997. Marketing Research, Addison-Wesley Educational Publishers, Inc. Reading. Massachussetts. 780 p.

Mela, F. C., and K. K. Praveen. 2002. The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations. Appl. Econ. 34: 667-677.

Neuhaus, J. M. 1998. Estimation efficiency with omitted covariates in generalized linear models. J. Am. Stat. Assoc. 93(443): 1124-1129.

Neuhaus, J. M., and N. P. Jewell. 1993. A geometric approach to assess bias due to omitted covariates in generalized linear models. Biometrika 80: 807-815.

Rao, P. 1971. Some notes on misspecification in multiple regressions. The Am. Stat. 25(5):37-39.

Rosenberg, S. H., and P.S. Levy. 1972. Note: a characterization on misspecification in the general linear regression. Biometrics 28(4):1129-1133.

Ross, S. M. 1999. Simulation. Academic Press. Fourth edition. 312 p.

Walls, R. C., and D.L. Weeks. 1969. A note on the variance of a predicted response in regression. The Am. Stat. 23(3): 24-26.

Wittink, D. R. 1988. The Application of Regression Analysis. Simon & Schuster. Needham Heights. Massachusetts. 350 p.