UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2009, Vol. 2, No. 1

Estimadores de regresión logística para tratamiento de no respuesta en el caso de cocientes de variables dicotómicas

Logistic Regression Estimators for the Treatment of Nonresponse for the Ratio of Dichotomic Variables

Pedro César Del Campo Neira[a]
pcampo@icfes.gov.co

Resumen

Para la estimación de un cociente de variables dicotómicas en un diseño MAS^2 se utiliza información auxiliar para los elementos de la segunda etapa. La información auxiliar se usa en el numerador y el denominador a través de regresión logística. Para distintas funciones de enlace en el modelo de regresión y tratamiento de la no respuesta, se compara la eficiencia del uso de distintos tipos de información auxiliar binaria en el numerador y el denominador. Usando el proceso de simulación de Monte Carlo se realiza la comparación de estos distintos escenarios.

***Palabras clave*:** muestreo, estimación de cocientes, modelos lineales generalizados, regresión logística, tratamiento de la no respuesta.

Abstract

To make the estimate of ratio with dichotomic variables in a $SISI$ design is used auxiliary information for the second stage elements. Auxiliary information is used in the numerator and denominator with the use of logistic regression. To different link functions in the logistic regression model and the treatment of the nonresponse, it is compared the efficiency of the use of different types of binary auxiliary information for numerator and denominator. Through of a Monte Carlo Simulation process is made the comparison of the different approaches.

***Key words*:** survey sampling, estimate of ratio, generalized linear model, logistic regression, treatment of nonresponse.

[a]Asesor estadístico. Subdirección académica. Instituto Colombiano para el Fomento de la Educación Superior (ICFES).

1. Introducción

Para muchos estudios de muestreo se tiene interés en estimar razones de totales de variables dicotómicas. Un ejemplo son las encuestas de tipo electoral, donde se desea estimar la razón entre el total de personas que votan por un determinado candidato sobre el total de personas que votan. También son un ejemplo las encuestas de tipo médico, donde se desea estimar la razón entre el total de personas que sufren de una determinada enfermedad y que viven en una región sobre el total de personas que viven en esa región.

En particular se desea formular estimadores de cocientes de variables dicotómicas que se obtienen con el uso de información auxiliar y modelos de regresión logística para un diseño MAS^2 con distintas funciones de enlace y teniendo en cuenta la no respuesta para elementos de la segunda etapa. Se trabaja con estimadores de cocientes sin información auxiliar, con información auxiliar para el denominador, información auxiliar para el numerador e información auxiliar para el denominador y el numerador. La información auxiliar se trabaja a través de regresión logística para las funciones de enlace *Logit* y *Probit*. Posteriormente, se comparará su eficiencia en términos de sesgo y varianza a través de procesos de simulación.

2. Estimación de cocientes de totales asistidos por modelos

Sea U un conjunto considerado como una población con N elementos $\{e_1, e_2, \ldots, e_N\}$. Se tienen dos variables dicotómicas y_k y z_k para cada elemento de U. Sea $U_y := \{e_k \,|\, y_k = 1\}$ y $U_z := \{e_k \,|\, z_k = 1\}$. En particular se tiene que $U_y \subseteq U_z \subseteq U$. El objetivo es estimar el cociente C representado en la ecuación (1).

$$C = \frac{\sum_U y_k}{\sum_U z_k} = \frac{N_y}{N_z} \tag{1}$$

Debido a que $U_y \subseteq U_z$ se puede ver claramente que C representa la proporción de elementos de U_y en U_z. Para el caso de diseño muestral MAS^2 la estimación del cociente C depende de la información disponible para la variables Y y Z. La varianza y el estimador de la varianza del estimador del cociente C se obtienen mediante el método de linealización de Taylor (Särndal et al. 1993)

2.1. Estimación de cocientes en un modelo MAS^2

Para un diseño muestral MAS^2 se tiene en particular que el estimador de un cociente de variables dicotómicas sin el uso de información auxiliar está dado por la ecuación (2).

$$\widehat{C}_\pi = \frac{\sum_{mI} \sum_{mi} \frac{y_k}{\pi_k}}{\sum_{mI} \sum_{mi} \frac{y_k}{\pi_k}} = \frac{\widehat{N}_y}{\widehat{N}_z} \tag{2}$$

Para la estimación de totales bajo un diseño muestral en dos etapas se debe tener en cuenta qué tipo de información auxiliar se tiene. En particular cuando se dispone de información auxiliar sólo para los elementos de la segunda etapa, y en particular de sus totales poblacionales, se plantea la estimación del total de una variable mediante la ecuación (3), según Särndal et al. (1993) muestra en el capítulo 8. Es decir se dispone de $\boldsymbol{t}_x = \sum_U \boldsymbol{x}_k$.

$$\widehat{t}_{yBr} = \sum_m \frac{g_{ksB} y_k}{\pi_k} \tag{3}$$

donde,

$$g_{ksB} = 1 + \left[\sum_{U_I} \boldsymbol{t}_{xi} - \sum_{m_I} \frac{\widehat{\boldsymbol{t}}_{xi\pi}}{\pi_{Ii}} \right]^t \widehat{\boldsymbol{T}}^{-1} \frac{\boldsymbol{x}_k}{\sigma_k^2} \tag{4}$$

Luego con el uso de información auxiliar se determina que el estimador de un cociente es como indica la ecuación (5), donde se dispone de información auxiliar para el numerador y el denominador.

$$\widehat{C}_{xw} = \frac{\sum_m \left(1 + \left[\sum_{U_I} \boldsymbol{t}_{xi} - \sum_{m_I} \frac{\widehat{\boldsymbol{t}}_{xi\pi}}{\pi_{Ii}} \right]^t \widehat{\boldsymbol{T}}_x^{-1} \frac{\boldsymbol{x}_k}{\sigma_k^2} \right) \frac{y_k}{\pi_k}}{\sum_m \left(1 + \left[\sum_{U_I} \boldsymbol{t}_{wi} - \sum_{m_I} \frac{\widehat{\boldsymbol{t}}_{wi\pi}}{\pi_{Ii}} \right]^t \widehat{\boldsymbol{T}}_w^{-1} \frac{\boldsymbol{w}_k}{\sigma_k^2} \right) \frac{y_k}{\pi_k}} \tag{5}$$

Se asume que el vector $\mathbf{x_k}$ es información auxiliar para los elementos y_k y el vector $\mathbf{w_k}$ es información auxiliar para los elementos z_k.

