Estimación en Areas Pequeñas

Muestreo II

Licenciatura en Estadística

2023

Introducción

A nivel general, el **enfoque tradicional basado en el diseño** logra buenas precisiones a nivel poblacional y aceptables cuando el tamaño de muestra en los dominios es "suficientemente" grande para utilizar estimadores tradicionales.

Sin embargo, esto último no siempre sucede, lo que lleva a estimaciones volátiles en los dominios en donde el tamaño de muestra efectivo no es suficiente.

El marco de **Estimación en Áreas Pequeñas** surge como alternativa al abordaje clásico y busca revertir esta situación.

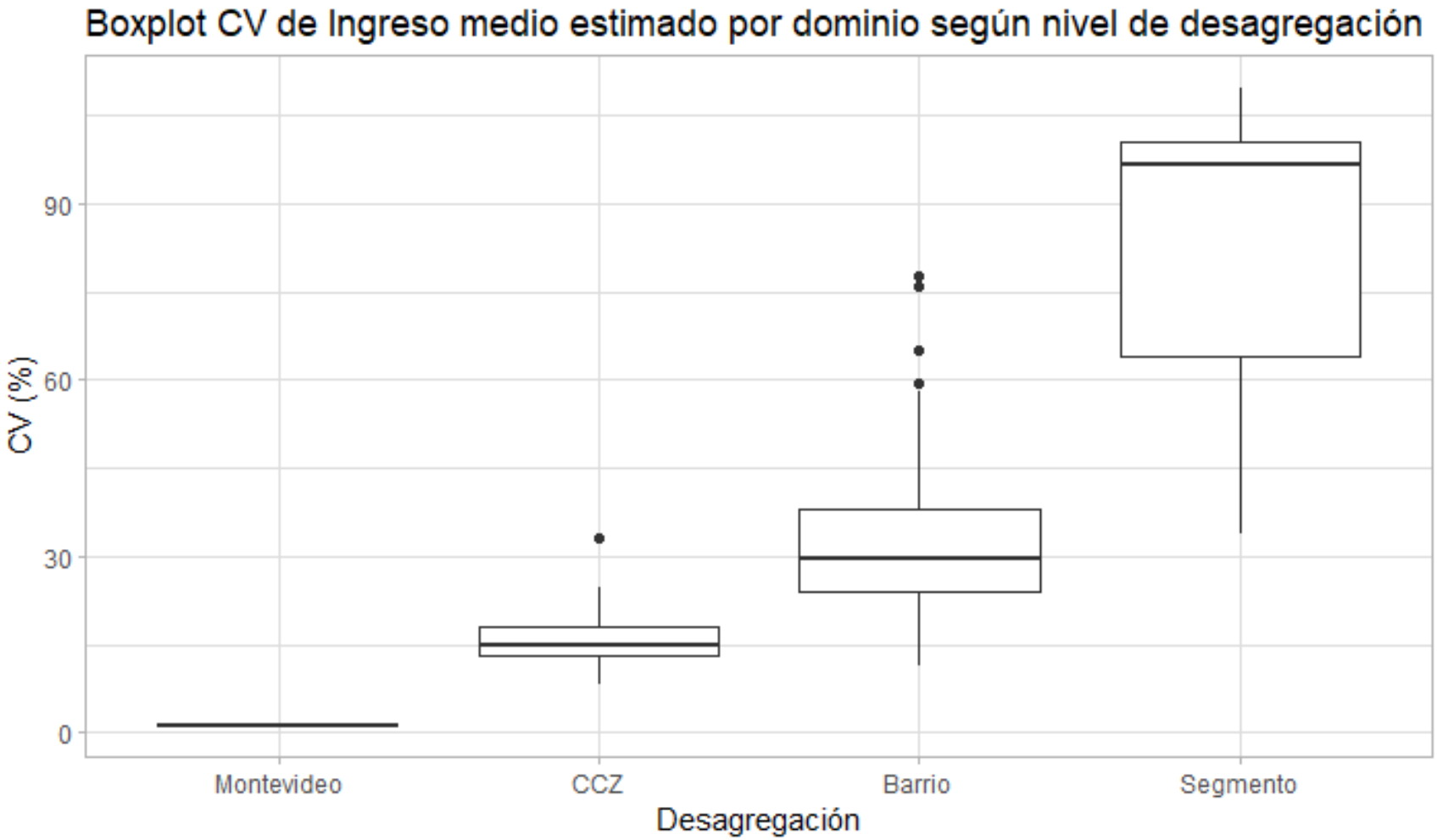

Figure 1: CV Ingresos medio estimados por nivel de desagregación

Área Pequeña

Área Pequeña

Un "área pequeña" es un dominio, es decir, un subconjunto específico de la población.

