Estimación en áreas pequeñas

Métodos indirectos basados en modelos: EBLUP basado en el modelo BHF y el método ELL

Andrés Gutiérrez
Comisión Económica para América Latina y el Caribe

2019

Referencias

Referencias

- (2018) Molina, Isabel. *Estudio de los límites de desagregación de dates en encuestas de hogares para subgrupos de población y áreas geográficas y los requerimientos para superarlos: Fase II*. CEPAL.

Referencias

(2018) Molina, Isabel. *Estudio de los límites de desagregación de dates en encuestas de hogares para subgrupos de población y áreas geográficas y los requerimientos para superarlos: Fase II*. CEPAL.

Referencias

- (2018) Molina, Isabel. *Estudio de los límites de desagregación de dates en encuestas de hogares para subgrupos de población y áreas geográficas y los requerimientos para superarlos: Fase II*. CEPAL.
- (2015) Rao, J.N.K y Isabel Molina. *Small Area Estimation*. Second ed. Wiley Series in Survey Methodology.

Introducción

- Nuevamente, los estimadores para áreas basados en modelos se consideran modelos indirectos porque usan información de otras áreas.

Introducción

- Nuevamente, los estimadores para áreas basados en modelos se consideran modelos indirectos porque usan información de otras áreas.
- Los estimadores basados en modelos incorporan la heterogeneidad que no puede ser explicada por las variables auxiliares coleccionadas.

Introducción

- Nuevamente, los estimadores para áreas basados en modelos se consideran modelos indirectos porque usan información de otras áreas.
- Los estimadores basados en modelos incorporan la heterogeneidad que no puede ser explicada por las variables auxiliares coleccionadas.
- Esto se realiza incorporando efectos aleatorios de las áreas en los modelos de interés.

BLUP/EBLUP basado en el modelos con errores anidados (BHF)

BLUP/EBLUP basado en el modelo BHF

- El modelo con errores anidados fue propuesto por Battese, Harter, and Fuller (1988) para explicar el crecimiento de varios cultivos en Estados Unidos.

BLUP/EBLUP basado en el modelo BHF

- El modelo con errores anidados fue propuesto por Battese, Harter, and Fuller (1988) para explicar el crecimiento de varios cultivos en Estados Unidos.
- El modelo relaciona en una forma lineal una variable Y_{di} para el individuo i en el área d con p variables auxiliares.

BLUP/EBLUP basado en el modelo BHF

- El modelo con errores anidados fue propuesto por Battese, Harter, and Fuller (1988) para explicar el crecimiento de varios cultivos en Estados Unidos.
- El modelo relaciona en una forma lineal una variable Y_{di} para el individuo i en el área d con p variables auxiliares.
- Es diferente del modelo Fay Herriot pues el modelo FH relaciona los estimadores directos a variables auxiliares.

BLUP/EBLUP basado en el modelo BHF

- El Modelo viene dado por

$$Y_{di} = \mathbf{x}'_{di}\beta + u_d + e_{di}, \quad i = 1, \ldots, N_d,\ d = 1, \ldots, D,$$

donde β es el vector de coeficientes, u_d es el *efecto aleatorio* a nivel de área que representa la heterogeneidad no explicada de los valores Y_{di}, y e_{di} es el error a nivel del individuo.

BLUP/EBLUP basado en el modelo BHF

- Los efectos aleatorios se consideran independientes de los errores, con

$$u_d \overset{iid}{\sim} (0, \sigma_u^2)$$

y

$$e_{di} \overset{ind}{\sim} (0, \sigma_e^2 k_{di}^2)$$

BLUP/EBLUP basado en el modelo BHF

- Los efectos aleatorios se consideran independientes de los errores, con

$$u_d \overset{iid}{\sim} (0, \sigma_u^2)$$

y

$$e_{di} \overset{ind}{\sim} (0, \sigma_e^2 k_{di}^2)$$

- siendo k_{di} constantes conocidas que representan la posible heteroscedasticidad.

BLUP/EBLUP basado en el modelo BHF

- La media del área d se puede escribir con la suma de los valores muestreados y los no muestreados, en esta forma:

$$\overline{Y}_d = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} Y_{di}\right)$$

BLUP/EBLUP basado en el modelo BHF

- La media del área d se puede escribir con la suma de los valores muestreados y los no muestreados, en esta forma:

$$\overline{Y}_d = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} Y_{di}\right)$$

- El estimador BLUP basado en nuestro modelo se obtiene ajustando el modelo con los valores muestreados para predecir los no muestreados:

$$\tilde{\overline{Y}}_d^{BLUP} = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} \tilde{Y}_{di}^{BLUP}\right)$$

BLUP/EBLUP basado en el modelo BHF

- Para estimar $\tilde{Y}_{di}^{BLUP}$, usamos

$$\tilde{Y}_{di}^{BLUP} = \mathbf{x}'_{di}\tilde{\beta} + \tilde{u}_d$$

donde

$$\tilde{u}_d = \gamma_d(\bar{y}_{da} - \bar{\mathbf{x}}'_{da}\tilde{\beta}),$$

y

$$\gamma_d = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/a_{d\cdot})$$

BLUP/EBLUP basado en el modelo BHF

- Para estimar $\tilde{Y}_{di}^{BLUP}$, usamos

$$\tilde{Y}_{di}^{BLUP} = \mathbf{x}'_{di}\tilde{\beta} + \tilde{u}_d$$

donde

$$\tilde{u}_d = \gamma_d(\bar{y}_{da} - \bar{\mathbf{x}}'_{da}\tilde{\beta}),$$

y

$$\gamma_d = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/a_{d\cdot})$$

- $\bar{y}_{da} = a_{d\cdot}^{-1}\sum_{i\in s_d} a_{di}Y_{di}$ y $\bar{\mathbf{x}}_{da} = a_{d\cdot}^{-1}\sum_{i\in s_d} a_{di}\mathbf{x}_{di}$ son las medias muestrales ponderadas con pesos $a_{di} = k_{di}^{-2}$, donde $a_{d\cdot} = \sum_{i\in s_d} a_{di}$

BLUP/EBLUP basado en el modelo BHF

- Definamos $\mathbf{y}_d = (Y_{d1}, \ldots, Y_{dN_d})'$ un vector de variables respuestas para el área d y $\mathbf{X}_d = (\mathbf{x}_{d1}, \ldots, \mathbf{x}_{dN_d})'$ la matriz de covariables en el área d.

