CONSIDERACIÓN DEL MUESTREO COMPLEJO EN EL ANÁLISIS DE DATOS DE LA EPH*

Servy, E.; Marí, G.; Hachuel, L.

Instituto de Investigaciones Teóricas y Aplicadas en Estadística, Escuela de Estadística

OBJETIVO

Realizar la estimación de regresiones logísticas con datos provenientes de la Encuesta Permanente de Hogares (EPH) 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.

DISEÑO MUESTRAL

- Se lleva a cabo dos veces al año en 28 aglomerados urbanos y uno urbano-rural.
- En cada dominio se selecciona una muestra probabilística en dos etapas de selección, con UPM estratificadas de acuerdo al porcentaje de jefes de hogar con educación primaria incompleta. Las UPM son Áreas seleccionadas con probabilidad proporcional al total de viviendas ocupadas de las mismas. En la segunda etapa se seleccionan viviendas en forma sistemática, de manera tal de obtener una muestra autoponderada.
- Este estudio se lleva a cabo en dos aglomerados, 1 y 2, para datos de la onda de mayo de 1998. Los totales muestrales se muestran en el cuadro

Totales Muestrales para los 2 aglomerados en estudio

	Aglomerado 1	Aglomerado 2
Personas	11807	2212
Hogares	3549	624
Viviendas	3433	614
Áreas	518	48
Estratos	12	5

RESULTADOS

MATRIZ DE VARIANCIAS Y COVARIANCIAS BAJO MUESTRAS COMPLEJAS

- Variable en estudio: Estado Ocupacional: Ocupado, Desocupado e Inactivo.
- Estimación de las matrices de variancias y covariancias de la proporción de Ocupados, Desocupados e Inactivos

Aglomerado 1

i) Muestreo Simple al Azar (MSA)

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

ii) Muestreo Complejo con Área como UPM (MCA)

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

iii) Muestreo Complejo con Hogar como UPM (MCH)

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

Aglomerado 2

i) Muestreo Simple al Azar (MSA)

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

ii) Muestreo Complejo con Área como UPM (MCA)

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

iii) Muestreo Complejo con Hogar como UPM (MCH)

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

- Efecto de diseño: se estima la matriz de efectos de diseño entre $\mathbf{V}_{VERD}$ y $\mathbf{V}_{MSA}$

$$\mathbf{deff}(\hat{\theta}, \mathbf{V}_{MSA}) = \Delta = E_{VERD}(\mathbf{V}_{MSA})^{-1}\mathbf{V}_{VERD}(\hat{\theta})$$

- Efecto de diseño generalizado: autovalores de Δ.

Aglomerado	Promedio de Deffs Generalizados	
	UPM=Área	UPM=Hogar
Aglomerado 1	1.19	1.03
Aglomerado 2	5.96	4.92

CONCLUSIONES

- En el Aglomerado 1 prácticamente no existe efecto de diseño.
- En el Aglomerado 2 el efecto de diseño es grande.
- La utilización del Área como UPM da estimaciones de variancias mayores que las obtenidas bajo el supuesto que el Hogar es la UPM, lo que lleva aparejado una subestimación de la verdadera variancia.
- 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
- En principio cabe señalar como causa posible de las diferencias existentes entre los deffs hallados entre los aglomerados se debería considerar el tamaño de muestra, específicamente el número de UPM seleccionadas, y la correlación intraclase en los conglomerados.

REGRESIÓN LOGÍSTICA CON VARIABLES DICOTÓMICAS

- *Variable Dependiente: Estado Ocupacional: 0 ocupado, 1 desocupado*
- *Variables Independientes:*
  - *Sexo: 0 masculino, 1 femenino*
  - *Edad*: X_2: variable continua medida en años
  - *Escolaridad*: X_3: hasta primaria incompleta – primaria completa – secundaria incompleta – secundaria completa – superior o universitaria incompleta – superior o universitaria completa.

Se consideraron los 3 enfoques de diseño muestral. Para los diseños complejos se utilizó el programa SUDAAN, aplicando el método de Linearización de Taylor (Lin) y el de Jackknife (Jack)

Aglomerado 1

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%

Aglomerado 2

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%

CONCLUSIONES

- En el aglomerado 1 las regresiones ajustadas cuando se consideran los tres diseños muestrales brindan resultados similares
- En el aglomerado 2, las conclusiones no varían entre el muestreo simple al azar y el complejo considerando los Hogares como UPM. Hay variación en los resultados entre el simple al azar y el diseño de complejo considerando Áreas como UPM.

Se necesita profundizar en:

- La magnitud de los efectos de diseño de las variables explicativas ya que sólo se estudiaron previamente los efectos de diseño de la variable considerada respuesta en las regresiones logísticas.
- Características de las variables explicativas: la variable de estratificación se encuentra relacionada con una variable independiente (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 resultados erróneos.

* Trabajo realizado en el marco del Proyecto de Investigación y Desarrollo (Nº19/E069): “Estudio de datos categóricos medidos a través del tiempo”. Secretaría de Ciencia y Tecnología de la UNR.