Estos dominios cumplen que el tamaños de muestra efectivo no es lo suficientemente grande como para obtener estimaciones basadas en el diseño de calidad.

- Esta subdivisión puede darse a partir de características más allá de lo geográfico, como pueden ser criterios sociodemográficos o combinaciones de ambos.

Inferencia basada en modelos

El marco de Small Area Estimation (**SAE**) se basa en la **inferencia basada en modelos** y en los **estimadores indirectos**.

- Estimación basada en modelos: las observaciones tienen una distribución de probabilidad asumida, reflejada en un modelo superpoblacional.
- Estimadores indirectos: Los estimadores indirectos son aquellos que no solo hacen uso de la información relevada asociada al área de la cual se desea obtener una estimación, sino que se apoya de la información asociada al resto de los dominios para así ganar eficiencia.

Modelos SAE

Los estimadores clásicos de SAE tienen un fuerte componente de Modelos Líneales Mixtos (**MLM**).

Los **MLM** son una extensión de los modelos líneales clásicos que incorporan tanto **efectos fijos** como **aleatorios**.

- Efectos fijos: un conjunto de covariables cuya relación con la variable de respuesta es homógenea a nivel poblacional.
- Efectos aleatorios: Son factores categóricos asociados a la unidad experimental. Estos parámetros son aleatorios.

Modelo líneal mixto

El mismo supone:

$$y = X\beta + Zu + \varepsilon$$

Donde $y \in \mathbb{R}^n$ es el vector de la variable de interés a nivel poblacional, $X \in \mathcal{M}_{n\times p}(\mathbb{R})$ y $Z \in \mathcal{M}_{n\times h}(\mathbb{R})$ son matrices (**fijas**) de **rango completo**.

A su vez:

- $u \sim \mathcal{N}_n(0, \mathbf{G})$.
- $\varepsilon \sim \mathcal{N}_n(0, \mathbf{R})$.
- u y ε son independientes.
- $\mathbf{G}$ y $\mathbf{R}$ dependen de $\delta = (\delta_1, ..., \delta_q)^t$.

Ejemplo: Modelo de interceptos aleatorios

Se busca modelar el comportamiento de una variable y a partir de una covariable x. Se conoce también que los individuos de la muestra se encuentran agrupados.

- Por ejemplo: resultados de un examen de un individuo a partir de las horas que dedicó a su preparación (la muestra proviene de distintos centros educativos).

$$y_{ij} = u_j + \beta_1 x_{ij} + \varepsilon_{ij}$$

Donde:

- y_{ij}: puntaje obtenido por el alumno i-ésimo del centro j.
- x_{ij}: horas dedicadas por el alumno i-ésimo del centro j.
- $u_j \sim \mathcal{N}(0, \sigma_u^2)$ es el efecto aleatorio asociado al centro j.
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2)$.