La información auxiliar también puede ser utilizada en solo uno de los términos del cociente, en el numerador o en el denominador, dando como resultados estimadores de los cocientes, como se representa en las ecuaciones (6) y (7).

$$\widehat{C}_x = \frac{\sum_m \left(1 + \left[\sum_{U_I} \boldsymbol{t}_{xi} - \sum_{m_I} \frac{\widehat{\boldsymbol{t}}_{xi\pi}}{\pi_{Ii}} \right]^t \widehat{\boldsymbol{T}}_x^{-1} \frac{\boldsymbol{x}_k}{\sigma_k^2} \right) \frac{y_k}{\pi_k}}{\widehat{N}_z} \tag{6}$$

$$\widehat{C}_w = \frac{\widehat{N}_y}{\sum_m \left(1 + \left[\sum_{U_I} \boldsymbol{t}_{wi} - \sum_{m_I} \frac{\widehat{\boldsymbol{t}}_{wi\pi}}{\pi_{Ii}}\right]^t \widehat{\boldsymbol{T}}_w^{-1} \frac{\boldsymbol{w}_k}{\sigma_k^2}\right) \frac{y_k}{\pi_k}} \tag{7}$$

En particular $\boldsymbol{x}_k^t = (1, x_k)$ y $\boldsymbol{w}_k^t = (1, w_k)$ son los vectores de información auxiliar. Las variables x_k y w_k son variables dicotómicas que poseen un grado de correlación con las variables y_k y z_k respectivamente.

2.2. Modelo Lineal Generalizado (MLG)

Los modelos lineales generalizados extienden el caso de los modelos lineales, incorporando de esta manera la posibilidad de modelar variables respuesta con distribuciones no necesariamente normales según estudios de Nelder y Wedderburn(1972, citado por Dobson, 1990). Para este trabajo en particular las variables provienen de una distribución bernoulli.

Sea $y_1, y_2, \ldots y_n$ valores muestrales de variables aleatorias que provienen de una distribución binomial. Para un modelo lineal generalizado se considera un conjunto de parámetros $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^t$ donde $p < n$ tal que una combinación lineal de estos parámetros sea igual a una función que dependa del valor esperado μ_i de y_i.

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^t \boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij}^t \beta_j \tag{8}$$

Los parámetros del modelo lineal generalizado, representados en la ecuación (8), se estiman a través de un proceso iterativo descrito en la ecuación (9). El proceso iterativo se da porque la matriz $\boldsymbol{W}$ de tamaño nxn y el vector $\boldsymbol{z}$ y el vector n dependen directamente de los parámetros del modelo.

$$\boldsymbol{X}^t \boldsymbol{W}^{(m)} \boldsymbol{X} \boldsymbol{b}^{(m+1)} = \boldsymbol{X}^t \boldsymbol{W}^{(m)} \boldsymbol{z}^{(m)} \tag{9}$$

La matriz $\boldsymbol{W}$ es una matriz diagonal con los elementos dados por la ecuación (10).

$$w_{ii} = \frac{1}{V(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \tag{10}$$

El vector $\boldsymbol{z}$ tiene los elementos representados en la ecuación 11.

$$z_i = \sum_{k=1}^{p} x_{ik} b_k^{(m)} + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i} \tag{11}$$

Para efectos de este trabajo se van a considerar grupos de un solo individuo que provienen de una distribución binomial $Z_i \sim bin(1, \mathfrak{p}_i)$ o distribución Bernoulli.

A continuación se va a presentar la forma de estimación de los parámetros del modelo de regresión lineal generalizado teniendo en cuenta los dos tipos distintos de función de enlace *Logit* y *Probit*.

2.2.1. Estimación de parámetros función de enlace Logit

Teniendo en cuenta que los datos $y_1, y_2, \ldots, y_G$ se distribuyen $Y_g \sim bin(n_g, \mathfrak{p}_g)$ se tiene que $E[y_g] = \mu_g = n_g\mathfrak{p}_\mathfrak{g}$ y $V[y_g] = n_g\mathfrak{p}_g(1-\mathfrak{p}_g)$. Luego se desea determinar los parámetros de la matriz diagonal $\boldsymbol{W}$ y del vector $\boldsymbol{z}$ según las ecuaciones (10) y (11) respectivamente. Como primera medida se determina la derivada $\partial\eta_g/\partial\mu_g$ donde η_g está representado con la función de enlace *Logit*.

$$\frac{\partial\eta_g}{\partial\mu_g} = \frac{\partial\eta_g}{\partial\mathfrak{p}_g}\frac{\partial\mathfrak{p}_g}{\partial\mu_g} = \left[\frac{1}{\mathfrak{p}_g} + \frac{1}{1-\mathfrak{p}_g}\right]\frac{1}{n_g} = \left[\frac{1-\mathfrak{p}_g+\mathfrak{p}_g}{\mathfrak{p}_g(1-\mathfrak{p}_g)}\right]\frac{1}{n_g} = \frac{1}{n_g\mathfrak{p}_g(1-\mathfrak{p}_g)} = \frac{1}{V(y_g)}$$

Luego, utilizando este resultado se obtiene el resultado de los elementos de la matriz diagonal $\boldsymbol{W}$ para la iteración m-ésima.

$$\begin{aligned}
w_{gg}^{(m)} &= \frac{1}{V(y_g)^{(m)}}\left(\frac{\partial\mu_g}{\partial\eta_g}\right)^2_{(m)} \\
&= \frac{(V(y_g))^2}{V(y_g)} \\
&= n_g\mathfrak{p}_g(1-\mathfrak{p}_g) \\
&= n_g\frac{\exp\left\{\boldsymbol{x}_g^t\boldsymbol{b}^{(m)}\right\}}{1+\exp\left\{\boldsymbol{x}_g^t\boldsymbol{b}^{(m)}\right\}}\left(1-\frac{\exp\left\{\boldsymbol{x}_g^t\boldsymbol{b}^{(m)}\right\}}{1+\exp\left\{\boldsymbol{x}_g^t\boldsymbol{b}^{(m)}\right\}}\right) \\
&= n_g\frac{\exp\left\{\boldsymbol{x}_g^t\boldsymbol{b}^{(m)}\right\}}{\left(1+\exp\left\{\boldsymbol{x}_g^t\boldsymbol{b}^{(m)}\right\}\right)^2}
\end{aligned} \tag{12}$$

Luego, utilizando en parte el resultado de la ecuación (12) se obtiene la forma de los elementos del vector $\boldsymbol{z}$ reflejados en la ecuación (13).

$$z_g^{(m)} = \boldsymbol{x}_g^t \boldsymbol{b}^{(m)} + \left(y_g - \mu_g^{(m)}\right)\left(\frac{\partial \eta_g}{\partial \mu_g}\right)_{(m)}$$
$$= \boldsymbol{x}_g^t \boldsymbol{b}^{(m)} + \left(y_g - n_g \frac{\exp\left\{\boldsymbol{x}_g^t \boldsymbol{b}^{(m)}\right\}}{1+\exp\left\{\boldsymbol{x}_g^t \boldsymbol{b}^{(m)}\right\}}\right) \frac{\left(1+\exp\left\{\boldsymbol{x}_g^t \boldsymbol{b}^{(m)}\right\}\right)^2}{n_g \exp\left\{\boldsymbol{x}_g^t \boldsymbol{b}^{(m)}\right\}} \quad (13)$$

Finalmente, se estima $\boldsymbol{\beta}$ utilizando las ecuaciones (12) y (13), y el método iterativo dado en la ecuación (9) partiendo de un valor $\boldsymbol{b}^{(0)}$.

