Servy, Elsa
Marí, Gonzalo
Hachuel, Leticia
Instituto de Investigaciones Teóricas y Aplicadas en Estadística, Escuela de Estadística

CONSIDERACIÓN DEL MUESTREO COMPLEJO EN EL ANÁLISIS DE DATOS DE LA EPH[1]

1. INTRODUCCIÓN

Una práctica común en el análisis de datos provenientes de encuestas por muestreo es considerar que los datos provienen de un muestreo simple al azar, o sea, observaciones independientes e igualmente distribuidas (iid). En muchos estudios este supuesto no se verifica en parte debido a la utilización de estratos, conglomerados y distintas probabilidades de selección. La utilización de los métodos clásicos bajo estos diseños muestrales puede conducir a inferencias erróneas. En general, los programas utilizados para llevar a cabo estos análisis no consideran la complejidad del diseño.

Este trabajo tiene como objetivo realizar la estimación de regresiones logísticas con datos provenientes de la Encuesta Permanente de Hogares (EPH) llevada a cabo por el Instituto Nacional de Estadística y Censos, así como el cálculo de matrices de variancias y covariancias considerando el diseño complejo de la Encuesta a fin de evaluar las posibles consecuencias de ignorar en dichos procedimientos el diseño muestral.

En una primera parte se compararon las estimaciones de las matrices de variancias y covariancias de un vector de variables categóricas suponiendo que el método de selección de la EPH fuese

i) completamente al azar,

ii) estratificado con Áreas como unidades primarias, y

iii) estratificado con Hogares como unidades primarias.

En la segunda etapa se analizó el comportamiento de los desvíos estándares de las estimaciones de los parámetros de las regresiones logísticas mediante la utilización de métodos de estimación de variancias que tienen en cuenta el diseño complejo, en particular, el método de linearización de Taylor y el método de replicación Jackknife. Se describe un programa ad-hoc desarrollado por el Research Triangle Institute, el SUDAAN, que permite incorporar las modificaciones necesarias para considerar el diseño muestral complejo.

[1] Trabajo realizado en el marco del Proyecto de Investigación y Desarrollo “Estudio de datos categóricos medidos a través del tiempo”. Secretaría de Ciencia y Tecnología de la UNR.

2. DISEÑO MUESTRAL DE LA EPH

La Encuesta Permanente de Hogares es llevada a cabo por el INDEC en 29 aglomerados, formados por los aglomerados urbanos de más de 100.000 habitantes, o menos pero que son capitales de provincias (con la excepción de San Nicolás), el aglomerado urbano-rural Alto Valle del Río Negro, y el aglomerado Gran Buenos Aires. La misma se lleva a cabo dos veces en el año, la primer onda en el mes de Mayo y la segunda en el mes de Octubre.

En cada uno de los dominios se selecciona una muestra siguiendo un diseño muestral probabilístico en dos etapas de selección con las unidades de primera etapa estratificadas, utilizando como variable de estratificación el porcentaje de jefes de hogar con educación primaria incompleta.

En la primera etapa, y dentro de cada estrato, se seleccionan radios censales que en la Encuesta se los denomina Áreas (UPM: Unidades Primarias de Muestreo) con probabilidad proporcional al tamaño medido en cantidad de viviendas ocupadas. Para la segunda etapa, se listan las viviendas particulares de las UPM seleccionadas. Sobre este listado, se seleccionan viviendas en forma sistemática, de manera tal de obtener una muestra auto ponderada, o sea, cada vivienda tiene igual probabilidad de selección.

La EPH utiliza un esquema de panel donde la rotación es tal que entre 2 ondas permanece el 75% de la muestra, o sea, cuando una vivienda es seleccionada, permanece en la muestra durante cuatro ondas, o sea, dos años.

Este estudio se realizó sobre dos aglomerados de distintos tamaños para medir el impacto del diseño muestral. El primero de ellos es el Gran Buenos Aires, llamado Aglomerado 1, que en la onda de mayo de 1998 estuvo compuesto por 11807 personas que constituían 3549 hogares, ocupando 3433 viviendas en 518 áreas estratificadas en 12 estratos. El segundo de ellos es un aglomerado de menor tamaño, llamado Aglomerado 2, y que está compuesto en la misma onda por 2212 personas que constituyen 624 hogares, que ocupan 614 viviendas en 48 áreas en 5 estratos.

