Curso Internacional de Desagregación de Estimaciones en Áreas Pequeñas usando R

El mejor predictor empírico en modelos de unidad (EB) y el método Censo EB

División de Estadísticas
Comisión Económica para América Latina y el Caribe

2020

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

- Como ya se ha mencionado, estimadores para áreas basados en modelos se consideran modelos indirectos porque usan información de otras áreas
- 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

Mejor predictor empírico (EB, empirical best) bajo el modelo con errores anidados

Mejor predictor empírico bajo el modelo con errores anidados: el modelo

- El mejor predictor (*best/Bayes predictor*, BP) basado en el modelo con errores anidados es para estimar indicadores no lineales generales (Molina y Rao 2010)
- Este método asume que las variables $Y_{di} = \log(E_{di} + c)$ siguen el modelo

$$Y_{di} = \mathbf{x}'_d i\beta + u_d + e_d$$

con normalidad para los efectos aleatorios de las áreas u_d y para los errores e_{di}.

Mejor predictor empírico bajo el modelo con errores anidados: el modelo

- Bajo este modelo, los vectores de la variable de interés para cada área $\mathbf{y}_d = (Y_{d1}, \ldots, Y_{dN_d})'$, $d = 1, \ldots, D$, son independientes y verifican

$$\mathbf{y}_d \overset{ind}{\sim} N(\boldsymbol{\mu}_d, \mathbf{V}_d)$$

donde $\boldsymbol{\mu}_d = \mathbf{X}_d\beta$, siendo $\mathbf{X}_d = (\mathbf{x}_{d1}, \ldots, \mathbf{x}_{dN_d})'$ y

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

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

Mejor predictor empírico bajo el modelo con errores anidados

- Para un indicador que es una función de $\mathbf{y}_d$, el mejor predictor es aquel que minimiza el ECM, dado por

$$\tilde{\delta}_d^B(\boldsymbol{\theta}) = E_{\mathbf{y}_{dr}}[\delta_d(\mathbf{y}_d)|\mathbf{y}_{ds}; \boldsymbol{\theta}]$$

- La esperanza se toma respecto de la distribución de los datos fuera de la muestra, $\mathbf{y}_{dr}$, dado los valores en la muestra $\mathbf{y}_{ds}$

Mejor predictor empírico bajo el modelo con errores anidados

- El estimador $\tilde{\delta}_d^B(\boldsymbol{\theta}) = E_{\mathbf{y}_{dr}}[\delta_d(\mathbf{y}_d)|\mathbf{y}_{ds}; \boldsymbol{\theta}]$ depende del valor de los parámetros del modelo, $\boldsymbol{\theta}$
- Cuando reemplazamos $\boldsymbol{\theta}$ con un estimador consistente $\hat{\boldsymbol{\theta}}$ (e.g.ML, REML, Henderson III), obtenemos el mejor predictor empírico (*empirical best/Bayes*, EB)

Mejor predictor empírico bajo el modelo con errores anidados

- Para obtener la distribución $\mathbf{y}_{dr}|\mathbf{y}_{ds}$, descomponemos $\mathbf{X}_d$ y $\mathbf{V}_d$ así:

$$\mathbf{y}_d = \begin{pmatrix} \mathbf{y}_{ds} \\ \mathbf{y}_{dr} \end{pmatrix}, \quad \mathbf{X}_d = \begin{pmatrix} \mathbf{X}_{ds} \\ \mathbf{X}_{dr} \end{pmatrix}, \quad \mathbf{V}_d = \begin{pmatrix} \mathbf{V}_{ds} & \mathbf{V}_{dsr} \\ \mathbf{V}_{drs} & \mathbf{V}_{dr} \end{pmatrix}$$

donde s representa la muestra, y r los individuos en área d afuera de la muestra

Mejor predictor empírico bajo el modelo con errores anidados

- Dado que $\mathbf{y}_d$ sigue una distribución normal, las condicionadas también la siguen, es decir

$$\mathbf{y}_{dr}|\mathbf{y}_{ds} \overset{ind}{\sim} N(\boldsymbol{\mu}_{dr|s}, \mathbf{V}_{dr|s}), \quad d = 1, \ldots, D$$

donde

$$\boldsymbol{\mu}_{dr|s} = \mathbf{X}_{dr}\boldsymbol{\beta} + \gamma_d(\bar{y}_{da} - \bar{\mathbf{x}}_{da}^T\boldsymbol{\beta})\mathbf{1}_{N_d - n_d}$$

y

$$\mathbf{V}_{dr|s} = \sigma_u^2(1-\gamma_d)\mathbf{1}_{N_d-n_d}\mathbf{1}_{N_d-n_d}^T + \sigma_e^2 \text{diag}_{i\in r_d}(k_{di}^2)$$

