ANÁLISIS DE DATOS PROVENIENTES DE DISEÑOS MUESTRALES COMPLEJOS: APLICACIONES A LA ENCUESTA DE HOGARES Y EMPLEO DE LA PROVINCIA DE BUENOS AIRES[1].

Autores: Damonte C[2], Monteverde M[3], Perez V[4] y Sotelo R[5]

(VERSIÓN DEL 23 DE AGOSTO DE 2012, PRELIMINAR, NO CITAR).

RESUMEN

Los datos de las encuestas a hogares para la obtención de estadísticas sociales, sobre mercado de trabajo, salud y epidemiología, entre otras, provienen mayoritariamente de estudios con diseños muestrales complejos y ello es cierto tanto para la Argentina como para otros países de la región y del mundo.

Este trabajo busca resumir la discusión existente entre tres enfoques de inferencia: a) el enfoque de inferencia basado en el *diseño de la muestra*, b) el basado en *modelos superpoblacionales* y c) la inferencia *asistida por modelos*, intentando clarificar las implicancias respecto a la utilización de la información de diseño de la muestra para realizar inferencia bajo cada esquema. Además, se discuten los efectos del diseño muestral complejo sobre las estimaciones de parámetros poblacionales y sobre las estimaciones de parámetros de modelos de regresión lineal y logística. Por último, se realiza un ejercicio con el fin de medir los efectos de diseño con datos reales, para lo cual se utilizan datos de la Encuesta de Hogares y Empleo (EHE), una encuesta a hogares, diseñada y desarrollada por la Dirección Provincial de Estadística de la provincia de Buenos Aires (DPE).

---

[1] Trabajo presentado en el V Congreso de la Asociación Latinoamericana de Población, Montevideo, Uruguay, del 23 al 26 de octubre de 2012.

[2] Dirección de Planificación, Metodología y Coordinación del Sistema Estadístico Provincial, Dirección Provincial de Estadística de la provincia de Buenos Aires (DPE). E-mail: cdamonte@estadistica.ec.gba.gov.ar.

[3] CIECS-CONICET, UNC. E-mail: montemale@yahoo.com

[4] Dirección de Estadísticas Sociales, DPE. E-mail: vperez@estadistica.ec.gba.gov.ar

[5] Dirección de Planificación, Metodología y Coordinación del Sistema Estadístico Provincial, DPE. E-mail: rsotelo@estadistica.ec.gba.gov.ar

La conclusión general a la que se arriba luego de la revisión de la literatura y los ejemplos analizados con datos reales, es que el costo de ignorar la información de diseño puede conducir a errores que pueden ser graves, en términos de modificar las conclusiones de inferencia, y en este sentido se recomienda contemplar dicha información, aún cuando implique un costo en términos de eficiencia de los estimadores por el aumento en el estimador de la varianza de diseño respecto al estimador basado en un modelo.

I- Introducción

Los datos de las encuestas a hogares para la obtención de estadísticas sociales, sobre mercado de trabajo, salud y epidemiología, entre otras, provienen mayoritariamente de estudios con diseños muestrales complejos y ello es cierto tanto para la Argentina como para otros países de la región y del mundo.

En general, puede afirmarse que es necesario tener en cuenta la información de diseño a la hora de realizar estimaciones, de modo tal que la propiedad de insesgadez o cuasi-insesgadez de los estimadores (de los parámetros y de sus varianzas) no se vea afectada cuando se trabaja con datos provenientes de diseños muestrales complejos (ver Medina 1998; Cañizares Pérez et al. 2004; Guillén et al. 2000, Lemeshow et al. 1998, entre otros). A pesar de lo anterior existe cierta discusión acerca de la necesidad de incorporar la información de diseño, al menos en ciertos casos, como es el de las estimaciones basadas puramente en modelos superpoblacionales.

Este trabajo busca resumir la discusión existente entre tres enfoques de inferencia: a) el enfoque de inferencia basado en el *diseño de la muestra*, b) el basado en *modelos superpoblacionales* y c) la inferencia *asistida por modelos*, intentando clarificar las implicancias respecto a la utilización de la información de diseño (de la muestra) para realizar inferencia bajo cada esquema. Además, se discuten los efectos del diseño muestral complejo sobre las estimaciones de parámetros poblacionales y sobre las estimaciones de parámetros de modelos de regresión lineal y logística. Por último, se realiza un ejercicio con el fin de medir los efectos de diseño con datos reales, para lo cual se utilizan datos de la Encuesta de Hogares y Empleo (EHE), una encuesta a hogares, diseñada y desarrollada por la Dirección Provincial de Estadística de la provincia de Buenos Aires (DPE).

II- Enfoques de inferencia

Para comprender los argumentos que hay detrás de la discusión antes mencionada, es preciso repasar brevemente los conceptos en los que se sustentan las llamadas "Inferencia Basada en Modelos Superpoblacionales" e "Inferencia Basada en el Diseño", que son los dos enfoques extremos de inferencia estadística.

De acuerdo a la perspectiva de la inferencia basada en modelos superpoblacionales (Royall 1970, Thompson 1988, Valliant, Dorfman and Royall 2000, entre otros), los valores de la población finita $y_1, y_2,\ldots, y_N$, se consideran realizaciones de un vector aleatorio $(Y_1, Y_2,\ldots, Y_N)$, donde Y_i es una variable aleatoria ligada al i-ésimo elemento de la población.

Para describir la distribución N-dimensional de estas variables se define un modelo ξ, llamado modelo de superpoblación. Los valores de la población de Y se suponen provenientes de una muestra aleatoria de una superpoblación que tienen asignada una distribución de probabilidad $p(Y \mid \theta)$ con parámetros fijos θ. La inferencia basada en este enfoque se basa en la frecuencia de distintas realizaciones del vector aleatorio (distribución del vector aleatorio) y la misma está condicionada a una sola muestra s, que es la realizada, y no a otras muestras posibles (como en el enfoque basado en el diseño descrito a continuación). Los estimadores utilizados bajo este enfoque tienen propiedades deseables bajo el modelo supuesto, sin importar que sean consistentes bajo la aleatorización inducida por la selección de la muestra, en caso de existir un diseño muestral. Por ejemplo, bajo un modelo de regresión lineal simple, el mejor estimador lineal insesgado de los parámetros es el de mínimos cuadrados ordinarios.

Por su parte, de acuerdo al esquema de inferencia basada en el diseño, descrito en textos tales como Hansen, Hurwitz and Madow (1953), Kish (1965) y Cochran (1975), los valores $y_1, y_2,\ldots, y_N$ de la característica Y evaluada en cada miembro de la población de tamaño N, son considerados valores constantes y no aleatorios. La aleatoriedad en este caso proviene de considerar el conjunto de todas las muestras posibles de tamaño n, y una distribución de probabilidad definida sobre este conjunto. Se define (S,p) como un espacio de probabilidad, donde S es el conjunto de todas las muestras posibles de tamaño n y p es la probabilidad definida sobre él, p es el llamado diseño muestral, de manera tal que la probabilidad de inclusión de un individuo k de la población es: $\pi_k = \sum_{s:k\in s} p(s)$.

