UNIVERSIDAD DE LA REPÚBLICA
Facultad de Ciencias Económicas y de Administración
Licenciatura en Estadística

Estimación en dominios

Juan Pablo Ferreira Neira

Tutor: Guillermo Zoppolo

Mayo de 2011

Índice general

Índice de cuadros

Capítulo 1

Introducción

Las encuestas por muestreo no solo se utilizan para obtener información a nivel del conjunto de la población total. Es posible, a su vez, realizar estimaciones para subconjuntos específicos de la población, a los que se les denomina *dominios*. Los dominios pueden estar definidos por áreas geográficas, grupos demográficos, u otro tipo de subpoblaciones. Por ejemplo, en una encuesta a personas, dichos dominios pueden estar definidos por grupos de edad, sexo, nivel educativo y región geográfica de residencia.

En la práctica, los tamaños de muestra, son más que suficientes para obtener una buena precisión para el total de la población. Sin embargo, cuando se requieren estimaciones para determinados dominios de interés puede ocurrir que se cuente con muy pocas observaciones (o incluso ninguna), de manera que no sea posible obtener precisiones aceptables utilizando los estimadores usuales. De lo anterior surge una pregunta muy frecuente en estimación en dominios, ¿es suficiente el tamaño de muestra en el dominio para obtener precisiones aceptables? En los últimos cuarenta años diferentes técnicas y recomendaciones han sido realizadas para intentar resolver dicho problema; desde recomendaciones para tener en cuenta los dominios en el diseño muestral (para controlar tamaños de muestra y precisiones) hasta diferentes métodos de estimación.

Los dominios pueden clasificarse según su tamaño relativo. Por ejemplo, Purcell y Kish (1979) distinguen cuatro categorías: grande, menor, pequeño y raro. Un dominio es considerado grande si representa mas del 10 % de la población, menor si representa entre el 1 % y el 10 %, pequeño entre 0,01 % y 1 %, y raro si su tamaño relativo es menor que 0,01 %. Estevao y Särndal (2004) no utilizan una clasificación tan específica, y distinguen dos casos, grandes y menores. Un dominio es considerado grande si representa más del 10 % de la población, y menor en otro caso.

De forma de complementar la clasificación realizada por Purcell y Kish, es necesario tener en cuenta el tamaño de la muestra, por ejemplo si un dominio representa el 5 % de la población y el tamaño de muestra total es 30.000 y se selecciona una muestra bajo un muestreo aleatorio simple, el tamaño de muestra esperado en el dominio es $30000 \times 0{,}05 = 1500$. En este caso, el tamaño de muestra esperado es lo suficientemente grande para realizar estimaciones con un nivel aceptable

de precisión.

El ejemplo anterior introduce el tamaño de muestra esperado para clasificar a los dominios. J.N.K Rao (2003) define un dominio como pequeño si el tamaño de muestra efectivo en el dominio no es lo suficientemente grande para realizar estimaciones con una precisión aceptable utilizando estimadores tradicionales.

Singh et al. (1994) clasifican a los dominios en planeados y/o no planeados. A la hora de definir el diseño muestral, pueden tenerse en cuenta los dominios para los cuales se requiere brindar estimaciones y se pueden calcular tamaños de muestra específicos para cumplir determinados requisitos de precisión. Por ejemplo, si el diseño muestral es estratificado, los dominios pueden coincidir con los estratos y se denominan dominios planeados (o identificados). El hecho de poder identificar el dominio a priori, no solo permite calcular el tamaño de muestra, sino que a su vez el mismo se puede controlar.

Un muestreo estratificado donde los estratos coinciden con los dominios junto con una asignación eficiente de la muestra entre los estratos, puede producir buenos resultados. Por ejemplo, utilizando la asignación de Bankier (1988), *Power Allocation*, la cual es un compromiso entre la asignación *óptima de Neyman* y una precisión constante en los estratos. La asignación *óptima de Neyman* proporciona excelentes precisiones para el total de la población y para los estratos grandes, en tanto para los estratos pequeños las precisiones pueden llegar a ser muy pobres. La *Power Allocation* intenta resolver dicho problema asignando tamaños de muestra más grandes a los estratos pequeños, ocasionando una reducción en las precisiones en los estratos grandes y para el total de la población, en comparación con la asignación óptima.

Los dominios no planeados son aquellos que no se tuvieron en cuenta en la especificación del diseño muestral, ya sea por no tener disponible una variable en el marco muestral que identifique a los individuos en los dominios de interés, por ser poco prácticos a la hora de definir los estratos en el diseño muestral, o por ser requeridos después de tomada la muestra. En la práctica, los dominios no planeados generalmente intersectan los estratos del diseño muestral. La diferencia principal entre un dominio no planeado y planeado, radica que en el primero, el tamaño de muestra es aleatorio, y en el segundo el tamaño de muestra es controlado y puede ser fijo si el diseño lo permite. Las precisiones de las estimaciones en los dominios no planeados, solo pueden ser conocidas una vez seleccionada la muestra, en donde el tamaño de muestra efectivo en el dominio junto con el método de estimación utilizado toman un rol determinante.

En el problema de estimación en dominios pueden distinguirse dos grandes enfoques: los basados en el diseño y los basados en modelos.

En el enfoque basado en el diseño la única aleatoriedad proviene del diseño muestral en donde los ponderadores muestrales tienen un rol crucial. Se buscan estimadores consistentes en el diseño y *very nearly design unbiased*, término utilizado en Estevao y Särndal (2004), para estimadores $\hat{\theta}$ de θ que cumplen que, $\frac{E(\hat{\theta})-\theta}{\sqrt{V(\hat{\theta})}}$ es $O(n^{-1/2})$.

Bajo el enfoque de los estimadores basados en modelos se agrega una aleatoriedad que proviene del modelo propuesto. Bajo este enfoque, los estimadores generalmente poseen una varianza pequeña, pero suelen ser sesgados si los supuestos del modelo no se cumplen. Si el sesgo es grande, dominará a la expresión del error cuadrático medio del estimador, y los intervalos de confianza basados en su cálculo no tendrán el nivel de cobertura deseado.

La disponibilidad de información auxiliar potente, como la proveniente de censos o registros administrativos es determinante para realizar estimaciones bajo ambos enfoques. La razón de la incorporación de información auxiliar en el proceso de estimación es evidente: mejorar la precisión de los estimadores siempre y cuando la información auxiliar disponible sea buena. Diferentes tipos de información auxiliar pueden ser utilizadas bajo los dos enfoques.

Los estimadores calibrados y de regresión tienen un rol preponderante en la estimación basada en el diseño muestral. Ambos utilizan información auxiliar en el proceso de estimación. La diferencia entre ellos radica en que la primer clase de estimadores no especifica ningún modelo explícito, mientras que los estimadores de regresión se apoyan en un modelo dado. Por otro lado, los estimadores dependientes de un modelo, utilizan la información de la variable de interés de otros dominios o de la población en su conjunto a través de un modelo que supone un vínculo con el dominio de interés.

Adicionalmente, los estimadores pueden ser clasificados como *directos* e *indirectos*. Según Schaible (1996) un estimador es *directo* si utiliza valores de la variable de interés solo del período de referencia y únicamente de las unidades de la muestra incluidas en el dominio de interés. Un caso simple de un estimador directo es el estimador *Horvitz - Thompson*. En las estimaciones asistidas por modelos, un estimador es directo si el modelo que lo asiste es específico del dominio. Por otro lado, un estimador *indirecto* utiliza valores de la variable de interés de otros individuos no pertenecientes al dominio de interés, o de otros períodos de tiempo. El objetivo es reducir la variabilidad de los estimadores cuando el tamaño de muestra efectivo en el domino es reducido. Por ejemplo, en la estimación asistida por modelos, un estimador es indirecto si el modelo que asiste al estimador, es definido a nivel de toda la población. Por otro lado, todos los estimadores basados en modelos son indirectos, por ejemplo, el estimador sintético, el cual puede ser calculado inclusive si el tamaño de muestra en el dominio es nulo.

Finalmente, independientemente del método utilizado para brindar estimaciones en los dominios

de interés, es importante (y requerido), que se cumpla la propiedad de aditividad. Si los dominios particionan a la población objeto de estudio, la suma de las estimaciones realizadas para cada uno de estos dominios, deben coincidir con la estimación realizada para el total de la población (bajo el mismo método de estimación).

Capítulo 2

Conceptos básicos de estimación en dominios

2.1. Notación

En este capítulo se presenta la notación a seguir y las herramientas básicas para la estimación en dominios. La notación y el contenido se basan en Särndal et al. (1992).

Sea $U = \{1, ..., k, ..., N\}$ la población finita objeto de estudio de tamaño N. De U se toma una muestra probabilística s, de tamaño n_s, según un diseño $p(.)$. El individuo k es incluido en la muestra con una probabilidad $\pi_k = P\{k \in s\} > 0 \; \forall k \in U$ (diseño aleatorio). El inverso de la probabilidad de inclusión $a_k = 1/\pi_k$ es el ponderador muestral o ponderador del diseño del individuo k. Los individuos k y l son incluidos en la muestra con probabilidad $\pi_{kl} = P\{k \;\; y \;\; l \in s\} > 0 \; \forall k \neq l \in U$ (diseño medible). La variable de interés se denota como y, y y_k es el valor que toma la variable y para el individuo k. En el contexto de los estimadores basados en el diseño la variable de interés se considera fija pero con valores desconocidos.

En lo que sigue no se consideran problemas de medición, no respuesta, ni marcos imperfectos.

En general, el objetivo es estimar el total de la variable de interés, $t = \sum_{k \in U} y_k$, o su media poblacional, $\bar{y}_U = \sum_{k \in U} y_k / N$.

En estimación en dominios, el interés recae en estimar totales o medias de la variable de interés y en subconjuntos de la población U. Sin pérdida de generalidad, supongamos que U es particionada en D dominios $U_1, ..U_d, ..., U_D$ y sea N_d el tamaño de U_d, el cual puede ser o no conocido.

Entonces, se tienen las siguientes ecuaciones

$$U = \bigcup_{d=1}^{D} U_d \quad \text{y} \quad N = \sum_{d=1}^{D} N_d. \tag{2.1}$$

Sea s_d el subconjunto de la muestra s perteneciente al dominio U_d, o sea, $s_d = s \cap U_d$ y n_{s_d} el tamaño de s_d.

De manera análoga a la ecuación (2.1) se tiene

$$s = \bigcup_{d=1}^{D} s_d \quad \text{y} \quad n_s = \sum_{d=1}^{D} n_{s_d}. \tag{2.2}$$

El diseño muestral puede estar basado en el conocimiento de los dominios, los cuales pueden ser considerados a la hora del diseño muestral y ser definidos como estratos (dominios planeados). En ese caso el tamaño de muestra en el dominio, n_{s_d}, es controlado y puede ser fijo (si el diseño lo permite). En tanto, si el dominio no es considerado en el diseño muestral (dominio no planeado), el tamaño de la muestra es aleatorio (y en algunas circunstancias puede ser nulo).

El tamaño absoluto de un dominio N_d, o su tamaño relativo, $P_d = N_d/N$, pueden ser vistos como un total y una media poblacional, respectivamente. En este sentido, es útil definir una variable indicadora de pertenencia al dominio, δ_d, que según el individuo k vale

$$\delta_{dk} = \begin{cases} 1 & \text{si} \quad k \in U_d \\ 0 & \text{si} \quad k \notin U_d \end{cases}. \tag{2.3}$$

Luego, se tiene que

$$\sum_{k \in U} \delta_{dk} = \sum_{k \in U_d} 1 = N_d, \tag{2.4}$$

y

$$\sum_{k \in U} \delta_{dk}/N = N_d/N = P_d. \tag{2.5}$$

El tamaño de muestra en el dominio puede escribirse como

$$n_{s_d} = \sum_{k \in U} \delta_{dk} I_k = \sum_{k \in U_d} I_k, \tag{2.6}$$

donde I_k es la variable indicadora de pertenencia a la muestra, o sea, $I_k = 1$ si $k \in s$ y 0 en otro caso.

El tamaño de muestra esperado en el dominio es

$$E(n_{s_d}) = \sum_{k \in U} \delta_{dk} \pi_k = \sum_{k \in U_d} \pi_k. \tag{2.7}$$

Ejemplo 2.1.1 Bajo un diseño simple (SI) de tamaño n de una población de N individuos, el tamaño de muestra esperado en el dominio U_d, es

$$E_{SI}(n_{s_d}) = N_d n/N = f P_d N, \tag{2.8}$$

donde $f = n/N$.

En la ecuación (2.8), la tasa de muestreo, f, el tamaño relativo del dominio, P_d, y el tamaño de la población, N, determinan el tamaño de muestra esperado en el dominio. □

Una herramienta útil para la estimación en dominios es la variable extendida y_d

$$y_{dk} = \begin{cases} y_k & \text{si} \quad k \in U_d \\ 0 & \text{si} \quad k \notin U_d \end{cases}. \tag{2.9}$$

En otras palabras, $y_{dk} = \delta_{dk} y_k$. Entonces, el total de la variable y en el dominio U_d, $t_d = \sum\limits_{k \in U_d} y_k = \sum\limits_{k \in U} y_{dk}$, se puede estimar como el total de la variable poblacional y_d.

Observación 2.1.1 Si $\delta_{dk} = 1 \ \forall k \in U$ entonces $y_d = y$ y $U_d = U$. □

2.2. El estimador π de Horvitz-Thompson

La primera aproximación para la estimación en dominios es utilizar el estimador π, que se basa solamente en las probababilidades de inclusión en la muestra.

El estimador π del total de la variable y en el dominio U_d, viene dado por

$$\hat{t}_{d\pi} = \sum_{k \in s_d} \frac{y_k}{\pi_k} = \sum_{k \in s_d} a_k y_k \tag{2.10}$$

$$= \sum_{k \in s} \frac{y_{dk}}{\pi_k} = \sum_{k \in s} a_k y_{dk}. \tag{2.11}$$

Luego se tiene que

$$E(\hat{t}_{d\pi}) = t_d \tag{2.12}$$

y

$$\begin{aligned} V(\hat{t}_{d\pi}) &= \sum_{k\in U_d}\sum_{l\in U_d}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} && (2.13)\\ &= \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_{dk}}{\pi_k}\frac{y_{dl}}{\pi_l}, && (2.14)\end{aligned}$$

donde $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$.

El estimador π de la varianza anterior es

$$\begin{aligned} \hat{V}(\hat{t}_{d\pi}) &= \sum_{k\in s_d}\sum_{l\in s_d}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} && (2.15)\\ &= \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_{dk}}{\pi_k}\frac{y_{dl}}{\pi_l}, && (2.16)\end{aligned}$$

El estimador π, por construcción, es un estimador directo, ya que solamente utiliza los valores de la variable de interés del dominio U_d.

Observación 2.2.1 El estimador $\hat{t}_{d\pi}$ cumple la propiedad de aditividad, o sea, $\hat{t}_\pi = \sum_{d=1}^{D}\hat{t}_{d\pi}$. □

Observación 2.2.2 De aquí en adelante, se anota $\sum_{k\in A} = \sum_A$ y $\sum_{k\in A}\sum_{l\in A} = \sum\sum_A \ \ \forall A \subseteq U$. □

Ejemplo 2.2.1 Bajo un diseño SI de n elementos tomados de N, el estimador π de la variable y en el dominio U_d es

$$\hat{t}_{d\pi} = \sum\nolimits_{s_d} a_k y_k = (N/n)\sum\nolimits_{s_d} y_k = N\bar{y}_{s_d}, \tag{2.17}$$

donde $\bar{y}_{s_d}$ es la media muestral en el dominio de interés U_d.

La varianza definida en (2.13) toma la forma de

$$\begin{aligned} V_{SI}(\hat{t}_{d\pi}) &= N^2\frac{1-f}{n}\frac{(N_d-1)S^2_{y_{U_d}} + N_dQ_d\bar{y}^2_{U_d}}{N-1}\\ &\doteq N^2\frac{1-f}{n}P_d(S^2_{y_{U_d}} + Q_d\bar{y}^2_{U_d}), && (2.18)\end{aligned}$$

donde $f = n/N$ es la tasa de muestreo, $\bar{y}_{U_d} = \sum_{U_d} y_k/N_d$ y $S^2_{y_{U_d}} = (N_d-1)^{-1}\sum_{U_d}(y_k-\bar{y}_{U_d})^2$ son la media y varianza poblacional en el dominio U_d, $P_d = N_d/N$ es el tamaño relativo del dominio U_d en la población y $Q_d = 1 - P_d$. □

Cuando el tamaño del dominio N_d es conocido, se puede utilizar el estimador de Hayek dado por

$$\tilde{t}_d = N_d\tilde{y}_{s_d}, \tag{2.19}$$

donde $\tilde{y}_{s_d} = \sum_{s_d} a_k y_k/\hat{N}_d$, con $\hat{N}_d = \sum_{s_d} a_k = \sum_s \delta_{dk}a_k$.

El estimador $\tilde{t}_d$ es un caso especial del estimador de razón, su varianza aproximada viene dada por

$$AV(\tilde{t}_d) = \sum\sum\nolimits_{U_d} \Delta_{kl}\left(\frac{y_k - \bar{y}_{U_d}}{\pi_k}\right)\left(\frac{y_l - \bar{y}_{U_d}}{\pi_l}\right), \tag{2.20}$$

y un estimador de la varianza es

$$\hat{V}(\tilde{t}_d) = \left(\frac{N_d}{\hat{N}_d}\right)^2 \sum\sum\nolimits_{s_d} \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_k - \tilde{y}_{s_d}}{\pi_k}\right)\left(\frac{y_l - \tilde{y}_{s_d}}{\pi_l}\right). \tag{2.21}$$

Observación 2.2.3 Las ecuaciones (2.20) y (2.21) pueden escribirse como sumas en U y s, utilizando las correspondientes variables indicadoras δ_d . □

Observación 2.2.4 El estimador $\tilde{t}_d$, usa como información auxiliar a la variable indicadora de pertenencia al dominio δ_d y su total en U, N_d . □

Observación 2.2.5 Para estimar la media del dominio $\bar{y}_{U_d} = t_d/N_d$, sea o no conocido el tamaño del dominio N_d, es preferible usar $\tilde{y}_{s_d}$. □

Ejemplo 2.2.2 Bajo un SI de tamaño n, el estimador de la ecuación (2.19) es

$$\tilde{t}_d = N_d\tilde{y}_{s_d} = N_d\bar{y}_{s_d}. \tag{2.22}$$

Su varianza aproximada definida en la ecuación (2.20) viene dada por

$$\begin{aligned} AV_{SI}(\tilde{t}_d) &= N^2\frac{1-f}{n}\frac{(N_d-1)S^2_{y_{U_d}}}{N-1} \\ &\doteq N^2\frac{1-f}{n}P_dS^2_{y_{U_d}}. \end{aligned} \tag{2.23}$$

El estimador de la varianza de $\tilde{t}_d$ es, según (2.21)

$$\begin{aligned} \hat{V}_{SI}(\tilde{t}_d) &= \left(\frac{N_d}{\hat{N}_d}\right)^2 N^2 \frac{1-f}{n} \frac{(n_{s_d}-1)S^2_{y_{s_d}}}{n-1} \\ &\doteq N_d^2 \left(\frac{1}{n_{s_d}} - \frac{1}{\hat{N}_d}\right) S^2_{y_{s_d}}, \end{aligned} \tag{2.24}$$

donde $S^2_{y_{s_d}} = (n_{s_d}-1)^{-1} \sum_{s_d} (y_k - \bar{y}_{s_d})^2$ es la varianza muestral en el dominio U_d . □

Observación 2.2.6 El cociente de las ecuaciones (2.18) y (2.23) permite analizar la eficiencia relativa de los estimadores $\hat{t}_{d\pi}$ y $\tilde{t}_d$

$$\frac{V_{SI}(\hat{t}_{d\pi})}{AV_{SI}(\tilde{t}_d)} \doteq 1 + \frac{Q_d}{(cv_{y_{U_d}})^2}, \tag{2.25}$$

donde $cv_{y_{U_d}} = S_{y_{U_d}}/\bar{y}_{U_d}$ es el coeficiente de variación de la variable de interés y en el dominio U_d. Por ejemplo, si $cv_{yU_d} = 0{,}5$; la varianza del estimador $\hat{t}_{d\pi}$, es aproximadamente cinco veces mayor que la del estimador $\tilde{t}_d$, cuando el tamaño del dominio es un porcentaje pequeño de la población (Q_d es casi 1). En cambio si el dominio es el 50 % de la población ($Q_d = 0{,}5$) y $cv_{yU_d} = 0{,}5$; la ineficiencia del $\hat{t}_{d\pi}$ es menos pronunciada, pero aún considerable (su varianza es cerca de tres veces más grande). □

Ejemplo 2.2.3 Supongamos que el dominio de interés U_d puede ser identificado a priori y el tamaño de muestra en el mismo puede ser fijo. Es esperable que lo anterior derive en un estimador con menor varianza en comparación con otro estimador, para el cual el tamaño de muestra no fue controlado. Supongamos el caso de un diseño SI.

El estimador π para el total t_d, es

$$N_d \bar{y}_{s_d} = N_d \sum\nolimits_{s_d} y_k/n_d,$$

y su varianza viene dada por

$$V_{SI}(N_d \bar{y}_{s_d}) = N_d^2 \left(\frac{1}{n_d} - \frac{1}{N_d}\right) S^2_{y_{U_d}}. \tag{2.26}$$

Comparemos la varianza anterior con la varianza aproximada del estimador $\tilde{t}_d$, de la ecuación (2.23) escrita como

$$AV_{SI}(\tilde{t}_d) \doteq N_d^2 \left(\frac{1}{n_d^0} - \frac{1}{N_d} \right) S^2_{y_{U_d}}, \tag{2.27}$$

donde $n_d^0 = nN_d/N$ es el tamaño de muestra esperado en el dominio U_d.

Entonces, si el tamaño de muestra en el dominio, n_d fijo, es igual al tamaño de muestra esperado en el dominio $E(n_{s_d}) = n_d^0$, las varianzas de los dos estimadores son aproximadamente iguales. Bajo la aproximación anterior, si el tamaño de muestra en el dominio no es controlado, no hay pérdida de precisión, siempre y cuando el tamaño del dominio, N_d, sea conocido. □

Siguiendo con el ejemplo anterior, a pesar de que el estimador $\tilde{t}_d = N_d \tilde{y}_{s_d}$, es aproximadamente igual de preciso que el estimador $N_d \bar{y}_{s_d}$, es esperable que el tamaño de muestra aleatorio en el dominio contribuya a aumentar la varianza del estimador. Para determinar lo anterior, es necesario encontrar una mejor aproximación a la varianza del estimador $\tilde{t}_d$ que la definida en la ecuación (2.26). En este sentido es útil condicionar al tamaño de muestra obtenido en el dominio, n_{s_d}.

Sea A_d el evento $\{n_{s_d} \geq 1\}$. Si el tamaño de muestra total, n, es considerablemente grande, es esperable que la probabilidad del evento A_d se encuentre cercana a uno, inclusive si el tamaño relativo del dominio, P_d, es pequeño. Para un valor fijo de n_{s_d}, tal que $n_{s_d} \geq 1$, la muestra, $s_d = s \cap U_d$, se comporta como un muestreo aleatorio simple de tamaño n_{s_d} de U_d.

Por lo tanto, para el estimador $\tilde{t}_d = N_d \bar{y}_{s_d}$, se tiene que

$$E_{SI}(\tilde{t}_d | A_d, n_{s_d}) = t_d, \tag{2.28}$$

$$V_{SI}(\tilde{t}_d | A_d, n_{s_d}) = N_d^2 \left(\frac{1}{n_{s_d}} - \frac{1}{N_d} \right) S^2_{y_{U_d}}. \tag{2.29}$$

El estimador $\tilde{t}_d$, es condicionalmente insesgado, dado cualquier tamaño de muestra en el dominio, siempre y cuando $n_{s_d} \geq 1$.

Promediando sobre todos los valores $n_{s_d} \geq 1$ se obtiene que

$$E_{SI}(\tilde{t}_d | A_d) = t_d, \tag{2.30}$$

$$V_{SI}(\tilde{t}_d | A_d) = N_d^2 \left(E\left(\frac{1}{n_{s_d}} | A_d \right) - \frac{1}{N_d} \right) S^2_{y_{U_d}}, \tag{2.31}$$

en donde para obtener (2.31), se utilizó que

$$V\left[E_{SI}(\tilde{t}_d|A_d, n_{s_d})\right] = V(t_d|A_d) = 0,$$

donde $V(.)$ denota la varianza respecto a la distribución de n_{s_d}. O sea, dado que la muestra contiene al menos un elemento, el estimador, $\tilde{t}_d$, es insesgado para t_d, bajo un diseño SI.

De todas formas, las ecuaciones (2.30) y (2.31) se encuentran condicionadas por el evento A_d. Supongamos que el tamaño de muestra n, es lo suficientemente grande, de maneara que es casi seguro que el evento A_d ocurra. Entonces, se concluye por las ecuaciones (2.30) y (2.31), que el estimador $\tilde{t}_d$, es insesgado para t_d con varianza incondicional dada por

$$V_{SI}(\tilde{t}_d) = N_d^2 \left[E_{SI}\left(\frac{1}{n_{s_d}}\right) - \frac{1}{N_d}\right] S_{y_{U_d}}. \tag{2.32}$$

La ecuación (2.32) se obtiene asumiendo que la probabilidad $P(n_{s_d} = 0) = 0$.

Luego, usando la aproximación del desarrollo de Taylor de segundo orden, se obtiene que

$$E_{SI}\left(\frac{1}{n_{s_d}}\right) \doteq \frac{1}{n_d^0} + \frac{(1-f)(1-P_d)}{(n_d^0)^2}, \tag{2.33}$$

donde $n_d^0 = E(n_{s_d}) = nN_d/N = nP_d$.

Por (2.32) y (2.33) se obtiene

$$V_{SI}(\tilde{t}_d) = N_d^2 \left(\frac{1}{n_d^0} - \frac{1}{N_d}\right)\left(1 + \frac{Q_d}{n_d^0}\right) S_{y_{U_d}}^2, \tag{2.34}$$

con $Q_d = 1 - P_d$.

Comparando la varianza (2.34) respecto a la varianza del estimador $N_d\bar{y}_{s_d}$ cuando el tamaño de muestra es fijo de la ecuación (2.26), se obtiene

$$\frac{V_{SI}(\tilde{t}_d)}{V_{SI}(N_d\bar{y}_{s_d})} = 1 + \frac{Q_d}{n_d^0}. \tag{2.35}$$

Si el tamaño de muestra n, es considerablemente más grande que $n_d^0 = nP_d$, la expresión anterior es aproximadamente $1 + 1/n_d^0$. Por lo tanto, existe una pérdida de precisión no despreciable a causa de no poder controlar el tamaño de la muestra en el dominio cuando el tamaño de muestra esperado es pequeño.

Finalmente, la varianza condicional de la ecuación (2.29) es estimada de manera insesgada (dado $n_{s_d} \geq 2$) por

$$\hat{V}^*_{SI} = N_d^2 \left(\frac{1}{n_{s_d}} - \frac{1}{N_d} \right) S^2_{y_{s_d}}. \tag{2.36}$$

Este estimador de la varianza condicionada, coincide básicamente con la ecuación (2.24). La diferencia entre $1/\hat{N}_d$ y $1/N_d$, no tiene repercusiones importantes en la práctica.

Conclusiones:

El tamaño de muestra aleatorio en el dominio introduce una fuente de variabilidad adicional en los estimadores y generalmente los mismos suelen ser menos eficientes que aquellos en donde el tamaño de muestra es controlado. Dicha pérdida de precisión es despreciable a medida que el tamaño de muestra esperado en el dominio aumenta.