2.2.2. Estimación de parámetros función de enlace Probit

Teniendo en cuenta que los datos $y_1, y_2, \ldots, y_G$ se distribuyen $Y_g \sim bin(n_g, \mathfrak{p}_g)$ se tiene que $E[y_g] = \mu_g = n_g \mathfrak{p}_\mathfrak{g}$ y $V[y_g] = n_g \mathfrak{p}_g (1 - \mathfrak{p}_g)$. Luego se desea determinar los parámetros de la matriz diagonal $\boldsymbol{W}$ y del vector $\boldsymbol{z}$ según las ecuaciones (10) y (11) respectivamente. Como primera medida se determina la derivada $\partial \mu_g / \partial \eta_g$ usando el teorema de diferencias de Lebesgue. η_g está representado con la función de enlace *Probit*.

$$\frac{\partial \mu_g}{\partial \eta_g} = \frac{\partial \mu_g}{\partial \mathfrak{p}_g} \frac{\partial \mathfrak{p}_g}{\partial \eta_g} = n_g \; f_N\left(\eta_g\right) = n_g \; f_N\left(\boldsymbol{x}_g^t \boldsymbol{b}\right)$$

Donde f_N representa la función de distribución normal estándar. Luego, utilizando este resultado se obtienen los elementos de la matriz diagonal $\boldsymbol{W}$ para la iteración m-ésima.

$$w_{gg}^{(m)} = \frac{1}{V(y_g)^{(m)}} \left(\frac{\partial \mu_g}{\partial \eta_g}\right)_{(m)}^2$$
$$= \frac{n_g^2 \; f_N\left(\boldsymbol{x}_g^t \boldsymbol{b}^{(m)}\right)^2}{\Phi\left(\boldsymbol{x}_g^t \boldsymbol{b}^{(m)}\right)\left(1 - \Phi\left(\boldsymbol{x}_g^t \boldsymbol{b}^{(m)}\right)\right)} \quad (14)$$

Luego se obtienen los elementos del vector $\boldsymbol{z}$ reflejados en la ecuación (15).

$$z_g^{(m)} = \boldsymbol{x}_g^t \boldsymbol{b}^{(m)} + \left(y_g - \mu_g^{(m)}\right)\left(\frac{\partial \eta_g}{\partial \mu_g}\right)_{(m)}$$
$$= \boldsymbol{x}_g^t \boldsymbol{b}^{(m)} + \frac{y_g - n_g \Phi\left(\boldsymbol{x}_g^t \boldsymbol{b}^{(m)}\right)}{n_g \; f_N\left(\boldsymbol{x}_g^t \boldsymbol{b}^{(m)}\right)} \quad (15)$$

Finalmente se estima $\boldsymbol{\beta}$ utilizando las ecuaciones (14) y (15), y el método iterativo dado en la ecuación (9) partiendo de un valor $\boldsymbol{b}^{(0)}$.

2.3. Estimación de cocientes bajo modelo logístico

Teniendo en cuenta los desarrollos de la estimación de un total asistido por regresión dado en la ecuación (3) se puede obtener otra manera de escribir esta estimación según la ecuación (16).

$$\begin{aligned}
\widehat{t}_{yBr} &= \sum\nolimits_m \frac{g_{ksB} y_k}{\pi_k} \\
&= \sum\nolimits_m \left(1 + \left[\sum\nolimits_{U_I} \boldsymbol{t}_{xi} - \sum\nolimits_{m_I} \frac{\widehat{\boldsymbol{t}}_{xi\pi}}{\pi_{Ii}} \right]^t \widehat{\boldsymbol{T}}^{-1} \frac{\boldsymbol{x}_k}{\sigma_k^2} \right) \frac{y_k}{\pi_k} \\
&= \sum\nolimits_m \frac{y_k}{\pi_k} + \sum\nolimits_m \left(\left[\sum\nolimits_{U_I} \boldsymbol{t}_{xi} - \sum\nolimits_{m_I} \frac{\widehat{\boldsymbol{t}}_{xi\pi}}{\pi_{Ii}} \right]^t \widehat{\boldsymbol{T}}^{-1} \frac{\boldsymbol{x}_k}{\sigma_k^2} \frac{y_k}{\pi_k} \right) \\
&= \sum\nolimits_m \frac{y_k}{\pi_k} + \left[\sum\nolimits_{U_I} \boldsymbol{t}_{xi} - \sum\nolimits_{m_I} \frac{\widehat{\boldsymbol{t}}_{xi\pi}}{\pi_{Ii}} \right]^t \widehat{\boldsymbol{T}}^{-1} \sum\nolimits_m \frac{\boldsymbol{x}_k}{\sigma_k^2} \frac{y_k}{\pi_k} \\
&= \sum\nolimits_m \frac{y_k}{\pi_k} + \left[\sum\nolimits_{U_I} \sum\nolimits_{U_i} \boldsymbol{x}_k - \sum\nolimits_m \frac{\boldsymbol{x}_k}{\pi_k} \right]^t \widehat{\boldsymbol{B}} \\
&= \sum\nolimits_m \frac{y_k}{\pi_k} + \sum\nolimits_{m_I} \sum\nolimits_{U_i} \boldsymbol{x}_k^t \widehat{\boldsymbol{B}} - \sum\nolimits_m \frac{\boldsymbol{x}_k^t \widehat{\boldsymbol{B}}}{\pi_k}
\end{aligned} \tag{16}$$

Teniendo en cuenta que en un modelo lineal generalizado se relaciona el predictor lineal $\boldsymbol{x}_k^t \boldsymbol{\beta}$ con el valor esperado de la variable respuesta $E[y_k] = p_k$ mediante la función $g(p_k) = \boldsymbol{x}_k^t \boldsymbol{\beta}$. Se obtiene que el estimador del total de la variable Y asistido por un modelo lineal generalizado está dado por la ecuación (17), el cual es una generalización de la ecuación (16).

