Curso Internacional de Desagregación de Estimaciones en Áreas Pequeñas usando R

Métodos indirectos con modelos de área: EBLUP basado en el modelo de Fay-Herriot

División de Estadísticas
Comisión Económica para América Latina y el Caribe

2020

Referencias

(2018) Molina, Isabel. *Estudio de los límites de desagregación de datos 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

- 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.

Introducción

- Como veremos, los efectos aleatorios ofrece a los estimadores la buena propiedad de poder escribirse como estimadores compuestos que tienden a un estimador directo con tamaño muestral suficiente.
- Como es muy difícil acceder a todas las variables auxiliares que expliquen la heterogeneidad entre las áreas, los estimadores con efectos aleatorios basados en modelos son más realistas que los modelos sintéticos.

BLUP/EBLUP basado en el modelo Fay-Herriot

BLUP/EBLUP basado en el modelo Fay-Herriot

- El modelo FH enlaza indicadores de las *áreas* δ_d, $d = 1, \ldots, D$, asumiendo que varían respeto a un vector de p covariables, $\mathbf{x}_d$, de forma constante.
- Viene dado por

$$\delta_d = \mathbf{x}_d'\beta + u_d, \quad d = 1, \ldots, D$$

- u_d es el término de error, o el efecto aleatorio, diferente para cada área dado por

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

BLUP/EBLUP basado en el modelo Fay-Herriot

- Sin embargo, los verdaderos valores de los indicadores δ_d no son observables.
- Entonces, usamos el estimador directo $\hat{\delta}_d^{DIR}$ para δ_d, lo que conlleva un error debido al muestro.
- $\hat{\delta}_d^{DIR}$ todavía se considera insesgado bajo el diseño muestral.

BLUP/EBLUP basado en el modelo Fay-Herriot

- Podemos definir, entonces,

$$\hat{\delta}_d^{DIR} = \delta_d + e_d, \quad d = 1, \ldots, D,$$

  donde e_d es el error debido al muestreo, $e_d \overset{ind}{\sim} (0, \psi_d)$.
- Dichas varianzas $\psi_d = \text{var}_\pi(\hat{\delta}_d^{DIR} | \delta_d)$, $d = 1, \ldots, D$, se estiman con los microdatos de la encuesta.
- Por tanto, el modelo se hace,

$$\hat{\delta}_d^{DIR} = \mathbf{x}_d' \boldsymbol{\beta} + u_d + e_d, \quad d = 1, \ldots, D$$

BLUP/EBLUP basado en el modelo Fay-Herriot

- Minimizando el ECM bajo el modelo, obtenemos el mejor predictor lineal insesgado (*best linear unbiased predictor, BLUP*) para $\delta_d = \mathbf{x}'_d\beta + u_d$.
- El BLUP bajo el modelo FH de δ_d viene dado por

$$\tilde{\delta}_d^{FH} = \mathbf{x}'_d\tilde{\beta} + \tilde{u}_d$$

- $\tilde{\beta}$ viene dado por

$$\tilde{\beta} = \left(\sum_{d=1}^{D}\left(\frac{1}{\sigma_u^2+\psi_d}\right)\mathbf{x}_d\mathbf{x}'_d\right)^{-1}\sum_{d=1}^{D}\left(\frac{1}{\sigma_u^2+\psi_d}\right)\mathbf{x}_d\hat{\delta}_d^{DIR}$$

Siendo

$$\gamma_d = \sigma_u^2/(\sigma_u^2+\psi_d)$$

BLUP/EBLUP basado en el modelo Fay-Herriot

- En el BLUP del modelo FH,

$$\tilde{u}_d = \gamma_d(\hat{\delta}_d^{DIR} - \mathbf{x}'_d\tilde{\boldsymbol{\beta}})$$

  es el *BLUP de* u_d.
- Si sustituimos $\tilde{u}_d = \gamma_d(\hat{\delta}_d^{DIR} - \mathbf{x}'_d\tilde{\boldsymbol{\beta}})$ en el BLUP bajo el modelo FH, obtenemos

$$\tilde{\delta}_d^{FH} = \gamma_d\hat{\delta}_d^{DIR} + (1-\gamma_d)\mathbf{x}'_d\tilde{\boldsymbol{\beta}}$$