BLUP/EBLUP basado en el modelo BHF

- Definamos $\mathbf{y}_d = (Y_{d1}, \ldots, Y_{dN_d})'$ un vector de variables respuestas para el área d y $\mathbf{X}_d = (\mathbf{x}_{d1}, \ldots, \mathbf{x}_{dN_d})'$ la matriz de covariables en el área d.
- Bajo el modelo de errores anidados, $\mathbf{y}_d \overset{ind}{\sim} N(\mathbf{X}_d\beta, \mathbf{V}_d)$, $d = 1, \ldots, D$, donde

$$\mathbf{V}_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}_{N_d}' + \sigma_e^2 \mathbf{A}_d,$$

BLUP/EBLUP basado en el modelo BHF

- Definamos $\mathbf{y}_d = (Y_{d1}, \ldots, Y_{dN_d})'$ un vector de variables respuestas para el área d y $\mathbf{X}_d = (\mathbf{x}_{d1}, \ldots, \mathbf{x}_{dN_d})'$ la matriz de covariables en el área d.
- Bajo el modelo de errores anidados, $\mathbf{y}_d \overset{ind}{\sim} N(\mathbf{X}_d\beta, \mathbf{V}_d)$, $d = 1, \ldots, D$, donde

$$\mathbf{V}_d = \sigma_u^2 \mathbf{1}_{N_d} \mathbf{1}'_{N_d} + \sigma_e^2 \mathbf{A}_d,$$

- $\mathbf{A}_d = \mathrm{diag}(k_{di}^2; i = 1, \ldots, N_d)$

BLUP/EBLUP basado en el modelo BHF

- Sea
$$\mathbf{V}_d = \sigma_u^2 \mathbf{1}_{N_d} \mathbf{1}'_{N_d} + \sigma_e^2 \mathbf{A}_d, = \begin{pmatrix} \mathbf{V}_{ds} & \mathbf{V}_{dsr} \\ \mathbf{V}_{drs} & \mathbf{V}_{dr} \end{pmatrix}$$
donde s representa los individuos muestreados y r representa los no muestreados.

BLUP/EBLUP basado en el modelo BHF

- Sea
$$\mathbf{V}_d = \sigma_u^2 \mathbf{1}_{N_d} \mathbf{1}'_{N_d} + \sigma_e^2 \mathbf{A}_d, = \begin{pmatrix} \mathbf{V}_{ds} & \mathbf{V}_{dsr} \\ \mathbf{V}_{drs} & \mathbf{V}_{dr} \end{pmatrix}$$
donde s representa los individuos muestreados y r representa los no muestreados.
- Entonces, el estimador de MMCC ponderados de β está dado por
$$\tilde{\beta} = \left(\sum_{d=1}^{D} \mathbf{X}_{ds} \mathbf{V}_{ds}^{-1} \mathbf{X}'_{ds} \right)^{-1} \sum_{d=1}^{D} \mathbf{X}_{ds} \mathbf{V}_{ds}^{-1} \mathbf{y}_{ds}$$

BLUP/EBLUP basado en el modelo BHF

- Para áreas donde $n_d/N_d \approx 0$, el BLUP de la media $\overline{Y}_d$ se puede escribir como

$$\tilde{\overline{Y}}_d^{BLUP} \approx \gamma_d \left\{ \bar{y}_{da} + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_{da})' \tilde{\beta} \right\} + (1-\gamma_d) \bar{\mathbf{X}}_d' \tilde{\beta},$$

lo que representa una suma ponderada entre $\bar{y}_{da} + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_{da})' \tilde{\beta}$, conocido como estimador *"survey regression"* y el estimador sintético de regresión, $\bar{\mathbf{X}}_d' \tilde{\beta}$.

BLUP/EBLUP basado en el modelo BHF

- Para áreas donde $n_d/N_d \approx 0$, el BLUP de la media $\overline{Y}_d$ se puede escribir como

$$\tilde{\overline{Y}}_d^{BLUP} \approx \gamma_d \left\{ \bar{y}_{da} + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_{da})' \tilde{\beta} \right\} + (1-\gamma_d) \bar{\mathbf{X}}_d' \tilde{\beta},$$

lo que representa una suma ponderada entre $\bar{y}_{da} + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_{da})' \tilde{\beta}$, conocido como estimador *"survey regression"* y el estimador sintético de regresión, $\bar{\mathbf{X}}_d' \tilde{\beta}$.

- El estimador "survey-regression" se obtiene de ajustar el mismo modelo de errores anidados, pero tomando los efectos de las áreas u_d como fijos en lugar de aleatorios.

BLUP/EBLUP basado en el modelo BHF

- Para interpretar

$$\tilde{\bar{Y}}_d^{BLUP} \approx \gamma_d \left\{ \bar{y}_{da} + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_{da})'\tilde{\beta} \right\} + (1-\gamma_d)\bar{\mathbf{X}}'_d\tilde{\beta},$$

consideremos un modelo homoscedástico, es decir $k_{di} = 1$.

BLUP/EBLUP basado en el modelo BHF

- Para interpretar

$$\tilde{\bar{Y}}_d^{BLUP} \approx \gamma_d \left\{ \bar{y}_{da} + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_{da})' \tilde{\beta} \right\} + (1-\gamma_d)\bar{\mathbf{X}}_d' \tilde{\beta},$$

consideremos un modelo homoscedástico, es decir $k_{di} = 1$.

