Estimación de parámetros

Andrés Gutiérrez

Asesor Regional en Estadísticas Sociales - Comisión Económica para América Latina y el Caribe (CEPAL) - andres.gutierrez@cepal.org

Tabla de contenidos I

Estimación de parámetros

Introducción

- **Estimador:** Una función de la muestra aleatoria que toma valores en los reales y depende únicamente de los elementos de la muestra.
- **Diseño de muestreo:** Determinado por el conjunto Q de todas las posibles muestras.
- Las propiedades estadísticas de un estimador están determinadas por la medida de probabilidad discreta inducida por el diseño de muestreo.

Introducción

- La esperanza de un estimador $\hat{\theta}$ se calcula como

$$E(\hat{\theta}) = \sum_{s \subset Q} \theta(s) p(s)$$

donde $p(s)$ es la probabilidad de selección de cada muestra $s \in Q$.

- Las propiedades más comúnmente buscadas en un estimador $\hat{\theta}$ son el insesgamiento, la eficiencia y la consistencias.

Introducción

El sesgo está definido por la siguiente expresión:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

y el error cuadrático medio, dado por

$$ECM(\hat{\theta}) = E[\hat{\theta} - \theta]^2 = Var(\hat{\theta}) + B^2(\hat{\theta}).$$

Propiedades de un estimador.

- **Insésgado:** Un estimador con un sesgo nulo.
- **Error cuadrático medio:** Se convierte en la varianza del estimador cuando este es insesgado.
- **Eficiente:** Un estimador cuya varianza es pequeña en relación con otros estimadores.
- **Consistente:** Cuando el valor del estimador se acerca al parámetro desconocido a medida que el tamaño de muestra crece.

Observaciones

- Särndal, Swensson, y Wretman (2003) afirman que el objetivo en un estudio por muestreo es estimar uno a más parámetros poblacionales.
- Gutiérrez (2016) resalta que las decisiones más importantes a la hora de abordar un problema de estimación por muestreo son
  1. Escoger un diseño de muestreo y un algoritmo de selección.
  2. Elegir un estimador que calcule una estimación del parámetro de interés en la muestra seleccionada.

Observaciones

- Las decisiones no son independientes; la elección del estimador depende del diseño de muestreo utilizado.
- Note que, si $\hat{\theta}$ es un estimador del parámetro θ y $p_s(\cdot)$ un diseño de muestreo definido sobre un soporte Q, entonces la estrategia de muestreo será la dupla $(p(\cdot), \hat{T})$
- Es indispensable estimar subgrupos poblacionales en las encuestas, ya que esto permite obtener resultados confiables mediante una estrategia de muestreo adecuada.

Definición de **Dominios**

- Sean $U_1, \ldots, U_g, \ldots, U_G$ que denotan los subgrupos poblacionales tales que $\bigcup_{g=1}^{G} U_g = U$.
- Sea N_g el tamaño absoluto del subgrupo U_g, entonces se tiene que $\sum_{g=1}^{G} N_g = N$.
  - El diseño de muestra debe asegurar cobertura adecuada en cada dominio.
  - El número de individuos en la muestra pertenecientes a un dominio es aleatorio.
  - El tamaño absoluto de cada dominio no se conoce previamente.
  - Ejemplos: condición de ocupación, pobreza, rama de actividad.

Definición de **Estratos**

- Se conoce la pertenencia de todos los individuos al subgrupo poblacional.
- El diseño estratificado considera diferencias en la característica de interés.
- Se controla el tamaño de muestra antes de la estimación.
- Ejemplos: zonas urbanas/rurales, regiones, municipios.

Definición de **Postestratos**

- En el diseño se conoce el tamaño del postestrato, pero no el número de individuos en la muestra.
- Ejemplos: grupos etarios, sexo, etnia.
- Situaciones de postestratificación:
  a. No se utiliza la información en el diseño, pero sí en la estimación.
  b. Se conocen los tamaños absolutos de los subgrupos, pero no su pertenencia a estos.

El estimador de Horvitz-Thompson para totales y tamaños poblacionales

Estimación para totales

- Los indicadores sociales a nivel nacional suelen basarse en totales de variables de interés.
- El estimador de Horvitz-Thompson (HT) se utiliza para estimar estos totales.
- El estimador se define como:

$$\hat{t}_{y,\pi} = \sum_{s} d_k y_k$$

- s es la muestra seleccionada bajo un diseño de muestreo probabilístico.
- d_k son los pesos de muestreo que expanden el valor de la variable de interés y_k para el individuo k.