3. MATRIZ DE VARIANCIAS Y COVARIANCIAS BAJO MUESTRAS COMPLEJAS

Una de las variables de mayor interés de la EPH es la variable Estado Ocupacional, la cual permite construir, entre otras, la tasa de desocupación. Dicha variable es categórica tomando los valores Ocupado, Desocupado e Inactivo, los cuales se pueden codificar con 1, 2, y 3 respectivamente.

Bajo el supuesto que los datos sean iid, o sea, que provengan de una muestra simple al azar, esta variable posee una distribución multinomial con vector de esperanzas $E(X)=(p_1,p_2,p_3)'$ y matriz de variancias y covariancias

$$Var(X)=\begin{bmatrix} \frac{p_1(1-p_1)}{n} & -\frac{p_1p_2}{n} & -\frac{p_1p_3}{n} \\ & \frac{p_2(1-p_2)}{n} & -\frac{p_2p_3}{n} \\ & & \frac{p_3(1-p_3)}{n} \end{bmatrix}$$

Cuando el diseño muestral es complejo, particularmente si pertenece a aquéllos que involucran conglomerados, las observaciones se apartan del supuesto de independencia y

de igualdad de distribución, lo que lleva implícito que la variable de interés no pueda ser considerada proveniente de una distribución multinomial.

Una de las metodologías para el cálculo de la matriz de variancias y covariancias que puede ser extendida a diseños muestrales complejos es la basada en los métodos de linearización por Series de Taylor. Este método es recomendado por su cualidad de robustez cuando no se cumplen los supuestos distribucionales clásicos.

Sea (hik) el elemento de la muestra que representa al k-ésimo individuo de la i-ésima unidad primaria del h-ésimo estrato (k=1,...,n_{hi}; i=1,...,n_h; h=1,...,L), y sean $y_{t(hik)}$ la variable que es igual a 1 si el elemento (hik) pertenece a la categoría *t* y 0 en otro caso, y w_{hik} es el peso muestral del elemento (hik) ajustado por no-respuesta. El parámetro p_t, con t=1,2,3, es estimado por el estadístico de razones combinado $\hat{p}_t = \hat{N}_t / \hat{N}$, donde $\hat{N}_t = \sum_{(hik)\in s} w_{hik} y_{t(hik)}$ es la estimación del total de individuos pertenecientes a la categoría *t*, $\hat{N} = \sum_{(hik)\in s} w_{hik}$ es la estimación del total de individuos en la población, y *s* representa el conjunto de individuos de la muestra.

Las estimaciones de las variancias y covariancias se calculan bajo el supuesto que las unidades de la primera etapa dentro de cada uno de los estratos estén seleccionadas con reposición, si bien este supuesto (que no se verifica en la EPH) conlleva una sobreestimación de las variancias estimadas. La covariancia de $\hat{p}_t$ y $\hat{p}_u$ se estima por

$$cov\,est(\hat{p}_t, \hat{p}_u) = \hat{N}^{-2} \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (z_{thi} - \bar{z}_{th})(z_{uhi} - \bar{z}_{uh})$$

donde $z_{thi} = \sum_k w_{hik} y_{t(hik)} - \hat{p}_t \sum_k w_{hik}$ y $\bar{z}_{th} = \sum_{i=1}^{n_h} z_{thi} / n_h$.

Una medida utilizada para evaluar el comportamiento de una estrategia compuesta por un diseño muestral y un estimador $\hat{\theta}$, es el efecto de diseño (deff). El mismo se define para el caso unidimensional como el cociente entre dos variancias: la obtenida considerando el diseño complejo y aquella que supone que la muestra es simple al azar (MSA). Un valor del deff mayor a uno indica una pérdida en la precisión por no utilizar MSA, un valor inferior a la unidad revela una ganancia en la precisión, mientras que un valor igual a uno indica que no existen diferencias en cuanto a la precisión entre los dos diseños considerados.

