ALGUNAS ALTERNATIVAS PARA ESTIMAR LA MATRIZ DE COVARIANZA DE UN ESTIMADOR EN UN MODELO DE REGRESION LINEAL*

Fernando Esquivel Bocanegra, Departamento de Estadística y Cálculo, Universidad Autónoma Agraria "Antonio Narro"
Ignacio Méndez Ramírez, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México
Gustavo Ramírez Valverde, Instituto de Socioeconomía. Estadística e Informática. Colegio de Postgraduados

RESUMEN

Para el análisis de datos en muestras complejas, comúnmente el usuario recurre a los procedimientos tradicionales (o estándar) que están implementados en los paquetes estadísticos computacionales convencionales, en donde, -por default- suponen que las observaciones obtenidas en la muestra son independientes e idénticamente distribuidas. Por lo cual, ignoran el diseño de muestreo complejo, o bien la forma estructural que presenta la población de interés, ocasionando graves distorsiones en las inferencias de tipo analítico. Sobre la base de lo anterior, se examinan algunas alternativas para estimar la matriz de covarianza. Además, se proponen algunas correcciones simples para obtener estimadores consistentes de los errores estándar del vector de coeficientes en un modelo de regresión lineal.

ABSTRACT

Commonly in data analysis with Complex samples, the user uses traditional procedures or standard which are implemented in conventional computing statistical packages, where, by default is assumed that the observations were obtained from an independently and identically distributed sample. Hence, the use of a complex desing or the structure of the population is ignored, generating important distortions in the development of analytical inferences. Based on these arguments, some alternatives for estimating the covariance matrix are examined. In addition, some simple corrections are proposed for obtaining consistent estimators of the standard errors of the vector of the linear regression coefficients.

Key words: linear regression, analytical surveys, complex sampling design, design effect.

MSC: 62J05

INTRODUCCION

Desde el punto de vista conceptual, las encuestas por muestreo pueden ser divididas en dos grandes ramas: encuestas *descriptivas* y encuestas *analíticas*. En lo que corresponde al uso de encuestas descriptivas, están dirigidas a la estimación -precisa y eficientemente- de ciertas características poblacionales, usualmente pocas, tales como medias, frecuencias, etc. Por otra parte, Deming (1950) describe las encuestas analíticas como …"dirigidas a las causas fundamentales que originan las frecuencias de varias clases de la población correspondiente, y que determinarán las frecuencias de estas clases en lo futuro". De

*This paper was presented at VIII Conferencia Latinoamericana de Probabilidades y Estadística Matemática.

acuerdo con esto, el uso analítico va más allá que medir parámetros poblacionales implicados en la descripción, considerando la explicación causal de los procesos que se extiende debajo de las medidas descriptivas.

En ambos casos, las encuestas descriptivas y analíticas pueden ser complejas, es decir, involucran un diseño de muestreo complejo tal y como un *muestreo por conglomerados estratificado polietápico*. Tener en cuenta las complejidades del muestreo es esencial para garantizar la validez de las estimaciones y análisis en ambos tipos de encuestas. Lo cual indica que, las complejidades que se presentan en el diseño de muestreo están frecuente e íntimamente conectadas con las complejidades propias del procedimiento de estimación. En ocasiones, consciente o inconscientemente se recurre a procedimientos tradicionales (o estándar), en la etapa de estimación, que no contemplan la forma estructural que presenta la población de interés. Precisamente esto último nos da la pauta para presentar las intenciones y los objetivos que se persiguen en el desarrollo de este trabajo.

De manera general, la meta es describir algunas opciones que permiten calcular estimaciones consistentes para la matriz de covarianza de un estimador del vector de coeficientes en un modelo de regresión lineal. Para esto, se mencionan métodos de estimación que ignoran, o que no ignoran, las complejidades del diseño de muestreo. Además, se proponen algunas correcciones simples, con la finalidad de ajustar la precisión de las estimaciones que se obtiene en un paquete computacional estadístico convencional. Para llevar a cabo estas correcciones se hace uso, principalmente, de lo que se conoce como *efecto del diseño.*

Por otra parte, y con la intención de realizar una Simulación para contrastar los resultados que más tarde serán presentados, se considera información censal publicada por INEGI[1] bajo el nombre de NIBA (Niveles de Bienestar por AGEB). Este producto contiene información sobre las características regionales de México, expresadas a través de indicadores relacionados con Demografía, Ocupación, Vivienda y, en general, con el Desarrollo Socioeconómico de las áreas geoestadísticas Básicas (AGEB), tanto urbanas como rurales del país, conforme a datos del XI Censo General de Población y Vivienda, 1990. En nuestro caso, solamente hemos considerado algunas de las áreas geoestadísticas básicas que están contempladas en el medio rural, para describir la variación sistemática relacionada con el alfabetismo, sobre la base de un modelo de regresión lineal. De acuerdo con esto, nuestra intención es comparar, a través de un estudio de simulación, diferentes métodos para estimar los errores estándar del vector de coeficientes del modelo. Así mismo, se sugieren algunas correcciones simples que ajusten la precisión de los errores estándar obtenidos a partir de un paquete estadístico computacional de uso común.

Especificación del modelo

En el análisis de regresión lineal múltiple, se supone un modelo en el cual una variable y (variable *respuesta*) está relacionada a un conjunto de variables explicativas. Sea y_i una variable dependiente observada. La manera más común para analizar la dependencia de y_i sobre las variables explicativas es el modelo de regresión lineal univariado

$$y_i = x_i'\beta + \varepsilon_i, \qquad i = 1,2,...,n \tag{1}$$

donde, x_i es un vector q x 1 de variables explicativas observadas, β es un vector q x 1 de coeficientes de regresión, y ε_i es una variable aleatoria no observada llamada el *error* o perturbación. El primer elemento de x_i es usualmente igual a 1, así que, β_1 es la constante de regresión o intercepto. Si $E(\varepsilon_i|x_i) = 0$ entonces

[1]Instituto Nacional de Estadística, Geografía e Informática.