Estimación para totales

- $d_k = \pi_k^{-1}$, donde π_k es la probabilidad de inclusión del individuo k en la muestra.
- En esquemas complejos de muestreo, π_k es el producto de probabilidades condicionales en cada etapa del proceso de selección.
- El peso final de muestreo es típicamente una multiplicación de factores de expansión en cada etapa del esquema de muestreo.

Muestreo aleatorio simple

Las probabilidades de inclusión son equivalentes para cada unidad incluida en la muestra,

$$\pi_k = \frac{n}{N}$$

Por tanto el estimador toma la siguiente forma:

$$\hat{t}_{y,\pi} = \frac{N}{n} \sum_s y_k$$

Muestreo proporcional al tamaño

- Este diseño de muestreo asigna probabilidades de inclusión proporcionales al tamaño de una característica de información auxiliar disponible en el marco de muestreo.

$$\pi_k = \frac{n\ x_k}{t_x} \qquad 0 < \pi_k \leq 1$$

Por tanto el estimador toma la siguiente forma:

$$\hat{t}_{y,\pi} = t_x \sum_s \frac{y_k}{n\ x_k}$$

- No hay sesgo en la encuesta debido a la asignación de probabilidades desiguales en las unidades de muestreo.

Muestreo estratificado

- Un estimador insesgado para el total poblacional t_y se calcula como la suma de los totales estimados de cada estrato.

$$\hat{t}_{y,\pi} = \sum_{h=1}^{H} \hat{t}_{yh,\pi}$$

- Para un diseño de muestreo aleatorio estratificado, las probabilidades de inclusión de primer orden están dadas por:

$$\pi_k = \frac{n_h}{N_h}, \quad \text{si } k \in U_h$$

- Si s_h es la muestra seleccionada en el estrato U_h, entonces el estimador insesgado del total t_y está dado por:

$$\hat{t}_{y,\pi} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in s_h} y_k$$

Muestreo de conglomerados

- En el muestreo por conglomerados, la probabilidad de que un elemento sea incluido en la muestra es igual a la probabilidad de inclusión del conglomerado al que pertenece.

$$\pi_k = \pi_{Ii}, \quad \text{si } k \in U_i$$

- Si la población está dividida en N_I conglomerados y se selecciona una muestra de conglomerados s_I de tamaño n_I, entonces para un diseño de muestreo aleatorio de conglomerados, el estimador de HT del total poblacional está dado por:

$$\hat{t}_{y,\pi} = \frac{N_I}{n_I} \sum_{s_I} t_{yi}$$

- Es importante crear conglomerados acotados, como a nivel de manzana o vereda, para reducir la varianza del estimador.

Muestreo en dos etapas

- La probabilidad de inclusión de primer orden del k-ésimo elemento en este diseño está determinada por:

$$\pi_k = \pi_{k|i} \cdot \pi_{Ii}$$

- s_i es la submuestra de elementos seleccionada en el conglomerado U_i.
- En un diseño de muestreo aleatorio simple en las dos etapas, donde para cada unidad primaria de muestreo seleccionada $i \in s_I$ de tamaño N_i se selecciona una muestra s_i de elementos de tamaño n_i, el estimador de HT toma la siguiente forma:

$$\hat{t}_{y,\pi} = \frac{N_I}{n_I} \sum_{i \in S_I} \frac{N_i}{n_i} \sum_{k \in s_i} y_k$$

Muestreo en dos fases

- Este tipo de muestreo implica seleccionar una muestra de elementos s_a en una primera fase para recolectar información y crear una versión reducida del marco de muestreo.
- En una segunda fase, se realiza una nueva selección basada en esta información para definir una submuestra s, donde se observa la característica de interés.
- La probabilidad de que un elemento esté en la submuestra de la segunda fase s depende de lo ocurrido en la muestra de la primera fase s_a.
- La probabilidad de inclusión de cualquier elemento en la muestra final no tiene una forma cerrada y es algebraicamente intratable.