La matriz de efectos de diseño (deff) es una generalización de la medida calculada para un parámetro escalar. Sea $\hat{\boldsymbol{\theta}}$ un vector *p x 1* y V_0 un estimador de la matriz de covariancias de $\hat{\boldsymbol{\theta}}$ basado en los supuestos clásicos (iid). Definimos $V_{VERD}(\hat{\boldsymbol{\theta}})$ al estimador de la matriz de covariancias que tiene en cuenta el diseño complejo. Luego la matriz de efectos de diseño viene dada por

$$\mathbf{deff}(\hat{\boldsymbol{\theta}}, \mathbf{V_0}) = \boldsymbol{\Delta} = E_{verd}(\mathbf{V_0})^{-1} \mathbf{Var_{verd}}(\hat{\boldsymbol{\theta}})$$

Los autovalores de la matriz de efectos de diseño: $\delta_1 \geq ... \geq \delta_p$ se denominan efectos de diseño generalizados y poseen la propiedad de que δ_1 y δ_p definen los límites para los deffs univariados de cualquier combinación $\mathbf{c}'\hat{\boldsymbol{\theta}}$ de $\hat{\boldsymbol{\theta}}$. Esto es:

$$\delta_1 = \max_{\mathbf{c}} deff(\mathbf{c}'\hat{\boldsymbol{\theta}}, \mathbf{c}'\mathbf{V_0}\mathbf{c})$$

$$\delta_p = \min_{\mathbf{c}} deff(\mathbf{c}'\hat{\boldsymbol{\theta}}, \mathbf{c}'\mathbf{V_0}\mathbf{c})$$

Para este estudio se calculó la matriz de variancias y covariancias de la variable Estado Ocupacional considerando que el muestreo fuese simple al azar (caso iid), y que el diseño fuese complejo (como es el caso de la Encuesta) con dos variantes, que consistieron en tomar como Unidades Primarias de Muestreo (UPM) a las Áreas (conglomerados) y a los Hogares, respectivamente.

A continuación se presentan los resultados para el Aglomerado 1:

$$\mathbf{V}_{MSA} = \begin{bmatrix} 0.0000201 & -2.134x10^{-6} & -0.000018 \\ & 5.1202x10^{-6} & -2.986x10^{-6} \\ & & 0.0000210 \end{bmatrix}$$

$$\mathbf{V}_{MCArea} = \begin{bmatrix} 0.0000217 & -1.584x10^{-6} & -0.000020 \\ & 6.4728x10^{-6} & -4.889x10^{-6} \\ & & 0.0000250 \end{bmatrix}$$

$$\mathbf{V}_{MCHogaer} = \begin{bmatrix} 0.0000190 & -2.039x10^{-6} & -0.000017 \\ & 5.7467x10^{-6} & -3.708x10^{-6} \\ & & 0.0000206 \end{bmatrix}$$

Las matrices de deffs se calculan eliminando la última fila y columna de las matrices de variancias y covariancias, dado que las mismas son singulares.

$$\Delta_{Area} = \begin{bmatrix} 1.0930 & 0.0578 \\ 0.1461 & 1.2882 \end{bmatrix}$$

$$\Delta_{Hogar} = \begin{bmatrix} 0.9409 & 0.0185 \\ -0.0061 & 1.13 \end{bmatrix}$$

Los autovalores de estas matrices son:

$$\delta_{1,Area} = 1.32 \qquad \delta_{2,Area} = 1.06$$

$$\delta_{1,Hogar} = 1.12 \qquad \delta_{2,Hogar} = 0.94$$

Se puede observar que si bien los resultados son similares para los dos enfoques del diseño complejo, la utilización del Área como UPM da estimaciones de variancias mayores que las obtenidas bajo el supuesto que el Hogar sea la UPM. Esto se ve reflejado en el promedio de los autovalores de las matrices de deffs para Área y Hogar como UPM. La primera da un promedio de 1.19 contra 1.03 de la segunda. Se puede observar que prácticamente no existe efecto de diseño, y que considerar al Hogar como UPM en lugar del Área lleva aparejado una subestimación de la variancia.

Para el diseño que considera los Hogares como UPM, y en los estados 1 y 3 de la variable, se obtuvieron variancias menores que con un muestreo simple al azar. Esta ganancia en precisión se debe en parte a la heterogeneidad que existe en los Hogares respecto a la variable en estudio, lo cual no ocurre cuando se consideran a las Áreas como UPM.

