

*Curso Internacional de Desagregación de
Estimaciones en Áreas Pequeñas usando R*

*Objetivos de Desarrollo Sostenible y limitaciones de las
encuestas*

División de Estadísticas
Comisión Económica para América Latina y el Caribe

2020

1 *Objetivos de Desarrollo Sostenible*

2 *Limitaciones de las encuestas*

3 *Uso de métodos SAE*

4 *Algunas aplicaciones actuales*

Objetivos de Desarrollo Sostenible

Agenda 2030



Figura 1: Los 17 ODS

Algunas metas del ODS1 (*Poner fin a la pobreza*)

- De aquí a 2030, erradicar para todas las personas y en todo el mundo la pobreza extrema (actualmente se considera que sufren pobreza extrema las personas que viven con menos de 1,25 dólares de los Estados Unidos al día).
- De aquí a 2030, reducir al menos a la mitad la proporción de hombres, mujeres y niños de todas las edades que viven en la pobreza en todas sus dimensiones con arreglo a las definiciones nacionales.

Algunas metas del ODS2 (Hambre cero)

- De aquí a 2030, poner fin al hambre y asegurar el acceso de todas las personas, en particular los pobres y las personas en situaciones de vulnerabilidad, incluidos los niños menores de 1 año, a una alimentación sana, nutritiva y suficiente durante todo el año.
 - Prevalencia de la subalimentación.
 - Prevalencia de la inseguridad alimentaria moderada o grave en la población, según la Escala de Experiencia de Inseguridad Alimentaria.

Algunas metas del ODS8 (Empleo decente)

- Promover políticas orientadas al desarrollo que apoyen las actividades productivas, la creación de puestos de trabajo decentes, el emprendimiento, la creatividad y la innovación, y fomentar la formalización y el crecimiento de las microempresas y las pequeñas y medianas empresas, incluso mediante el acceso a servicios financieros.
 - Proporción del empleo informal en el empleo no agrícola, desglosada por sexo.

Algunas metas del ODS8 (Empleo decente)

- De aquí a 2030, lograr el empleo pleno y productivo y el trabajo decente para todas las mujeres y los hombres, incluidos los jóvenes y las personas con discapacidad, así como la igualdad de remuneración por trabajo de igual valor.
 - Ingreso medio por hora de mujeres y hombres empleados, desglosado por ocupación, edad y personas con discapacidad.

Algunas metas del ODS8 (Empleo decente)

- De aquí a 2030, lograr el empleo pleno y productivo y el trabajo decente para todas las mujeres y los hombres, incluidos los jóvenes y las personas con discapacidad, así como la igualdad de remuneración por trabajo de igual valor.
 - Tasa de desempleo, desglosada por sexo, edad y personas con discapacidad.

Principio fundamental de la desagregación de datos

Los indicadores de los Objetivos de Desarrollo Sostenible deberán desglosarse, siempre que sea pertinente, por ingreso, sexo, edad, raza, etnicidad, estado migratorio, discapacidad y ubicación geográfica, u otras características, de conformidad con los Principios Fundamentales de las Estadísticas Oficiales.

Resolución de la Asamblea General - 68/261

No dejar a nadie atrás



Figura 2: Desagregación de indicadores en los 17 ODS

Principios fundamentales de las estadísticas oficiales

La confianza esencial del público en la integridad de los sistemas estadísticos oficiales y la credibilidad que este otorga a las estadísticas dependen en gran medida del respeto de los valores y principios fundamentales que son la base de toda sociedad que procura entenderse a sí misma y respetar los derechos de sus miembros y que, en este contexto, son cruciales la independencia profesional y la rendición de cuentas de los organismos de estadística.

Resolución de la Asamblea General - 68/261

No dejar a nadie atrás

Share of households per « Basic Unmet Needs » index, Colombia

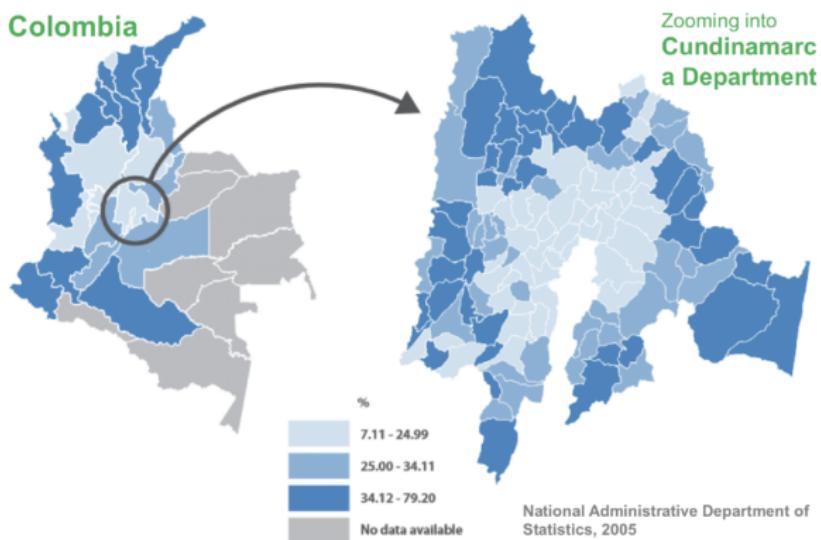


Figura3: Desagregación de un indicador en áreas pequeñas en Colombia.
Fuente: UNSD

Algunas metas del ODS17 (Alianzas para lograr los objetivos)

- De aquí a 2020, mejorar el apoyo a la creación de capacidad prestado a los países en desarrollo, incluidos los países menos adelantados y los pequeños Estados insulares en desarrollo, para aumentar significativamente la disponibilidad de datos oportunos, fiables y de gran calidad desglosados por ingresos, sexo, edad, raza, origen étnico, estatus migratorio, discapacidad, ubicación geográfica y otras características pertinentes en los contextos nacionales.

Limitaciones de las encuestas

¿Qué es el coeficiente de variación?

El coeficiente de variación es una medida de error relativo a un estimador, se define como:

$$cve(\hat{\theta}) = \frac{se(\hat{\theta})}{\hat{\theta}}$$

Muchas veces se expresa como un porcentaje, aunque no está acotado a la derecha, y por eso es conveniente a la hora de hablar de la precisión de una estadística que viene de una encuesta.

Uso del coeficiente de variación

Sarndal et. al.(2003) afirma que un estadístico puede expresar su opinión acerca de que *un coeficiente de variación del 2% es bueno, considerando las restricciones de la encuesta, mientras que un valor del coeficiente de variación de 9% puede ser considerado inaceptable.*

De esta forma, muchos institutos nacionales de estadística alrededor del mundo han considerado que las precisiones de las estadísticas resultantes de una encuesta estén supeditadas al comportamiento de su coeficiente de variación.

Alertas sobre el coeficiente de variación

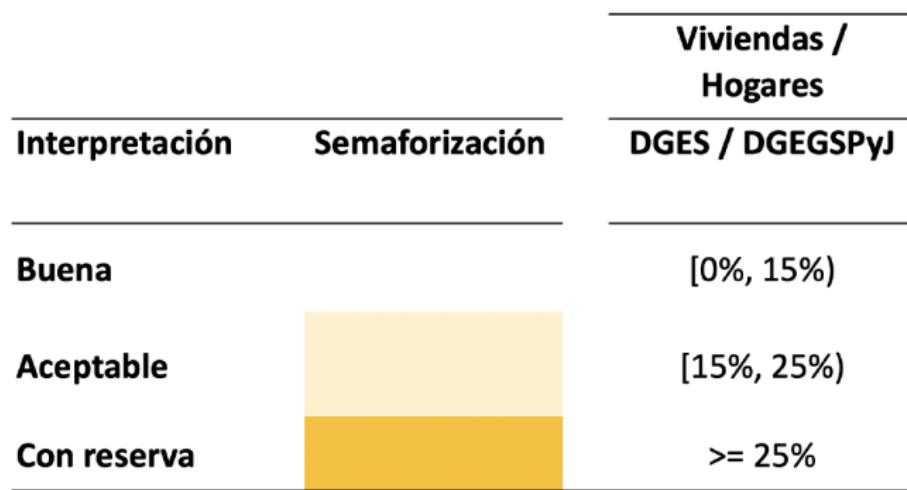


Figura 4: Fuente: INEGI

Alertas sobre el coeficiente de variación

Coeficiente de variación (%)	Número de Observaciones	
	Bajo	Alto
[20 , 100]	Estimador no confiable	Estimador no confiable
[15 , 20)	Estimador no confiable	Descriptivo
[5 , 15)	Descriptivo	Estimador confiable
(0 , 5)	Estimador confiable	Estimador confiable

Figura5: Fuente: INE - Chile

Estándares de alerta en algunos países (encuestas de hogares)

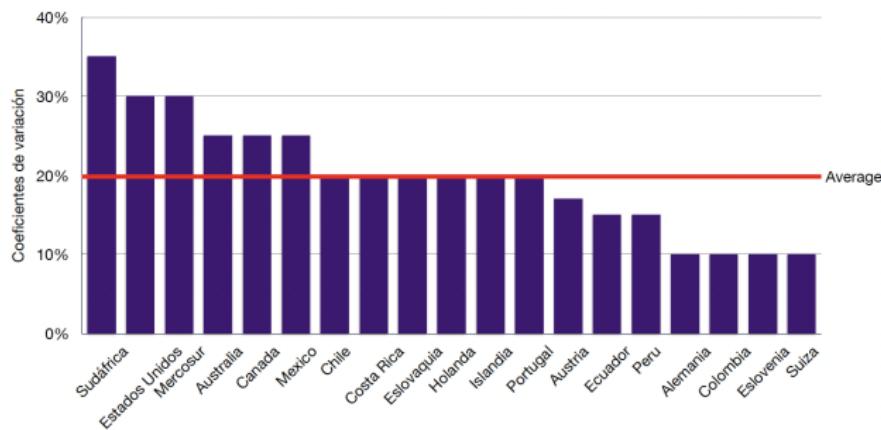


Figura 6: Alertas sobre los coeficientes de variación

Algunas alertas definidas en la publicación

Cuando se sobrepasa el umbral del coeficiente de variación aparecen algunas de las siguientes alertas:

- No se publica
- Usar con precaución.
- Las estimaciones requieren revisiones, no son precisas y se deben usar con precaución.
- Poco confiable, menos preciso.
- No cumple con los estándares de publicación.
- Con reserva, referencial, cuestionable.
- Valores muy aleatorios, estimación pobre.

Dominios de estudio y subpoblaciones de interés

Una encuesta se planea con el fin de generar información precisa y confiable en los dominios de estudio que se han predefinido. Sin embargo, existen subgrupos poblacionales que la encuesta no abordó en su diseño, y sobre los cuales se quisiera una mayor precisión.

- Incidencia de la pobreza desagregado por departamento o provincia (tamaño de muestra conocido y planificado).
- Tasa de desocupación desagregada por sexo (tamaño de muestra aleatorio, pero planificado).
- Tasa de asistencia neta estudiantil en primaria desagregada por quintiles de ingreso (tamaño de muestra aleatorio).

Precisión de los estimadores

Debido a que una encuesta es una investigación parcial sobre una población finita, es necesario saber que:

- A partir de una encuesta, no se calculan indicadores, sino que se estiman con ayuda de los datos de la encuesta.
- Es necesario calcular el grado de error que se comete al no poder realizar una investigación exhaustiva. Este error es conocido como el error de muestreo.
- La precisión de un estimador está supeditada al intervalo de confianza.

Entre más angosto sea el intervalo, más precisión se genera y por ende se tiene un menor error de muestreo.

El intervalo de confianza en subpoblaciones

Si el parámetro de interés sobre el cual se busca realizar la inferencia es θ_d , y se ha definido una subpoblación de interés U_d , entonces un intervalo del 95 % de confianza sobre esa subpoblación está dado por la siguiente expresión:

$$\left(\hat{\theta} - t_{0975,gI} \times se(\hat{\theta}) , \quad \hat{\theta} + t_{0975,gI} \times se(\hat{\theta}) \right)$$

El uso del coeficiente de variación como indicador de la confiabilidad de las estadísticas provenientes de encuestas de hogares debería ser complementado con algunas otras medidas que permitan crear reglas de confiabilidad y precisión.

El intervalo de confianza

Nótese que la longitud de los intervalos de confianza induce la seguridad de que un estimador es preciso:

- La incidencia de la pobreza en el departamento del país se estimó en 5.2 %, con un intervalo de confianza de (5.15 %, 5.25 %).
- La tasa de desocupación en el país para los hombres se ubicó en 7.5 %, con un intervalo de confianza de (7.1 %, 7.9 %); mientras que para las mujeres se ubicó en 9.2 %, con intervalo de confianza de (8.8 %, 9.6 %).
- La tasa de asistencia neta estudiantil en primaria para el último quintil de ingreso se estimó en 85 %, con un intervalo de confianza de (48.2 %, 100.0 %).

El tamaño de muestra

- El tamaño de muestra afecta de manera indirecta la amplitud del intervalo de confianza, a través del error estándar que generalmente decrece a medida que el tamaño de muestra se hace más grande.
- Un tamaño de muestra adecuado garantiza la convergencia en distribución de los estimadores a la distribución teórica de donde se calculan los percentiles.
- Por ejemplo, es posible plantear que todas las estimaciones basadas en un tamaño de muestra menor a un umbral predefinido deberían ser suprimidas o marcadas como no confiables.

El tamaño de muestra efectivo

- En las encuestas de hogares, con diseños de muestreo complejos, no existe una sucesión de variables que sean independientes e identicamente distribuidas.
- La muestra y_1, \dots, y_n no es un vector en el espacio n -dimensional, donde se asume que cada componente del vector puede variar por sí mismo.
- La dimensión final del vector (y_1, \dots, y_n) es mucho menor que n , puesto que existe una forma jerárquica en la selección de los hogares y a la interrelación de la variable de interés con las UPMs

El tamaño de muestra efectivo

El tamaño de muestra efectivo se define como sigue:

$$n_{\text{efectivo}} = \frac{n}{Deff}$$

En donde $Deff$ es el efecto de diseño que depende de: 1. El número de encuestas promedio que se realizaron en cada UPM. 2. La correlación existente entre la variable de interés y las mismas UPMs.

Es posible considerar que, si el tamaño de muestra efectivo no es mayor a un umbral, entonces la cifra no debería ser considerada para publicación.

Grados de libertad

Son una medida de cuántas unidades independientes de información se tienen en la inferencia. Nótese que:

- En el caso extremo de realizar un censo en cada UPMs, sin importar el número de individuos que componen el conglomerado, el número de unidades independientes será únicamente el número de UPMs seleccionadas en la primera etapa de muestreo.
- En las encuestas de hogares, la variabilidad de la estimación es la contribución del conglomerado a la gran media más una contribución (considerada insignificante) de la segunda etapa de muestreo.

Grados de libertad

En las subpoblaciones los grados de libertad no se consideran fijos sino variables.

$$gl = \sum_{h=1}^H \nu_h \times (n_{lh} - 1)$$

Note que ν_h es una variable indicadora que toma el valor uno si el estrato h contiene uno o mas casos de la subpoblación de interés, n_{lh} es el número de UPMs en el estrato. En el caso más general, los grados de libertad se reducen a la siguiente expresión:

$$gl = \#UPMs - \#Estratos$$

Grados de libertad

Por ejemplo, considere por ejemplo el percentil 0.975 para el cual los valores críticos de la distribución t varían con respecto a sus grados de libertad

- $t - student_{gl=1} = 127$
- $t - student_{gl=2} = 430$
- $t - student_{gl=5} = 257$
- $t - student_{gl=40} = 202$
- $t - student_{gl=\infty} = 196$

Es posible considerar que si los grados de libertad inducidos por la subpoblación son menores a un umbral predefinido, la cifra debería ser suprimida.