Bajo este enfoque, la inferencia sobre una cantidad finita de la población Q=Q(Y) envuelve los siguientes pasos (Little, 2003):

a) La elección de un estimador $\hat{q} = \hat{q}(Y_{inc}, I)$, una función de la parte observada Y_{inc} de Y, que es insesgado o aproximadamente insesgado de Q con respecto a la distribución p. Donde *I* es la variable aleatoria que indica la pertenencia o no de un individuo de la población en la muestra.

b) La elección de un estimador de la varianza $\hat{v} = \hat{v}(Y_{inc}, I)$, que es insesgado o aproximadamente insesgado de la varianza de $\hat{q}$ con respecto a la distribución p.

Estas inferencias se basan generalmente en aproximaciones normales de muestras grandes. Por ejemplo, un intervalo de confianza de 95% para Q es $\hat{q} \pm 1{,}96\sqrt{\hat{v}}$.

En cuanto a los estimadores utilizados bajo este enfoque, reflejan las características del diseño muestral a través de la utilización de las probabilidades de selección (π).

A diferencia del enfoque de inferencia basada en el diseño, la inferencia basada en modelo requiere de la definición de un modelo específico para las variables "y" producto de la encuesta, el cual es utilizado para predecir los valores no muestreados de la población y por tanto las cantidades Q de la población finita. Dentro de este enfoque a su vez, existen dos grandes variantes: los modelos de superpoblación y los modelos bayesianos (Little, 2003). Si bien dichos enfoques difieren, son similares en el hecho de que ambos requieren de la especificación de la distribución de probabilidad $p(Y \mid \theta)$ (mientras que para el enfoque de diseño los valores de Y, poblacionales, son fijos).

Además de incidir en el tipo de estimadores a ser utilizados, la elección de un esquema u otro de inferencia tiene implicaciones directas e indirectas en el diseño y el método de selección de las muestras. Mientras que para el esquema de inferencia basado en el diseño resulta esencial que la muestra provenga de un diseño probabilístico, los más acérrimos defensores del enfoque basado en modelos, indican que no es necesaria la selección de la muestra en forma aleatoria. En este sentido, el enfoque basado en modelos, representa una ventaja, en particular para aquellos estudios en los que no resulta posible la obtención de una muestra probabilística, ya sea por la falta de un marco muestral más o menos completo, ya sea por restricciones de recursos (Liseras, 2004).

Sin embargo, en el caso de los datos provenientes de encuestas a hogares realizadas por las oficinas o institutos de estadísticas oficiales, en las que se utilizan diseños complejos basados en marcos muestrales relativamente completos y actualizados y en los que es posible calcular de forma razonable las pesos muestrales, cabe preguntarse ¿qué relevancia tiene la

discusión entre el esquema de inferencia basado en un modelo y el basado en el diseño? Específicamente, ¿es posible ignorar la aleatoriedad proveniente del diseño para la obtención de estimaciones insesgadas o cuasi-insesgadas de los parámetros poblacionales de interés? ¿Y cuáles son las consecuencias de ignorar dicha información de diseño?

Los trabajos de Hansen, Madow y Tepping (1983), Särndal, Swensson and Wretman (1992) y Little (2003), ayudan a responder estos interrogantes.

En el libro del año 1992 titulado *Model Assisted Survey Sampling*, Särndal y otros plantean que hay dos tipos de inferencias posibles a partir de una muestra de una población finita:

1) Inferencias sobre parámetros descriptivos de la población, que la caracterizan en determinado momento (parámetros que podrían conocerse exactamente si se llevara a cabo un censo sin errores de medición y con 100% de respuesta).

2) Inferencias acerca del proceso que genera la población finita: para ello es necesario plantear un modelo superpoblacional del cual interesará estimar sus parámetros (pero que nunca podrán conocerse exactamente).

Para analizar las consecuencias del diseño muestral en los dos esquemas de inferencia, los autores plantean el siguiente caso de regresión lineal:

Supongamos que nuestro interés es estimar los coeficientes de regresión que surgen de plantear la siguiente relación entre las variables y, z_1, z_2,…, z_q, que se recogen en la encuesta:

$$y = B_0 + B_1 z_1 + B_2 z_2 + \ldots + B_q z_q \qquad (*)$$

Y que el objetivo es realizar inferencia de tipo 1) y, por tanto, la idea es encontrar el "mejor" hiperplano de la forma (*). Si conociéramos los valores de las variables y, z_1,…, z_q para todos los individuos de la población finita, dicho hiperplano se obtendría por mínimos cuadrados ordinarios, es decir el vector de parámetros $\mathbf{B} = (B_0, B_1, ..., B_q)'$ que minimiza la expresión:

$$\sum_{k=1}^{N} (y_k - \mathbf{B}'\mathbf{z_k})^2 \qquad (1)$$

Donde:

y_k es el valor correspondiente de la variable y para el individuo k (=1,…, N) de la población finita.

$z_k = (1, z_{k1}, z_{k2}, \ldots, z_{kq})'$ es un vector columna donde z_{ki} es el valor de la variable z_i para el individuo k de la población.

La solución para B es:

$$B = (\sum_{k=1}^{N} \mathbf{z_k z'_k})^{-1} \sum_{k=1}^{N} \mathbf{z_k} y_k \qquad (2)$$

O en forma matricial:

$$\mathbf{B} = (\mathbf{z'z})^{-1}\mathbf{z'y},$$

siendo $\mathbf{z}$ la matriz cuyas filas son los vectores $\mathbf{z'_k}$.

Con lo cual cada componente del vector **B**, es decir cada parámetro de la regresión, es una función de totales de los valores de toda la población finita. Estos totales pueden estimarse a partir de una muestra de la población objeto de estudio, reemplazando cada total por su estimador de Horvitz Thompson (que tiene en cuenta lo pesos asociados al diseño muestral). En Särndal (1997, pp: 175) se demuestra que dicho estimador de razones de totales es aproximadamente insesgado.

Además, en el mismo trabajo se presenta un estimador de la matriz de covarianza aproximada del estimador de B bajo este esquema de inferencia:

$$\hat{V}(\hat{B}) = (\sum_{s} \mathbf{z_k z'_k} / \pi_k)^{-1} \hat{V} (\sum_{s} \mathbf{z_k z'_k} / \pi_k)^{-1} \qquad (3)$$

Donde $\hat{V}$ es una matriz qxq donde el elemento ij de la matriz es igual a:

$$\hat{v}_{ij} = \sum_{k \in s} \sum_{l \in s} (\Delta_{kl} / \pi_{kl})(z_{ik} e_k / \pi_k)(z_{jl} e_l / \pi_l)$$

Donde:

z_{ij} el elemento ij de la martiz z

π_k es la probabilidad de inclusión del individuo k

π_{kl} es la probabilidad de inclusión conjunta de las individuos k y l

$\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$, es la covarianza entre las variables indicadoras I_k e I_l

$e_k = y_k - \mathbf{z'_k} \hat{B}$, es el residuo muestral

Los elementos de la diagonal principal de la matriz $\hat{V}(\hat{B})$ son los estimadores de las varianzas aproximadas de los estimadores de los parámetros $\hat{B}_i$. Estas aproximaciones se obtienen a partir del desarrollo de Taylor de primer grado.

El método de linealización de Taylor tiende a subestimar las varianzas pero si los tamaños de muestras son grandes, este sesgo es despreciable, sobretodo para el cálculo de varianzas de estadísticos complejos (Särndal, 1992).