$E(y_i|x_i) = \mu_i = x_i'\beta$, lo cual constituye la parte sistemática del modelo. La función μ_i es llamada estructura de la media de y_i dado x_i . El modelo de regresión lineal para la muestra total puede ser escrito en notación matricial como

$$y = X\beta + \varepsilon \tag{2}$$

con $y = (y_1,...,y_i,...,y_n)'$ como el vector de variables respuesta, X_{nxq} la matriz de regresores; β_{qxl} y ε_{nxl} son el vector de coeficientes y el vector de errores, respectivamente. La matriz X es también conocida como la *matriz del modelo* ya que desempeña un papel importante en la especificación de la estructura de la media, $\mu_i = x_i'\beta$. En el ámbito experimental, X es también llamada *matriz diseño*.

Para la estimación del parámetro β en un modelo de regresión lineal, la teoría estadística clásica se basa, principalmente en los siguientes supuestos:

Supuesto 1. $E(\varepsilon_i) = 0 \Rightarrow$ descripción correcta de la distribución de la media.

Supuesto 2. x_i es no-estocástica; X es (columnas) de rango completo.

Supuesto 3. $V(\varepsilon_i) = \sigma 2 \Rightarrow$ homoscedasticidad.

Supuesto 4. $\varepsilon_i \sim N(0, \sigma^2)$; distribución normal de los errores.

Estas suposiciones clásicas para la estimación de β, en la práctica, difícilmente se cumplen. En primer lugar, las variables explicativas x_i pueden ser estocásticas; en donde de supone que los parámetros de la densidad de x_i son distintos respecto a los parámetros de la densidad condicional de y_i dado x_i, (Arminger, **et al**., 1995). Por tal motivo, los supuestos $E(\varepsilon_i) = 0$ y $V(\varepsilon_i) = \sigma^2$ son reemplazados por $E(\varepsilon_i|x_i) = 0$ y $V(\varepsilon_i|x_i) = \sigma^2$, respectivamente. La suposición de que $E(\varepsilon_i|x_i) = 0$, implica

$$\begin{aligned} Cov(\varepsilon_i|x_i) &= E(\varepsilon_i|x_i) - E(\varepsilon_i)E(x_i) \\ &= E(\varepsilon_i|x_i) - E[E(\varepsilon_i|x_i)]E(x_i) \\ &= E(\varepsilon_i|x_i) \end{aligned}$$

A partir de esto, se puede ver que

$$\begin{aligned} E(\varepsilon_i, x_i) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varepsilon_i x_i f_{\varepsilon_i|x_i} f_{x_i} d\varepsilon_i dx_i \\ &= \int_{-\infty}^{\infty} x_i f_{x_i} [E(\varepsilon_i|x_i)] dx_i \\ &= 0 \end{aligned}$$

lo cual significa que, el término del error y las variables explicativas no están correlacionadas.

Para la discusión que a continuación abordaremos, suponga que el modelo de regresión para los datos muestrales (y, X) está dado por

$$y = X\beta + \varepsilon \qquad (3)$$

donde:

$$E(\varepsilon|X) = 0, \quad V(\varepsilon|X) = \sigma^2\Omega$$

con lo cual, E(ylX) = Xβ y V (y|X) = $\sigma^2\Omega$, siendo Ω la matriz de varianzas y covarianzas del vector de observaciones. Sobre la base de este modelo (3), enseguida se presentan diferentes métodos de estimación para el vector de coeficientes en un modelo de regresión lineal.

ESTIMACION DE LOS COEFICIENTES DE REGRESION

• Método de estimación OLS

Considerando el modelo (3), el estimador de *cuadrados mínimos ordinarios* (OLS) del vector β_{qXl} se obtiene minimizando la función objetivo

$$Q(\beta) = (y - X\beta)'(y - X\beta) \qquad (4)$$

dando

$$\hat{\beta} = (X'X)^{-1}X'y \qquad (5)$$

cuya matriz de covarianza -verdadera- del estimador $\hat{\beta}$, es

$$V_{real}(\hat{\beta}|X) = (X'X)^{-1}X'\Omega X(X'X)^{-1}\sigma^2 \qquad (6)$$

El procedimiento de estimación estándar de cuadrados mínimos ordinarios, implementado en los paquetes estadísticos computacionales convencionales, supone implícitamente que los errores ε_i (i =1,2,...,n) son independientes e idénticamente distribuidos (lID), tal y como los considera el modelo clásico; adoptando como estimador para la matriz de covarianza del vector (5), la expresión

$$\hat{V}_{pc}(\hat{\beta}|X) = (X'X)^{-1}\hat{\sigma}^2 \qquad (7)$$

siendo $\hat{\sigma}^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/(n - q)$. Con lo cual, el método OLS no incorpora en el análisis la estructura compleja del diseño de muestreo (*conglomerados, estratos, o medidas del tamaño*), tampoco las probababilidades diferentes de inclusión de las observaciones en la muestra. Entonces, el uso de la opción OLS bajo el modelo (3), conduce a un estimador puntual insesgado de β, sin embargo, las inferencias respecto a la precisión de $\hat{\beta}$ son incorrectas, (Skinner **et al**., 1989).

Generalmente, las encuestas que se realizan a gran escala están basadas en un esquema de muestreo, complejo. En estos casos, las observaciones incluidas en la muestra -usualmente- son seleccionadas con probabilidades diferentes. Por tal razón, las mediciones que se obtienen sobre una variable aleatoria en este tipo de encuestas, no deben ser consideradas como una muestra de observaciones independientes e idénticamente distribuidas, como frecuentemente se supone en la teoría estadística tradicional y en una gran variedad de paquetes computacionales diseñados para el análisis estadístico. Entonces, llevar a cabo un

análisis sin tomar en cuenta las diferentes probabilidades de selección puede conducir a sesgos en los parámetros de regresión estimados, (Nathan y Smith, 1989).