Ejemplo

Quintil Urbano	sexo	n	Deff	n.eff	gl	Desocupación%	Li %	Ls %	cv %	Alerta
Quinto	Mujer	2055	1.2	1757	309	1.0	0.4	1.6	30.6	*
Quinto	Hombre	1969	1.1	1738	335	1.1	0.5	1.7	26.3	*
Cuarto	Hombre	2245	1.2	1807	347	2.2	1.4	3.0	19.3	
Cuarto	Mujer	2301	1.6	1466	357	4.1	2.7	5.5	17.5	
Tercero	Mujer	2421	1.5	1646	336	6.1	4.3	7.9	15.1	
Segundo	Hombre	2280	1.4	1654	295	5.9	4.3	7.5	13.8	
Tercero	Hombre	2351	1.2	2025	331	4.6	3.4	5.8	13.3	
Segundo	Mujer	2541	1.6	1547	310	10.8	8.0	13.6	13.1	
Primero	Mujer	2862	2.0	1466	266	20.0	15.4	24.6	11.8	
Primero	Hombre	2562	1.6	1610	263	11.9	9.4	14.5	10.9	

Figura 7: Desocupación urbana por quintiles de ingreso y sexo

Ejemplo

Quintil Rural	sexo	n	Deff	n.eff	gl	Desocupación%	Li %	Ls %	cv %	Alerta
Primer	Mujer	1788	0.6	2754	140	0.8	0.1	1.5	44.5	*
Cuarto	Hombre	2112	1.7	1223	178	1.8	0.8	2.7	26.7	*
Segundo	Hombre	2281	1.8	1236	156	2.6	1.3	3.9	25.7	*
Primer	Hombre	1704	1.3	1324	137	2.8	1.4	4.2	25.5	*
Quinto	Mujer	1780	1.3	1391	166	2.6	1.4	3.9	23.7	*
Segundo	Mujer	2195	1.4	1579	158	5.3	2.9	7.6	22.7	*
Quinto	Hombre	2127	1.4	1553	171	1.8	1.0	2.6	22.1	*
Tercero	Hombre	2180	2.0	1068	169	3.7	2.1	5.2	21.4	*
Tercero	Mujer	2023	1.4	1411	164	6.5	3.8	9.2	20.8	*
Cuarto	Mujer	1942	1.6	1225	174	7.3	4.6	10.0	19.1	

Figura8: Desocupación rural por quintiles de ingreso y sexo

Uso de métodos SAE

Justificación

- Los estimadores directos, basados solo en unidades de muestreo observadas para cada área pequeña, no son suficientemente confiables.
- Tamaño de muestra pequeño o incluso ninguna unidad observada (falta de información).
- El coeficiente de variación (CV) es demasiado alto para el indicador objetivo a nivel de área.

Incremento del coeficiente de variación

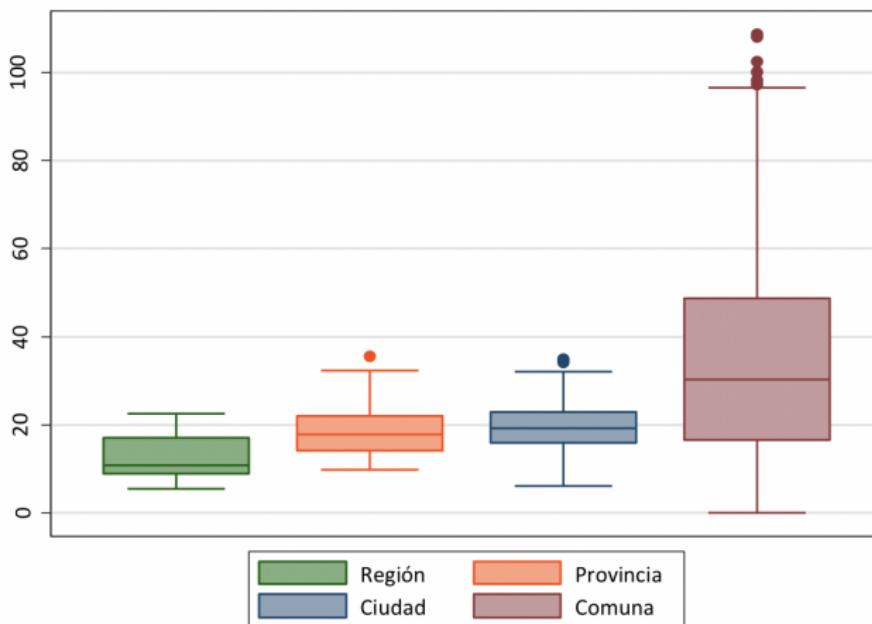


Figura 9: Distribución de los coeficientes de variación en Chile

Justificación

Área	Obs	Mean	Std. Dev.	Min	Max
region	15	12,0	4,9	5,5	22,5
provincia	26	19,5	7,1	9,8	35,5
ciudad	33	20,0	6,5	6,0	34,9
comuna	237	34,8	26,0	0,0	108,7

Figura10: Coeficientes de variación en Chile

Justificación

Cuando los estimadores directos no son confiables para algunos dominios de interés, existen dos opciones:

- ① Sobremuestreo: aumentar el tamaño de la muestra en los dominios de interés (aumento de los costos).
- ② Aplicar técnicas estadísticas que permitan estimaciones confiables en esos dominios, métodos SAE.

Justificación

- Durante la última década ha habido una demanda creciente de (objetivo y subjetivo) indicadores de progreso y bienestar.
- Estas medidas desempeñan un papel central para los responsables de las políticas, para planificar y verificar la efectividad de las mismas.

Ejemplos

- Indicadores de pobreza: en riesgo de pobreza, ingreso de los hogares.
- Indicadores del mercado de trabajo: Tasa de desempleo, Satisfacción con el trabajo, etc.
- Indicadores de salud: esperanza de vida media, porcentaje de población con conductas peligrosas (obesidad, fumadores, etc.)

Justificación

- Para ser informativos y efectivos, estos indicadores deben elegirse a el nivel apropiado de desagregación.
- Los indicadores pueden ser desagregados a lo largo de varias dimensiones, incluyendo áreas geográficas, grupos demográficos, grupos de ingresos / consumo y grupos sociales.

¿Qué es un área pequeña?

- La mayoría de las encuestas nacionales están planificadas para entregar estimaciones confiables a nivel nacional y regional pero a niveles más bajos se reduce la precisión.
- Un área pequeña es un dominio para el cual el tamaño de muestra específico no es suficientemente grande para obtener estimaciones confiables.
- Habitualmente son dominios no planificados y su tamaño de muestra esperado es aleatorio y es más grande a medida que aumenta el tamaño de la población del área.

¿Qué es un área pequeña?

La subpoblación de interés puede ser una zona geográfica o subgrupos socioeconómicos.

- Geográfico: provincias, áreas del mercado de trabajo, municipios, sectores censales para medir por ejemplo la tasa de desempleo a nivel comunal.
- Dominio de subgrupos específicos: edad \times sexo \times raza dentro del ámbito geográfico de una zona, para medir por ejemplo la tasa de desempleo por sexo o edad específica en las zonas urbanas.

¿Qué es un área pequeña?

- La solución es **tomar prestada fuerza** de otras áreas y/o en diferentes ocasiones mediante modelos explícitos o implícitos que explotan la relación entre variables aumentando el tamaño efectivo de la muestra.
- El modelo proporciona un enlace a áreas relacionadas y/o períodos de tiempo a través de información complementaria tales como recuentos de censos (recientes o actuales) o registros administrativos relacionados con la variable objetivo.

Algunos métodos

- Estimador sintético: En el contexto de subpoblaciones, los estimadores se llaman sintéticos cuando éstos se basan en un estimador directo y se estiman a partir de información auxiliar a través de un modelo.
- Estimador compuesto: es una combinación lineal entre un estimador directo y un estimador sintético. Representa un buen compromiso entre las características de los dos componentes.

Algunos métodos

- El estimador compuesto está dado por una combinación lineal de estimador sintético y estimador directo equilibrando el sesgo potencial del estimador sintético contra la inestabilidad del estimador directo (compensación entre precisión y sesgo).
- Las estimaciones más grandes de áreas pequeñas están más cerca de las estimaciones directas mientras que las más pequeñas están más cerca de las estimaciones sintéticas.

Algunos métodos

- Los estimadores SAE se dividen en dos tipos principales dependiendo de cómo se aplican los modelos a los datos dentro de las áreas pequeñas: nivel de área y nivel de unidad.
- Los estimadores de área pequeña se basan en cálculos de nivel de área si los modelos vinculan la variable de interés y con variables auxiliares x específicas del área.

Algunos métodos

- Se llaman modelos a nivel de unidad si se vinculan valores individuales para las variables auxiliares específicas de la unidad.
- Los estimadores basados en áreas pequeñas se calculan a nivel de área si los datos de la unidad no están disponibles.
- También pueden ser calculados si los datos de nivel de unidad están disponibles resumiéndolos en el nivel de área apropiado.

Proceso de estimación



Figura11: Producción de estadísticas con SAE

Algunos riesgos

La producción de estimaciones en área pequeña involucra riesgos que se deben tener en consideración:

- El tamaño de las áreas pequeñas en los términos del número de unidades que les pertenecen es también una consideración importante. Áreas que son demasiado pequeñas pueden presentar problemas de confidencialidad.
- Las estimaciones de área pequeña pueden diferir demasiado de las estadísticas basadas en el conocimiento local.
- Las fuentes de información y el diseño utilizado pueden ser no sustentables en el tiempo.

Algunos riesgos

- El compromiso y la voluntad de la agencia para apoyar estas metodologías a través de sistemas y personal capacitado en la materia.
- Disponibilidad de datos auxiliares correlacionados con la variable de interés.
- Tamaño de muestra debe ser suficientemente grande para permitir estimaciones confiables mediante el uso de los datos de la encuesta y los datos auxiliares existentes.

Consideraciones

- Todos los métodos SAE requieren datos auxiliares a nivel del área pequeña desde el cual **toman prestada la fuerza**.
- La efectividad de los métodos SAE depende del grado de asociación entre la variable de interés y los datos auxiliares.
- La búsqueda de buenas variables auxiliares es crítica, incluida la construcción imaginativa de tales variables.
- Los datos auxiliares deben medirse de manera consistente a través de las áreas pequeñas, pero pueden incluir estimaciones de muestras grandes con error de muestreo conocido.

Desafíos

- Aumento de las tasas de no respuesta.
- Aumento de costos, menos financiación.
- Aumento de la demanda de estimaciones para dominios pequeños como por raza, etnia o pobreza.
- Aumento de la demanda de estimaciones de áreas pequeñas.
- Aumento de la complejidad en los contenidos de los cuestionarios y por lo tanto la carga de respuesta.
- Aumento de la demanda de análisis secundarios, uso público y archivos de datos de uso restringido.

Algunas aplicaciones actuales

Mapas de pobreza

Estimados por el Banco Mundial en diversos países: Perú, Brasil, Guatemala, Nicaragua, Panamá, Ecuador. Combinan datos provenientes de encuestas con datos censales.

Small Area Income and Poverty Estimates (SAIPE)

El Small Area Income and Poverty Estimates (SAIPE) es un programa del Census Bureau de los Estados Unidos. El programa SAIPE produce estimaciones de pobreza para el total de la población y la media de los ingresos por hogares estimada anualmente para todos los condados y estados.

Estimaciones mensuales de empleo local y estatal

A través del programa de Estadísticas de desempleo del área local, la BLS produce estimaciones mensuales del empleo y desempleo total para aproximadamente 7300 áreas, incluido regiones censales, divisiones, estados, condados y ciudades.

Las estimaciones son basadas en datos de varias fuentes, incluyendo el CPS, el programa de estadísticas de empleo actual (CES), los sistemas de seguro de desempleo de los estados (UI) y el censo decenal.

La experiencia canadiense

Utilizaron una serie de fuentes primarias de datos: Censo Canadiense de población y vivienda, La encuesta de fuerza laboral (LFS) Canadiense y el sistema de seguro gubernamental federal de desempleo (UI). Otros datos de áreas pequeñas de mercado laboral fueron utilizados como información auxiliar.

Dado que el objetivo principal era minimizar el error de estimación del modelo, probaron tres técnicas de estimación: Estimación sintética, SPREE y regresión.

En el Reino Unido

En conjunto con la universidad de Southampton, se realizó una estimación del desempleo basado en la definición de la OIT. Se exploraron distintos enfoques utilizando la alta correlación entre las estimaciones de desempleo de la LFS y el número de demandantes de beneficios de subsidios para solicitantes de empleo.

El enfoque principal fue desarrollar modelos de regresión que vinculen las estimaciones de desempleo con información de los solicitantes otorgada por las autoridades locales de los distritos (ONS, 2001b).

En Italia

La experiencia de Italia fue un intento de mejorar el rendimiento de la estimación de la tasa de desocupación a nivel sub-regional a través de la LFS del ITSTAT (Italy statistics) utilizando como referencia los dominios que abarcan los estratos que se pueden agregar a los municipios.

La tasa de desempleo se estimó utilizando un modelo lineal mixto con efectos de área espacialmente correlacionados y covariables como el sexo, edad y el desempleo a nivel área del censo anterior.

¡Gracias!

¡Gracias!

*Curso Internacional de Desagregación de
Estimaciones en Áreas Pequeñas usando R*

Indicadores de pobreza y métodos directos

División de Estadísticas
Comisión Económica para América Latina y el Caribe

2020

- 1 *Introducción*
- 2 *Indicadores comunes de pobreza y desigualdad*
- 3 *Métodos directos para la desagregación de datos de pobreza*
- 4 *Métodos directos: Estimadores Horvitz-Thompson y Hájek*
- 5 *Métodos directos: Estimadores GREG y de calibración*
- 6 *Resultados: Estimación de ingreso medio en sectores de Montevideo*

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

Introducción

- Una encuesta es realizada con un tamaño muestral establecido.
- Después de una encuesta realizada, a menudo se produce una demanda para estimaciones en áreas más desagregadas.
- Por ejemplo, se realiza un muestreo para estimar niveles de pobreza en departamentos, pero después, el cliente quiere que se realicen estas estimaciones a nivel de municipio.

Introducción

- Cuando eso pasa, se puede aumentar los tamaños muestrales en las áreas en las que sea necesario.
- Hay varios métodos para mejorar el diseño muestral.
- No obstante, esto podría ser caro, y el cliente podría pedir más de lo que es posible.

Introducción

- Las subdivisiones para las cuales se desean estimaciones se llaman “áreas” o “dominios”.
- “áreas” pueden ser no solo áreas geográficas, sino también grupos socioeconómicos, o un cruce de ambos tipos.
- A la hora de estimar indicadores en estas áreas, se puede usar un *estimador directo*, lo que usa solamente los datos de la encuesta para esa área.
- Habitualmente son insesgados o prácticamente insesgados con respecto al diseño muestral.
- En esta presentación nos enfocaremos en estos estimadores.

Introducción

- Como se ha dicho, en algunas áreas, el tamaño muestral es demasiado pequeño, lo que incrementa errores de muestreo en los estimadores directos para esas áreas.
- Cuando esto pasa, estas áreas se llaman *areas pequeñas*.
- Esto no refiere al tamaño poblacional del área, sino áreas para las que no se disponen estimadores directos eficientes debido a tamaños muestrales pequeños.

Indicadores comunes de pobreza y desigualdad

Indicadores comunes de pobreza y desigualdad

- El indicador más común para medir pobreza es *la incidencia o tasa de pobreza*, también se conoce como tasa en riesgo de pobreza.
- Otro indicador es la *brecha de la pobreza*, que mide la magnitud de pobreza en lugar de frecuencia.
- Estos dos son parte de una familia de indicadores más amplia definidos por Foster, Greer y Thorbecke (1984), que llamaremos *indicadores FGT*.
- Ambos indicadores tienen la ventaja de ser aditivos.