Para el caso particular de dos variables, supongamos que la relación a estudiar es:

$$y = B_1 z$$

Donde:

$B_1 = \sum_{k=1}^{N} z_k y_k / \sum_{k=1}^{N} z_k^2$, es un cociente de totales

Mientras que la expresión del estimador propuesto para el parámetro es:

$$\hat{B}_1 = (\sum_{k=1}^{n} z_k y_k / \pi_k) / (\sum_{k=1}^{n} z_k^2 / \pi_k)$$

Y el estimador correspondiente de la varianza aproximada del parámetro es:

$$\hat{V}(\hat{B}_1) = [\sum_{k=1}^{n} \sum_{l=1}^{n} \breve{\Delta}_{kl} (z_k e_k / \pi_k)(z_l e_l / \pi_l)] / [\sum_{k=1}^{n} z_k^2 / \pi_k]^2$$

con $\breve{\Delta}_{kl} = \Delta_{kl} / \pi_{kl}$

Es decir, las expresiones de los estimadores (tanto del parámetro como de la varianza) dependen del diseño de muestral utilizado.

Sólo en el caso de tratarse de un muestreo aleatorio simple, las expresiones de estos estimadores de coeficientes de regresión coinciden con los obtenidos por mínimos cuadrados ordinarios de la expresión (2).

Por otra parte, si el objetivo es realizar inferencia del tipo 2), es necesario plantear un modelo. Sea el modelo tradicional de regresión:

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_q z_q + \varepsilon_k \qquad k = 1, \ldots, N \qquad (**)$$

Con $\varepsilon_1, \ldots, \varepsilon_N$ variables aleatorias independientes $N(0, \sigma^2)$

La cuestión que surge aquí es qué estimador de los parámetros utilizar.

Si el estimador basado en el modelo: $\hat{\beta}_M = (\sum_{k=1}^{n} \mathbf{z_k z_k}')^{-1} (\sum_{k=1}^{n} \mathbf{z_k} y_k)$, que ignora el diseño de la muestra y que está probado que, si el modelo es correcto, es el mejor estimador lineal insesgado. O utilizar el estimador $\hat{\beta}_s = (\sum_{k=1}^{n} \frac{\mathbf{z_k z_k}'}{\pi_k})^{-1} (\sum_{k=1}^{n} \frac{\mathbf{z_k} y_k}{\pi_k})$, que surge de reemplazar cada suma en $\hat{\beta}_M$ por su estimador de Horvitz Thompson.

Särndal y colegas (1992) señalan que si el modelo es correcto, $\hat{\beta}_M$ es mejor estimador que $\hat{\beta}_s$, pues es el estimador lineal insesgado de menor varianza (y por lo tanto de menor error cuadrático medio) bajo el modelo:

$$E_\xi((\beta_M - \beta)^2 / s) \leq E_\xi((\beta_s - \beta)^2 / s), \text{ para cualquier muestra s.}$$

Con lo cual:

$$E_\xi E_p((\beta_M - \beta)^2 / s) \leq E_\xi E_p((\beta_s - \beta)^2 / s)$$

Donde:

E_ξ es la esperanza bajo las hipótesis del modelo

E_p es la esperanza bajo las hipótesis del diseño

Es decir que, $\hat{\beta}_M$ es mejor estimador que $\hat{\beta}_s$ bajo ambas distribuciones (modelo y diseño muestral).

Pero $\hat{\beta}_s$ tiene la ventaja de ser más robusto, en el sentido que si el modelo no es del todo correcto conserva sus propiedades de consistencia, y además es insesgado bajo el modelo.

Además, señala que aún considerando un modelo más realista que (**), que incorpore varianzas no constantes y errores correlacionados (no independientes), dentro de aglomeraciones naturales de la población, el cálculo del mejor estimador para β y su varianza, aunque exista en teoría, sería imposible de calcular ya que la composición exacta de las aglomeraciones y la correlación interna de los errores son, muy probablemente, desconocidos.

Por otra parte, para describir formalmente el efecto del diseño de la muestra en un modelo logístico (que es utilizado para el ejercicio con datos reales más adelante), se presenta el siguiente caso.

Supongamos que el objeto de estudio sea una variable categórica y variables independientes $z_1, z_2,\ldots, z_q$. Para cada categoría "c" de la variable, definimos una variable binaria:

$$Y_c = \begin{cases} 1 & \text{para los individuos de la categoria c} \\ 0 & \text{caso contrario} \end{cases}$$

Sea $P = P(Y_c = 1/\mathbf{z})$, la probabilidad de que la variable Y_c sea 1 dado el vector de variables $\mathbf{z} = (z_1, z_2,\ldots, z_q)'$.

Como en el caso del modelo de regresión lineal, la idea es estimar los coeficientes, pero ahora para la relación entre la función logit de P y el vector de variables **z**:

$$\ln\left[\frac{P}{(1-P)}\right] = B_0 + B_1 z_1 + ... + B_q z_q$$

Encontrados los valores de $\mathbf{B} = (B_0, B_1,..., B_q)'$, el valor de P se obtiene como:

$$P(\mathbf{B}) = \frac{e^{B_0+...+B_q z_q}}{1+e^{B_0+...+B_q z_q}}$$

Si conociéramos los valores de Y_c, $z_1,\ldots,z_q$ para todos los individuos de la población finita, el vector de parámetros **B**, se obtendría, aplicando el método de máxima verosimilitud, de manera que maximice la función:

$$\prod_{i=1}^{N} P(\mathbf{B}/\mathbf{z_i})^{y_i}(1-P(\mathbf{B}/\mathbf{z_i}))^{1-y_i}$$

Donde:

y_i es el valor de Y_c para el $i-$ésimo individuo
$z_i = (1, z_{1i},...,z_{qi})$ son los valores de las z_j en el individuo i

Como no se conocen los valores de las variables para toda la población, la función que se maximiza es una estimación de la anterior, a partir de los valores de la muestra:

$$\prod_{i=1}^{n} \{P(\mathbf{B}/\mathbf{z_i})^{y_i}(1-P(\mathbf{B}/\mathbf{z_i}))^{1-y_i}\}^{1/\pi_i}$$

Donde:

π_i es la probabilidad de inclusión del individuo i- ésimo en la muestra.

La función anterior se conoce como función de pseudo verosimilitud ponderada.

La solución para **B** que maximiza dicha función, se obtiene resolviendo un sistema de q+1 ecuaciones obtenido de derivar el logaritmo natural de la función respecto de los parámetros. Teniendo en cuenta que $P(B/z_i) = \frac{e^{B'z}}{1+e^{B'z}}$, al derivar respecto de los B_j el sistema a resolver resulta:

$$\hat{S}(\hat{\mathbf{B}}) = \sum_{i=1}^{n}\left(\frac{1}{\pi_i}\right)\left(y_i - P(\frac{\hat{\mathbf{B}}}{\mathbf{z_i}})\right)\mathbf{z_i} = \mathbf{0}$$

Esta solución no es directa, sino que se obtiene aplicando un procedimiento iterativo.

A partir de la distribución límite de este estimador, Binder (1983) probó su insesgadez asintótica y consistencia.