Para el Aglomerado 2 los resultados fueron los siguientes.

$$\mathbf{V}_{MSA} = \begin{bmatrix} 0.0000197 & -5.672x10^{-7} & -0.000019 \\ & 1.5054x10^{-6} & -9.383x10^{-7} \\ & & 0.0000201 \end{bmatrix}$$

$$\mathbf{V}_{MCArea} = \begin{bmatrix} 0.0001088 & -1.294x10^{-6} & -0.000107 \\ & 9.5444x10^{-6} & -8.25x10^{-6} \\ & & 0.0001157 \end{bmatrix}$$

$$\mathbf{V}_{MCHogaer} = \begin{bmatrix} 0.0000878 & -2.625x10^{-6} & -0.000085 \\ & 8.1033x10^{-6} & -5.479x10^{-6} \\ & & 0.0000907 \end{bmatrix}$$

Las matrices de deffs se calculan eliminando las últimas filas y columnas de las matrices de covariancias dado que las mismas son singulares.

$$\Delta_{Area} = \begin{bmatrix} 5.5447 & 0.1179 \\ 1.2292 & 6.3844 \end{bmatrix}$$

$$\Delta_{Hogar} = \begin{bmatrix} 4.4463 & 0.0219 \\ -0.0683 & 5.3911 \end{bmatrix}$$

Los autovalores de estas matrices son:

$$\delta_{1,Area} = 6.53 \qquad \delta_{2,Area} = 5.39$$

$$\delta_{1,Hogar} = 5.39 \qquad \delta_{2,Hogar} = 4.45$$

Se observa que en este aglomerado el efecto de utilizar un diseño muestral complejo es muy alto, de 5.96 considerando las Áreas como UPM, y de 4.92 cuando consideramos los Hogares. Es por este motivo que la utilización de estimadores que consideren el verdadero diseño muestral utilizado es aconsejable para no incurrir en subestimaciones severas debidas al hecho de suponer un diseño muestral simple al azar, sobre todo en muestras de tamaño no muy grande.

Si bien el tema necesita profundizar el estudio, en principio cabe señalar que como causa posible de las diferencias existentes entre los deffs hallados para los 2 aglomerados se debería considerar el tamaño de muestra, específicamente el número de UPM seleccionadas en cada uno de ellos, y la correlación intraclase en los conglomerados.

4. REGRESIÓN LOGÍSTICA CON VARIABLES DICOTÓMICAS

Sea **x** un vector de variables explicativas e *y* una variable respuesta dicotómica, y sean ($\mathbf{x}_k$, y_k) los valores que toman esas variables en la población. Se supone que para un $\mathbf{x}_k$ dado, y_k proviene de un modelo con media $E(y_t)=\mu_t(\theta)$ y variancia $var(y_t)=v_{ot}$, y que $E(y_t)$ está relacionada con las variables explicativas a través del modelo de regresión logística

$$log\left[\mu_k(\boldsymbol{\theta})/(1-\mu_k(\boldsymbol{\theta}))\right] = \mathbf{x}_k'\boldsymbol{\theta}$$

con $v_{ok}=\mu_k(1-\mu_k)$. Cuando la muestra es de diseño complejo, se obtiene un estimador denominado pseudo máximo verosímil $\hat{\boldsymbol{\theta}}$ del parámetro $\boldsymbol{\theta}$ a través de los datos muestrales $\{(\mathbf{x}_k, y_k), k \in s\}$ resolviendo

$$\hat{\mathbf{T}}(\boldsymbol{\theta}) = \sum_{k \in s} w_k \mathbf{u}_k(\boldsymbol{\theta}) = 0$$

donde w_k son los pesos muestrales del diseño muestral particular que se este considerando, y $u_k(\theta)=[y_k-\mu_k(\theta)]x_k$.