BLUP/EBLUP basado en el modelo Fay-Herriot

- Note que

$$\tilde{\delta}_d^{FH} = \gamma_d \hat{\delta}_d^{DIR} + (1 - \gamma_d)\mathbf{x}_d' \tilde{\boldsymbol{\beta}},$$

  es una combinación lineal convexa del estimador directo y del estimador sintético de regresión a nivel de área.
- Si la varianza muestral ψ_d es pequeña comparada con la heterogeneidad no explicada σ_u^2, $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d)$ es cercano a uno.
- Entonces, cuando el tamaño muestral del área es grande (ψ_d pequeña), el BLUP $\tilde{\delta}_d^{FH}$ se acerca al estimador directo.
- Por tanto, no necesitamos saber si el área es pequeña para usar este estimador.

BLUP/EBLUP basado en el modelo Fay-Herriot

- Habitualmente, no sabemos el verdadero valor de σ_u^2 de los efectos aleatorios u_d.
- Sea $\hat{\sigma}_u^2$ un estimador consistente para σ_u^2.
- Entonces, obtenemos el BLUP empírico (*empirical BLUP, EBLUP*) de δ_d,

$$\hat{\delta}_d^{FH} = \hat{\gamma}_d \hat{\delta}_d^{DIR} + (1 - \hat{\gamma}_d)\mathbf{x}_d'\hat{\boldsymbol{\beta}}$$

donde

$$\hat{\gamma}_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \psi_d)$$

y

$$\hat{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D}\left(\frac{1}{\hat{\sigma}_u^2 + \psi_d}\right)\mathbf{x}_d\mathbf{x}_d'\right)^{-1}\sum_{d=1}^{D}\left(\frac{1}{\hat{\sigma}_u^2 + \psi_d}\right)\mathbf{x}_d\hat{\delta}_d^{DIR}$$

BLUP/EBLUP basado en el modelo Fay-Herriot

- En un área no muestreada, la varianza del estimador directo ψ_d tiende a infinito y γ_d tiende a cero
- Tomando el valor límite $\gamma_d = 0$, obtenemos el estimador sintético de regresión,

$$\hat{\delta}_d^{FH} = \mathbf{x}_d' \hat{\beta}$$

BLUP/EBLUP basado en el modelo Fay-Herriot: Sesgo y ECM

- Si se conocen los parámetros del modelo $\boldsymbol{\beta}$ y σ_u^2, el ECM del BLUP $\tilde{\delta}_d^{FH}$ viene dado por

$$\text{MSE}(\tilde{\delta}_d^{FH}) = \gamma_d^2 \psi_d \leq \psi_d = \text{var}_\pi(\hat{\delta}_d^{DIR} | \delta_d)$$

- En ese caso, el BLUP bajo el modelo FH no puede ser menos eficiente que el estimador directo.
- En la práctica, no se dispone de estos valores, y el ECM crece.
- Sin embargo, este crecimiento tiende a cero con un aumento en el número de áreas D.

BLUP/EBLUP basado en el modelo Fay-Herriot: Sesgo y ECM

- Este estimador usa los pesos del diseño muestral a través del estimador directo.
- Entonces, es consistente bajo el diseño muestral cuando n_d crece.
- Su sesgo absoluto bajo el diseño muestral viene dado por:

$$(1-\gamma_d)|\delta_d - \mathbf{x}_d'\boldsymbol{\beta}| \leq |\delta_d - \mathbf{x}_d'\boldsymbol{\beta}|$$

BLUP/EBLUP basado en el modelo Fay-Herriot: Sesgo y ECM

- Un estimador insesgado de segundo orden del ECM (llamado el estimador Prasad-Rao) viene dado por

$$\text{mse}_{PR}(\hat{\delta}_d^{FH}) = g_{d1}(\hat{\sigma}_u^2) + g_{d2}(\hat{\sigma}_u^2) + 2g_{d3}(\hat{\sigma}_u^2)$$

donde

$$g_{1d}(\sigma_u^2) = \gamma_d \psi_d$$

BLUP/EBLUP basado en el modelo Fay-Herriot: Sesgo y ECM

- Las otras ecuaciones incluidas en el estimador vienen dadas por

$$g_{2d}(\sigma_u^2) = (1-\gamma_d)^2 \mathbf{x}_d' \left(\sum_{d=1}^{D} (\sigma_u^2 + \psi_d) \mathbf{x}_d \mathbf{x}_d' \right)^{-1} \mathbf{x}_d,$$