El uso de información auxiliar es de vital importancia para producir estimadores con mayor precisión. Hasta ahora, la única información auxiliar utilizada fue la variable indicadora de pertenencia al dominio que implica conocer el tamaño del dominio N_d ($\sum_U \delta_{dk} = N_d$) para construir el estimador $\tilde{t}_d$, el cual tiene una menor varianza respecto al estimador π (que no utiliza ningún tipo de información auxiliar). A su vez, el desempeño del estimador $\tilde{t}_d$, mejora considerablemente respecto al estimador π, en dominios pequeños. Por lo tanto, la disponibilidad de información auxiliar potente es esencial en aquellos dominios en donde el tamaño de muestra es reducido.

Capítulo 3

Estimador de regresión generalizado

3.1. Introducción

El uso de algún tipo de información auxiliar es fundamental para la obtención de estimadores con mayor precisión que el estimador π, sobre todo cuando el tamaño de muestra esperado en el dominio es pequeño.

Las variables auxiliares pueden ser utilizadas a la hora de definir el diseño muestral o posteriormente en la etapa de estimación. Si las variables auxiliares se encuentran en el marco muestral (son conocidas para todos los individuos), las mismas pueden ser utilizadas para definir probabilidades de inclusión y/o para la construcción de estratos.

En la etapa de estimación, la información auxiliar puede ser conocida solamente a nivel de totales. Dichos totales pueden provenir de registros administrativos o de otras encuestas. Los estimadores de regresión lineal, utilizan la información auxiliar por medio de un modelo de regresión que asiste al estimador de forma de producir estimaciones más eficientes.

A continuación se hace una breve reseña de los estimadores de regresión lineal y su aplicación al problema de estimación en dominios.

Supongamos que el interés se centra en estimar el total poblacional, $t = \sum_U y_k$. Para ello se selecciona una muestra, s, bajo un diseño $p(.)$ medible. El valor y_k de la variable de interés es observado para todos los individuos incluidos en la muestra, por otro lado, para aquellos individuos que no han sido seleccionados en la muestra, el valor y_k es desconocido, pero se puede encontrar un valor μ_k que se aproxime al valor desconocido y_k para todos los individuos de la población.

Entonces se puede reescribir el total poblacional $t = \sum_U y_k$, de la forma

$$t = \sum\nolimits_U \mu_k + \sum\nolimits_U (y_k - \mu_k),$$

en donde el segundo sumando de la ecuación, $\sum_U (y_k - \mu_k)$, es desconocido y requiere ser estimado. Entonces, se deben tomar dos decisiones:

1. Se debe elegir un estimador para la suma $\sum_U (y_k - \mu_k)$.
2. Como elegir los valores μ_k cercanos a los valores y_k. Esta decisión consta de dos partes (i) el modelo que relacione a y_k con μ_k y (ii) la técnica a utilizar para ajustar dicho modelo.

La opción usual para 1. es el estimador π

$$\hat{t} = \sum\nolimits_U \mu_k + \sum\nolimits_s a_k (y_k - \mu_k). \tag{3.1}$$

En la construcción de los valores μ_k de 2. es importante considerar la información auxiliar disponible. Consideremos $\mathbf{x}$ un vector de información auxiliar de dimensión $J \geq 1$ y $\mathbf{x}_k$ el valor que toma $\mathbf{x}$ para el individuo k. Supongamos que $\mathbf{x}_k$ se encuentra disponible para todos los individuos de la población. Los valores predichos $\hat{y}_k$ son obtenidos utilizando la información auxiliar, ajustando un modelo, m, de forma que $E_m(y_k|\mathbf{x}_k, \boldsymbol{\beta}) = f(\mathbf{x}_k|\boldsymbol{\beta})$, donde E_m es la esperanza bajo el modelo m, $f(.\,|\boldsymbol{\beta})$ es una función conocida y $\boldsymbol{\beta}$ un vector de parámetros desconocidos. El modelo es *lineal* si la función $f(\mathbf{x}_k\,|\boldsymbol{\beta}) = \mathbf{x'}_k\boldsymbol{\beta}$, en otro caso es *no lineal*.

El rol del modelo m es simplemente describir el comportamiento de la población. En ningún momento se supone que la población es realmente generada por el modelo. Por lo tanto las conclusiones que se obtengan sobre los parámetros de la población serán independientes de la validez del modelo.

Utilizando los datos de la muestra $\{(y_k, \mathbf{x}_k) : k \in s\}$, se obtiene una estimación del vector de parámetros $\boldsymbol{\beta}$, la cual se denota como $\hat{\mathbf{B}}$. Posteriormente se calculan los valores predichos $\hat{y}_k = f(\mathbf{x}_k|\hat{\mathbf{B}})$ para todos los individuos de la población. Finalmente utilizando $\hat{y}_k$ y el estimador π de la suma $\sum_U (y_k - \hat{y}_k)$, se obtiene el estimador de regresión generalizado

$$\hat{t}_{greg} = \sum\nolimits_U \hat{y}_k + \sum\nolimits_s a_k (y_k - \hat{y}_k). \tag{3.2}$$

El estimador $\hat{t}_{greg}$ es aproximadamente insesgado sin tener en cuenta si el modelo m elegido es "verdadero". Así, el estimador $\hat{t}_{greg}$ es un estimador *asistido* por el modelo y no basado en el modelo.

Si el modelo que asiste a el estimador es lineal, $E_m(y_k) = \mathbf{x'}_k\boldsymbol{\beta},\ V(y_k) = c_k\ \forall k \in U$ y se utiliza el método de mínimos cuadrados generalizados para obtener una estimación de $\mathbf{B}$, se obtiene

$$\mathbf{B} = \left(\sum\nolimits_U \mathbf{x}_k \mathbf{x'}_k / c_k\right)^{-1} \sum\nolimits_U \mathbf{x}_k y_k / c_k, \tag{3.3}$$

donde para la estimación de $\mathbf{B}$, es necesario conocer los valores de la variable de interés y para todos los individuos de la población U.

Luego, el estimador π de $\hat{\mathbf{B}}$ viene dado por

$$\hat{\mathbf{B}} = \left(\sum_s a_k \mathbf{x}_k \mathbf{x'}_k / c_k\right)^{-1} \sum_s a_k \mathbf{x}_k y_k / c_k. \tag{3.4}$$

Finalmente, utilizando los valores ajustados por el modelo, $\hat{y}_k = \mathbf{x'}_k \hat{\mathbf{B}}$, para toda la población y siguiendo la ecuación (3.2) se obtiene el estimador *greg* lineal

$$\hat{t}_{greg} = \sum_U \mathbf{x'}_k \hat{\mathbf{B}} + \sum_s a_k e_k, \tag{3.5}$$

donde $e_k = y_k - \mathbf{x'}_k \hat{\mathbf{B}}$ son los residuos muestrales.

El estimador de regresión puede expresarse como

$$\hat{t}_{greg} = \hat{t}_\pi + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \hat{\mathbf{B}}, \tag{3.6}$$

donde $\mathbf{t}_x = \sum_U \mathbf{x}_k$, es el vector de totales de las variables auxiliares, $\hat{t}_\pi = \sum_s a_k y_k$ es el estimador π para el total de la variable de interés y $\hat{\mathbf{t}}_{x\pi} = \sum_s a_k \mathbf{x}_k$ es el estimador π del vector de totales de las variables auxiliares utilizadas para la construcción del estimador de regresión.

Notemos que para el cálculo del estimador de la ecuación (3.6), no es necesario disponer de la información auxiliar a nivel de todos los individuos de la población. Simplemente basta con conocer los totales para las variables auxiliares y relevar en la muestra los valores $\mathbf{x}_k$ para los elementos seleccionados. Lo anterior, tiene la ventaja, que dichos totales pueden no encontrarse disponibles en el marco muestral y los mismos pueden ser obtenidos de otras encuestas o de registros administrativos.

Observación 3.1.1 A su vez el estimador $\hat{t}_{greg}$ es un estimador homogéneo

$$\hat{t}_{greg} = \sum_s w_k y_k, \tag{3.7}$$

donde $w_k = a_k g_{ks}$,

$$g_{ks} = 1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right)' \hat{\mathbf{T}}^{-1} \mathbf{x}_k / c_k, \tag{3.8}$$

y $\hat{\mathbf{T}} = \sum_s a_k \mathbf{x}_k \mathbf{x}'_k / c_k$. □

Observación 3.1.2 Los ponderadores $w_k = a_k g_{ks}$ estiman sin error los totales poblacionales de las variables auxiliares utilizadas en el modelo, o sea, $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. □

Observación 3.1.3 El subíndice s, en los ponderadores g, hace referencia a que los mismos dependen de la muestra s. Para alivianar la notación, de aquí en adelante se omite explicitar dicha dependencia de la muestra. □

La varianza del estimador de regresión puede ser aproximada utilizando linealización de Taylor por

$$AV(\hat{t}_{greg}) = \sum \sum_U \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}, \tag{3.9}$$

donde $E_k = y_k - \mathbf{x}'_k \mathbf{B}$, son los residuos a nivel poblaciónal.

Entonces, un estimador de la varianza aproximada del estimador $\hat{t}_{greg}$ puede ser calculado utilizando los residuos muestrales $e_k = y_k - \mathbf{x}'_k \hat{\mathbf{B}}$.

$$\widehat{AV}(\hat{t}_{greg}) = \sum \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}. \tag{3.10}$$

Un estimador alternativo para la varianza de (3.9), propuesto por Särndal (ver Särndal et al. (1992) Cap 6) viene dado por

$$\hat{V}(\hat{t}_{greg}) = \sum \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_k e_k}{\pi_k} \frac{g_l e_l}{\pi_l}. \tag{3.11}$$

En la práctica, ambas expresiones producen similares resultados, pero en general se prefiere (3.11).

El modelo utilizado para la construcción del estimador de regresión es determinante para obtener una varianza pequeña. Si el modelo tiene un buen poder de ajuste, esto deriva en que los residuos E_k sean pequeños, dando como resultado que la varianza del estimador de regresión sea pequeña. Si todos los residuos son cero, o sea, $y_k = \mathbf{x}'_k \mathbf{B}, \ \ \forall k \in U$, la varianza del estimador de regresión es cero. Por lo tanto, si el modelo no ajusta bien, la varianza del estimador de regresión puede ser considerablemente grande. En cualquier caso, la varianza de (3.9) es estimada, de manera aproximadamente insesgada, por (3.11) o (3.10).

3.2. Estimadores de regresión en dominios

Como ya se dijo, en el problema de estimación en dominios, los estimadores pueden ser sustancialmente mejorados utilizando algún tipo de información auxiliar. La estimación básica en dominios se logra utilizando el estimador π, el cual solo utiliza información específica de la variable de interés y para aquellos individuos pertenecientes al dominio. Si se cuenta con información auxiliar se pueden utilizar estimadores de regresión, los cuales pueden ser fácilmente adaptados al problema de estimación en dominios.

Supongamos que se dispone de la siguiente información auxiliar:

- El vector de información auxiliar $\mathbf{x}$ es conocido para todos los individuos incluidos en la muestra, junto con la variable indicadora de pertenencia al dominio δ_d.
- Se conocen los totales del vector de información auxiliar a nivel del dominio U_d, o sea, $\mathbf{t}_{dx} = \sum_{U_d} \mathbf{x}_k = \sum_U \delta_{dk}\mathbf{x}_k$ es conocido.

Dentro de los modelos posibles pueden distinguirse:

- Casos en que el dominio de interés posee sus propias características y que estas difieren de la población en su conjunto. Así, se utiliza solo información de los individuos pertenecientes al dominio, y el modelo que asiste al estimador de regresión es específico del dominio.
- Casos en que el dominio de interés puede asimilarse a un subconjunto más amplio de la población, el cual incluye al dominio de interés, y el modelo que asiste al estimador es igual para todos los dominios de interés incluidos dentro de ese subconjunto.

Lo anterior da lugar a la clasificación en estimadores de regresión directos e indirectos, dependiendo si se utiliza o no información de la variable de interés y de los individuos no incluidos en el dominio U_d para la estimación de los parámetros del modelo que asiste al estimador.

3.2.1. Estimadores directos de regresión

Un estimador de regresión es directo, si para estimar el total del dominio, $t_d = \sum_{U_d} y_k$, es asistido por un modelo de regresión de la forma $E_m(y_k) = \mathbf{x'}_k\boldsymbol{\beta}_d, \;\; V_m(y_k) = c_k \;\; \forall k \in U_d$, donde el parámetro del modelo, es específico del dominio, $\mathbf{B}_d$. En este caso es estimado como en (3.4), pero sustituyendo la variable y por la variable extendida y_d y el vector de información auxiliar $\mathbf{x}$ por $\mathbf{x}_d$, donde

$$\mathbf{x}_{dk} = \delta_{dk}\mathbf{x}_k = \begin{cases} \mathbf{x}_k & \text{si} \quad k \in U_d \\ 0 & \text{si} \quad k \notin U_d \end{cases} . \tag{3.12}$$

Así

$$\hat{\mathbf{B}}_d = \left(\sum\nolimits_s a_k \mathbf{x}_{dk}\mathbf{x}'_{dk}/c_k\right)^{-1} \sum\nolimits_s a_k \mathbf{x}_{dk} y_{dk}/c_k \tag{3.13}$$

$$= \left(\sum\nolimits_{s_d} a_k \mathbf{x}_k \mathbf{x}'_k/c_k\right)^{-1} \sum\nolimits_{s_d} a_k \mathbf{x}_k y_k/c_k. \tag{3.14}$$

Los valores ajustados por el modelo para la variable de interés, $\hat{y}_{dk} = \mathbf{x}'_{dk}\hat{\mathbf{B}}_d$, y los residuos muestrales

$$e_{dk} = \begin{cases} y_k - \mathbf{x}'_k\hat{\mathbf{B}}_d & \text{si} \quad k \in U_d \\ 0 & \text{si} \quad k \notin U_d \end{cases}, \tag{3.15}$$

son utilizados para construir el estimador directo de regresión

$$\hat{t}_{d,gregD} = \sum\nolimits_U \hat{y}_{dk} + \sum\nolimits_s a_k e_{dk} \tag{3.16}$$

$$= \sum\nolimits_{U_d} \hat{y}_k + \sum\nolimits_{s_d} a_k e_k. \tag{3.17}$$

Observación 3.2.1 El estimador directo de regresión no cumple la propiedad de aditividad, $\hat{t}_{greg} \neq \sum_{d=1}^{D} \hat{t}_{d,gregD}$. □

El estimador $\hat{t}_{d,gregD}$, puede ser expresado como

$$\hat{t}_{d,gregD} = \hat{t}_{d\pi} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx\pi})'\hat{\mathbf{B}}_d, \tag{3.18}$$

donde $\mathbf{t}_{dx} = \sum_{U_d} \mathbf{x}_k = \sum_U \delta_{dk}\mathbf{x}_k = \sum_U \mathbf{x}_{dk}$ y $\hat{\mathbf{t}}_{dx\pi} = \sum_{s_d} a_k \mathbf{x}_k = \sum_s a_k \delta_{dk}\mathbf{x}_k = \sum_s a_k \mathbf{x}_{dk}$.

Observación 3.2.2 El estimador $\hat{t}_{d,gregD}$ es homogéneo

$$\hat{t}_{d,gregD} = \sum\nolimits_{s_d} a_k g_{dk} y_k = \sum\nolimits_s w_{dk} y_{dk}, \tag{3.19}$$

donde $w_{dk} = a_k g_{dk}$, con

$$g_{dk} = \delta_{dk} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx\pi})'\hat{\mathbf{T}}_d^{-1}\frac{\mathbf{x}_{dk}}{c_k}, \tag{3.20}$$

y $\hat{\mathbf{T}}_d = \sum_s a_k \mathbf{x}_{dk}\mathbf{x}'_{dk}/c_k$. □

Una aproximación para la varianza del $\hat{t}_{d,gregD}$, es

$$AV(\hat{t}_{d,gregD}) = \sum\sum\nolimits_{U} \Delta_{kl} \frac{E_{dk}}{\pi_k} \frac{E_{dl}}{\pi_l} \tag{3.21}$$

$$= \sum\sum\nolimits_{U_d} \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}, \tag{3.22}$$

donde

$$E_{dk} = \begin{cases} y_k - \mathbf{x}'_k \mathbf{B}_d & \text{si} \quad k \in U_d \\ 0 & \text{si} \quad k \notin U_d \end{cases}. \tag{3.23}$$

Un estimador de la varianza viene dado por

$$\hat{V}(\hat{t}_{d,gregD}) = \sum\sum\nolimits_{s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_{dk} e_{dk}}{\pi_k} \frac{g_{dl} e_{dl}}{\pi_l} \tag{3.24}$$

$$= \sum\sum\nolimits_{s_d} \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_{dk} e_k}{\pi_k} \frac{g_{dl} e_l}{\pi_l}. \tag{3.25}$$

Ejemplo 3.2.1 Un caso sencillo es si se considera una única variable auxiliar x, en donde el modelo es $E_m(y_k) = \beta_d x_k, \;\; V_m(y_k) = c_k = \lambda' x_k \;\; \forall k \in U_d$, con λ un vector de constantes conocidas.

Bajo dicho modelo se obtiene el estimador de razón en dominios

$$\hat{t}_{d,raD} = \sum\nolimits_{U_d} x_k \frac{\sum_{s_d} a_k y_k}{\sum_{s_d} a_k x_k} = \sum\nolimits_{U_d} x_k \hat{B}_d. \tag{3.26}$$

Observación 3.2.3 Si la variable auxiliar x es la indicadora de pertenencia del dominio, δ_d, el estimador $\hat{t}_{d,raD}$ coincide con el estimador de Hayek, $\tilde{t}_d = N_d \tilde{y}_{s_d}$. □

Los residuos muestrales e_{dk} y los ponderadores w_{dk} vienen dados por

$$e_{dk} = y_{dk} - \hat{B}_d x_{dk}, \tag{3.27}$$

y

$$w_{dk} = a_k \left(\frac{\sum_{U_d} x_k}{\sum_{s_d} a_k x_k} \right) \delta_{dk}. \tag{3.28}$$

El estimador la varianza de $\hat{t}_{d,raD}$ viene dado por

$$\hat{V}(\hat{t}_{d,raD}) = \left(\frac{\sum_{U_d} x_k}{\sum_{s_d} a_k x_k}\right)^2 \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_{dk}}{\pi_k} \frac{e_{dl}}{\pi_l} \tag{3.29}$$

$$= \left(\frac{\sum_{U_d} x_k}{\sum_{s_d} a_k x_k}\right)^2 \sum\sum_{s_d} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}. \tag{3.30}$$

□

Ejemplo 3.2.2 Bajo un diseño SI de tamaño n de una población de N individuos, el estimador directo de razón $\hat{t}_{d,raD}$, queda expresado de la forma

$$\hat{t}_{d,raD} = \sum_{U_d} x_k \hat{B}_d = \left(\sum_{U_d} x_k\right) \frac{\bar{y}_{s_d}}{\bar{x}_{s_d}}, \tag{3.31}$$

donde $\bar{y}_{s_d}$ y $\bar{x}_{s_d}$ son las medias muestrales del dominio U_d para la variable de interés y la variable auxiliar respectivamente.

En este caso, el estimador de la varianza de la ecuación (3.29) es

$$\hat{V}(\hat{t}_{d,raD}) = \left(\frac{n(n_{s_d}-1)}{(n-1)n_{s_d}}\right)\left(\frac{\bar{x}_{U_d}}{\bar{x}_{s_d}}\right)^2 N_d^2 \left(\frac{1}{n_{s_d}} - \frac{1}{\hat{N}_d}\right) S_{e_{s_d}}^2$$

$$\doteq \left(\frac{\bar{x}_{U_d}}{\bar{x}_{s_d}}\right)^2 N_d^2 \left(\frac{1}{n_{s_d}} - \frac{1}{\hat{N}_d}\right) S_{e_{s_d}}^2, \tag{3.32}$$

donde $\hat{N}_d = N n_{s_d}/n$, $\bar{x}_{U_d}$ y $\bar{x}_{s_d}$ son las medias poblacional y muestral del dominio U_d respectivamente, $S_{e_{s_d}}^2 = (n_{s_d}-1)^{-1} \sum_{s_d} (y_k - \hat{B}_d x_k)^2$. □

3.2.2. Estimadores indirectos de regresión

En algunas situaciones el tamaño de muestra efectivo en el dominio puede ser muy pequeño, produciendo que las estimaciones de los parámetros del modelo específico del dominio sean inestables. Una manera posible de lograr estimaciones estables, es utilizando información de un subconjunto más amplio de la población para definir el modelo que asiste al estimador de regresión. De esta forma se aumenta el tamaño de muestra efectivo utilizado para estimar los parámetros del modelo.

Un estimador indirecto de regresión para estimar el total del dominio, t_d, es asistido, por un modelo de regresión a nivel de toda la población de la forma $E_m(y_k) = \mathbf{x'}_k \boldsymbol{\beta},\ V_m(y_k) = c_k\ \forall k \in U$.

La construcción del estimador es la siguiente:

- Bajo el modelo a nivel poblacional anterior, se obtienen los valores ajustados $\hat{y}_k = \mathbf{x}'_k\hat{\mathbf{B}}$ para todos los individuos de la población.
- Se calculan los residuos muestrales $e_k = y_k - \hat{y}_k$ para todos los individuos de la muestra.
- Posteriormente, sólo se utilizan los valores ajustados $\hat{y}_k$ para los individuos incluidos en el dominio U_d y sólo se tiene en cuenta el ajuste del modelo en el domino U_d.

Entonces, el estimador indirecto de regresión viene dado por

$$\begin{aligned} \hat{t}_{d,gregP} &= \sum\nolimits_{U_d} \hat{y}_k + \sum\nolimits_{s_d} a_k e_k \\ &= \hat{t}_{d\pi} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx\pi})'\hat{\mathbf{B}}. \end{aligned} \tag{3.33}$$

Este estimador es aproximadamente insesgado, aún para tamaños de muestras modestos.

Observación 3.2.4 El estimador indirecto de regresión $\hat{t}_{d,gregP}$, cumple la propiedad de aditividad

$$\begin{aligned} \hat{t}_{greg} &= \sum_{d=1}^{D} \hat{t}_{d,gregP} = \sum_{d=1}^{D} \left\{ \hat{t}_{d\pi} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx\pi})'\hat{\mathbf{B}} \right\} \\ &= \sum_{d=1}^{D} t_{d\pi} + \left(\sum_{d=1}^{D} \mathbf{t}_{dx} - \sum_{d=1}^{D} \hat{\mathbf{t}}_{dx\pi} \right)' \hat{\mathbf{B}} \\ &= \hat{t}_{\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})'\hat{\mathbf{B}}. \end{aligned}$$

□

Observación 3.2.5 El estimador indirecto de regresión $\hat{t}_{d,gregP}$, es homogéneo

$$\begin{aligned} \hat{t}_{d,gregP} &= \hat{t}_{d\pi} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx\pi}\right)' \hat{\mathbf{B}} \\ &= \sum\nolimits_s a_k y_{dk} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx\pi}\right)' \hat{\mathbf{T}}^{-1} \sum\nolimits_s \frac{a_k \mathbf{x}_k y_k}{c_k} \\ &= \sum\nolimits_s \left[\delta_{dk} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx\pi}\right)' \hat{\mathbf{T}}^{-1} \frac{\mathbf{x}_k}{c_k} \right] a_k y_k \\ &= \sum\nolimits_s g_{dk} a_k y_k = \sum\nolimits_s w_{dk} y_k, \end{aligned} \tag{3.34}$$

con $w_{dk} = g_{dk}a_k$ y $g_{dk} = \delta_{dk} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx\pi}\right)' \hat{\mathbf{T}}^{-1}\mathbf{x}_k/c_k$, en donde los ponderadores g_{dk} son generalmente pequeños para aquellos individuos que no pertenecen al dominio de interés ($\delta_{dk} = 0$) y dependen de la información auxiliar de toda la muestra.

Todos los individuos incluidos en la muestra son ponderados e intervienen en la estimación, tanto aquellos pertenecientes al dominio, como aquellos que no pertenecen al mismo. Por construcción el estimador es indirecto . □

La aproximación de la varianza del estimador $\hat{t}_{d,gregP}$ es

$$AV(\hat{t}_{d,gregP}) = \sum\sum\nolimits_{U_d} \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}, \tag{3.35}$$

con $E_k = y_k - \mathbf{x'}_k\mathbf{B}$.

El estimador de la varianza viene dada por

$$\hat{V}(\hat{t}_{d,gregP}) = \sum\sum\nolimits_{s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_{dk}e_k}{\pi_k} \frac{g_{dl}e_l}{\pi_l}. \tag{3.36}$$

donde la doble suma en (3.36) es sobre toda la muestra s, y se debe a que se utiliza información de otros dominios para estimar el modelo.

Observación 3.2.6 De forma alternativa a la ecuación (3.36), Hidiroglou y Patak (2004) utilizan la doble suma en s_d . □

El estimador indirecto de regresión de la ecuación (3.33) es simple de calcular. Con los valores ajustados $\hat{y}_k$ y los errores e_k, se puede construir el estimador de $\hat{t}_{d,gregP}$ para cada uno de los dominios de interés. A su vez, el estimador cumple la propiedad de aditividad (deseable para todo estimador de dominios).

El problema se encuentra en la práctica, por ejemplo, en encuestas de gran escala en donde generalmente el mismo sistema de ponderadores es utilizado para brindar estimaciones de todas las variables y dominios de interés de la encuesta. Al ser un estimador indirecto, se generan tantos sistemas de ponderadores diferentes como la cantidad de dominios a estimar, lo cual produce que sea poco práctico, debido a que es poco eficaz y engorroso trabajar con distintos sistemas de ponderadores.

Ejemplo 3.2.3 Si se considera el caso de una única variable auxiliar x, en donde el modelo es $E_m(y_k) = \beta x_k,\ \ V(y_k) = c_k = \lambda' x_k\ \ \forall k \in U$, el estimador de razón queda expresado como

$$\hat{t}_{d,raP} = \hat{t}_{d\pi} + \frac{\hat{t}_\pi}{\hat{t}_{x\pi}}(t_{dx} - \hat{t}_{dx\pi}), \tag{3.37}$$

o,

$$\hat{t}_{d,raP} = \sum\nolimits_s a_k g_{dk} y_k, \tag{3.38}$$

con

$$g_{dk} = \delta_{dk} + \frac{(t_{dx} - \hat{t}_{dx\pi})}{\hat{t}_{x\pi}}. \tag{3.39}$$

El estimador de la varianza del estimador de la ecuación (3.37) se obtiene utilizando los ponderadores de la ecuación (3.39) en la ecuación (3.36).