Mejor predictor empírico bajo el modelo con errores anidados

- Para un individuo $i \in r_d$:

$$Y_{di}|\mathbf{y}_{ds} \sim N(\mu_{di|s}, \sigma^2_{di|s}),$$

donde la media y la varianza condicionadas vienen dadas por

$$\mu_{di|s} = \mathbf{x}'_{di}\boldsymbol{\beta} + \gamma_d(\bar{y}_{da} - \bar{\mathbf{x}}^T_{da}\boldsymbol{\beta})$$

y

$$\sigma^2_{di|s} = \sigma^2_u(1-\gamma_d) + \sigma^2_e k^2_{di}$$

Mejor predictor empírico bajo el modelo con errores anidados: indicadores FGT

- Para estimar un indicador FGT, $\delta_d = F_{\alpha d}$, asumiendo que $Y_{di} = log E_{di} + c$ para $c > 0$, primero reescribimos el indicador FGT en cuestión como una función de las variables respuesta en el modelo Y_{di}, es decir

$$F_{\alpha d} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\frac{z + c - \exp(Y_{di})}{z} \right)^{\alpha} I(\exp(Y_{di}) < z + c)$$

- Entonces, calculamos la esperanza del mejor predictor,

$$\tilde{F}^{B}_{\alpha d} = E_{\mathbf{y}_{dr}}[F_{\alpha d} | \mathbf{y}_{ds}; \boldsymbol{\theta}]$$

Mejor predictor empírico bajo el modelo con errores anidados: indicadores FGT

- Para este mejor predictor, separamos la suma que define el indicador FGT, es decir

$$\tilde{F}^B_{\alpha d}(\boldsymbol{\theta}) = \frac{1}{N_d}\left(\sum_{i\in s_d} F_{\alpha,di} + \sum_{i\in r_d} \tilde{F}^B_{\alpha,di}(\boldsymbol{\theta})\right)$$

- Aquí,

$$\tilde{F}^B_{\alpha,di}(\boldsymbol{\theta}) = E[F_{\alpha,di}|\mathbf{y}_{ds};\boldsymbol{\theta}]$$

donde la esperanza se toma respecto de la distribución de $Y_{di}|\mathbf{y}_{ds}$, $i \in r_d$, dada en la diapositiva anterior

Mejor predictor empírico bajo el modelo con errores anidados: indicadores FGT

- Para $\alpha = 0, 1$ puede probar que

$$\tilde{F}^{B}_{0,di}(\boldsymbol{\theta}) = \Phi(\alpha_{di}),$$

y

$$\tilde{F}^{B}_{1,di}(\boldsymbol{\theta}) = \Phi(\alpha_{di}) \left\{ 1 - \frac{1}{z} \left[\exp\left(\mu_{di|s} + \frac{\sigma^2_{di|s}}{2} \right) \frac{\Phi(\alpha_{di} - \sigma_{di|s})}{\Phi(\alpha_{di})} - c \right] \right\}$$

- Con indicadores $\delta_d = \delta_d \mathbf{y}_d$ más complejos, incluyendo indicadores $\alpha > 1$, el mejor predictor se puede calcular usando un proceso de simulación Monte Carlo

Mejor predictor empírico bajo el modelo con errores anidados: proceso Monte-Carlo

1) Obtener un estimador $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ para los verdaderos parámetros ajustando del modelo de errores anidados a los datos $(\mathbf{y}_s, \mathbf{X}_s)$

2) Generar $a = 1, \ldots, A$ vectores de variables respuesta para individuos que no están en la muestra de área d, $\mathbf{y}_{dr}^{(a)}$, usando la distribución $\mathbf{y}_{dr}|\mathbf{y}_{ds}$