$$\widehat{t}_{yGLM} = \sum\nolimits_m \frac{y_k}{\pi_k} + \sum\nolimits_{m_I} \sum\nolimits_{U_i} g^{-1}\left(\boldsymbol{x}_k^t \widehat{\boldsymbol{B}} \right) - \sum\nolimits_m \frac{g^{-1}\left(\boldsymbol{x}_k^t \widehat{\boldsymbol{B}} \right)}{\pi_k} \tag{17}$$

En particular $g^{-1}\left(\boldsymbol{x}_k^t \widehat{\boldsymbol{B}} \right)$ es la probabilidad estimada $\widehat{p}_k$ de que un individuo tome el valor $y_k = 1$, dado que posee un vector de información auxiliar $\boldsymbol{x}_k$ bajo el uso de las funciones de enlace *Logit* y *Probit*. La estimación de β dada por $\widehat{\boldsymbol{B}}$ se obtiene mediante la ecuación (18).

$$\boldsymbol{X}^t \widehat{\boldsymbol{W}}^{(m)} \boldsymbol{\Pi}^{-1} \boldsymbol{X} \widehat{\boldsymbol{B}}^{(m+1)} = \boldsymbol{X}^t \boldsymbol{W}^{(m)} \boldsymbol{\Pi}^{-1} \widehat{\boldsymbol{z}}^{(m)} \tag{18}$$

En un modelo lineal generalizado la estimación de los parámetros del modelo se obtiene mediante un proceso iterativo como se refleja en la ecuación (9). La ecuación (18) es un caso general de la ecuación (9), donde $\mathbf{\Pi}$ es una matriz cuadrada diagonal con las probabilidades de inclusión π_k para los elementos de la muestra. Los elementos de la matriz diagonal $\widehat{\boldsymbol{W}}^{(m)}$ para la iteración m se calculan sobre los elementos de la muestra como indica la ecuación (10), y los elementos del vector $\widehat{\boldsymbol{z}}^{(m)}$ para la iteración m se calculan sobre los elementos de la muestra como indica la ecuación (11).

Luego para estimar el cociente de dos variables dicotómicas en un diseño muestral de dos etapas se propone el estimador dado por la ecuación (19), el cual utiliza toda la información auxiliar del universo..

$$\widehat{C}_{GLM} = \frac{\widehat{N}_{yGLM}}{\widehat{N}_{zGLM}} = \frac{\sum_m \frac{y_k}{\pi_k} + \sum_{m_I} \sum_{U_i} g^{-1}\left(\boldsymbol{x}_k^t \widehat{\boldsymbol{B}}\right) - \sum_m \frac{g^{-1}\left(\boldsymbol{x}_k^t \widehat{\boldsymbol{B}}\right)}{\pi_k}}{\sum_m \frac{z_k}{\pi_k} + \sum_{m_I} \sum_{U_i} g^{-1}\left(\boldsymbol{w}_k^t \widehat{\boldsymbol{B}}\right) - \sum_m \frac{g^{-1}\left(\boldsymbol{w}_k^t \widehat{\boldsymbol{B}}\right)}{\pi_k}} \tag{19}$$

3. Medidas de asociación de variables binarias

En estimación de totales, el nivel de asociación de las variable de interés y las variables auxiliares es importante para determinar la precisión del modelo. Por lo tanto, es de interés conocer cuál es el nivel de asociación de las variables y_k y z_k con las variables x_k y w_k respectivamente. Teniendo en cuenta que el conjunto de variables es de carácter dicotómico, se establece su nivel de asociación a través de un distinto conjunto de medidas. La siguiente tabla ayuda a determinar los coeficientes que se explican a continuación.

		x_k = 0	x_k = 1	
y_k	0	n_{00}	n_{01}	$n_{0\bullet}$
	1	n_{10}	n_{11}	$n_{1\bullet}$
		$n_{\bullet 0}$	$n_{\bullet 1}$	n

Tabla 1: *Tabla de contingencia de dos variables binarias y_k y x_k*

3.1. Coeficiente de contingencia

En el caso de variables dicotómicas se obtiene el coeficiente de contingencia a partir de la Tabla 1. En primer lugar, se calcula el estadístico de Ji-cuadrado como se

muestra en la ecuación (20), y finalmente se obtiene el coeficiente de contingencia como indica la ecuación (21).

$$\chi^2 = \sum_{i=0}^{1}\sum_{j=0}^{1} \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}} \tag{20}$$

$$C_C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{21}$$

En el caso de variables dicotómicas, el coeficiente de contingencia tiene un rango entre $0 \leq CC \leq \sqrt{1/2}$. El coeficiente de contingencia es un parámetro importante en el que se busca un estimador óptimo en este trabajo.

3.2. Coeficiente ϕ

El coeficiente de asociación ϕ es una medida de asociación derivada del estadístico de Pearson Ji-cuadrado. El coeficiente ϕ tiene un rango en $[-1, 1]$ para tablas $2x2$. El coeficiente ϕ es calculado como:

$$\phi = \frac{n_{00}n_{11} - n_{01}n_{10}}{\sqrt{n_{\bullet 0}n_{\bullet 1}n_{0\bullet}n_{1\bullet}}} \tag{22}$$

3.3. Coeficiente de Yule

El coeficiente de Yule se basa en el número de concordantes y discordances en los pares de observaciones. El coeficiente de Yule es apropiado solo cuando las variables son de escala ordinal. El rango del coeficiente de Yule está entre -1 y 1. Si las dos variables son independientes, el coeficiente de Yule se acerca a cero. El coeficiente de Yule se calcula como:

$$C_Y = \frac{n_{00}n_{11} - n_{01}n_{10}}{n_{11}n_{00} + n_{01}n_{10}} \tag{23}$$

En este trabajo se utiliza en particular la medida de asociación dada por el coeficiente de contingencia, ya que esta medida tiene en cuenta la asociación indiferentemente si la tendencia es creciente o decreciente.

4. No respuesta

La no respuesta es una característica frecuente pero indeseable de un estudio y puede afectar severamente la calidad de los estimativos que se calculan y se publi-

can. La idea es utilizar técnicas estadísticas que se desempeñen bien en presencia de no respuesta, ya que las inferencias estadísticas que se hacen basadas en datos afectados por tal fenómeno, corren un alto riesgo de ser inválidas. Hay dos caminos: evitar la no respuesta antes de que ocurra, o utilizar las técnicas de estimación adecuadas para manejar la no respuesta que ocurre. Esto último se conoce como ajuste por no respuesta (Särndal & Lundström 2005).

