Curso Internacional de Desagregación de Estimaciones en Áreas Pequeñas usando R

Métodos indirectos básicos

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

- Como ya se ha mencionado, los métodos directos utilizan solamente información del área para el indicador que se desea estimar.
- Los *Métodos indirectos* para indicadores en un área usan información de otras áreas, asumiendo algún tipo de homogeneidad entre ellas.
- Esto conlleva un aumento de la eficiencia de los estimadores.

Introducción

- Un tipo de estimadores indirectos se llama *estimadores sintéticos*.
- Estos estimadores consideran que las áreas son homogéneas, es decir que poseen parámetros comunes.
- Esta hipótesis es poco probable en la práctica y por consiguiente, los estimadores pueden tener sesgo grande.

Estimador post-estratificado sintético

Estimador post-estratificado sintético

- Este estimador no es muy utilizado en aplicaciones reales.
- Para este estimador, se dispone de una variable relacionada con la variable Y_{di} que tiene J categóricas posibles.
- La población U es dividida en J grupos $U^1, \ldots, U^J$ con tamaños poblacionales $N^1, \ldots, N^J$

Estimador post-estratificado sintético

- El área d, U_d también es dividida en J grupos, llamados post-estratos, $U_d^1, \ldots, U_d^J$ de tamaño $N_d^1, \ldots, N_d^J$.
- Tienen medias $\bar{Y}_d^1, \ldots, \bar{Y}_d^J$, donde $\bar{Y}_d^j = \sum_{i \in U_d^j} Y_{di}/N_d^j$, $j = 1, \ldots, J$.
- Dado que las medias son indicadores aditivos, podemos descomponerlos en sumas para los J estratos, de la forma

$$\bar{Y}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} Y_{di} = \frac{1}{N_d} \sum_{j=1}^{J} N_d^j \, \bar{Y}_d^j$$

Estimador post-estratificado sintético

- Se asume que los individuos en cada post-estrato se comportan de la misma manera.
- Es decir,

$$\bar{Y}_d^j = \bar{Y}^j, \quad j = 1, \ldots, J,$$

con $\bar{Y}^j = \sum_{i \in U^j} Y_i / N^j$ siendo la media del estrato j.

Estimador post-estratificado sintético

- Con esta homogeneidad, podemos escribir $\overline{Y}_d$ así:

$$\overline{Y}_d = \frac{1}{N_d} \sum_{j=1}^{J} N_d^j \, \overline{Y}^j$$

- Por tanto, se estima la media de un área estimando las medias de los post-estratos.
- El estimador post-estratificado sintético (PS-SYN) de $\bar{Y}_d$ se obtiene utilizando estimadores de Hájek para cada estrato. Es decir,

$$\hat{\bar{Y}}_d^{PS-SYN} = \frac{1}{N_d} \sum_{j=1}^{J} N_d^j \, \hat{\bar{Y}}^{j,HA}$$

Estimador post-estratificado sintético

- Se supone que el número de estratos J es pequeño y que los grupos tienen muestras suficientes.
- Por eso, la varianza del estimador $\hat{\bar{Y}}^{j,HA}$ es pequeña.
- Dado que estimamos $\bar{Y}_d$ usando el estimador de Hájek para los estratos, el estimador PS-SYN también debiese tener una varianza pequeña.
- Como la hipótesis de homogeneidad entre estratos es poco probable, es mejor usar el error cuadrático medio.

Estimador post-estratificado sintético

- Es posible usar el estimador PS-SYN para un estimador FGT.
- Todavía se usaría la hipótesis que el indicador es igual dentro de los estratos, es decir

$$F^j_{\alpha d} = F^j_\alpha, \quad j = 1, \ldots, J$$

donde F^j_α es el indicador FGT en estrato J.

Resumen del estimador PS-SYN

- Indicadores objetivos: Medias/Totales de la variable de interés.
- Requerimientos de datos:
  - Pesos muestrales w_{di} para los individuos en la muestra.
  - Tamaño poblacional N_d y los tamaños poblacionales de las intersecciones (post-estrato), $N_d^j, j = 1, \ldots, J$
  - Una variable cualitativa (o varias) relacionada a la variable de interés y observada en la misma encuesta.

Resumen del estimador PS-SYN

- Ventajas:
  - Si los estratos tienen suficientes observaciones en la muestra, la varianza puede disminuir en comparación con los estimadores directos.
- Desventajas:
  - La hipótesis de homogeneidad para las variables Y_{di} es poco probable. Si esto no se verifica, el estimador puede tener un sesgo considerable.
  - Por eso, es difícil encontrar un estimador del ECM bajo el diseño que sea estable.

Estimador sintético de regresión a nivel de área (REG1-SYN)

Estimador sintético de regresión a nivel de área

- Los estimadores sintéticos de regresión asumen un modelo de regresión lineal utilizando información auxiliar.
- Este estimador (estimador REG1-SYN) se usa cuando solo se dispone de información auxiliar a nivel de área.
- Llamamos $\mathbf{x}_d$ al vector de p variables auxiliares y se asume que el indicador que queremos estimar, δ_d (e.g.la media del área), varía respecto a estos datos $\mathbf{x}_d$ de forma constante para todas las áreas.