Mejor predictor empírico bajo el modelo con errores anidados: proceso Monte-Carlo

- En el segundo paso, puede ser costoso o imposible generar $\mathbf{y}_{dr}^{(a)}$, lo que tiene $N_d - n_d$ valores
- Podemos observar que la matriz de covarianzas de este vector, $\mathbf{V}_{dr|s}$ corresponde a la matriz de covarianzas de un vector aleatorio $\mathbf{y}_{dr}^{(a)}$ generado del modelo

$$\mathbf{y}_{dr}^{(a)} = \boldsymbol{\mu}_{dr|s} + v_d^{(a)} \mathbf{1}_{N_d - n_d} + \boldsymbol{\epsilon}_{dr}^{(a)},$$

donde

$$v_d^{(a)} \sim N(0, \sigma_u^2(1-\gamma_d)), \quad \boldsymbol{\epsilon}_{dr}^{(a)} \sim N(\mathbf{0}_{N_d - n_d}, \sigma_e^2 \text{diag}_{i \in r_d}(k_{di}^2));$$

- Ahora, solo es necesario calcular $1 + N_d - n_d$ variables normales independientes en lugar del vector normal multivariante, $\mathbf{y}_{dr}^{(a)}$

Mejor predictor empírico bajo el modelo con errores anidados: proceso Monte-Carlo

3) Formar el vector censal $\mathbf{y}_d^{(a)} = (\mathbf{y}'_{ds}, (\mathbf{y}_{dr}^{(a)})')'$ y usarlo para calcular

$$\delta_d^{(a)} = \delta_d(\mathbf{y}_d^{(a)})$$

- El estimador viene dado por

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

Mejor predictor empírico bajo el modelo con errores anidados: proceso bootstrap para ECM

Molina y Rao (2010) ofrece una manera de calcular el ECM del mejor predictor usando un método bootstrap

1) Ajustar el modelo de errores anidados a los datos $\mathbf{y}_s = (\mathbf{y}'_{1s}, \ldots, \mathbf{y}'_{Ds})'$ para obtener estimaciones de los parámetros del modelo, $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2_u$ y $\hat{\sigma}^2_e$
2) Generar efectos bootstap para las áreas con

$$u_d^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}^2_u), \quad d = 1, \ldots, D$$

3) Generar errores bootstap para los individuos con

$$e_{di}^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}^2_e), \quad i = 1, \ldots, N_d, d = 1, \ldots, D$$

Mejor predictor empírico bajo el modelo con errores anidados: proceso bootstrap para ECM

4) Generar el censo bootstrap de la variable respuesta con el modelo

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

5) Calcular los indicadores de interés, $\delta_d^{*(b)} = \delta_d(\mathbf{y}_d^{*(b)})$, $d = 1, \ldots, D$, donde $\mathbf{y}_d^{*(b)} = (Y_{d1}^{*(b)}, \ldots, Y_{dN_d}^{*(b)})'$

6) Sea $\mathbf{y}_s^{*(b)} = ((\mathbf{y}_{1s}^{*(b)})', \ldots, (\mathbf{y}_{Ds}^{*(b)})')'$ el vector que contiene las observaciones bootstrap cuyos índices están en la muestra. Ajustar el modelo de errores anidados con $\mathbf{y}_s^{*(b)}$ y obtener los predictores EB bootstrap para el indicador $\hat{\delta}_d^{EB*(b)}$, $d = 1, \ldots, D$

Mejor predictor empírico bajo el modelo con errores anidados: proceso bootstrap para ECM

7) Repetir los pasos 2-6 para obtener $\hat{\delta}_d^{EB*(b)}$ y $\delta_d^{*(b)}$ para cada área.

8) El estimador "naive bootstrap" del ECM del mejor predictor, $\hat{\delta}_d^{EB}$ viene dado por:

$$\text{mse}_B(\hat{\delta}_d^{EB}) = B^{-1}\sum_{b=1}^{B}\left(\hat{\delta}_d^{EB*(b)} - \delta_d^{*(b)}\right)^2, \quad d = 1, \ldots, D$$

Mejor predictor empírico bajo el modelo con errores anidados: predictor "Census best"

- Para estimar indicadores complejos, tanto el método ELL como el EB presentado en esta sección, requiere datos para todas las áreas $\{(E_{di}, \mathbf{x}_{di}); i \in s_d,\ d = 1, \ldots, D\}$.
- Porque se necesita datos de las mismas variables auxiliares para todas las área, a menudo se usa un censo ("census") o registro administrativo.
- Tiene que vincular los datos del censo con los individuos en la muestra de las áreas, lo que a veces puede ser difícil.