Bajo un diseño SI de tamaño n, el estimador de razón de la ecuación (3.37), se expresa como

$$\hat{t}_{d,raP} = N \left\{ \bar{y}_{s_d} + (\bar{x}_{U_d} - \bar{x}_{s_d}) \frac{\bar{y}_s}{\bar{x}_s} \right\}. \tag{3.40}$$

□

3.2.3. Estimadores Hayek de regresión

Los estimadores de regresión directos e indirectos de las ecuaciones (3.18) y (3.33), pertenecen a la familia de estimadores π. Särndal y Hidiroglou (1989) proponen modificar los estimadores de regresión en dominios si se conoce el tamaño del dominio N_d, el cual es incorporado en el proceso de estimación. Hidiroglou y Patak (2004), denominan a estos estimadores Hayek de regresión, los cuales se obtienen de remplazar $\hat{t}_{d\pi}$ y $\hat{\mathbf{t}}_{dx\pi}$ por los correspondientes estimadores de Hayek en las ecuaciones (3.18) y (3.33)

$$\tilde{t}_d = N_d \tilde{y}_{s_d}, \qquad \tilde{\mathbf{t}}_{dx} = \frac{N_d}{\hat{N}_d} \hat{\mathbf{t}}_{dx\pi}.$$

Los estimadores Hayek de regresión directos e indirectos vienen dados respectivamente como

$$\tilde{t}_{d,gregD} = \tilde{t}_d + (\mathbf{t}_{dx} - \tilde{\mathbf{t}}_{dx})' \hat{\mathbf{B}}_d = \sum\nolimits_U \hat{y}_{dk} + (N_d/\hat{N}_d) \sum\nolimits_s a_k e_{dk}, \tag{3.41}$$

$$\tilde{t}_{d,gregP} = \tilde{t}_d + (\mathbf{t}_{dx} - \tilde{\mathbf{t}}_{dx})' \hat{\mathbf{B}} = \sum\nolimits_{U_d} \hat{y}_k + (N_d/\hat{N}_d) \sum\nolimits_{s_d} a_k e_k. \tag{3.42}$$

Särndal y Hidiroglou (1989) demuestran que el estimador $\tilde{t}_{d,gregP}$ es mas preciso en comparación con el estimador $\hat{t}_{d,gregP}$, debido a que la suma ponderada de los residuos es más estable. Por otro lado, este estimador no cumple la propiedad de aditividad $\hat{t}_{greg} \neq \sum_{d=1}^{D} \tilde{t}_{d,gregP}$ a menos en los casos que $\sum_{s_d} a_k e_k = 0$ para todos los dominios de la población.

Observación 3.2.7 Los estimadores de las ecuaciones (3.41) y (3.42) son estimadores homogéneos

$$\tilde{t}_{d,gregD} = \sum\nolimits_{s_d} \tilde{g}_{dk} a_k y_k, \tag{3.43}$$

con

$$\tilde{g}_{dk} = \frac{N_d}{\hat{N}_d}\delta_{dk} + (\mathbf{t}_{dx} - \tilde{\mathbf{t}}_{dx})' \hat{\mathbf{T}}_d^{-1} \frac{\mathbf{x}_{dk}}{c_k}, \tag{3.44}$$

y

$$\tilde{t}_{d,gregP} = \sum\nolimits_{s} \tilde{g}_{dk} a_k y_k, \tag{3.45}$$

donde

$$\tilde{g}_{dk} = \frac{N_d}{\hat{N}_d}\delta_{dk} + \left(\mathbf{t}_{dx} - \tilde{\mathbf{t}}_{dx}\right)' \hat{\mathbf{T}}^{-1} \frac{\mathbf{x}_k}{c_k}. \tag{3.46}$$

□

Särndal et al. (1992) definen la aproximación de la varianza del estimador $\tilde{t}_{d,gregP}$ como

$$AV(\tilde{t}_{d,gregP}) = \sum \sum\nolimits_{U_d} \Delta_{kl} \frac{E_k - \bar{E}_{U_d}}{\pi_k} \frac{E_l - \bar{E}_{U_d}}{\pi_l}, \tag{3.47}$$

donde $E_k = y_k - \mathbf{x}'_k \mathbf{B}$ y $\bar{E}_{U_d} = \sum_{U_d} E_k / N_d$.

Observación 3.2.8 La aproximación de la varianza de la ecuación (3.47) se obtiene de escribir el error del estimador $\tilde{t}_{d,gregP}$ como

$$\tilde{t}_{d,gregP} - t_d = N_d \left[\left(\tilde{E}_{s_d} - \bar{E}_{U_d} \right) - \left(\mathbf{B} - \hat{\mathbf{B}} \right)' \left(\tilde{\mathbf{x}}_{s_d} - \bar{\mathbf{x}}_{U_d} \right) \right],$$

donde $\tilde{E}_{s_d} = \sum_{s_d} a_k E_k / \hat{N}_d$, $\tilde{\mathbf{x}}_{s_d} = \hat{\mathbf{t}}_{dx\pi} / \hat{N}_d$ y $\bar{\mathbf{x}}_{U_d} = \mathbf{t}_{dx} / N_d$.

El término $\left(\mathbf{B} - \hat{\mathbf{B}} \right)' \left(\tilde{\mathbf{x}}_{s_d} - \bar{\mathbf{x}}_{U_d} \right)$ tiende a cero y es de menor orden en probabilidad que el término $N_d \left(\tilde{E}_{s_d} - \bar{E}_{U_d} \right)$. Este último por si solo provee la aproximación

$$\tilde{t}_{d,gregP} - t_d \doteq N_d \left(\tilde{E}_{s_d} - \bar{E}_{U_d} \right). \tag{3.48}$$

Finalmente $N_d \tilde{E}_{s_d}$, tiene la misma estructura que el estimador $\tilde{t}_d$ de la ecuación (2.19). Por lo tanto, la aproximación de la varianza del estimador $\tilde{t}_{d,gregP}$, se obtiene de remplazar en la ecuación (2.20), y_k, por E_k . □

Observación 3.2.9 La aproximación de la varianza del estimador $\tilde{t}_{d,gregD}$, se obtiene de cambiar en la ecuación (3.47), los residuos poblacionales E_k, por $E_{dk} = y_{dk} - \mathbf{x}'_{dk}\mathbf{B}_d$. □

Observación 3.2.10 Los estimadores de las varianzas de $\tilde{t}_{d,gregD}$ y $\tilde{t}_{d,gregP}$ se obtienen de reemplazar respectivamente los ponderadores $\tilde{g}$ de (3.44) y (3.46) en las ecuaciones (3.29) y (3.36). □

3.3. Modelos de grupos para la estimación en dominios

En vez de utilizar un modelo común para toda la población, en algunas circunstancias puede resultar conveniente, considerar un conjunto de modelos de regresión definidos en subconjuntos de la población, denominados modelos de grupos.

La idea central, es que los grupos son un factor poderoso para explicar la variabilidad de la variable de interés, mientras que quizás los dominios por si solos no lo sean. Por ejemplo, en una encuesta a personas, los grupos pueden ser estratos geográficos o grupos de sexo/edad. En la práctica, los grupos pueden coincidir con los estratos y en esos casos el tamaño de muestra en el grupo es controlado (y puede ser fijo si el diseño lo permite).

Consideremos que la población es particionada en G grupos, $U_1, ..., U_g, .., U_G$, en donde los límites de los grupos no tienen porque coincidir con los límites de los dominios de interés.

Sin pérdida de generalidad, se analiza el caso en donde los G grupos intersectan los D dominios para formar una grilla de DG celdas, U_{dg}, $d = 1, ..., D$; $g = 1, ..., G$. Sea N_{dg} el tamaño de la celda dg, o sea, la intersección del dominio U_d con el grupo U_g.

Cuadro 3.1: *Partición de la población U*

	$U_{\cdot 1}$	$\cdots$	$U_{\cdot g}$	$\cdots$	$U_{\cdot G}$
$U_{1\cdot}$	U_{11}	$\cdots$	U_{1g}	$\cdots$	U_{1G}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$U_{d\cdot}$	U_{d1}	$\cdots$	U_{dg}	$\cdots$	U_{dG}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$U_{D\cdot}$	U_{D1}	$\cdots$	U_{Dg}	$\cdots$	U_{DG}

En consecuencia, son válidas las siguientes ecuaciones

$$U = \bigcup_{d=1}^{D} U_{d\cdot} = \bigcup_{g=1}^{G} U_{\cdot g} = \bigcup_{d=1}^{D}\bigcup_{g=1}^{G} U_{dg}, \tag{3.49}$$

y

$$N = \sum_{d=1}^{D} N_{d\cdot} = \sum_{g=1}^{G} N_{\cdot g} = \sum_{d=1}^{D}\sum_{g=1}^{G} N_{dg}, \tag{3.50}$$

donde $N_{d\cdot}$ indica que se suma todas las celdas de la fila d y de forma análoga $N_{\cdot g}$ indica que se suma todas las celdas de la columna g.

Análogamente a la ecuación (3.49) y (3.50) se tiene

$$s = \bigcup_{d=1}^{D} s_{d\cdot} = \bigcup_{g=1}^{G} s_{\cdot g} = \bigcup_{d=1}^{D}\bigcup_{g=1}^{G} s_{dg}, \tag{3.51}$$

y

$$n = \sum_{d=1}^{D} n_{d\cdot} = \sum_{g=1}^{G} n_{\cdot g} = \sum_{d=1}^{D}\sum_{g=1}^{G} n_{dg}, \tag{3.52}$$

donde los tamaños muestrales en las celdas $n_{s_{dg}}$ son aleatorios. Usualmente, $n_{s_{d\cdot}}$ y $n_{s_{\cdot g}}$, también son aleatorios, aunque, circunstancialmente, el tamaño $n_{s_{\cdot g}}$ puede ser fijo, si el grupo g es un estrato, donde se selecciona un número predeterminado de individuos.

En la práctica los dominios de interés pueden ser numerosos, por ejemplo, cien o más. En tanto, los grupos son un número pequeño, digamos diez o menos. Särndal et al. (1992) indica que trabajar con un número mayor de grupos no genera una ganancia relativa de eficiencia. La reducción de la varianza del estimador puede ser marginal si se aumenta el número de grupos a más de diez.

Existen distintas alternativas según se utilicen estimadores directos o indirectos de regresión. La elección dependen de los tamaños de muestras en las celdas y a la información auxiliar disponible.

3.3.1. Modelo a nivel de celda

Cuando el tamaño de muestra en el dominio es lo suficientemente grande, se puede definir un modelo de regresión para cada celda dg, como $E_m(y_k) = \mathbf{x'}_k\boldsymbol{\beta}_{dg},\ \ V_m(y_k) = c_k,\ \ \forall k \in U_{dg}$.

Una ventaja de poder modelar por celda, se encuentra en el vector auxiliar, el cual puede contener distintas variables para cada una de las celdas. En algunos casos, se puede tener disponible más información auxiliar para ciertas celdas o tamaños de muestras más grandes, por otro lado, en otras circunstancias puede no suceder lo mismo (por disponibilidad de información auxiliar o tamaño de muestra reducido), lo cual obliga a definir un modelo más parsimonioso.

Supongamos que la información auxiliar disponible es la misma para cada una de las celdas. Entonces, se requiere conocer los totales de las variables auxiliares a nivel de la celda dg, o sea, $\mathbf{t}_{dgx} = \sum_U \delta_{dgk}\mathbf{x}_k = \sum_{U_{dg}} \mathbf{x}_k$ es conocido, $\forall g = 1, ..., G, \ \ \forall d = 1, ..., D$, en donde $\delta_{dgk} = 1$ si $k \in U_{dg}$ y 0 en otro caso.

El estimador de regresión queda definido como

$$\hat{t}_{d,gregDG} = \sum_{g=1}^{G} \sum\nolimits_{U_{dg}} \hat{y}_k + \sum_{g=1}^{G} \sum\nolimits_{s_{dg}} a_k(y_k - \hat{y}_k) \tag{3.53}$$

$$= \hat{t}_{d\pi} + \sum_{g=1}^{G} (\mathbf{t}_{dgx} - \hat{\mathbf{t}}_{dgx\pi})'\hat{\mathbf{B}}_{dg}, \tag{3.54}$$

en donde los parámetros específicos de la celda, $\mathbf{B}_{dg}$, son estimados como en (3.4) sustituyendo la variable y por la variable $y_{dg} = \delta_{dg}y$ y el vector de información auxiliar $\mathbf{x}$ por $\mathbf{x}_{dg} = \delta_{dg}\mathbf{x}$.

Este estimador, necesita que el tamaño de muestra para cada celda dg sea lo suficientemente grande de manera de evitar estimaciones inestables de los parámetros $\mathbf{B}_{dg}$, por lo tanto, su uso se encuentra restringido para dominios con tamaños de muestra lo suficientemente grandes.

Observación 3.3.1 El estimador obtenido es homogéneo

$$\hat{t}_{d,gregDG} = \sum_{g=1}^{G} \sum\nolimits_{s_g} g_{dgk} a_k y_k = \sum\nolimits_{s_d} w_{dgk} y_k, \tag{3.55}$$

donde

$$g_{dgk} = \delta_{dgk} + (\mathbf{t}_{dgx} - \hat{\mathbf{t}}_{dgx\pi})'\hat{\mathbf{T}}_{dg}^{-1}\mathbf{x}_{dgk}/c_k, \tag{3.56}$$

y $\hat{\mathbf{T}}_{dg} = \sum_s a_k \mathbf{x}_{dgk}\mathbf{x}'_{dgk}/c_k$. □

Por construcción el estimador es directo, solo los individuos pertenecientes a las celdas U_{dg} intervienen en el proceso de estimación.

La aproximación de la varianza del estimador es

$$AV(\hat{t}_{d,gregDG}) = \sum \sum\nolimits_U \Delta_{kl} \frac{E_{dgk}}{\pi_k} \frac{E_{dgl}}{\pi_l}, \tag{3.57}$$

donde $E_{dgk} = y_k - \mathbf{x'}_{dgk}\mathbf{B}_{dg}$ si $k \in U_{dg}$ y 0 en otro caso, $\forall g = 1, ..., G$.

El estimador de la varianza viene dado por

$$\hat{V}(\hat{t}_{d,gregDG}) = \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_{dgk}e_{dgk}}{\pi_k} \frac{g_{dgl}e_{dgl}}{\pi_l}, \tag{3.58}$$

donde $e_{dgk} = y_k - \mathbf{x'}_{dgk}\hat{\mathbf{B}}_{dg}$ si $k \in s_{dg}$ y 0 en otro caso, $\forall g = 1, ..., G$.

Ejemplo 3.3.1 Un caso sencillo es cuando existe una sola variable auxiliar x y la misma es la variable indicadora de pertenencia a la celda dg, δ_{dg}.

El modelo de medias por grupo es $E_m(y_k) = \beta_{dg}, \;\; V_m(y_k) = c_{dg} \;\; \forall k \in U_{dg}$ y la estimación del parámetro B_{dg} viene dada por

$$\hat{B}_{dg} = \frac{\sum_{s_{dg}} a_k y_k}{\hat{N}_{dg}} = \tilde{y}_{s_{dg}}$$

donde $\hat{N}_{dg} = \sum_{s_{dg}} a_k$.

Lo anterior, produce el estimador post-estratificado para dominios que se obtiene como una suma ponderada de las medias muestrales por celda

$$\hat{t}_{d,post} = \sum_{g=1}^{G} N_{dg}\tilde{y}_{s_{dg}}. \tag{3.59}$$

Observación 3.3.2 El único requisito para calcular este estimador, es conocer el tamaño de las celdas N_{dg} . □

En el estimador de la ecuación (3.59), se requiere que ninguno de los tamaños de muestra por celda sea extremadamente pequeño. Si alguna celda se encuentra vacía, o sea, el tamaño de muestra es nulo, el estimador es imposible de calcular. Por otro lado, si los tamaños de muestras para algunas celdas son extremadamente pequeños, el estimador puede ser muy inestable y no se debería usar en estos casos. Una alternativa para estos casos es colapsar grupos, de forma de obtener tamaños de muestra mas grandes, y así asegurar obtener estimaciones más estables.

La varianza aproximada y el estimador de la varianza de (3.59), se obtiene de la ecuación (3.57) y (3.58) respectivamente, en donde los residuos poblacionales, muestrales, y los ponderadores g vienen dados respectivamente por $E_{dgk} = y_{dgk} - \delta_{dgk}\bar{y}_{U_{dg}}$, $e_{dgk} = y_{dgk} - \delta_{dgk}\tilde{y}_{s_{dg}}$ y $g_{dgk} = \delta_{dgk}N_{dg}/\hat{N}_{dg}$. □

Observación 3.3.3 Bajo un SI de tamaño n, el estimador post-estratificado de la ecuación (3.59) se expresa como

$$\hat{t}_{d,post} = \sum_{g=1}^{G} N_{dg} \bar{y}_{s_{dg}}.$$

El estimador es construido como una suma ponderada de las medias muestrales por celda, $\bar{y}_{s_{dg}} = \sum_{s_{dg}} y_k / n_{s_{dg}}$ con los totales N_{dg} como ponderadores.

El estimador de la varianza es

$$\begin{aligned} \hat{V}_{SI}(\hat{t}_{d,post}) &= \sum_{g=1}^{G} \left(\frac{n}{n-1} \frac{n_{s_{dg}} - 1}{n_{s_{dg}}} \right) N_{dg}^2 \left(\frac{1}{n_{s_{dg}}} - \frac{1}{\hat{N}_{dg}^2} \right) S_{ys_{dg}}^2 \\ &\doteq \sum_{g=1}^{G} N_{dg}^2 \left(\frac{1}{n_{s_{dg}}} - \frac{1}{\hat{N}_{dg}} \right) S_{ys_{dg}}^2, \end{aligned} \tag{3.60}$$

donde $\hat{N}_{dg} = N n_{s_{dg}} / n$ y $S_{ys_{dg}}^2 = \left(n_{s_{dg}} - 1\right)^{-1} \sum_{s_{dg}} (y_k - \bar{y}_{s_{dg}})^2$. □

3.3.2. Modelo a nivel de grupo

Utilizar un modelo para cada celda puede ser excesivo, sobre todo cuando se trabaja con muestras de tamaño modesto. Una alternativa es definir un modelo para cada grupo de la forma, $E_m(y_k) = \mathbf{x'}_k \boldsymbol{\beta}_g$, $V_m(y_k) = c_k \ \forall k \in U_g$.

Entonces, el estimador queda definido como

$$\begin{aligned} \hat{t}_{d,gregPG} &= \sum_{g=1}^{G} \sum\nolimits_{U_d} \hat{y}_k + \sum_{g=1}^{G} \sum\nolimits_{s_{dg}} a_k (y_k - \hat{y}_k) \\ &= \hat{t}_{d\pi} + \sum_{g=1}^{G} (\mathbf{t}_{dgx} - \hat{\mathbf{t}}_{dgx\pi})' \hat{\mathbf{B}}_g, \end{aligned} \tag{3.61}$$

donde $\hat{\mathbf{B}}_g$ se obtiene de ajustar un único modelo a cada grupo U_g

$$\begin{aligned} \hat{\mathbf{B}}_g &= \left(\sum\nolimits_s a_k \mathbf{x}_{gk} \mathbf{x'}_{gk} / c_k \right)^{-1} \left(\sum\nolimits_s a_k \mathbf{x}_{gk} y_{gk} / c_k \right) \\ &= \left(\sum\nolimits_{s_g} a_k \mathbf{x}_k \mathbf{x'}_k / c_k \right)^{-1} \left(\sum\nolimits_{s_g} a_k \mathbf{x}_k y_k / c_k \right), \end{aligned} \tag{3.62}$$

donde $\mathbf{x}_{gk} = \delta_{gk} \mathbf{x}_k$ y $y_{gk} = \delta_{gk} y_k$, con $\delta_{gk} = 1$ si $k \in U_g$ y 0 si $k \notin U_g$.

La información auxiliar necesaria para poder construir el estimador es la misma que para el caso del estimador de la ecuación (3.53). La diferencia radica que de esta manera se evita obtener estimaciones inestables en los parámetros del modelo.

Observación 3.3.4 El estimador de la ecuación (3.61) puede ser expresado de manera homogénea

$$\begin{aligned}
\hat{t}_{d,gregPG} &= \sum_{g=1}^{G} \left\{ \hat{t}_{dg\pi} + (\mathbf{t}_{dgx} - \hat{\mathbf{t}}_{dgx\pi})' \hat{\mathbf{B}}_g \right\} \\
&= \sum_{g=1}^{G} \left\{ \sum\nolimits_s a_k \delta_{dk} y_{gk} + (\mathbf{t}_{dgx} - \hat{\mathbf{t}}_{dgx\pi})' \hat{\mathbf{T}}_g^{-1} \left(\sum\nolimits_s a_k \mathbf{x}_{gk} y_{gk}/c_k \right) \right\} \\
&= \sum_{g=1}^{G} \left\{ \sum\nolimits_s \left[\delta_{dk} + (\mathbf{t}_{dgx} - \hat{\mathbf{t}}_{dgx\pi})' \hat{\mathbf{T}}_g^{-1} \mathbf{x}_{gk}/c_k \right] a_k y_{gk} \right\} \\
&= \sum_{g=1}^{G} \sum\nolimits_s g_{dgk} a_k y_{gk} = \sum\nolimits_s g_{dgk} a_k y_k
\end{aligned} \tag{3.63}$$

donde $\hat{\mathbf{T}}_g = \sum_s a_k \mathbf{x}_{gk} \mathbf{x}'_{gk}/c_k$. □

El estimador por construcción es indirecto, todos los individuos de la población son ponderados e intervienen en el proceso de estimación.

La varianza aproximada del estimador es

$$AV(\hat{t}_{d,gregPG}) = \sum \sum\nolimits_{U_d} \Delta_{kl} \frac{E_{gk}}{\pi_k} \frac{E_{gl}}{\pi_l}, \tag{3.64}$$

donde $E_{gk} = y_k - \mathbf{x}'_{gk} \mathbf{B}_g$ si $k \in U_{\cdot g} \ \ \forall g = 1, ..., G$.

El estimador de la varianza viene dado por

$$\hat{V}(\hat{t}_{d,gregPG}) = \sum \sum\nolimits_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_{dgk} e_{gk}}{\pi_k} \frac{g_{dgl} e_{gl}}{\pi_l}, \tag{3.65}$$

donde $e_{gk} = y_k - \mathbf{x}'_{gk} \hat{\mathbf{B}}_g$ si $k \in s_{\cdot g} \ \ \forall g = 1, ..., G$.

Ejemplo 3.3.2 Se considera el modelo de medias común por grupo $E_m(y_k) = \beta_g, \ \ V(y_k) = c_g, \ \ \forall k \in U_g$, para $g = 1, .., G$.

El estimador indirecto de regresión para el total del dominio t_d, es

$$\hat{t}_{d,gregPG} = \sum_{g=1}^{G} N_{dg}\tilde{y}_{s_{\cdot g}} + \sum_{g=1}^{G} \hat{N}_{dg}\left(\tilde{y}_{s_{dg}} - \tilde{y}_{s_{\cdot g}}\right), \tag{3.66}$$

donde

$$\tilde{y}_{s_{\cdot g}} = \sum\nolimits_{s_{\cdot g}} y_k/\hat{N}_{\cdot g}, \quad \tilde{y}_{s_{dg}} = \sum\nolimits_{s_{dg}} y_k/\hat{N}_{dg},$$

$$\hat{N}_{\cdot g} = \sum\nolimits_{s_{\cdot g}} a_k \quad \text{y} \quad \hat{N}_{dg} = \sum\nolimits_{s_{dg}} a_k.$$

En general, es prudente trabajar con un número reducido de parámetros ya que esto estabiliza las estimaciones y, si los grupos son el principal factor de variabilidad de la variable de interés y, no genera una pérdida sustancial de eficiencia si se compara con el estimador post-estratificado de la ecuación (3.59).

De todas formas, el sumando $\sum_{g=1}^{G} \hat{N}_{dg}\left(\tilde{y}_{s_{dg}} - \tilde{y}_{s_{\cdot g}}\right)$ puede ser muy inestable, debido al tamaño de muestra pequeño en las celdas. Entonces, se puede utilizar el estimador Hayek de regresión de la ecuación (3.42), que en este caso queda

$$\tilde{t}_{d,gregPG} = \sum_{g=1}^{G} N_{dg}\tilde{y}_{s_{\cdot g}} + \left(N_{d\cdot}/\hat{N}_{d\cdot}\right)\sum_{g=1}^{G} \hat{N}_{dg}\left(\tilde{y}_{s_{dg}} - \tilde{y}_{s_{\cdot g}}\right). \tag{3.67}$$

Las cantidades requeridas para el cálculo del estimador de la varianza están dadas por

$$e_{gk} = y_{gk} - \tilde{y}_{s_{\cdot g}} \tag{3.68}$$

y

$$g_{dgk} = N_{d\cdot}\left\{\frac{\delta_{dk}}{\hat{N}_{d\cdot}} + \left(\frac{N_{dg}}{N_{d\cdot}} - \frac{\hat{N}_{dg}}{\hat{N}_{d\cdot}}\right)\frac{1}{\hat{N}_{\cdot g}}\right\}. \tag{3.69}$$

□

Capítulo 4

Estimadores calibrados

4.1. Introducción

La calibración en el problema de estimación en dominios es una alternativa para producir estimaciones. Al igual que en las estrategias anteriores la clave se encuentra en disponer de información auxiliar potente, la cual es utilizada para construir un sistema de ponderadores, llamados ponderadores calibrados. En encuestas de gran escala, el mismo sistema de ponderadores calibrados puede ser utilizado para realizar estimaciones de distintas variables en distintas subpoblaciones.

El vector de información auxiliar utilizado para la calibración, siempre que sea posible, debe contener información específica del dominio o de subpoblaciones que contengan a dichos dominios. Si la información auxiliar se encuentra disponible en el marco muestral junto con la variable indicadora de pertenencia al dominio, δ_d, este requisito se cumple. En tanto, si la información auxiliar proviene de fuentes externas, registros administrativos o estimaciones provenientes de otras encuestas, el requisito de información específica del dominio puede verse comprometido y generalmente hay que conformarse con la información a nivel de subpoblaciones más amplias.

A continuación se hace una breve reseña de los estimadores calibrados y su aplicación al problema de estimación en dominios.

4.2. Estimadores calibrados

Para el total de la variable y en la población U, $t = \sum_U y_k$, el estimador calibrado toma la forma

$$\hat{t}_{cal} = \sum\nolimits_s w_k y_k, \tag{4.1}$$

donde $\{w_k \in s\}$ es el sistema de ponderadores calibrados, los cuales dependen de la información auxiliar disponible y cumplen

$$\sum\nolimits_s w_k \mathbf{x}_k = \sum\nolimits_U \mathbf{x}_k = \mathbf{t}_x, \tag{4.2}$$

con $\mathbf{t}_x$ el vector de totales poblacionales de las variables auxiliares.

La ecuación (4.2) es llamada ecuación de calibración.

De esta manera, el sistema de ponderadores calibrados estima sin error a los totales de las variables auxiliares. Esta es una propiedad deseable, debido a que brinda coherencia a las estimaciones.

La calibración solo hace referencia a la información auxiliar a utilizar para calcular el nuevo sistema de ponderadores y no hace explícito ningún modelo. De esta manera, la calibración se diferencia del enfoque de regresión, en donde su construcción se basa en encontrar predicciones, $\hat{y}_k$, de la variable de interés a través de un modelo que lo asiste.

Existen dos métodos comúnmente utilizados para construir el nuevo sistema de ponderadores:

- El de minimización de la distancia.
- El enfoque funcional.

El método de minimización de la distancia, consiste en definir una medida apropiada de distancia entre los ponderadores originales, a_k y los nuevos ponderadores calibrados, w_k. Dicha distancia es posteriormente minimizada sujeta a la restricción proveniente de la ecuación de calibración (4.2). Existen muchas medidas de distancias usadas en la práctica, una de ellas es la distancia de mínimos cuadrados generalizados, también llamada distancia chi-cuadrado que viene dada por

$$(1/2)\sum\nolimits_s c_k(w_k - a_k)^2/a_k = (1/2)\sum\nolimits_s c_k a_k(w_k/a_k - 1)^2. \tag{4.3}$$

Minimizando (4.3) sujeto a la ecuación (4.2), se obtiene

$$w_k = a_k\left[1 + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})'\hat{\mathbf{T}}^{-1}\mathbf{x}_k/c_k\right]. \tag{4.4}$$

Bajo esta distancia, el estimador calibrado $\hat{t}_{cal}$ es idéntico al estimador de regresión $\hat{t}_{greg}$.

Los valores c_k tienen el rol de moderar la importancia de los términos en la ecuación (4.3). Una observación con un valor alto en c_k tendrá un ponderador calibrado w_k más cercano al ponderador original a_k, que una observación con un valor c_k más pequeño.