Ponderar las observaciones, es decir, asignar un valor positivo a cada elemento del conjunto de datos, tiene como finalidad compensar la desproporcionalidad de la muestra respecto a la población de interés. En la inferencia descriptiva (inferencia respecto a funciones conocidas de los valores en poblaciones finitas), la ponderación de los datos en la muestra es ampliamente aceptada; mientras que, en la inferencia analítica (inferencia respecto a parámetros en un modelo) se percibe en la literatura una gran diversidad de opiniones sobre la relevancia o no relevancia que refleja la incorporación de variables ponderadas al proceso de inferencia, (Pfeffermann, 1993). En nuestro caso, seguiremos la línea de quienes comparten la idea de incorporar elementos ponderados a los estimadores.

• Método de estimación WLS

La ponderación llevada a cabo a través de las probabilidades de inclusión muestral $\pi = (\pi_1,\ldots,\pi_n)'$ conduce a un análisis del modelo descrito en la ecuación (3), mediante el método de *cuadrados mínimos ponderados* (WLS). Si

$$\Pi = \text{diag}(\pi) = \begin{pmatrix} \pi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \pi_n \end{pmatrix}$$

entonces el estimador WLS de β es

$$\hat{\beta}_W = (X'\Pi^{-1}X)^{-1}X'\Pi^{-1}y \qquad (8)$$

el cual, bajo el modelo (3), es también un estimador insesgado de β. La opción WLS en un paquete computacional estadístico, usualmente estima la matriz de covarianza del vector $\hat{\beta}_W$ mediante la ecuación

$$\hat{V}_{pc}(\hat{\beta}_W|X) = (X'\Pi^{-1}X)^{-1}\tilde{\sigma}^2 \qquad (9)$$

siendo $\tilde{\sigma}^2 = (y - X\hat{\beta})'\Pi^{-1}(y - X\hat{\beta})/(n-q)$; la cual, bajo el modelo (3), es diferente de la verdadera matriz de covarianza del vector $\hat{\beta}_W$, dada por

$$\hat{V}_{real}(\hat{\beta}_W|X) = (X'\Pi^{-1}X)^{-1}X'\Pi^{-1}\Omega\Pi^{-1}X(X'\Pi^{-1}X)^{-1}\sigma^2 \qquad (10)$$

Así que, las inferencias respecto a la precisión de $\hat{\beta}_W$ serán incorrectas; excepto cuando $\Omega = \Pi$, lo cual es poco probable en la práctica, (Skinner **et al**., 1989). De esta manera, es posible advertir que en la etapa de estimación de la matriz de covarianza del vector estimador de coeficientes de regresión, algunos paquetes computacionales se basan en procedimientos simples, aunque poco convincentes.

Uno de los procedimientos - extensamente utilizado - para estimar la matriz de covarianza de un estimador no-lineal es conocido como **método de linealización**. Sobre la base de este método, se presentan algunas

opciones que permiten calcular estimaciones consistentes para la matriz de covarianza de un estimador del vector de coeficientes en un modelo de regresión lineal.

Para estimar las varianzas, también pueden ser usados otros métodos, tales como *bootstrap* o cualquier otro método de remuestreo (ver: Good, 1999); sin embargo, éstos métodos de remuestreo no serán considerados en este trabajo; en lugar de esto, sugerimos comparar nuestros resultados con los métodos que a continuación serán presentados.

• Estimadores Horvitz-Thompson

De manera equivalente, el vector $\hat{\beta}_W$ puede ser expresado como

$$\hat{\beta}_W = \hat{\beta}_\pi = \hat{T}^{-1}\hat{t} \tag{11}$$

donde

$$\hat{T} = \sum_k x_k x_k' / \pi_k \qquad \text{y} \qquad \hat{t} = \sum_k x_k y_k / \pi_k$$

siendo $x_k = (1, x_{2k},\ldots,x_{qk})'$ el vector de predictores de la variable respuesta y_k, para todo k ε s, s = {1, 2,..., n} y, π_k es la probabilidad de inclusión del k-ésimo elemento en la muestra s de tamaño n. Usando el método de linealización, el estimador de la matriz de covarianza del vector estimador[2] $\hat{\beta}_\pi$, está dado por

$$\hat{V} = (\hat{\beta}_\pi) = \hat{T}^{-1}\hat{\Sigma}_\pi\hat{T}^{-1} \tag{12}$$

donde $\hat{\Sigma}_\pi$ es la matriz q × q cuyo elemento ij está dado por $\Sigma_k\Sigma_l\Delta_{kl}(u_{ik}/\pi_k)(u_{jl}/\pi_l)$ con $u_{ik} = x_{ik}(y_k - x_k'\hat{\beta}_\pi)$; i,j = 1,2,...,q; k,l ∈ s, y $\breve{\Delta}_{kl} = \dfrac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}$, donde π_{kl} es la probabilidad de inclusión conjunta de los elementos k y l en la muestra s (Särndal **et al**., 1992).

Para un muestreo por conglomerados bietápico, la población de elementos U = {1,..., k,...,N} es particionada en N_I *unidades primarias de muestreo* (UPM) denotadas por $U_1,\ldots,U_i,\ldots,U_{NI}$. La población de UPM está dada por $U_I = \{I,\ldots,i,\ldots,N_I\}$, en donde el número de elementos en U_i es N_i, tal que $N = \Sigma_{i\in U_I}N_i$. En este caso, las probabilidades de inclusión conjunta π_{kl}, se definen de la manera siguiente:

$$\pi_{kl} = \begin{cases} \pi_{kl} = \pi_{li}\pi_{k|i} & \text{si } k = l \in U_i \\ \pi_{li}\pi_{kl|i} & \text{si } k \,\&\, l \in U_i, k \neq l \\ \pi_{lij}\pi_{k|i}\pi_{l|j} & \text{si } k \in U_i \text{ y } l \in U_j, i \neq j \end{cases}$$

Además, para un muestreo por conglomerados estratificado bietápico: $\Delta_{kl} = 0$, para k y 1 que pertenecen a diferentes estratos, ya que $\pi_{kl} = \pi_k\pi_l$.

[2]Este estimador también es conocido como estimador de Horvitz- Thompson.

Entre otras alternativas (ver: Skinner e**t al**., 1989) para estimar la matriz de covarianza del vector (11), puede aplicarse un procedimento basado en el método de muestreo llamado **esquema A**, (Raj, 1984).