Indicadores comunes de pobreza y desigualdad

- Llamemos U a la población objetivo de tamaño N , la cual se divide en D subpoblaciones de tamaños N_1, \dots, N_D .
- Llamemos E_{di} al poder adquisitivo (e.g.medida de ingresos o gastos) del individuo i en área d .
- Llamamos z al umbral predefinido de pobreza, por debajo del cual un individuo se considera en riesgo de pobreza.

Indicadores comunes de pobreza y desigualdad

- Los indicadores FGT para el área d pueden ser definidos por:

$$F_{\alpha d} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\frac{z - E_{di}}{z} \right)^\alpha I(E_{di} < z), \quad d = 1, \dots, D, \quad \alpha \geq 0$$

donde $I(E_{di} < z)$ es una función indicadora que toma el valor 1 si $E_{di} < z$ y 0 en caso contrario. Note que:

- Con $\alpha = 0$, obtenemos la *tasa de pobreza*
- Con $\alpha = 1$, obtenemos la *brecha de pobreza*

Métodos directos para la desagregación de datos de pobreza

Métodos directos

- En esta sección, se describirán estimadores directos para la media de una variable en un área, dada por:

$$\bar{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} Y_{di}$$

donde Y_{di} es el valor de la variable de individuo i en área d .

Métodos directos

- Los indicadores FGT,

$$F_{\alpha,di} = \left(\frac{z - E_{di}}{z} \right)^\alpha I(E_{di} < z),$$

también se pueden escribir en la forma de la diapositiva anterior.

- Llámemos $F_{\alpha d}$ a la media de $Y_{di} = F_{\alpha,di}$ en el dominio d .
- Entonces,

$$F_{\alpha d} = N_d^{-1} \sum_{i=1}^{N_d} F_{\alpha,di}$$

- Este parámetro solo usa los datos del dominio d en cuestión.

Métodos directos: Estimadores Horvitz-Thompson y Hájek

Métodos directos: Horvitz-Thompson (HT)

- El estimador de Horvitz-Thompson es insesgado con respecto al diseño muestral para la media de área d , \hat{Y}_d .
- El estimador HT está definido como

$$\hat{Y}_d = N_d^{-1} \sum_{i \in s_d} w_{di} Y_{di}$$

- En donde w_{di} es el peso de muestreo dado por la siguiente expresión

$$w_{di} = \frac{1}{\pi_{d,i}}$$

- $\pi_{d,i} = Pr(i \in s)$ es la probabilidad de inclusión del elemento a la muestra.

Métodos directos: Horvitz-Thompson (HT)

- El estimador de Horvitz-Thompson también es insesgado con respecto al diseño muestral para el total de área d ,
$$Y_d = \sum_{i=1}^{N_d} Y_{di}.$$
- Está dado por la siguiente expresión

$$\hat{Y}_d = \sum_{i \in s_d} w_{di} Y_{di}$$

- Al contrario que en la estimación para medias, para este caso no se necesita conocer el tamaño poblacional, N_d .

Métodos directos: Horvitz-Thompson (HT)

- Un estimador para la varianza del estimador HT viene dado por

$$\widehat{\text{var}}_{\pi}(\hat{Y}_d) = N_d^{-2} \left\{ \sum_{i \in s_d} \sum_{j \in s_d} (w_{di} w_{dj} - w_{d,ij}) Y_{di} Y_{dj} \right\}$$

En donde $w_{d,ij} = \frac{1}{\pi_{d,ij}}$ y $\pi_{d,ij} = Pr(i, j \in s)$.

- Este estimador es insesgado si $\pi_{di} > 0$ y

$$\pi_{d,ij} > 0$$

para todo i, j .

- Si se supone que $w_{d,ij} \approx w_{di} w_{dj}$, el estimador queda definido por:

$$\widehat{\text{var}}_{\pi}(\hat{Y}_d) = N_d^{-2} \sum_{i \in s_d} w_{di} (w_{di} - 1) Y_{di}^2$$

Métodos directos: Horvitz-Thompson (HT)

- Como se ha mencionado, los indicadores FGT se pueden escribir como una media para individuos en un área,

$$F_{\alpha d} = N_d^{-1} \sum_{i=1}^{N_d} F_{\alpha,di}$$

- Por consiguiente, el estimador HT de $F_{\alpha d}$ es,

$$\hat{F}_{\alpha d} = N_d^{-1} \sum_{i \in s_d} w_{di} F_{\alpha,di}$$

Métodos directos: Horvitz-Thompson (HT)

- Podemos usar el estimador HT, $\hat{Y}_d = \sum_{i \in s_d} w_{di} Y_{di}$, para estimar el total poblacional, es decir,

$$\hat{Y} = \sum_{d=1}^D \hat{Y}_d = \sum_{d=1}^D \sum_{i \in s_d} w_{di} Y_{di}$$

- Esta propiedad se llama *benchmarking*, donde los estimadores para áreas desagregadas suman al estimador para el total.

Métodos directos: Horvitz-Thompson (HT), comentario sobre benchmarking

- Cuando no se cumple la propiedad de benchmarking, es común ajustar de la siguiente manera:

$$\hat{Y}_d^{AEST} = \hat{Y}_d^{EST} \frac{\hat{Y}}{\sum_{d=1}^D \hat{Y}_d^{EST}}, \quad d = 1, \dots, D$$

Métodos directos: Hájek

- Aunque el estimador HT es insesgado, puede tener una varianza muy grande bajo el diseño muestral.
- El estimador de Hájek es ligeramente sesgado pero con una varianza menor que la de HT, escrito de la siguiente forma,

$$\hat{Y}_d^{HA} = \hat{N}_d^{-1} \sum_{i \in s_d} w_{di} Y_{di}, \text{ donde } \hat{N}_d = \sum_{i \in s_d} w_{di}$$

- Observe que no se necesita conocer el tamaño poblacional como con el estimador de Horvitz-Thompson.

Métodos directos: Hájek

- Un estimador de la varianza de Hájek, \hat{Y}_d^{HA} , se obtiene con un proceso de linealización de Taylor.
- Si suponemos que $\pi_{d,ij} \approx \pi_{di}\pi_{dj}$ para todo $j \neq i$, y que todo $\pi_{di} > 0$, obtenemos:

$$\widehat{\text{var}}_\pi(\hat{Y}_d^{HA}) = \hat{N}_d^{-2} \sum_{i \in s_d} w_{di}(w_{di} - 1)(Y_{di} - \hat{Y}_d^{HA})^2$$

Métodos directos: Hájek

- Como se ha mencionado, variables FGT se pueden escribir como una media para individuos en un área.
- Por consiguiente, el estimador de Hájek de $F_{\alpha d}$ es,

$$\hat{F}_{\alpha d}^{HA} = \hat{N}_d^{-1} \sum_{i \in s_d} w_{di} F_{\alpha, di}$$

Resumen de estimadores HT y Hájek

- Indicadores objetivos:

- Parámetros aditivos (que son sumas de ciertas variables para cada individuo del área).
- Pueden ser funciones de variables de interés, por ejemplo, $F_{\alpha,di} = f(E_{di})$.

- Requerimientos de datos:

- Pesos muestrales w_{di} para individuos en grupo d .
- Para algunos estimadores se necesita conocer el tamaño poblacional del área N_d .

Resumen de estimadores HT y Hájek

- Ventajas:

- El estimador HT es insesgado y el de Hájek es ligeramente sesgado.
- Ambos son consistentes cuando n_d crece.
- Son no paramétricos porque no se supone nada de la distribución de Y_{di} .

Resumen de estimadores HT y Hájek

- Desventajas:
 - Son muy inefficientes para áreas con tamaños de muestra pequeños.
 - No se puede calcular un estimador cuando $n_d = 0$, o cuando el área no es muestreada.

Métodos directos: Estimadores GREG y de calibración

Métodos directos: Estimador GREG

- El estimador generalizado de regresión (*generalized regression*), GREG, utiliza información auxiliar.
- Este estimador requiere el total $\mathbf{X}_d = \sum_{i=1}^{N_d} \mathbf{x}_{di}$, o la media $\bar{\mathbf{X}}_d = N_d^{-1} \sum_{i=1}^{N_d} \mathbf{x}_{di}$, para el área d .
- El vector \mathbf{x}_{di} consiste de valores de p variables auxiliares relacionadas con Y_{di} , para el individuo i en el área d .

Métodos directos: Estimador GREG

- Asumamos que existe un modelo de la forma

$$Y_{di} = \mathbf{x}'_{di}\beta_d + \epsilon_{di}, \quad i = 1, \dots, N_d$$

- Entonces, podemos definir un estimador

$$\hat{\mathbf{B}}_d = \left(\sum_{i \in s_d} w_{di} \mathbf{x}_{di} \mathbf{x}'_{di} / c_{di} \right)^{-1} \sum_{i \in s_d} w_{di} \mathbf{x}_{di} Y_{di} / c_{di}$$

- En el modelo, los errores ϵ_{di} son independientes con esperanza igual a 0 y varianza $\sigma^2 c_{di}$, con $c_{di} > 0$ siendo constantes que representan la posible heteroscedasticidad, $i = 1, \dots, N_d$.

Métodos directos: Estimador GREG

- $\hat{\bar{\mathbf{X}}}_d = N_d^{-1} \sum_{i \in s_d} w_{di} \mathbf{x}_{di}$ es el estimador de HT de $\bar{\mathbf{X}}_d$
- Podemos usar la regresión mencionada para estimar $\hat{\bar{Y}}_d$
- Este estimador está dado por:

$$\hat{\bar{Y}}_d^{GREG} = \hat{\bar{Y}}_d + (\bar{\mathbf{X}}_d - \hat{\bar{\mathbf{X}}}_d)' \hat{\mathbf{B}}_d$$

Métodos directos: Estimador GREG

- El estimador GREG es más eficiente que el estimador directo \hat{Y} si las variables auxiliares x_{di} están linealmente relacionadas con Y_{di} ,
- No es fácil encontrar auxiliares x_{di} relacionadas con $F_{\alpha,di} = I\{(z - E_{di})/z\}^{\alpha} I(E_{di} < z)$, porque es una función compleja.

Métodos directos: Estimador GREG

- Si $\pi_{d,ij} \approx \pi_{di}\pi_{dj}$, para $j \neq i$, el estimador de varianza para GREG viene dado por:

$$\widehat{\text{var}}_{\pi}(\hat{Y}_d^{GREG}) = N_d^{-2} \sum_{i \in s_d} w_{di}(w_{di} - 1) \tilde{e}_{di}^2$$

donde $\tilde{e}_{di} = Y_{di} - \mathbf{x}'_{di} \hat{\mathbf{B}}_d$.

Métodos directos: Estimador de calibración

- Este método utiliza los pesos calibrados h_{di} para estimar el total de una variable de interés usando p variables auxiliares.
- h_{di} son los pesos más cercanos a los pesos originales, w_{di} , sujeto a

$$\sum_{i \in s_d} h_{di} \mathbf{x}_{di} = \mathbf{X}_d$$

- Una posibilidad viene dada por

$$h_{di} = w_{di} \left\{ 1 + \mathbf{x}'_{di} \left(\sum_{i \in s_d} w_{di} \mathbf{x}_{di} \mathbf{x}'_{di} / c_{di} \right)^{-1} \left(\mathbf{X}_d - \sum_{i \in s_d} w_{di} \mathbf{x}_{di} / c_{di} \right) \right\}, i \in s_d$$

Métodos directos: Estimador de calibración

- El estimador de calibración de \bar{Y}_d se obtiene igual que el estimador de HT

$$\hat{\bar{Y}}_d^{CAL} = N_d^{-1} \sum_{i \in s_d} h_{di} Y_{di}$$

- Se puede mostrar que, bajo ciertas condiciones de regularidad, el estimador de calibración es asintóticamente igual al GREG y comparten la misma varianza asintótica.

Resumen de estimadores GREG y de calibración

- Indicadores objetivo: Medias/totales de la variable de interés.
- Requerimientos de datos:
 - Pesos muestrales w_{di} para individuos de la muestra en el área d .
 - Para el estimador de la media, tamaño poblacional del área N_d .
 - Observaciones muestrales de las p variables auxiliares.
 - Totales \mathbf{X}_d o medias $\bar{\mathbf{X}}_d$ poblacionales de las p variables auxiliares.

Resumen de estimadores GREG y de calibración

- Ventajas:

- Son aproximadamente insensibles con respecto al diseño muestral.
- Pueden mejorar a los estimadores directos básicos si el modelo de regresión tiene buen poder predictivo.
- No requieren la verificación del modelo considerado para las variables de interés Y_{di} ; son no paramétricos.

Resumen de estimadores GREG y de calibración

- Desventajas:
 - Pueden ser ineficientes para áreas pequeñas.
 - No se pueden calcular en áreas con un tamaño muestro n_d igual a 0.

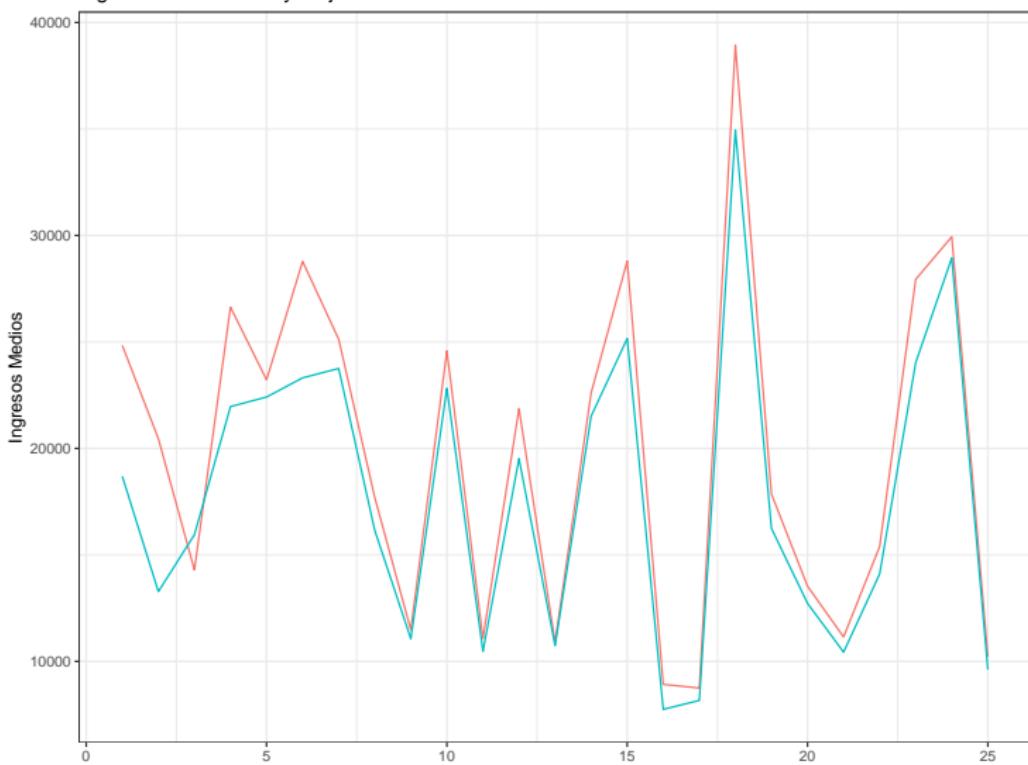
Resultados: Estimación de ingreso medio en sectores de Montevideo