y

$$g_{3d}(\sigma_u^2) = (1-\gamma_d)^2 (\sigma_u^2 + \psi_d)^{-1} \overline{\text{var}}(\hat{\sigma}_u^2),$$

donde

$$\overline{\text{var}}(\hat{\sigma}_u^2) = \mathcal{I}^{-1}(\sigma_u^2) = 2 \left\{ \sum_{d=1}^{D} (\sigma_u^2 + \psi_d)^{-2} \right\}^{-1}$$

para un estimador REML y $\mathcal{I}$ es la información Fisher

- $g_{2d}(\sigma_u^2)$ y $g_{3d}(\sigma_u^2)$ tienden a *cero* cuando el número de áreas D suficientemente grande.

BLUP/EBLUP de $F_{\alpha d}$ basado en el modelo Fay-Herriot

- También podemos escribir el modelo FH en términos del estimador $\hat{F}_{\alpha d}^{DIR}$, donde

$$\hat{F}_{\alpha d}^{DIR} = \mathbf{x}_d'\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \ldots, D$$

- Como ya se ha mencionado, u_d es el efecto aleatorio del grupo d y e_d es la diferencia que viene de $\hat{F}_{\alpha d}^{DIR} = F_{\alpha d} + e_d$.
- Por tanto, el BLUP de $F_{\alpha d} = \mathbf{x}_d'\boldsymbol{\beta} + u_d$ sería

$$\tilde{F}_{\alpha d}^{FH} = \mathbf{x}_d'\tilde{\boldsymbol{\beta}} + \tilde{u}_d$$

BLUP/EBLUP $F_{\alpha d}$ basado en el modelo Fay-Herriot

- En el BLUP $\tilde{F}_{\alpha d}^{FH} = \mathbf{x}_d'\tilde{\boldsymbol{\beta}} + \tilde{u}_d$,

$$\tilde{u}_d = \gamma_d(\hat{F}_{\alpha d}^{DIR} - \mathbf{x}_d'\tilde{\boldsymbol{\beta}})$$

y

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \gamma_d \mathbf{x}_d \mathbf{x}_d'\right)^{-1} \sum_{d=1}^{D} \gamma_d \mathbf{x}_d \hat{F}_{\alpha d}^{DIR}$$

Resumen del estimador FH

- Indicadores objetivos: Parámetros generales.
- Requerimientos de datos:
  - Datos agregados, e.g.medias poblacionales de las p covariables para las áreas $d = 1, \ldots, D$

Resumen del estimador FH

- Ventajas:
  - Suele mejorar la eficiencia del estimador directo.
  - Incorpora heterogeneidad no explicada entre las áreas.
  - Es un estimador compuesto que tiende al estimador directo cuando el tamaño muestral es suficientemente grande.
  - Usan datos agregados, por lo que no se ve excesivamente afectado por datos atípicos aislados.

Resumen del estimador FH

- Ventajas:
  - Con datos agregados, hay un beneficio de confidencialidad de los microdatos.
  - Si para un área d, el peso dado al estimador directo $\hat{\delta}_d^{DIR}$ es positivo, se usan los pesos muestrales w_{di} a través del estimador directo. Como consecuencia, es consistente bajo el diseño.

Resumen del estimador FH

- Ventajas:
  - Para estimadores directos lineales, se aplica el Teorema Central del Límite para las áreas con tamaño muestral suficiente. Por tanto, el modelo siempre tendrá una mínima bondad de ajuste para áreas de tamaño muestral suficiente.
  - El estimador Prasad-Rao que vimos para el ECM es eficiente e insesgado bajo el diseño cuando se promedia a lo largo de muchas áreas.
  - Se puede estimar en áreas no muestreadas.

Resumen del estimador FH

- Desventajas:
  - Se basan en un modelo lineal y es necesario analizar dicho modelo.
  - Las varianzas muestrales de los estimadores directos, ψ_d, se asumen conocidas, pero en la práctica es necesario estimarlas. Esto puede tener el mismo problema de áreas pequeñas. El estimador del ECM no incluye el error asociado a ψ_d.
  - El número de observaciones es el número de áreas, lo que suele ser menor que el número de individuos. Esto reduce la eficiencia.