Estimador sintético de regresión a nivel de área

- Los valores verdaderos del indicador en las áreas no están disponibles (son los parámetros objetivo).
- En lugar de estos, se consideran estimadores directos, $\hat{\delta}_d, d = 1, \ldots, D$
- El modelo se asume entonces,

$$\hat{\delta}_d = \mathbf{x}_d'\boldsymbol{\alpha} + \varepsilon_d, \quad d = 1, \ldots, D$$

Estimador sintético de regresión a nivel de área

- En nuestro modelo, $\hat{\delta}_d = \mathbf{x}'_d\boldsymbol{\alpha} + \varepsilon_d$, x_d son valores poblacionales y por tanto tienen varianza cero.
- ε_d tiene esperanza cero y varianza ψ_d conocida igual a $var(\hat{\delta}_d), d = 1, \ldots, D$.
- En la práctica, estas varianzas se estiman usando microdatos.

Estimador sintético de regresión a nivel de área

- Podemos escribir el estimador sintético de regresión a nivel de área como

$$\hat{\delta}_d^{REG1-SYN} = \mathbf{x}'_d \hat{\boldsymbol{\alpha}}$$

- Donde

$$\hat{\boldsymbol{\alpha}} = (\sum_{d=1}^{D} \psi_d^{-1} \mathbf{x}_d \mathbf{x}'_d)^{-1} \sum_{d=1}^{D} \psi_d^{-1} \mathbf{x}_d \hat{\delta}_d$$

Estimador sintético de regresión a nivel de área

- Para $\boldsymbol{\alpha}$, el sesgo bajo el diseño de $\hat{\delta}_d^{REG1-SYN}$ viene dado por $\mathbf{x}'_d\boldsymbol{\alpha} - \delta_d$.
- Como este sesgo no depende del tamaño muestral del área n_d, no disminuye al aumentar el tamaño muestral del área.

Estimador sintético de regresión a nivel de área

- Si $\delta_d = F_{\alpha d}$, el modelo a nivel de área viene dado por, $\hat{F}_{\alpha d} = \mathbf{x}'_d\boldsymbol{\alpha} + \varepsilon_d, \quad d = 1, \ldots, D$ y se estima de la misma manera.
- Observe que con el estimador sintético de regresión, si sabemos los parámetros $\boldsymbol{\alpha}$, el estimador REG1-SYN sería $\mathbf{x}'_d\boldsymbol{\alpha}$, es decir, no se estarían utilizando los datos de la variable de interés.

Resumen del estimador indirecto REG1-SYN

- Indicadores objetivos: Parámetros generales (no solo la media o totales)
- Requerimientos de datos:
  - Datos agregados (e.g.medias poblacionales) de las p variables auxiliares en las áreas $d = 1, \ldots, D$, $\mathbf{x}_d$.

Resumen del estimador indirecto REG1-SYN

- Ventajas:
  - Se puede disminuir la varianza considerablemente en comparación con los estimadores directos.
  - Se puede estimar en áreas *no muestreadas*.

Resumen del estimador indirecto REG1-SYN

- Desventajas:
  - El modelo de regresión sintético no representa los casos en los que no se dispone de las variables auxiliares.
  - No se usarían los datos de la variable de interés para un área si ya se conoce el modelo.
  - No tiende al estimador directo cuando aumenta el tamaño muestral.

Resumen del estimador indirecto REG1-SYN

- Desventajas:
  - No se conocen estimadores del ECM que sean estables bajo el diseño muestral y distintos para las áreas al mismo tiempo.
  - Es importante verificar el modelo (e.g.con residuos) porque no considera efectos de las áreas.
  - Requiere un reajuste para verificar la propiedad "benchmarking" de que la suma de los totales estimados en las áreas de una región coincida con el estimador directo para dicha región.

Estimador sintético de regresión a nivel de individuo (REG2-SYN)

Estimador sintético de regresión a nivel de individuo

- Ahora, imaginemos que tenemos datos a nivel de individuo (*microdatos*) de las p covariables de la encuesta, $\mathbf{x}_{di}$, $i \in s_d$, $d = 1, \ldots, D$.
- Se puede obtener por tanto un modelo lineal a nivel de individuo para Y_{di}.
- Llamamos $\mathbf{y}_d = (Y_{d1}, \ldots, Y_{dN_d})$ a la variable de la encuesta en cuestión para el área d.
- Digamos que el indicador que queremos estimar es una función de $\mathbf{y}_d$, es decir $\delta_d = \delta_d(\mathbf{y}_d)$.

Estimador sintético de regresión a nivel de individuo

- El modelo sintético considera que las variables $\mathbf{y}_d$ siguen el modelo,

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

- Los errores ε_{di} tienen una esperanza cero y varianza $\sigma^2 k_{di}^2$ para representar posible heteroscedasticidad.
- Podemos estimar β de la siguiente forma

$$\hat{\beta} = (\sum_{d=1}^{D} \sum_{i \in s_d} a_{di} \mathbf{x}_{di} \mathbf{x}'_{di})^{-1} \sum_{d=1}^{D} \sum_{i \in s_d} a_{di} \mathbf{x}_{di} Y_{di},$$

siendo $a_{di} = k_{di}^{-2}$.

Estimador sintético de regresión a nivel de individuo

- El vector de predicciones para el área d es entonces $\hat{\mathbf{y}}_d = (\hat{Y}_{d1}, \ldots, \hat{Y}_{dN_d})'$ donde $\hat{Y}_{di} = \mathbf{x}'_{di}\hat{\beta}$, $i = 1, \ldots, N_d$
- El estimador de regresión sintético a nivel de individuo, REG2-SYN, de δ_d viene dado por

$$\hat{\delta}_d^{REG2-SYN} = \delta_d(\hat{\mathbf{y}}_d)$$

Estimador sintético de regresión a nivel de individuo

- Por ejemplo, para la media del área d, $\delta_d = \overline{Y}_d$, si $\overline{\mathbf{X}}_d$ es el vector de medias poblacionales de la p variables auxiliares, $\hat{\overline{Y}}_d^{REG2-SYN}$ sería

$$\hat{\overline{Y}}_d^{REG2-SYN} = \overline{\mathbf{X}}_d' \hat{\beta}$$

- Se obtiene el estimador para un área no muestrada de la misma forma.
- Para β conocido, el sesgo bajo el diseño de la media es $\overline{\mathbf{X}}_d' \beta - \overline{Y}_d$, lo que no depende del tamaño muestral del área n_d.