Muestreo en dos fases

- Se define el estimador de Horvitz-Thompson condicionado, que toma la siguiente forma:

$$\hat{t}_{y,\pi^*} = \sum_s \frac{y_k}{\pi_k^*} = \sum_s \frac{y_k}{\pi_{ak}\pi_{k|s_a}}$$

- Donde π_{ak} es la probabilidad de inclusión del elemento en la muestra de la primera fase, y $\pi_{k|s_a}$ es la probabilidad de inclusión del elemento en la submuestra de la segunda fase, condicionada a que haya sido incluido en la primera fase.

Estimador HT en una encuesta de hogares regular

- Supongamos un diseño regular en una encuesta de hogares, como un esquema estratificado de H estratos, con dos etapas de selección dentro de cada estrato.
- En la *primera etapa*, se seleccionan UPM dentro del estrato, y en la *segunda etapa*, se seleccionan hogares.
- El peso de muestreo final y el estimador del total se expresan como:

$$\hat{t}_{y,\pi} = \sum_{s} d_k y_k = \sum_{h} \sum_{i \in s_{Ih}} \sum_{k \in s_{hi}} w_{hik} y_{hik}$$

Estimador HT en una encuesta de hogares regular

Por ejemplo, si dentro de cada estrato U_h $h = 1, \ldots, H$ hay N_{Ih} unidades primarias de muestreo, de las cuales se selecciona una muestra s_{Ih} de n_{Ih} unidades mediante un diseño de muestreo aleatorio simple, y se considera un sub-muestreo aleatorio simple dentro de cada unidad primaria seleccionada:

$$\hat{t}_{y,\pi} = \sum_{h=1}^{H} \hat{t}_{yh,\pi} = \sum_{h=1}^{H} \left[\frac{N_{Ih}}{n_{Ih}} \sum_{i \in S_{Ih}} \frac{N_i}{n_i} \sum_{k \in s_i} y_k \right]$$

Estimación para tamaños y totales en dominios

- Las expresiones para totales también son aplicables para tamaños poblacionales, donde la variable $y_k = 1$ para todo $k \in s$.
- El estimador HT para un tamaño está dado por la suma de los factores de expansión:

$$\hat{N} = \sum_s d_k$$

- Bajo un diseño regular en una encuesta de hogares, con un esquema estratificado y dos etapas de selección, el estimador del tamaño poblacional está dado por:

$$\hat{N} = \sum_s d_k = \sum_h \sum_{i \in s_{Ih}} \sum_{k \in s_{hi}} w_{hik}$$

Estimación para tamaños y totales en dominios

- Si se asume un diseño de muestreo estratificado bietápico, con selección aleatoria simple en cada etapa, entonces la forma final del estimador HT para el tamaño poblacional es:

$$\hat{N}_{\pi} = \sum_{h=1}^{H} \left[\frac{N_{Ih}}{n_{Ih}} \sum_{i \in S_{Ih}} \frac{N_i}{n_i} \sum_{k \in s_i} 1 \right]$$

Estimación para tamaños y totales en dominios

- Los dominios deben cumplir con ciertas características:
  1. Ningún elemento de la población puede pertenecer a dos dominios.
  2. Todo elemento de la población debe pertenecer a un dominio.
  3. La unión de todos los dominios constituye la población del estudio.
- La estimación por dominios se caracteriza por el desconocimiento previo de la pertenencia de las unidades poblacionales al dominio.
- Se construye una función indicadora z_{dk} de la pertenencia del elemento al dominio, que toma el valor 1 si el elemento k pertenece al dominio U_d $(k \in U_d)$, y 0 en otro caso.

Estimación para tamaños y totales en dominios

- Utilizando los principios del estimador de Horvitz-Thompson, se puede obtener un estimador insesgado del tamaño del dominio U_d:

$$\hat{N}_d = \sum_{s_d} d_k$$

- Al multiplicar la variable de pertenencia z_{dk} por el valor de la característica de interés y_k, se crea una nueva variable y_{dk} dada por $y_{dk} = z_{dk} \cdot y_k$. Con esta nueva variable, se define el estimador insesgado del total de la característica de interés en el dominio U_d:

$$\hat{t}_{y_d,\pi} = \sum_{s} d_k \cdot y_{dk} = \sum_{S_d} d_k y_k$$

El estimador de Hájek para medias y proporciones

El estimador de Hájek para medias y proporciones

- El estimador de Hájek se utiliza para estimar medias y proporciones cuando no se conoce exactamente el tamaño poblacional.
- Para la estimación de la media, se utiliza el estimador de Hájek de la siguiente manera:

$$\hat{\bar{y}}_s = \frac{\hat{t}_y}{\hat{N}} = \frac{\sum_s d_k \; y_k}{\sum_s d_k}$$