• Estimador de la matriz de covarianza bajo el esquema A de Raj

Para lo que a continuación será presentado, es conveniente tener expresiones equivalentes para la ecuación (11); en las que implícitamente, se supone un esquema de muestreo estratificado bietápico.

Considere que una muestra s, de n elementos, es seleccionada usando un muestreo por conglomerados estratificado en dos etapas, tal que en la primera etapa, m_h conglomerados son extraídos (sin reemplazo y con probabilidades iguales) de cada uno de los h = 1,2,...,H estratos, con M_h UPM en el estrato correspondiente. En la segunda etapa, una muestra de n_{hd} elementos es seleccionada[3] a partir del d-ésimo (d = 1,2,...,m_h) conglomerado muestreado en el estrato h, donde $\sum_{h=1}^{H}\sum_{d=1}^{m_h} n_{hd} = n.$ De acuerdo con esto, sea

$$\sum_{h=1}^{H}\sum_{d=1}^{m_h} w_{1hd} = \sum_{h=1}^{H}\sum_{d=1}^{m_h}\sum_{c=1}^{n_{hd}} w_{1hdc} = \hat{t} \qquad \text{y} \qquad \sum_{h=1}^{H}\sum_{d=1}^{m_h} w_{2hd} = \sum_{h=1}^{H}\sum_{d=1}^{m_h}\sum_{c=1}^{n_{hd}} w_{2hdc} = \hat{T}$$

donde, el vector $w_{1hd} = \sum_{c=1}^{n_{hd}} w_{1hdc}$ y $w_{2hd} = \sum_{c=1}^{n_{hd}} w_{2hdc}$, siendo w_{ihdc} el valor asociado con la c-ésima unidad (c = 1,2,...,n_{hd}) del d-ésimo conglomerado muestreado en el estrato h, para i = 1,2; es decir, $w_{1hdc} = x_k y_k/\pi_k$ y $w_{2hdc} = x_k x_k'/\pi_k, \forall_{k\in s}$.

Entonces, bajo el esquema A y haciendo uso del método de linealización, el estimador de la matriz de covarianza $\hat{V}_A(\cdot)$ del vector $\hat{\beta}_\pi$, está dado por

$$\hat{V}_A(\hat{\beta}_\pi) = \hat{T}^{-1}\hat{\Sigma}_A\hat{T}^{-1} \tag{13}$$

donde

$$\hat{\Sigma}_A = \sum_{h=1}^{H}\left(1-\frac{m_h}{M_h}\right)\left(\frac{m_h}{m_h-1}\right)\sum_{d=1}^{m_h}(w_{hd}-\overline{w}_h)(w_{hd}-\overline{w}_h)'$$

siendo, $w_{hd} = w_{1hd} - w_{2hd}\hat{\beta}_\pi$ y $\overline{w}_h = \sum_{d=1}^{m_h} w_{hd}/m_h$; $w_{1hd} = \sum_{c=1}^{n_{hd}} w_{1hdc}$ y $w_{2hd} = \sum_{c=1}^{n_{hd}} w_{2hdc}$, tal y como anteriormente fueron definidos.

Hemos mencionado que los procedimientos tradicionales presentan ciertas desventajas, en la etapa de estimación de la matriz de covarianza, al ignorar las complejidades inherentes al diseño de muestreo; ocasionando distorsiones en las inferencias. Como posibles alternativas, se han presentado dos métodos que sí contemplan e incorporan, en el proceso de estimación, el diseño complejo de la muestra. Desafortunadamente, estos métodos de estimación no están disponibles en los paquetes computacionales

[3]Cada UPM seleccionada es submuestreada de una manera conocida, cualquiera que sea el número de etapas del submuestreo.

convencionales diseñados para el análisis estadístico. Por otra parte, y con el propósito de llevar a cabo una *Simulación*, programar la matriz $\hat{\Sigma}_{\pi}$ no es trivial. Para estos mismos fines, la matriz $\hat{\Sigma}_{A}$ es mucho más noble para programar.

Comúnmente el usuario, consciente o inconscientemente, recurre a los paquetes convencionales para llevar a cabo un análisis estadístico dado, obteniendo -en la mayoría de los casos- resultados erróneos, (sobre la base de los argumentos antes mencionados). Por esta razón, surge la idea de efectuar ciertas correcciones a los errores estándar del vector $\hat{\beta}_{(\cdot)}$ (calculados a partir del paquete computacional estadístico elegido) que puedan compensar las deficiencias provocadas por ignorar el diseño de muestreo. Estos ajustes a los errrores estándar del vector $\hat{\beta}_{(\cdot)}$, se lleva a cabo usando el efecto del diseño, (ver: Kish, 1965). Para un estimador $\hat{\theta}$ del parámetro θ, el efecto del diseño denotado por DEFF, se define como

$$\mathrm{DEFF}_{p(s)}(\hat{\theta}) = \frac{V_{p(s)}(\hat{\theta})}{V_{SRS}(\hat{\theta})} \tag{14}$$

en donde p(.) se refiere al diseño de muestreo complejo que se asume; mientras que, SRS se refiere a un esquema de muestreo aletorio simple. Es muy claro que, para obtener un DEEF se requieren los valores de ambas varianzas del diseño, los cuales están escasamente disponibles en la práctica. Sin embargo, es posible calcular una estimación del efecto de un diseño complejo usando los correspondientes estimadores de la varianzas con el conjunto de datos muestrales. Entonces, un estimador para el DEFF es,

$$\mathrm{deff}_{p(s)}(\hat{\theta}) = \frac{\hat{V}_{p(s)}(\hat{\theta})}{V_{SRS}(\hat{\theta})} \tag{15}$$

Es conveniente hacer notar que, la comparación de las dos varianzas del diseño se puede llevar a cabo usando -no necesariamente- el mismo estimador $\hat{\theta}$. En tal caso, el DEFF puede ser definido por una estrategia, es decir, una combinación de un diseño de muestreo p(.) y un estimador especificado $\hat{\theta}^{*}$:

$$\mathrm{DEFF}_{p(s)}(\hat{\theta}^{*}) = \frac{\hat{V}_{p(s)}(\hat{\theta}^{*})}{V_{SRS}(\hat{\theta})} \tag{16}$$

donde $\hat{\theta}^{*}$ es un estimador de θ bajo p(s) y $\hat{\theta}$ es el estimador de θ asumiendo un diseño SRS; $V_{p(s)}$ y V_{SRS} son las varianzas respectivas. Recordando la regla, se sabe que un diseño de muestreo complejo es igualmente tan eficiente como el esquema SRS si el DEFF es igual al valor uno, más eficiente si el DEFF es menor que el valor uno y menos eficiente si el DEFF es mayor que el valor uno.