Mejor predictor empírico bajo el modelo con errores anidados: predictor "Census best"

- El estimador "census best" se obtiene calculando las esperanzas $\tilde{F}^{B}_{\alpha,di}(\boldsymbol{\theta}) = F^{B}_{\alpha,di}(\boldsymbol{\theta}) = E[F_{\alpha,di}|\mathbf{y}_{ds};\boldsymbol{\theta}]$, también para los individuos de la muestra como si no se observaran.
- el predictor Census best de $F_{\alpha d}$ viene dado por

$$\tilde{F}^{CB}_{\alpha d}(\boldsymbol{\theta}) = \frac{1}{N_d}\sum_{i=1}^{N_d}\tilde{F}^{B}_{\alpha,di}(\boldsymbol{\theta})$$

- Se la esperanza no se puede calcular de una forma analítica, se usa un procedimiento Monte Carlo ya descrito.

Resumen del mejor predictor empírico (EB)

- Note que lo siguiente es aproximadamente igual para el Census EB si n_d/N_d es pequeña.
- Indicadores objetivos: Parámetros generales.
- Requerimientos de datos:
  - Microdatos de las p variables auxiliares de la misma encuesta de la variable de interés.
  - Área de interés obtenida de la misma encuesta.
  - Microdatos de las p variables auxiliares a partir de un censo o un registro administrativo.

Resumen del mejor predictor empírico (EB)

- Ventajas:
  - Basado en datos a nivel de individuo, lo que proporciona información más detallada
  - Permite la estimación de cualquier indicador que es una función de $Y_d i$
  - Son insesgados bajo el modelo si los parámetros son conocidos
  - Son óptimos en el sentido de que minimizan el ECM bajo el modelo para valores conocidos de los parámetros

Resumen del mejor predictor empírico (EB)

- Ventajas:
  - Se comportan mucho mejor que los estimadores ELL en términos de ECM bajo el modelo cuando la heterogeneidad no explicada entre áreas es significativa
  - Una vez que se ajusta el modelo, se puede estimar en subáreas sin tener que reajustar el modelo
  - Una vez que se ajusta el modelo, se puede estimar cualquier indicador que es una función de Y_{di} sin tener que reajustar el modelo

Resumen del mejor predictor empírico (EB)

- Desventajas:
  - Son basados en un modelo y tiene que comprobar que el modelo se ajusta correctamente
  - No tienen en cuenta el diseño muestral, y por eso pueden conllevar sesgo bajo el diseño
  - Pueden ser seriamente afectados por atípicos aislados

Resumen del mejor predictor empírico (EB)