Deville y Särndal (1992) analizan diferentes distancias que llevan a diferentes sistemas de ponderados calibrados, por ejemplo la distancia llamada Raking Ratio

$$\sum\nolimits_s c_k a_k \left\{(w_k/a_k)Log(w_k/a_k) - w_k/a_k + 1\right\}. \tag{4.5}$$

Otras distancias aseguran obtener ponderadores calibrados, w_k, de tal forma que se cumpla que $A_k \leq w_k \leq B_k \quad \forall k \in s$, para unos límites específicos A_k y B_k. De esta forma, se evitan ponderadores w_k, muy grandes (influyentes) o muy pequeños (negativos).

Estevao y Särndal (2000), proponen como alternativa al método de minimización de la distancia, el enfoque funcional, el cual permite generar diferentes opciones para construir sistemas de ponderadores calibrados utilizando la misma información auxiliar.

El enfoque funcional parte de considerar un nuevo sistema de ponderados de la forma

$$w_k = a_k F(\lambda' \mathbf{z}_k), \tag{4.6}$$

donde $\mathbf{z}_k$, es un vector con valores definidos para todos los individuos de la muestra con la misma dimensión que el vector de información auxiliar $\mathbf{x}_k$ y el vector λ es determinado usando la ecuación de calibración, $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$.

Existen diferentes elecciones de la función $F(.)$, por ejemplo, para el caso de una función lineal $F(u) = 1 + u$, los ponderadores calibrados quedan definidos de la forma $w_k = a_k(1 + \lambda' \mathbf{z}_k)$.

El estimador calibrado queda definido como

$$\hat{t}_{cal} = \sum\nolimits_s w_k y_k = \sum\nolimits_s a_k(1 + \lambda' \mathbf{z}_k) y_k, \tag{4.7}$$

donde

$$\lambda' = (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x}'_k\right)^{-1}. \tag{4.8}$$

Para cualquier elección del vector $\mathbf{z}_k$, los ponderadores calibrados $w_k = a_k(1 + \lambda' \mathbf{z}_k)$ cumplen la ecuación de calibración. El vector $\mathbf{z}_k$ puede tomar cualquier valor, inclusive cero, siempre y cuando no sea $0 \quad \forall k \in s$. Por otro lado la matriz $\sum_s a_k \mathbf{z}_k \mathbf{x}'_k$ debe no ser singular.

El estimador calibrado, para el caso de la función lineal, puede expresarse como la suma del estimador π más un término de ajuste

$$\hat{t}_{cal} = \hat{t}_\pi + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \hat{\mathbf{Q}}, \tag{4.9}$$

donde

$$\hat{\mathbf{Q}} = \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x'}_k\right)^{-1} \left(\sum\nolimits_s a_k \mathbf{z}_k y_k\right). \tag{4.10}$$

En la práctica, el vector $\mathbf{z}_k$ coincide con el vector de información auxiliar, o sea, $\mathbf{z}_k = \mathbf{x}_k$.

Observación 4.2.1 Si $\mathbf{z}_k = \mathbf{x}_k / c_k$, entonces $\hat{\mathbf{Q}}$ coincide con $\hat{\mathbf{B}}$ de la ecuación (3.4) y el estimador calibrado es idéntico al estimador de regresión, $\hat{t}_{greg}$, de la ecuación (3.6) . □

La aproximación de la varianza del estimador calibrado, $\hat{t}_{cal}$, viene dada por

$$AV(\hat{t}_{cal}) = \sum \sum\nolimits_U \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}, \tag{4.11}$$

con $E_k = y_k - \mathbf{x'}_k \mathbf{Q}$ y

$$\mathbf{Q} = \left(\sum\nolimits_U \mathbf{z}_k \mathbf{x'}_k\right)^{-1} \left(\sum\nolimits_U \mathbf{z}_k y_k\right). \tag{4.12}$$

El estimador para la varianza de $\hat{t}_{cal}$ es

$$\hat{V}(\hat{t}_{cal}) = \sum \sum\nolimits_s \frac{\Delta_{kl}}{\pi_{kl}} (w_k e_k)(w_l e_l), \tag{4.13}$$

donde $e_k = y_k - \mathbf{x'}_k \hat{\mathbf{Q}}$ y

$$w_k = a_k \left[1 + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x'}_k\right)^{-1} \mathbf{z}_k\right].$$

4.3. Calibración en dominios

Para el caso de un total, el estimador calibrado $\hat{t}_{cal} = \sum_s w_k y_k$, se basa en los ponderadores w_k, que son determinados para todos los individuos de la muestra en base al vector de totales, $\mathbf{t}_x = \sum_U \mathbf{x}_k$. En el problema de estimación en dominios, se calcula un sistema de ponderadores calibrados w_{dk} para todos los individuos pertenecientes al dominio U_d, esto se realiza en base al vector de totales específicos del dominio $\mathbf{t}_{dx} = \sum_{U_d} x_k = \sum_U \delta_{dk} \mathbf{x}_k = \sum_U \mathbf{x}_{dk}$.

$$\begin{aligned} \hat{t}_{d,calD} &= \sum\nolimits_{s_d} w_{dk} y_k \\ &= \sum\nolimits_{s_d} a_k (1 + \lambda_d' \mathbf{z}_k) y_k, \end{aligned} \tag{4.14}$$

donde

$$\lambda'_d = (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx\pi})' \left(\sum\nolimits_{s_d} a_k \mathbf{z}_k \mathbf{x}'_k \right)^{-1}. \tag{4.15}$$

El estimador $\hat{t}_{d,calD}$ es directo y puede expresarse, al igual que en la ecuación (4.9), como la suma del estimador π para el total del dominio U_d más un término de ajuste

$$\hat{t}_{d,calD} = \hat{t}_{d\pi} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx\pi})' \hat{\mathbf{Q}}_d, \tag{4.16}$$

con

$$\hat{\mathbf{Q}}_d = \left(\sum\nolimits_{s_d} a_k \mathbf{z}_k \mathbf{x}'_k \right)^{-1} \left(\sum\nolimits_{s_d} a_k \mathbf{z}_k y_k \right). \tag{4.17}$$

Como se dijo el sistema de ponderadores calibrados $\{w_{dk} = a_k(1 + \lambda'_d \mathbf{z}_k)\ \forall k \in s_d\}$ estima sin error los totales de las variables auxiliares en el dominio U_d.

Observación 4.3.1 Si el vector $\mathbf{z}_k = \mathbf{x}_k / c_k$, el estimador $\hat{t}_{d,calD}$ es idéntico al estimador directo de regresión, $\hat{t}_{d,gregD}$, de la ecuación (3.18) . □

El estimador de la varianza del estimador $\hat{t}_{d,calD}$, como se verá más adelante, viene dado por

$$\hat{V}(\hat{t}_{d,calD}) = \sum \sum\nolimits_{s_d} \frac{\Delta_{dk}}{\pi_{kl}} (w_{dk} e_{dk})(w_{dl} e_{dl}), \tag{4.18}$$

donde

$$e_{dk} = \begin{cases} y_k - \mathbf{x}'_k \hat{\mathbf{Q}}_d & \text{si} \quad k \in U_d \\ 0 & \text{si} \quad k \notin U_d \end{cases}. \tag{4.19}$$

Si es necesario obtener estimaciones para un conjunto numeroso de dominios de la población puede ser poco práctico calcular un sistema de ponderadores calibrados para cada dominio, incluso pueda ocurrir que la información auxiliar específica del dominio, necesaria para la calibración, puede no encontrarse disponible. Una alternativa es utilizar un único sistema de ponderadores $\{w_k = a_k(1 + \lambda' \mathbf{z}_k)\ \forall k \in s\}$, que cumplan con la ecuación de calibración (4.2), la cual se encuentra definida para la población U.

Los ponderadores w_k pueden ser utilizados para

- Obtener estimaciones de todas las variables de interés.

- Obtener estimaciones en todos los dominios de la población.

Si los ponderadores calibrados w_k son aplicados a los individuos de la muestra pertenecientes al dominio U_d, el estimador del total del dominio queda definido como

$$\hat{t}_{d,calU} = \sum\nolimits_s w_k y_{dk} = \sum\nolimits_{s_d} w_k y_k. \tag{4.20}$$

donde

$$w_k = a_k \left[1 + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x'}_k \right)^{-1} \mathbf{z}_k \right].$$

Para el cálculo del estimador $\hat{t}_{d,calU}$ todos los individuos de la muestra incluidos en el dominio U_d son ponderados por el mismo sistema de ponderadores. Por construcción el estimador es directo.

El estimador $\hat{t}_{d,calU}$, puede escribirse como el estimador π para el dominio U_d, más un término de ajuste a nivel de toda la población

$$\begin{aligned}
\hat{t}_{d,calU} &= \sum\nolimits_s w_k y_{dk} = \sum\nolimits_s a_k (1 + \lambda' \mathbf{z}_k) y_{dk} \\
&= \sum\nolimits_s a_k \left[1 + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x'}_k \right)^{-1} \mathbf{z}_k \right] y_{dk} \\
&= \sum\nolimits_s a_k y_{dk} + \sum\nolimits_s a_k (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x'}_k \right)^{-1} \mathbf{z}_k y_{dk} \\
&= \hat{t}_{d\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x'}_k \right)^{-1} \left(\sum\nolimits_s a_k \mathbf{z}_k y_{dk} \right) \\
&= \hat{t}_{d\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \hat{\mathbf{R}}_U.
\end{aligned} \tag{4.21}$$

Un estimador para la varianza del estimador $\hat{t}_{d,calU}$ es

$$\hat{V}(\hat{t}_{d,calU}) = \sum \sum\nolimits_s \frac{\Delta_{kl}}{\pi_{kl}} (w_k e_{Uk})(w_l e_{Ul}), \tag{4.22}$$

donde

$$e_{Uk} = \begin{cases} y_k - \mathbf{x'}_k \hat{\mathbf{R}}_U & \text{si} \quad k \in U_d \\ -\mathbf{x'}_k \hat{\mathbf{R}}_U & \text{si} \quad k \notin U_d \end{cases}. \tag{4.23}$$

El estimador $\hat{t}_{d,calU}$, puede ser poco eficiente debido a la cantidad de residuos negativos para todas los individuos que no pertenecen al dominio U_d.

Estevao y Särndal (1999) denominan al estimador de la ecuación (4.20) como estimador *uni-weight*, de forma de hacer énfasis en que el estimador para el dominio U_d, es construido como una

suma ponderada, en donde los ponderadores utilizados provienen de un único sistema, que ha sido calculado para proporcionar estimaciones para el total y para todos los dominios de la población.

En encuestas de gran escala, en donde es necesario brindar estimaciones para un conjunto amplio de dominios, o en aquellas en donde la periodicidad juega un rol importante, el estimador $\hat{t}_{d,calU}$, es una alternativa útil. El uso de un único sistema de ponderadores genera estimadores con las siguientes características:

- *Very nearly design unbiased* para todos los dominios de la población.
- Cumple la propiedad de aditividad.
- Crea economías de escala.
- No posee como requisito disponer de información específica del domino para su construcción, a diferencia de los estimadores de regresión vistos antes, ya sea directos o indirectos.

El estimador *uni-weight*, no tiene por que ser la opción más eficiente para cada uno de los dominios de la población, pero permite obtener buenas estimaciones en tiempo y forma.

La eficiencia del estimador $\hat{t}_{d,calU}$, depende de la información auxiliar utilizada para calcular los ponderadores w_k. Dicha información, siempre y cuando sea posible, puede estar definida a un nivel más desagregado de la población, llamados grupos de calibración.

En la práctica, los dominios de interés pueden intersectar a varios grupos de calibración. Por ejemplo, en una encuesta a hogares, los grupos de calibración pueden estar definidos por regiones geográficas, en donde la información auxiliar corresponde al número de personas por sexo y tramo etario provenientes de las proyecciones de población. Si el interés es estimar el ingreso promedio para las mujeres de un determinado tramo etario, en este caso, el dominio intersecta a todos los grupos de calibración y a su vez dichos grupos de calibración particionan a la población objetivo. En algunos casos, la intersección del dominio con algunos grupos de calibración puede ser vacía.

De forma general, la población finita U, se encuentra particionada por I grupos de calibración, denotados como U_{C_i}, $(i = 1, ..., I)$ y en donde el dominio, U_d, puede intersectar a varios de ellos.

Se define la variable indicadora de pertenencia al $i-$ésimo grupo de calibración como

$$\delta_{C_{ik}} = \begin{cases} 1 & \text{si} \quad k \in U_{C_i} \\ 0 & \text{si} \quad k \notin U_{C_i} \end{cases} . \tag{4.24}$$

y el vector de información auxiliar viene dado como $\mathbf{x}_{C_i k} = \delta_{C_i k}\mathbf{x}_k$.

Se requiere, para la información auxiliar que:

- El vector de totales de las variables auxiliares $\mathbf{t}_{C_i x} = \sum_{U_{C_i}} \mathbf{x}_k = \sum_U \delta_{C_i k}\mathbf{x}_k = \sum_U \mathbf{x}_{C_i k}$ sea conocido para los I grupos de calibración.
- Para todo $k \in s$, el vector $\mathbf{x}_k$ y las I variables indicadoras de pertenencia a los grupos de calibración son conocidos.

Los vectores $\mathbf{x}_{C_i}$, conforman un vector, $\mathbf{x}_0$, el cual tiene una dimensión $J \times I$. Sea, $\mathbf{x}_{0k}$, el valor que toma $\mathbf{x}_0$ para el individuo k, el cual viene dado por

$$\mathbf{x}_{0k} = (\mathbf{x}_{C_1 k}, ..., \mathbf{x}_{C_i k}, ..., \mathbf{x}_{C_I k})'. \tag{4.25}$$

Por otro lado, el vector de totales poblaciones conocidos es

$$\mathbf{t}_{0x} = \sum_U \mathbf{x}_{0k} = \left(\sum_{U_{C_1}} \mathbf{x}_k, ..., \sum_{U_{C_i}} \mathbf{x}_k, ..., \sum_{U_{C_I}} \mathbf{x}_k\right)'. \tag{4.26}$$

Finalmente, el estimador *uni-weight*, queda definido como

$$\hat{t}_{d,calU} = \sum_s w_{0k} y_{dk} = \sum_{s_d} w_{0k} y_k, \tag{4.27}$$

donde

$$w_{0k} = a_k \left[1 + (\mathbf{t}_{0x} - \hat{\mathbf{t}}_{0x\pi})' \left(\sum_s a_k \mathbf{z}_k \mathbf{x'}_{0k}\right)^{-1} \mathbf{z}_k\right].$$

Capítulo 5

Una clase general de estimadores en dominios (basados en el diseño)

En este capítulo se presenta una clase general de estimadores de dominios, que comprende a los estimadores bajo el enfoque de calibración y regresión presentados en los capítulos previos.

Para ello es necesario adecuar los supuestos respecto a la información auxiliar disponible. De aquí en adelante se supone que se dispone de la siguiente información auxiliar:

1. Se conocen los totales del vector $\mathbf{x}_k$ de información auxiliar a nivel del subconjunto (o grupo) de calibración, $U_C \subset U$. O sea, $\sum_{U_C} \mathbf{x}_k$ es conocido.

2. El vector $\mathbf{x}_k$ es conocido para todos los individuos incluidos en la muestra s. A su vez, la variable indicadora de pertenencia al subconjunto U_C

$$\delta_{Ck} = \begin{cases} 1 & \text{si} \quad k \in U_C \\ 0 & \text{si} \quad k \notin U_C \end{cases}, \tag{5.1}$$

   es conocida $\forall k \in s$.

Los totales poblacionales de las variables auxiliares $\mathbf{t}_{Cx} = \sum_{U_C} \mathbf{x}_k = \sum_U \delta_{Ck}\mathbf{x}_k = \sum_U \mathbf{x}_{Ck}$, son conocidos mientras que los totales a nivel del dominio, $\mathbf{t}_{dx} = \sum_{U_d} \mathbf{x}_k = \sum_U \delta_{dk}\mathbf{x}_k = \sum_U \mathbf{x}_{dk}$, son desconocidos, a menos en la situación que $U_C = U_d$.

De lo anterior se distinguen dos casos especiales:

1. El dominio en sí es un grupo de calibración, $U_d = U_C \subset U$.

2. Toda la población es un grupo de calibración y el dominio de interés U_d se encuentra contenido en el grupo de calibración, $U_d \subset U_C = U$.

El problema puede resumirse como sigue. Para el dominio $U_d \subseteq U$ se busca estimar el total desconocido de la variable y en el dominio U_d, $t_d = \sum_{U_d} y_k = \sum_U y_{dk}$. Para dicho propósito

se encuentra disponible información de la variable de interés y de las variables auxiliares para las unidades incluidas en la muestra, $(\mathbf{x}_k, y_k) \ \ \forall k \in s$, y los totales poblacionales de las variables auxiliares a nivel del grupo de calibración, $\mathbf{t}_{Cx}$.

5.1. Enfoque de los estimadores calibrados

Bajo este enfoque se construye un sistema de ponderadores calibrados de la forma, $w_{Ck} = a_k(1 + \lambda'_C \mathbf{z}_k) \ \ \forall k \in s$, donde λ'_C se determina para que se cumpla la ecuación de calibración, $\sum_{s_C} w_{Ck}\mathbf{x}_k = \sum_{U_C} \mathbf{x}_k$ y $\mathbf{z}_k$, es un vector con valores definidos para todos los individuos de la muestra y tiene la misma dimensión que el vector de información auxiliar $\mathbf{x}_k$.

Dicho sistema de ponderadores calibrados son aplicados a la variable de dominio extendida $y_{dk} = \delta_{dk} y_k$.

El estimador calibrado queda definido como

$$\hat{t}_{d,calC} = \sum\nolimits_s w_{Ck} y_{dk} = \sum\nolimits_{s_d} w_{Ck} y_k, \tag{5.2}$$

donde $w_{Ck} = a_k g_{Ck}$ y $g_{Ck} = 1 + \lambda'_C \mathbf{z}_k = 1 + \left[(\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})' \left(\sum_s a_k \mathbf{z}_k \mathbf{x}'_{Ck} \right)^{-1} \right] \mathbf{z}_k$, con $\mathbf{x}_{Ck} = \delta_{Ck}\mathbf{x}_k$, $\mathbf{t}_{Cx} = \sum_U \mathbf{x}_{Ck}$ y $\hat{\mathbf{t}}_{Cx\pi} = \sum_U a_k \mathbf{x}_{Ck}$.

El estimador calibrado $\hat{t}_{d,calC}$ puede escribirse como el estimador π más un término de ajuste

$$\hat{t}_{d,calC} = \hat{t}_{d\pi} + (\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})' \hat{\mathbf{R}}_C, \tag{5.3}$$

donde

$$\hat{\mathbf{R}}_C = \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x}'_{Ck} \right)^{-1} \left(\sum\nolimits_s a_k \mathbf{z}_k y_{dk} \right). \tag{5.4}$$

El estimador calibrado $\hat{t}_{d,calC}$ posee las siguientes propiedades:

1. Consistente en el diseño y *very nearly design unbiased* [1].
2. Es directo, ya que solo valores y_k del dominio U_d son utilizados para la estimación.
3. Diferentes elecciones del vector $\mathbf{z}_k$ producen diferentes sistemas de ponderadores. La alternativa usual es seleccionar $\mathbf{z}_k$ como $\mathbf{x}_k$.

[1] Estevao y Särndal (2004) utilizan el término *very nearly design unbiased* si a medida que el tamaño de muestra n tiende a infinito, la razón de sesgo (sesgo dividido por la desviación estándar del estimador) es $O(n^{-1/2})$.

Observación 5.1.1 Si el grupo de calibración es toda la población, o sea $U_C = U$, el estimador $\hat{t}_{d,calC}$ es idéntico al estimador *uni-weight* de Estevao y Särndal (1999) de la ecuación (4.21). En tanto si el grupo de calibración es el dominio, o sea $U_C = U_d$, entonces $\hat{t}_{d,calC}$ es idéntico al estimador $\hat{t}_{d,calD}$ de la ecuación (4.16) . □

5.2. Enfoque de regresión

En el capítulo tres se introdujeron los estimadores de regresión ya sea directos o indirectos, en donde la diferencia entre ambos recae en el tipo de modelo que asiste al estimador. Ambos estimadores, independientemente del modelo propuesto, necesitan información auxiliar específica del dominio para su construcción. Si la información auxiliar no se encuentra disponible a un nivel tan desagregado igualmente es posible construir estimadores de regresión, si se modifican los requerimientos sobre la información auxiliar disponible.

En este enfoque el primer paso consiste en estimar el vector de los parámetros de regresión $\hat{\mathbf{B}}$. Este cálculo puede llevarse a cabo con diferentes niveles de desagregación de la población U, lo cual conduce a que exista una gama de diferentes especificaciones del modelo que implica distintos $\mathbf{B}$ (derivando en estimadores directos o indirectos).

El estimador para el total del dominio U_d, t_d, se construye como

$$\hat{t}_{d,gregC} = \hat{t}_{d\pi} + (\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})'\hat{\mathbf{B}}. \tag{5.5}$$

Lo anterior produce un estimador con menor varianza respecto al estimador $\hat{t}_{d\pi}$, siempre y cuando exista una correlación negativa entre el término correspondiente al estimador π, $\hat{t}_{d\pi}$, y el término de ajuste $(\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx})'\hat{\mathbf{B}}$, es *very nearly design unbiased* de cero. Así, la reducción de la varianza de este estimador depende de:

1. El grupo de calibración U_C utilizado para el cálculo del estimador.
2. La especificación del modelo, o sea, el nivel del coeficiente de regresión $\mathbf{B}$.

Estevao y Särndal (2004) indican que siempre y cuando sea posible el grupo de calibración U_C debe ser cercano al dominio de interés U_d. Lo ideal es que $U_C = U_d$. Si esto se cumple la información auxiliar disponible se encuentra a nivel del dominio, o sea $\mathbf{t}_{Cx} = \mathbf{t}_{dx}$. Estevao y Särndal (2004) indican que si el grupo de calibración U_C es un subconjunto más amplio, el efecto del término de ajuste puede llegar a ser muy pequeño y ocasionalmente puede conducir a que la varianza del estimador sea mas grande que la del estimador $\hat{t}_{d\pi}$ (el cual no utiliza ningún tipo de información auxiliar).

Si se supone que el dominio posee sus propias características y que estas difieren de la población en su conjunto se puede estimar el coeficiente de regresión a nivel de dominio. Para ello consideremos

$$\hat{\mathbf{B}}_d = \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x}'_{dk}\right)^{-1} \left(\sum\nolimits_s a_k \mathbf{z}_k y_{dk}\right), \tag{5.6}$$

luego, el estimador, basado en $\hat{\mathbf{B}}_d$, queda definido como

$$\hat{t}_{d,gregD_C} = \hat{t}_{d\pi} + (\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})'\hat{\mathbf{B}}_d. \tag{5.7}$$

Observación 5.2.1 Si el grupo de calibración coincide con el dominio, $U_C = U_d$ y $\mathbf{z}_k = \mathbf{x}_k/c_k$ el estimador $\hat{t}_{d,gregD_C}$, coincide con el estimador directo de regresión $\hat{t}_{d,gregD}$ de la ecuación (3.18). □

Observación 5.2.2 El estimador $\hat{t}_{d,gregD_C}$ es homogéneo

$$\hat{t}_{dgregD_C} = \sum\nolimits_s a_k g_{dkC} y_{dk}, \tag{5.8}$$

con

$$g_{dkC} = 1 + (\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})' \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x}'_{dk}\right)^{-1} \mathbf{z}_k. \tag{5.9}$$

□

Por construcción el estimador $\hat{t}_{d,gregD_C}$ es directo. Por otro lado, este estimador coincide con el estimador $\hat{t}_{d,calC}$ de la ecuación (5.3) cuando el dominio U_d es el grupo de calibración U_C.

En algunas ocasiones el tamaño de muestra efectivo en el dominio puede ser muy pequeño, lo que conlleva a que la estimación de los parámetros del modelo en el dominio puedan ser inestables. Lo anterior motiva a utilizar información de toda la muestra o de otros dominios, con el objetivo de poder realizar estimaciones más estables de los parámetros del modelo.

En este sentido, se puede plantear un modelo a nivel de toda la población, al igual que el empleado para el estimador indirecto de regresión $\hat{t}_{d,gregP}$. La estimación de los coeficientes de regresión, así como el estimador resultante, quedan definidos por

$$\hat{\mathbf{B}}_s = \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x}'_k\right)^{-1} \left(\sum\nolimits_s a_k \mathbf{z}_k y_k\right), \tag{5.10}$$

$$\begin{aligned}
\hat{t}_{d,gregP_C} &= \hat{t}_{d\pi} + (\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})'\hat{\mathbf{B}}_s \\
&= \sum\nolimits_s \left[\delta_{dk} + (\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})' \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x}'_k\right)^{-1} \mathbf{z}_k\right] a_k y_k \\
&= \sum\nolimits_s g_{dkC} a_k y_k.
\end{aligned}$$

El estimador $\hat{t}_{d,gregP_C}$ utiliza todas las unidades de la muestra, tanto aquellas incluidas en el dominio U_d, como aquellas que no pertenecen al mismo, por lo tanto el estimador es indirecto.

Observación 5.2.3 Si $\mathbf{z}_k = \mathbf{x}_k/c_k$ y $U_C = U_d$ el estimador $\hat{t}_{d,gregP_C}$ es idéntico al estimador indirecto de regresión, $\hat{t}_{d,gregP}$, de la ecuación (3.33) . □

5.3. Una clase general de estimadores

Bajo el enfoque de calibración y regresión, los estimadores son construidos en base a diferentes argumentos. Sin embargo, ambos son consistentes y aproximadamente insesgados, pero sus respectivas varianzas pueden diferir de forma considerable. A continuación se presenta una clase general de estimadores en dominios, que generaliza los estimadores presentados anteriormente.

El estimador se define como

$$\hat{t}_{d,gral} = \hat{t}_{d\pi} + (\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})'\hat{\mathbf{Q}}_{MLz}, \tag{5.11}$$

con

$$\hat{\mathbf{Q}}_{MLz} = \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x'}_{Mk}\right)^{-1} \left(\sum\nolimits_s a_k \mathbf{z}_k y_{Lk}\right), \tag{5.12}$$

donde $\mathbf{x}_{Mk} = \delta_{Mk}\mathbf{x}_k$, $y_{Lk} = \delta_{Lk}y_k$ y δ_{Mk} y δ_{Lk} son las variables indicadoras de pertenencia a los subconjuntos de la población, $U_M \subseteq U$ y $U_L \subseteq U$. La población objetivo U, el grupo de calibración, U_C y el dominio U_d se encuentran fijos. Mientras que $\hat{\mathbf{Q}}_{MLz}$ depende de el vector $\mathbf{z}_k$ y las subpoblaciones U_M y U_L.

Esta clase de estimadores comprende como casos particulares a los estimadores de los enfoques de calibración y de regresión:

- Si se fija $U_M = U_C$ y $U_L = U_d$, se obtiene el estimador bajo el enfoque de calibración $\hat{t}_{d,calC}$.
- Si se fija $U_L = U_M$ se obtiene el estimador bajo el enfoque de regresión $\hat{t}_{d,gregC}$.
- Si se fija $U_L = U_M = U_d$ se obtiene el estimador $\hat{t}_{d,gregD_C}$.
- Si se fija $U_L = U_M = U$ se obtiene el estimador $\hat{t}_{d,gregP_C}$.

El primer sumando del estimador de la ecuación (5.11) corresponde al estimador π, $\hat{t}_{d\pi}$, el cual es insesgado para estimar el total t_d y el segundo término $(\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})'\hat{\mathbf{Q}}_{MLz}$ es *very nearly design*

unbiased de cero. Lo anterior lleva a que el estimador $\hat{t}_{d,gral}$ sea *very nearly design unbiased*.