- Para estimar una proporción P_d, el estimador de Hájek se define como:

$$\hat{P}_d = \frac{\hat{N}_d}{\hat{N}} = \frac{\sum_s d_k \; z_{dk}}{\sum_s d_k} = \frac{\sum_{s_d} d_k}{\sum_s d_k}$$

El estimador de Hájek para medias y proporciones

- Para la estimación de una media en una subpoblación, como por ejemplo la media del gasto en el área urbana, el estimador de Hájek se expresa como:

$$\hat{\bar{y}}_d = \frac{\hat{t}_{y_d}}{\hat{N}_d} = \frac{\sum_s d_k \; y_k \; z_{dk}}{\sum_s d_k \; z_{dk}} = \frac{\sum_{s_d} d_k \; y_k}{\sum_{s_d} d_k}$$

- Estos estimadores son no lineales y sus propiedades estadísticas son complejas, requiriendo la verificación de supuestos relacionados con el tamaño de la población y de la muestra.

El estimador de Hájek para medias y proporciones

- Bajo un diseño de muestreo que incluye un esquema estratificado y tres etapas de selección, el estimador de la media poblacional se define como:

$$\hat{\bar{y}}_s = \frac{\sum_h \sum_{i \in s_{Ih}} \sum_{j \in s_{hi}} \sum_{k \in s_{hij}} w_{hijk}\, y_{hijk}}{\sum_h \sum_{i \in s_{Ih}} \sum_{j \in s_{hi}} \sum_{k \in s_{hij}} w_{hijk}}$$

Otros estimadores de muestreo

Estimador de Razón

En presencia de información auxiliar, es posible mejorar la eficiencia de la estimación utilizando diferentes formas funcionales.

$$\hat{t}_{y,r} = t_x \frac{\sum_s d_k y_k}{\sum_s d_k x_k}$$

donde t_x denota el total poblacional de una variable auxiliar x, que se supone conocido para toda la población.

Estimación de la Razón Poblacional

- En el análisis de encuestas de hogares, es común realizar inferencias sobre parámetros que tienen una forma no lineal. Uno de los más básicos es la razón poblacional $R_U = \frac{t_{y_1}}{t_{y_2}}$, cuya estimación se lleva a cabo estimando ambos componentes de la fracción:

$$\hat{R} = \frac{\hat{t}_{y_1}}{\hat{t}_{y_2}} = \frac{\sum_s d_k y_{1k}}{\sum_s d_k y_{2k}}$$

- La estimación de un promedio poblacional $\bar{y}_U = t_y/N$ se puede ver como un caso particular de la estimación de una razón.

Estimación del Cambio de Indicadores

En las encuestas de hogares con diseños panel o rotativos, el interés suele estar en la estimación del cambio de indicadores en dos períodos de tiempo $\Delta = t_{y^{(t)}} - t_{y^{(t-1)}}$. Un estimador de este parámetro está dado por:

$$\hat{\Delta} = \hat{t}_{y^{(t)}} - \hat{t}_{y^{(t-1)}}$$

Estimadores Complejos

Si se desea estimar características asociadas con la pobreza, se pueden utilizar estimadores más complejos. Por ejemplo, para el ingreso y_k del individuo k y un umbral de pobreza l, el siguiente estimador puede ser utilizado:

$$\hat{F}_\alpha = \frac{1}{N} \sum_{k \in s} d_k \left(\frac{l - y_k}{l} \right)^\alpha I(y_k < l)$$

donde $I(y_k < l)$ es una variable indicadora que toma el valor uno si $y_k < l$ o cero en cualquier otro caso. Si $\alpha = 0$, se obtiene una estimación de la incidencia de la pobreza, y si $\alpha = 1$, se obtiene una estimación de la brecha de la pobreza (Foster, Greer, y Thorbecke 1984).

Observaciones

- La elección del estimador está estrechamente relacionada con el diseño de la encuesta.
- Si se busca estimar un indicador para un período específico, evita un esquema de rotación con traslape de hogares.
- El traslape de hogares puede aumentar la varianza del indicador y afectar la eficiencia de la estimación.
- Si el objetivo es estimar el cambio del indicador entre dos momentos en el tiempo, es necesario un esquema de rotación que garantice un tamaño de muestra adecuado.
- Según Cochran (1977) (sección 12.13), cuando te interesa tanto la estimación actual como el cambio entre periodos, se sugiere una tasa de traslape de 2/3, 3/4 o 4/5 de una ronda a otra.