Para los fines que nos hemos propuesto, suponemos estimadores de razón ponderados para el parámetro θ que corresponde a la media $\overline{Y}$ de la variable repuesta o a $\overline{x}$, la media del vector de predictores. En el caso de la variable respuesta, el estimador de razón ponderado $\overline{y}_{w}$ para el parámetro de la media $\overline{Y}$ se define como

$$\overline{y}_{w} = \frac{\hat{t}_{y}}{\hat{n}} \tag{17}$$

donde, $\hat{t}_y = \sum_{h=1}^{H}\sum_{d=1}^{m_h} y_{hd} = \sum_{h=1}^{H}\sum_{d=1}^{m_h}\sum_{c=1}^{n_{hd}} w_{hdc} y_{hdc}$ es la suma muestral ponderada de la variable respuesta y para los elementos maestrales que pertenecen a s = {1,...,n} , y $\hat{n} = \sum_{h=1}^{H}\sum_{d=1}^{m_h}\hat{n}_{hd} = \sum_{h=1}^{H}\sum_{d=1}^{m_h}\sum_{c=1}^{n_{hd}} w_{hdc}$ es el tamaño muestral ponderado. Las ponderaciones w_{hdc} son los pesos relativos[4] de los elementos en la muestra s; en Lehtonen **et al**., (1995), se discuten en detalle diferentes enfoques en los que las observaciones pueden ser ponderadas. La varianza estimada de $\bar{y}_w, \hat{V}_A(\bar{y}_w)$, se obtiene sobre la base del método de linearización y bajo el esquema A de Raj, es decir, en una versión equivalente a la ecuación (13); luego, el estimador de la varianza del promedio de la variable respuesta se puede expresar como

$$\hat{V}_A(\bar{y}_w) = \sum_{h=1}^{H}\left(1-\frac{m_h}{M_h}\right)\left(\frac{m_h}{m_h-1}\right)\sum_{d=1}^{m_h}(u_{hd}-\bar{u}_h)^2 \qquad (18)$$

donde $u_{hd} = (y_{hd} - \hat{n}_{hd}\bar{y}_w)/\hat{n}$ y $\bar{u}_h = \sum_{d=1}^{m_h} u_{hd}/m_h$.

En lo que corresponde al vector de predictores, y de manera similar a la ecuación (17), el estimador de razón ponderado $\hat{\bar{x}}$ del vector paramétrico de la media $\bar{x}$, está dado por

$$\hat{\bar{x}} = \frac{\hat{t}_x}{\hat{n}} \qquad (19)$$

donde, $\hat{t}_x = \sum_{h=1}^{H}\sum_{d=1}^{m_h} x_{hd} = \sum_{h=1}^{H}\sum_{d=1}^{m_h}\sum_{c=1}^{n_{hd}} w_{hdc} x_{hdc}$ es la suma muestral ponderada del vector de predictores x para los elementos muestrales que pertenecen a s = {1,...,n}. En cuanto a la matriz de covarianza estimada del vector $\hat{\bar{x}}, \hat{V}_A(\hat{\bar{x}})$, se obtiene de manera similar a la ecuación (18); luego, el estimador de la matriz de covarianza del vector promedio de predictores, es

$$\hat{V}_A(\hat{\bar{x}}) = \sum_{h=1}^{H}\left(1-\frac{m_h}{M_h}\right)\left(\frac{m_h}{m_h-1}\right)\sum_{d=1}^{m_h}(v_{hd}-\bar{v}_h)(v_{hd}-\bar{v}_h)' \qquad (20)$$

donde $v_{hd} = (x_{hd} - \hat{n}_{hd}\hat{\bar{x}})/\hat{n}$ y $\bar{v}_h = \sum_{d=1}^{m_h} v_{hd}/m_h$.

La intención de haber presentado estos últimos resultados, se justifica por el hecho de ser el entorno sobre el que giran las correcciones que se proponen seguidamente, con el afán de ajustar la precisión de las estimaciones que se obtienen en un paquete computacional estadístico de uso común.

---

[4]El peso relativo del elemento $k \in s$, se define como $w_k = (n/N)(1/\pi_k)$, luego $\sum_{k\in s} w_k = (n/N)\hat{N} = \hat{n}$; siendo $\hat{N} = \sum_{k\in s} 1/\ _k$, el tamaño estimado de la población N.

• **Corrección C_1 deff constante**

Considerando que el efecto del diseño del estimador $\bar{y}_w$, se puede expresar como $\text{deff}(\bar{y}_w) = \hat{V}_A(\bar{y}_w)/\hat{V}_{SRS}(\bar{y})$, siendo $\bar{y} = \sum_{k\in s} y_k/n$, la media aritmética. Para realizar el ajuste C_1, se sugiere: multiplicar la desviación estándar de cada coeficiente de regresión por el *factor del diseño*[5] del estimador $\bar{y}_w$; esto es

$$C_1 = \text{deft}(\bar{y}_w)\sqrt{\hat{V}_{pc}(\beta_j)}, \quad j = 1,2,...,q. \tag{21}$$

Las componentes β_j pueden ser los elementos del vector $\hat{\beta}$ o del vector $\hat{\beta}_W$, según sea el caso. Aunque el estimador $\hat{\beta}_W$ es más deseable, ya que al menos incorpora una de las complejidades del diseño de muestreo, la ponderación de las observaciones. Mientras que, el vector $\hat{\beta}$ ignora totalmente la estructura compleja de la población.