Aplicando el procedimiento de Taylor en varias variables, dicho autor propuso el siguiente estimador de la matriz de covarianza aproximada de $\hat{\mathbf{B}}$:

$$\hat{V}(\hat{\mathbf{B}}) = \hat{J}^{-1}\, \hat{var}(\hat{S}(\hat{\mathbf{B}}))\ \hat{J}^{-1}$$

Donde:

J es la matriz de las derivadas segundas del logaritmo de la función de pseudo verosimilitud ponderada, respecto de los parámetros, y $\hat{var}(\hat{S}(\hat{\mathbf{B}}))$ es un estimador de la matriz de covarianzas de $\hat{S}(\mathbf{B})$.

Como los elementos del vector $\hat{S}(\mathbf{B})$ son totales de variables, los elementos de su función de covarianza también son totales y por lo tanto $\hat{var}\left(\hat{S}(\hat{\mathbf{B}})\right)$, obtenido al reemplazar $\mathbf{B}$ por el estimador consistente $\hat{\mathbf{B}}$ es un estimador consistente de $var(\hat{S}(\mathbf{B}))$ (Särndal 1992, pág. 168). Además como los elementos de J son totales de variables, existe un estimador consistente de J, cuyos elementos son los estimadores de Horvitz-Thompson de los elementos de J.

Por lo tanto el estimador $\hat{V}(\hat{\mathbf{B}})$ es un estimador consistente (ver Anexo I).

Nuevamente, las expresiones de los estimadores (tanto del parámetro como de la varianza) dependen del diseño de muestra utilizado.

Esta estimación es la que utilizan algunos softwares estadísticos, otros utilizan métodos de replicación de muestras, como Jakknife y muestras replicadas balanceadas.

III- Estimación de los Efectos de Diseño

A continuación se presenta un ejercicio para la población de la localidad de Olavarría, en la Provincia de Buenos Aires, con el objetivo de:

a) Medir los efectos de la información de diseño en la inferencia de parámetros descriptivos de una población (Inferencia de tipo 1) y

b) Realizar inferencia acerca del proceso que genera una de las variables ouput de interés de la encuesta (la condición de "ocupado"), para lo cual se asume un modelo (teórico) de superpoblación (Inferencia de tipo 2). En este caso se evalúa en qué medida los resultados de la estimación del modelo son robustos manteniéndose las conclusiones obtenidas con la muestra sin considerar la información del diseño (bajo el supuesto de que la misma proviene de un muestreo simple al azar) respecto a la muestra considerando dicha información. Notar que bajo el esquema de modelos superpoblacionales, si el modelo está perfectamente especificado, entonces la inferencia (acerca de los parámetros de dicho modelo) no debería verse afectada de manera significativa por el tipo de muestra.

III.1 Materiales y Métodos

La encuesta

Los datos que se utilizan para el desarrollo del ejercicio provienen de la Encuesta de Hogares y Empleo (EHE) y son obtenidos mediante un muestreo complejo. Esta encuesta es un programa de la Dirección Provincial de Estadística (DPE) de la provincia de Buenos Aires de Argentina que releva, en el ámbito municipal urbano[6], datos referidos a las características socioeconómicas de la población, siendo los hogares particulares su unidad de análisis. Esta encuesta está pensada como un sistema integrado de indicadores sociales, lo que la hace una encuesta de objetivos múltiples. Enfatizando una perspectiva socioeconómica, representa a la situación de las personas y de los hogares, según su lugar en la estructura social.

El diseño de la muestra

Encuestas como la EHE enfrentan restricciones prácticas que hacen que el muestreo simple aleatorio (MSA) no sea factible o no sea conveniente y por tanto sea necesario recurrir a otras alternativas de muestreo como el muestreo estratificado, el muestreo por conglomerados, el muestreo en etapas o el muestreo con probabilidades desiguales. Las

[6] Localidades urbanas de la provincia de Buenos Aires.

técnicas de muestreo que emplean una combinación de estas alternativas se denominan complejos.

El diseño de la muestra de la EHE contempla dos etapas de selección. Las unidades de primera etapa, Unidades Primarias de Muestreo (UPM), son las áreas, definidas en base a radios censales, y las de segunda etapa son las viviendas.

El procedimiento de selección de las unidades en la primera etapa es probabilístico y estratificado, y en la segunda solo probabilístico.

Las probabilidades de selección en la primera etapa son proporcionales al total de viviendas particulares, según los listados de viviendas previos al Censo Nacional de Población, Hogares y Viviendas 2010.

La selección de las viviendas en la segunda etapa se realiza en forma sistemática, dentro de cada UPM seleccionada en la primera etapa.

Para realizar la estratificación, los radios censales se clasifican según características de población y viviendas, utilizando información del Censo Nacional de Población, Hogares y Viviendas del año 2001, último disponible.

La selección de las UPM se realiza en forma sistemática y proporcional a la cantidad de viviendas particulares de las mismas, utilizando el procedimiento SURVEYSELECT del SAS.

Las probabilidades de selección de esta primera etapa, los define el procedimiento del SAS a partir de la indicación del tipo de selección aplicada, como:

$$P_{1i\,h} = V_i \times n_h / V_h$$

Donde,

P_{1ih}= probabilidad de selección de la i-ésima UPM perteneciente al estrato h,

V_i = total de viviendas particulares de la i-ésima UPM,

n_h = total de UPMs seleccionadas en el estrato h,

V_h = total de viviendas particulares del estrato h.

De los Listados de Viviendas confeccionados para el Censo Nacional de Población, Hogares y Viviendas 2010, se toman los registros correspondientes a las UPM que fueron seleccionadas. De esta manera se define la base de datos que se utiliza para la selección de las unidades de segunda etapa del diseño, las viviendas.

Para la selección de viviendas también se utiliza el procedimiento SURVEYSELECT del SAS con el método de selección sistemática. La probabilidad de selección de la vivienda j del área i-ésima es:

$$P_{2ij} = k/V_i$$

Donde:

k= total de viviendas seleccionadas en cada área,

V_i =total de viviendas particulares del área i.

Para el caso que se analiza, localidad de Olavarría año 2011, el tamaño muestral calculado es de 880 viviendas, distribuidas en 80 áreas seleccionadas (11 viviendas por área) entre 8 estratos (11 áreas por estrato).

Las ponderaciones

El procedimiento de estimación de resultados de la EHE contempla, en una primer etapa, la ponderación de los datos por el producto de las inversas de las fracciones de muestreo de cada etapa de diseño, es decir el factor de ponderación de las viviendas de un área es el producto entre la inversa de la probabilidad de selección del área y la inversa de la probabilidad de selección de la vivienda.

En una segunda etapa se calcula un factor de corrección por no respuesta global. Para realizar este ajuste aplicando un modelo de grupos homogéneos de respuesta, previamente definidos. Para estos grupos se utilizan los estratos, definidos a partir de características socioeconómicas, que ya fueran utilizados en el diseño de la muestra.

En una tercera y última etapa, el procedimiento contempla el ajuste de las primeras ponderaciones utilizando datos censales proyectados al año de realización de la encuesta. La idea de este ajuste que se basa en información auxiliar, consiste en reemplazar los pesos iniciales (en nuestro caso los pesos corregidos por no respuesta), por otros no muy diferentes de éstos y de manera que la distribución de las variables auxiliares deducidas de la muestra coincidan con la distribución poblacional de las mismas. Las variables auxiliares que se utilizan en el caso de la EHE son tamaño del hogar, edad y sexo.