El error del estimador $\hat{t}_{d,gral}$, se define como

$$\hat{t}_{d,gral} - t_d = \hat{t}_{d\pi} - t_d + (\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})'\hat{\mathbf{Q}}_{MLz}. \tag{5.13}$$

El obstáculo que se presenta en el análisis de la ecuación (5.13) se encuentra en el último término, en donde $(\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})'\hat{\mathbf{Q}}_{MLz}$ es un término no lineal. La no linealidad de $(\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})'\hat{\mathbf{Q}}_{MLz}$ deja de ser un obstáculo si se consigue reemplazar, con un pequeño error, el vector aleatorio $\hat{\mathbf{Q}}_{MLz}$ por un vector constante.

Lo anterior se puede lograr centrando $\hat{\mathbf{Q}}_{MLz}$ en el vector de constantes

$$\mathbf{Q}_{MLz} = \left(\sum\nolimits_U \mathbf{z}_k \mathbf{x}'_{Mk}\right)^{-1} \left(\sum\nolimits_U \mathbf{z}_k y_{Lk}\right). \tag{5.14}$$

Reemplazando $\hat{\mathbf{Q}}_{MLz}$ por $\mathbf{Q}_{MLz} + \left(\hat{\mathbf{Q}}_{MLz} - \mathbf{Q}_{MLz}\right)$ y reordenando los términos de la ecuación (5.13) se obtiene

$$\hat{t}_{d,gral} - t_d = \hat{t}_{CE\pi} - t_{CE} - (\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})' \left(\hat{\mathbf{Q}}_{MLz} - \mathbf{Q}_{MLz}\right), \tag{5.15}$$

donde $\hat{t}_{CE\pi} = \sum_s a_k E_{Ck}$ y $t_{CE} = \sum_U E_{Ck}$, con

$$E_{Ck} = \begin{cases} y_k - \mathbf{x}'_k \mathbf{Q}_{MLz} & \text{si} \quad k \in U_d \\ -\mathbf{x}'_k \mathbf{Q}_{MLz} & \text{si} \quad k \in U_C - U_d \\ 0 & \text{si} \quad k \notin U_C \end{cases}. \tag{5.16}$$

En la ecuación (5.15) las diferencias $\hat{t}_{CE\pi} - t_{CE}$ y $(\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx})$ tienden a cero y tienen el mismo orden en probabilidad, debido a que cuando se multiplican por N^{-1}, bajo condiciones generales, cada una de ellas es $O_p(n^{-1/2})$.

Por otro lado el término $\left(\hat{\mathbf{Q}}_{MLz} - \mathbf{Q}_{MLz}\right)$ es cercano a cero y con el mismo orden en probabilidad, $O_p(n^{-1/2})$.

Finalmente, el producto $N^{-1}(\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})' \left(\hat{\mathbf{Q}}_{MLz} - \mathbf{Q}_{MLz}\right)$ es $O_p(n^{-1})$, y de menor orden que el término $N^{-1}(\hat{t}_{CE\pi} - t_{CE})$. Este último término, por sí solo, provee la aproximación lineal buscada

$$N^{-1}(\hat{t}_{d,gral} - t_d) = N^{-1}(\hat{t}_{CE\pi} - t_{CE}) + O_p(n^{-1}) \doteq N^{-1}(\hat{t}_{CE\pi} - t_{CE}). \tag{5.17}$$

El sesgo del estimador $\hat{t}_{d,gral}$, es aproximadamente cero, lo anterior se debe a que $E(\hat{t}_{CE\pi}) = t_{CE}$, una expresión exacta del mismo viene dada por

$$B(\hat{t}_{d,gral}) = E(\hat{t}_{d,gral}) - t_{d,gral} = -E\left[(\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})'\left(\hat{\mathbf{Q}}_{MLz} - \mathbf{Q}_{MLz}\right)\right]. \tag{5.18}$$

Para el sesgo se tiene que $N^{-1}B(\hat{t}_{d,gral}) = O(n^{-1})$ y para la varianza $N^{-2}V(\hat{t}_{d,gral}) = O(n^{-1})$. Entonces, la razón de sesgo del estimador $\hat{t}_{d,gral}$ es $O(n^{-1/2})$, y los requerimientos para realizar inferencias son alcanzados inclusive con tamaños de muestras modestos no perturban seriamente la validez de los intervalos de confianza. Así, $\hat{t}_{d,gral}$ es consistente en el diseño y *very nearly design unbiased*.

La varianza asintótica del estimador $\hat{t}_{d,gral}$, es

$$AV(\hat{t}_{d,gral}) = V(\hat{t}_{CE\pi}) = \sum\sum\nolimits_U \Delta_{kl}\frac{E_{Ck}}{\pi_k}\frac{E_{Cl}}{\pi_l}. \tag{5.19}$$

Un estimador para la varianza del estimador $\hat{t}_{d,gral}$, se obtiene de

$$\hat{V}(\hat{t}_{d,gral}) = \sum\sum\nolimits_s \frac{\Delta_{xl}}{\pi_{kl}}(w_{Ck}e_{Ck})(w_{Cl}e_{Cl}), \tag{5.20}$$

donde los residuos muestrales vienen dados por

$$e_{Ck} = \begin{cases} y_k - \mathbf{x'}_k\hat{\mathbf{Q}}_{MLz} & \text{si} \quad k \in U_d \\ -\mathbf{x'}_k\hat{\mathbf{Q}}_{MLz} & \text{si} \quad k \in U_C - U_d \\ 0 & \text{si} \quad k \notin U_C \end{cases}, \tag{5.21}$$

y los ponderadores son

$$w_{Ck} = a_k\left[\delta_{dk} + \left(\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi}\right)'\left(\sum\nolimits_s a_k\mathbf{z}_k\mathbf{x'}_{Mk}\right)^{-1}\delta_{Lk}\mathbf{z}_k\right]. \tag{5.22}$$

5.3.1. Mínima varianza asintótica

En la ecuación (5.11) se utiliza la información auxiliar respecto al grupo de calibración U_C para la construcción del estimador, por su parte $\hat{\mathbf{Q}}_{MLz}$ depende de las dos subpoblaciones U_M y U_L, y el vector $\mathbf{z}_k$ los cuales deben ser especificados. El problema entonces es encontrar el vector $\mathbf{Q}_{MLz}$ que minimice la varianza $V(\hat{t}_{CE\pi})$, lo que se traduce en encontrar las mejores opciones para las subpoblacionales U_M y U_L, y el vector $\mathbf{z}_k$.

Estevao y Särndal (2004), demostraron que el estimador asintotícamente óptimo bajo esta clase de estimadores, se obtiene eligiendo al grupo de calibración U_C como la subpoblación U_M, al dominio U_d como la subpoblación U_L y al vector $\mathbf{z}_k = \pi_k \sum_{l \in s}(a_k a_l - 1/\pi_{kl})\mathbf{x}_{C_l}$.

Se obtiene, entonces,

$$\hat{t}_{d,gral} = \hat{t}_{d\pi} + (\mathbf{t}_{Cx} - \hat{\mathbf{t}}_{Cx\pi})'\hat{\mathbf{Q}}_{U_C U_d z}, \tag{5.23}$$

donde

$$\hat{\mathbf{Q}}_{U_C U_d z} = \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x}'_{Ck}\right)^{-1} \left(\sum\nolimits_s a_k \mathbf{z}_k y_{dk}\right). \tag{5.24}$$

De esta manera, el estimador $\hat{t}_{d,gral}$ coincide con el estimador calibrado $\hat{t}_{d,calC}$ de la ecuación (5.3). Ninguno de los estimadores bajo el enfoque de regresión posee menor varianza asintótica, al menos en el caso en donde el grupo de calibración U_C es el dominio de interés U_d y a su vez en esa situación, los estimadores $\hat{t}_{d,gregD_C}$ y $\hat{t}_{d,calC}$ son iguales y por lo tanto poseen la misma varianza asintótica. La elección del vector instrumental $\mathbf{z}_k = \pi_k \sum_{l \in s}(a_k a_l - 1/\pi_{kl})\mathbf{x}_{C_l}$ puede ser en algunos casos muy inestable, una alternativa simple sin necesariamente pérdida de precisión puede ser eligiendo $\mathbf{z}_k = \mathbf{x}_{Ck}$.

Capítulo 6

Estimadores sintéticos

6.1. Introducción

En las situaciones donde el tamaño de muestra en un dominio es pequeño los estimadores basados en el diseño pueden presentar problemas. Si bien bajo dicho enfoque los estimadores son aproximadamente insesgados, la varianza puede ser excesiva, de manera de no permitir intervalos de confianza con niveles de confianza y precisión razonables. A su vez, en el caso extremo que el tamaño de muestra efectivo en un dominio sea nulo, no es posible obtener estimaciones. En estas situaciones es necesario recurrir a los estimadores basados en modelos, los cuales suponen un modelo que relaciona el dominio de interés con otros subconjuntos (o dominios) de la población. De esta forma, se utiliza información de la variable de interés de individuos incluidos en la muestra, que no pertenecen al dominio de interés, con el objetivo de aumentar el tamaño de muestra utilizado para realizar las estimaciones. Este concepto también es utilizado en los estimadores basados en el diseño, por ejemplo si se utiliza un modelo a nivel de toda la población para asistir al estimador de regresión, en donde para la estimación de los parámetros del modelo intervienen todos los individuos de la muestra.

Según Gonzalez (1973), un estimador es llamado sintético si utiliza un estimador confiable de un dominio suficientemente amplio (o todo el universo) el cual incluye otros dominios más pequeños y dicho estimador se utiliza para estimar indirectamente el dominio pequeño, bajo el supuesto de que el dominio pequeño posee las mismas características que el dominio más amplio.

Los estimadores sintéticos son utilizados en la práctica debido a su fácil implementación y adaptación a cualquier tipo de diseño de muestreo y su potencial para reducir la variabilidad de las estimaciones basándose en información de otros dominios similares.

La varianza de los estimadores sintéticos generalmente es pequeña en relación a los estimadores basados en el diseño. El costo de la reducción de la variabilidad del estimador, deriva en un aumento del sesgo del estimador. De esta manera, los estimadores tienen una varianza pequeña respecto a aquellos basados en el diseño pero mayor sesgo. Los estimadores sintéticos son basados en el

modelo y por lo tanto si el modelo no es verdadero, no tendrá buenas propiedades.

Para introducir el estimador sintético, supongamos que el objetivo es estimar la media de la variable de interés y en el dominio U_d y no se cuenta con ningún tipo de información auxiliar. La primera opción es utilizar el estimador, $\tilde{y}_{s_d} = \sum_{s_d} a_k y_k / \hat{N}_d$, siempre y cuando exista al menos un individuo en la muestra perteneciente al dominio U_d. Por otro lado, si el tamaño de muestra es muy pequeño, por ejemplo, uno o dos casos efectivos, el estimador puede ser muy inestable. Ahora bien, supongamos que existe la creencia que la media desconocida del dominio es similar a la media de un subconjunto de la población. Sin pérdida de generalidad, supongamos que dicho conjunto se trata de toda la población, y se considera que, $\tilde{y}_s = \sum_s a_k y_k / \hat{N}$, como estimador de la media poblacional, $\bar{y}_U$. Entonces, de forma de aumentar el tamaño de muestra efectivo para calcular la estimación de la media del dominio, se puede utilizar un modelo implícito, el cual supone que la media del dominio es similar a la media de la población. Bajo este modelo el estimador sintético es

$$\hat{\bar{y}}_{d,S} = \tilde{y}_s = \sum\nolimits_s a_k y_k / \hat{N}.$$

El sesgo del estimador sintético $\hat{\bar{y}}_{d,S}$ es aproximadamente igual a $B(\hat{\bar{y}}_{d,S}) \doteq \bar{y}_U - \bar{y}_{U_d}$, el cual puede ser relativamente pequeño si el modelo es verdadero. Si lo anterior se cumple, el estimador sintético será muy eficiente debido a que su error cuadrático medio será pequeño, esto se debe a que la varianza del estimador $\tilde{y}_s$ es relativamente pequeña debido a que para su cálculo se utiliza toda la muestra. Por otro lado, si el modelo es falso, el estimador sintético será sesgado y el sesgo dominará al error cuadrático medio y los intervalos de confianza basados en su cálculo no tendrán el nivel de cobertura deseado.

6.2. Estimador sintético en el contexto de los estimadores de regresión.

Al igual que en los capítulos anteriores, se supone que se encuentra disponible información auxiliar específica del dominio. El objetivo es obtener una estimación del total del dominio, $t_d = \sum_{s_d} y_k$. El modelo utilizado para la construcción del estimador sintético es definido a nivel de una subpoblación que incluye al dominio de interés supongamos que dicho subconjunto se trata de toda la población. Supongamos entonces el modelo $E_m(y_k) = \mathbf{x'}_k \boldsymbol{\beta}, \;\; V_m(y_k) = c_k \;\; \forall k \in U$. Con el modelo estimado, se calculan las predicciones de la variable objetivo, y, para todos los individuos pertenecientes al dominio U_d y la suma de todos estos valores definen al estimador sintético de regresión

$$\hat{t}_{d,gregS} = \sum\nolimits_{U_d} \mathbf{x'}_k \hat{\mathbf{B}} = \sum\nolimits_{U_d} \hat{y}_k. \tag{6.1}$$

Por construcción, y al igual que todos los estimadores basados en modelos, este estimador es indirecto, lo cual se debe a que en el proceso de estimación de los parámetros del modelo, intervienen todos los individuos incluidos en la muestra y no solo aquellos que pertenecen al dominio. Por lo tanto el estimador sintético puede ser calculado incluso si el tamaño de muestra en el dominio es nulo.

Observación 6.2.1 El estimador sintético $\hat{t}_{d,gregS}$ cumple la propiedad de aditividad cuando $c_k = \lambda' \mathbf{x}_k$, con λ un vector de constantes conocidas . □

Observación 6.2.2 Si el tamaño de muestra en el dominio es cero o $\sum_{s_d} a_k e_k = 0$, el estimador $\hat{t}_{d,gregS}$ es idéntico al estimador indirecto de regresión $\hat{t}_{d,gregP}$ de la ecuación (3.33) . □

El rol del modelo propuesto en el estimador sintético difiere al del estimador de regresión, en este último el fin es asistir al estimador y como consecuencia el mismo es aproximadamente insesgado independiente si el modelo propuesto es verdadero o no. Si el modelo tiene un pobre poder de ajuste o el tamaño de muestra en el dominio es reducido, deriva en que la varianza del estimador de regresión sea grande. En el caso del estimador sintético si el modelo no es verdadero el estimador será sesgado, debido a que el término $\sum_{s_d} a_k e_k$, el cual proteje al estimador de regresión si el modelo utilizado no es verdadero, no se encuentra presente en el estimador sintético.

Si bien el requisito de un tamaño de muestra determinado para el dominio no es necesario, la información auxiliar disponible deber ser poderosa y es importante (aún más que en los estimadores basados en el diseño) para que el sesgo del estimador sea pequeño. En la práctica, dicha situación es poco común por lo que el modelo utilizado generalmente no tiene un buen poder de ajuste, derivando en que el estimador sea sesgado.

El sesgo del estimador sintético de regresión es

$$B(\hat{t}_{d,gregS}) = E(\hat{t}_{d,gregS}) - t_d.$$

Reescribiendo $t_d = \sum_{U_d} \mathbf{x'}_k \mathbf{B} + \sum_{U_d} E_k$, donde $E_k = y_k - \mathbf{x'}_k \mathbf{B}$ son los errores a nivel poblacional, el sesgo del estimador sintético viene dado por

$$B(\hat{t}_{d,gregS}) = E\left(\sum_{U_d} \mathbf{x'}_k \hat{\mathbf{B}}\right) - \sum_{U_d} \mathbf{x'}_k \mathbf{B} - \sum_{U_d} E_k \doteq - \sum_{U_d} E_k. \tag{6.2}$$

El sesgo puede ser estimado por

$$\hat{B}(\hat{t}_{d,gregS}) = - \sum_{s_d} a_k e_k, \tag{6.3}$$

donde $e_k = y_k - \mathbf{x'}_k\hat{\mathbf{B}}$.

El uso del estimador sintético se da en los casos donde el tamaño de muestra en el dominio es nulo o muy pequeño. Si el tamaño es nulo, no es posible calcular $\hat{B}(\hat{t}_{d,gregS})$ y si el tamaño de muestra es muy pequeño, la estimación del mismo puede ser muy inestable.

Un estimador de la varianza del estimador sintético $\hat{t}_{d,gregS}$, viene dado por

$$\hat{V}(\hat{t}_{d,gregS}) = \sum\nolimits_{U_d} \mathbf{x'}_k \hat{V}(\hat{\mathbf{B}})\mathbf{x}_k, \tag{6.4}$$

donde

$$\hat{V}(\hat{\mathbf{B}}) = \left(\sum\nolimits_s a_k \mathbf{x}_k \mathbf{x'}_k\right)^{-1} \hat{\mathbf{V}} \left(\sum\nolimits_s a_k \mathbf{x}_k \mathbf{x'}_k\right)^{-1}, \tag{6.5}$$

y $\hat{\mathbf{V}}$ es una matriz simétrica de $J \times J$ de elemento genérico

$$\hat{v}_{jj'} = \sum\sum\nolimits_s \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{x_{jk}e_k}{\pi_k}\right)\left(\frac{x_{j'l}e_l}{\pi_l}\right). \tag{6.6}$$

Ver Särndal et al. (1992).

6.3. Casos particulares del estimador sintético

A continuación se presentan casos particulares del estimador sintético junto con una expresión del sesgo si el modelo utilizado para su construcción no es correcto.

Ejemplo 6.3.1 Un caso sencillo es considerar una única variable auxiliar x, en donde el modelo a nivel poblacional cumple que $E_m(y_k) = \beta x_k,\ \ V_m(y_k) = c_k = \lambda' x_k\ \ \forall k \in U$.

El estimador sintético de razón es definido como

$$\hat{t}_{d,raS} = \sum\nolimits_{U_d} \hat{y}_k = \sum\nolimits_{U_d} x_k \hat{B} \tag{6.7}$$

$$= \left(\sum\nolimits_{U_d} x_k\right) \frac{\sum_s a_k y_k}{\sum_s a_k x_k}. \tag{6.8}$$

La varianza de la pendiente $\hat{B}$ es generalmente pequeña debido a que para su estimación se utilizan los individuos de toda la muestra. Por lo tanto, por construcción, el estimador es indirecto.

Este estimador es generalmente sesgado y su sesgo aproximado viene dado por

$$B(\hat{t}_{d,raS}) = E(\hat{t}_{d,raS}) - t_d \doteq -\sum\nolimits_{U_d} E_k = -\left(\sum\nolimits_{U_d} x_k\right)(B_d - B)\,, \tag{6.9}$$

donde $B_d = \sum_{U_d} y_k / \sum_{U_d} x_k$ es la pendiente específica del dominio U_d y $B = \sum_U y_k / \sum_U x_k$ es la pendiente para toda la población U.

En el caso que la pendiente específica del dominio B_d sea aproximadamente igual a B, el sesgo del estimador será pequeño, en tanto si las diferencias entre las pendientes son considerables el sesgo puede ser sustancialmente grande y dominará en la expresión del error cuadrático medio. □

Observación 6.3.1 El estimador de razón sintético cumple la propiedad de aditividad

$$\hat{t}_{ra} = \sum_{d=1}^{D} \hat{t}_{d,raS} = \sum_{d=1}^{D} \sum\nolimits_{U_d} \hat{y}_k = \hat{B} \sum_{d=1}^{D} \sum\nolimits_{U_d} x_k = \frac{\sum_s y_k}{\sum_s x_k} \sum\nolimits_U x_k.$$

□

Ejemplo 6.3.2 Si se considera el modelo de medias por grupo $E_m(y_k) = \beta_g,\ V(y_k) = c_g\ \forall k \in U_g,\ \ \forall g = 1, ..., G$. El estimador sintético de un modelo de media común, queda expresado de la forma

$$\hat{t}_{d,gregS} = \sum_{g=1}^{G} N_{dg} \tilde{y}_{s_g}. \tag{6.10}$$

La esperanza del estimador es $E(\hat{t}_{d,gregS}) \doteq \sum_{g=1}^{G} N_{dg} \bar{y}_{U_g}$ y el sesgo aproximado viene dado por

$$B(\hat{t}_{d,gregS}) \doteq -\sum\nolimits_{U_d} E_k = -\sum_{g=1}^{G} N_{dg}(\bar{y}_{U_{dg}} - \bar{y}_{U.g}). \tag{6.11}$$

La varianza del estimador, generalmente es pequeña en relación a la varianza del estimador de la ecuación (4.18) . Lo anterior se debe a que la estimación de la media del grupo $\tilde{y}_{s.g}$ es determinada con buena precisión suponiendo que el tamaño muestral del grupo es grande y quizás por la pequeña varianza de la variable de interés y, dentro del grupo. En tanto, el término que proteje al estimador si el modelo no es verdadero

$$\sum_{g=1}^{G} \hat{N}_{dg}(\tilde{y}_{s_{dg}} - \tilde{y}_{s.g}),$$

no se encuentra presente y generalmente posee una proporción considerable de la varianza del estimador. De todas formas, un estimador con un sesgo relativo alto no permite calcular intervalos de confianza con niveles de cobertura apropiados ya que su validez queda sujeta al cumplimiento

del modelo supuesto. □

Ejemplo 6.3.3 De manera general, si se considera el modelo de razón por grupo $E_m(y_k) = \beta_g x_k,\ \ V_m(y_k) = c_k = \lambda' x_k,\ \ \forall k \in U_g,\ \ \forall g = 1, ..., G$ y el estimador sintético queda expresado de la forma

$$\hat{t}_{d,raPGS} = \sum_{g=1}^{G} \sum\nolimits_{s_{dg}} \hat{y}_k = \sum_{g=1}^{G} \sum\nolimits_{s_{dg}} x_k \hat{B}_g = \sum_{g=1}^{G} t_{dgx} \frac{\hat{t}_{g\pi}}{\hat{t}_{gx\pi}}, \tag{6.12}$$

y el sesgo es

$$B(\hat{t}_{d,raPGS}) \doteq -\sum\nolimits_{U_d} E_k = -\sum_{g=1}^{G} t_{dgx}(B_{dg} - B_g). \tag{6.13}$$

□

6.4. Estimación del error cuadrático medio

Los estimadores basados en el diseño son aproximadamente insesgados y el error cuadrático medio (*ECM*) se reduce a su varianza. Por otro lado, los estimadores sintéticos son sesgados y por tanto es relevante el análisis de su *ECM*.

Así, para los estimadores sintéticos, el problema se centra en encontrar un estimador del error cuadrático medio

$$ECM(\hat{t}_{d,S}) = E(\hat{t}_{d,S} - t_d)^2 = V(\hat{t}_{d,S}) + B^2(\hat{t}_{d,S}).$$

con $B^2(\hat{t}_{d,S}) = (E(\hat{t}_{d,S}) - t_d)^2$.

El primer sumando correspondiente a la varianza del estimador, puede ser estimado utilizando métodos clásicos como linealización de Taylor o métodos de remuestreo, por ejemplo, *Bootstrap* o *Jackknife*.

El *ECM* se puede escribir utilizando un estimador insesgado del total del dominio, el cual puede ser cualquiera basado en el diseño, sin pérdida de generalidad, utilizamos el estimador π, $\hat{t}_{d\pi} = \sum_{s_d} a_k y_k$.

Luego

$$\begin{aligned} ECM(\hat{t}_{d,S}) &= E(\hat{t}_{d,S} - \hat{t}_{d\pi} + \hat{t}_{d\pi} - t_d)^2 \\ &= E(\hat{t}_{d,S} - \hat{t}_{d\pi})^2 - V(\hat{t}_{d\pi}) + 2COV(\hat{t}_{d,S}, \hat{t}_{d\pi}) \\ &= E(\hat{t}_{d,S} - \hat{t}_{d\pi})^2 - V(\hat{t}_{d,S} - \hat{t}_{d\pi}) + V(\hat{t}_{d,S}). \end{aligned} \tag{6.14}$$

Un estimador aproximadamente insesgado de la expresión anterior viene dado por

$$\widehat{ECM}(\hat{t}_{d,S}) = (\hat{t}_{d,S} - \hat{t}_{d\pi})^2 - \hat{V}(\hat{t}_{d,S} - \hat{t}_{d\pi}) + \hat{V}(\hat{t}_{d,S}). \tag{6.15}$$

La estimación del *ECM* de la ecuación (6.15) se calcula utilizando generalmente técnicas de remuestreo. Sin embargo, el problema se encuentra en la inestabilidad que posee la ecuación (6.15), dado que en algunos casos el término $\hat{V}(\hat{t}_{d,S} - \hat{t}_{d\pi})$ puede ser muy grande.

Existen muchas propuestas para intentar resolver el problema de la inestabilidad de la ecuación (6.15). Por ejemplo, Rao (2003), menciona a Gonzalez y Waksberg (1973), que proponen tomar una media de los errores cuadráticos medios de los dominios en el caso que se utilice el mismo estimador sintético para estimar un conjunto de dominios. Supongamos que se estima la media $\bar{y}_{U_d}$, para D dominios de la población, entonces el estimador sintético se expresa como $\hat{\bar{y}}_{d,S} = \hat{t}_{d,S}/N_d$, en donde el tamaño del dominio es conocido y la estimación del error cuadrático medio es

$$\widehat{ECM}(\hat{\bar{y}}_{d,S}) = \widehat{ECM}(\hat{t}_{d,S})/N_d^2.$$

Una aproximación viene dada por

$$\widehat{ECM}_w(\hat{\bar{y}}_{d,S}) = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{N_d^2}(\hat{t}_{d,S} - \hat{t}_{d\pi})^2 - \frac{1}{D}\sum_{d=1}^{D}\frac{1}{N_d^2}\hat{V}(\hat{t}_{d,S} - \hat{t}_{d\pi}) + \frac{1}{D}\sum_{d=1}^{D}\frac{1}{N_d^2}\hat{V}(\hat{t}_{d,S}). \tag{6.16}$$

Esta medida global de incertidumbre puede ser engañosa ya que se refiere a un promedio de los errores cuadráticos medios y no a los de un dominio específico.

Teniendo en cuenta que la varianza del estimador sintético es generalmente pequeña respecto a la varianza del estimador π, se puede aproximar la ecuación (6.15) por

$$\widehat{ECM}(\hat{t}_{d,S}) \doteq (\hat{t}_{d,S} - \hat{t}_{d\pi})^2 - \hat{V}(\hat{t}_{d\pi}). \tag{6.17}$$

Utilizando la aproximación de la ecuación (6.17) en la ecuación (6.16) se obtiene

$$\widehat{ECM}_w(\hat{\bar{y}}_{d,S}) \doteq \frac{1}{D}\sum_{d=1}^{D}\frac{1}{N_d^2}(\hat{t}_{d,S}-\hat{t}_{d\pi})^2 - \frac{1}{D}\sum_{d=1}^{D}\frac{1}{N_d^2}\hat{V}(\hat{t}_{d\pi}). \tag{6.18}$$

Por otro lado, Rao (2003) menciona a Marker (1995), el cual, propuso un método simple para obtener una estimación del *ECM* para un dominio U_d, bajo la hipótesis que el sesgo al cuadrado del estimador, $B^2(\hat{\bar{y}}_{d,S})$, es aproximadamente igual al promedio de los sesgos cuadrados en los dominios. Entonces, se tiene que

$$B_w^2(\hat{\bar{y}}_{d,S}) = \frac{1}{D}\sum_{d=1}^{D}B^2(\hat{\bar{y}}_{d,S}). \tag{6.19}$$

La estimación del promedio de los sesgos al cuadrado viene dada por

$$\hat{B}_w^2(\hat{\bar{y}}_{d,S}) = \widehat{ECM}_w(\hat{\bar{y}}_{d,S}) - \frac{1}{D}\sum_{d=1}^{D}\hat{V}(\hat{\bar{y}}_{d,S}). \tag{6.20}$$

Finalmente, bajo las hipótesis de la ecuación (6.19) el *ECM* del estimador sintético para el total del dominio U_d puede ser estimado como

$$\widehat{ECM}(\hat{t}_{d,S}) = \hat{V}(\hat{t}_{d,S}) + N_d^2\hat{B}_w^2(\hat{\bar{y}}_{d,S}). \tag{6.21}$$

Capítulo 7

Estimadores compuestos

7.1. Introducción

Los estimadores compuestos intentan abarcar los beneficios de los estimadores basados en el diseño y los basados en modelos. Si el tamaño de muestra en el dominio es nulo, la única alternativa vista hasta ahora es utilizar un estimador sintético. En cambio, si el tamaño de muestra es reducido (pero no nulo), se puede construir un estimador como combinación lineal convexa de un estimador basado en el diseño ($\hat{t}_d$) y un estimador sintético ($\hat{t}_{d,S}$).