• **Corrección C2: deff específico del coeficiente**

Ahora, considere que el efecto del diseño de cada una de las componentes del vector estimador $\hat{\bar{x}} = (\hat{\bar{x}}_1,\ldots,\hat{\bar{x}}_j,\ldots,\hat{\bar{x}}_q)'$, está dado por $\text{deff}(\hat{\bar{x}}_j) = \hat{V}_A(\hat{\bar{x}}_j)/\hat{V}_{SRS}(\bar{x}_j)$; donde los estimadores de las varianzas de las $(\hat{\bar{x}}_j)$ son los respectivos elementos diagonales de la matriz de covarianza estimada $\hat{V}_A(\hat{\bar{x}})$, y $\bar{x}_j$, es la contra- parte de $(\hat{\bar{x}}_j)$, en un esquema de muestreo aleatorio simple. Para llevar a cabo la corrección C_2, y de manera parecida a la ecuación (21), se sugiere: multiplicar la desviación estándar de cada coeficiente de regresión por el factor del diseño del estimador $(\hat{\bar{x}}_j)$ correspondiente; es decir

$$C_2 = \text{deft}(\hat{\bar{x}}_j)\sqrt{\hat{V}_{pc}(\beta_j)}, \quad j = 1,2,...,q. \tag{22}$$

Cuando el primer elemento del vector x_i (i = 1,...,n) es igual a 1, el $\text{deft}(\hat{\bar{x}}_j)$ no está definido, ya que $\hat{V}_{SRS}(\bar{x}_j) = 0$.

Diseño de la Muestra

Para nuestro caso, el marco de muestreo que se ha considerado, basado en datos del XI Censo General de Población y Vivienda 1990, está conformado por 10981 AGEB'S Rurales[6] (N = 10981). Estos AGEB'S han

[5]El factor del diseño: $\text{deft}(\cdot) = \sqrt{\text{deff}(\cdot)}$, es el factor de inflación apropiado para los errores estándar e intervalos de confianza. [Esta cantidad es extensamente usada (ver: Verma **et al**., 1980)].

[6] Area Geoestadística Básica (AGEB): Espacio geográfico delimitado mediante rasgos naturales o culturales, permanentes y reconocibles en el terreno, con extensión convencional al interior de cada municipio del país o delegación del Distrito Federal que facilita la captación y referenciación geográfica de la información. AGEB Rural: Es la AGEB que delimita una superficie donde no existen localidades de 2500 habitantes o más, ni cabeceras municipales.

sido divididos en cinco estratos: Región Noreste (estrato 1), Región Noroeste (estrato 2), Región Centro (estrato 3), Región Sureste (estrato 4), Región Suroeste (estrato 5). En los estratos 1,...,5 los 7 AGEB'S se agruparon en 191, 175, 164, 147, 103 conglomerados[7] (UPM), respectivamente. En cada estrato se seleccionaron 50 conglomerados mediante el esquema de muestreo A de Raj; y en cada UPM seleccionada, son extraídos aleatoriamente 5 AGEB'S con probabilidades iguales y sin reemplazo, con lo que el tamaño de la muestra es n = 1250.

El Modelo

De acuerdo con los propósitos de este trabajo, suponga que nuestro objetivo es modelar la variación sistemática de la variable respuesta y: = *alfabetismo*, en función de las variables predictoras o explicativas x_1: = *escolaridad*, x_2:= *dependientes económicos*, x_3:= *ingresos menores al salario mínimo*, x_4: = disponibilidad *de electricidad* y x_5 := *hacinamiento*, en donde se supone el modelo de regresión lineal para los datos muestrales (y, X) , dado por la ecuación (3):

$$y = X\beta + \varepsilon$$

donde, $E(\varepsilon|X) = 0$, $V(\varepsilon|X) = \sigma^2\Omega$; $E(y|X) = X\beta$ y $V(y|X) = \sigma^2\Omega$.

Los vectores de coeficientes de regresión y sus correspondientes matrices de covarianza, son estimados por los diferentes métodos antes descritos.

La Simulación

Para llevar a cabo el estudio de simulación, un programa (en S-plus) fue escrito para que ejecutara las siguientes rutinas:

- En la primera etapa, y a partir de los datos del XI Censo General de Población y Vivienda 1990, una muestra de 50 conglomerados (UPM) es extraída de cada uno de los 5 estratos antes descritos.

- En la segunda etapa, de cada UPM que aparece en la muestra, 5 AGEB'S son seleccionados. Con lo cual, el tamaño de la muestra considerada es n = 1250. Este procedimiento es ejecutado 1000 veces.

- Finalmente, con las 1000 muestras obtenidas mediante el proceso anterior, fueron calculados empíricamente: la media del vector $\hat{\beta}$ (OLS y WLS), el error cuadrático medio del vector $\hat{\beta}$ (OLS y WLS), el sesgo del vector $\hat{\beta}$ (OLS y WLS); además, la media de los errores estándar del vector $\hat{\beta}$ bajo los métodos: OLS, WLS, Esquema A de Raj (E.A. Raj), corrección C_1 y corrección C_2; el error estándar de los errores estándar del vector $\hat{\beta}$ (OLS, WLS, E.A.RAJ, C_1 y C_2), el error cuadrático medio de los errores estándar del vector $\hat{\beta}$ (OLS, WLS, C_1 y C_2), y por último, el sesgo de los errores estándar del vector $\hat{\beta}$ (OLS, WLS, C_1 y C_2).

RESULTADOS

[7]Estos conglomerados corresponden a los 780 municipios que han sido considerados para nuestro estudio.

A continuación se presentan los resultados obtenidos mediante el proceso de simulación, donde se incluyen las estimaciones de los errores estándar del vector $\hat{\beta}$ arrojadas por los diferentes métodos anteriormente descritos. Además, con la finalidad de hacer comparaciones, se menciona el valor **a nivel poblacional** del vector β obtenido, se utilizó el paquete computacional S-plus.

En la Tabla 1, aparecen los resultados relacionados con la media, el error cuadrático medio, así como el sesgo del vector $\hat{\beta}$, calculados empíricamente bajo los métodos de estimación OLS y WLS. También aparece el valor poblacional del vector β.