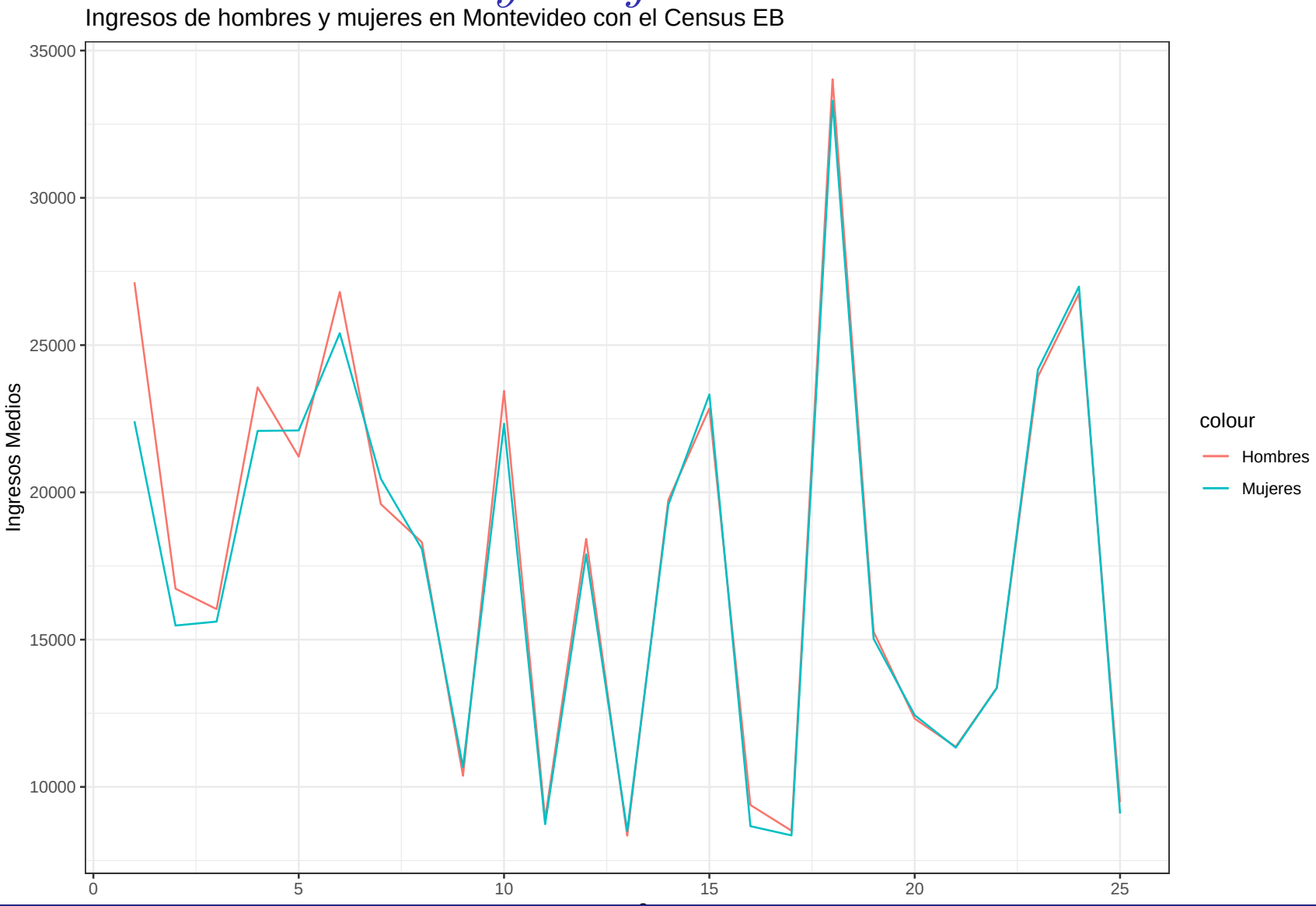
- Desventajas:
  - Estimadores de ECM usando el bootstrap son computacionalmente intensivos:
  - Para la aproximación Monte-Carlo del estimador EB, se necesita A censos $\mathbf{y}^{(a)}$ de grande tamaño
  - Para el estimar el ECM a través del proceso bootstrap requiere que se repita la aproximación Monte Carlo para cada réplica bootstrap

Resultados: Estimación de ingreso medio en sectores de Montevideo

Census EB: Hombres y Mujeres en Montevideo

sec2	ntotal	Hombres	Mujeres
2	121	16729	15479
1	167	27135	22415
3	186	16036	15612
4	319	23565	22086
6	320	26804	25405
5	495	21211	22102
21	3165	11363	11333
13	3556	8344	8502
18	3950	34024	33292
11	3963	8906	8736
17	4373	8513	8354
10	6302	23443	22328