Estimador sintético de regresión a nivel de individuo

- Si queremos estimar un indicador FGT, el modelo sería

$$F_{\alpha,di} = \mathbf{x}'_{di}\beta + \varepsilon_{di}, \quad i = 1, \ldots, N_d, \; d = 1, \ldots, D$$

- No obstante, es difícil encontrar variables relacionadas linealmente a $F_{\alpha,di}$.
- Para evitar esto, a menudo se usa la variable para medir el poder adquisitivo E_{di} con otra transformación.
- Por ejemplo, $Y_{di} = \log(E_{di} + c)$, lo que tendría una distribución más simétrica que la de E_{di}.

Resumen del estimador indirecto REG2-SYN

- Indicadores objetivos: Parámetros generales.
- Requerimientos de datos:
  - Observaciones muestrales de las p covariables relacionadas con el indicador de interés a nivel de individuo.
  - Para indicadores de totales o medias, se necesitan totales o medias poblaciones de las p variables auxiliares en las áreas, $\bar{\mathbf{X}}_d$, $d = 1, \ldots, D$

Resumen del estimador indirecto REG2-SYN

- Ventajas:
  - La varianza puede ser reducida en comparación con estimadores directos y modelos a nivel de área.
  - Se puede estimar en áreas no muestreadas.

Resumen del estimador indirecto REG2-SYN

- Desventajas:
  - El modelo no representa los casos en los que no se dispone de todas las variables auxiliares.
  - Es importante comprobar si existe efecto del área porque el modelo no considera esto.
  - Si se conoce exactamente el modelo, no se usaría la variable de interés para esa área.

Resumen del estimador indirecto REG2-SYN

- Desventajas:
  - No converge al estimador directo cuando aumenta el tamaño muestral.
  - No se conocen estimadores del ECM que sean estables bajo el diseño muestral y distintos al mismo tiempo para las áreas.
  - Requiere un reajuste para verificar la propiedad de "benchmarking".

Estimadores compuestos

Estimadores compuestos

- Como se ha mencionado en las otras secciones, los estimadores directos son aproximadamente insesgados, pero pueden tener varianza grande.
- Los estimadores indirectos, sin embargo, tienen una varianza pequeñas pero pueden ser sesgados bajo el diseño muestral.
- Los estimadores compuestos se usan con el objetivo de reducir la varianza del estimador directo a cambio de un aumento de sesgo inducido por el estimador indirecto.

Estimadores compuestos

- Un estimador compuesto para $\bar{Y}_d$ tiene la forma

$$\hat{\bar{Y}}_d^C = \phi_d \hat{\bar{Y}}_d^{DIR} + (1 - \phi_d) \hat{\bar{Y}}_d^{SYN}, \quad 0 \leq \phi_d \leq 1$$

- El peso ϕ_d puede ser establecido al minimizar una aproximación del ECM bajo el diseño muestral, o fijándolo de una manera arbitraria.

Estimadores compuestos

- Drew, Singh y Choudhry (1982) proponen un peso ϕ_d que depende del tamaño muestral del área d (*sample size dependent*, SSD).
- Tomando un valor $\delta > 0$ predeterminado, el peso SSD tiene la forma

$$\phi_d = \begin{cases} 1 \text{ si } \hat{N}_d \geq \delta N_d \\ \hat{N}_d/(\delta N_d) \text{ si } \hat{N}_d < \delta N_d \end{cases}$$

Estimadores compuestos

- Se puede mostrar que

$$\text{MSE}_\pi(\hat{\bar{Y}}_d^C) \approx \phi_d^2 \text{var}_\pi(\hat{\bar{Y}}_d^{DIR}) + (1-\phi_d)^2 \text{MSE}_\pi(\hat{\bar{Y}}_d^{SYN})$$

- Minimizando este valor, obtenemos un peso óptimo estimado que viene dado por

$$\hat{\phi}^* = 1 - \sum_{\ell=1}^{D} \widehat{\text{var}}_\pi(\hat{\bar{Y}}_\ell^{DIR}) / \sum_{\ell=1}^{D} (\hat{\bar{Y}}_\ell^{SYN} - \hat{\bar{Y}}_\ell^{DIR})^2$$

Resumen de estimadores compuestos

- Indicadores objetivos: parámetros aditivos.
- Requerimientos de datos:
  - Pesos muestrales w_{di} para individuos en el área d para poder estimar $\hat{N}_d$.
  - Tamaño poblacional del área, N_d, si se usa un estimador de HT de la media o el estimador de Hájek del total.

Resumen de estimadores compuestos

- Ventajas:
  - Provee una forma de encontrar un equilibrio entre la varianza de estimadores directos y el sesgo de estimadores indirectos.
- Desventajas:
  - Para un área de tamaño muestral pequeño que no es inferior al tamaño muestral esperado, no se utiliza información de las otras áreas a través del estimador sintético. En ese caso, no se ganará eficiencia respecto del estimador directo considerado.
  - No se pueden calcular para áreas no muestreadas.