Tabla 1. Comparación entre los métodos de cuadrados mínimos ordinarios (OLS) y cuadrados mínimos ponderados (WLS).

		MEDIA DEL VECTOR $\hat{}$		**ERROR CUADRATICO MEDIO DEL VECTOR $\hat{}$**		SESGO DEL VECTOR $\hat{}$	
Coeficiente	**Valor de B**	**OLS**	**WLS**	**OLS**	**WLS**	**OLS**	WLS
Intercepto	52,316	56,3795105	58,4219546	25,9027451	56,6811659	4,063998	**6,10641209**
X1	8,338	9,01715116	9,25209578	0,7080691	1,28324502	0,67522821	**0,91373902**
X2	0,314	0,6483784	0,78359471	0,17459268	0,35178867	0,33405426	**0,48927057**
X3	0,007	0,02558209	0,03197909	0,00050943	0,00096259	0,01811253	**0,02450653**
X4	-0,006	0,00506498	0,00996985	0,00017846	0,0003848	0,010686	**0,01559188**
X5	**-3,924**	**-2,6766859**	**-2,7351566**	**2,2171191**	**2,19497127**	**1,24711925**	**1,18863758**

En cuanto a la media del vector $\hat{\beta}$, así como en el sesgo, es posible observar (a partir de los resultados de la Tabla 1), diferencias relativamente pequeñas entre los métodos en cuestión. Sin embargo, las diferencias entre los métodos son más notorias en el error cuadrático medio, principalmente en el coeficiente $\hat{\beta}_0$ (INTERCEPTO). De acuerdo con estos resultados, se puede apreciar menos precisión en el estimador WLS respecto del estimador OLS. Esto último puede ser por el hecho de que el estimador OLS subestima los errores estándar, lo cual se ve en los sesgos negativos grandes que se muestran en la Tabla 5.

En la Tabla 2, se presentan los resultados que contemplan la media de los errores estándar del vector $\hat{\beta}$, calculados empíricamente bajo los métodos de estimación OLS, WLS, E.A.RAJ, así como las correcciones C_1 y C_2. Tanto en el E.A.RAJ y en las correcciones C_1 y C_2, se adopta a $\hat{\beta}_W$ como el vector estimador de β.

En estos últimos resultados (Tabla 2), los métodos OLS y WLS presentan entre ellos diferencias relativamente pequeñas. Un comportamiento similar se observa entre las correcciones C_1 y C_2. Sin embargo, bajo el esquema A de Raj, los errores estándar de los coeficientes $\hat{\beta}_0$ y $\hat{\beta}_1$ presentan valores más grandes que los otros métodos; aunque también es cierto que, el resto de los coeficientes en E.A.RAJ son aproximadamente iguales a sus contrapartes en la corrección C_1. La inflación de los errores estándar tanto en el E.A.RAJ como en las correcciones C_1 y C_2, es debido a los efectos de conglomerado, es decir, a la correlación positiva intra-conglomerado de la variable respuesta.

Tabla 2. Comparación entre los métodos de cuadrados mínimos ordinarios (OLS),

cuadrados mínimos ponderados (WLS), esquema A de Raj, correcciones C_1 y C_2.

	MEDIA DE LOS ERRORES ESTANDAR DEL VECTOR $\hat{}$				
Coeficiente	**OLS**	**WLS**	**E.A.RAJ**	**C_1**	C_2
Intercepto	1,82937793	1,76612603	5,83303367	3,46020471	**inf**
X1	0,29318323	0,28639698	0,90900608	0,56092619	**0,49190653**
X2	0,25387484	0,28397021	0,48082454	0,55663843	**0,53886015**
X3	0,01272965	0,01322101	0,02588654	0,02592076	**0,02850116**
X4	0,00927886	0,00972149	0,0169925	0,0190467	**0,01710562**
X5	**0,49454117**	**0,55172503**	**1,15041041**	**1,08104442**	**0,90786698**

En lo que corresponde a la tabla siguiente (Tabla 3), se presentan los resultados relacionados con el error estándar de los errores estándar del vector $\hat{\beta}$, bajo los métodos de estimación OLS, WLS, E.A.RAJ, así como las correcciones C_1 y C_2 .

Tabla 3. Comparación entre los métodos de cuadrados mínimos ordinarios (OLS), cuadrados mínimos ponderados (WLS), esquema A de Raj, correcciones C_1 y C_2 .

	ERROR ESTANDAR DE LOS ERRORES ESTANDAR DEL VECTOR $\hat{}$				
Coeficiente	**OLS**	**WLS**	**E.A.RAJ**	**C_1**	C_2
Intercepto	0,00765686	0,01059501	0,13548264	0,04948035	**NA**
X1	0,00119719	0,00167909	0,01976401	0,00785764	**0,00740899**
X2	0,00163574	0,00242401	0,00952067	0,00886517	**0,00897295**
X3	6,6943E-05	9,3531E-05	0,00058805	0,00039018	**0,00041594**
X4	4,7177E-05	6,8926E-05	0,00036853	0,00028866	**0,00023217**
X5	**0,00310979**	**0,00474805**	**0,02400889**	**0,01692129**	**0,01507842**

Los resultados que aparecen en la Tabla 3, muestran que los errores estándar de los errores estándar del vector $\hat{\beta}$, en las correcciones C_1 y C_2 son prácticamente iguales, y no muy alejados de los valores obtenidos mediante el E.A.RAJ. Mientras que los métodos de estimación OLS (que ignora todas las complejidades del diseño) y WLS (que incorpora solamente la ponderación de las observaciones) presentan valores inferiores a los otros métodos.

En la Tabla 4, se muestran los resultados obtenidos al calcular empíricamente el error cuadrático medio de los errores estándar del vector $\hat{\beta}$, bajo los métodos de estimación OLS, WLS, y las correcciones C_1 y C_2. Además, para obtener los resultados, tanto de la Tabla 4 así como los de la Tabla 5, se supone conocido el "verdadero" valor de los errores estándar del vector $\hat{\beta}$. Para esto, se tomó el valor promedio (vía simulación)

de los errores estándar del vector $\hat{\beta}$ bajo el método conocido como E.A.RAJ. Por esta razón, en dichas tablas no aparece el método en cuestión.