Resumen del estimador FH

- Desventajas:
  - A la hora de estimar indicadores que dependen de una variable común (e.g.$F_{\alpha d}$), se requiere una modelización y búsqueda de variables auxiliares para cada uno de los indicadores por separado.
  - El estimador Prasad-Rao de ECM es correcto bajo normalidad de e_d y u_d, no es insesgado bajo el diseño para el ECM bajo el diseño en un área concreta.

Resumen del estimador FH

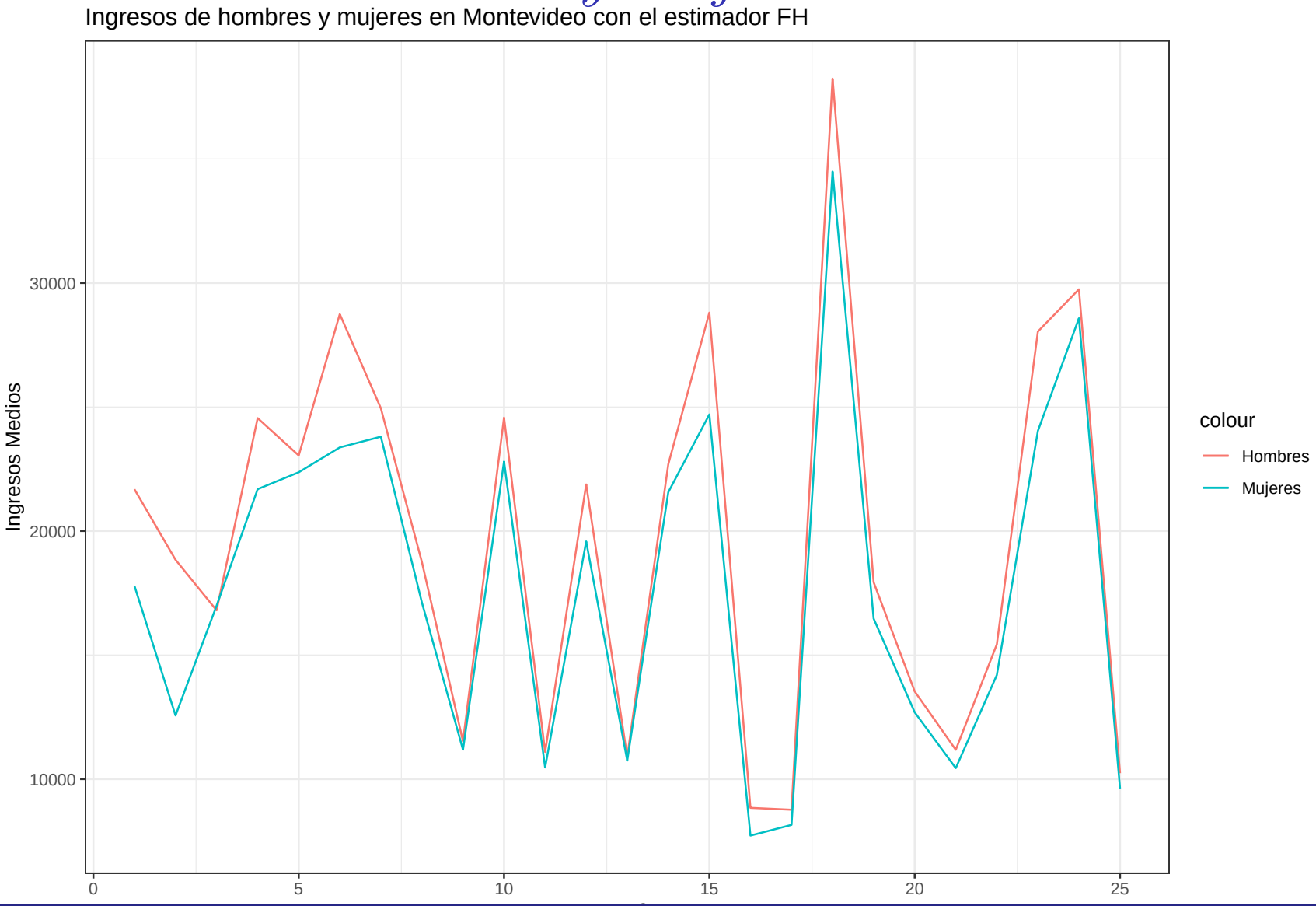
- Desventajas:
  - Una vez se ha ajustado el modelo a nivel de área, los estimadores $\hat{\delta}_d^{FH}$ no se pueden desagregar para subáreas dentro de las áreas.
  - Requiere un reajuste para verificar la propiedad de "benchmarking".