Método de análisis

Con el fin de abordar el objetivo a) (inferencia tipo 1), se estima y analiza la tasa de desocupación. La misma se calcula como la proporción de desocupados dentro de la población económicamente activa (PEA). Se consideran desocupadas a todas las personas que

buscan activamente trabajo y no lo encuentran. La PEA se define como la población que se encuentra ocupada más la población que se encuentra desocupada.

Para ilustrar lo descrito en el objetivo b) (inferencia tipo 2) se realiza un análisis multivariado de la relación existente entre la condición de ocupación de un individuo y algunas variables sociodemográficas. El mismo consiste en un modelo de regresión logística que permite estudiar cómo varía la probabilidad de un individuo de estar ocupado, estimando la variable dicotómica *Ocupado* como función de las variables independientes.

Dado que la variable dependiente a analizar en este ejemplo es dicotómica, la regresión logística provee un análisis superior a la regresión lineal tradicional. De esta manera los coeficientes son estimados por el método de máxima verosimilitud.

La variable dependiente o explicada *Ocupado* toma el valor 1 si el encuestado reúne las condiciones para ser considerado ocupado y 0 si pertenece a la PEA pero está desocupado (la categoría base es desocupado).

Dentro de las posibles variables explicativas a considerar en el modelo se encuentran la edad, el sexo y la educación del individuo.

La variable *Edad* está medida en años cumplidos. Se consideran en el análisis a las personas de 18 años o más.

La variable *Sexo* es una variable dicotómica que toma el valor 1 si la persona es varón y 0 si es mujer (categoría base mujer).

Para incorporar la variable Educación, se crea una variable proxy de los años de educación formal recibidos denominada *Años_Educación*. Para su construcción se consideran los años o grados de estudios que aprobó el individuo en el nivel más alto alcanzado. Dependiendo del nivel y los años aprobados, se calcula la cantidad total de años estudiados adicionando los años de estudio de los niveles previos. Para el cálculo de estos últimos se toma como base el sistema de Educación Primaria (7 años) - Secundaria (5 años) por ser el que representa a la mayor proporción de personas en las edades consideradas en este análisis.

Por ejemplo: si la persona encuestada declara haber asistido hasta Secundaria Incompleta y que el último año aprobado en dicho nivel es 4º, entonces los años de estudios totales (*Años_Educación*) son 11, 7 de primaria completa + 4 de secundaria incompleta.

En los casos de educación para adultos, tanto de nivel básico como nivel medio, si bien la cantidad de años es menor que en los sistemas tradicionales, los contenidos son

aproximadamente los mismos y por tanto se toma en cuenta la cantidad de años equivalente al sistema Primaria - Secundaria. De esta manera 1 año de nivel básico para adultos equivale a 7/3 años de primaria y 1 año de nivel medio para adultos 5/3 años de secundaria.

Para el caso del nivel post-universitario, para computar la cantidad de años previos se asumen 5 años para el nivel universitario considerando que la mayoría de las carreras tienen dicha duración, es decir, en estos casos la cantidad de años totales previos a adicionar es 17, 12 (nivel básico + nivel medio) + 5 (superior universitario).

Los años de educación se comienzan a contar desde el primer año de educación obligatoria, es decir desde los niveles básicos (Primario, EGB, EPB), de manera que para los casos en que el nivel más alto alcanzado pertenezca a niveles iniciales (Jardín y Preescolar) se computa cero años de educación.

Se excluyen del análisis a las personas que asisten o asistieron a nivel educación especial, y al igual que a los ignorados se los considera como *missing*. La cantidad de casos en esta situación es prácticamente despreciable.

El modelo de regresión logística a estimar para la variable *ocupado* en función de *Edad*, *Sexo* y *Años_Educación* sería:

$$\text{Logit}(P) = \beta_0 + \beta_1 * \text{Edad} + \beta_2 * \text{Sexo} + \beta_3 * \text{Años\_Educación}$$

Para los objetivos a) y b) se comparan los resultados de emplear tres métodos estadísticos: i) el análisis suponiendo MSA (análisis convencional), en el que se supone que los datos provienen de un muestreo simple aleatorio y por tanto no considera la información del diseño complejo de la muestra; ii) el análisis con ponderaciones, en el que solo se toman en cuenta las ponderaciones muestrales para obtener las estimaciones pero las varianzas se obtienen sin considerar la información del diseño de la muestra[7]; y iii) el análisis con información del diseño, en el que, además de utilizar las ponderaciones muestrales al estimar el parámetro, se toman en cuenta todos los aspectos del diseño muestral complejo al estimar las varianzas.

El programa estadístico que se utiliza para realizar las estimaciones es el STATA 11.1.

[7] A pesar de los problemas obvios de este caso, se incluye en el análisis porque en ocasiones los microdatos incluyen los ponderadores pero no el resto de la información de diseño.

III.2. Resultados

A continuación, en la Tabla 1, se muestran los resultados de estimar un parámetro de la población, como es la tasa de desocupación, con sus respectivos errores estándar, para los tres métodos mencionados en el apartado anterior, y el efecto de diseño correspondiente.

El efecto del diseño (*deff*) se obtiene como la razón entre la varianza del estimador utilizando toda la información del diseño de la muestra y la varianza del estimador suponiendo MSA, y permite así mostrar el efecto de no tomar en cuenta el diseño en el análisis estadístico (Cañizares Pérez, et al. 2004; Naciones Unidas, 2009). La raíz cuadrada del efecto del diseño da la razón entre los errores estándar y se denomina *deft.*

deft = Error estándar diseño complejo / Error estándar MSA

En este ejemplo se presenta el *deft* que indica la eficiencia del diseño empleado en relación a un diseño MSA. Un valor de 1 indica que el error estándar obtenido por ambos diseños (complejo y MSA) es igual; es decir, el muestreo complejo es tan eficiente como un MSA con el mismo tamaño de muestra. Si el valor es superior a 1, el muestreo complejo produjo un error estándar mayor al obtenido con un MSA (Naciones Unidas, 2009).

Tabla 1: Estimación de la tasa de desocupación y errores estándar según cada método de análisis. Olavarría 2011

Indicador	Análisis suponiendo MSA		Análisis con ponderaciones		Análisis con información del diseño		Deft
	Estimador	Error estándar	Estimador	Error estándar	Estimador	Error estándar	
Tasa de desocupación	0,0721	0,0085	0,0710	0,0013	0,0710	0,0095	1,12

Fuente: Encuesta de Hogares y Empleo (EHE). Dirección Provincial de Estadística.

Las estimaciones puntuales obtenidas para la tasa de desocupación por los tres métodos no difieren demasiado, aunque existe un pequeño sesgo positivo (del 1,55%) al estimar asumiendo MSA sin ponderar respecto de los otros dos métodos. Ello sugiere que, dado el diseño y las respuestas efectivas, los desocupados quedan sobrerrepresentados en la muestra. Las ponderaciones muestrales permiten compensar las distorsiones de la representatividad de la muestra, ya que toman en cuenta las probabilidades de selección, la disminución de la muestra por no respuesta y las diferencias con la población.