- En este caso, $\gamma_d = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2 / n_d)$.

BLUP/EBLUP basado en el modelo BHF

- Para interpretar

$$\tilde{\bar{Y}}_d^{BLUP} \approx \gamma_d \left\{ \bar{y}_{da} + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_{da})' \tilde{\beta} \right\} + (1 - \gamma_d) \bar{\mathbf{X}}_d' \tilde{\beta},$$

  consideremos un modelo homoscedástico, es decir $k_{di} = 1$.
- En este caso, $\gamma_d = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2 / n_d)$.
- Para un tamaño n_d pequeño, γ_d es cercano a uno y el BLUP se acerca al estimador "survey regression".

BLUP/EBLUP basado en el modelo BHF

- Para interpretar

$$\tilde{\bar{Y}}_d^{BLUP} \approx \gamma_d \left\{ \bar{y}_{da} + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_{da})' \tilde{\beta} \right\} + (1 - \gamma_d) \bar{\mathbf{X}}_d' \tilde{\beta},$$

  consideremos un modelo homoscedástico, es decir $k_{di} = 1$.
- En este caso, $\gamma_d = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2 / n_d)$.
- Para un tamaño n_d pequeño, γ_d es cercano a uno y el BLUP se acerca al estimador "survey regression".
- También si σ_u^2 es grande comparada con σ_e^2 / n_d, el BLUP acerca al estimador "survey regression".

BLUP/EBLUP basado en el modelo BHF

- Si sustituimos los verdaderos valores, $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)'$ con $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$, obtenemos el estimador EBIUP:

$$\hat{\bar{Y}}_d^{EBLUP} = N_d^{-1}\left(\sum_{i \in s_d} Y_{di} + \sum_{i \in r_d} \hat{Y}_{di}^{EBLUP}\right)$$

donde

$$Y_{di}^{EBLUP} = \mathbf{x}_{di}'\hat{\boldsymbol{\beta}} + \hat{u}_d$$

BLUP/EBLUP basado en el modelo BHF

- En el EBLUP $Y_{di}^{EBLUP} = \mathbf{x}'_{di}\hat{\beta} + \hat{u}_d$, $\hat{\beta}$ es el resultado de sustituir $\boldsymbol{\theta}$ por $\hat{\boldsymbol{\theta}}$ en $\tilde{\beta}$.

BLUP/EBLUP basado en el modelo BHF

- En el EBLUP $Y_{di}^{EBLUP} = \mathbf{x}'_{di}\hat{\beta} + \hat{u}_d$, $\hat{\beta}$ es el resultado de sustituir $\boldsymbol{\theta}$ por $\hat{\boldsymbol{\theta}}$ en $\tilde{\beta}$.
- Donde

$$\hat{u}_d = \hat{\gamma}_d(\bar{y}_{da} - \bar{\mathbf{x}}'_{da}\hat{\beta})$$

y

$$\hat{\gamma}_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_e^2/a_{d\cdot})$$

BLUP/EBLUP basado en el modelo BHF: Sesgo y ECM

- El EBLUP, al igual que el BLUP, sigue siendo insesgado bajo el modelo.

BLUP/EBLUP basado en el modelo BHF: Sesgo y ECM

- El EBLUP, al igual que el BLUP, sigue siendo insesgado bajo el modelo.
- Ni el BLUP ni el EBLUP son insesgados bajo el diseño muestral.

BLUP/EBLUP basado en el modelo BHF: Sesgo y ECM

- El EBLUP, al igual que el BLUP, sigue siendo insesgado bajo el modelo.
- Ni el BLUP ni el EBLUP son insesgados bajo el diseño muestral.
- No obstante, los estimadores BLUP y EBLUP aumenten la eficiencia respecto de los estimadores directos y respecto de los estimadores FH porque usan información mucho más detallada.

BLUP/EBLUP basado en el modelo BHF: Sesgo y ECM

- Para un área no muestreada, fijamos $\gamma_d = 0$, obtenemos el estimador sintético de regresión $\bar{\mathbf{X}}'_d \hat{\beta}$.

BLUP/EBLUP basado en el modelo BHF: Sesgo y ECM

- Para un área no muestreada, fijamos $\gamma_d = 0$, obtenemos el estimador sintético de regresión $\bar{\mathbf{X}}'_d\hat{\beta}$.
- Bajo MAS (muestreo aleatorio simple) y $k_{di} = 1$ para todas los i y d, y $n_d/N_d \approx 0$, el sesgo absoluto relativo (SAR) bajo el diseño es igual a

$$(1-\gamma_d)\left|\frac{\bar{Y}_d - \bar{\mathbf{X}}'_d\beta}{\bar{Y}_d}\right| \leq \left|\frac{\bar{Y}_d - \bar{\mathbf{X}}'_d\beta}{\bar{Y}_d}\right|,$$

BLUP/EBLUP basado en el modelo BHF: Sesgo y ECM

- Para un área no muestreada, fijamos $\gamma_d = 0$, obtenemos el estimador sintético de regresión $\bar{\mathbf{X}}'_d\hat{\beta}$.
- Bajo MAS (muestreo aleatorio simple) y $k_{di} = 1$ para todas los i y d, y $n_d/N_d \approx 0$, el sesgo absoluto relativo (SAR) bajo el diseño es igual a

$$(1-\gamma_d)\left|\frac{\bar{Y}_d - \bar{\mathbf{X}}'_d\beta}{\bar{Y}_d}\right| \leq \left|\frac{\bar{Y}_d - \bar{\mathbf{X}}'_d\beta}{\bar{Y}_d}\right|,$$

- es decir, es menor que el sesgo absoluto relativo bajo el diseño del estimador sintético de regresión $\bar{\mathbf{X}}'_d\beta$ para el mismo vector de coeficientes β, $|(\bar{Y}_d - \bar{\mathbf{X}}'_d\beta)/\bar{Y}_d|$, mientras $\gamma_d > 0$.