En particular para las variables y_k y z_k se tiene no respuesta por *item*, bajo distintos modelos de distribución de la respuesta. El tratamiento de la no respuesta se realiza mediante el proceso de imputación.

Dada una muestra m se define r como el conjunto de los datos respondientes. Para todo $e_k \in U$ se define θ_k como la probabilidad de respuesta del individuo e_k, Donde $\theta_k \in [0, 1]$. Generalmente la probabilidad de respuesta θ_k es desconocida. La distribución de la respuesta generalmente depende de variables de factores desconocidos. En particular Särndal & Lundström (2005) muestran dos formas de modelar la respuesta: $\theta_k = cte$ para todo $e_k \in U$ o distribución constante de la respuesta, e incremento exponencial de la distribución de la respuesta evaluado como $\theta_k = 1 - \exp(-c\, b_k)$, donde b_k es una variable desconocida. La tasa de respuesta se evalúa como el promedio de los valores θ_k en el universo.

La imputación es el procedimiento mediante el cual los valores faltantes de una o más variables de estudio son sustituidas. Esta sustitución puede ser construida en gran variedad de formas. En particular, entre la imputación por reglas estadísticas se utiliza la imputación por regresión. Si $e_k \in m - r$ su imputación es de la forma que indica la ecuación (24).

$$\widehat{y}_k = h(\boldsymbol{x}_k^t \widehat{\boldsymbol{\beta}}_{\boldsymbol{m}}) \tag{24}$$

El coeficiente de regresión $\widehat{\beta}_m$ es el resultado de regresión usando $(y_k, \boldsymbol{x}_k)$ disponible para $e_k \in m$.

5. Comportamiento empírico del sesgo y la varianza de estimadores asistidos por modelos

Con el fin de evaluar la precisión y la exactitud de los estimadores de cocientes, se genera una población con un diseño de dos etapas. La primera etapa con $N_I = 240$ conglomerados de muestreo. Cada conglomerado posee aproximadamente $N_i = 1900$ elementos para generar una población de $N = 450.000$ elementos aproximadamente. En cada población se generan variables binarias y_k y z_k tal que si $zk = 0$ entonces $y_k = 0$. Se fija un tamaño muestral de $n_I = 15$ y $n_i = 0.2N_i$ para todo $U_i \in U_I$, buscando obtener valores comparables de la precisión de los estimadores de cocientes.

Se realizan dos simulaciones de Monte Carlo con las siguientes características: la primera para obtener un escenario deseable según las asociación de las variables

auxiliares utilizadas, la segunda incluye la no respuesta para dos escenarios de distribución de la no respuesta para numerador y denominador.

5.1. Escenarios de variables auxiliares según coeficiente de contingencia

Se generan variables auxiliares binarias x_k y w_k que guardan un cierto grado de asociación con las variables y_k y z_k respectivamente. La asociación entre las variables se define como alto, medio y bajo según el coeficiente de contingencia dado en la sección 3. Siendo $C_{C_A} \approx 0.7$, $C_{C_M} \approx 0.3$ y $C_{C_B} \approx 0.001$.

Lo anterior se realiza para establecer la asociación entre y_k, x_k y z_k, w_k, originando nueve escenarios posibles para comparar. El parámetro de interés para el universo generado es $C = N_y/N_z = 0.321$. Para este universo se extrajeron 1500 muestras mediante diseño muestral MAS^2.

En particular para cada uno de los nueve escenarios de información auxiliar se trabajó con 6 estimadores de cocientes que se presentan a continuación.

1. Estimador convencional del cociente $\widehat{C}_\pi = N_{y\pi}/N_{z\pi}$

2. Estimador con uso de información auxiliar en el numerador dado por el modelo $y_k = \beta_0 + \beta_1\, x_k$. Cociente $\widehat{C}_x = \widehat{N}_{yreg}/\widehat{N}_{z\pi}$.

3. Estimador con uso de información auxiliar en el denominador dado por el modelo $z_k = \beta_0 + \beta_1\, w_k$. Cociente $\widehat{C}_w = \widehat{N}_{y\pi}/\widehat{N}_{zreg}$.

4. Estimador con uso de información auxiliar en el numerador y el denominador dado por los modelos $y_k = \beta_0 + \beta_1\, x_k$ y $z_k = \beta_0 + \beta_1\, w_k$ respectivamente. Cociente $\widehat{C}_{xw} = \widehat{N}_{yreg}/\widehat{N}_{zreg}$.

5. Estimador con uso de información auxiliar tanto en el numerador como en el denominador dado por los modelos lineales generalizados con función de enlace Logit, $Logit(p_{yk}) = \beta_0 + \beta_1\, x_k$ y $Logit(p_{zk}) = \beta_0 + \beta_1\, w_k$ respectivamente. Cociente $\widehat{C}_{Logit} = \widehat{N}_{yGLMl}/\widehat{N}_{zGLMl}$.

6. Estimador con uso de información auxiliar en el numerador y el denominador dado por los modelos lineales generalizados con función de enlace *Probit*, $\Phi(p_{yk}) = \beta_0 + \beta_1\, x_k$ y $\Phi(p_{zk}) = \beta_0 + \beta_1\, w_k$ respectivamente. Cociente $\widehat{C}_{Probit} = \widehat{N}_{yGLMp}/\widehat{N}_{zGLMp}$.

Para las 1500 simulaciones dadas por el método de Monte Carlo en los seis estimadores, se toma el promedio de las estimaciones como estimador de la esperanza del estimador del cociente dado en la Tabla 2, la varianza muestral como estimador de la varianza del estimador del cociente. Se obtiene el cálculo del sesgo relativo estimado en la Tabla 3 y el coeficiente de variación estimado en la Tabla 4.

Asociación numerador	Asociación denominador	$\widehat{C}_\pi$	$\widehat{C}_x$	$\widehat{C}_w$	$\widehat{C}_{xw}$	$\widehat{C}_{Logit}$	$\widehat{C}_{Probit}$
Alta	Alta	0.323	0.323	0.323	0.321	0.321	0.321
	Media	0.323	0.322	0.323	0.321	0.321	0.321
	Baja	0.321	0.323	0.320	0.321	0.321	0.321
Media	Alta	0.320	0.321	0.320	0.320	0.320	0.319
	Media	0.321	0.322	0.322	0.321	0.321	0.320
	Baja	0.320	0.322	0.319	0.319	0.319	0.318
Baja	Alta	0.322	0.323	0.323	0.322	0.322	0.322
	Media	0.323	0.325	0.322	0.323	0.323	0.323
	Baja	0.322	0.325	0.322	0.322	0.322	0.322

Tabla 2: *Estimaciones de la esperanza de Cocientes de método Monte Carlo por tipo de información auxiliar, según estimador de Cociente ($C = 0.321$)*

En la Tabla 2, los promedios de las estimaciones son prácticamente iguales al valor poblacional del cociente $C = 0.321$. No hay cambios significativos según el tipo de información auxiliar en el denominador y el numerador. El promedio de las estimaciones no tiene cambio por el tipo de estimador utilizado.