Resumen de estimadores compuestos

- Desventajas:
  - El peso que se da al estimador sintético no depende de lo bien explicada que esté la variable de interés por las covariables auxiliares.
  - No se pueden calcular para áreas no muestreadas.
  - No se conocen estimadores del ECM que sean estables bajo el diseño muestral y distintos para las áreas al mismo tiempo.
  - Requiere un reajuste para verificar la propiedad de "benchmarking".

Resultados: Estimación de ingreso medio en sectores de Montevideo

Estimador post-estratificado sintético: Hombres y Mujeres en Montevideo

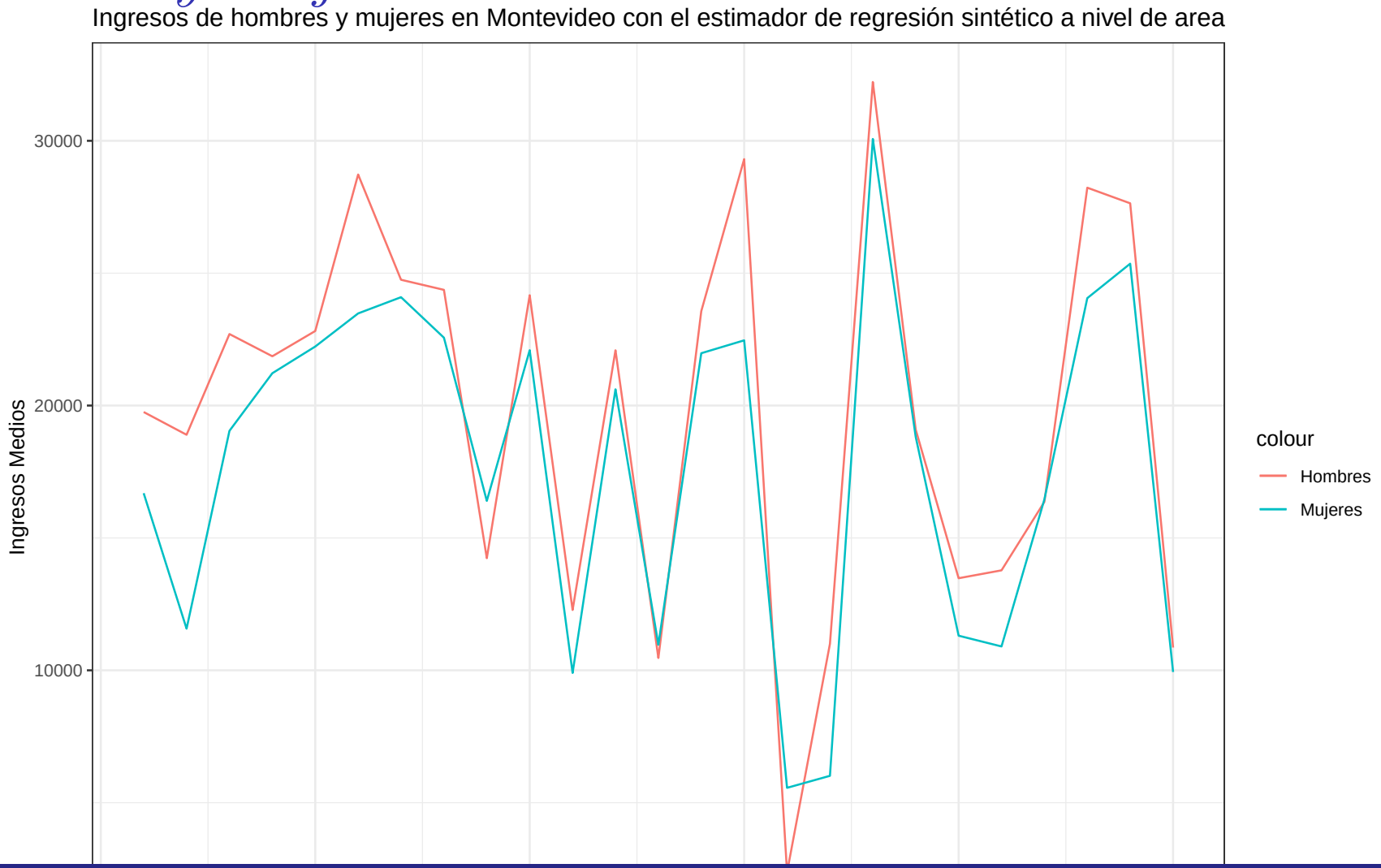
sec2	ntotal	Hombres	Mujeres
2	121	16430	15980
1	167	17093	16760
3	186	16838	16582
4	319	17071	16971
6	320	17186	16935
5	495	16935	16885
21	3165	16117	16017
13	3556	15635	15447
18	3950	17030	16945
11	3963	15799	15579
17	4373	15567	15361
10	6302	16555	16483

Estimador post-estratificado sintético: Hombres y Mujeres en Montevideo

Estimador de regresión sintético a nivel de área: Hombres y Mujeres en Montevideo

sec2	ntotal	Hombres	Mujeres
2	121	18896	11573
1	167	19756	16689
3	186	22701	19045
4	319	21862	21221
6	320	28726	23480
5	495	22816	22229
21	3165	13776	10901
13	3556	10466	10966
18	3950	32221	30070
11	3963	12278	9901
17	4373	11001	6015
10	6302	24167	22091