BLUP/EBLUP basado en el modelo BHF: Sesgo y ECM

Para estimar el ECM del EBLUP $\hat{\bar{Y}}_d^{EBLUP}$ de $\bar{Y}_d$, podemos usar un procedimiento bootstrap:

1) Ajustar el modelo de errores anidados $Y_{di} = \mathbf{x}'_{di}\beta + u_d + e_{di}$ a los datos de la muestra para obtener estimadores de los parámetros $\hat{\beta}$, $\hat{\sigma}_u^2$ y $\hat{\sigma}_e^2$.

BLUP/EBLUP basado en el modelo BHF: Sesgo y ECM

Para estimar el ECM del EBLUP $\hat{\bar{Y}}_d^{EBLUP}$ de $\bar{Y}_d$, podemos usar un procedimiento bootstrap:

1) Ajustar el modelo de errores anidados $Y_{di} = \mathbf{x}'_{di}\beta + u_d + e_{di}$ a los datos de la muestra para obtener estimadores de los parámetros $\hat{\beta}$, $\hat{\sigma}_u^2$ y $\hat{\sigma}_e^2$.
2) Generar los efectos de las áreas de la forma $u_d^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2)$, $d = 1, \ldots, D$

BLUP/EBLUP basado en el modelo BHF: Sesgo y ECM

3) Generar errores bootstrap para las unidades de la muestra en el área, $e_{di}^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$, $i \in s_d$. Generar también las medias poblacionales de los errores en las áreas, $\bar{E}_d^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2/N_d)$, $d = 1, \ldots, D$

BLUP/EBLUP basado en el modelo BHF: Sesgo y ECM

3) Generar errores bootstrap para las unidades de la muestra en el área, $e_{di}^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$, $i \in s_d$. Generar también las medias poblacionales de los errores en las áreas, $\bar{E}_d^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2 / N_d)$, $d = 1, \ldots, D$
4) Calcular las verdaderas medias bootstrap de las áreas,

$$\bar{Y}_d^{*(b)} = \bar{\mathbf{X}}'_d \hat{\beta} + u_d^{*(b)} + \bar{E}_d^{*(b)}, \quad d = 1, \ldots, D$$

Nótese que que este cómputo no requiere los valores individuales de $\mathbf{x}_d i$ para unidades fuera de la muestra.

BLUP/EBLUP basado en el modelo BHF: Sesgo y ECM

5) Usando los valores de las p variables auxiliares, generar las variables respuestas

$$Y_{di}^{*(b)} = \mathbf{x}'_{di}\hat{\beta} + u_d^{*(b)} + e_{di}^{*(b)}, \quad i \in s_d, \quad d = 1, \ldots, D$$

BLUP/EBLUP basado en el modelo BHF: Sesgo y ECM

5) Usando los valores de las p variables auxiliares, generar las variables respuestas

$$Y_{di}^{*(b)} = \mathbf{x}_{di}'\hat{\beta} + u_d^{*(b)} + e_{di}^{*(b)}, \quad i \in s_d, \quad d = 1, \ldots, D$$

6) Para la muestra original $s = s_1 \cup \cdots \cup s_D$, sea $\mathbf{y}_s^{*(b)} = ((\mathbf{y}_{1s}^{*(b)})', \ldots, (\mathbf{y}_{Ds}^{*(b)})')'$ el vector bootstrap de valores en la muestra. Ajustar el modelo de errores anidados a los datos bootstrap $\mathbf{y}_s^{*(b)}$ y calcular los EBLUPs bootstrap $\hat{\bar{Y}}_d^{EBLUP*(b)}$, $d = 1, \ldots, D$.

BLUP/EBLUP basado en el modelo BHF: Sesgo y ECM

7) Repetir los pasos 2-6 para $b = 1, \ldots, B$. El estimador "naive bootstrap" del ECM de los EBLUP $\hat{\bar{Y}}_d^{EBLUP}$ viene dado por:

$$mse_B(\hat{\bar{Y}}_d^{EBLUP}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\bar{Y}}_d^{EBLUP*(b)} - \bar{Y}_d^{*(b)}\right)^2, \quad d = 1, \ldots, D$$

Este estimador no es insesgado de segundo orden, sino de primer orden; es decir, su sesgo no decrece más rápido que D^{-1} cuando el número de áreas D crece.

Resumen del EBLUP basado en modelo BHF

- Indicadores objetivos: Medias/Totales de la variable de interés

Resumen del EBLUP basado en modelo BHF

- Indicadores objetivos: Medias/Totales de la variable de interés
- Requerimientos de datos:

Resumen del EBLUP basado en modelo BHF

- Indicadores objetivos: Medias/Totales de la variable de interés
- Requerimientos de datos:
  - Microdatos de las p variables auxiliares de la encuesta con la variable de interés.

Resumen del EBLUP basado en modelo BHF

- Indicadores objetivos: Medias/Totales de la variable de interés
- Requerimientos de datos:
  - Microdatos de las p variables auxiliares de la encuesta con la variable de interés.
  - Área de interés obtenida de la misma encuesta.

Resumen del EBLUP basado en modelo BHF

- Indicadores objetivos: Medias/Totales de la variable de interés
- Requerimientos de datos:
  - Microdatos de las p variables auxiliares de la encuesta con la variable de interés.
  - Área de interés obtenida de la misma encuesta.
  - Medias poblacionales de las p variables auxiliares en las áreas, $\overline{\mathbf{X}}_d$.

Resumen del EBLUP basado en modelo BHF

- Ventajas:

Resumen del EBLUP basado en modelo BHF

- Ventajas:
  - El tamaño muestral es de todos los individuos, y por eso, tiene más eficiencia que el estimador FH.