Horvitz Thompson: Hombres y Mujeres en Montevideo

sec2	ntotal	Hombres	Mujeres
2	121	20461	13277
1	167	24837	18694
3	186	14299	15951
4	319	26635	21965
6	320	28784	23314
5	495	23223	22414
21	3165	11148	10435
13	3556	10897	10742
18	3950	38932	34943
11	3963	11080	10473
17	4373	8750	8167
10	6302	24576	22823

Horvitz Thompson: Hombres y Mujeres en Montevideo

Ingresos de hombres y mujeres en Montevideo con el estimador HT

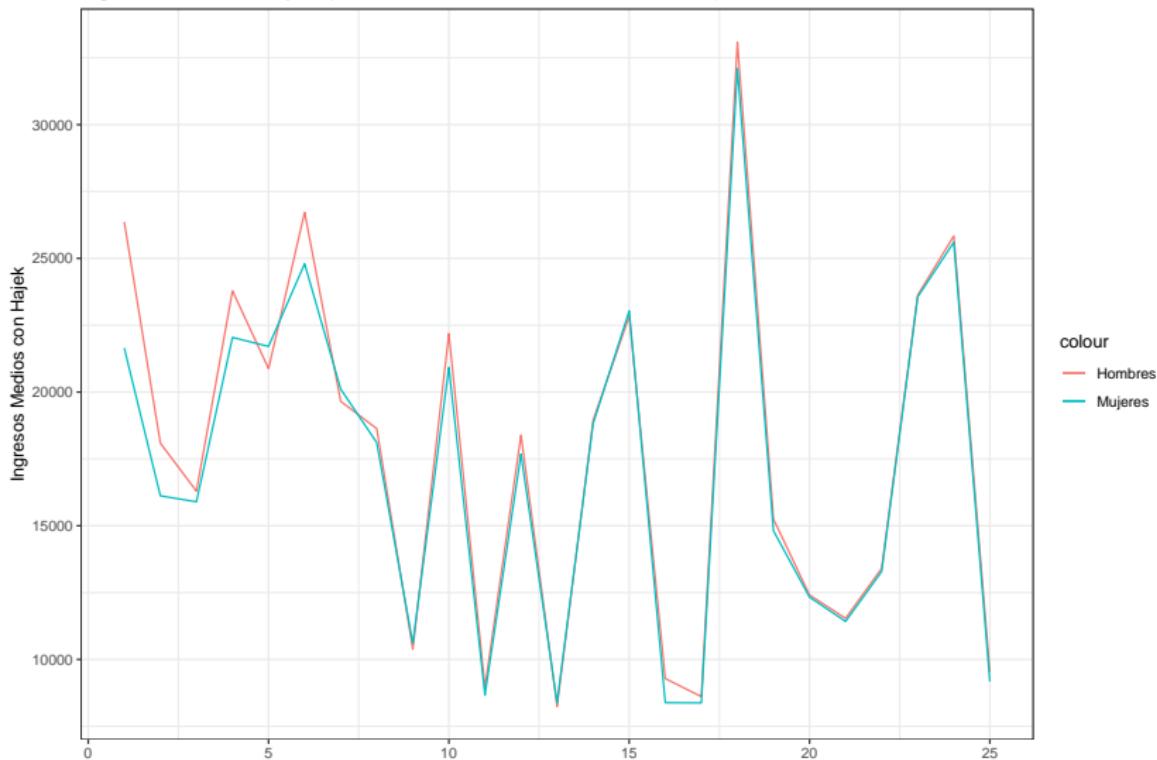


Hájek: Hombres y Mujeres en Montevideo

sec2	ntotal	Hombres	Mujeres
2	121	18088	16120
1	167	26363	21644
3	186	16294	15896
4	319	23786	22044
6	320	26723	24798
5	495	20874	21706
21	3165	11539	11424
13	3556	8248	8384
18	3950	33081	32103
11	3963	8954	8675
17	4373	8612	8377
10	6302	22186	20929

Hájek: Hombres y Mujeres en Montevideo

Ingresos de hombres y mujeres en Montevideo con el estimador Hájek

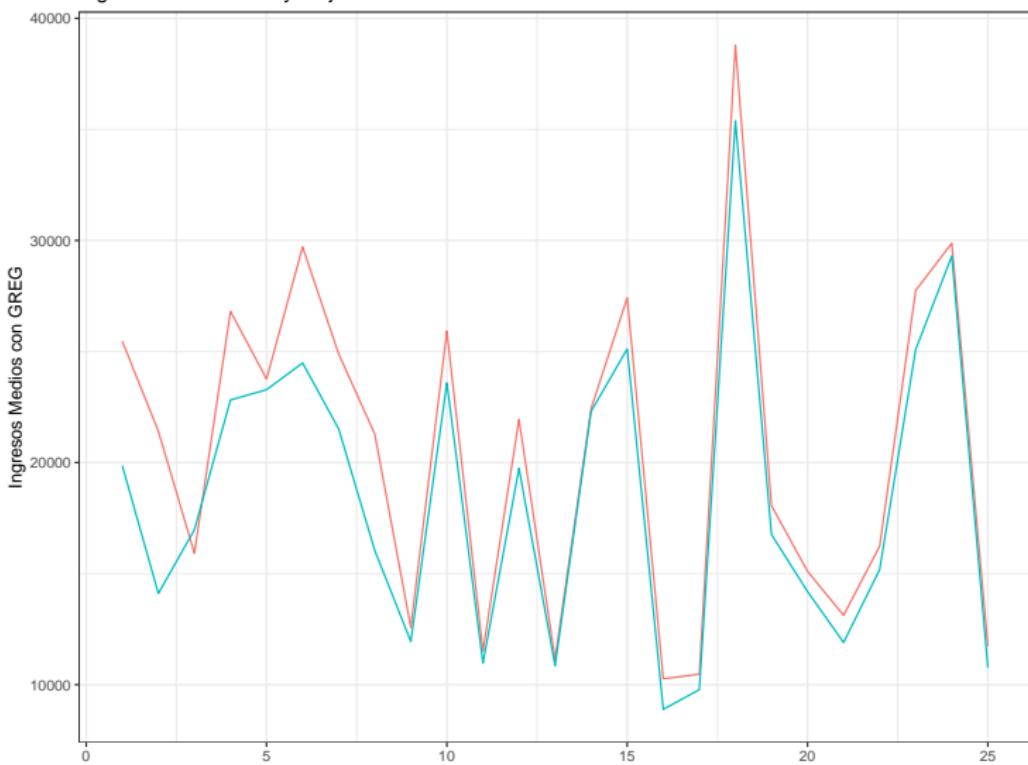


GREG: Hombres y Mujeres en Montevideo

sec2	ntotal	Hombres	Mujeres
2	121	21410	14107
1	167	25468	19861
3	186	15921	16981
4	319	26809	22819
6	320	29710	24484
5	495	23763	23282
21	3165	13125	11901
13	3556	11156	10862
18	3950	38789	35391
11	3963	11510	10977
17	4373	10473	9777
10	6302	25921	23589

GREG: Hombres y Mujeres en Montevideo

Ingresos de hombres y mujeres en Montevideo con el estimador GREG

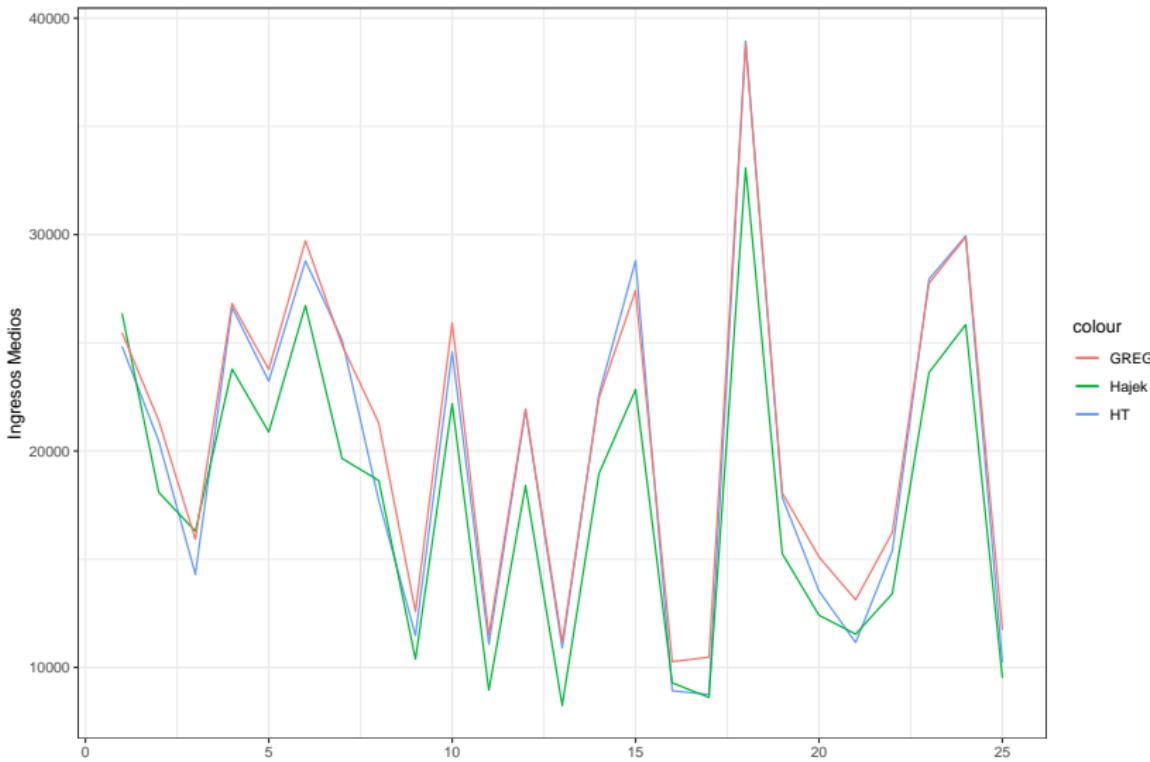


Comparando los estimadores: Hombres

sec2	ntotal	HT	Hajek	GREG
2	121	20461	18088	21410
1	167	24837	26363	25468
3	186	14299	16294	15921
4	319	26635	23786	26809
6	320	28784	26723	29710
5	495	23223	20874	23763
21	3165	11148	11539	13125
13	3556	10897	8248	11156
18	3950	38932	33081	38789
11	3963	11080	8954	11510
17	4373	8750	8612	10473
10	6302	24576	22186	25921

Comparando los estimadores: Hombres

Ingresos de hombres en Montevideo con estimadores directos

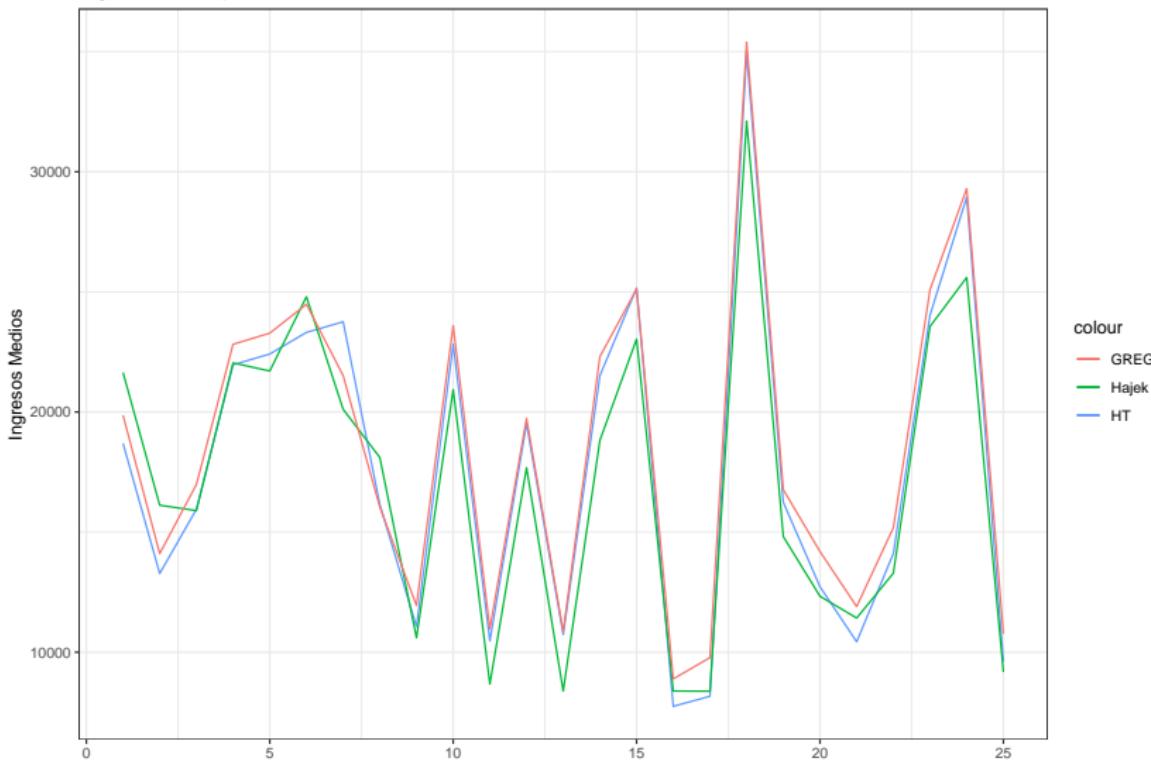


Comparando los estimadores: Mujeres

sec2	ntotal	HT	Hajek	GREG
2	121	13277	16120	14107
1	167	18694	21644	19861
3	186	15951	15896	16981
4	319	21965	22044	22819
6	320	23314	24798	24484
5	495	22414	21706	23282
21	3165	10435	11424	11901
13	3556	10742	8384	10862
18	3950	34943	32103	35391
11	3963	10473	8675	10977
17	4373	8167	8377	9777
10	6302	22823	20929	23589

Comparando los estimadores: Mujeres

Ingresa de mujeres en Montevideo con estimadores directos



¡Gracias!

¡Gracias!

*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

- 1 *Estimador post-estratificado sintético*
- 2 *Estimador sintético de regresión a nivel de área (REG1-SYN)*
- 3 *Estimador sintético de regresión a nivel de individuo (REG2-SYN)*
- 4 *Estimadores compuestos*
- 5 *Resultados: Estimación de ingreso medio en sectores de Montevideo*

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, \dots, U^J con tamaños poblacionales N^1, \dots, N^J

Estimador post-estratificado sintético

- El área d , U_d también es dividida en J grupos, llamados post-estratos, U_d^1, \dots, U_d^J de tamaño N_d^1, \dots, N_d^J .
- Tienen medias $\bar{Y}_d^1, \dots, \bar{Y}_d^J$, donde $\bar{Y}_d^j = \sum_{i \in U_d^j} Y_{di} / N_d^j$, $j = 1, \dots, J$.
- 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,

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

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 \bar{Y}_d así:

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

- 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,

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

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 $\hat{Y}^{j, HA}$ 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

$$F_{\alpha d}^j = F_\alpha^j, \quad j = 1, \dots, J$$

donde F_α^j 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), $N_d^j, j = 1, \dots, J$
 - 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, $\hat{\delta}_d, d = 1, \dots, D$
- El modelo se asume entonces,

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

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 $\text{var}(\hat{\delta}_d)$, $d = 1, \dots, D$.
- 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

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

- Donde

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

Estimador sintético de regresión a nivel de área

- Para α , el sesgo bajo el diseño de $\hat{\delta}_d^{REG1-SYN}$ viene dado por $\mathbf{x}'_d \alpha - \delta_d$.
- 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,
 $\hat{F}_{\alpha d} = \mathbf{x}'_d \boldsymbol{\alpha} + \varepsilon_d$, $d = 1, \dots, D$ 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 $d = 1, \dots, D$, \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} , $i \in s_d$, $d = 1, \dots, D$.
- Se puede obtener por tanto un modelo lineal a nivel de individuo para Y_{di} .
- Llamamos $\mathbf{y}_d = (Y_{d1}, \dots, Y_{dN_d})$ 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,

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

- 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

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

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

Estimador sintético de regresión a nivel de individuo

- El vector de predicciones para el área d es entonces
 $\hat{\mathbf{y}}_d = (\hat{Y}_{d1}, \dots, \hat{Y}_{dN_d})'$ donde $\hat{Y}_{di} = \mathbf{x}'_{di}\hat{\beta}$, $i = 1, \dots, N_d$
- 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 , $\delta_d = \bar{Y}_d$, si $\bar{\mathbf{X}}_d$ es el vector de medias poblacionales de las p variables auxiliares, $\hat{\bar{Y}}_d^{REG2-SYN}$ sería

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

- Se obtiene el estimador para un área no muestreada de la misma forma.
- Para β conocido, el sesgo bajo el diseño de la media es $\bar{\mathbf{X}}_d' \beta - \bar{Y}_d$, 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

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

- 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, $\bar{\mathbf{X}}_d$, $d = 1, \dots, D$

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”.