Estimadores de calibración

Metodología

Gutiérrez (2016) da una breve descripción de este método:

1. Uso de un vector de información auxiliar conocido para la muestra y totales poblacionales: $\mathbf{x}_k = (x_{1k}, x_{2k}, ..., x_{pk})$ y $\mathbf{t}_\mathbf{X} = \sum_{k \in U} \mathbf{x}_k$.
2. Estimación del total de la característica de interés incorporando la información auxiliar: $\sum_{k \in s} w_k \mathbf{x}_k = \mathbf{t}_\mathbf{X}$.
3. Se requiere que los pesos resultantes cumplan con la siguiente restricción

$$\sum_{k \in s} w_k \mathbf{x}_k = \mathbf{t}_\mathbf{X}$$

   la cual es conocida como la ecuación de calibración.
4. El resultado de la calibración es un nuevo conjunto de pesos w_k que son muy cercanos al inverso de la probabilidad de inclusión del k-ésimo elemento $d_k = 1/\pi_k$

Ventajas y aplicaciones

- En América Latina, los INE usan la calibración sobre proyecciones poblacionales en dominios representativos de encuestas.
- Ventajas incluyen sesgo despreciable, errores estándar más pequeños y reproducción de información auxiliar.
- Puede aplicarse a características de hogares y personas simultáneamente, controlando desagregaciones y esquemas de muestreo.

Beneficios

- Garantiza consistencia estética con conteos censales y registros administrativos.
- Aumenta la precisión al buscar estrategias de muestreo para intervalos de variación más estrechos y menores errores de muestreo.
- Reduce el sesgo por ausencia de respuesta o falta de cobertura mediante integración de información auxiliar.

Ganancia en eficiencia

- Se realizaron cuatro conjuntos de datos con una relación específica entre la variable de interés y las variables auxiliares.
- Se empleó la metodología de calibración y se compararon, en mil iteraciones, las medidas de variabilidad.

Conjunto de datos 1: Relación lineal

- Se supuso una relación lineal entre la variable de interés y una variable auxiliar continua.
- Homoscedasticidad en el modelo y residuos coherentes.
- Ambos estimadores son insesgados, pero el estimador de calibración es menos disperso y más eficiente.

Conjunto de datos 1: Relación lineal

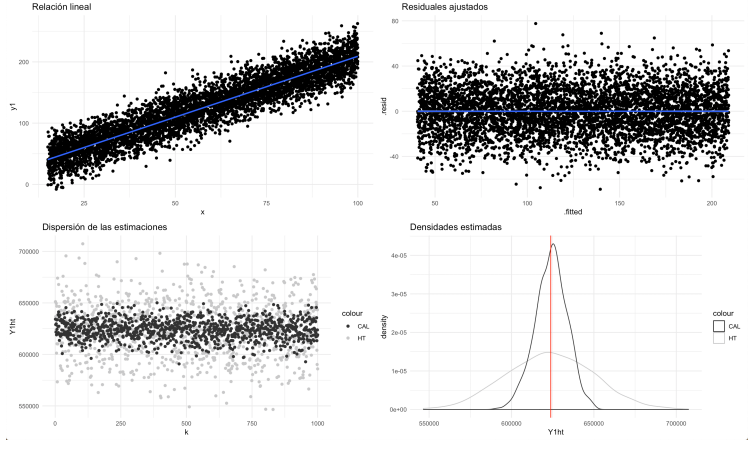

Figura 1: Comportamiento del estimador de calibración en una relación de dependencia lineal

Conjunto de datos 2: Relación lineal

- Se supuso una relación lineal entre la variable de interés y una variable auxiliar continua.
- Heteroscedasticidad en el modelo.
- Ambos estimadores se muestran insesgados para el parámetro de interés, el estimador de calibración es un poco más eficiente que el de Horvitz-Thompson.