Resumen del EBLUP basado en modelo BHF

- Ventajas:
  - El tamaño muestral es de todos los individuos, y por eso, tiene más eficiencia que el estimador FH.
  - El modelo incluye heterogeneidad no explicada entre las áreas.

Resumen del EBLUP basado en modelo BHF

- Ventajas:
  - El tamaño muestral es de todos los individuos, y por eso, tiene más eficiencia que el estimador FH.
  - El modelo incluye heterogeneidad no explicada entre las áreas.
  - Es un estimador compuesto que toma prestada información del resto de áreas y da mayor peso al estimador sintético de regresión cuando el tamaño muestral es pequeño.

Resumen del EBLUP basado en modelo BHF

- Ventajas:

Resumen del EBLUP basado en modelo BHF

- Ventajas:
  - Al contrario que el modelo FH, no se necesita ninguna varianza.

Resumen del EBLUP basado en modelo BHF

- Ventajas:
  - Al contrario que el modelo FH, no se necesita ninguna varianza.
  - El estimador del ECM bajo el modelo es estable bajo el diseño e insesgado bajo el diseño cuando se promedia a lo largo de muchas áreas.

Resumen del EBLUP basado en modelo BHF

- Ventajas:
  - Al contrario que el modelo FH, no se necesita ninguna varianza.
  - El estimador del ECM bajo el modelo es estable bajo el diseño e insesgado bajo el diseño cuando se promedia a lo largo de muchas áreas.
  - Se pueden desagregar las estimaciones para cualquier subárea dentro de las áreas.

Resumen del EBLUP basado en modelo BHF

- Ventajas:
  - Al contrario que el modelo FH, no se necesita ninguna varianza.
  - El estimador del ECM bajo el modelo es estable bajo el diseño e insesgado bajo el diseño cuando se promedia a lo largo de muchas áreas.
  - Se pueden desagregar las estimaciones para cualquier subárea dentro de las áreas.
  - Se puede estimar en áreas no muestreadas.

Resumen del EBLUP basado en modelo BHF

- Desventajas:

Resumen del EBLUP basado en modelo BHF

- Desventajas:
  - Es basado en un modelo y es necesario analizar ese modelo.

Resumen del EBLUP basado en modelo BHF

- Desventajas:
  - Es basado en un modelo y es necesario analizar ese modelo.
  - No tiene en cuenta el diseño muestral y no es insesgado bajo el modelo. Por eso, es más apropiado usarlo en un MAS.

Resumen del EBLUP basado en modelo BHF

- Desventajas:
  - Es basado en un modelo y es necesario analizar ese modelo.
  - No tiene en cuenta el diseño muestral y no es insesgado bajo el modelo. Por eso, es más apropiado usarlo en un MAS.
  - Se ve afectado por observaciones atípicas aisladas o la falta de normalidad.

Resumen del EBLUP basado en modelo BHF

- Desventajas:

Resumen del EBLUP basado en modelo BHF

- Desventajas:
  - Los microdatos suelen ser obtenidos de un censo, lo que conlleva problemas de confidencialidad.

Resumen del EBLUP basado en modelo BHF

- Desventajas:
  - Los microdatos suelen ser obtenidos de un censo, lo que conlleva problemas de confidencialidad.
  - El estimador Prasad-Rao de ECM que vimos con el estimador FH igualmente es correcto bajo normalidad de los errores, pero no es insesgado bajo el diseño para el ECM bajo el diseño en un área concreta.

Resumen del EBLUP basado en modelo BHF

- Desventajas:
  - Los microdatos suelen ser obtenidos de un censo, lo que conlleva problemas de confidencialidad.
  - El estimador Prasad-Rao de ECM que vimos con el estimador FH igualmente es correcto bajo normalidad de los errores, pero no es insesgado bajo el diseño para el ECM bajo el diseño en un área concreta.
  - Requiere un reajuste para verificar la propiedad de "benchmarking".

Método ELL

Método ELL

- El método de Elbers, Lanjouw y Lanjouw (2003) asume un modelo con errores anidados para la transformación logaritmo de la variable de interés.

Método ELL

- El método de Elbers, Lanjouw y Lanjouw (2003) asume un modelo con errores anidados para la transformación logaritmo de la variable de interés.
- Los efectos aleatorios son de las unidades de primera etapa del diseño muestral, no las áreas de interés.

Método ELL

- El método de Elbers, Lanjouw y Lanjouw (2003) asume un modelo con errores anidados para la transformación logaritmo de la variable de interés.
- Los efectos aleatorios son de las unidades de primera etapa del diseño muestral, no las áreas de interés.
- Para propósitos de notación, consideramos que estas unidades son las áreas.

Método ELL

- El método de Elbers, Lanjouw y Lanjouw (2003) asume un modelo con errores anidados para la transformación logaritmo de la variable de interés.
- Los efectos aleatorios son de las unidades de primera etapa del diseño muestral, no las áreas de interés.
- Para propósitos de notación, consideramos que estas unidades son las áreas.
- Este método es usado por el Banco Mundial.

Método ELL

- Tomando $Y_{di} = \log(E_{di} + c)$, donde $c > 0$ es una constante, el modelo ELL es

$$Y_{di} = \mathbf{x}'_{di}\beta + u_d + e_{di}, \quad i = 1, \dots, N_d,\ d = 1, \dots, D$$

$$u_d \overset{iid}{\sim} (0, \sigma_u^2)$$

y

$$e_{di} \overset{ind}{\sim} (0, \sigma_e^2 k_{di}^2),$$

siendo u_d y e_{di} independientes, y k_{di} constantes conocidas que representan heteroscedasticidad.