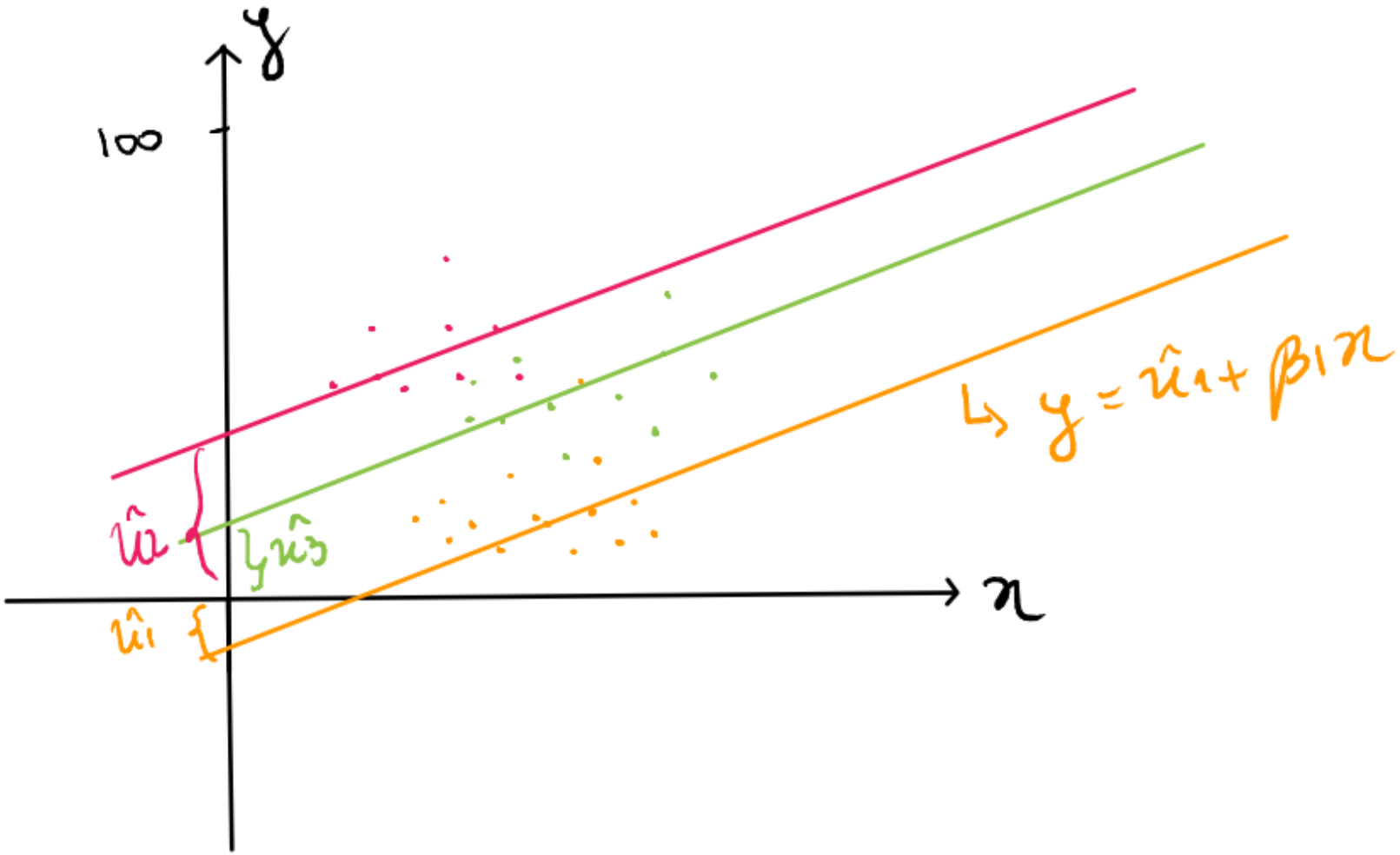

Figure 2: Boceto modelo interceptos aleatorios ajustado

BLUP: Best Linear Unbiased Predictor.

Suponiendo que la variable de interés se distribuye siguiendo el **modelo líneal mixto general** y que a su vez **se conoce a los parámetros de varianza** δ, se buscará estimar una combinación líneal del tipo $\mu = l'\beta + m'u$.

Se desea que el estimador sea:

- **Lineal:** $\hat{\mu} = a'y + b$.
- **Insesgado bajo el modelo:** $\mathbb{E}(\mu) = \mathbb{E}(\hat{\mu})$.
- **Minimice el error cuadrático medio:** $\mathbb{ECM}(\hat{\mu}) = \mathbb{E}([\hat{\mu} - \mu]^2) \leq \mathbb{E}([\hat{\mu}^{alt} - \mu]^2)$.

Resultado:

$$\tilde{\beta} = \tilde{\beta}(\delta) = \left(X'V^{-1}X\right)^{-1} X'V^{-1}y$$

$$\tilde{u} = \tilde{u}(\delta) = GZ'V^{-1}\left(y - X\tilde{\beta}\right)$$

$$\tilde{\mu}^{\text{BLUP}} = l'\tilde{\beta} + m'\tilde{u} = l'\tilde{\beta} + m'GZ'V^{-1}\left(y - X\tilde{\beta}\right)$$

Con:

$$V = \mathbb{VAR}(y) = ZGZ' + R$$

Esquema de demostración:

- Se plantean condiciones para lograr la **insesgadez**.
- Se minimiza el **ECM** aplicando el teorema de extremos condicionados de Lagrange.

Error cuadrático medio del BLUP

$$\mathbb{ECM}(\hat{\mu}^{\text{BLUP}}) = g_1(\delta) + g_2(\delta)$$

$$g_1(\delta) = m' \left(G - GZ'V^{-1}ZG \right) m$$

$$g_2(\delta) = d' \left(X'V^{-1}X \right)^{-1} d$$

Con $d = l' - b'X$ y $b' = m'GZ'V^{-1}$

Empirical Best Lineal Unbiased Predictor

En la práctica, los parámetros de varianza δ **no** son conocidos.

Por lo tanto, no se tiene una especificación completa de:
$V = V(\delta) = ZGZ' + R$.