La precisión de la estimación de la tasa de desocupación difiere según el método de cálculo utilizado. Los errores estándar son subestimados en aproximadamente un 12% cuando se asume MSA respecto al error que contempla toda la información del diseño (el cual resulta aproximadamente insesgado, de acuerdo a lo indicado en el apartado II).

Se aprecia también que en el análisis con ponderaciones pero asumiendo MSA en el cálculo de la varianza, los errores estándar son mucho menores. Esto se debe en gran parte a que el denominador en el cálculo de la varianza resulta de la suma de los ponderadores, reduciendo el estimador del error estándar, induciendo a que diferencias no significativas cobren significatividad. Con este ejemplo queda en evidencia que el uso de ponderadores si bien ayuda a corregir el sesgo de los estimadores de los parámetros, incrementa sustancialmente el sesgo del estimador de los desvíos estándar.

A continuación, en la Tabla 2 se presentan los resultados de las estimaciones de los coeficientes y de los desvíos estándar de la regresión logística para la variable *ocupado* por los tres métodos de análisis.

Tabla 2: Estimación de modelo de regresión logística para la variable ocupado, según cada método de análisis. Olavarría 2011

	Análisis suponiendo MSA		Análisis con ponderaciones		Análisis con información del diseño		
Ocupado	Coeficiente	Error estándar	Coeficiente	Error estándar	Coeficiente	Error estándar	Deft
Edad	0,025 (**)	0,011	0,026(***)	0,002	0,026(**)	0,013	1,04
Sexo	1,078 (***)	0,278	0,920(***)	0,042	0,920(***)	0,282	1,03
Años_Educación	0,084 (**)	0,034	0,067(***)	0,005	0,067	0,042	1,29
Constante	0,288	0,598	0,502	0,089	0,502	0,694	1,14

Fuente: Encuesta de Hogares y Empleo (EHE). Dirección Provincial de Estadística.

(*) Estadísticamente significativo al 10%

(**) Estadísticamente significativo al 5%

(***) Estadísticamente significativo al 1%

Al estimar los modelos de regresión se observan diferencias en las estimaciones de los coeficientes, indicando la existencia de sesgos en los parámetros estimados bajo el supuesto MSA. Los sesgos mayores se ven en el caso de las variables *Sexo* y *Años_ Educación*.

Para el coeficiente estimado de la variable Edad se observa un comportamiento aproximadamente igual cuando se supone MSA que cuando se incorpora toda la información del diseño. Sin embargo, para la variable sexo, bajo el supuesto de MSA se observa que el logit de la probabilidad de estar ocupado de un individuo aumentaría en 1,078 si es varón (lo que equivale a un aumento de la probabilidad de 0,254), mientras que dicho aumento sería 0,920 cuando se considera la información del diseño (lo que equivale a un incremento de la probabilidad de 0,285). Del mismo modo, cada año adicional de educación incrementa el logit de la probabilidad de estar empleado en 0,084 bajo MSA (0,479 en términos de probabilidad), pero al contemplar la información del diseño dicho incremento se reduce a 0,067 (una probabilidad de 0,483).

Además, se observa un aumento en el desvío estándar para los parámetros de las variables al incorporar la información del diseño, sugiriendo que dicho desvío se subestima cuando se hace el supuesto de MSA.

Por último, cabe resaltar el cambio en la significatividad estadística del parámetro estimado para la variable *Años_Educación*: los resultados del modelo estimado bajo el supuesto de MSA indican que la variable años de educación tiene un efecto positivo y estadísticamente significativo (al 5%) sobre la probabilidad de estar ocupado, significatividad que desaparece al contemplar la información de diseño.

IV- Consideraciones finales

El presente trabajo tuvo por objeto el estudio de los efectos del diseño muestral sobre la inferencia estadística realizada a partir de datos provenientes de encuestas con diseños complejos.

Se pretendió abordar el problema desde una perspectiva práctica, con el objeto que fuese de utilidad tanto para los usuarios de datos provenientes de encuestas con diseños muestrales complejos (como son la mayoría de las encuestas a hogares provenientes de las oficinas de estadísticas oficiales en la actualidad), como para los responsables de elaborar y difundir este tipo de estadísticas.

Sin embargo, y en el intento de dar respuesta a preguntas prácticas puntuales tales como: ¿Por qué es necesario utilizar la información del diseño muestral? ¿Cuándo es razonable asumir el supuesto de que las observaciones de la muestra están independiente e igualmente distribuidas? ¿Cuál es el costo de ignorar los efectos de diseño? ¿Qué significa suponer un

modelo de superpoblación? ¿Qué consecuencias tienen los errores de especificación del modelo?, resultó imposible no entrar en la amplia discusión existente en la literatura sobre inferencia estadística. Dar respuesta a los interrogantes antes mencionados, resultaba muy difícil si no se ponían en el contexto de las diferentes concepciones de inferencia y de los problemas teóricos y prácticos subyacentes en cada uno, lo que a su vez involucra de forma directa decisiones acerca del uso de la información de diseño para la obtención de conclusiones acerca del objeto de estudio (ya sea una población en un momento dado o un modelo de superpoblación más general).

Si bien es cierto que en los años 70 y durante las siguientes cuatro décadas, hubieron posiciones muy duras a favor y en contra de cada uno de los enfoques de inferencia (el basado en el diseño y el basado en el modelo), a lo largo de los años, y gracias a estudios y contribuciones hechas en el tema por diferentes estadísticos, ambas posiciones se han ido acercando, al grado que ha surgido un punto de vista armonizador entre la inferencia basada en modelos y la basada en el diseño, que se denomina Inferencia Asistida por Modelos.

Este nuevo enfoque basado en el diseño y asistido por modelos fue desarrollado fundamentalmente por Carl Erik Särndal, quien señala la importancia del uso de los modelos para afrontar problemas reales de las encuestas, como son por ejemplo, la existencia de datos faltantes y estimación en pequeñas áreas, pero sin dejar de contemplar la aleatoriedad inducida por el diseño muestral. Es posible afirmar que en la actualidad gran parte de los defensores de cada enfoque reconocen la importancia y la complementariedad del otro.

De acuerdo a Särndal (2010) *"hoy en día, dado que las dos líneas de pensamiento fundamentalmente diferentes existen, ¿pueden los que proponen la visión basada en el diseño proceder de manera efectiva sin referencia a modelos? ¿Pueden los que proponen el enfoque basado en modelos trabajar sin ninguna referencia a las características del muestreo de selección aleatoria? La respuesta es "no" en ambos casos. En cada enfoque se siente la necesidad de contemplar al otro, de integrar aspectos del otro. Para los defensores del enfoque basado en el diseño, el desafío fue, desde los '80 en adelante, hacer explícitos los modelos en presentaciones formales. Para los defensores del enfoque basado en modelos el desafío fue, y aún es, plantear formalmente lo referido a la selección aleatoria de la muestra."*

La idea de complementariedad de los enfoques también puede ser visualizada en trabajos como los de Graubard and Korn (2002) y Little (2003), en los que se observa un esfuerzo por

contemplar las dos posibles fuentes de aleatoriedad: la proveniente del diseño y la proveniente del modelo, a la hora de inferir el comportamiento de parámetros poblacionales.