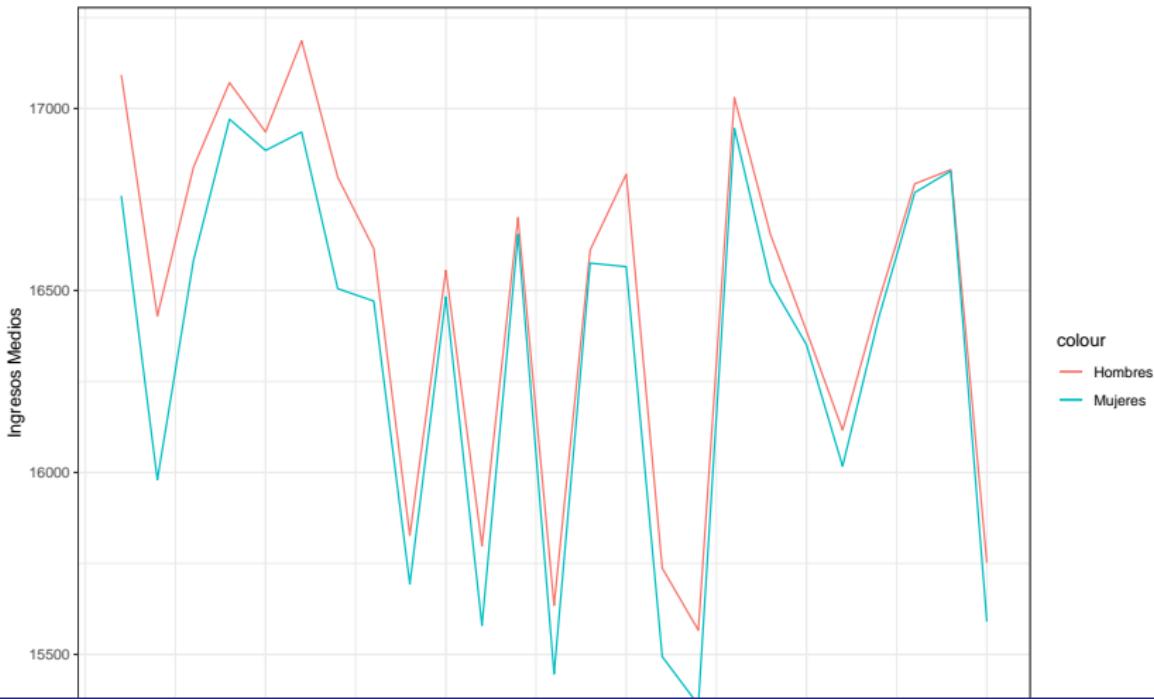
Resultados: Estimación de ingreso medio en sectores de Montevideo

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

Ingresos de hombres y mujeres en Montevideo con el estimador post–estratificado sintético

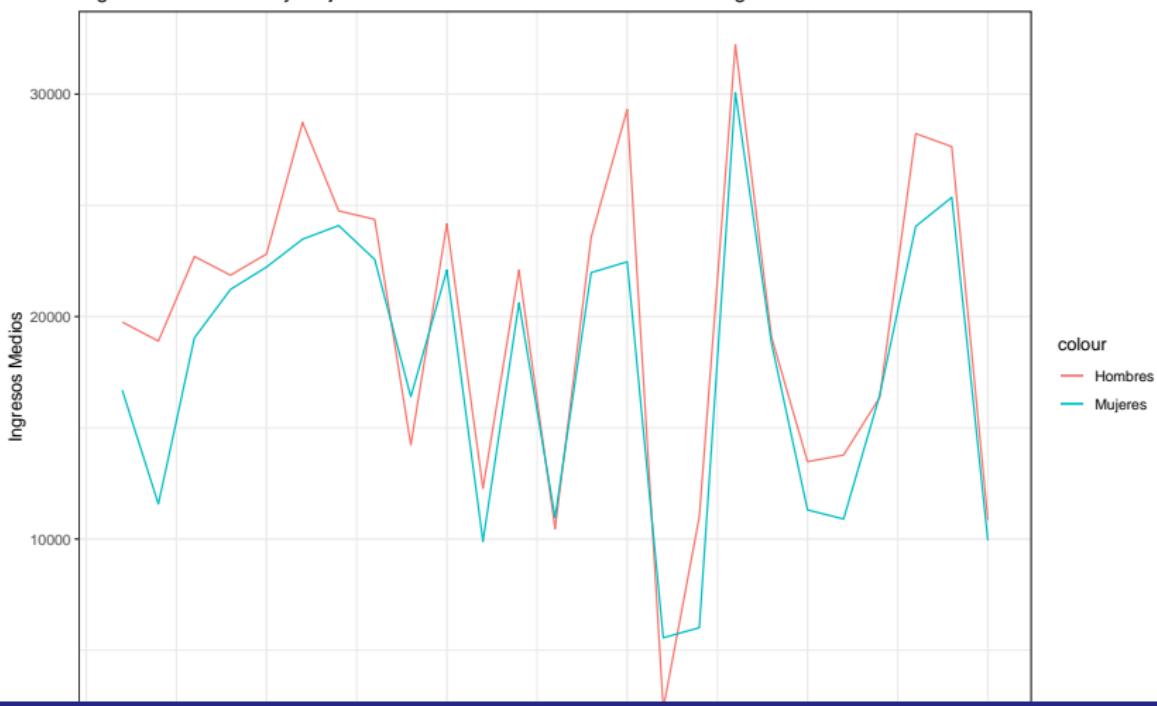


Estimador de regresión sintético a nivel de área: Hombres y Mujeres en Montevideo

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

Ingresos de hombres y mujeres en Montevideo con el estimador de regresión sintético a nivel de área

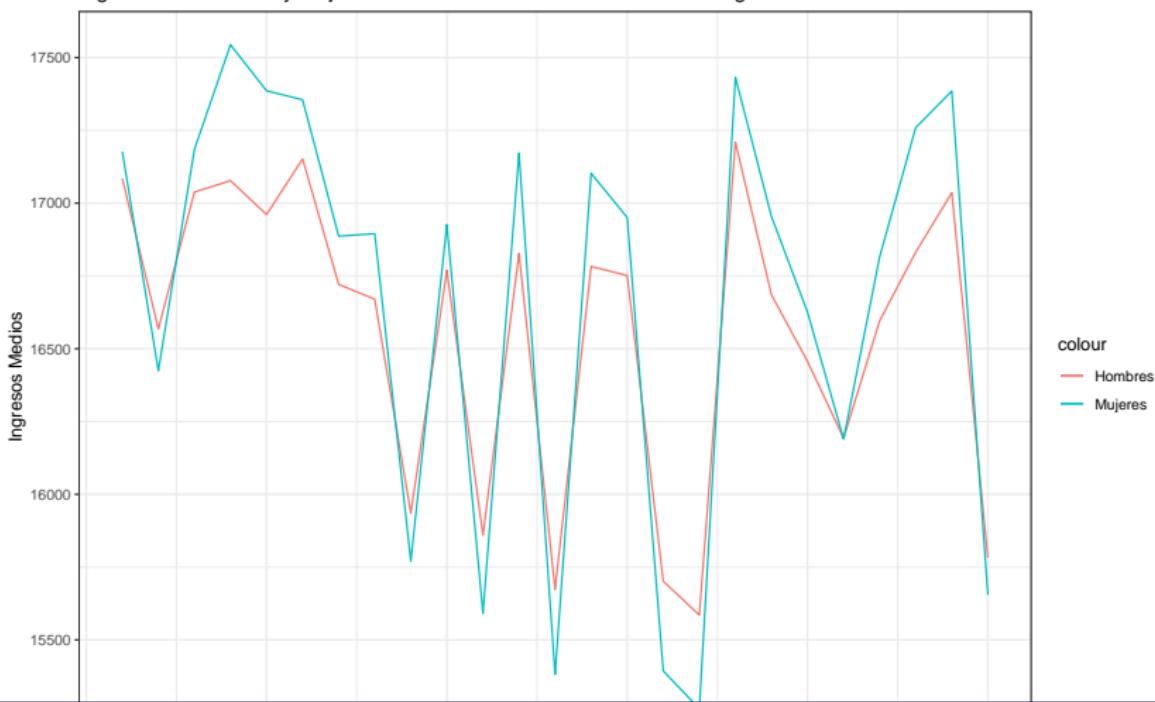


*Estimador de regresión sintético a nivel de individuo:
Hombres y Mujeres en Montevideo*

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

Ingresos de hombres y mujeres en Montevideo con el estimador de regresión sintético a nivel de individuo