Asociación numerador	Asociación denominador	$\widehat{C}_\pi$	$\widehat{C}_x$	$\widehat{C}_w$	$\widehat{C}_{xw}$	$\widehat{C}_{Logit}$	$\widehat{C}_{Probit}$
Alta	Alta	−3.46	−7.72	−3.13	2.60	2.60	−0.10
	Media	−3.38	−5.45	−3.20	7.29	7.29	4.47
	Baja	0.47	−8.33	1.55	4.45	4.45	1.17
Media	Alta	2.74	−0.15	1.73	4.27	4.27	7.39
	Media	−0.73	−2.32	−0.96	1.90	1.90	5.50
	Baja	2.74	−1.63	3.78	5.59	5.59	9.67
Baja	Alta	−1.25	−3.37	−2.92	−1.95	−1.95	−2.83
	Media	−3.62	−7.81	−2.27	−3.71	−3.71	−3.32
	Baja	−2.94	−7.52	−0.85	−2.93	−2.93	−2.50

Tabla 3: *Estimaciones del sesgo relativo porcentual de método Monte Carlo por tipo de información auxiliar, según estimador de Cociente*

En la Tabla 3 de sesgo relativo se puede observar que hay una ligera sobre-estimación por parte de los estimadores que no utilizan regresión logística y para los que utilizan información de mala calidad en el numerador, observado en los sesgos relativos negativos. Los estimadores asistidos por regresión logística con uso de información auxiliar alta y media para el numerador muestran una ligera sub-estimación observada en los signos positivos del sesgo relativo. Sin embargo, ningún sesgo relativo porcentual sobrepasa el 10 %, por lo que se considera que el sesgo es insignificante.

En la Tabla 4 se puede ver que los estimadores asistidos por regresión logística son más adecuados que el convencional C_π y los asistidos por regresión lineal. En particular, los mejores escenarios en cuanto a precisión se observan cuando existe la información auxiliar para asistir el total del numerador, ya que ésta es alta.

Asociación numerador	Asociación denominador	$\widehat{C}_\pi$	$\widehat{C}_x$	$\widehat{C}_w$	$\widehat{C}_{xw}$	$\widehat{C}_{Logit}$	$\widehat{C}_{Probit}$
Alta	Alta	12.94	7.52	15.54	0.80	0.80	0.64
	Media	12.90	7.62	14.81	1.21	1.21	1.68
	Baja	13.04	7.67	14.83	1.38	1.38	2.01
Media	Alta	13.15	12.76	15.47	10.66	10.66	9.75
	Media	13.37	12.62	15.10	9.98	9.98	9.21
	Baja	12.95	12.26	14.56	9.46	9.46	8.86
Baja	Alta	13.03	15.44	15.50	14.09	14.09	14.29
	Media	13.10	15.36	15.20	13.31	13.31	13.68
	Baja	13.27	15.61	14.62	13.27	13.27	13.56

Tabla 4: *Estimaciones del coeficiente de variación porcentual de método Monte Carlo por tipo de información auxiliar, según estimador de Cociente*

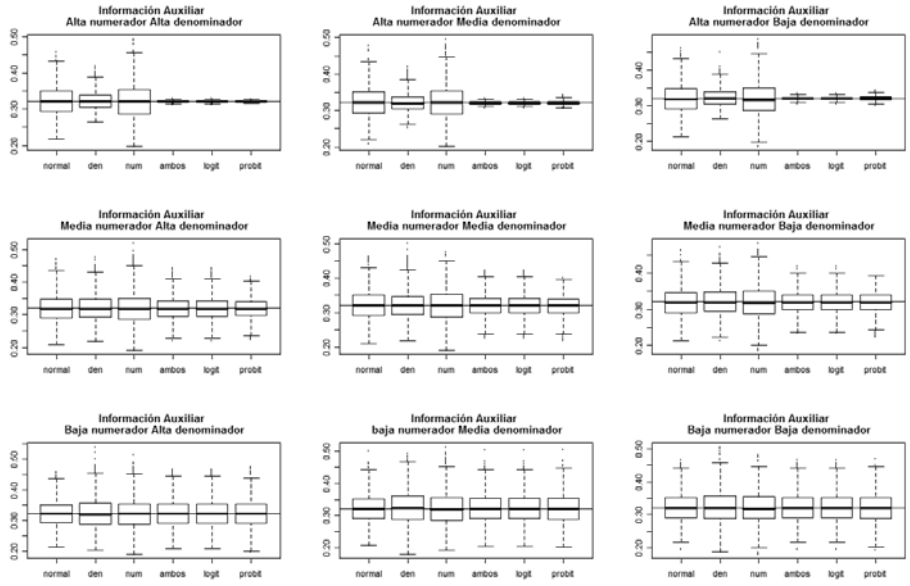

Figura 1: *Comparación de estimadores de cocientes según tipo de información auxiliar.*

La Figura 1 ilustra claramente que el sesgo no es relevante para ninguno de los estimadores en ninguno de los escenarios de uso de información auxiliar. Sin embargo, se observa una clara diferencia de variación en los estimadores asistidos por regresión logística cuando la información auxiliar para el denominador es alta. Bajo estas condiciones se selecciona de los mejores escenarios el más simple: información auxiliar de alta calidad para el numerador e información auxiliar de baja calidad para el denominador.

5.2. Imputación por regresión logística

El objetivo de esta simulación es dar un ilustración del comportamiento de los estimadores asistidos por regresión logística bajo condiciones de no respuesta. El tratamiento de no respuesta elegido es la imputación mediante uso de regresión logística como lo indica la ecuación (24).

A partir del universo generado en la sección 5.1, en la vida real la distribución de la respuesta es desconocida. Por lo tanto, la simulación cubre dos tipos de distribución de la respuesta posibles.

- *Distribución de respuesta constante*: esta se basa en que la probabilidad que un individuo conteste es constante para todos los individuos del universo $\theta_k = cte$.
- *Incremento exponencial de la respuesta*: esta probabilidad es calculada por $\theta_k = 1 - \exp(-c\, b_k)$. Bajo esta estructura con c y b_k mayores que cero, se ve claramente que la probabilidad de ser respondiente es mayor a medida que la variable b_k aumente. La variable b_k es un valor desconocido en las encuestas reales.