Resultados: Estimación de ingreso medio en sectores de Montevideo

Estimador FH: Hombres y Mujeres en Montevideo

sec2	ntotal	Hombres	Mujeres
2	121	18839	12564
1	167	21682	17790
3	186	16801	16987
4	319	24552	21687
6	320	28744	23370
5	495	23046	22366
21	3165	11180	10441
13	3556	10892	10744
18	3950	38237	34490
11	3963	11092	10467
17	4373	8763	8154
10	6302	24574	22802

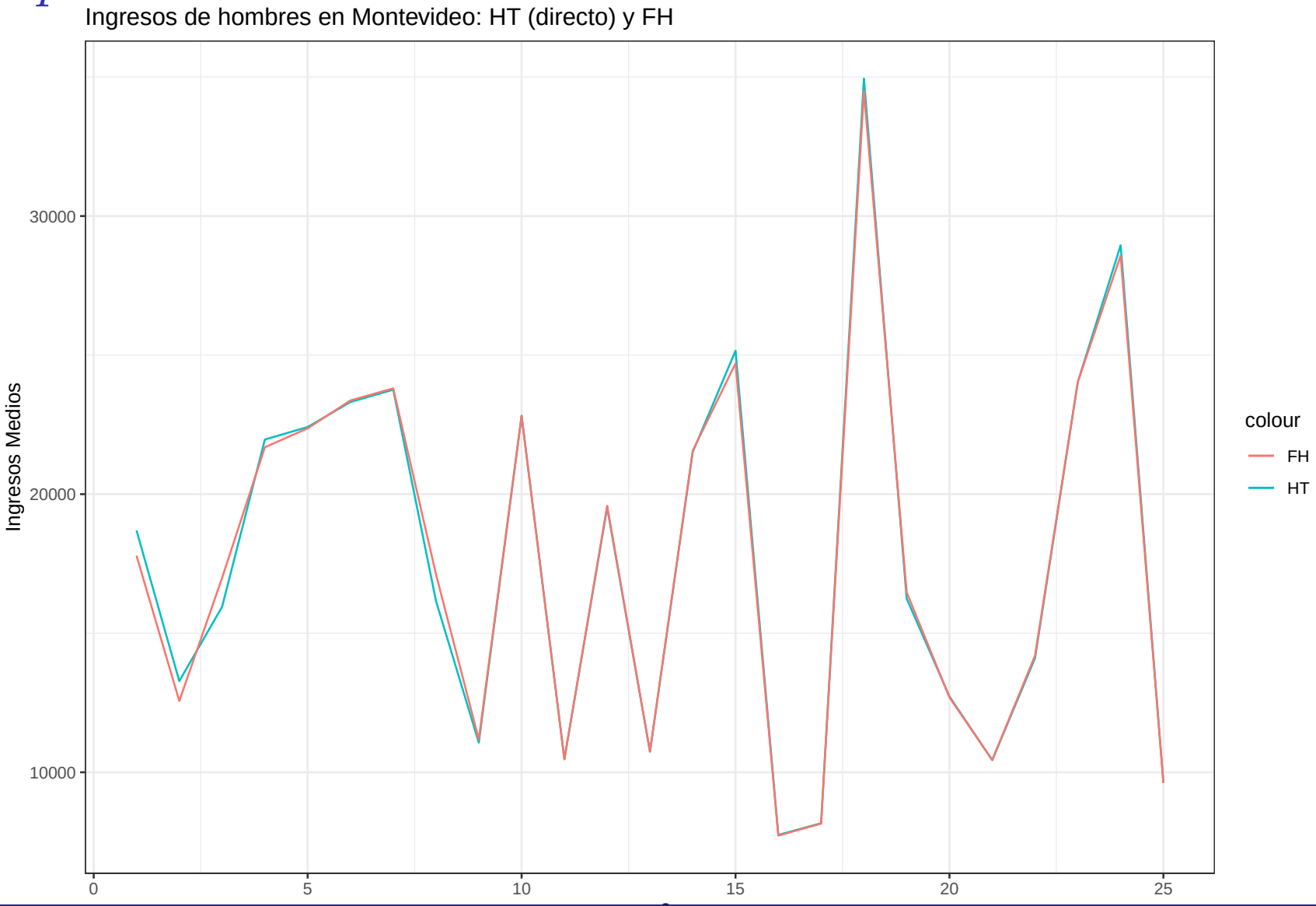
Estimador FH: Hombres y Mujeres en Montevideo