Método ELL

- Tomando $Y_{di} = \log(E_{di} + c)$, donde $c > 0$ es una constante, el modelo ELL es

$$Y_{di} = \mathbf{x}'_{di}\beta + u_d + e_{di}, \quad i = 1, \ldots, N_d,\ d = 1, \ldots, D$$

$$u_d \overset{iid}{\sim} (0, \sigma_u^2)$$

y

$$e_{di} \overset{ind}{\sim} (0, \sigma_e^2 k_{di}^2),$$

siendo u_d y e_{di} independientes, y k_{di} constantes conocidas que representan heteroscedasticidad.

- El estimador ELL de un parámetro general $\delta_d = \delta_d(\mathbf{y}_d)$ bajo este modelo se obtiene mediante un procedimiento bootstrap.

Método ELL

1) A partir de los residuos del modelo ajustado a los datos, se generan efectos aleatorios u_d^* para cada área $d = 1, \ldots, D$, y errores e_{di}^*, para cada individuo $i = 1, \ldots, N_d$, $d = 1, \ldots, D$

Método ELL

1) A partir de los residuos del modelo ajustado a los datos, se generan efectos aleatorios u_d^* para cada área $d = 1, \ldots, D$, y errores e_{di}^*, para cada individuo $i = 1, \ldots, N_d$, $d = 1, \ldots, D$

2) Se generan valores bootstrap de la variable respuesta

$$Y_{di}^* = \mathbf{x}_{di}'\hat{\beta} + u_d^* + e_{di}^*, \quad i = 1, \ldots, N_d,\ d = 1, \ldots, D$$

Método ELL

3) Con esto vector de variables respuestas $\mathbf{y}_d^{*(a)}$, o censo, podemos calcular cualquier indicador de interés.

Método ELL

3) Con esto vector de variables respuestas $\mathbf{y}_d^{*(a)}$, o censo, podemos calcular cualquier indicador de interés.
4) Generar A censos completos y A indicadores $\delta_d^{*(a)} = \delta_d(\mathbf{y}_d^{*(a)})$.

Método ELL

3) Con esto vector de variables respuestas $\mathbf{y}_d^{*(a)}$, o censo, podemos calcular cualquier indicador de interés.
4) Generar A censos completos y A indicadores $\delta_d^{*(a)} = \delta_d(\mathbf{y}_d^{*(a)})$.
5) Finalmente, nuestro estimador ELL viene dado por

$$\hat{\delta}_d^{ELL} = \frac{1}{A}\sum_{a=1}^{A}\delta_d^{*(a)}$$

Método ELL

3) Con esto vector de variables respuestas $\mathbf{y}_d^{*(a)}$, o censo, podemos calcular cualquier indicador de interés.
4) Generar A censos completos y A indicadores $\delta_d^{*(a)} = \delta_d(\mathbf{y}_d^{*(a)})$.
5) Finalmente, nuestro estimador ELL viene dado por

$$\hat{\delta}_d^{ELL} = \frac{1}{A}\sum_{a=1}^{A}\delta_d^{*(a)}$$

- El ECM del estimador se estima de la forma

$$\text{mse}_{ELL}(\hat{\delta}_d^{ELL}) = \frac{1}{A}\sum_{a=1}^{A}(\delta_d^{*(a)} - \hat{\delta}_d^{ELL})^2$$

Método ELL

- Podemos sustituir $E_{di} = \exp(Y_{di}) - c$ en la fórmula del indicador FGT.

Método ELL

- Podemos sustituir $E_{di} = \exp(Y_{di}) - c$ en la fórmula del indicador FGT.
- Obtenemos el indicador de $F_{\alpha d}$ con los valores Y^*_{di} generados para cada censo a de la forma

$$F^{*(a)}_{\alpha d} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\frac{z + c - \exp(Y^{*(a)}_{di})}{z} \right)^{\alpha} I(\exp(Y^{*(a)}_{di}) < z+c),$$

Método ELL

- Podemos sustituir $E_{di} = \exp(Y_{di}) - c$ en la fórmula del indicador FGT.
- Obtenemos el indicador de $F_{\alpha d}$ con los valores Y^*_{di} generados para cada censo a de la forma

$$F^{*(a)}_{\alpha d} = \frac{1}{N_d}\sum_{i=1}^{N_d}\left(\frac{z + c - \exp(Y^{*(a)}_{di})}{z}\right)^{\alpha} I(\exp(Y^{*(a)}_{di}) < z+c),$$

- El estimador ELL de $F_{\alpha d}$ viene dado en la forma:

$$\hat{F}^{ELL}_{\alpha d} = \frac{1}{A}\sum_{a=1}^{A} F^{*(a)}_{\alpha d}$$

Método ELL

- Calculamos la media del área d en el censo a con

$$\bar{Y}_d^{*(a)} \approx \overline{\mathbf{X}}'_d \hat{\beta} + u_d^{*(a)}$$

Método ELL

- Calculamos la media del área d en el censo a con

$$\bar{Y}_d^{*(a)} \approx \overline{\mathbf{X}}_d' \hat{\beta} + u_d^{*(a)}$$

- A lo largo de las réplicas boostrap,

$$A^{-1} \sum_{a=1}^{A} u_d^{*(a)} \approx E(u_d) = 0$$

Método ELL

- Calculamos la media del área d en el censo a con

$$\bar{Y}_d^{*(a)} \approx \overline{\mathbf{X}}'_d \hat{\beta} + u_d^{*(a)}$$

- A lo largo de las réplicas boostrap,

$$A^{-1} \sum_{a=1}^{A} u_d^{*(a)} \approx E(u_d) = 0$$

- Por tanto, el estimador ELL para una media resulta ser el estimador sintético de regresión

$$\hat{\bar{Y}}_d^{ELL} = \bar{\mathbf{X}}'_d \hat{\beta}$$

Método ELL

- Calculamos la media del área d en el censo a con

$$\bar{Y}_d^{*(a)} \approx \overline{\mathbf{X}}'_d \hat{\beta} + u_d^{*(a)}$$

- A lo largo de las réplicas boostrap,

$$A^{-1} \sum_{a=1}^{A} u_d^{*(a)} \approx E(u_d) = 0$$

- Por tanto, el estimador ELL para una media resulta ser el estimador sintético de regresión

$$\hat{\bar{Y}}_d^{ELL} = \bar{\mathbf{X}}'_d \hat{\beta}$$

- Esto puede ser muy sesgado si el modelo de regresión sin efectos aleatorios no se verifica.