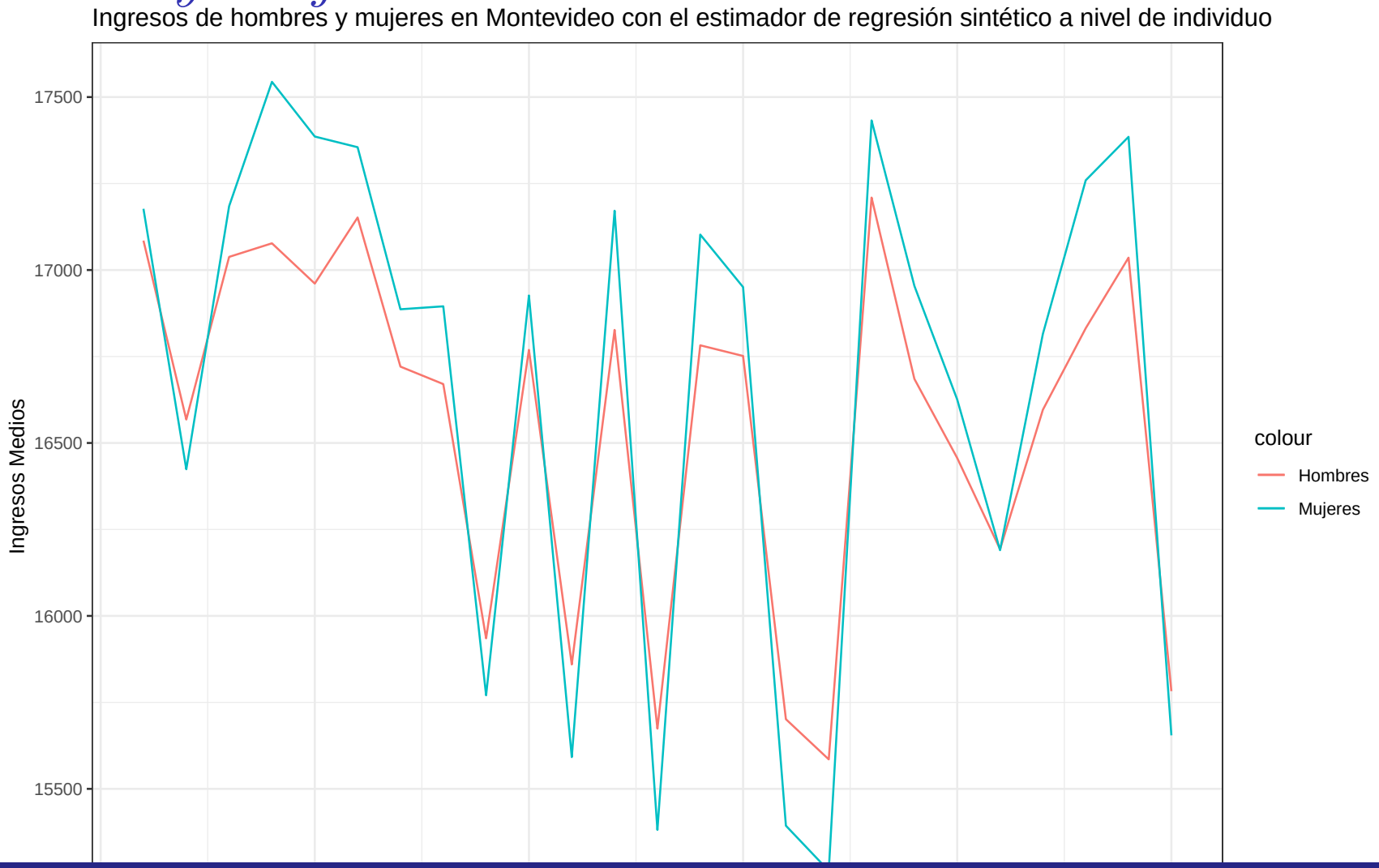
Estimador de regresión sintético a nivel de área: Hombres y Mujeres en Montevideo



Estimador de regresión sintético a nivel de individuo: Hombres y Mujeres en Montevideo

sec2	ntotal	Hombres	Mujeres
2	121	16568	16424
1	167	17085	17177
3	186	17038	17185
4	319	17077	17544
6	320	17152	17355
5	495	16961	17386
21	3165	16193	16190
13	3556	15674	15381
18	3950	17209	17433
11	3963	15860	15592
17	4373	15585	15263
10	6302	16769	16926