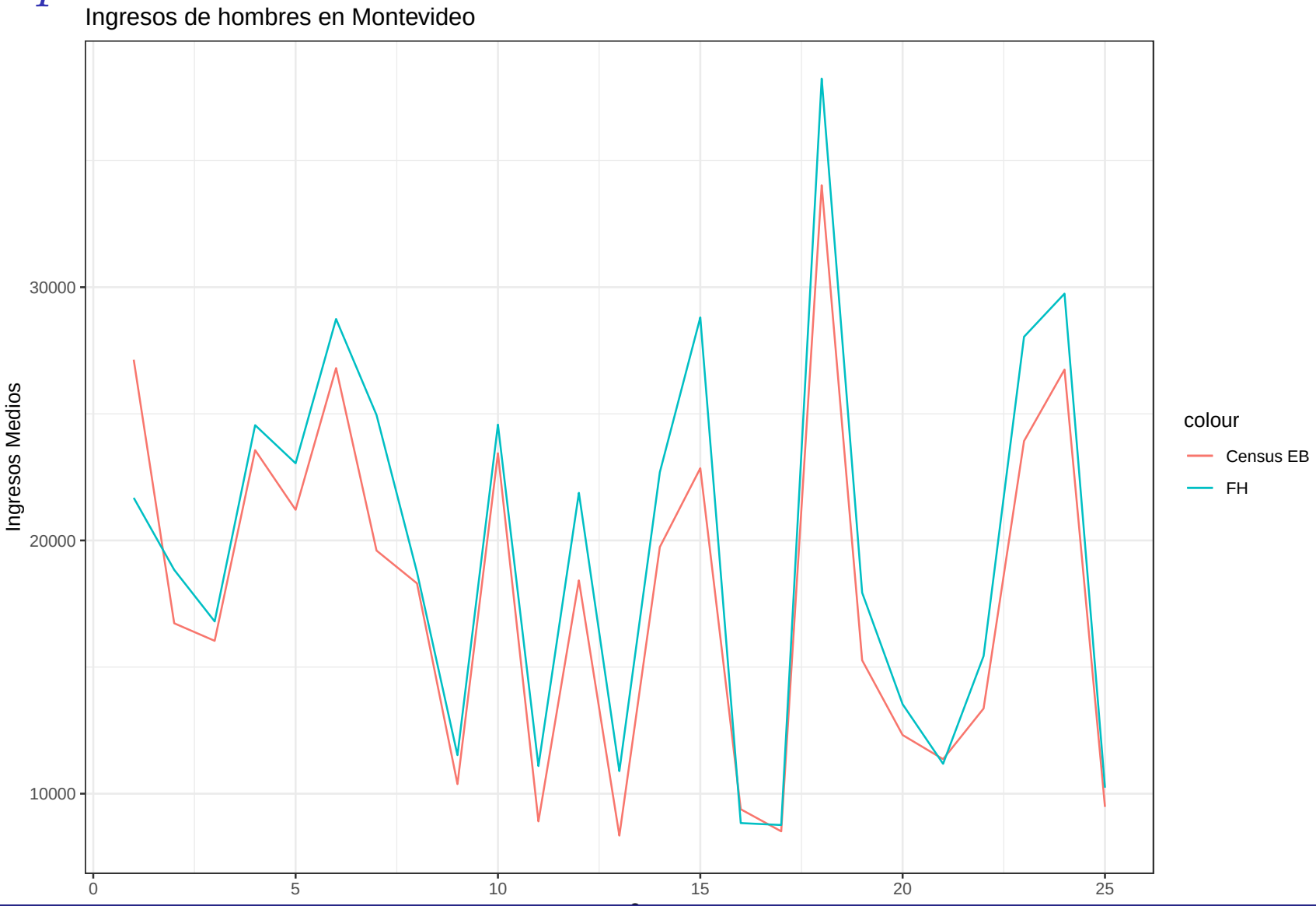
Census EB: Hombres y Mujeres en Montevideo



Comparando los estimadores: Hombres

sec2	ntotal	HT	FH	CensusEB
2	121	20461	18839	16729
1	167	24837	21682	27135
3	186	14299	16801	16036
4	319	26635	24552	23565
6	320	28784	28744	26804
5	495	23223	23046	21211
21	3165	11148	11180	11363
13	3556	10897	10892	8344
18	3950	38932	38237	34024
11	3963	11080	11092	8906
17	4373	8750	8763	8513
10	6302	24576	24574	23443