Comparando los estimadores: Hombres

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

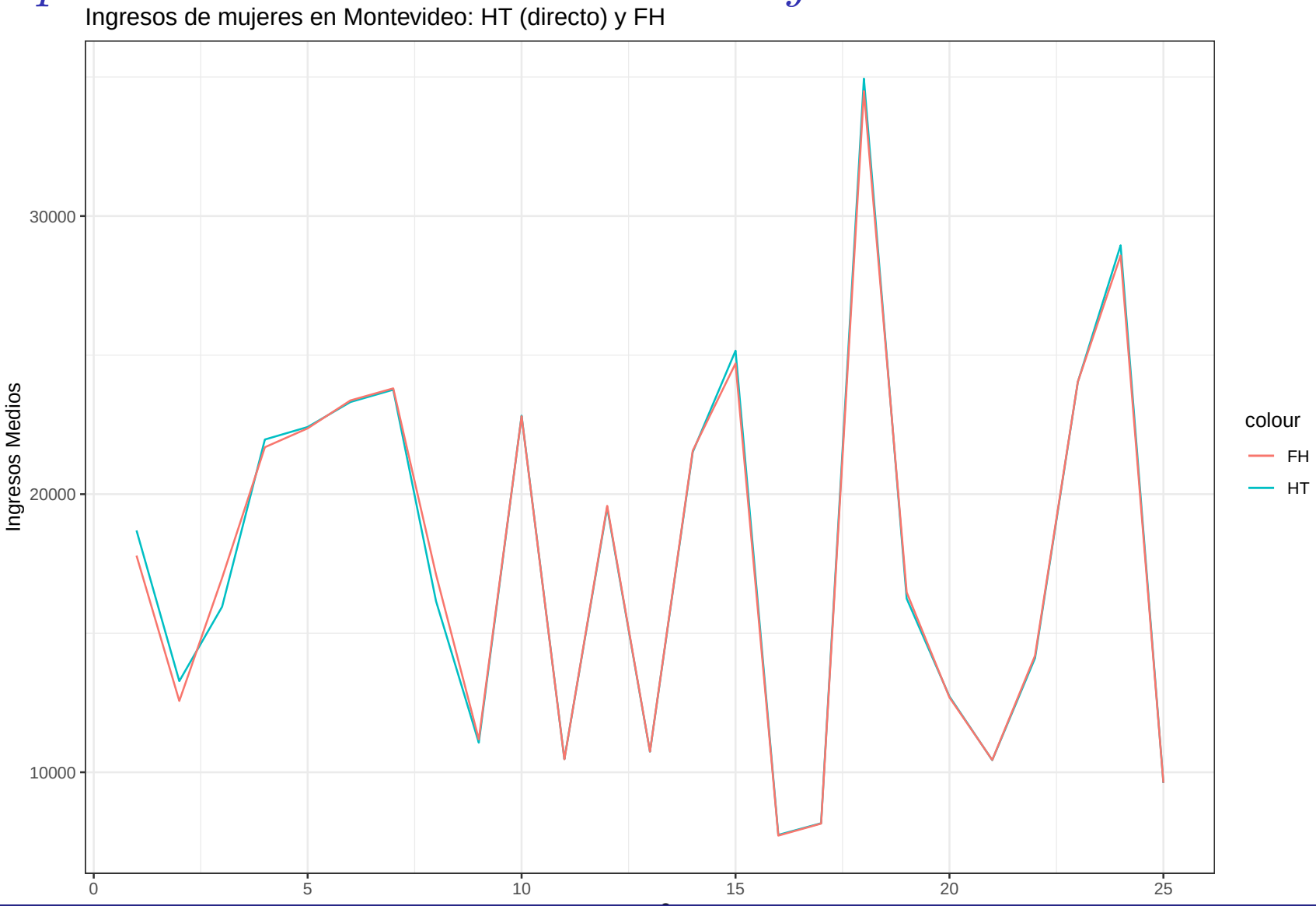
Comparando los estimadores: Hombres



Comparando los estimadores: Mujeres

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

Comparando los estimadores: Mujeres



¡Gracias!

¡Gracias!