De acuerdo con los resultados que aparecen en la Tabla 4, es posible advertir una mayor exactitud en las estimaciones de los errores estándar del vector $\hat{\beta}$, bajo las correcciones C_1 y C_2 sobre la base de los métodos OLS y WLS. Por lo cual, ajustar los errores estándar (que se obtienen en algunos paquetes estadísticos computacionales convencionales) mediante dichas correcciones, garantizan mayor precisión en las estimaciones.

Tabla 4. Comparación entre los métodos de cuadrados mínimos ordinarios (OLS), cuadrados mínimos ponderados (WLS) , correcciones C_1 y C_2.

	ERROR CUADRATICO MEDIO DE LOS ERRORES ESTANDAR DEL VECTOR $\hat{}$			
Coeficiente	**OLS**	**WLS**	**C_1**	C_2
Intercepto	1,61E+01	1,66E+01	5,88E+00	**inf**
X1	3,80E-01	3,88E-01	1,27E-01	**1,80E-01**
X2	5,18E-02	3,94E-02	1,36E-02	**1,13E-02**
X3	1,7 4E-04	1,61 E-04	1,53E-05	**2,42E-05**
X4	6,01 E-05	5,37E-05	1,26E-05	**5,48E-06**
X5	**4,32E-01**	**3,61E-01**	**3,38E-02**	**8,18E-02**

En la Tabla 5, que a continuación se presenta, aparecen los resultados relacionados con el sesgo de los errores estándar del vector $\hat{\beta}$, bajo los métodos de estimación OLS, WLS, y las correcciones C_1 y C_2 .

Tabla 5. Comparación entre los métodos de cuadrados mínimos ordinarios (OLS), cuadrados mínimos ponderados (WLS), correcciones C_1 y C_2.

.

	SESGO DE LOS ERRORES ESTANDAR DEL VECTOR $\hat{}$			
Coeficiente	**OLS**	**WLS**	**C_1**	C_2
Intercepto	-4,0036677	-4,0669176	-2,3728565	**inf**
X1	-0,6158339	-0,6226181	-0,3480793	**-0,4171151**
X2	-0,2269507	-0,1968523	0,07581639	**0,05803419**
X3	-0,0131529	-0,0126615	2,672E-05	**0,00261654**
X4	-0,0077276	-0,007285	0,00204871	**0,00010338**
X5	**-0,6558632**	**-0,5986794**	**-0,0693574**	**-0,2425283**

Los resultados que aparecen en la Tabla 5, muestran que los valores del sesgo de los errores estándar del vector $\hat{\beta}$ son menores cuando se hace el ajuste a dichos errores a través de C_1 o C_2, respecto a los valores obtenidos bajo OLS o WLS. Es decir, al usar las correcciones C_1 o C_2, disminuye el sesgo que se obtiene bajo OLS o WLS. Esto, junto con los resultados de la Tabla 4, rectifica -de manera empírica- el aumento de precisión en las estimaciones, ajustando a través de C_1 o C_2, los errores estándar del vector $\hat{\beta}$ que pueden ser obtenidos a partir de un paquete estadístico computacional de uso común. Además, la baja precisión obtenida mediante los métodos OLS y WLS que se observa a partir de la Tablas 4 y 5, se manifiesta al no considerar en dichos métodos el efecto de conglomeración.

CONCLUSIONES

Los datos de encuestas son frecuentemente el resultado de un diseño de muestreo por conglomerados polietápico estratificado. Ignorar estas complejidades del diseño en la etapa de estimación, puede provocar graves distorsiones en la inferencias de tipo analítico. Por esta razón, el análisis de datos que provienen de

encuestas complejas, podría ser inadecuado si se lleva a cabo usando procedimientos convencionales, tales como aquellos que se suponen en la teoría estadística clásica y en una gran variedad de paquetes computacionales diseñados para el análisis estadístico. Sobre la base de esto, y de acuerdo con nuestros resultados obtenidos vía simulación, se sugiere efectuar ajustes a los errores estándar (que se obtienen en algún paquete computacional), mediante simples correcciones tales como C_1 o C_2; ya que éstas correcciones, incrementan la exactitud de las estimaciones de los errores estándar del vector $\hat{\beta}$ en una encuesta compleja. Además, se puede apreciar que dichas correcciones disminuyen el sesgo, de los errores estándar del vector $\hat{\beta}$, que se obtendría en un paquete estadístico computacional de uso común.

REFERENCIAS

ARMINGER, G.; C.C. CLOGG and M.E. SABAL (1995): **Handbook of Statistical Modeling for the Social and Behavioral Sciences**, New York, Plenum Press.

DEMING, W.E. (1950): **Some Theory of Sampling**, New York: Dover

GOOD, P. (1999): "Resampling Methods", **A practical guide to data analysis**. Boston: Birhauser.

LEHTONEN, R. and E.J. PAHKINEN (1995): **Practical Methods for Design and Analysis of Complex Surveys**, Chichester, Wiley.

NATHAN, G.and T.M.F. SMITH (1989): "The effect of selection on regression analysis", In: Skinner C.J., Holt D. and Smith T.M.F. (eds) **Analysis of ComplexSurveys**. Chichester: Wiley, 149-163.

PFEFFERMANN (1993): "The role of sampling weights when modeling survey data", **International Statistical Review** 61, 317-337.

RAJ, D. (1984): **Sampling Theory**, New York: McGraw-Hill, Inc.

SARNDAL,C.E.; B. SWENSSON and J. WRETMAN (1992): **Model Assited Survey Sampling**, New York: Springer.

SKINNER, C.J.; D. HOLT and T.M.F. SMITH (1989): **Analysis of Complex Surveys**, Chichester: Wiley.

VERMA, V.; C. SCOTT and C. O'MUIRCHEARTAIGH (1980): "Sample designs and sampling errors for fue World Fertility Survey", **Journal of the Royal Statistical Society A**, 143, 431-473.