Resumen del método ELL

- Indicadores objetivos: Parámetros generales

Resumen del método ELL

- Indicadores objetivos: Parámetros generales
- Requerimientos de datos:

Resumen del método ELL

- Indicadores objetivos: Parámetros generales
- Requerimientos de datos:
    - Microdatos de las p variables auxiliares de la encuesta.

Resumen del método ELL

- Indicadores objetivos: Parámetros generales
- Requerimientos de datos:
  - Microdatos de las p variables auxiliares de la encuesta.
  - Área de interés obtenida de la misma encuesta.

Resumen del método ELL

- Indicadores objetivos: Parámetros generales
- Requerimientos de datos:
    - Microdatos de las p variables auxiliares de la encuesta.
    - Área de interés obtenida de la misma encuesta.
    - Datos de las p variables auxiliares consideradas en las áreas de un censo o registro.

Resumen del método ELL

- Ventajas:

Resumen del método ELL

- Ventajas:
  - Basado en datos a nivel de individuo (incluye mucho más información).

Resumen del método ELL

- Ventajas:
  - Basado en datos a nivel de individuo (incluye mucho más información).
  - Permite estimar indicadores cualesquiera que estén definidos como una función de la variable respuesta Y_{di}.

Resumen del método ELL

- Ventajas:
  - Basado en datos a nivel de individuo (incluye mucho más información).
  - Permite estimar indicadores cualesquiera que estén definidos como una función de la variable respuesta Y_{di}.
  - Son insesgados bajo el modelo si los parámetros son conocidos.

Resumen del método ELL

- Ventajas:
  - Basado en datos a nivel de individuo (incluye mucho más información).
  - Permite estimar indicadores cualesquiera que estén definidos como una función de la variable respuesta Y_{di}.
  - Son insesgados bajo el modelo si los parámetros son conocidos.
  - Se puede estimar para cualquier subarea o subdominio, incluso a nivel de individuo.

Resumen del método ELL

- Ventajas:
  - Basado en datos a nivel de individuo (incluye mucho más información).
  - Permite estimar indicadores cualesquiera que estén definidos como una función de la variable respuesta Y_{di}.
  - Son insesgados bajo el modelo si los parámetros son conocidos.
  - Se puede estimar para cualquier subarea o subdominio, incluso a nivel de individuo.
  - Una vez se ajusta el modelo, se pueden estimar indicadores sin necesidad de ajustar modelos distintos para cada indicador.

Resumen del método ELL

- Desventajas:

Resumen del método ELL

- Desventajas:
  - Los estimadores ELL pueden presentar un alto ECM bajo el modelo y pueden comportarse peor que estimadores directos.

Resumen del método ELL

- Desventajas:
  - Los estimadores ELL pueden presentar un alto ECM bajo el modelo y pueden comportarse peor que estimadores directos.
  - Los estimadores están basados en un modelo y se necesita por tanto, comprobar que el modelo se ajusta correctamente a los datos.

Resumen del método ELL

- Desventajas:
  - Los estimadores ELL pueden presentar un alto ECM bajo el modelo y pueden comportarse peor que estimadores directos.
  - Los estimadores están basados en un modelo y se necesita por tanto, comprobar que el modelo se ajusta correctamente a los datos.
  - No son insesgados bajo el diseño.

Resumen del método ELL

- Desventajas:
  - Los estimadores ELL pueden presentar un alto ECM bajo el modelo y pueden comportarse peor que estimadores directos.
  - Los estimadores están basados en un modelo y se necesita por tanto, comprobar que el modelo se ajusta correctamente a los datos.
  - No son insesgados bajo el diseño.
  - Pueden verse afectados seriamente por datos atípicos aislados.

Resumen del método ELL

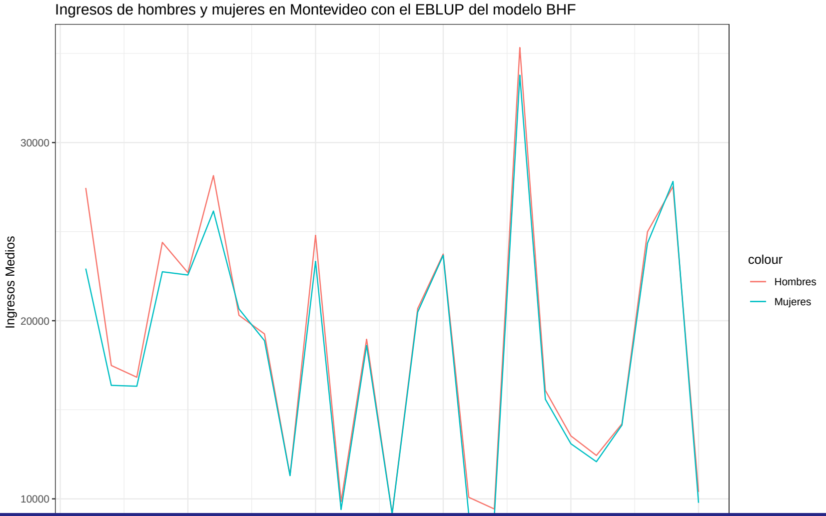
- Desventajas:
  - Los estimadores ELL pueden presentar un alto ECM bajo el modelo y pueden comportarse peor que estimadores directos.
  - Los estimadores están basados en un modelo y se necesita por tanto, comprobar que el modelo se ajusta correctamente a los datos.
  - No son insesgados bajo el diseño.
  - Pueden verse afectados seriamente por datos atípicos aislados.
  - Si incluye efectos de conglomerados y no de áreas cuando hay heterogeneidad entre las áreas, los estimadores ELL del ECM no estiman el verdadero ECM de los estimadores ELL para cada área.