La dificultad de trabajar con datos provenientes de diseños muestrales complejos radica mayormente en la estimación de la matriz de covariancias $\mathbf{V}(\hat{\boldsymbol{\theta}})$ de $\hat{\boldsymbol{\theta}}$, dado que la mayoría de los programas computacionales, por ejemplo el procedimiento LOGISTIC del programa SAS, calculan la misma haciendo el supuesto que las observaciones son independientes e igualmente distribuidas, o sea, provenientes de una muestra simple al azar. Es por ello que se presenta un programa alternativo que considera en la estimación de dicha matriz la estructura compleja de la muestra seleccionada.

4.1. DESCRIPCIÓN DEL SOFTWARE SUDAAN

El programa SUDAAN desarrollado por el Research Triangle Institute permite el análisis de datos provenientes de diseños muestrales complejos. Puede ser ejecutado en forma independiente o dentro del programa SAS.

Está constituido por un conjunto de procedimientos dentro de los cuales encontramos el LOGISTIC que permite ajustar modelos de regresión logística.

Dentro de las metodologías que pueden ser aplicadas para el cálculo de las estimaciones de las variancias el paquete presenta no sólo el método de linearización por Series de Taylor, sino también la posibilidad de aplicar métodos de replicación como el Jackknife y el BRR (Balanced Repeated Replications).

En este trabajo se utiliza el método de linearización por Series de Taylor y el de Jackknife.

El primero estima la matriz de covariancia $V(\hat{\boldsymbol{\theta}})$ de $\hat{\boldsymbol{\theta}}$ a través de

$$\hat{\mathbf{V}}_{\mathbf{L}}(\hat{\boldsymbol{\theta}}) = \left[\mathbf{J}(\hat{\boldsymbol{\theta}})\right]^{-1} \hat{\mathbf{V}}(\hat{\mathbf{T}}) \left[\mathbf{J}(\hat{\boldsymbol{\theta}})'\right]^{-1}$$

donde $\mathbf{J}(\hat{\boldsymbol{\theta}}) = -\sum_{k \in s} w_k \partial \mathbf{u}_k(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$ es la matriz de información observada y $\hat{\mathbf{V}}(\hat{\mathbf{T}})$ es la matriz de covariancias estimada del total estimado $\hat{\mathbf{T}}(\boldsymbol{\theta})$ cuando $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

Suponiendo un muestreo multietápico con UPM estratificadas, la técnica Jackknife permite obtener la estimación de la matriz de covariancia $V(\hat{\boldsymbol{\theta}})$ de $\hat{\boldsymbol{\theta}}$ mediante el estimador

$$\hat{\mathbf{V}}_{\mathbf{J}}(\hat{\boldsymbol{\theta}}) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} \left(\hat{\boldsymbol{\theta}}_{(\mathbf{hi})} - \hat{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}_{(\mathbf{hi})} - \hat{\boldsymbol{\theta}}\right)'$$

donde *L* es el número de estratos, n_h es el número de UPM del estrato *h* y $\hat{\boldsymbol{\theta}}_{(hi)}$ es obtenido de igual forma que $\hat{\boldsymbol{\theta}}$ luego de haber eliminado el conglomerado (*hi*) y utilizando los pesos muestrales Jackknife que vienen dados por $w_{hik(gj)} = w_{hik} b_{gj}$ donde

$$b_{gj} = \begin{cases} 0 & si\ (hi) = (gj) \\ \dfrac{n_g}{n_g - 1} & si\ h = g\ \wedge\ i \neq j \\ 1 & si\ h \neq g \end{cases}$$

La utilización del subíndice (hik) refleja el diseño muestral complejo, representando al k-ésimo elemento (k=1,...,n_{hi}), del i-ésimo conglomerado o UPM (i=1,...,n_h) del h-ésimo estrato (h=1,..,L).

4.2. APLICACIÓN

En los dos aglomerados definidos anteriormente se ajustó un modelo de regresión logística. Sólo se consideró la población económicamente activa. Las variables intervinientes en el modelo son las siguientes:

Variable Dependiente

$$\textit{Estado Ocupacional}:\ Y = \begin{cases} 0 & \text{ocupado} \\ 1 & \text{desocupado} \end{cases}$$

Variables Independientes

$$\textit{Sexo}:\ X_1 = \begin{cases} 0 & \text{masculino} \\ 1 & \text{femenino} \end{cases}$$

Edad: X_2*:* variable continua medida en años

Escolaridad: X_3*:* variable ordinal con los siguientes niveles: hasta primaria incompleta – primaria completa – secundaria incompleta – secundaria completa – superior o universitaria incompleta – superior o universitaria completa.