Un estimador compuesto para el total de un dominio, $t_d = \sum_{U_d} y_k$, se define como

$$\hat{t}_{d,C} = \phi_d \hat{t}_d + (1 - \phi_d)\hat{t}_{d,S}, \tag{7.1}$$

con $0 \leq \phi_d \leq 1$.

Para adoptar la estrategia anterior, hay que resolver dos problemas (no necesariamente independientes):

(i) Cuales son los estimadores a considerar.

(ii) Como elegir ϕ_d.

Una solución para el punto (ii), es elegir los ponderadores ϕ_d, de forma de minimizar el error cuadrático medio del estimador compuesto.

Como se trata de estimadores sesgados, la comparación entre posibles competidores debe basarse en el *ECM* del estimador $\hat{t}_{d,C}$,

$$ECM(\hat{t}_{d,C}) = \phi_d^2 ECM(\hat{t}_d) + (1 - \phi_d)^2 ECM(\hat{t}_{d,S}) + 2\phi_d(1 - \phi_d)E\left[(\hat{t}_d - t_d)(\hat{t}_{d,S} - t_d)\right]. \tag{7.2}$$

Minimizando (7.2) respecto a ϕ_d, se obtiene

$$\phi_d^* = \frac{ECM(\hat{t}_{d,S}) - E\left[(\hat{t}_d - t_d)(\hat{t}_{d,S} - t_d)\right]}{ECM(\hat{t}_d) + ECM(\hat{t}_{d,S}) - 2E\left[(\hat{t}_d - t_d)(\hat{t}_{d,S} - t_d)\right]}. \tag{7.3}$$

Asumiendo que el término $E\left[(\hat{t}_d - t_d)(\hat{t}_{d,S} - t_d)\right]$ es despreciable (de orden de magnitud inferior) respecto a $ECM(\hat{t}_{d,S})$, entonces, una solución aproximada viene dada por

$$\phi_d^* \doteq \frac{ECM(\hat{t}_{d,S})}{ECM(\hat{t}_d) + ECM(\hat{t}_{d,S})}, \tag{7.4}$$

en donde la participación del estimador basado en el diseño, $\hat{t}_d$, está sujeta al *ECM* del estimador sintético. Si el *ECM* del estimador sintético es pequeño en comparación al *ECM* del estimador basado en el diseño, el ponderador ϕ_d sera pequeño, aumentando así la participación del estimador sintético $\hat{t}_{d,S}$ en el estimador compuesto.

En la práctica, una estimación del ponderador ϕ_d^*, de la ecuación (7.4), utilizando el resultado de la ecuación (6.17) viene dada por

$$\hat{\phi}_d^* = \frac{\widehat{ECM}(\hat{t}_{d,S})}{\left(\hat{t}_{d,S} - \hat{t}_d\right)^2}. \tag{7.5}$$

Los estimadores compuestos son utilizados en los casos en donde el tamaño de muestra en el dominio es reducido, por lo tanto los ponderadores ϕ_d, deben ser elegidos de tal manera, que a medida que el tamaño de muestra, n_{s_d}, en el dominio, U_d, crezca, la participación del estimador basado en el diseño aumente. De esta manera, el sesgo del estimador compuesto tiende a cero, cuando n_{s_d} aumenta. Cuando el tamaño de muestra en el dominio es reducido, es necesario asignarle una mayor ponderación al estimador sintético (debido a que $\hat{t}_d$ puede ser muy inestable) y a medida que el tamaño de muestra aumente, dicha ponderación debe ser gradualmente reducida hasta llegar al punto en donde se puede utilizar únicamente un estimador basado en el diseño.

Otra alternativa para los ponderadores del estimador compuesto, es utilizar, ponderadores iguales $\phi_d = \phi$, para todos los dominios de interés. Purcell y Kish (1979) proponen minimizar el *ECM* agregado, o sea, $\sum_{d=1}^{D} ECM(\hat{t}_{d,C})$, respecto a ϕ. Lo anterior asegura obtener buenas estimaciones para el agregado pero no necesariamente para cada uno de los dominios en particular.

En el caso de que los ponderadores sea iguales por dominio, se tiene que

$$\sum_{d=1}^{D} ECM(\hat{t}_{d,C}) \doteq \phi^2 \sum_{d=1}^{D} ECM(\hat{t}_d) + (1-\phi)^2 \sum_{d=1}^{D} ECM(\hat{t}_{d,S}). \tag{7.6}$$

Minimizando (7.6) respecto a ϕ, se obtiene

$$\phi^* \doteq \frac{\sum_{d=1}^{D} ECM(\hat{t}_{d,S})}{\sum_{d=1}^{D} \left[ECM(\hat{t}_d) + ECM(\hat{t}_{d,S})\right]}, \tag{7.7}$$

en donde la participación del estimador basado en el diseño para todos los dominios esta sujeta a la suma de los *ECM* del estimador sintético. Si en total los *ECM* de los estimadores sintéticos son pequeños, en comparación a los *ECM* de los estimadores basados en el diseño, el ponderador ϕ sera pequeño para todos los dominios, aumentando así la participación de los estimadores sintéticos, en el estimador compuesto en todos los dominios.

Teniendo en cuenta la aproximación para estimar el *ECM* de la ecuación (6.17), el ponderador ϕ^*, puede ser estimado como

$$\hat{\phi}^* = \frac{\sum_{d=1}^{D} \left[(\hat{t}_{d,S} - \hat{t}_d)^2 - \hat{V}(\hat{t}_d)\right]}{\sum_{d=1}^{D} (\hat{t}_{d,S} - \hat{t}_d)^2} = 1 - \frac{\sum_{d=1}^{D} \hat{V}(\hat{t}_d)}{\sum_{d=1}^{D} (\hat{t}_{d,S} - \hat{t}_d)^2}. \tag{7.8}$$

El estimador $\hat{\phi}^*$, es mas estable que el estimador ϕ_d^*, de la ecuación (7.5), debido a que se están utilizando todos los dominios para estimar el ponderador. De todas formas, el uso de ponderadores comunes para todos los estimadores compuestos puede no ser efectivo, si las varianzas de de los estimadores basados en el diseño en cada uno de los dominios difieren considerablemente entre si.

Existen otros criterios para elegir los ponderadores del estimador compuesto. Por ejemplo, Pfefferman (2002), sugiere utilizar como ponderadores las tasas de muestreo efectivas en el dominio, o sea, $\phi_d = f_d$, donde $f_d = n_d/N_d$. Bajo estos ponderadores, la participación del estimador basado en el diseño ($\hat{t}_d$) en el estimador compuesto, aumenta a medida que la tasa de muestreo en el dominio crece. De esta forma, dado que los estimadores compuestos son utilizados en los casos en donde un estimador basado en el diseño puede ser muy inestable, utilizando estos ponderadores, generalmente se le asigna más participación al estimador sintético. Lo anterior se debe a que la tasa de muestreo en el dominio es muy pequeña o despreciable, derivando, prácticamente, en el uso únicamente del estimador sintético.

7.2. Ejemplos estimadores compuestos

7.2.1. Estimadores dependientes del tamaño de muestra

Los estimadores dependientes del tamaño de muestra (*sample size dependent*), son estimadores compuestos con ponderadores ϕ_d, que dependen únicamente de los tamaños del dominio N_d, y de

su estimación $\hat{N}_d$, o de los totales de una variable auxiliar x del dominio, t_{dx}, y de su estimación $\hat{t}_{dx\pi}$, en donde la variable auxiliar x se supone que se encuentra correlacionada con la variable de interés y.

Este tipo de estimadores fueron planteados para controlar el efecto del tamaño de muestra aleatorio en un dominio, en donde el tamaño esperado de muestra es lo suficientemente grande para utilizar estimadores basados en el diseño.

Drew, Singh y Choudhry (1982) propusieron un estimador para aquellos casos en donde el tamaño de muestra efectivo en el dominio no supere el tamaño de muestra esperado. El estimador queda definido como

$$\hat{t}_{d,SSD} = \phi_{d,s}\hat{t}_d + (1 - \phi_{d,s})\hat{t}_{d,S}, \tag{7.9}$$

con

$$\phi_{d,s} = \begin{cases} 1 & \text{si} \quad \hat{N}_d/N_d \geq \alpha \\ \hat{N}_d/(\alpha N_d) & \text{si} \quad \hat{N}_d/N_d < \alpha \end{cases}; \tag{7.10}$$

donde $\hat{N}_d = \sum_{s_d} a_k$ es el estimador π del tamaño del dominio N_d y α es una constante subjetivamente elegida para poder controlar la contribución del estimador sintético.

De forma general, se puede utilizar el estimador indirecto de regresión, $\hat{t}_d = \hat{t}_{d,gregP}$, y el estimador sintético de regresión, $\hat{t}_{d,S} = \hat{t}_{d,gregS}$, con $\alpha = 1$.

Otra forma de obtener los ponderadores ϕ_d es sustituir en (7.10) $\hat{N}_d/N_d$ por $\hat{t}_{dx\pi}/t_{dx}$, en donde x es una variable correlacionada con la variable de interés y.

7.2.2. Estimador de regresión amortiguado

El estimador de regresión amortiguado (*dampened regression estimator*), se obtiene al modificar el estimador indirecto de regresión $\hat{t}_{d,gregP}$ de la ecuación (3.33).

El objetivo es lograr "amortiguar" el efecto de la suma ponderada de los residuos $\sum_{s_d} a_k e_k$, la cual en algunos casos en donde el tamaño de muestra en el dominio, n_{s_d}, es muy pequeño (por ejemplo, cinco o menos), puede ser muy inestable. En algunas circunstancias, tanto el estimador indirecto de regresión $\hat{t}_{d,gregP}$, como el estimador $\tilde{t}_{d,gregP}$, pueden derivar en estimaciones fuera del rango de la variable de interés, ambos estimadores puede dar como resultado estimaciones negativas si algunos residuos $e_k = y_k - \mathbf{x'}_k\hat{\mathbf{B}}$ son extremadamente negativos.

De manera de poder controlar el término $\sum_{s_d} a_k e_k$ y reducir el riesgo de estimaciones inaceptables, Hidiroglou y Särndal (1989) sugieren la aplicación de un factor de amortiguación, para los casos en donde, la estimación del tamaño del dominio es menor que el verdadero valor (el cual es conocido), o sea, $\hat{N}_d < N_d$.

El resultado de esta corrección da lugar al estimador de regresión amortiguado

$$\hat{t}_{d,DRE} = \sum_{U_d} \hat{y}_k + (\hat{N}_d/N_d)^{H-1} \sum_{s_d} a_k e_k, \tag{7.11}$$

donde $\sum_{U_d} \hat{y}_k = \sum_{U_d} \mathbf{x'}_k \hat{\mathbf{B}}$, con $H = 0$ si $\hat{N}_d \geq N_d$ y $H = h$ si $\hat{N}_d < N_d$ y donde h es una constante positiva convenientemente elegida.

El estimador amortiguado de regresión, $\hat{t}_{dDRE}$, puede expresarse como un estimador compuesto, utilizando el estimador Hayek indirecto de regresión $\tilde{t}_{d,greg}$ de la ecuación (3.42) y el estimador sintético de regresión de la ecuación (6.1), con los siguientes ponderadores

$$\phi_{d,s} = \begin{cases} 1 & \text{si} \quad \hat{N}_d/N_d \geq 1 \\ (\hat{N}_d/N_d)^h & \text{si} \quad \hat{N}_d/N_d < 1 \end{cases}. \tag{7.12}$$

Observación 7.2.1 Si se elije $\alpha = 1$ en la ecuación (7.10) y $h = 2$ en (7.12), el estimador $\hat{t}_{d,DRE}$ es idéntico al estimador $\hat{t}_{d,SSD}$. □

7.3. Estimadores compuestos en el contexto de los estimadores de regresión

Si se cuenta con información específica del dominio, es posible utilizar tanto el estimador indirecto de regresión $\hat{t}_{d,gregP}$ como el estimador sintético de regresión, $\hat{t}_{d,gregS}$. Ambos estimadores, utilizan el mismo modelo, $E_m(y_k) = \mathbf{x'}_k\boldsymbol{\beta}$, $V_m(y_k) = c_k \ \ \forall k \in U$. El rol del modelo depende del estimador utilizado. Con el modelo estimado se calculan las predicciones para la variable y para todos los individuos del dominio y las mismas son utilizadas para ambos estimadores. El estimador indirecto de regresión es aproximadamente insesgado, pero en los casos donde el tamaño de muestra en el dominio es pequeño, el estimador puede ser inestable. Por otro lado, el estimador sintético de regresión, posee una varianza pequeña en relación al estimador $\hat{t}_{d,gregP}$, pero el mismo es generalmente sesgado, a menos que el modelo utilizado para su construcción sea verdadero.

Una manera para reducir la variabilidad del estimador indirecto de regresión y el sesgo del estimador sintético de regresión, es utilizar un ponderador de manera que el estimador compuesto, sea

$$\hat{t}_{d,gregComp} = \phi_d \hat{t}_{d,gregP} + (1 - \phi_d)\hat{t}_{d,gregS}. \tag{7.13}$$

En este caso, el estimador compuesto de regresión puede expresarse como

$$\begin{aligned} \hat{t}_{d,gregComp} &= \sum\nolimits_{U_d} y_k + \phi_d \sum\nolimits_{s_d} a_k e_k \quad (7.14) \\ &= \phi_d \hat{t}_{d\pi} + \left(\mathbf{t}_{dx} - \phi_d \hat{\mathbf{t}}_{dx\pi}\right) \hat{\mathbf{B}}. \quad (7.15) \end{aligned}$$

Observación 7.3.1 El estimador compuesto de la ecuación (7.13) puede expresarse como un estimador homogéneo

$$\begin{aligned} \hat{t}_{d,gregComp} &= \phi_d \hat{t}_{d\pi} + \left(\mathbf{t}_{dx} - \phi_d \hat{\mathbf{t}}_{dx\pi}\right)' \hat{\mathbf{T}}^{-1} \sum\nolimits_s a_k \mathbf{x}_k y_k / c_k \\ &= \sum\nolimits_s \left[\phi_d \delta_{dk} + \left(\mathbf{t}_{dx} - \phi_d \hat{\mathbf{t}}_{dx\pi}\right)' \hat{\mathbf{T}}^{-1} \mathbf{x}_k / c_k\right] a_k y_k \\ &= \sum\nolimits_s a_k g_{dk\phi} y_k = \sum\nolimits_s w_{k\phi} y_k. \end{aligned}$$

donde $w_{k\phi} = a_k g_{dk\phi}$ y $g_{dk\phi} = \phi_d \delta_{dk} + \left(\mathbf{t}_{dx} - \phi_d \hat{\mathbf{t}}_{dx\pi}\right)' \hat{\mathbf{T}}^{-1} \mathbf{x}_k / c_k$. □

El sesgo del estimador compuesto, $\hat{t}_{d,gregComp}$ viene dado por

$$B(\hat{t}_{d,gregComp}) = E(\hat{t}_{d,gregComp}) - t_d.$$

Reescribiendo el total del dominio, t_d como $t_d = \sum_{U_d} \mathbf{x'}_k \mathbf{B} + \sum_{U_d} E_k$, el sesgo del estimador es

$$\begin{aligned} B(\hat{t}_{d,gregComp}) &= E\left(\sum\nolimits_{U_d} \mathbf{x'}_k \hat{\mathbf{B}}\right) + E\left(\phi_d \sum\nolimits_{s_d} a_k e_k\right) - \sum\nolimits_{U_d} \mathbf{x'}_k \mathbf{B} - \sum\nolimits_{U_d} E_k \\ &= E\left(\phi_d \sum\nolimits_{s_d} a_k e_k\right) - \sum\nolimits_{U_d} E_k. \quad (7.16) \end{aligned}$$

Si, por simplicidad, se supone que el ponderador ϕ_d es constante, se obtiene

$$B(\hat{t}_{d,gregComp}) \doteq (\phi_d - 1) \sum\nolimits_{U_d} E_k. \quad (7.17)$$

Si el modelo utilizado para la construcción del estimador tiene un buen poder de ajuste en el domino y a su vez el ponderador ϕ_d es cercano a uno, el estimador compuesto tendrá un sesgo reducido.

Si el tamaño de muestra en el dominio, n_{s_d}, es lo suficientemente grande, el ponderador ϕ_d debe ser cercano a uno, debido a que la varianza del estimador de regresión sería lo suficientemente

pequeña para utilizar el estimador $\hat{t}_{d,gregP}$.

En tanto, cuando el tamaño de muestra en el domino es reducido, el término $\sum_{s_d} a_k e_k$ puede ser muy volátil y su varianza muy grande, por lo que se debería reducir la participación del mismo en el estimador $\hat{t}_{d,gregComp}$. Lo anterior se debe a que la varianza del estimador sintético tiende a ser muy pequeña.

Entonces, al elegir el ponderador ϕ_d se está eligiendo una forma de compromiso entre sesgo y varianza. El objetivo es encontrar un balance de los dos términos que integran el *ECM* del estimador $\hat{t}_{d,gregComp}$. El precio que se paga por reducir la ponderación del término $\sum_{s_d} a_k e_k$, produce que el estimador sea sesgado, debido a que el estimador sintético lo es, a menos que el modelo seleccionado sea verdadero. El *ECM* del estimador compuesto será más pequeño que el *ECM* del estimador de regresión si el modelo seleccionado no es muy malo, pero de no ser así el sesgo puede dominar el *ECM* y los intervalos de confianza construidos pueden ser inválidos.

Capítulo 8

Aplicación

8.1. Introducción

En este capítulo se presenta la aplicación de algunos de los métodos descritos anteriormente. La muestra utilizada es la Encuesta Continua de Hogares (ECH) del año 2009 que realiza el Instituto Nacional de Estadística (INE) y la información auxiliar utilizada son las proyecciones de población realizadas por el INE. De esta manera, todos los insumos utilizados están disponibles en la página web del INE[1].

El objetivo es evaluar la precisión que presentan algunos de los diferentes métodos de estimación para las principales variables que releva la ECH, en dominios definidos por agrupaciones geográficas y para distintos períodos de tiempo (mes, trimestre y año).

La ECH tiene como objetivo entre otros, proporcionar estimaciones para las tasas de actividad, empleo y desempleo. Estos indicadores son presentados en forma mensual, trimestral y anual. La desagregación geográfica utilizada para estos indicadores se encuentra sujeta al período de tiempo. Por ejemplo, de forma mensual se publican las tasas de actividad, empleo y desempleo para el total país y para dos dominios, los cuales corresponden a Montevideo y al resto del país. Para el trimestre y para el año, como consecuencia del aumento del tamaño de muestra efectivo, se presentan estos indicadores para todos los departamentos del país (19 dominios). A partir del año 2010, el INE anexa en sus publicaciones los intervalos de confianza para dichas estimaciones.

Los tamaños de muestra por departamento (teniendo en cuenta el período de tiempo) pueden no ser suficientes para obtener estimaciones con un nivel de precisión aceptable. Para algunos de estos dominios, por ejemplo Montevideo, el tamaño de muestra es suficiente para realizar estimaciones con un nivel de precisión aceptable (mensual, trimestral y anual), en tanto, para otros departamentos, el tamaño de muestra efectivo es reducido y los estimadores π no permiten obtener niveles de precisión aceptables, ni siquiera a nivel anual.

[1] http://www.ine.gub.uy/microdatos/microdatosnew2008.asp

En la aplicación se utilizan estimadores calibrados y de regresión.

8.2. Diseño Muestral de la ECH

El diseño muestral de la ECH para el año 2009 es estratificado, con dos o tres etapas de selección dependiendo del tipo de estrato. Los estratos son 58 y se definen en términos geográficos. El departamento de Montevideo se encuentra dividido en cuatro estratos socioeconómicos : bajo, medio bajo, medio alto y alto, los cuales son definidos en base al ingreso per cápita de los hogares a nivel de segmento censal. El anillo periférico (periferia) es un estrato y el mismo incluye parte de los departamentos de Canelones y San José en un radio de aproximadamente 30 kilómetros desde el centro de la ciudad de Montevideo. Para los 18 departamentos restantes, se definen en cada uno de ellos tres estratos: localidades de más de 5.000 habitantes, localidades de menos de 5.000 y zonas rurales.

El diseño es en dos etapas, a excepción de aquellos estratos conformados por localidades de menos de 5.000 habitantes, en donde se realizan tres etapas de selección. Para el resto de los estratos, la unidad primaria de muestreo (*PSUs*) es la zona censal (manzanas o territorio identificable), seleccionadas con probabilidad proporcional al tamaño medido en número de viviendas particulares. Las unidades secundarias de muestreo (*SSUs*) son las viviendas particulares dentro de cada *PSUs*, las viviendas son seleccionadas bajo un diseño aleatorio simple. Se seleccionan 3 viviendas en cada *PSU* seleccionada. En los estratos donde se realizan tres etapas de selección (localidades de menos de 5.000 habitantes), las *PSUs* son las localidad, las *SSUs* son las zonas y las *TSUs* son las viviendas particulares ocupadas.

Para el cálculo de los estimadores y sus varianzas se utilizó una aproximación del diseño muestral de la ECH debido a que no se conocen las probabilidades de inclusión de las diferentes etapas del muestreo. Por lo tanto, el diseño muestral aproximado, corresponde a un diseño estratificado por conglomerados en una etapa, en donde los estratos son idénticos a los del diseño muestral de la ECH y las *PSUs* son los hogares y los ponderadores muestrales son los que provee la encuesta (pesomen, pesotri y pesoano según el período de referencia).

8.3. Parámetros y dominios de interés

8.3.1. Parámetros de interés

Los parámetros de interés en esta aplicación son todos razones entre dos totales poblacionales desconocidos:

- **Tasa de actividad:** Se calcula como la razón entre la población económicamente activa

(PEA) y la población total en edad de trabajar (14 o más años de edad)

$$\frac{\sum_U I_k\{2\leq \text{pobpcoac} \leq 5\}}{\sum_U I_k\{\text{pobpcoac} \geq 2\}}$$

donde pobpcoac es una variable categórica con etiquetas

$$\text{pobpcoac}_k = \begin{cases} 1 & \text{si } k \text{ es menor de 14 años} \\ 2 & \text{si } k \text{ es ocupado} \\ 3 & \text{si } k \text{ es desocupado que busca trabajo por 1era vez} \\ 4 & \text{si } k \text{ es desocupado propiamente dicho} \\ 5 & \text{si } k \text{ es desocupado en seguro de paro} \\ 6 & \text{si } k \text{ es inactivo (realiza quehaceres del hogar)} \\ 7 & \text{si } k \text{ es inactivo (estudiante)} \\ 8 & \text{si } k \text{ es inactivo (rentista)} \\ 9 & \text{si } k \text{ es inactivo (pensionista)} \\ 10 & \text{si } k \text{ es inactivo (jubilado)} \\ 11 & \text{si } k \text{ es inactivo (otro)} \end{cases}.$$

- **Tasa de empleo:** Se calcula como la razón entre la población ocupada y la población total en edad de trabajar

$$\frac{\sum_U I_k\{\text{pobpcoac}=2\}}{\sum_U I_k\{\text{pobpcoac} \geq 2\}}.$$

- **Tasa de desempleo:** Se calcula como la razón entre la población desempleada y la PEA

$$\frac{\sum_U I_k\{3\leq \text{pobpcoac} \leq 5\}}{\sum_U I_k\{2\leq \text{pobpcoac} \leq 5\}}.$$

8.3.2. Dominios de interés

Los dominios considerados corresponden mayoritariamente a particiones geográficas de la población, los mismos están conformados por los departamentos (a excepción de Montevideo), el anillo periférico y los cuatro estratos del diseño muestral de la ECH de Montevideo. El total de dominios para los cuales se quiere brindar estimaciones para las tasas de actividad, empleo y desempleo, es de 23. En muchos casos dichos dominios coinciden con los estratos del diseño muestral, o se encuentran conformados por varios estratos, lo cual implica que los mismos son planeados y su tamaño de muestra es controlado.

8.4. Variables auxiliares

La variables auxiliares utilizadas para el cálculo de los estimadores, corresponden a las proyecciones de población. Dichas proyecciones son realizas por el INE[2].

Las proyecciones de población se encuentran disponibles según la siguiente desagregación:

- Total del país y de las áreas urbanas y rurales desagregadas por sexo y edad, en edades simples o grupos quinquenales, para el periodo 1996-2025.
- Total de la población para cada uno de los departamentos del país, desagregada en cada uno de ellos, en urbana y rural por sexo y edad (edades simples o grupos quinquenales).

Para esta aplicación, se utilizaron las proyecciones de población a nivel total país y para cada uno de los departamentos, en donde los tramos etarios utilizados son iguales por sexo y para cada una de las diferentes aperturas. A su vez, dichas celdas no deben estar vacías o con tamaños de muestra pequeños (se exigió un tamaño mínimo de 10 para el trimestre) para los distintos niveles de desagregación y teniendo en cuenta distintos momentos del tiempo, lo que obliga a conformar un conjunto de celdas no muy numerosas.

La construcción de las celdas (tramo etario-sexo) independientemente de la desagregación utilizada (departamental o total país), son 16. En los cuadros 8.1 y 8.2 se presentan los totales poblacionales y los totales muestrales para todo el año y de forma trimestral para cada una de estas celdas para todo el país.

Cuadro 8.1: *Proyecciones de población y totales muestrales por trimestre según tramo etario para hombres.*

Tramo etario	Proyecciones de población	Trimestre			
		1er	2do	3er	4to
0 a 13	362795	3503	3511	3505	3404
14 a 19	164119	1561	1646	1660	1648
20 a 24	130254	1143	1104	1172	1074
25 a 34	236763	2065	2083	2016	2025
35 a 44	204461	1891	1945	1959	1954
45 a 54	191367	1970	1973	1923	1911
55 a 64	146614	1513	1538	1545	1576
65 o +	179336	1947	1948	1927	1983
Total	1615709	15593	15748	15707	15575

[2] http://www.ine.gub.uy/socio-demograficos

Cuadro 8.2: *Proyecciones de población y totales muestrales por trimestre según tramo etario para mujeres.*

Tramo etario	Proyecciones de población	Trimestre			
		1er	2do	3er	4to
0 a 13	347068	3276	3397	3359	3316
14 a 19	157285	1595	1610	1642	1502
20 a 24	126380	1147	1185	1188	1144
25 a 34	240395	2288	2283	2310	2279
35 a 44	213684	2192	2142	2256	2201
45 a 54	203770	2167	2204	2180	2191
55 a 64	166324	1798	1707	1770	1811
65 o +	274323	2987	2959	2923	2967
Total	1729229	17450	17487	17628	17411

Debido a la desagregación de la información auxiliar disponible, para algunos de los dominios fijados anteriormente, no se tiene información específica, en donde el máximo nivel de apertura es departamental (los estratos de Montevideo y el anillo periférico). A los efectos de completar las estimaciones para todo el país se ópto por utilizar las proyecciones de población a nivel del departamento de Montevideo para los cuatro estratos del mismo, en tanto, para el anillo periférico, se optó por utilizar las proyecciones de población de los departamentos de Canelones y San José (de forma agregada). Esta opción no es necesariamente la más apropiada.