Para "materializar" una estimación de μ se deberá trabajar en dos etapas. Primero estimar δ para finalmente evaluar a $\hat{\mu}^{\text{BLUP}}$ en δ.

$$\Rightarrow \hat{\mu}^{\text{EBLUP}} = \hat{\mu}^{\text{BLUP}}(\hat{\delta})$$

Los dos métodos principales para estimar δ son:

- Máxima verosimilitud.
- Máxima verosimilitud restringida.

Estimación de δ a partir de MV

Teniendo en cuenta que:

$$y \sim \mathcal{N}(X\beta, ZGZ' + R)$$

Se tiene que la función de log-verosimilitud del problema es:

$$l(\beta, \delta) = c - \frac{1}{2}\left(\ln(|V|) + (y - X\beta)'V^{-1}(y - X\beta)\right)$$

Finalmente se aplica algún método computacional para obtener:

- $\hat{\delta}^{\text{ML}}$
- $\hat{\beta}^{\text{ML}} = \tilde{\beta}(\hat{\delta}^{\text{ML}})$
- $\hat{u}^{\text{ML}} = \tilde{u}(\hat{\delta}^{\text{ML}})$
- $\hat{\mu}^{\text{EBLUP}}$

Estimación de δ a partir de MVRE

El método de verosimilitud obtiene estimaciones sesgadas de δ, una alternativa es aplicar el método de **máxima verosimilitud restringida**.

Para ello se aplica la transformación líneal $y^* = A'y$ donde A es cualquier matriz $n \times (n-p)$ **ortogonal** a X.

$$y^* \sim \mathcal{N}_{n-p}(0, A'VA)$$

Maximizando la verosimilitud de y^* se llega a δ^{MVRE}.

Error cuadrático medio del EBLUP

$$\mathbb{ECM}(\hat{\mu}^{\text{EBLUP}}) \approx g_1(\delta) + g_2(\delta) + g_3(\delta).$$

Donde:

$$g_3(\delta) = \text{tr}\left(\frac{\partial b'}{\partial \delta} V \left(\frac{\partial b'}{\partial \delta}\right)' \bar{V}(\hat{\delta})\right)$$

Donde $\bar{V}(\hat{\delta})$ es la matriz de covarianza asintótica de $\hat{\delta}$.

Un estimador insesgado del $\mathbb{ECM}$ (cuando δ se estima usando MVRE) es:

$$\widehat{\mathbb{ECM}}(\hat{\mu}^{\text{EBLUP}}) = g_1(\hat{\delta}) + g_2(\hat{\delta}) + 2g_3(\hat{\delta})$$

Poblaciones con estructura de covarianza Block-Diagonal

Un caso particular del modelo que hemos estudiando hasta ahora es aquel en donde las matrices y vectores estan particionados en **m** componentes.

$$y = ({y_1}', {y_2}', ..., {y_m}')'$$

$$X = \text{COL}_{1 \leq i \leq m}(X_i)$$

$$Z = \text{DIAG}_{1 \leq i \leq m}(Z_i)$$

$$v = \text{COL}_{1 \leq i \leq m}(v_i)$$

$$\varepsilon = \text{COL}_{1 \leq i \leq m}(\varepsilon_i)$$

$$R = \text{DIAG}_{1 \leq i \leq m}(R_i)$$

$$G = \text{DIAG}_{1 \leq i \leq m}(G_i)$$

Donde $X_i \in \mathcal{M}_{n_i \times p}$, $Z_i \in \mathcal{M}_{n_i \times h_i}$, $y_i \in \mathbb{R}^{n_i}$.

Bajo esta especificación y aplicando las fórmulas de las diapositivas anteriores:

$$\hat{\mu}_i^{\text{EBLUP}} = I_i'\hat{\beta} + m_i'\hat{u}_i$$

$$\hat{u}_i = G_i Z_i' V_i^{-1}(y_i - X_i\hat{\beta})$$

$$\hat{\beta} = \left(\sum_{i=1}^{m} X_i' V_i^{-1} X_i\right)^{-1} \sum_{i=1}^{m} X_i' V_i^{-1} y_i$$

ECM

Las fórmulas asociadas al $\mathbb{ECM}$ en cada dominio pueden expresarse como:

$$g_{1i}(\delta) = m_i'(G_i - G_i Z_i' V_i^{-1} Z_i G_i) m_i$$

$$g_{2i}(\delta) = d_i' \left(\sum_{i=1}^{D} X_i' V_i^1 X_i \right) d_i$$

$$g_{3i}(\delta) = \text{tr} \left(\frac{\partial b_i'}{\partial \delta} V_i \left(\frac{\partial b_i'}{\partial \delta} \right)' \bar{V}(\hat{\delta}) \right)$$