Conjunto de datos 2: Relación lineal

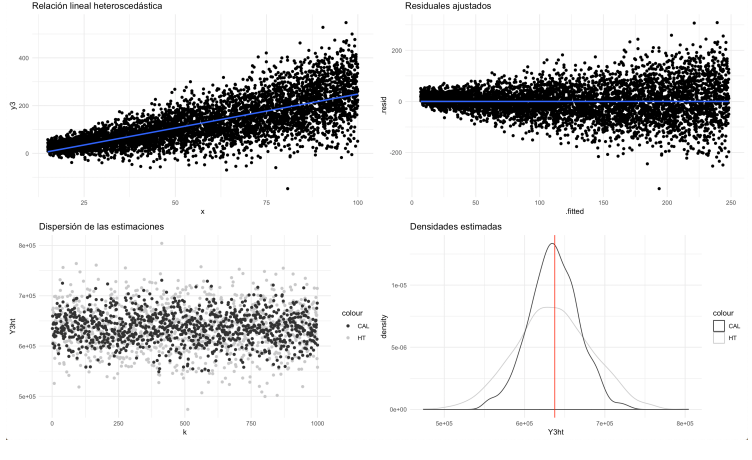

Figura 2: Comportamiento del estimador de calibración en una relación de dependencia lineal con heteroscedasticidad

Conjunto de datos 3: Relación cuadrática

- Se supuso una relación cuadrática entre la característica de interés y una variable de información auxiliar continua
- Residuales muestran un comportamiento inapropiado.
- Ambos estimadores se muestran insesgados para el parámetro de interés, pero el estimador de calibración es más eficiente que el de Horvitz-Thompson.

Conjunto de datos 3: Relación cuadrática

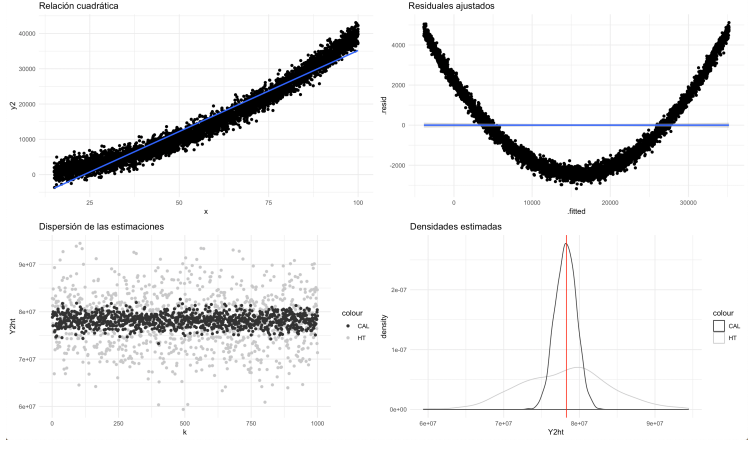

Figura 3: Comportamiento del estimador de calibración en una relación de dependencia cuadrática

Conjunto de datos 4: Relación logística

- Se supuso que existe una relación logística entre la característica de interés y una variable de información auxiliar dicotómica Al utilizar un estimador de calibración lineal.
- Residuales muestran un comportamiento inapropiado.
- Ambos estimadores se muestran insesgados para el parámetro de interés e igual de eficientes.

Conjunto de datos 4: Relación logística

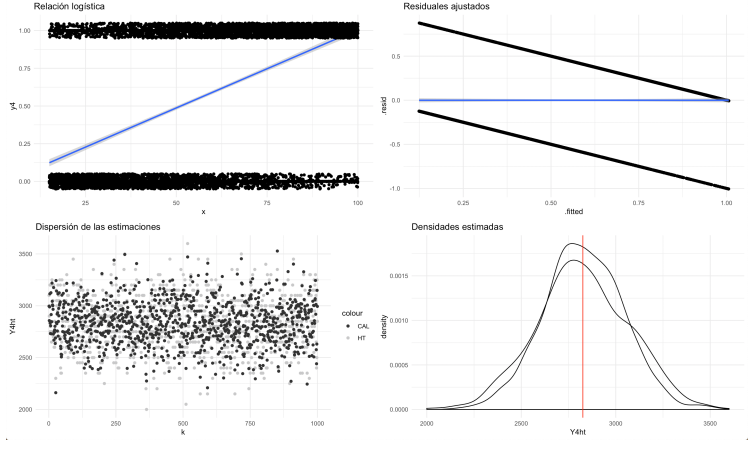

Figura 4: Comportamiento del estimador de calibración en una relación de dependencia logística