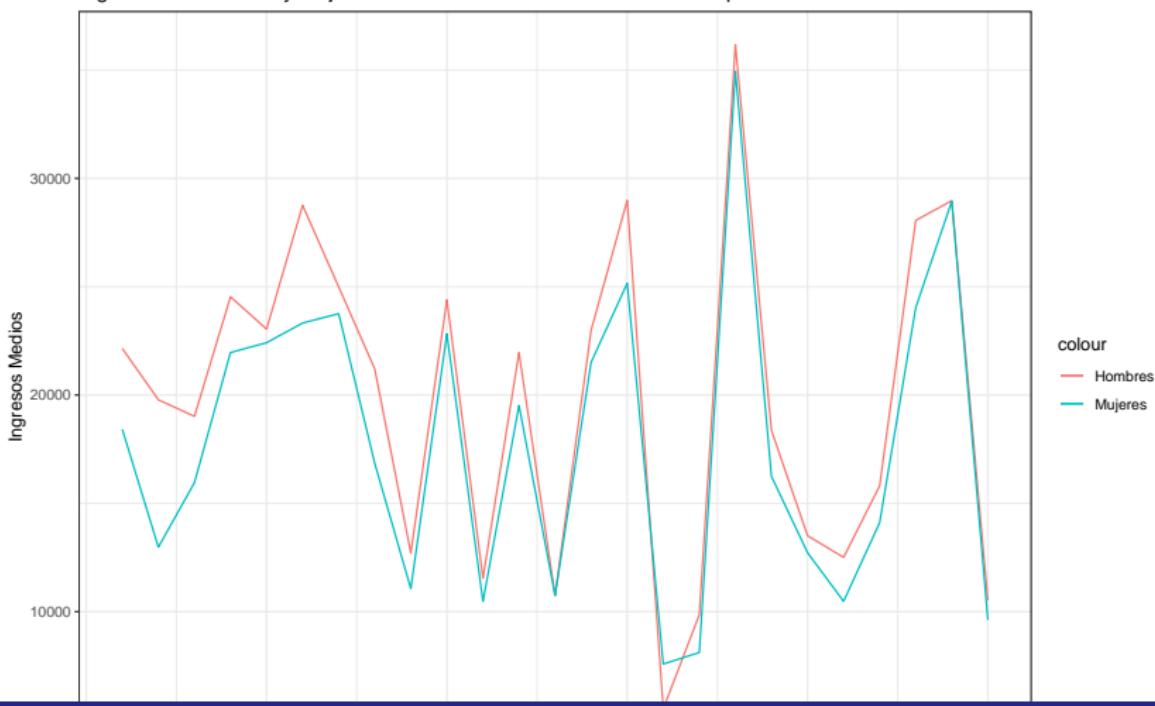


Estimador compuesto: Hombres y Mujeres en Montevideo

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

Ingresos de hombres y mujeres en Montevideo con el estimador compuesto

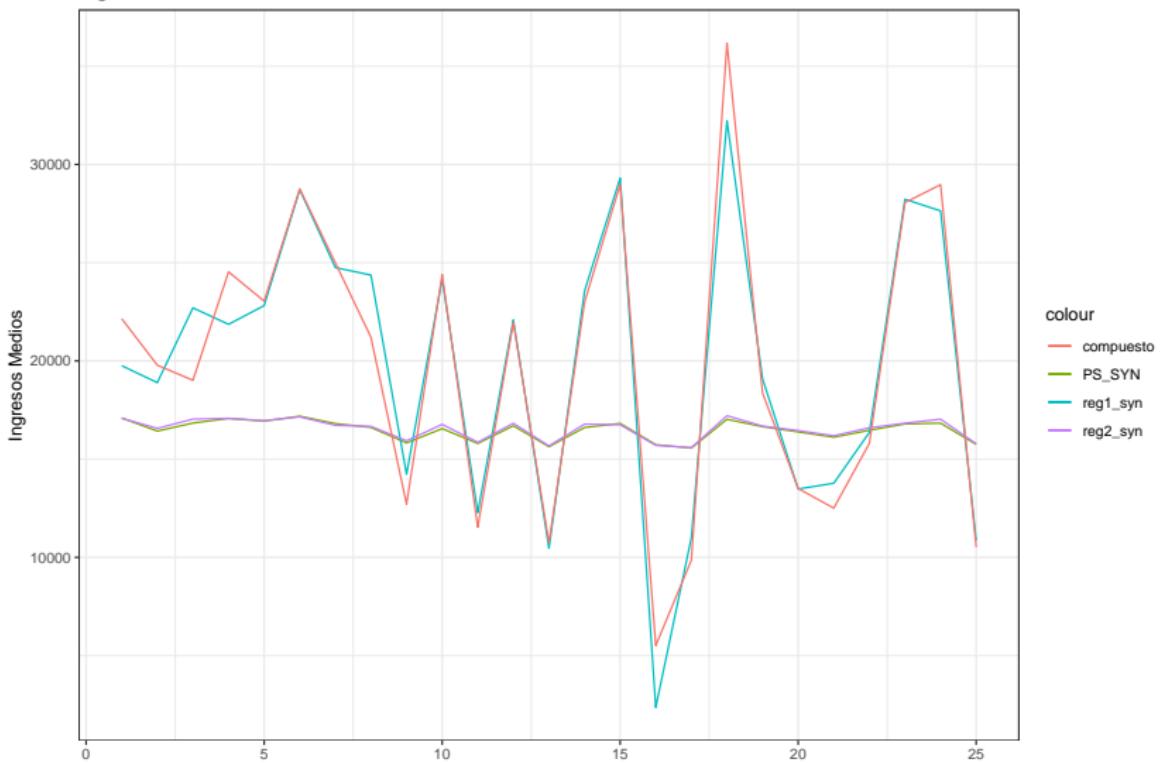


Comparando los estimadores: Hombres

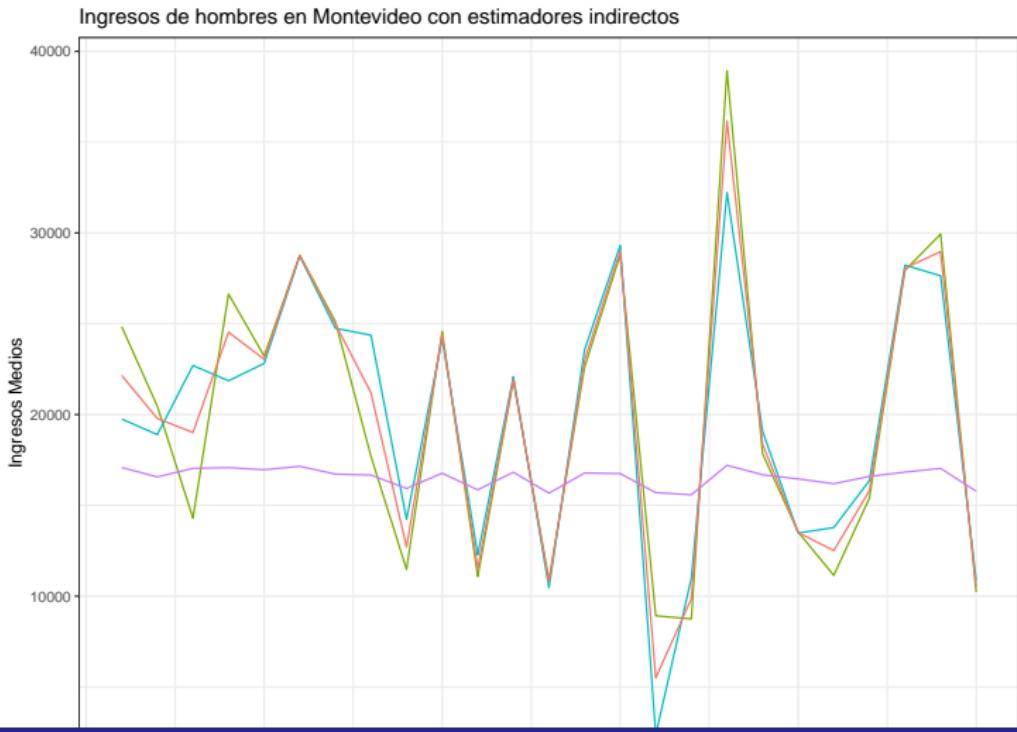
sec2	ntotal	HT	PS_SYN	reg1_syn	reg2_syn	compuesto
2	121	20461	16430	18896	16568	19781
1	167	24837	17093	19756	17085	22149
3	186	14299	16838	22701	17038	19015
4	319	26635	17071	21862	17077	24534
6	320	28784	17186	28726	17152	28757
5	495	23223	16935	22816	16961	23042
21	3165	11148	16117	13776	16193	12506
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

Ingresa de hombres en Montevideo con estimadores indirectos



Comparando los estimadores: Hombres, usando HT para referencia

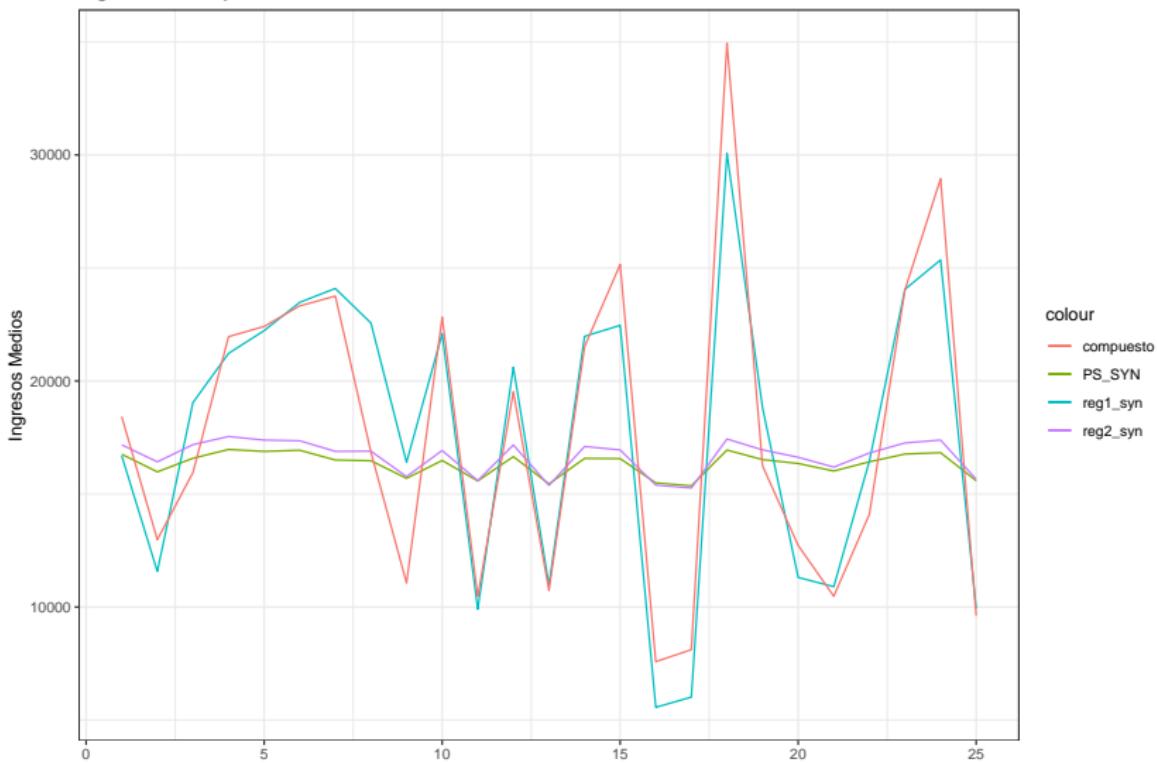


Comparando los estimadores: Mujeres

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

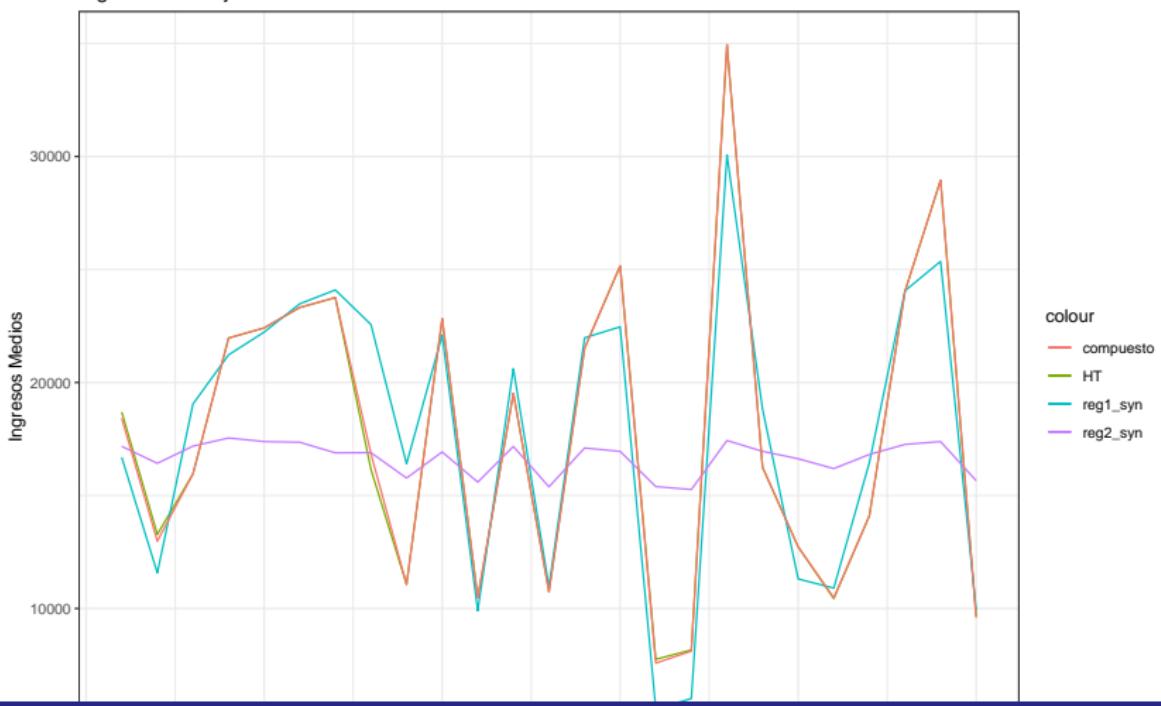
Comparando los estimadores: Mujeres

Ingresa de mujeres en Montevideo con estimadores indirectos



Comparando los estimadores: Mujeres, usando HT para referencia

Ingresos de mujeres en Montevideo con estimadores indirectos



¡Gracias!

¡Gracias!

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

- 1 BLUP/EBLUP basado en el modelo Fay-Herriot
- 2 Resultados: Estimación de ingreso medio en sectores de Montevideo

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 , $d = 1, \dots, D$, asumiendo que varían respecto a un vector de p covariables, \mathbf{x}_d , de forma constante.
- Viene dado por

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

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

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

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 muestreo.
- $\hat{\delta}_d^{DIR}$ todavía se considera insesgado bajo el diseño muestral.

BLUP/EBLUP basado en el modelo Fay-Herriot

- Podemos definir, entonces,

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

donde e_d es el error debido al muestreo, $e_d \stackrel{ind}{\sim} (0, \psi_d)$.

- Dichas varianzas $\psi_d = \text{var}_{\pi}(\hat{\delta}_d^{DIR} | \delta_d)$, $d = 1, \dots, D$, se estiman con los microdatos de la encuesta.
- Por tanto, el modelo se hace,

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

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,

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

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

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

BLUP/EBLUP basado en el modelo Fay-Herriot

- Note que

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

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 ,

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

donde

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

y

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

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 β 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:

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

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

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

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

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

y

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

Resumen del estimador FH

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

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”.

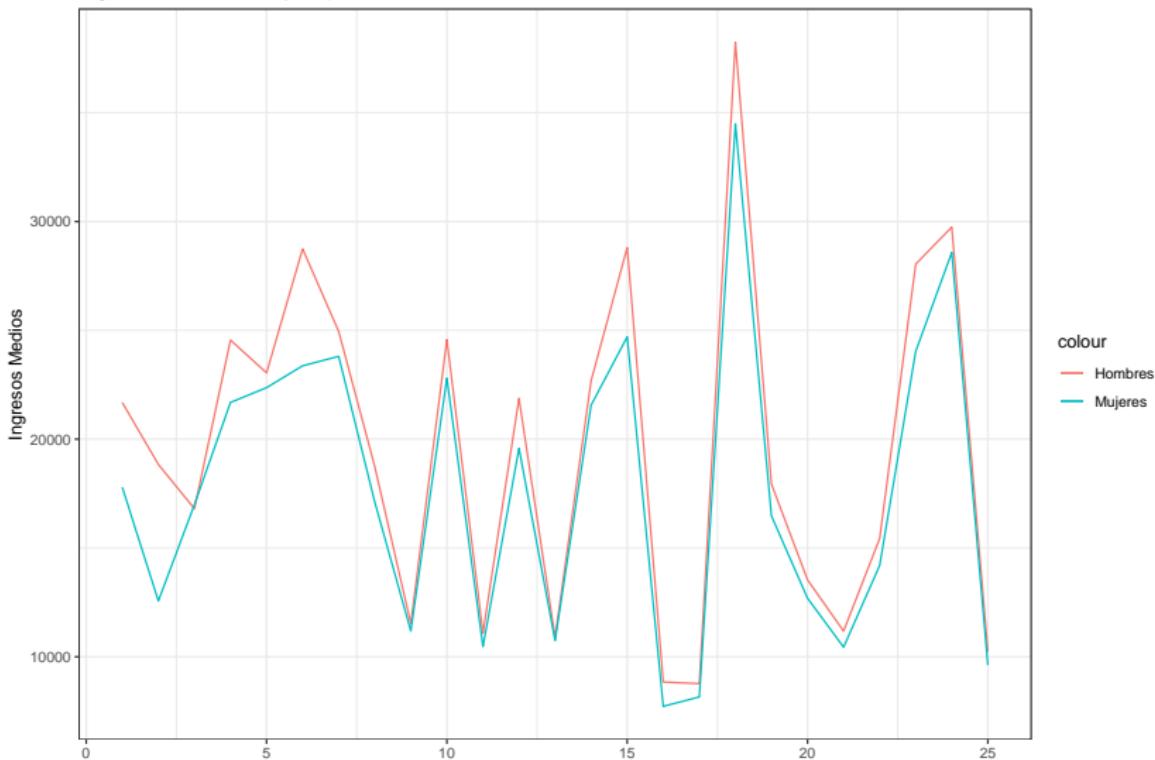
Resultados: Estimación de ingreso medio en sectores de Montevideo