Se ajustó un modelo de regresión logística con los siguientes efectos: sexo, edad, edad al cuadrado, escolaridad, edad*sexo, edad al cuadrado*sexo, escolaridad*sexo. Nuevamente se consideraron tres diseños muestrales: muestreo simple al azar, diseño muestral complejo con Área como UPM, y diseño muestral complejo con Hogar como UPM. Para el primero se utilizó el procedimiento LOGISTIC de SAS, para el segundo y tercer diseño se utilizó el programa SUDAAN aplicando el método de Linearización de Taylor y el de Jackknife. Cabe señalar que la consideración del diseño de la muestra afecta sólo el valor del error estándar del estimador, no así el valor de la estimación puntual. Por lo tanto en los cuadros siguientes se presentan la estimación puntual de cada efecto, el desvío estándar de dicha estimación, con Muestreo Simple al Azar (MSA), y por los dos métodos presentados, Linearización (Lin) y Jackknife (Jack), considerando Área (MCA) y Hogar (MCH) como UPM respectivamente. Además, la significación de cada coeficiente se determina mediante el test de Wald, que consiste en calcular el cociente entre el estimador y su desvío estándar. Al calcular éstos últimos de diferentes maneras puede ocurrir que se arribe a distintas conclusiones según el desvío estándar que se tenga en cuenta. Por tal

razón al lado de los desvíos estándares se agrega un * en caso en que el test de Wald que lo utiliza en su cálculo resulte con una probabilidad asociada de $p<0.05$.

Los resultados para el aglomerado 1 fueron los siguientes.

Tabla 4.2.1 Estimación de los parámetros, y de los errores estándares considerando MSA y Muestreo Complejo con Área y Hogar como UPM

Parámetro	Estim	Error Estándar				
		MSA	MCA-Lin	MCA-Jack	MCH-Lin	MCH-Jack
Constante	1.8732	0.3598 *	0.3647 *	0.3670 *	0.3622 *	0.3645 *
Sexo (X_1)	-0.2173	0.5788	0.5829	0.5867	0.5828	0.5867
Edad (X_2)	-0.1538	0.0181 *	0.0182 *	0.0183 *	0.0179 *	0.0180 *
Edadcuad (X_2^2)	0.0017	0.0002 *	0.0002 *	0.0002 *	0.0002 *	0.0002 *
Escolaridad (X_3)	-0.2219	0.0400 *	0.0409 *	0.0410 *	0.0404 *	0.0405 *
Sexo*Edad ($X_1 * X_2$)	0.0303	0.0301	0.0301	0.0303	0.0294	0.0296
Sexo*Edadcuad ($X_1 * X_2^2$)	-0.0005	0.0004	0.0004	0.0004	0.0003	0.0004
Sexo*Escolaridad ($X_1 * X_3$)	0.0402	0.0574	0.0550	0.0552	0.0553	0.0556

* El test de Wad que utiliza este valor como desvío estándar resulta significativo al nivel del 5%

Se puede observar que tanto los errores estándares como la significancia de los coeficientes no difieren para las distintas estimaciones de las variancias. Este resultado, similar al encontrado para las estimaciones de matrices de covariancias en el punto 3, puede deberse al hecho de trabajar con una muestra con un gran número de pequeñas UPM, lo que elimina en parte el efecto de la conglomeración.

En el cuadro siguiente se presentan los resultados para el Aglomerado 2.