8.5. Estimadores y sus varianzas

Como ya se dijo, todos los parámetros de interés corresponden a razones (tasa de actividad, empleo y desempleo).

La razón entre dos variables y, z, para el dominio U_d, se define como

$$R_d = \frac{t_{dy}}{t_{dz}} = \frac{\sum_U y_{dk}}{\sum_U z_{dk}} = \frac{\sum_{U_d} y_k}{\sum_{U_d} z_k}.$$

Para estimar dichas razones, se utilizan estimadores calibrados, de regresión y el estimador π.

Los estimadores elegidos son:

1. $\hat{R}_{d,calU}$, denota a un estimador calibrado en donde la información auxiliar utilizada se encuentra definida a nivel de toda la población, o sea, a nivel total país.

2. $\hat{R}_{d,calU_D}$, denota a un estimador calibrado en donde la información auxiliar es a nivel departamental (para el cual el dominio de interés se encuentra incluido).

3. $\hat{R}_{d,gregU_D}$, denota un estimador de regresión, en donde el parámetro del modelo $\mathbf{B}$ que asiste a dicho estimador se encuentra definido a nivel de toda la población y el término de ajuste es definido a nivel departamental (para el cual el dominio de interés se encuentra incluido).

4. $\hat{R}_{d,\pi}$, denota a un estimador π, con ponderadores, a_k, provienen de la base de la ECH.

Todos los estimadores anteriores (a excepción del estimador π), son casos particulares del estimador general, presentado en la sección 5.3.

A su vez todos los estimadores son directos, a excepción del estimador $\hat{R}_{d,gregU_D}$. Lo anterior genera que el estimador, $\hat{R}_{d,gregU_D}$, produzca 23 sistemas de ponderadores diferentes (cantidad de dominios a estimar). Debido a que los dominios conforman una partición de la población, los estimadores calibrados (independientemente del nivel de desagregación de la información auxiliar), producen un sistema único de ponderadores.

El estimador π de la razón R_d, para el dominio U_d, viene dado por

$$\hat{R}_{d,\pi} = \frac{\hat{t}_{dy,\pi}}{\hat{t}_{dx,\pi}} = \frac{\sum_s a_k y_{dk}}{\sum_s a_k z_{dk}} = \frac{\sum_{s_d} a_k y_k}{\sum_{s_d} a_k z_k}. \tag{8.1}$$

Utilizando el desarrollo de Taylor de primer order, la razón $\hat{R}_{d,\pi} = \hat{t}_{dy,\pi}/\hat{t}_{dz,\pi}$ es aproximada por

$$\hat{R}_{d,\pi} \doteq \hat{R}_{d,\pi 0} = R_d + \frac{1}{t_{dz}} \sum\nolimits_s a_k (y_{dk} - R_d z_{dk}),$$

donde $R_d = t_{dy}/t_{dz}$.

El estimador $\hat{R}_{d,\pi}$ es aproximadamente insesgado para R_d y su varianza aproximada es

$$AV(\hat{R}_{d,\pi}) = \frac{1}{t_{dz}^2} \sum\sum\nolimits_U \Delta_{kl} \frac{y_{dk} - R_d z_{dk}}{\pi_k} \frac{y_{dl} - R_d z_{dl}}{\pi_l}. \tag{8.2}$$

El estimador de la varianza viene dado por

$$\hat{V}(\hat{R}_{d,\pi}) = \frac{1}{\hat{t}_{dz,\pi}^2} \sum\sum\nolimits_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_{dk} - \hat{R}_{d,\pi} z_{dk}}{\pi_k} \frac{y_{dl} - \hat{R}_{d,\pi} z_{dl}}{\pi_l}. \tag{8.3}$$

Para el cálculo de la fórmula (8.3) es necesario conocer $\Delta_{kl} \ \forall k$ y $l \in s$, que no está disponible en la base de la ECH. Entonces se considera el siguiente estimador

$$\hat{V}_0(\hat{R}_{d,\pi}) = \left(\frac{1}{\hat{t}^2_{dz,\pi}}\right)\left(\frac{1}{n(n-1)}\right)\sum\nolimits_s (r_{dk}a_k n - \hat{t}_{dr,\pi})^2, \tag{8.4}$$

donde $r_{dk} = y_{dk} - \hat{R}_{d,\pi}z_{dk}$ y $\hat{t}_{dr,\pi} = \sum_s a_k r_{dk}$.

Al utilizar (8.4), se supone que la muestra se obtuvo mediante un diseño con remplazo (ver Särndal et al. (1992)).

En esta aplicación, en donde el diseño muestral aproximado de la ECH es estratificado en conglomerados, el estimador de la varianza de (8.4) , toma la forma

$$\hat{V}_0(\hat{R}_{d,\pi}) = \frac{1}{\hat{t}_{dz,\pi}}\sum_{h=1}^{H}\frac{1}{n_{I_h}(n_{I_h}-1)}\sum\nolimits_{s_{I_h}}(\hat{t}_{dr_i,\pi}n_{I_h} - \hat{t}_{dr_h,\pi})^2, \tag{8.5}$$

donde s_{I_h} es la muestra de hogares en el estrato h, n_{I_h} el tamaño de la misma, $\hat{t}_{dr_h\pi} = \sum_{s_h} a_k r_{dk}$, y $\hat{t}_{dr_i\pi} = \sum_{s_i} a_k r_{dk}$ es el estimador π de la variable extendida r_d en el $i-$ésimo hogar.

Por otro lado, el estimador de la razón R_d, para el dominio U_d, utilizando el estimador general de la ecuación (5.11), viene dado por

$$\hat{R}_{d,gral} = \frac{\hat{t}_{dy,gral}}{\hat{t}_{dz,gral}} = \frac{\sum_s w_{Ck}y_{dk}}{\sum_s w_{Ck}z_{dk}} = \frac{\sum_{s_d} w_{Ck}y_k}{\sum_{s_d} w_{Ck}z_k}, \tag{8.6}$$

donde los ponderadores w_{Ck}, provienen de la ecuación (5.22) y considerando el vector $\mathbf{z}$ igual al vector de información auxiliar $\mathbf{x}$.

La aproximación de la varianza del estimador $\hat{R}_{d,gral}$ es

$$AV(\hat{R}_{d,gral}) = \frac{1}{t^2_{dz}}\sum\sum\nolimits_U \Delta_{kl}\frac{E_{yCk} - R_d E_{zCk}}{\pi_k}\frac{E_{yCl} - R_d E_{zCl}}{\pi_l}, \tag{8.7}$$

donde $E_{yCk} = y_{dk} - \mathbf{x}'_{Ck}\mathbf{Q}_{yMlz}$, $E_{zCk} = z_{dk} - \mathbf{x}'_{Ck}\mathbf{Q}_{zMlz}$, en donde, $\mathbf{Q}_{yMlz}$ proviene de la ecuación (5.14) y $\mathbf{Q}_{zMlz}$ se obtiene de cambiar la variable extendida y_L, por z_L, en la ecuación (5.14).

La aproximación de la varianza de (8.7) se obtiene de reemplazar en (8.2) y_{dk} y z_{dk} por E_{yCk} y E_{zCk} respectivamente y $\hat{R}_{d,\pi}$ por $\hat{R}_{d,gral}$ (ver Särndal et al. (1992) o Lehtonen y Veijanen (2009)).

Un estimador de la varianza del estimador $\hat{R}_{d,gral}$ se obtiene como

$$\hat{V}(\hat{R}_{d,gral}) = \frac{1}{\hat{t}^2_{dz,gral}} \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} w_{Ck} \left(e_{yCk} - \hat{R}_{d,gral} e_{zCk}\right) w_{Cl} \left(e_{yCl} - \hat{R}_{d,gral} e_{zCl}\right), \tag{8.8}$$

donde $e_{yCk} = y_{dk} - \mathbf{x'}_{Ck}\hat{\mathbf{Q}}_{yMlz}$, $e_{zCk} = z_{dk} - \mathbf{x'}_{Ck}\hat{\mathbf{Q}}_{zMlz}$, en donde $\mathbf{Q}_{yMlz}$, proviene de la ecuación (5.12) y $\hat{\mathbf{Q}}_{zMlz}$ se obtiene de cambiar la variable y_L por z_L en la ecuación (5.12).

El estimador de la varianza utilizado viene dado por

$$\hat{V}_0(\hat{R}_{d,gral}) = \frac{1}{\hat{t}^2_{dz,gral}} \sum_{h=1}^{H} \frac{1}{n_{I_h}(n_{I_h}-1)} \sum\nolimits_{s_{I_h}} (\hat{t}_{dr_i,gral} n_{I_h} - \hat{t}_{dr_h,gral})^2, \tag{8.9}$$

donde $\hat{t}_{dr_h,gral} = \sum_{s_h} w_{Ck} r_{dk}$, y $\hat{t}_{dr_i,gral} = \sum_{s_i} w_{Ck} r_{dk}$.

Las estimaciones puntuales utilizando (8.6) y (8.1), y las estimaciones de las varianzas, utilizando las ecuaciones (8.9) y (8.5), se realizaron con un código propio implementado en el software *R*. Los cálculos obtenidos replican los que se obtienen utilizando la librería *Survey* del *R* con la especificación del diseño aproximado detallado anteriormente.

8.6. Resultados

La comparación entre los cuatro estimadores propuestos para brindar estimaciones en los 23 dominios definidos anteriormente para las tasas de actividad, empleo y desempleo, se realiza en términos de sus coeficientes de variación estimados.

En los cuadros 8.3, 8.4 y 8.5 se presentan las estimaciones puntuales y los coeficientes de variación para las tasas de actividad, empleo y desempleo anual.

En los cuadros 8.6, 8.7 y 8.8 se presentan los coeficientes de variación de los estimadores $\hat{R}_{d,\pi}$, $\hat{R}_{d,calU}$, $\hat{R}_{d,calU_D}$ y $\hat{R}_{d,gregU_D}$ para las tasas de actividad, empleo y desempleo para los cuatro trimestres.

Finalmente en el cuadro 8.9 se presentan el promedio mensual de los coeficientes de variación para los cuatro estimadores para las tasas de actividad, empleo y desempleo.

Cuadro 8.3: *Estimaciones puntuales y Coeficiente de Variación (%) de los estimadores $\hat{R}_{d,\pi}$, $\hat{R}_{d,calU}$, $\hat{R}_{d,calU_D}$ y $\hat{R}_{d,gregU_D}$ para la tasa de **actividad** anual, según dominio de interés.*

	$\hat{R}_{d,\pi}$		$\hat{R}_{d,calU}$		$\hat{R}_{d,calU_D}$		$\hat{R}_{d,gregU_D}$	
Dominio	Est	CV (%)	Est	CV (%)	Est	CV (%)	Est	CV (%)
Artigas	61,28	1,82	62,26	1,76	63,31	1,49	63,52	1,40
Canelones	61,48	1,41	62,75	1,34	62,75	1,19	64,93	0,77
Cerro Largo	55,96	1,92	56,96	1,86	59,01	1,53	59,58	1,49
Colonia	59,70	1,54	61,01	1,45	61,54	1,09	61,51	1,03
Durazno	61,15	1,92	62,32	1,84	63,41	1,45	63,45	1,32
Flores	61,86	2,03	62,94	1,97	62,40	1,74	62,42	1,61
Florida	60,00	2,69	61,22	2,59	61,89	1,70	62,22	1,73
Lavalleja	62,48	1,89	63,61	1,80	63,40	1,68	63,35	1,51
Maldonado	64,72	1,67	65,81	1,59	66,39	1,25	66,46	1,14
Paysandú	60,97	1,46	62,10	1,40	61,97	1,17	62,22	1,04
Río Negro	61,01	2,04	62,19	1,97	63,48	1,59	63,85	1,37
Rivera	61,70	1,52	62,75	1,47	62,93	1,27	63,17	1,20
Rocha	61,00	2,14	62,10	2,05	62,35	1,58	62,58	1,48
Salto	61,60	1,44	62,61	1,39	62,56	1,22	62,96	1,09
San José	62,94	1,59	64,18	1,51	64,31	1,13	64,70	0,97
Soriano	66,36	1,55	67,44	1,49	66,72	1,32	66,46	1,25
Tacuarembó	57,47	1,84	58,58	1,78	59,97	1,42	60,53	1,28
Treinta y tres	58,93	2,01	60,04	1,95	60,57	1,59	61,44	1,15
Mvdeo Bajo	65,67	0,80	66,79	0,75	66,33	0,74	68,00	0,49
Mvdeo Mbajo	65,36	0,72	66,66	0,66	66,10	0,63	67,15	0,44
Mvdeo Malto	65,10	0,73	66,47	0,65	65,85	0,61	66,77	0,43
Mvdeo Alto	63,13	0,87	64,25	0,81	63,69	0,77	65,73	0,49
Periferia	64,77	0,81	65,96	0,75	66,09	0,66	66,64	0,51
Promedio	-	**1.58**	-	**1.51**	-	**1.25**	-	**1.10**

Cuadro 8.4: *Estimaciones puntuales y Coeficiente de Variación (%) de los estimadores $\hat{R}_{d,\pi}$, $\hat{R}_{d,calU}$, $\hat{R}_{d,calU_D}$ y $\hat{R}_{d,gregU_D}$ para la tasa de* ***empleo*** *anual, según dominio de interés.*

	$\hat{R}_{d,\pi}$		$\hat{R}_{d,calU}$		$\hat{R}_{d,calU_D}$		$\hat{R}_{d,gregU_D}$	
Dominio	Est	CV (%)	Est	CV (%)	Est	CV (%)	Est	CV (%)
Artigas	56,39	1,93	57,22	1,88	58,14	1,64	58,34	1,53
Canelones	57,50	1,49	58,65	1,42	58,58	1,28	60,37	0,86
Cerro Largo	52,83	2,04	53,74	1,99	55,74	1,67	56,12	1,60
Colonia	57,19	1,57	58,40	1,50	58,87	1,16	58,60	1,09
Durazno	56,93	2,15	57,91	2,09	58,85	1,76	58,90	1,56
Flores	58,64	2,16	59,61	2,12	59,13	1,94	59,01	1,78
Florida	55,58	3,05	56,66	2,98	57,29	2,01	57,64	2,02
Lavalleja	57,61	2,01	58,61	1,93	58,46	1,84	58,47	1,65
Maldonado	60,63	1,78	61,58	1,71	62,00	1,37	62,02	1,25
Paysandú	56,10	1,60	57,08	1,55	56,91	1,35	57,12	1,20
Río Negro	55,38	2,19	56,38	2,13	57,55	1,80	58,15	1,53
Rivera	56,85	1,66	57,77	1,61	57,88	1,40	58,15	1,31
Rocha	55,78	2,38	56,72	2,29	57,07	1,84	57,36	1,74
Salto	57,00	1,54	57,89	1,49	57,71	1,33	58,08	1,18
San José	60,25	1,67	61,39	1,60	61,50	1,22	61,53	1,05
Soriano	61,37	1,64	62,35	1,59	61,76	1,44	61,40	1,33
Tacuarembó	53,69	1,93	54,67	1,87	55,90	1,50	56,35	1,35
Treinta y tres	54,01	2,21	54,93	2,16	55,46	1,81	56,53	1,29
Mvdeo Bajo	59,51	0,91	60,51	0,87	60,12	0,85	61,87	0,58
Mvdeo Mbajo	59,83	0,79	60,96	0,74	60,47	0,71	61,52	0,52
Mvdeo Malto	60,36	0,79	61,53	0,72	61,01	0,68	61,81	0,50
Mvdeo Alto	59,69	0,91	60,66	0,86	60,16	0,82	61,71	0,53
Periferia	59,59	0,88	60,61	0,83	60,69	0,75	61,30	0,58
Promedio	-	**1.71**	-	**1.65**	-	**1.40**	-	**1.22**

Cuadro 8.5: *Estimaciones puntuales y Coeficiente de Variación (%) de los estimadores $\hat{R}_{d,\pi}$, $\hat{R}_{d,calU}$, $\hat{R}_{d,calU_D}$ y $\hat{R}_{d,gregU_D}$ para la tasa de* ***desempleo*** *anual, según dominio de interés.*

	$\hat{R}_{d,\pi}$		$\hat{R}_{d,calU}$		$\hat{R}_{d,calU_D}$		$\hat{R}_{d,gregU_D}$	
Dominio	Est	CV (%)	Est	CV (%)	Est	CV (%)	Est	CV (%)
Artigas	7,97	8,70	8,09	8,69	8,17	8,69	8,15	7,71
Canelones	6,48	8,52	6,54	8,50	6,64	8,41	7,02	5,46
Cerro Largo	5,61	12,14	5,65	12,15	5,54	12,16	5,81	10,37
Colonia	4,20	9,93	4,28	9,94	4,34	9,96	4,72	8,05
Durazno	6,91	11,18	7,06	11,10	7,20	11,12	7,18	9,00
Flores	5,20	14,22	5,29	14,44	5,24	14,30	5,46	12,50
Florida	7,37	13,95	7,45	13,88	7,44	13,22	7,37	12,58
Lavalleja	7,80	9,10	7,87	9,09	7,79	8,95	7,71	7,75
Maldonado	6,32	10,26	6,44	10,32	6,62	10,16	6,69	8,38
Paysandú	7,98	8,80	8,08	8,82	8,16	8,82	8,18	7,46
Río Negro	9,23	9,92	9,35	9,88	9,34	9,74	8,93	8,43
Rivera	7,85	8,11	7,94	8,11	8,02	8,06	7,94	7,08
Rocha	8,54	10,73	8,67	10,71	8,47	10,25	8,34	9,52
Salto	7,46	8,55	7,53	8,54	7,76	8,49	7,75	7,16
San José	4,27	9,74	4,35	9,77	4,38	9,70	4,90	7,41
Soriano	7,52	10,04	7,56	10,19	7,44	9,78	7,60	8,71
Tacuarembó	6,58	10,54	6,67	10,59	6,79	10,37	6,90	8,20
Treinta y tres	8,35	10,92	8,51	10,91	8,43	10,82	7,99	7,86
Mvdeo Bajo	9,38	4,26	9,41	4,25	9,37	4,24	9,01	3,31
Mvdeo Mbajo	8,47	3,95	8,56	3,94	8,51	3,93	8,38	3,25
Mvdeo Malto	7,28	4,29	7,43	4,26	7,35	4,24	7,43	3,49
Mvdeo Alto	5,44	5,90	5,59	5,86	5,54	5,85	6,10	4,10
Periferia	7,99	4,61	8,10	4,60	8,17	4,54	8,03	3,63
Promedio	-	**9.06**	-	**9.07**	-	**8.95**	-	**7.45**

Cuadro 8.6: *Coeficientes de Variación (%) de los estimadores $\hat{R}_{d,\pi}$, $\hat{R}_{d,calU}$, $\hat{R}_{d,calU_D}$ y $\hat{R}_{d,gregU_D}$ para la tasa de* ***actividad*** *por dominio para los cuatro trimestres*

Dominio	1er trimestre				2do trimestre				3er trimestre				4to trimestre			
	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$
Artigas	3,51	3,38	2,70	2,95	3,59	3,47	2,83	2,77	3,32	3,19	2,49	2,73	3,49	3,40	2,86	3,23
Canelones	2,67	2,49	2,16	1,84	2,85	2,70	2,40	2,10	2,55	2,42	2,16	1,62	2,46	2,30	2,02	1,71
Cerro Largo	3,63	3,46	2,95	3,09	3,77	3,68	3,16	3,29	3,29	3,18	2,81	3,18	3,47	3,33	2,58	2,73
Colonia	2,76	2,61	1,97	2,13	2,94	2,79	2,18	2,19	3,00	2,82	2,05	2,21	2,95	2,80	2,10	2,25
Durazno	3,28	3,10	2,20	2,41	3,86	3,71	2,88	2,90	3,75	3,60	2,85	2,69	3,52	3,39	2,65	2,56
Flores	3,80	3,70	3,04	3,06	4,39	4,27	3,45	3,87	3,95	3,80	3,91	4,60	4,72	4,48	3,58	3,53
Florida	6,11	5,76	3,77	4,30	5,51	5,22	2,93	3,75	4,59	4,39	3,45	4,00	3,16	3,05	2,45	2,81
Lavalleja	4,17	4,00	3,15	3,30	3,48	3,32	2,78	2,78	3,71	3,58	2,87	3,03	3,63	3,53	2,48	2,51
Maldonado	3,02	2,84	2,32	2,24	3,71	3,54	2,58	2,52	3,21	3,02	2,30	2,44	3,38	3,20	2,62	2,73
Paysandú	3,15	3,02	2,40	2,34	2,98	2,85	2,14	2,12	2,89	2,78	2,40	2,40	2,77	2,62	2,28	2,17
Río Negro	3,13	3,03	2,32	2,49	3,75	3,61	2,71	2,36	4,08	3,96	3,07	2,87	4,03	3,88	2,91	2,84
Rivera	2,62	2,55	2,18	2,31	2,77	2,66	2,11	2,08	3,09	2,98	2,59	2,51	3,22	3,09	2,44	2,75
Rocha	3,82	3,65	2,76	2,51	4,40	4,18	3,22	3,52	5,72	5,42	3,50	5,32	3,49	3,35	3,14	3,26
Salto	2,92	2,81	2,54	2,50	2,74	2,65	2,33	2,39	3,28	3,17	3,36	3,62	2,77	2,66	2,20	2,14
San José	2,94	2,80	2,19	2,08	3,14	2,99	2,19	2,02	2,66	2,56	1,90	2,23	3,09	2,89	2,26	2,20
Soriano	3,69	3,49	3,35	3,29	2,83	2,74	2,06	2,20	2,97	2,84	2,27	2,48	3,03	2,93	2,79	2,96
Tacuarembó	3,54	3,45	2,72	2,64	3,50	3,35	2,51	2,55	3,18	3,05	2,17	2,19	4,19	3,95	2,66	3,07
Treinta y tres	3,88	3,68	2,99	2,43	4,04	3,89	3,14	2,51	3,97	3,89	2,85	2,23	4,97	4,96	4,14	3,80
Mvdeo Bajo	1,58	1,48	1,46	1,55	1,59	1,51	1,48	1,44	1,53	1,45	1,43	1,46	1,55	1,45	1,43	1,38
Mvdeo Mbajo	1,36	1,24	1,20	1,18	1,50	1,37	1,31	1,17	1,36	1,25	1,21	1,10	1,42	1,29	1,24	1,16
Mvdeo Malto	1,35	1,20	1,13	1,07	1,38	1,24	1,15	1,01	1,37	1,24	1,17	1,06	1,47	1,30	1,20	1,04
Mvdeo Alto	1,68	1,56	1,48	1,44	1,64	1,51	1,43	1,23	1,56	1,43	1,37	1,20	1,56	1,44	1,36	1,18
Periferia	1,70	1,60	1,41	1,25	1,51	1,41	1,24	1,16	1,58	1,48	1,29	1,15	1,64	1,52	1,35	1,23
Promedio	**3,06**	**2,91**	**2,37**	**2,37**	**3,12**	**2,99**	**2,36**	**2,35**	**3,07**	**2,93**	**2,41**	**2,54**	**3,04**	**2,90**	**2,38**	**2,40**

79

Cuadro 8.7: *Coeficientes de Variación (%) de los estimadores $\hat{R}_{d,\pi}$, $\hat{R}_{d,calU}$, $\hat{R}_{d,calU_D}$ y $\hat{R}_{d,gregU_D}$ para la tasa de* ***empleo*** *por dominio para los cuatro trimestres*

Dominio	1er trimestre				2do trimestre				3er trimestre				4to trimestre			
	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$
Artigas	3,70	3,60	3,07	3,24	3,76	3,66	3,04	2,94	3,53	3,41	2,81	3,05	3,76	3,69	3,10	3,53
Canelones	2,90	2,74	2,45	2,08	3,11	2,97	2,68	2,45	2,72	2,59	2,34	1,78	2,60	2,45	2,20	1,89
Cerro Largo	3,87	3,71	3,18	3,36	4,02	3,93	3,40	3,53	3,58	3,48	3,10	3,46	3,59	3,47	2,75	2,82
Colonia	2,84	2,70	2,12	2,28	3,00	2,85	2,29	2,32	3,14	2,97	2,22	2,40	2,99	2,84	2,25	2,36
Durazno	3,91	3,79	2,91	3,14	4,15	4,03	3,29	3,25	4,06	3,93	3,10	2,87	3,86	3,74	3,10	2,90
Flores	4,10	4,01	3,39	3,36	4,66	4,59	4,07	4,40	4,12	3,97	3,97	4,71	4,29	4,14	3,62	3,58
Florida	6,40	6,07	4,39	4,70	6,34	6,07	3,44	4,74	4,77	4,58	3,65	4,18	3,59	3,51	2,80	3,24
Lavalleja	4,33	4,17	3,53	3,66	3,87	3,75	3,24	3,18	3,79	3,66	2,90	2,99	3,77	3,70	2,64	2,67
Maldonado	3,21	3,05	2,58	2,51	3,92	3,78	2,83	2,73	3,50	3,34	2,62	2,71	3,70	3,55	2,86	3,01
Paysandú	3,61	3,53	3,00	2,88	3,22	3,10	2,40	2,39	3,09	3,00	2,69	2,69	3,08	2,93	2,63	2,46
Río Negro	3,61	3,52	2,94	2,99	4,06	3,97	3,12	2,65	4,32	4,24	3,53	3,21	4,25	4,09	3,35	3,21
Rivera	2,95	2,88	2,53	2,64	3,10	2,99	2,42	2,38	3,40	3,30	2,93	2,80	3,33	3,20	2,63	2,90
Rocha	4,32	4,20	3,22	3,20	4,94	4,72	3,81	4,03	5,73	5,44	3,67	5,05	3,84	3,68	3,48	3,55
Salto	3,15	3,06	2,75	2,75	2,91	2,83	2,50	2,55	3,41	3,29	3,53	3,88	2,96	2,86	2,42	2,33
San José	3,04	2,90	2,29	2,20	3,34	3,23	2,47	2,23	2,83	2,73	2,08	2,48	3,19	2,99	2,36	2,33
Soriano	3,82	3,62	3,50	3,46	3,12	3,04	2,50	2,60	3,47	3,36	2,82	3,04	3,12	3,01	2,87	2,95
Tacuarembó	3,72	3,63	2,82	2,77	3,66	3,51	2,69	2,69	3,32	3,19	2,45	2,39	4,38	4,13	2,94	3,12
Treinta y tres	4,28	4,10	3,66	2,89	4,59	4,47	3,62	2,85	4,27	4,18	3,32	2,61	5,22	5,21	4,05	3,74
Mvdeo Bajo	1,78	1,69	1,66	1,83	1,79	1,72	1,70	1,70	1,78	1,71	1,68	1,74	1,76	1,67	1,64	1,63
Mvdeo Mbajo	1,50	1,40	1,36	1,37	1,70	1,59	1,53	1,43	1,49	1,39	1,34	1,26	1,56	1,44	1,39	1,34
Mvdeo Malto	1,47	1,34	1,27	1,23	1,53	1,41	1,33	1,22	1,48	1,36	1,29	1,20	1,59	1,45	1,35	1,22
Mvdeo Alto	1,78	1,66	1,59	1,58	1,76	1,65	1,58	1,40	1,65	1,54	1,47	1,30	1,62	1,50	1,42	1,25
Periferia	1,84	1,74	1,57	1,39	1,71	1,63	1,47	1,38	1,73	1,64	1,48	1,34	1,72	1,61	1,45	1,33
Promedio	**3,31**	**3,18**	**2,68**	**2,67**	**3,40**	**3,28**	**2,67**	**2,65**	**3,27**	**3,14**	**2,65**	**2,74**	**3,21**	**3,08**	**2,58**	**2,58**