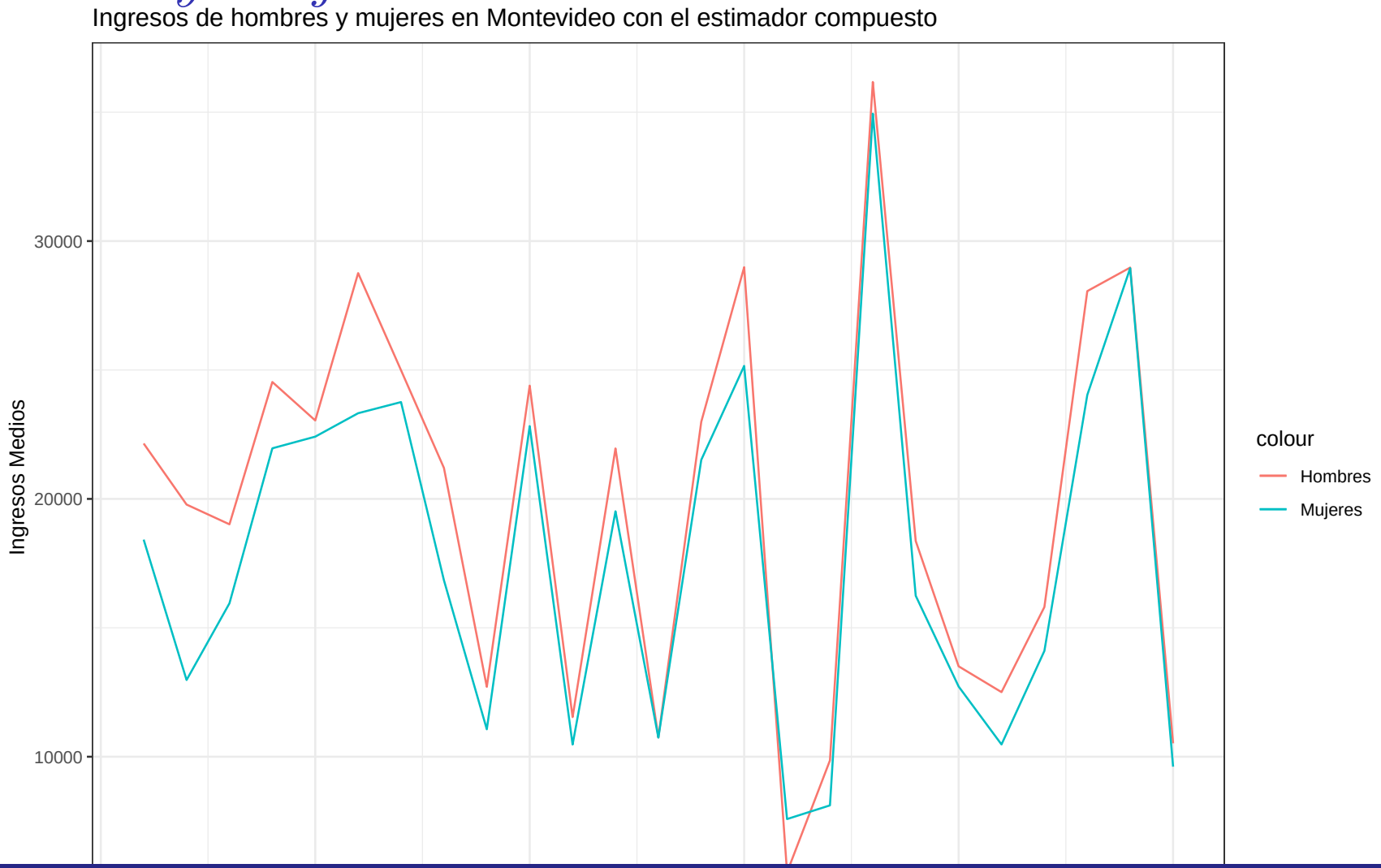
Estimador de regresión sintético a nivel de individuo: Hombres y Mujeres en Montevideo



Estimador compuesto: Hombres y Mujeres en Montevideo

sec2	ntotal	Hombres	Mujeres
2	121	19781	12977
1	167	22149	18420
3	186	19015	15951
4	319	24534	21962
6	320	28757	23324
5	495	23042	22414
21	3165	12506	10476
13	3556	10751	10742
18	3950	36170	34943
11	3963	11537	10473
17	4373	9857	8113
10	6302	24393	22823

Estimador de regresión sintético a nivel de individuo: Hombres y Mujeres en Montevideo