Tabla 4.2.2 Estimación de los parámetros, y de los errores estándares considerando MSA y Muestreo Complejo con Área y Hogar como UPM

Parámetro	Estim	Error Estándar				
		MSA	MCA-Lin	MCA-Jack	MCH-Lin	MCH-Jack
Constante	1.8979	1.5851	1.8333	1.9992	1.4983	1.5791
Sexo (X_1)	-8.3913	5.0807	3.5830 *	5.1487	4.3561	6.3092
Edad (X_2)	-0.2071	0.0852 *	0.0874 *	0.0942 *	0.0760 *	0.0803 *
Edadcuad (X_2^2)	0.0022	0.0011 *	0.0010 *	0.0011	0.0009 *	0.0009 *
Escolaridad (X_3)	-0.1487	0.1660	0.1619	0.1741	0.1455	0.1529
Sexo*Edad ($X_1 * X_2$)	0.5261	0.2972	0.2053 *	0.3095	0.2683	0.4010
Sexo*Edadcuad ($X_1 * X_2^2$)	-0.0073	0.0042	0.0031 *	0.0048	0.0039	0.0062
Sexo*Escolaridad ($X_1 * X_3$)	-0.2231	0.2748	0.1489	0.1606	0.1989	0.2141

* El test de Wad que utiliza este valor como desvío estándar resulta significativo al nivel del 5%

El modelo considerando los coeficientes significativos cuando se tienen en cuenta el diseño muestral complejo con Área como UPM es diferente al obtenido cuando se

consideran un muestreo simple al azar y un diseño complejo con Hogar como UPM. La utilización de distintas técnicas, Linearización y Jackknife, también ofrece resultados diferentes para el diseño complejo con Áreas como conglomerados.

5. CONCLUSIONES

La consideración del diseño complejo de la muestra en la Encuesta Permanente de Hogares tuvo fundamentalmente diferentes resultados según el tamaño del Aglomerado analizado.

De acuerdo a las diferentes estrategias utilizadas se puede concluir:

- El efecto de diseño es mayor cuando se consideran las Áreas como UPM que cuando se consideran a los Hogares.
- En el aglomerado 1, que está formado por un gran número de UPM, los efectos de diseño son semejantes a uno, y se obtiene resultados similares en las regresiones ajustadas cuando se considera o no el diseño complejo de la muestra.
- En el aglomerado 2, el cual está compuesto por un número menor de UPM, los efectos de diseño son grandes. En cuanto a la regresión logística las conclusiones no varían entre el muestreo simple al azar y el complejo considerando los Hogares como UPM. Pero sí cambian entre el simple al azar y el diseño complejo que considera las Áreas como UPM.

El estudio posee además ciertas limitaciones, como por ejemplo:

- Sólo se estudiaron previamente los efectos de diseño de la variable considerada respuesta en las regresiones logísticas, pero no se estudió la magnitud de los efectos de diseño de las variables explicativas.
- La variable de estratificación, porcentaje de jefes del hogar con primaria incompleta, se encuentra relacionada con una de las variables independientes (escolaridad), y los pesos muestrales se encuentran post-estratificados (generalmente por sexo y grupos de edad). Ajustar modelos con variables independientes que se encuentren correlacionadas con las probabilidades de inclusión lleva a obtener en muchas oportunidades resultados erróneos.

Se hace necesario, por lo tanto, profundizar el estudio considerando variables explicativas alternativas a las presentadas en este trabajo incluyendo la consideración de sus posibles efectos de diseño.

BIBLIOGRAFÍA

BINDER, D. A. “On the Variances of Asymptotically Normal Estimators from Complex Surveys”. International Statistical Review, 51 pp. 279-292. 1983

HIDIROGLOU, M.A. y RAO, J.N.K. “Chi-Squared Tests with Categorical Data from Complex Surveys”. Journal of Official Statistics, Vol 3, Nº 2, pp. 117-132. 1987.

LEHTONEN, R. Y PAHKINEN, E. “Practical Methods for Design and Analysis of Complex Surveys”. John Wiley & Sons. 1996.

RAO, J.N.K. y SCOTT, A.J. “The Analysis of Categorical Data From Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables”. Journal of the American Statistical Association, Vol 76, Nº 374. 1981.

RAO, J.N.K. y THOMAS, D.R. “Analysis of Categorical Response Data From Complex Surveys: An Appraisal and Update”. International Conference on Analysis of Survey Data, Southampton, UK , Agosto 24-26, 1999.

SKINNER, C.J., HOLT, D. y SMITH, T.M.F. “Analysis of Complex Surveys”. John Wiley & Sons. 1989.

SUDAAN User Manual, Release 8.0. Volumes I and II. Research Triangle Institute, 2001.