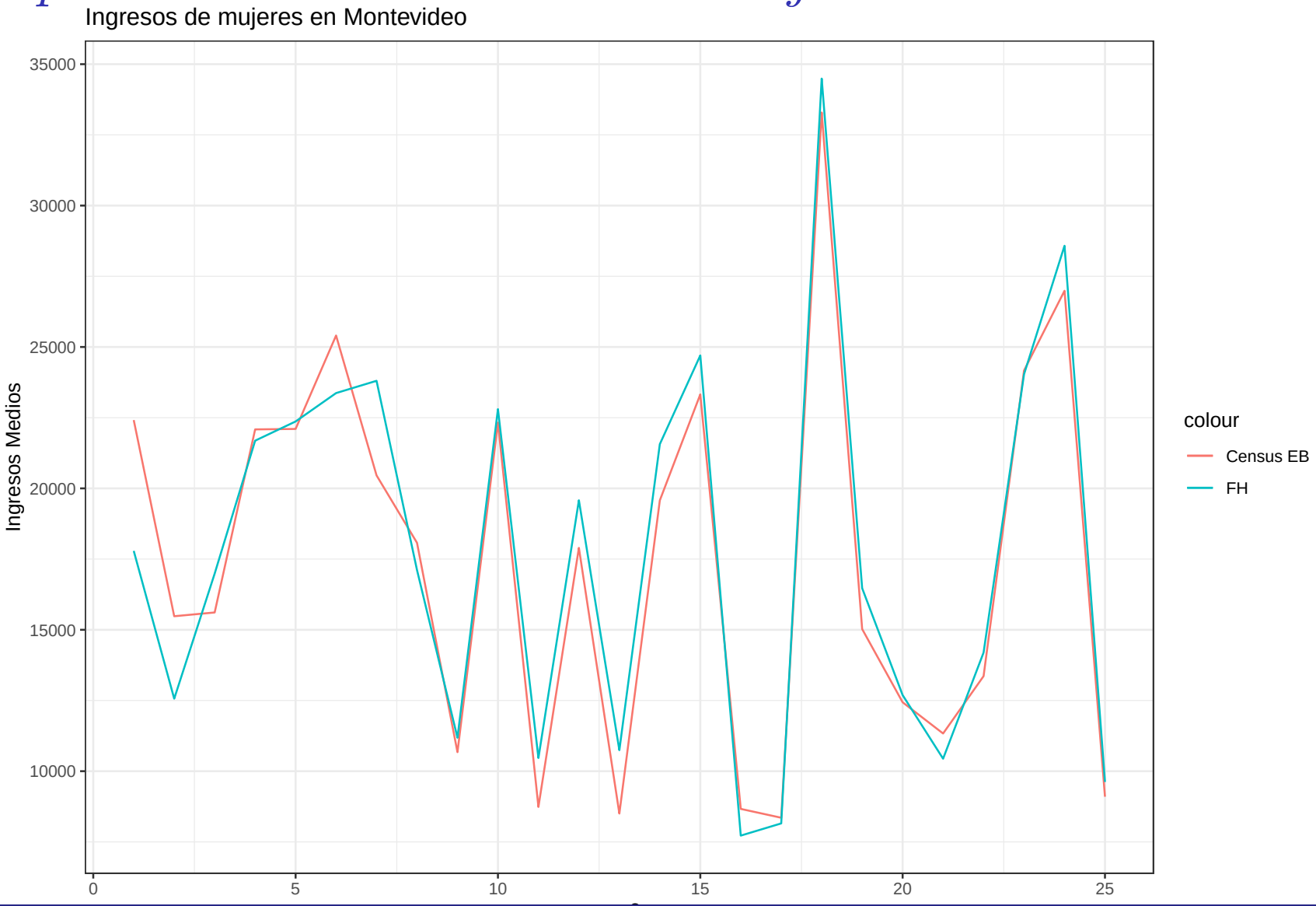
Comparando los estimadores: Hombres



Comparando los estimadores: Mujeres

sec2	ntotal	HT	FH	CensusEB
2	121	13277	12564	15479
1	167	18694	17790	22415
3	186	15951	16987	15612
4	319	21965	21687	22086
6	320	23314	23370	25405
5	495	22414	22366	22102
21	3165	10435	10441	11333
13	3556	10742	10744	8502
18	3950	34943	34490	33292
11	3963	10473	10467	8736
17	4373	8167	8154	8354
10	6302	22823	22802	22328

Comparando los estimadores: Mujeres



¡Gracias!

¡Gracias!