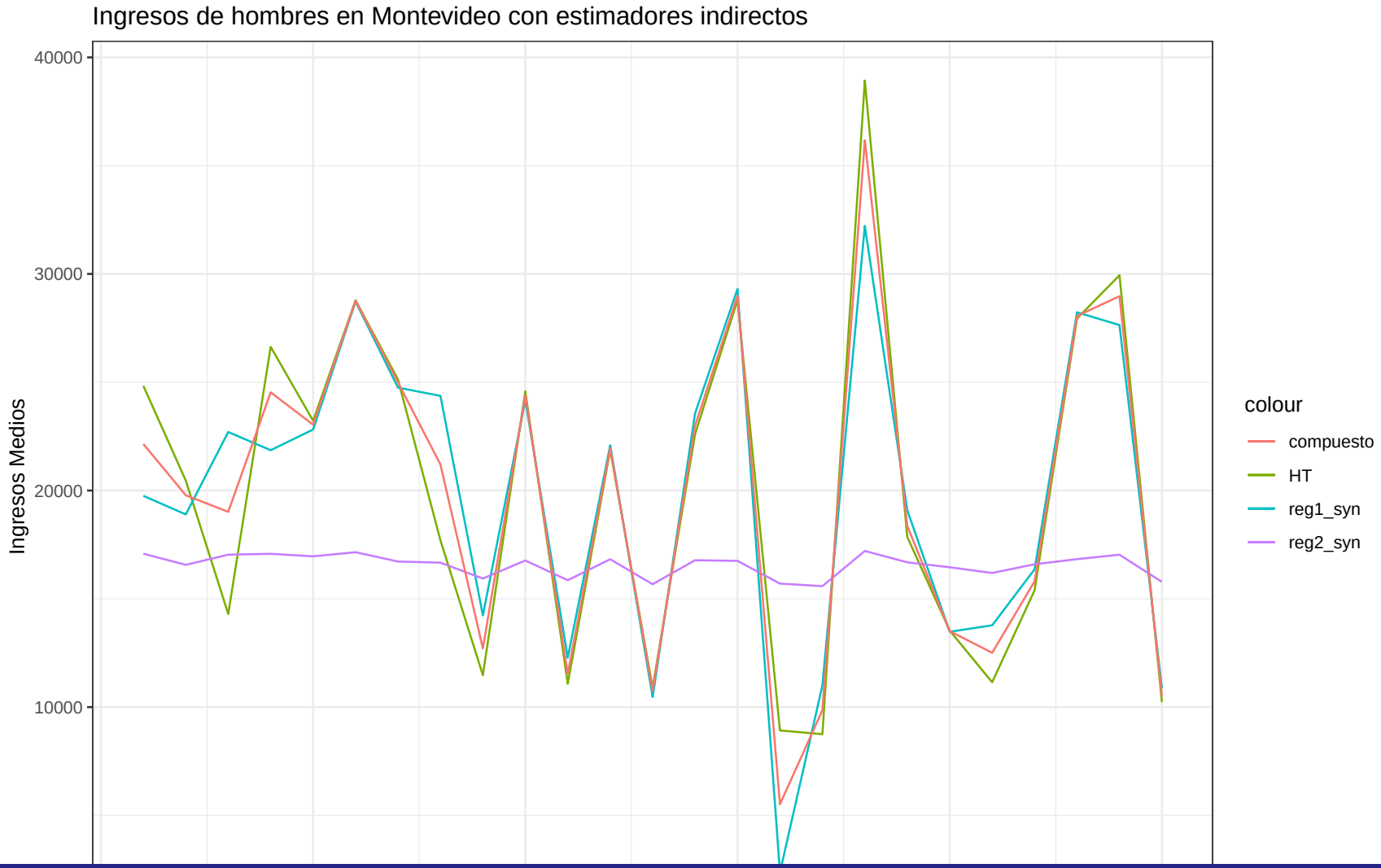


Comparando los estimadores: Hombres

sec2	ntotal	HT	PS_SYN	reg1_syn	reg2_syn	compuesto
2	121	20461	16430	18896	16568	19781
1	167	24837	17093	19756	17085	22149
3	186	14299	16838	22701	17038	19015
4	319	26635	17071	21862	17077	24534
6	320	28784	17186	28726	17152	28757
5	495	23223	16935	22816	16961	23042
21	3165	11148	16117	13776	16193	12506
13	3556	10897	15635	10466	15674	10751
18	3950	38932	17030	32221	17209	36170
11	3963	11080	15799	12278	15860	11537
17	4373	8750	15567	11001	15585	9857
10	6302	24576	16555	24167	16769	24393

Comparando los estimadores: Hombres

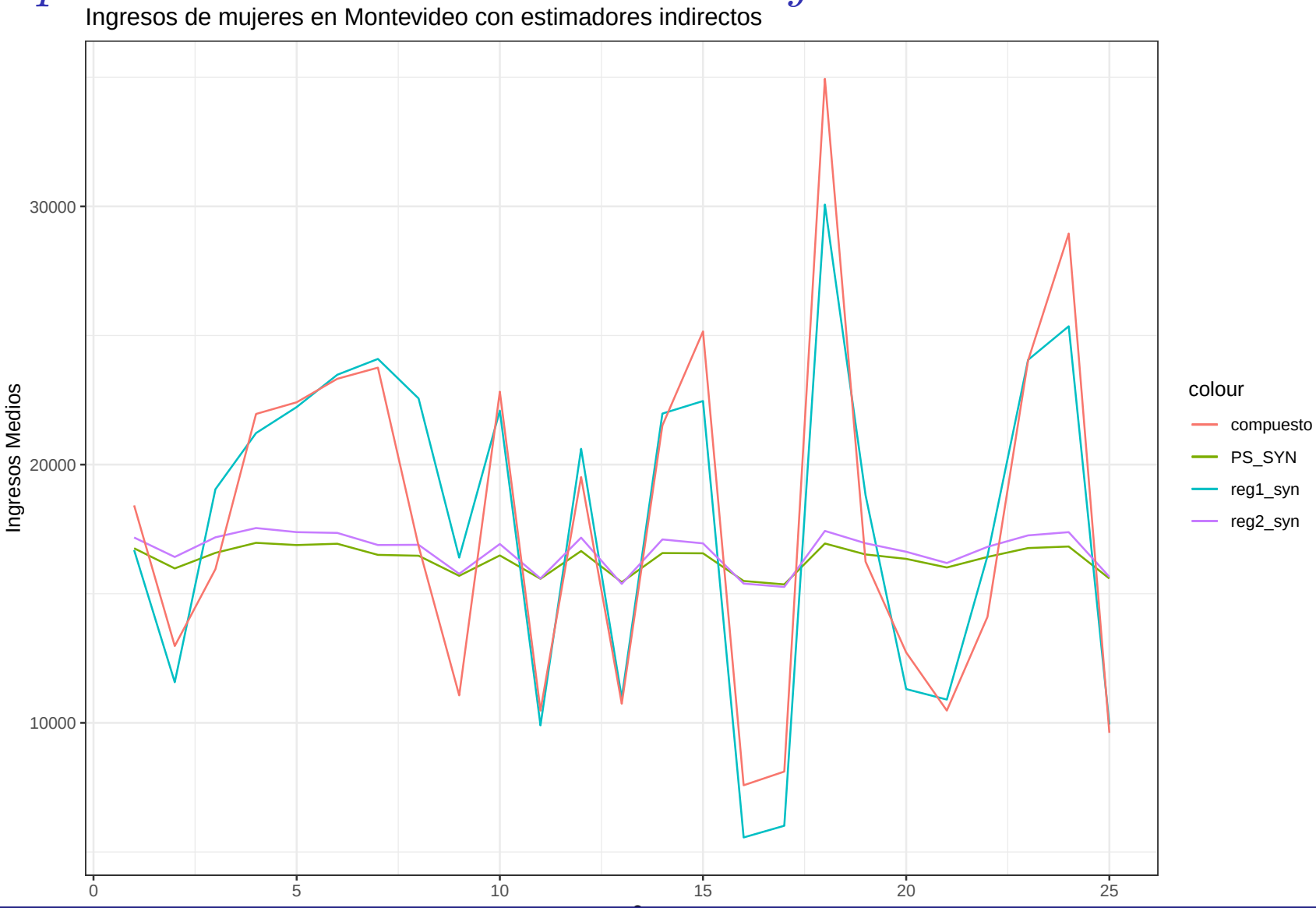
Comparando los estimadores: Hombres, usando HT para referencia