Cuadro 8.8: *Coeficientes de Variación (%) de los estimadores $\hat{R}_{d,\pi}$, $\hat{R}_{d,calU}$, $\hat{R}_{d,calU_D}$ y $\hat{R}_{d,gregU_D}$ para la tasa de **desempleo** por dominio para los cuatro trimestres*

Dominio	1er trimestre				2do trimestre				3er trimestre				4to trimestre			
	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$
Artigas	17,16	17,13	17,14	16,51	20,64	20,68	20,82	18,71	14,42	14,38	14,13	14,68	16,95	16,95	17,28	17,23
Canelones	14,68	14,66	14,63	11,86	15,85	16,17	15,42	15,47	16,12	15,93	15,69	11,51	16,47	16,20	16,01	13,66
Cerro Largo	23,31	23,16	22,89	22,71	21,52	21,50	21,31	20,68	25,49	25,52	25,51	27,68	21,25	21,10	20,87	20,58
Colonia	18,63	18,61	18,53	18,54	22,27	22,20	22,13	20,42	19,35	19,33	19,13	19,34	19,01	19,23	19,51	18,71
Durazno	21,74	21,80	20,71	21,18	19,01	18,91	18,45	16,36	18,77	18,87	18,30	15,81	26,57	26,17	25,52	22,38
Flores	33,30	33,06	32,98	32,41	25,24	25,90	25,66	26,95	25,53	25,53	25,81	25,20	33,54	32,85	31,18	30,76
Florida	22,70	22,48	21,34	20,08	26,43	26,22	24,29	33,11	18,75	18,43	18,23	19,59	25,27	25,19	22,25	25,39
Lavalleja	16,31	16,36	16,46	14,99	14,95	15,00	14,56	13,60	26,10	25,85	25,03	28,19	17,93	17,78	17,84	18,43
Maldonado	26,01	26,11	25,61	19,97	20,71	20,63	19,75	18,15	19,34	19,35	19,11	18,88	18,25	18,82	17,64	17,24
Paysandú	18,01	18,12	18,09	16,22	17,75	17,70	17,37	15,99	15,54	15,55	15,70	15,30	17,61	17,57	17,23	16,00
Río Negro	15,15	15,10	15,05	15,39	22,07	22,01	21,39	16,73	19,46	19,47	19,34	17,23	21,67	21,85	21,77	20,63
Rivera	13,52	13,44	13,45	13,21	16,40	16,36	16,34	15,32	16,54	16,52	16,33	14,84	17,38	17,56	17,80	17,75
Rocha	21,64	21,75	16,97	19,46	23,72	23,44	21,92	24,88	24,29	23,57	22,68	23,29	20,24	20,24	20,67	21,50
Salto	15,62	15,57	15,28	15,03	15,16	15,13	15,30	14,68	18,46	18,48	18,71	17,14	18,93	18,81	18,64	16,46
San José	20,70	20,69	20,93	16,26	19,86	19,55	18,97	16,14	17,12	17,16	17,09	23,45	20,21	20,22	19,77	18,53
Soriano	27,27	27,61	26,28	23,74	17,61	18,18	17,15	17,32	16,97	17,02	16,89	16,61	18,15	18,02	17,56	17,96
Tacuarembó	18,84	18,95	18,71	16,38	19,42	19,32	18,75	17,52	17,15	17,13	16,91	15,51	24,48	24,60	24,87	20,73
Treinta y tres	21,18	21,16	19,98	17,33	21,03	21,02	20,78	16,46	21,71	21,68	21,73	17,43	24,77	24,80	23,17	22,06
Mvdeo Bajo	8,11	8,11	8,06	8,95	8,39	8,38	8,30	7,95	8,63	8,61	8,56	9,13	8,44	8,42	8,38	8,56
Mvdeo Mbajo	7,77	7,74	7,70	8,26	7,33	7,32	7,30	7,06	7,91	7,95	7,90	8,39	8,32	8,27	8,22	8,37
Mvdeo Malto	7,53	7,52	7,48	8,08	7,73	7,68	7,60	7,37	8,83	8,78	8,75	9,27	9,74	9,72	9,62	9,79
Mvdeo Alto	9,74	9,71	9,70	11,41	10,02	9,93	9,91	9,40	10,73	10,65	10,65	12,21	13,21	13,25	13,24	14,37
Periferia	9,30	9,23	9,08	8,03	8,33	8,30	8,15	7,99	9,22	9,21	9,23	8,30	10,14	10,13	9,98	8,99
Promedio	**17,75**	**17,74**	**17,26**	**16,35**	**17,45**	**17,46**	**17,03**	**16,45**	**17,24**	**17,17**	**17,02**	**16,91**	**18,63**	**18,60**	**18,22**	**17,66**

81

Cuadro 8.9: *Promedio mensual de los Coeficientes de variación (%) para los estimadores $\hat{R}_{d,\pi}$, $\hat{R}_{d,calU}$, $\hat{R}_{d,calU_D}$ y $\hat{R}_{d,gregU_D}$ por dominio de interés para las tasas de actividad, empleo y desempleo mensual.*

Dominio	Tasa de actividad				Tasa de Ocupación				Tasa de Desempleo			
	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$	$\hat{R}_{d,\pi}$	$\hat{R}_{d,calU}$	$\hat{R}_{d,calU_D}$	$\hat{R}_{d,gregU_D}$
Artigas	5,89	5,67	4,43	4,72	6,26	6,07	4,83	5,19	30,64	30,64	29,84	29,77
Canelones	4,44	4,16	3,65	3,18	4,77	4,51	4,04	3,58	26,65	26,53	25,96	22,64
Cerro Largo	6,12	5,89	4,85	5,17	6,52	6,31	5,24	5,55	41,14	40,92	40,03	41,11
Colonia	4,81	4,55	3,46	3,73	4,93	4,68	3,72	3,99	34,40	34,45	34,48	35,00
Durazno	6,16	5,90	4,47	4,51	6,81	6,58	5,23	5,18	37,85	37,64	34,99	34,06
Flores	6,93	6,70	5,63	5,86	7,33	7,15	6,21	6,41	53,51	53,61	50,46	54,02
Florida	6,52	6,22	4,56	5,16	6,94	6,66	5,03	5,67	36,59	36,35	33,02	36,37
Lavalleja	6,16	5,91	4,40	4,65	6,50	6,29	4,98	5,06	32,63	32,58	31,78	31,00
Maldonado	5,65	5,35	4,06	4,17	6,06	5,79	4,53	4,62	35,82	35,74	34,10	31,59
Paysandú	5,10	4,86	3,85	3,88	5,60	5,39	4,51	4,48	30,34	30,30	29,44	27,22
Río Negro	6,22	6,01	4,59	4,48	6,86	6,70	5,55	5,23	36,25	36,33	35,03	30,55
Rivera	4,88	4,67	3,83	3,94	5,34	5,14	4,32	4,44	28,67	28,64	28,06	27,50
Rocha	7,25	6,91	4,57	6,01	7,47	7,18	5,18	6,19	34,44	34,35	31,04	38,46
Salto	4,89	4,72	4,16	4,10	5,20	5,03	4,52	4,43	29,19	29,12	28,42	26,97
San José	5,14	4,87	3,68	3,99	5,38	5,13	3,98	4,28	33,88	33,98	33,37	35,41
Soriano	5,18	4,97	4,22	4,40	5,59	5,40	4,82	4,93	33,17	33,12	31,63	30,89
Tacuarembó	5,95	5,70	4,35	4,36	6,22	5,97	4,63	4,65	34,45	34,40	34,01	29,45
Treinta y tres	6,94	6,73	5,37	4,31	7,62	7,44	6,09	4,85	41,20	41,15	39,15	31,21
Mvdeo Bajo	2,70	2,54	2,50	2,53	3,08	2,94	2,88	3,01	14,50	14,47	14,38	15,03
Mvdeo M. Bajo	2,44	2,23	2,14	2,00	2,70	2,51	2,42	2,33	13,61	13,59	13,49	13,99
Mvdeo M. Alto	2,40	2,14	1,99	1,80	2,60	2,37	2,24	2,09	14,47	14,37	14,26	14,87
Mvdeo Alto	2,75	2,52	2,39	2,15	2,90	2,69	2,57	2,36	18,78	18,73	18,66	21,11
Periferia	2,79	2,60	2,28	2,11	3,04	2,86	2,57	2,39	16,16	16,09	15,81	14,72
Promedio	**5,10**	**4,86**	**3,89**	**3,97**	**5,47**	**5,25**	**4,35**	**4,39**	**30,80**	**30,74**	**29,63**	**29,26**

La eficiencia de los estimadores calibrados y de regresión, en comparación al estimador π, es más evidente a medida que el tamaño de muestra en el dominio disminuye.

Para las estimaciones anuales, los tamaños de muestra por dominio son suficientes para obtener una precisión aceptable utilizando el estimador $\hat{R}_{d,\pi}$ en donde el coeficiente de variación promedio en los dominios se sitúa en 1,58 % para la tasa de actividad, 1,71 % para la tasa de empleo y 9,06 % para la tasa de desempleo.

El desempeño del estimador $\hat{R}_{d,\pi}$ es similar al del estimador calibrado, $\hat{R}_{d,calU}$, el cual utiliza información auxiliar a nivel de toda la población.

Por otra parte, para los estimadores que utilizan información auxiliar a nivel departamental ($\hat{R}_{d,calU_D}$ y $\hat{R}_{d,gregU_D}$), los coeficientes de variación obtenidos son mas pequeños en todos los dominios para la tasa de actividad y de empleo, respecto a los estimadores $\hat{R}_{d,\pi}$ y $\hat{R}_{d,calU}$.

Para la tasas de actividad y de empleo los *CV* en promedio son 1,25 % y 1,40 % para el estimador calibrado $\hat{R}_{d,calU_D}$, y 1,10 % y 1,22 % para el estimador de regresión $\hat{R}_{d,gregU_D}$.

Este resultado es consecuencia de que ambos estimadores, $\hat{R}_{d,calU_D}$ y $\hat{R}_{d,gregU_D}$ estiman sin error al total de personas en edad de trabajar (14 o más) para todos los departamentos, denominador en ambas tasas. En tanto, el estimador calibrado, $\hat{R}_{d,calU}$, estima sin error únicamente el total de personas en edad de trabajar para el total del país.

Las estimaciones puntuales de los estimadores $\hat{R}_{d,calU}$, $\hat{R}_{d,calU_D}$ y $\hat{R}_{d,gregU_D}$ para las tasas de actividad y de empleo en los dominios considerados, son más altas que las estimaciones puntuales obtenidas utilizando el estimador $\hat{R}_{d,\pi}$. Esto puede deberse a que por ejemplo, las personas de más de 65 años se encuentran sobre-representadas en la muestra en la mayoría de los dominios. Al utilizar los estimadores $\hat{R}_{d,calU_D}$ y $\hat{R}_{d,gregU_D}$ se estiman sin error al total de personas en este tramo de edad para todos los departamentos, en tanto, utilizando $\hat{R}_{d,calU}$ se estiman sin error únicamente para todo el país.

Por otro lado, para la tasa de desempleo, el *CV* promedio del estimador de regresión, $\hat{R}_{d,gregU_D}$, es de 7,45 % y los *CV* obtenidos son menores en todos los dominios respecto a los otros estimadores propuestos. Para el resto de los estimadores los *CV* son del orden de 9,02 % en promedio.

Para las estimaciones trimestrales, si se observan los valores de los *CV* para las distintas tasas y los distintos estimadores, se puede llegar a la misma conclusión que para las estimaciones anuales.

En el caso de las estimaciones mensuales, todos los estimadores propuestos tienen menor *CV* que el estimador $\hat{R}_{d,\pi}$ (para las tasas de actividad y empleo), en todos los dominios de interés.

Al igual que para el caso de las estimaciones anuales, la ganancia en eficiencia del estimador calibrado $\hat{R}_{d,calU}$, es muy pequeña respecto al estimador $\hat{R}_{d,\pi}$, del orden de un 5 % menos para las tasas de actividad y empleo (en términos del *CV*).

Por otro lado, los *CV* de los estimadores que utilizan información auxiliar más desagregada $\hat{R}_{d,gregU_D}$ y $\hat{R}_{d,calU_D}$, son un 22 % menor que el estimador $\hat{R}_{d,\pi}$ para la tasa de actividad y un $19\,\%$ para la tasa de empleo.

Finalmente para la tasa de desempleo mensual, ninguno de los estimadores propuestos anteriormente permite obtener estimaciones con un nivel de precisión aceptable, y a su vez, la diferencias entre los *CV* de los cuatro estimadores, es mínima. Esto se debe a que la edad y sexo del individuo, explican poco la condición de desempleo, y a su vez, los individuos que presentan dicha característica, representan una proporción muy pequeña en la población. Además, los tamaños de muestra mensuales en cada uno de estos dominios son considerablemente pequeños, lo que hace que no sea posible obtener estimaciones para la tasa de desempleo para cada uno de los dominios propuestos, con niveles de precisión aceptable.

Capítulo 9

Conclusiones

El problema de estimación en dominios se encuentra presente en cualquier encuesta por muestreo. Cada vez más los usuarios exigen tener información más desagregada y no solo para la población en su conjunto. En la práctica es imposible satisfacer todos los requerimientos para disponer estimaciones con buenos niveles de precisión usando estimadores convencionales.

La ventaja del muestreo radica en obtener una información acertada observando una pequeña fracción de la población, de esta manera, el costo de una información razonablemente aproximada, observando un 1, 2 o 5 % de la población es 99, 50 o 20 veces más barato. Los requisitos de información precisa para dominios muy reducidos se contraponen a esta idea. Surge entonces la necesidad de seguir logrando informaciones acertadas a bajo costo, esto solo se puede lograr apelando a algún tipo de información auxiliar que potencie las ventajas del muestreo.

El uso de información auxiliar, tanto a la hora de definir el diseño muestral o en el proceso de estimación, es de vital de importancia en el problema de estimación en dominios. Utilizando un muestreo estratificado donde los estratos coinciden con los dominios (planeados), junto con una asignación eficiente de la muestra entre los estratos, por ejemplo utilizando *Power Allocation*, puede producir buenos resultados. El hecho de poder definir al dominio como un estrato, permite calcular tamaños de muestra específicos para cumplir determinados requisitos de precisión, a su vez, el tamaño de muestra puede ser controlado y fijo si el diseño lo permite.

Lo anterior, puede llegar a ser restrictivo en la práctica, debido a que es necesario conocer la variable indicadora de pertenecencia al dominio δ_d, para todos los individuos de la población. El incremento de la variabilidad de los estimadores, por no controlar el tamaño de muestra, toma importancia en aquellos dominios con un tamaño de muestra esperado pequeño. En tanto, si el tamaño de muestra esperado en el dominio es suficientemente grande, la pérdida de precisión por no haberlo controlado es despreciable.

Por otro lado, el uso de información auxiliar en el proceso de estimación es fundamental para obtener estimaciones más precisas y más aún en aquellos dominios en donde el tamaño de muestra

efectivo es reducido. Dicha información auxiliar puede provenir del marco muestral, de registros administrativos o de encuestas anteriores, y no necesariamente se debe conocer la información para todos los individuos del dominio, basta simplemente con conocer los totales de las variables auxiliares. Siempre y cuando sea posible, se debe utilizar información auxiliar específica del dominio, lo cual puede llegar a ser restrictivo, si lo anterior no se cumple, el investigador debe conformarse con información de subpoblaciones más amplias, lo cual implica que las estimaciones obtenidas puedan llegar a ser menos precisas.

Se presentaron una serie de estimadores bajo dos grandes enfoques, los basados en el diseño y los basados en modelos, en donde la aleatoriedad de los primeros proviene del diseño muestral y del modelo propuesto para los últimos.

En los estimadores basados en el diseño se presentaron dos clases de estimadores, calibrados y de regresión, los cuales utilizan información auxiliar en el proceso de estimación. Si la información es potente se obtendrán buenos resultados. La calibración solo hace referencia a la información auxiliar, a utilizar para calcular el nuevo sistema de ponderadores y no hace explícito ningún modelo. Los estimadores de regresión se apoyan en un modelo dado y la construcción de los mismos se basa en encontrar predicciones de la variable de interés para todos los individuos de la población (o del dominio). Ambas clases de estimadores son consistentes en el diseño y *very nearly design unbiased*.

Se hizo referencia en el nivel de desagregación de la información auxiliar para la construcción de los estimadores calibrados, la cual (siempre y cuando sea posible) debe ser específica del dominio. En la práctica, dicha situación no es común y el investigador debe conformarse con utilizar información a nivel de subpoblaciones más amplias (grupos de calibración). A su vez, se hizo referencia al estimador *uni-weight*, en donde un único sistema de ponderadores calibrados es utilizado para brindar estimaciones en todos los dominios de interés. Si bien este estimador es muy práctico, producto de la comodidad de trabajar con un único sistema de ponderadores, en algunos dominios puede ocurrir que las estimaciones obtenidas no posean un buen nivel de precisión.

En tanto, para los estimadores de regresión, se presentaron diferentes alternativas para definir el modelo que asiste al estimador de regresión. Si se supone que el dominio posee sus propias características y que estas difieren de la población en su conjunto, el modelo que asiste al estimador de regresión es específico del dominio y el estimador es directo. En cambio si el tamaño de muestra efectivo en el dominio es muy pequeño, produciendo que las estimaciones de los parámetros del modelo específico del dominio sean inestables, se puede utilizar un modelo definido en una subpoblación más amplia (por ejemplo toda la población), con el objetivo de poder realizar estimaciones más estables de los parámetros del modelo. De esta manera, se aumenta el tamaño de muestra efectivo utilizado para estimar los parámetros del modelo, derivando en que el estimador

sea indirecto. A su vez, se presentaron estimadores Hayek de regresión, los cuales son más eficientes, en comparación a los estimadores de regresión π en el problema de estimación en dominios. El requisito adicional para su construcción es conocer el tamaño del dominio N_d.

Los estimadores de regresión vistos en este documento, son asistidos por modelos lineales de efectos fijos. Otra estrategia puede ser utilizar modelos no lineales, los cuales pueden tener un mejor poder de ajuste, especialmente si la variable de interés es por ejemplo binaria, en donde el modelo que asiste al estimador de regresión es logístico. Al utilizar un modelo no lineal para asistir al estimador de regresión, es necesario que la información auxiliar se encuentre disponible para todos los individuos de la población, a diferencia de los modelos lineales, para los cuales, simplemente basta con conocer los totales de las variables auxiliares. A su vez, bajo este tipo de modelos, el estimador de regresión no es homogéneo. Otra alternativa es utilizar modelos mixtos para asistir al estimador de regresión, los cuales, a parte de los efectos fijos, introducen un efecto aleatorio (ver por ejemplo Lehtonen, Särndal y Veijanen (2003)).

Por otro lado, se hizo hincapié en la propiedad de aditividad (deseable en todo estimador en dominios). Los estimadores calibrados cumplen la propiedad de adivitidad dentro del grupo de calibración utilizado para su construcción. Si el grupo de calibración es toda la población (estimador *uni-weight*), los estimadores calibrados cumplen la propiedad de aditividad para cualquier subconjunto de la población. En tanto, los estimadores de regresión, cumplen la propiedad de aditividad, únicamente si el modelo que asiste al estimador se encuentra definido a nivel de toda la población, de todas formas, bajo esta elección, no es posible obtener un sistema único de ponderadores.

En el enfoque de los estimadores basados en el modelo, se presentaron estimadores sintéticos y compuestos, ambos (en mayor o menor medida) son dependientes del modelo propuesto, por lo tanto, si los supuestos del modelo no se cumplen los estimadores no tendrán buenas propiedades. Desde nuestro punto de vista su uso se justifica en los casos en donde el tamaño de muestra efectivo en el dominio es muy pequeño (o nulo) y los estimadores basados en el diseño pueden fallar (o ser imposibles de calcular). El precio por reducir la variabilidad de los estimadores conlleva a un aumento en el sesgo. Si el modelo utilizado para la construcción del estimador no es verdadero, el sesgo puede dominar la expresión del *ECM*, y los intervalos de confianza obtenidos pueden no tener los niveles de cobertura deseados.

Los estimadores sintéticos son utilizados en la practica debido a su fácil implementación y adaptación a cualquier diseño de muestreo, a su vez, no tienen como requisito un tamaño de muestra en el dominio determinado, por lo tanto pueden ser calculados inclusive si el tamaño de muestra en el dominio es nulo. El requisito para la construcción de los mismos, es disponer de información auxiliar específica del dominio, la cual a su vez, debe ser potente (aún más que para el caso de los estimadores basados en el diseño). Disponer de información auxiliar poderosa y a su vez específi-

ca del dominio puede ser muy restrictivo, ocasionando así en la práctica, que el modelo utilizado para su construcción no posea un buen poder de ajuste, derivando en que el estimador sintético sea sesgado.

Los estimadores sintéticos y compuestos presentados en este documento, forman parte de una gama extensa de estimadores basados en el modelo. Se destacan por ejemplo, los estimadores que utilizan modelos mixtos, los cuales han recibido mucha atención en los ultimo años, estos hacen hincapié en la variación entre dominios, incluyendo efectos fijos y aleatorios, y los mismos pueden ser utilizados a nivel de elemento o a nivel de totales (ver por ejemplo, Rao (2003) o Fuller (2009)).

En la aplicación realizada en este documento, se evaluaron distintos estimadores basados en el diseño para estimar las tasas de actividad, empleo y desempleo en 23 dominios de la población, los cuales están determinados mayoritariamente por regiones geográficas. La no inclusión de estimadores basados en modelos, se debe a la no disponibilidad de información auxiliar potente y específica del dominio (requisito fundamental para este tipo de estimadores). A su vez, desde nuestro punto de vista, en estadísticas oficiales (como por ejemplo las que realiza el INE) es necesario poder brindar estimaciones sin asumir ningún tipo de modelo, lo cual brinda transparencia a las estimaciones.

Los resultados obtenidos muestran que aquellos estimadores que utilizan información específica del dominio de interés proporcionan estimaciones más precisas, entre los que se encuentra el estimador indirecto de regresión y el estimador calibrado (utilizando información a nivel departamental). Las diferencias en precisión entre ambos estimadores es muy pequeña por lo que optamos por el estimador calibrado, esto se debe a que el mismo es directo y genera un único sistema de ponderadores (dado que los dominios particionan a la población), lo cual es atractivo y fácil de manejar. A su vez, este nuevo sistema de ponderadores calibrados brinda consistencia a las estimaciones, debido a que dicho sistema, estima sin error las proyecciones de población a nivel departamental. Por otro lado el estimador calibrado que utiliza información a nivel de las proyecciones total país, genera estimaciones igualmente precisas que el estimador π, el cual fue utilizado en la ECH para el año 2009.

De todas formas se esperaba que el aumento de precisión en las estimaciones utilizando estimadores calibrados y de regresión fuera mayor que el obtenido, el problema radica en la información auxiliar utilizada para la construcción de los mismos (la cual no es potente). Una estrategia puede ser utilizar información de la Encuesta Continua de Hogares de periodos anteriores, por ejemplo, colapsando varios años de forma de obtener tamaños de muestra más grandes en los dominios de interés (calculando un nuevo sistema de ponderadores) y posteriormente estimar totales poblacionales de variables auxiliares que se encuentren más asociadas con las tasas de actividad, empleo y desempleo, para distintas subpoblaciones, y las cuales a su vez, sean estables en tiempo, por

ejemplo nivel de educación por departamento.

Finalmente, investigaciones futuras pueden ser llevadas a cabo una vez finalizado el censo del año 2011, en donde se tendrá información actualizada, para la construcción del marco muestral de la ECH (el cual actualmente se encuentra desactualizado) y a su vez se dispondrá de información auxiliar más potente que las proyecciones de población para diferentes subpoblaciones (aún mas desagregado que el departamento) las cuales se encuentren más relacionadas con las variables de interés en esta aplicación. A su vez, debido a la cantidad de información auxiliar que se encontrará disponible, se podrán probar diferentes estimadores basados en el modelo, por ejemplo aquellos que utilizan modelos mixtos y en dominios que conformen particiones aún más finas que los departamentos como ser ciudades del interior o zonas rurales, en donde en la actualidad no se tiene estimaciones con niveles de precisión aceptables.

Bibliografía

Bankier, M.D. (1988). Power Allocation: determining sample sizes for subnational areas. *The American Statistician* **42**, 174-177.

Deville, J.C. y Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statical Association* **87**, 376-382.

Drew, D., Singh, M.P. y Choundhry, G.H. (1982). Evaluation os Small Area Estimation Techniques for the Canadian Labour Force Survey. *Survey Methodology* **8**, 17-47.

Ghosh. M. Rao. J.N.K (1994). Small Area Estimation: An Appraisal. *Statistical Science* **9**, 55-76.

Estevao, V.M. y Särndal. C.E.(1999). The use of auxiliary information in design-based estimation for domains. *Survey Methodology* **25**, 213-221.

Estevao, V.M. y Särndal. C.E. (2000). A functional form approach to calibration. *Journal of Official Statistics* **16**, 379-399.

Estevao, V.M. y Särndal. C.E. (2004). Borrowing Strength Is Not the Best Technique Within a Wide Class of Design-Consistent Domain Estimators. *Journal of Official Statistics* **20**, 645-669.

Gonzalez, M.E. (1973). Use and Evaluation of Synthetic Estimates. *Proceedings of the Social Statistics Section, American Statistical Association* 33-36.

Hidiroglou, M.A. y Patak, Z. (2004). Domain Estimation Using Linear Regression. *Survey Methodology*, **30**, 67-78.

Holt, D., Smith, T. y Tomberlin. T. (1979). A model-based approach to estimation for small subgroups of population. *American Statistical Association* **74**, 405-410.

Lehtonen, R., Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys (2nd ed.)*. JohnWiley & Sons, Chichester, UK.

Lehtonen, R., Särndal. C.E. y Veijanen. A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology* **29**, 33-44.

Lehtonen, R., Särndal, C.E. y Veijanen, A. (2005). Does the model matter? Comparing model-assisted and modeldependent estimators of class frequencies for domains. *Statistics in Transition* **7**, 649-673.

Lehtonen, R., Särndal C.E. y Veijanen, A. (2008). Generalized regression and model-calibration estimation for domains. *Invited paper, NORDSTAT 2008 Conference, Vilnius, June 2008*.

Lehtonen, R. y Veijanen. A. (2009). Design-based Methods of Estimation for Domains and Small Areas. *Sample Surveys: Inference and Analysis. Vol. 29B*, Elsevier B.V.

Lumley, T. (2004) Analysis of complex survey samples. Journal of Statistical Software 9(1): 1-19

Lumley, T. (2009) survey: analysis of complex survey samples. R package version 3.11-2.

Pfeffermann, D. (2002). New Important Developments in Small Area Estimation. *International Statistical Review*. **70**, 125-143.

Purcell, N.J. y Kish, L. (1979). Estimation for Small Domains. *Biometrics*, **35**, 365-384.

R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.Rproject.org.

Rao, J.N.K. (2003). Practical issues in model-based small area estimation. *Statistics Canada international symposium.* **11**.

Rao, J.N.K (2003). *Small Area Estimation*. JohnWiley & Sons, Hoboken, New Jersey.

Särndal, C.E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association* **79**, 624-631.

Särndal, C.E. (2007). The calibration approach in survey theory and practice. *Survey Methodology* **33**, 99-119.

Särndal, C.E. y Hidiroglou, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association* **84**, 266-275.

Särndal. C.E., Swensson. B. y Wretman. J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer - Verlag.

Schaible. W. L. (1996). *Indirect estimators in U.S. Federal programs*. New York: Springer - Verlag.

Singh, M.P., Gambino, J. y Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, **20**, 3-14.