Observación

Se puede demostrar que:

$$\frac{Var(\hat{t}_{y,cal})}{Var(\hat{t}_{y,HT})} = (1 - R^2_{\xi} + o(\sqrt{n})) \approx (1 - R^2_{\xi})$$

Por ende, usar la metodología de calibración supone casi siempre una ganancia en la eficiencia de la estrategia de muestreo

Diferentes Formas del Estimador de Calibración

1. **Calibración con Variables Continuas**: En este caso, la calibración se aplica a totales de variables continuas como ingresos o gastos.
2. **Post-estratificación con Variables Categóricas**: Aquí, la calibración se basa en los tamaños poblacionales de subgrupos de interés.
3. **Raking con Variables Categóricas**: En este método, se calibra utilizando los tamaños marginales de tablas de contingencia de subgrupos. A diferencia del enfoque anterior, el raking no considera los tamaños de los cruces, sino solo los tamaños marginales.

Postestratificación

- La postestratificación es una técnica común para ajustar los pesos de muestreo mediante categorías poblacionales definidas posteriormente a la selección de la muestra.
- Requiere información auxiliar sobre todos los subgrupos definidos por las covariables de interés, como proyecciones demográficas.
- Los post-estratos se forman después de seleccionar la muestra y recolectar datos, lo que permite incluir variables difíciles de estratificar en el diseño de muestreo (raza, sexo, edad).

Postestratificación

Sea $g = 1, \dots, G$ el indicador del cruce poblacional (post-estrato), el estimador de postestratifiación queda definido de la siguiente manera:

$$\hat{t}_{y,pos} = \sum_{g=1}^{G} \frac{N_g}{\hat{N}_g} \hat{t}_{y_g}$$

En donde N_g corresponde al tamaño poblacional del post-estrato, $\hat{N}_g = \sum_{s_g} d_k$ y $\hat{t}_{y_g} = \sum_{s_g} d_k y_k$.

El factor de expansión de estimador postestratificado queda definido como sigue:

$$w_k = d_k \frac{N_g}{\hat{N}_g} \qquad (k \in s_g)$$

Postestratificación

- El factor de expansión del estimador postestratificado se calcula considerando los tamaños poblacionales de los post-estratos y los pesos de diseño ajustados.
- La cantidad de postestratos está determinada por la cantidad de interacciones en las variables auxiliares, pudiendo resultar en cientos de postestratos en algunos casos.
- Muchas variables e interacciones pueden disminuir la eficiencia de la calibración y causar estimaciones inestables, especialmente si hay celdas vacías o datos atípicos en los pesos calibrados resultantes.

Raking

- En el caso de que los conteos poblacionales no estén disponibles para todos los cruces de las variables de calibración, se puede utilizar el método de raking para ajustar los marginales de la tabla cruzada.
- El raking permite calibrar los marginales de la tabla cruzada sin necesidad de calibrar todas sus entradas, lo que reduce el número de restricciones en comparación con la postestratificación.
- Se utiliza un procedimiento iterativo (IPFP) para ajustar los marginales de la tabla cruzada, comenzando con las filas y luego las columnas hasta alcanzar la convergencia de los pesos calibrados.

Raking

- Los pesos calibrados se escriben de la siguiente manera:

$$w_k = d_k \times exp(u_h) \times exp(v_g)$$

En donde u_h es una función de los totales marginales de las filas de la tabla cruzada y v_g es una función de los totales marginales de las columnas.

- El raking permite incorporar variables predictoras de las variables de interés y mitigar los efectos adversos de las bajas tasas de cobertura del marco de muestreo en la inferencia.

La calibración como un cambio de paradigma en una teoría de estimación exhaustiva

1. *La calibración no se puede separar de la práctica.*
2. *La calibración no puede separarse de la consistencia estética*[1].
3. *La calibración debe ser de fácil interpretación.*
4. Representa un enfoque exhaustivo y unificado que aborda problemas comunes en encuestas de hogares, como la ausencia de respuesta, deficiencias del marco muestral y errores de medición, ofreciendo una teoría coherente para inferencia en poblaciones finitas.

[1] En este apartado la palabra consistente se da en el sentido de que el estimador reproduce exactamente los totales de la información auxiliar.