Resultados: Estimación de ingreso medio en sectores de Montevideo

EBLUP basado en el modelo BHF: Hombres y Mujeres en Montevideo

sec2	ntotal	Hombres	Mujeres
2	121	17486	16370
1	167	27455	22927
3	186	16826	16322
4	319	24397	22747
6	320	28146	26156
5	495	22697	22566
21	3165	12436	12085
13	3556	9169	9153
18	3950	35335	33784
11	3963	9850	9398
17	4373	9432	8967
10	6302	24793	23327

EBLUP basado en el modelo BHF: Hombres y Mujeres en Montevideo



Método ELL: Hombres y Mujeres en Montevideo

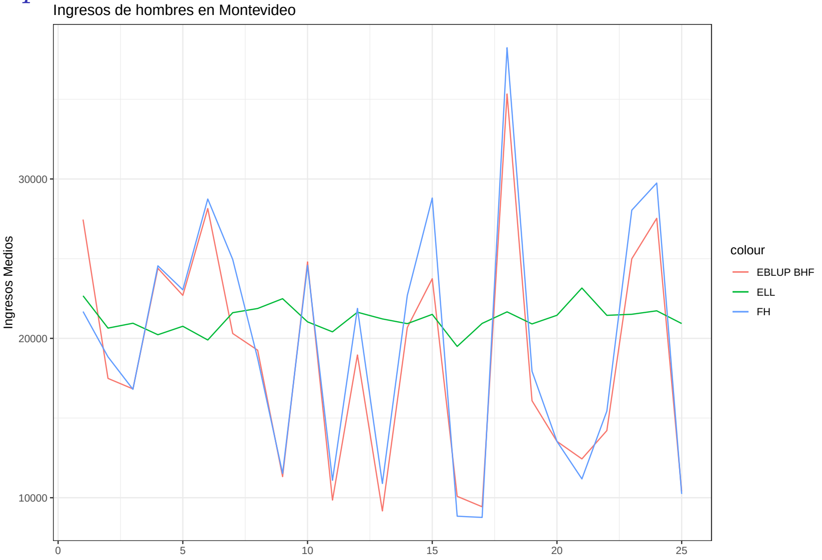
sec2	ntotal	Hombres	Mujeres
2	121	20642	20538
1	167	22669	20524
3	186	20949	21131
4	319	20226	19590
6	320	19897	21920
5	495	20758	21674
21	3165	23158	21411
13	3556	21220	21650
18	3950	21663	21547
11	3963	20409	21028
17	4373	20937	22705
10	6302	21035	20859

Método ELL: Hombres y Mujeres en Montevideo

Comparando los estimadores: Hombres

sec2	ntotal	HT	FH	EBLUP	ELL
2	121	20461	18839	17486	20642
1	167	24837	21682	27455	22669
3	186	14299	16801	16826	20949
4	319	26635	24552	24397	20226
6	320	28784	28744	28146	19897
5	495	23223	23046	22697	20758
21	3165	11148	11180	12436	23158
13	3556	10897	10892	9169	21220
18	3950	38932	38237	35335	21663
11	3963	11080	11092	9850	20409
17	4373	8750	8763	9432	20937
10	6302	24576	24574	24793	21035

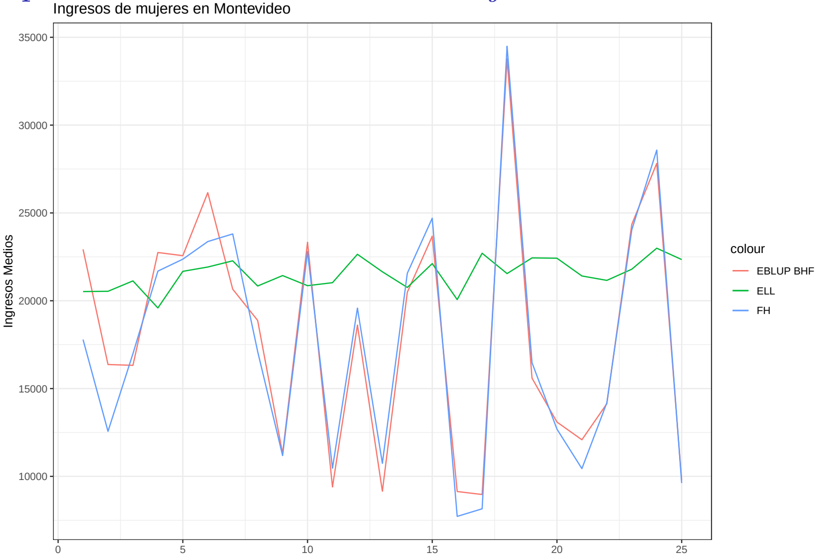
Comparando los estimadores: Hombres



Comparando los estimadores: Mujeres

sec2	totaln	HT	FH	EBLUP	ELL
2	121	13277	12564	16370	20538
1	167	18694	17790	22927	20524
3	186	15951	16987	16322	21131
4	319	21965	21687	22747	19590
6	320	23314	23370	26156	21920
5	495	22414	22366	22566	21674
21	3165	10435	10441	12085	21411
13	3556	10742	10744	9153	21650
18	3950	34943	34490	33784	21547
11	3963	10473	10467	9398	21028
17	4373	8167	8154	8967	22705
10	6302	22823	22802	23327	20859

Comparando los estimadores: Mujeres



¡Gracias!

Email: andres.gutierrez@cepal.org