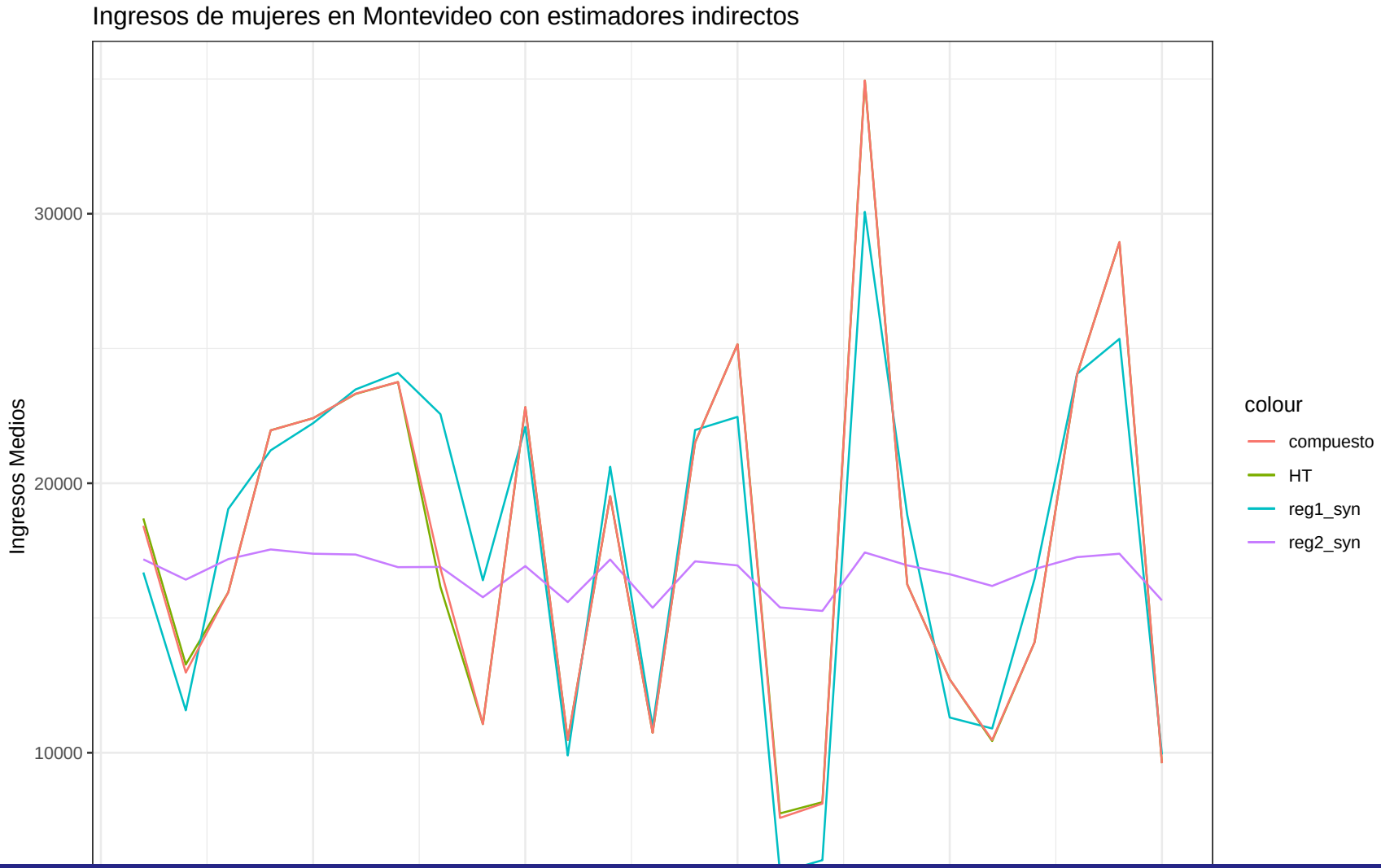
Comparando los estimadores: Mujeres

sec2	ntotal	HT	PS_SYN	reg1_syn	reg2_syn	compuesto
2	121	13277	15980	11573	16424	12977
1	167	18694	16760	16689	17177	18420
3	186	15951	16582	19045	17185	15951
4	319	21965	16971	21221	17544	21962
6	320	23314	16935	23480	17355	23324
5	495	22414	16885	22229	17386	22414
21	3165	10435	16017	10901	16190	10476
13	3556	10742	15447	10966	15381	10742
18	3950	34943	16945	30070	17433	34943
11	3963	10473	15579	9901	15592	10473
17	4373	8167	15361	6015	15263	8113
10	6302	22823	16483	22091	16926	22823

Comparando los estimadores: Mujeres



Comparando los estimadores: Mujeres, usando HT para referencia



¡Gracias!

¡Gracias!