Estimadores compuestos

- Se propone mejorar la estimación del total actual $t_y^{(t)}$ al considerar la información del traslape de la encuesta en el segundo período.

$$\hat{t}^K_{y^{(t)}} = (1-K)\hat{t}_{y^{(t)}} + K(\hat{t}_{y^{(t-1)}} + \hat{\Delta}_M)$$

- Se utiliza un factor K para combinar la estimación actual con la estimación del período anterior ajustada por la diferencia en la muestra traslapada.
- Se define $\hat{\Delta}_M$ como la diferencia entre las estimaciones actual y previa en la muestra traslapada.

$$\hat{\Delta}_M = \hat{t}^M_{y^{(t)}} - \hat{t}^M_{y^{(t-1)}}$$

Estimadores compuestos

- Se introduce un término adicional A que refleja la diferencia entre las estimaciones actuales de las muestras con y sin traslape.
- El estimador compuesto $\hat{t}_{y^{(t)}}^{AK}$ se calcula sumando el estimador $\hat{t}_{y^{(t)}}^{K}$ con el producto de A y la diferencia entre las estimaciones de las muestras sin traslape y con traslape.

$$\hat{t}_{y^{(t)}}^{AK} = \hat{t}_{y^{(t)}}^{K} + A(\hat{t}_{y^{(t)}}^{U} - \hat{t}_{y^{(t)}}^{M})$$

Estimadores compuestos

- Los estimadores compuestos aprovechan la correlación entre estimaciones del mismo panel a lo largo del tiempo para mejorar la eficiencia estadística.
- Se pueden ajustar los valores de las constantes A y K para minimizar la varianza del estimador o aumentar su eficiencia en comparación con otros estimadores, como los de niveles o cambios temporales.
- Se pueden escribir como estimadores de regresión o calibración, incorporando covariables relacionadas con la información del periodo anterior para estimar indicadores de nivel o cambios.

Estimadores compuestos

Gambino, Kennedy, y Singh (2001) proponen usar la siguiente covariable de calibración para estimar indicadores de nivel:

$$x_k^{(l)} = \begin{cases} \hat{\bar{y}}^{(t-1)}, & k \in s_U \\ y_k^{(t-1)}, & k \in s_M \end{cases}$$

Estimadores compuestos

Gambino, Kennedy, y Singh (2001) proponen la siguiente covariable:

$$x_k^{(c)} = \begin{cases} y_k^{(t)}, & k \in s_U \\ y_k^{(t)} - R(y_k^{(t)} - y_k^{(t-1)}), & k \in s_M \end{cases}$$

En donde $R = \sum_s w_k / \sum_{s_M} w_k$ es el inverso de la proporción de traslape real en el levantamiento.

Estimadores compuestos

- Se propone una combinación lineal convexa de covariables para evitar problemas numéricos.

$$x_k = (1-\alpha)x_k^{(l)} + \alpha x_k^{(c)}$$

- Fuller y Rao (2001) mencionan que en algunas aplicaciones particulares, se ha encontrado mayor eficiencia (menor varianza) al utilizar valores de $\alpha = 0.65$ o $\alpha = 0.75$; aunque se recomienda usar $\alpha = 2/3$ en los sistemas de producción de las ONE.

- Un ejemplo de aplicación en el rediseño de la Encuesta Continua de Hogares en Uruguay, donde se implementó un sistema de panel rotativo con mejoras en la eficiencia estadística utilizando estimadores compuestos.

¡Gracias!

Email: andres.gutierrez@cepal.org

Referencias

Cochran, W. G. 1977. *Sampling Techniques*. Third Edition. Wiley.

Foster, James, Joel Greer, y Erik Thorbecke. 1984. «A Class of Decomposable Poverty Measures». *Econometrica* 52 (3): 761-66. https://doi.org/10.2307/1913475.

Fuller, Wayne A, y J N K Rao. 2001. «A Regression Composite Estimator with Application to the Canadian Labour Force Survey». *Survey Methodology*, n.º 12: 7.

Gambino, Jack, Brian Kennedy, y Mangala P Singh. 2001. «Regression Composite Estimation for the Canadian Labour Force Survey: Evaluation and Implementation». *Survey Methodology*, n.º 12: 10.

Gutiérrez, H. A. 2016. *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*. Segunda edición. Ediciones de la U.

Särndal, Carl-Erik, Bengt Swensson, y Jan Wretman. 2003. *Model Assisted Survey Sampling*. Springer Science; Business Media.