De todo el análisis surge claramente que dado que bajo el enfoque de modelos la fuente de la aleatoriedad proviene del modelo que se asume, la clave para la obtención de inferencia satisfactoria bajo este esquema yace en la correcta especificación del mismo (cuando ello es posible). En este sentido, en línea con Särndal y otros (1992) y con Little (2003) cabe citar a Hansen, Madow and Tepping (1983) quienes afirman que *"un diseño modelo- dependiente poco razonable puede conducir a la obtención de inferencia no satisfactoria, problema que puede ser evitado con el uso de diseños muestrales probabilísticos".*

El ejercicio práctico desarrollado en el presente trabajo es un ejemplo concreto de los errores de inferencia que se pueden cometer si se ignora la información de diseño. Para nuestro ejemplo, se utilizaron dos tipos de estimaciones:

a) La estimación de un parámetro, en este caso la tasa de desocupación, con el objeto de realizar inferencia acerca de una característica de la población en un momento dado (inferencia de tipo 1).

b) La estimación de los parámetros del modelo (supuesto) que explicaría el comportamiento, es decir el proceso que determina la distribución de la variable "ocupado".

La conclusión general a la que se arriba luego de la revisión de la literatura y los ejemplos analizados con datos reales, es que el costo de ignorar la información de diseño puede conducir a errores que pueden ser graves (en términos de modificar las conclusiones de inferencia) y en este sentido se recomienda contemplar la información de diseño en el análisis, aún cuando implique un costo en términos de eficiencia de los estimadores (por el aumento en el estimador de la varianza de diseño respecto al estimador del esquema de modelo).

Al mismo tiempo, además de reconocer el riesgo en el que se puede incurrir si se ignora la información de diseño, resultaría de interés estudiar en qué casos, o bajo qué diseños, los efectos resultarían mayores. Lo que puede resultar especialmente relevante cuando se encuentran disponibles los microdatos, los pesos muestrales y una descripción general del diseño de la encuesta, pero no toda la información de diseño necesaria.

Bibliografía

Cañizares Pérez M, Barroso Utra I, Alfonso León A, García Roche R, Alfonso Sagué K, Chang de la Rosa M, Bonet Gorbea M y León EM. 2004. Estimaciones usadas en diseños muestrales complejos: aplicaciones en la encuesta de salud cubana del año 2001. *Rev Panam Salud Publica/Pan Am J Public* Health 15(3).

Cochran WG. 1977. *Sampling Techniques, 3rd*

Binder, D.A. 1983. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review,* 51, pp:279–292.

Guillén M, Juncà S, Rué M y Aragay J. M. Efecto del diseño muestral en el análisis de encuestas de diseño complejo. Aplicación a la encuesta de salud de Catalunya. *Gac Sanit* 14(5), pp:399-402.

Graubard B. I and Korn E.L. 2002. Inference for Superpopulation Parameters using Sample Surveys. *Statistical Science,* 17(1), pp:73-96.

Hansen, M. H., Madow, W. G. and Tepping, B. J. 1983. An evaluation of model-dependent and probability-sampling inferences in sample surveys (with discussion). *J. Amer. Statist. Assoc.* 78, pp: 776–807.

Hansen MH, Hurwitz WN and Madow WG. 1953. *Sampling Survey Mathods and Theory*, vols i and II, New Cork: Wiley.

Kish, L. 1965. *Survey Sampling*, New York: Wiley

Lemeshow S, Letenneur L, Dartigues JF, Lafont S, Orgogozo JM and Commenges D. 1998. Illustration of Analysis Taking into Account Complex Survey Considerations: The Association between Wine Consumption and Dementia in the PAQUID Study. *Am J Epidemiol.* 148(3).

Liseras N. 2004. Análisis de encuestas basado en diseño y modelos muestrales: una comparación entre métodos de inferencia aplicados al estudio de la vocación emprendedora en alumnos universitarios. Tesis de Maestría en Estadística Aplicada. Universidad Nacional de Córdoba.

Little R. 2003. To Model or Not to Model? Competing Modes of Inference for Finite Pupulation Sampling. The University of Michigan Department of Biostatistics Working Paper Series. Paper 4.

Medina F. 1998. Los Errores de Muestreo en las Encuestas Complejas, en Memoria del 1er. Taller Regional: Planificación y Desarrollo de Encuestas en Hogares para la Medición de las Condiciones de Vida, Proyecto MECOVI, ONU-CEPAL, Santiago de Chile, pp. 135-348.

Naciones Unidas, 2009. Diseño de muestras para encuestas de hogares: directrices prácticas. Estudios de métodos, serie F, nº 98. Disponible en: http://unstats.un.org/unsd/publication/seriesf/Seriesf_98s.pdf

Royall RM. 1970. On Finite Population Sampling Ander Certain Linear Regresión Models. *Biometrika* 57, pp. 377-387.

Särndal, C.E. 1978. Design-based and Model-based Inference in Survey Sampling. *Scandinavian Journal of Statistics*, 5, pp:27–52.

Särndal C-E, Swensson B and Wretman J. 1992. *Model Assisted Survey Sampling.* New York: Springer-Verlag. ISBN 0-387-40620-4.

Särndal, C.-E. 2010. Models in survey sampling. In *Model Assisted Survey Sumpling*. Carlson, Nyquist and Villani (eds), Official Statistics, Methodology and Applications in Honour of Daniel Thorburn, pp. 15-28. Available at officialstatistics.wordpress.com.

Sul Lee, E., Forthofer, R.N. 2006. *Analyzing complex survey data*, second edition. Sage Publications, Inc.

Steeven Heeringa, Brady West and Patricia Berglund. 2010. *Applied Survey Data Analysis*. New York: CRC Press.

Thompson ME. 1988. Superpopulations Models. *Enciclopedia of Statistical Sciences*, vol. I, 9, pp: 93-99.

Valliant R, Dorfman AH and Royall RM. 2000. *Finite Population Sampling and Inference: A Prediction Approach*, Wiley, New York.

Anexo I: Demostración de la propiedad de consistencia del estimador de la varianza bajo diseño muestral complejo en el caso de un modelo Logístico.-

A continuación replicamos la demostración que Binder (1983) desarrolló para modelos lineales generalizados, para el caso especial de regresión logística.

La función de pseudo-verosimilitud ponderada es:

$$\prod_{i=1}^{n} \left\{ \left[p(B/z_i) \right]^{y_i} \left[1 - p(B/z_i) \right]^{1-y_i} \right\}^{1/\pi_i} \qquad (1)$$

Donde:

$$p(B/z_i) = \frac{e^{B'z_i}}{1+e^{B'z_i}} = \frac{e^{B_0 + \sum_{j=1}^{q} B_j z_{ij}}}{1+e^{B_0 + \sum_{j=1}^{q} B_j z_{ij}}}$$

$$z_i = (1, z_{1i}, z_{2i}, ..., z_{qi})'$$

Aplicando ln a (1):

$$\sum_{i=1}^{n} \frac{1}{\pi_i} \left\{ y_i \ln(p(B/z_i)) + (1-y_i) \ln\left[1 - p(B/z_i)\right] \right\} =$$

$$= \sum_{i=1}^{n} \frac{1}{\pi_i} \left\{ y_i \ln\left(\frac{p(B/z_i)}{1-p(B/z_i)} \right) + \ln\left[1 - p(B/z_i)\right] \right\} =$$