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

Ingresaos de hombres y mujeres en Montevideo con el estimador FH

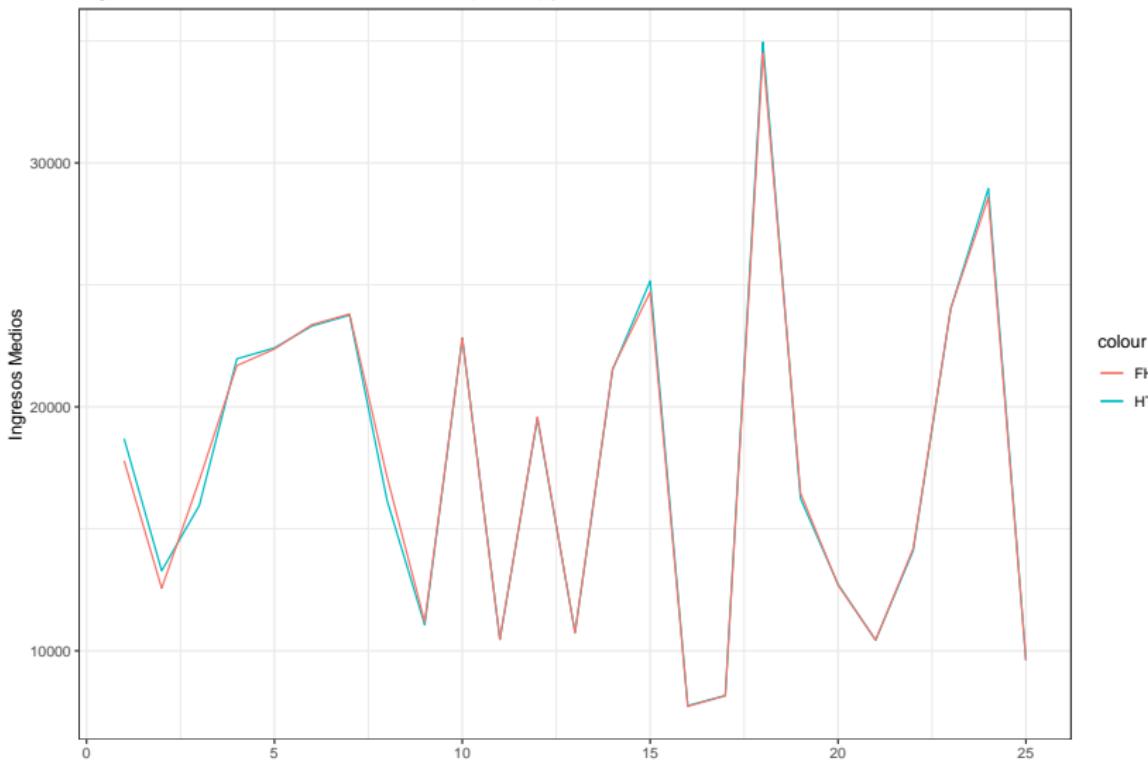


Comparando los estimadores: Hombres

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

Ingresaos de hombres en Montevideo: HT (directo) y FH

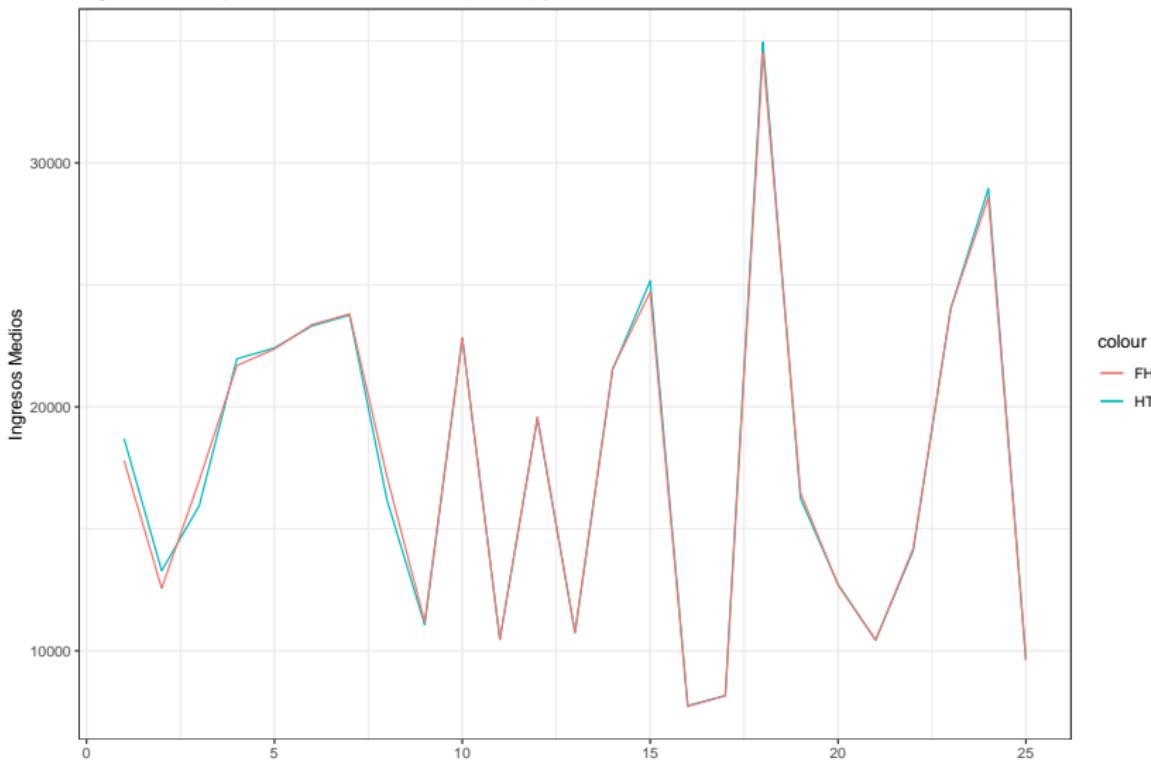


Comparando los estimadores: Mujeres

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

Ingresos de mujeres en Montevideo: HT (directo) y FH



¡Gracias!

¡Gracias!

*Curso Internacional de Desagregación de
Estimaciones en Áreas Pequeñas usando R*

Indicadores de pobreza y métodos directos

División de Estadísticas
Comisión Económica para América Latina y el Caribe

2020

- 1 *Introducción*
- 2 *Indicadores comunes de pobreza y desigualdad*
- 3 *Métodos directos para la desagregación de datos de pobreza*
- 4 *Métodos directos: Estimadores Horvitz-Thompson y Hájek*
- 5 *Métodos directos: Estimadores GREG y de calibración*
- 6 *Resultados: Estimación de ingreso medio en sectores de Montevideo*

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

Introducción

- Una encuesta es realizada con un tamaño muestral establecido.
- Después de una encuesta realizada, a menudo se produce una demanda para estimaciones en áreas más desagregadas.
- Por ejemplo, se realiza un muestreo para estimar niveles de pobreza en departamentos, pero después, el cliente quiere que se realicen estas estimaciones a nivel de municipio.

Introducción

- Cuando eso pasa, se puede aumentar los tamaños muestrales en las áreas en las que sea necesario.
- Hay varios métodos para mejorar el diseño muestral.
- No obstante, esto podría ser caro, y el cliente podría pedir más de lo que es posible.

Introducción

- Las subdivisiones para las cuales se desean estimaciones se llaman “áreas” o “dominios”.
- “áreas” pueden ser no solo áreas geográficas, sino también grupos socioeconómicos, o un cruce de ambos tipos.
- A la hora de estimar indicadores en estas áreas, se puede usar un *estimador directo*, lo que usa solamente los datos de la encuesta para esa área.
- Habitualmente son insesgados o prácticamente insesgados con respecto al diseño muestral.
- En esta presentación nos enfocaremos en estos estimadores.

Introducción

- Como se ha dicho, en algunas áreas, el tamaño muestral es demasiado pequeño, lo que incrementa errores de muestreo en los estimadores directos para esas áreas.
- Cuando esto pasa, estas áreas se llaman *areas pequeñas*.
- Esto no refiere al tamaño poblacional del área, sino áreas para las que no se disponen estimadores directos eficientes debido a tamaños muestrales pequeños.

Indicadores comunes de pobreza y desigualdad

Indicadores comunes de pobreza y desigualdad

- El indicador más común para medir pobreza es *la incidencia o tasa de pobreza*, también se conoce como tasa en riesgo de pobreza.
- Otro indicador es la *brecha de la pobreza*, que mide la magnitud de pobreza en lugar de frecuencia.
- Estos dos son parte de una familia de indicadores más amplia definidos por Foster, Greer y Thorbecke (1984), que llamaremos *indicadores FGT*.
- Ambos indicadores tienen la ventaja de ser aditivos.

Indicadores comunes de pobreza y desigualdad

- Llamemos U a la población objetivo de tamaño N , la cual se divide en D subpoblaciones de tamaños N_1, \dots, N_D .
- Llamemos E_{di} al poder adquisitivo (e.g.medida de ingresos o gastos) del individuo i en área d .
- Llamamos z al umbral predefinido de pobreza, por debajo del cual un individuo se considera en riesgo de pobreza.

Indicadores comunes de pobreza y desigualdad

- Los indicadores FGT para el área d pueden ser definidos por:

$$F_{\alpha d} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\frac{z - E_{di}}{z} \right)^\alpha I(E_{di} < z), \quad d = 1, \dots, D, \quad \alpha \geq 0$$

donde $I(E_{di} < z)$ es una función indicadora que toma el valor 1 si $E_{di} < z$ y 0 en caso contrario. Note que:

- Con $\alpha = 0$, obtenemos la *tasa de pobreza*
- Con $\alpha = 1$, obtenemos la *brecha de pobreza*

Métodos directos para la desagregación de datos de pobreza

Métodos directos

- En esta sección, se describirán estimadores directos para la media de una variable en un área, dada por:

$$\bar{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} Y_{di}$$

donde Y_{di} es el valor de la variable de individuo i en área d .

Métodos directos

- Los indicadores FGT,

$$F_{\alpha,di} = \left(\frac{z - E_{di}}{z} \right)^\alpha I(E_{di} < z),$$

también se pueden escribir en la forma de la diapositiva anterior.

- Llamemos $F_{\alpha d}$ a la media de $Y_{di} = F_{\alpha,di}$ en el dominio d .
- Entonces,

$$F_{\alpha d} = N_d^{-1} \sum_{i=1}^{N_d} F_{\alpha,di}$$

- Este parámetro solo usa los datos del dominio d en cuestión.

Métodos directos: Estimadores Horvitz-Thompson y Hájek

Métodos directos: Horvitz-Thompson (HT)

- El estimador de Horvitz-Thompson es insesgado con respecto al diseño muestral para la media de área d , \hat{Y}_d .
- El estimador HT está definido como

$$\hat{Y}_d = N_d^{-1} \sum_{i \in s_d} w_{di} Y_{di}$$

- En donde w_{di} es el peso de muestreo dado por la siguiente expresión

$$w_{di} = \frac{1}{\pi_{d,i}}$$

- $\pi_{d,i} = Pr(i \in s)$ es la probabilidad de inclusión del elemento a la muestra.

Métodos directos: Horvitz-Thompson (HT)

- El estimador de Horvitz-Thompson también es insesgado con respecto al diseño muestral para el total de área d ,
$$Y_d = \sum_{i=1}^{N_d} Y_{di}.$$
- Está dado por la siguiente expresión

$$\hat{Y}_d = \sum_{i \in s_d} w_{di} Y_{di}$$

- Al contrario que en la estimación para medias, para este caso no se necesita conocer el tamaño poblacional, N_d .

Métodos directos: Horvitz-Thompson (HT)

- Un estimador para la varianza del estimador HT viene dado por

$$\widehat{\text{var}}_{\pi}(\hat{Y}_d) = N_d^{-2} \left\{ \sum_{i \in s_d} \sum_{j \in s_d} (w_{di} w_{dj} - w_{d,ij}) Y_{di} Y_{dj} \right\}$$

En donde $w_{d,ij} = \frac{1}{\pi_{d,ij}}$ y $\pi_{d,ij} = Pr(i, j \in s)$.

- Este estimador es insesgado si $\pi_{di} > 0$ y

$$\pi_{d,ij} > 0$$

para todo i, j .

- Si se supone que $w_{d,ij} \approx w_{di} w_{dj}$, el estimador queda definido por:

$$\widehat{\text{var}}_{\pi}(\hat{Y}_d) = N_d^{-2} \sum_{i \in s_d} w_{di} (w_{di} - 1) Y_{di}^2$$

Métodos directos: Horvitz-Thompson (HT)

- Como se ha mencionado, los indicadores FGT se pueden escribir como una media para individuos en un área,

$$F_{\alpha d} = N_d^{-1} \sum_{i=1}^{N_d} F_{\alpha,di}$$

- Por consiguiente, el estimador HT de $F_{\alpha d}$ es,

$$\hat{F}_{\alpha d} = N_d^{-1} \sum_{i \in s_d} w_{di} F_{\alpha,di}$$

Métodos directos: Horvitz-Thompson (HT)

- Podemos usar el estimador HT, $\hat{Y}_d = \sum_{i \in s_d} w_{di} Y_{di}$, para estimar el total poblacional, es decir,

$$\hat{Y} = \sum_{d=1}^D \hat{Y}_d = \sum_{d=1}^D \sum_{i \in s_d} w_{di} Y_{di}$$

- Esta propiedad se llama *benchmarking*, donde los estimadores para áreas desagregadas suman al estimador para el total.

Métodos directos: Horvitz-Thompson (HT), comentario sobre benchmarking

- Cuando no se cumple la propiedad de benchmarking, es común ajustar de la siguiente manera:

$$\hat{Y}_d^{AEST} = \hat{Y}_d^{EST} \frac{\hat{Y}}{\sum_{d=1}^D \hat{Y}_d^{EST}}, \quad d = 1, \dots, D$$

Métodos directos: Hájek

- Aunque el estimador HT es insesgado, puede tener una varianza muy grande bajo el diseño muestral.
- El estimador de Hájek es ligeramente sesgado pero con una varianza menor que la de HT, escrito de la siguiente forma,

$$\hat{Y}_d^{HA} = \hat{N}_d^{-1} \sum_{i \in s_d} w_{di} Y_{di}, \text{ donde } \hat{N}_d = \sum_{i \in s_d} w_{di}$$

- Observe que no se necesita conocer el tamaño poblacional como con el estimador de Horvitz-Thompson.

Métodos directos: Hájek

- Un estimador de la varianza de Hájek, \hat{Y}_d^{HA} , se obtiene con un proceso de linealización de Taylor.
- Si suponemos que $\pi_{d,ij} \approx \pi_{di}\pi_{dj}$ para todo $j \neq i$, y que todo $\pi_{di} > 0$, obtenemos:

$$\widehat{\text{var}}_\pi(\hat{Y}_d^{HA}) = \hat{N}_d^{-2} \sum_{i \in s_d} w_{di}(w_{di} - 1)(Y_{di} - \hat{Y}_d^{HA})^2$$

Métodos directos: Hájek

- Como se ha mencionado, variables FGT se pueden escribir como una media para individuos en un área.
- Por consiguiente, el estimador de Hájek de $F_{\alpha d}$ es,

$$\hat{F}_{\alpha d}^{HA} = \hat{N}_d^{-1} \sum_{i \in s_d} w_{di} F_{\alpha, di}$$

Resumen de estimadores HT y Hájek

- Indicadores objetivos:

- Parámetros aditivos (que son sumas de ciertas variables para cada individuo del área).
- Pueden ser funciones de variables de interés, por ejemplo, $F_{\alpha,di} = f(E_{di})$.

- Requerimientos de datos:

- Pesos muestrales w_{di} para individuos en grupo d .
- Para algunos estimadores se necesita conocer el tamaño poblacional del área N_d .

Resumen de estimadores HT y Hájek

- Ventajas:

- El estimador HT es insesgado y el de Hájek es ligeramente sesgado.
- Ambos son consistentes cuando n_d crece.
- Son no paramétricos porque no se supone nada de la distribución de Y_{di} .

Resumen de estimadores HT y Hájek

- Desventajas:
 - Son muy inefficientes para áreas con tamaños de muestra pequeños.
 - No se puede calcular un estimador cuando $n_d = 0$, o cuando el área no es muestreada.

Métodos directos: Estimadores GREG y de calibración

Métodos directos: Estimador GREG

- El estimador generalizado de regresión (*generalized regression*), GREG, utiliza información auxiliar.
- Este estimador requiere el total $\mathbf{X}_d = \sum_{i=1}^{N_d} \mathbf{x}_{di}$, o la media $\bar{\mathbf{X}}_d = N_d^{-1} \sum_{i=1}^{N_d} \mathbf{x}_{di}$, para el área d .
- El vector \mathbf{x}_{di} consiste de valores de p variables auxiliares relacionadas con Y_{di} , para el individuo i en el área d .

Métodos directos: Estimador GREG

- Asumamos que existe un modelo de la forma

$$Y_{di} = \mathbf{x}'_{di}\beta_d + \epsilon_{di}, \quad i = 1, \dots, N_d$$

- Entonces, podemos definir un estimador

$$\hat{\mathbf{B}}_d = \left(\sum_{i \in s_d} w_{di} \mathbf{x}_{di} \mathbf{x}'_{di} / c_{di} \right)^{-1} \sum_{i \in s_d} w_{di} \mathbf{x}_{di} Y_{di} / c_{di}$$

- En el modelo, los errores ϵ_{di} son independientes con esperanza igual a 0 y varianza $\sigma^2 c_{di}$, con $c_{di} > 0$ siendo constantes que representan la posible heteroscedasticidad, $i = 1, \dots, N_d$.

Métodos directos: Estimador GREG

- $\hat{\bar{\mathbf{X}}}_d = N_d^{-1} \sum_{i \in s_d} w_{di} \mathbf{x}_{di}$ es el estimador de HT de $\bar{\mathbf{X}}_d$
- Podemos usar la regresión mencionada para estimar $\hat{\bar{Y}}_d$
- Este estimador está dado por:

$$\hat{\bar{Y}}_d^{GREG} = \hat{\bar{Y}}_d + (\bar{\mathbf{X}}_d - \hat{\bar{\mathbf{X}}}_d)' \hat{\mathbf{B}}_d$$

Métodos directos: Estimador GREG

- El estimador GREG es más eficiente que el estimador directo \hat{Y} si las variables auxiliares x_{di} están linealmente relacionadas con Y_{di} ,
- No es fácil encontrar auxiliares x_{di} relacionadas con $F_{\alpha,di} = I\{(z - E_{di})/z\}^{\alpha} I(E_{di} < z)$, porque es una función compleja.