Para la no respuesta de z_k se tomaron las dos distribuciones de la respuesta tal que la taza de respuesta sea del 82 %. En particular para la distribución de respuesta constante $\theta_{kz} = 0.82$ para todo $e_k \in U$, para el Incremento exponencial de la respuesta $\theta_{kz} = 1 - exp(-c_z\, b_{kz})$ se genera una variable ordinal v_{zk} de seis categorías tal que si $z_k = 1$ la variable categórica toma con mayor probabilidad los valores 4, 5 y 6 que 1, 2 y 3; cuando $z_k = 0$ la variable categórica v_{zk} toma con mayor probabilidad los valores 1, 2 y 3 que 4, 5 y 6. Luego b_k es una variable continua de valores positivos que toma valores grandes a medida que la categoría toma valores numéricos más altos. La Tabla 5 da una idea de la relación de las seis categorías y la variable z_k.

	1	2	3	4	5	6	Total z_k
0	18313	14058	13994	13801	4606	4553	69325
	26.42 %	20.28 %	20.19 %	19.91 %	6.64 %	6.57 %	100 %
1	26030	26205	78409	78084	78187	104366	391281
	6.65 %	6.70 %	20.04 %	19.96 %	19.98 %	26.67 %	100 %
Total	44343	40263	92403	91885	82793	108919	460606
Categorías	9.63 %	8.74 %	20.06 %	19.95 %	17.97 %	23.65 %	100 %

Tabla 5: *Tabla de contingencia variable v_{zk} contra z_k*

Luego con el uso de método de Newton-Raphson para solución de ecuaciones se obtienen que $c_z = 0.0113$. Análogamente se establecen dos modelos de distribución de la respuesta para la variable y_k esperando una taza de respuesta del 70 % bajo el modelo de distribución constante e incremento exponencial de la respuesta. Obteniendo un valor $c_y = 0.0188$.

En particular se manejan cuatro escenarios que dependen de la distribución de la respuesta para el numerador y el denominador.

Se toman 1500 muestras con $n_I = 15$ y $n_i = 0.2\,N_i$ para cada conglomerado $U_i \in m_I$. Para cada muestra se crea la no respuesta para el numerador y el denominador. Para cada elemento e_k de la muestra m se origino un experimento Bernoulli con parámetro θ_k. Esto es, el elemento pertenece a la muestra con probabilidad θ_k y a los no respondientes con probabilidad $1 - \theta_k$, donde θ_k es conocido para todo $e_k \in U$ en ambas distribuciones. Así, se realizó un experimento Bernoulli para cada individuo de la muestra m, obteniendo un conjunto de 1500 muestras de respondientes; para cada conjunto escenario de distribución de respondientes en el numerador y el denominador, se calculan cinco diferentes estimadores.

- $\widehat{C}_{\pi 0}$ estimador convencional de un cociente con no respuesta imputada de la forma simple $\widehat{y}_k = 0$ y $\widehat{z}_k = 0$.
- $\widehat{C}_{\pi 1}$ estimador convencional de un cociente con no respuesta imputada mediante regresión logística, función de enlace *Logit*.
- $\widehat{C}_{\pi 2}$ estimador convencional de un cociente con no respuesta imputada mediante regresión logística, función de enlace *Probit*.
- $\widehat{C}_{Logit}$ estimador con uso de información auxiliar en el numerador y el denominador dado por los modelos lineales generalizados con función de enlace *Logit*, $Logit(p_{yk}) = \beta_0 + \beta_1\, x_k$ y $Logit(p_{zk}) = \beta_0 + \beta_1\, w_k$ respectivamente. Estimador con no respuesta imputada mediante regresión logística, función de enlace *Logit*.
- $\widehat{C}_{Probit}$ estimador con uso de información auxiliar en el numerador y el denominador dado por los modelos lineales generalizados con función de enlace *Probit*, $\Phi(p_{yk}) = \beta_0 + \beta_1\, x_k$ y $\Phi(p_{zk}) = \beta_0 + \beta_1\, w_k$ respectivamente. Estimador con no respuesta imputada mediante regresión logística, función de enlace *Probit*.

Análogamente a la simulación de Monte Carlo bajo distintos escenarios de información auxiliar, se obtienen resultados de esperanza del cociente estimada (Tabla 6), sesgo relativo estimado (Tabla 7) y coeficiente de variación estimado (Tabla 8).

Distribución respuesta numerador	Distribución respuesta denominador	$\widehat{C}_{\pi 0}$	$\widehat{C}_{\pi 1}$	$\widehat{C}_{\pi 2}$	$\widehat{C}_{Logit}$	$\widehat{C}_{Probit}$
Constante	Constante	0.263	0.320	0.320	0.321	0.321
Exponencial		0.244	0.321	0.321	0.324	0.324
Constante	Exponencial	0.264	0.312	0.312	0.311	0.311
Exponencial		0.245	0.313	0.313	0.313	0.314

Tabla 6: *Estimaciones de la esperanza de Cocientes de método Monte Carlo por tipo de distribución de respuesta, según estimador de Cociente ($C = 0.321$)*

En los cuatro estimadores y los cuatro escenarios de distribución de la respuesta, el estimador con imputación simple subestima el cociente de las variables dicotómicas.

Distribución respuesta numerador	Distribución respuesta denominador	$\widehat{C}_{\pi 0}$	$\widehat{C}_{\pi 1}$	$\widehat{C}_{\pi 2}$	$\widehat{C}_{Logit}$	$\widehat{C}_{Probit}$
Constante	Constante	171.04	2.04	2.04	−0.58	−3.21
Exponencial		241.70	1.20	1.20	−50.99	−40.86
Constante	Exponencial	165.06	22.97	22.97	253.72	178.92
Exponencial		232.85	20.07	20.07	191.08	130.33

Tabla 7: *Estimaciones del sesgo relativo porcentual de método Monte Carlo por distribución de la respuesta, según estimador de Cociente*

El sesgo relativo es mayor de diez en todos los escenarios y estimadores. Aunque es menor en los estimadores que utilizan imputación por modelo logístico y estimadores convencionales.

Distribución respuesta numerador	Distribución respuesta denominador	$\widehat{C}_{\pi 0}$	$\widehat{C}_{\pi 1}$	$\widehat{C}_{\pi 2}$	$\widehat{C}_{Logit}$	$\widehat{C}_{Probit}$
Constante	Constante	13.03	12.94	12.94	1.49	2.07
Exponencial		13.10	12.85	12.85	1.52	2.14
Constante	Exponencial	13.25	13.39	13.39	1.28	1.80
Exponencial		13.26	13.26	13.26	1.28	1.84

Tabla 8: *Estimaciones del coeficiente de variación porcentual de método Monte Carlo por distribución de la respuesta, según estimador de Cociente*

En la Tabla 8 se puede ver que la precisión es mejor para estimadores asistidos por regresión logística que sin el uso de información auxiliar. Bajo imputación logística la precisión no cambia en comparación con el estimador con imputación simple.