Como $1 - p(B/z_i) = \frac{1}{1+e^{B'z_i}}$

La igualdad anterior se puede escribir como sigue:

$$\sum_{i=1}^{n} \frac{1}{\pi_i} \left\{ y_i \ln(e^{B'z_i}) + \ln\left(1/(1+e^{B'z_i})\right) \right\} =$$

$$= \sum_{i=1}^{n} \frac{1}{\pi_i} \left\{ y_i B'z_i - \ln\left(1+e^{B'Z_i}\right) \right\} \qquad (2)$$

Derivando (2) respecto de B para hallar los estimadores de máxima verosimilitud y teniendo en cuenta que $B'z_i = B_0 + B_1 z_{1i} + B_2 z_{2i} + ... + B_q z_{qi}$:

$$\frac{\partial}{\partial B_0} = \sum_{i=1}^{n} \frac{1}{\pi_i} \left\{ y_i - \frac{e^{\hat{B}'Z_i}}{1+e^{\hat{B}'Z_i}} \right\} = \sum_{i=1}^{n} \frac{1}{\pi_i} \left\{ y_i - p(\hat{B}/z_i) \right\} = 0$$

$$\frac{\partial}{\partial B_1} = \sum_{i=1}^{n} \frac{1}{\pi_i} \left\{ y_i z_{1i} - \frac{e^{\hat{B}'Z_i}}{1+e^{\hat{B}'Z_i}} z_{1i} \right\} = \sum_{i=1}^{n} \frac{1}{\pi_i} \left\{ y_i - p(\hat{B}/z_i) \right\} z_{1i} = 0$$

.

.

$$\frac{\partial}{\partial B_q} = \sum_{i=1}^{n} \frac{1}{\pi_i} \left\{ y_i z_{qi} - \frac{e^{\hat{B}' Z_i}}{1 + e^{\hat{B}' Z_i}} z_{qi} \right\} = \sum_{i=1}^{n} \frac{1}{\pi_i} \left\{ y_i - p(\hat{B} / z_i) \right\} z_{qi} = 0$$

Son q+1 ecuaciones que pueden escribirse en forma vectorial como:

$$\sum_{i=1}^{n} \frac{1}{\pi_i} \left\{ y_i - p(\hat{B} / z_i) \right\} z_i = 0$$

Llamando $\hat{S}(\hat{B})$ a la suma anterior:

$$\hat{S}(\hat{B}) = 0$$

Nota: S, sin el sombrero, sería la misma función para toda la población y sin ponderar:

$$S(\hat{B}) = \sum_{i=1}^{N} \left\{ y_i - p(\hat{B} / z_i) \right\} z_i = 0 .$$

Desarrollando $\hat{S}(\hat{B})$ por Taylor de orden 1 y alrededor de B, el verdadero valor del vector de parámetros:

$$0 = \hat{S}(\hat{B}) \cong \hat{S}(B) + \left. \frac{\partial \hat{S}}{\partial \hat{B}} \right|_{\hat{B}=B} * (\hat{B} - B)$$

Despejando $\hat{S}(B)$:

$$\hat{S}(B) \cong - \frac{\partial \hat{S}}{\partial \hat{B}} (B) * (\hat{B} - B)$$

Tomando varianzas en ambos miembros en el límite y considerando que $\hat{S}$ es un estimador consistente de S:

$$\mathrm{Var}(\hat{S}(B)) \cong \left[\frac{\partial S}{\partial B} (B) \right] * \mathrm{Var}(\hat{B}) * \left[\frac{\partial S}{\partial B} (B) \right]^{|}$$

Despejando $\mathrm{Var}(\hat{B})$:

$$\mathrm{Var}(\hat{B}) \cong \left[\frac{\partial S}{\partial B} (B) \right]^{-1} * \mathrm{Var}(\hat{S}(B)) * \left[\left(\frac{\partial S}{\partial B} (B) \right)^{|} \right]^{-1}$$

Sea J = $\frac{\partial S}{\partial B}(B)$, es una matriz qxq

$$\mathrm{Var}(\hat{B}) \cong J^{-1} * \mathrm{Var}(\hat{S}(B)) * \left[J^{|} \right]^{-1}$$

Analicemos la igualdad anterior:

1) $\hat{S}(B)$ es un vector de dimensión q+1 donde el j-ésimo componente es de la forma:

$$\sum_{i=1}^{n} \frac{1}{\pi_i} y_i z_{ji} - \sum_{i=1}^{n} \frac{1}{\pi_i} p(\hat{B} / z_i) z_{ji}$$

Sean: $\mu_{ij} = y_i z_{ji}$; $\nu_{ij} = p(\hat{B} / z_i) z_{ji}$

Entonces la componente j de $\hat{S}(B)$ es:

$$\sum_{i=1}^{n}\frac{1}{\pi_i}\mu_{ij}-\sum_{i=1}^{n}\frac{1}{\pi_i}\nu_{ij}$$

La expresión anterior es una función de totales de variables, por lo tanto, su matriz de covarianzas, también tiene elementos que son totales.
Entonces, reemplazando B por $\hat{B}$ (consistente), se obtiene un estimador consistente de $Var(\hat{S}(B))$, al que se llama: $Var(\hat{S}(\hat{B}))$ (Sarndal, 1992 pp: 168).

2) J = $\frac{\partial S}{\partial B}(B)$ es una matriz cuyos elementos también son funciones de totales, por lo tanto, también existe un estimador consistente de J y de J^{-1}. Veamos por ejemplo la expresión del elemento (2,2) de la matriz J, es decir, $\frac{\partial S_2}{\partial B_1}(B)$:

Donde: S_2, es la segunda componente del vector S.

$$\frac{\partial S_2}{\partial B_1}(B)=-\sum_{i=1}^{N}z_{1i}\frac{\partial p(B/z_i)}{\partial B_1}=-\sum_{i=1}^{N}z_{1i}\frac{\left[e^{B'z_i}(1+e^{B'z_i})z_{1i}-e^{B'z_i}e^{B'z_i}z_{1i}\right]}{(1+e^{B'z_i})^2}=$$

$$=-\sum_{i=1}^{N}z_{1i}{}^2\frac{e^{B'z_i}}{(1+e^{B'z_i})^2}=-\sum_{i=1}^{N}z_{1i}{}^2\frac{e^{B'z_i}}{1+e^{B'z_i}}\frac{1}{1+e^{B'z_i}}=-\sum_{i=1}^{N}z_{1i}{}^2p(B/z_i)(1-p(B/z_i))$$

Sean $\eta_i=z_{1i}{}^2p(B/z_i)(1-p(B/z_i))$, entonces:

$\partial S_2/\partial B_1=-\sum_{i=1}^{N}\eta_i$, es un total.

El estimador se obtendrá reemplazando cada elemento de J por su estimador de HT y reemplazando B por $\hat{B}$ que es consistente.

De 1) y 2) resulta que el estimador de la aproximación de la $Var(\hat{B})$, dado por:

$$\hat{var}(\hat{B})=\hat{J}^{-1}*var(\hat{S}(\hat{B}))*\left[\hat{J}'\right]^{-1}$$

$$\hat{var}(\hat{B})=\hat{J}^{-1}*var(\hat{S}(\hat{B}))*\hat{J}^{-1}$$

Donde: J=J´ (pues es una matriz simétrica)

al ser un producto de estimadores consistentes es consistente.