En la Figura 2 el estimador que utiliza imputación simple subestima el cociente. La alta variabilidad de los estimadores que no usan información auxiliar amortigua el efecto causado por el sesgo; sin embargo, los estimadores asistidos por regresión logística tiene un sesgo visible debido a su baja variabilidad.

6. Resultados

Este trabajo compara inicialmente distintos estimadores de un cociente de variables dicotómicas, asistidos por regresión lineal simple y modelos lineales generalizados. Estos estimadores se utilizan bajo distintos escenarios de información auxiliar para el numerador y el denominador. Luego, para el mejor escenario se generan distintos escenarios de no respuesta para el numerador y el denominador con el uso de estimadores de cocientes asistidos por modelos lineales generalizados. El proceso

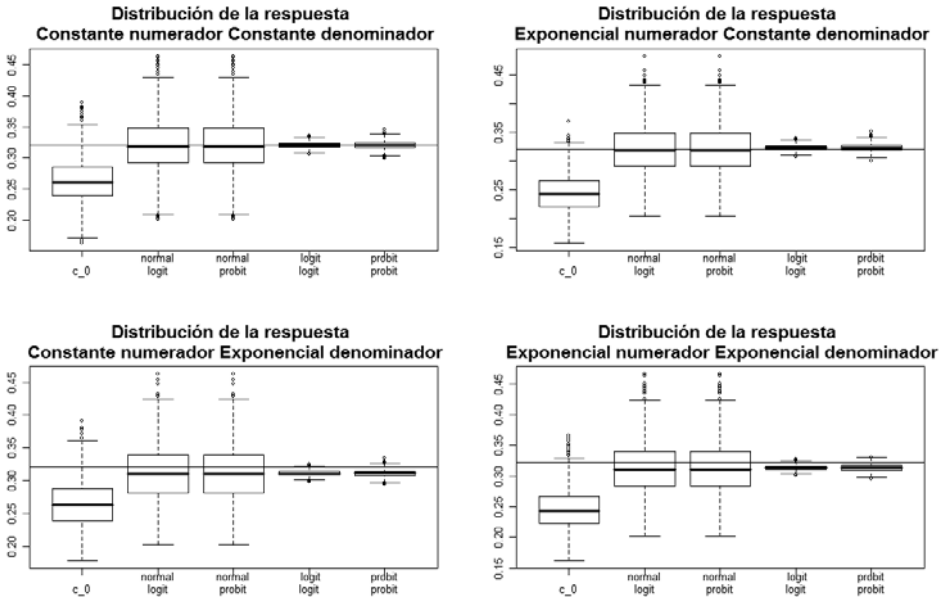

Figura 2: *Comparación de estimadores de cocientes según tipo de distribución de la respuesta*

de programación de generación y obtención de estimadores mediante simulación de Monte Carlo se realizó en SAS (2004), el proceso de análisis de resultados se realizó en lenguaje de programación R Development Core Team (2006) versión 2.5.1.

Como resultado de los ejercicios realizados en este trabajo, se propone, para la estimación de un cociente de variables dicotómicas con un diseño muestral MAS^2, el uso de un estimador que utilice información auxiliar de tipo binario en el numerador tal que discrimine prácticamente la variable y_k. Se recomienda el uso de información auxiliar para el denominador sin importar su nivel de asociación con la variable z_k. Esto reducirá drásticamente el coeficiente de variación del estimador como se puede ver en la Tabla 4, donde los coeficientes de variación más bajos se encuentran con el uso de buena información auxiliar en el numerador. También la Figura 1 muestra que los cambios de variabilidad de los estimadores se ven reflejados en estimadores con buen uso de información auxiliar en el denominador. Según la Tabla 3, el sesgo relativo porcentual no es mayor de 10 %, en valor absoluto, para ninguno de los estimadores. Además la Figura 1 muestra que la media de todos los estimadores está muy cercana al cociente verdadero a estimar, reflejado en la linea horizontal que atraviesa cada gráfico. Por lo tanto, el sesgo relativo es insignificante para cualquiera de los escenarios propuestos de información auxiliar en el numerador y el denominador.

Bajo no respuesta el estimador con imputación simple muestra una subestimación significativa que se puede ver en la Tabla 6 y la Figura 2. La imputación soluciona problemas en la exactitud del estimador. La asistencia de un modelo logístico soluciona problemas de precisión. El estimador que utiliza función de enlace *Logit* es ligeramente mejor que el estimador que utiliza función de enlace *Probit* en términos de precisión.

El estimador asistido por regresión logística con función de enlace *Logit* es ligeramente mejor que el de función de enlace *Probit* bajo condiciones de respuesta completa como de no respuesta. El sesgo relativo es relativamente el mismo bajo las condiciones de respuesta completa y bajo la no respuesta.

Agradecimientos

Gracias al profesor Leonardo Bautista por introducirme en el mundo del muestreo, por colaborarme y sobrellevarme en el desarrollo de este trabajo. Agradezco a la Universidad Nacional de Colombia por darme las herramientas para mi desarrollo como profesional y a la Universidad Santo Tomás por invitarme a ser parte en este volumen.[1]

Recibido: 5 de febrero de 2009
Aceptado: 18 de mayo de 2009

Referencias

Dobson, A. (1990), *An introduction to Generalized Linear Models*, Chapman & Hall, Australia.

Lehtonen, R. & Pahkinen, E. (2004), *Practical Methods for design and Analisys of Complex Surveys*, Jhon wiley & Sons, Ltd, Finland.

R Development Core Team (2006), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org

Särndal, C.-E. & Lundström, S. (2005), *Estimation in Surveys with nonresponse*, Jhon wiley & Sons, Ltd, Sweden.

[1]Este artículo es resultado del trabajo de grado para optar por el titulo de estadístico titulado *Estimadores de regresión logística para tratamiento de no respuesta en el caso de cocientes de variables dicotómicas*, dirigido por el maestro Leonardo Bautista y presentado en la Universidad Nacional de Colombia en 2007.

Särndal, C.-E., Swensson, B. & Wretman, J. (1993), *Model Assisted Survey Sampling*, Springer, New York.

SAS (2004), Sas onlinedoc 9.1.3., SAS Institute Inc, Cary, NC.
*http://www.sas.com