Métodos directos: Estimador GREG

- Si $\pi_{d,ij} \approx \pi_{di}\pi_{dj}$, para $j \neq i$, el estimador de varianza para GREG viene dado por:

$$\widehat{\text{var}}_{\pi}(\hat{Y}_d^{GREG}) = N_d^{-2} \sum_{i \in s_d} w_{di}(w_{di} - 1) \tilde{e}_{di}^2$$

donde $\tilde{e}_{di} = Y_{di} - \mathbf{x}'_{di} \hat{\mathbf{B}}_d$.

Métodos directos: Estimador de calibración

- Este método utiliza los pesos calibrados h_{di} para estimar el total de una variable de interés usando p variables auxiliares.
- h_{di} son los pesos más cercanos a los pesos originales, w_{di} , sujeto a

$$\sum_{i \in s_d} h_{di} \mathbf{x}_{di} = \mathbf{X}_d$$

- Una posibilidad viene dada por

$$h_{di} = w_{di} \left\{ 1 + \mathbf{x}'_{di} \left(\sum_{i \in s_d} w_{di} \mathbf{x}_{di} \mathbf{x}'_{di} / c_{di} \right)^{-1} \left(\mathbf{X}_d - \sum_{i \in s_d} w_{di} \mathbf{x}_{di} / c_{di} \right) \right\}, i \in s_d$$

Métodos directos: Estimador de calibración

- El estimador de calibración de \bar{Y}_d se obtiene igual que el estimador de HT

$$\hat{\bar{Y}}_d^{CAL} = N_d^{-1} \sum_{i \in s_d} h_{di} Y_{di}$$

- Se puede mostrar que, bajo ciertas condiciones de regularidad, el estimador de calibración es asintóticamente igual al GREG y comparten la misma varianza asintótica.

Resumen de estimadores GREG y de calibración

- Indicadores objetivo: Medias/totales de la variable de interés.
- Requerimientos de datos:
 - Pesos muestrales w_{di} para individuos de la muestra en el área d .
 - Para el estimador de la media, tamaño poblacional del área N_d .
 - Observaciones muestrales de las p variables auxiliares.
 - Totales \mathbf{X}_d o medias $\bar{\mathbf{X}}_d$ poblacionales de las p variables auxiliares.

Resumen de estimadores GREG y de calibración

- Ventajas:

- Son aproximadamente insensibles con respecto al diseño muestral.
- Pueden mejorar a los estimadores directos básicos si el modelo de regresión tiene buen poder predictivo.
- No requieren la verificación del modelo considerado para las variables de interés Y_{di} ; son no paramétricos.

Resumen de estimadores GREG y de calibración

- Desventajas:
 - Pueden ser ineficientes para áreas pequeñas.
 - No se pueden calcular en áreas con un tamaño muestro n_d igual a 0.

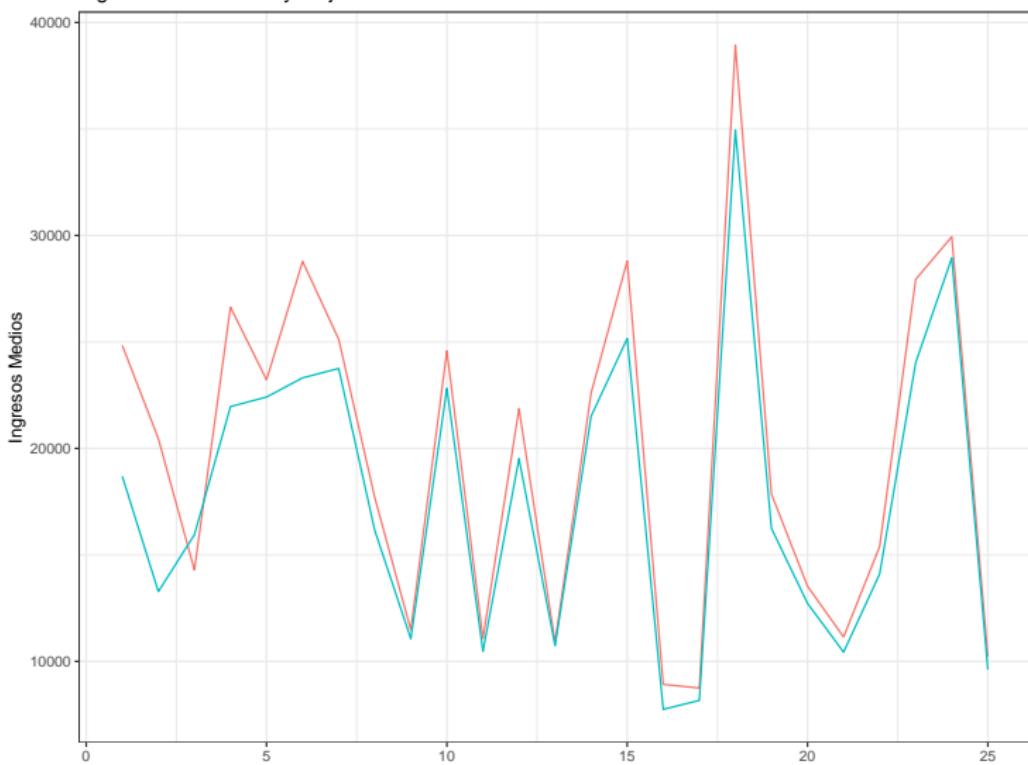
Resultados: Estimación de ingreso medio en sectores de Montevideo

Horvitz Thompson: Hombres y Mujeres en Montevideo

sec2	ntotal	Hombres	Mujeres
2	121	20461	13277
1	167	24837	18694
3	186	14299	15951
4	319	26635	21965
6	320	28784	23314
5	495	23223	22414
21	3165	11148	10435
13	3556	10897	10742
18	3950	38932	34943
11	3963	11080	10473
17	4373	8750	8167
10	6302	24576	22823

Horvitz Thompson: Hombres y Mujeres en Montevideo

Ingresos de hombres y mujeres en Montevideo con el estimador HT

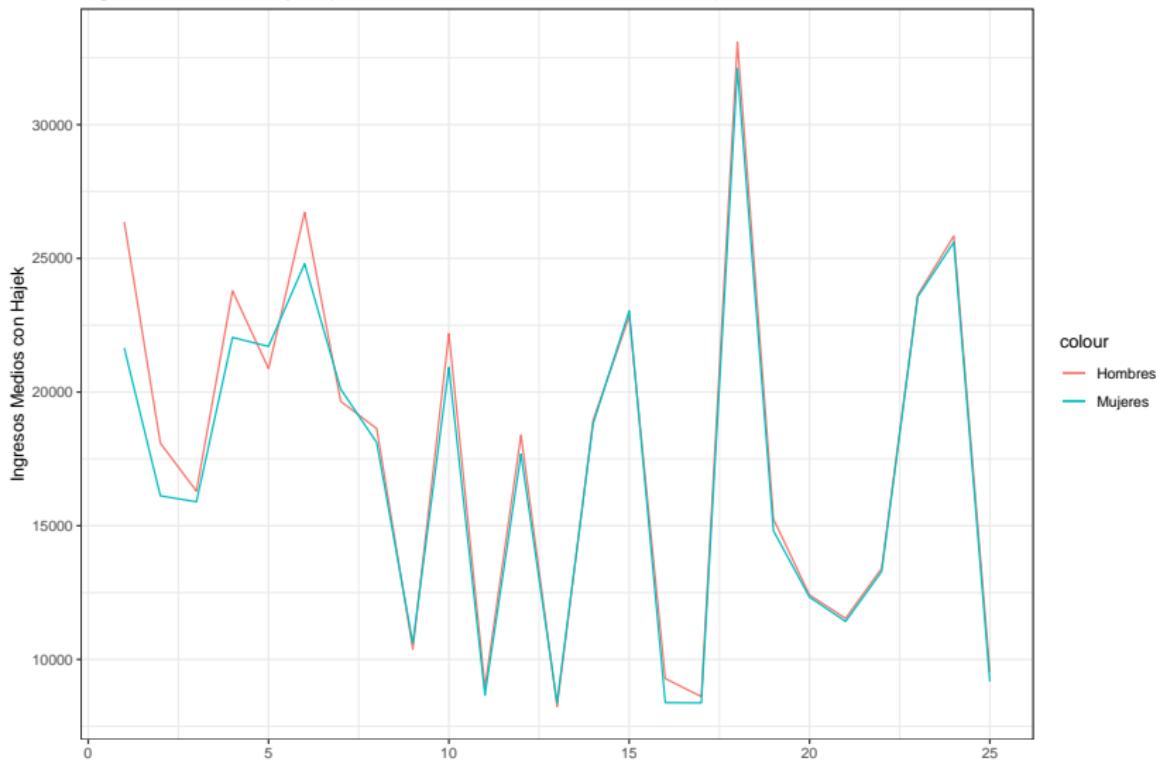


Hájek: Hombres y Mujeres en Montevideo

sec2	ntotal	Hombres	Mujeres
2	121	18088	16120
1	167	26363	21644
3	186	16294	15896
4	319	23786	22044
6	320	26723	24798
5	495	20874	21706
21	3165	11539	11424
13	3556	8248	8384
18	3950	33081	32103
11	3963	8954	8675
17	4373	8612	8377
10	6302	22186	20929

Hájek: Hombres y Mujeres en Montevideo

Ingresos de hombres y mujeres en Montevideo con el estimador Hájek

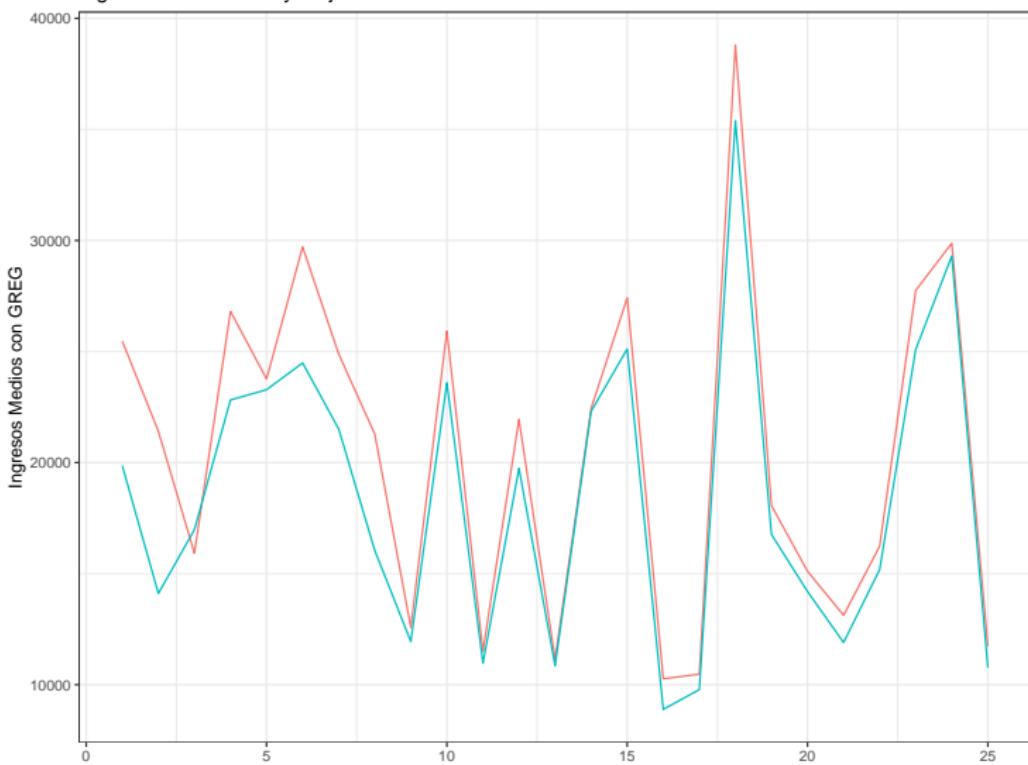


GREG: Hombres y Mujeres en Montevideo

sec2	ntotal	Hombres	Mujeres
2	121	21410	14107
1	167	25468	19861
3	186	15921	16981
4	319	26809	22819
6	320	29710	24484
5	495	23763	23282
21	3165	13125	11901
13	3556	11156	10862
18	3950	38789	35391
11	3963	11510	10977
17	4373	10473	9777
10	6302	25921	23589

GREG: Hombres y Mujeres en Montevideo

Ingresos de hombres y mujeres en Montevideo con el estimador GREG

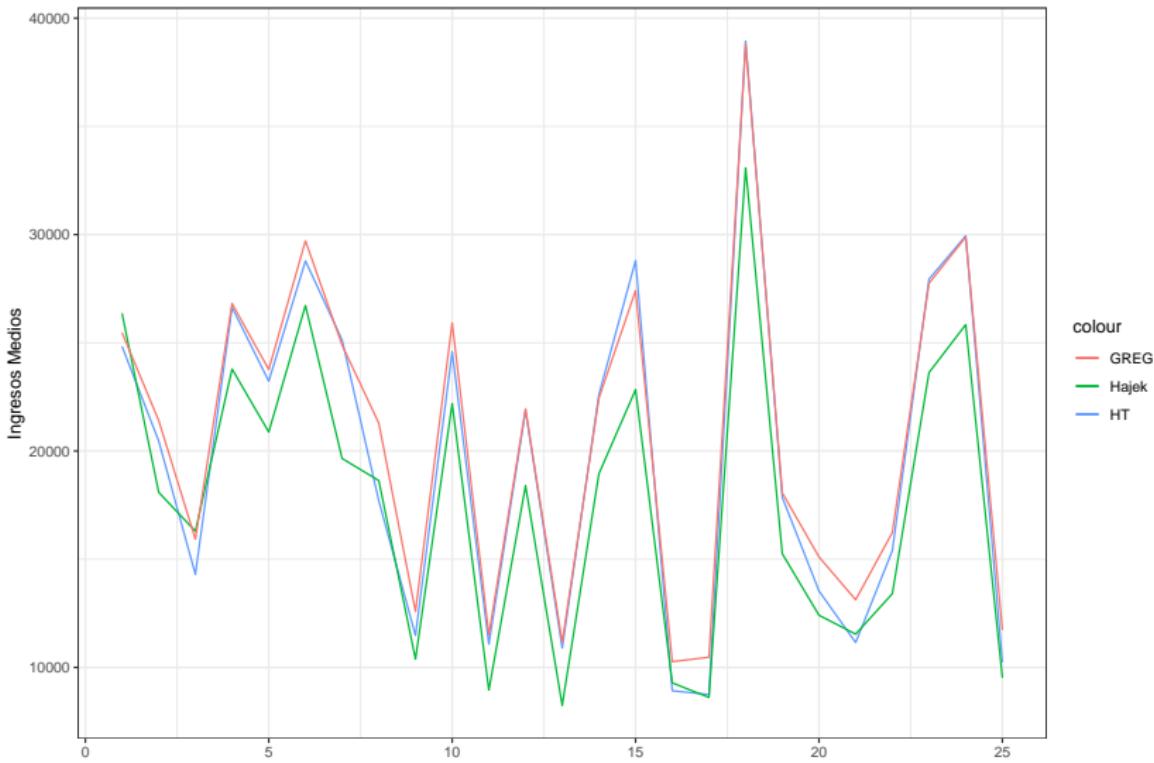


Comparando los estimadores: Hombres

sec2	ntotal	HT	Hajek	GREG
2	121	20461	18088	21410
1	167	24837	26363	25468
3	186	14299	16294	15921
4	319	26635	23786	26809
6	320	28784	26723	29710
5	495	23223	20874	23763
21	3165	11148	11539	13125
13	3556	10897	8248	11156
18	3950	38932	33081	38789
11	3963	11080	8954	11510
17	4373	8750	8612	10473
10	6302	24576	22186	25921

Comparando los estimadores: Hombres

Ingresos de hombres en Montevideo con estimadores directos

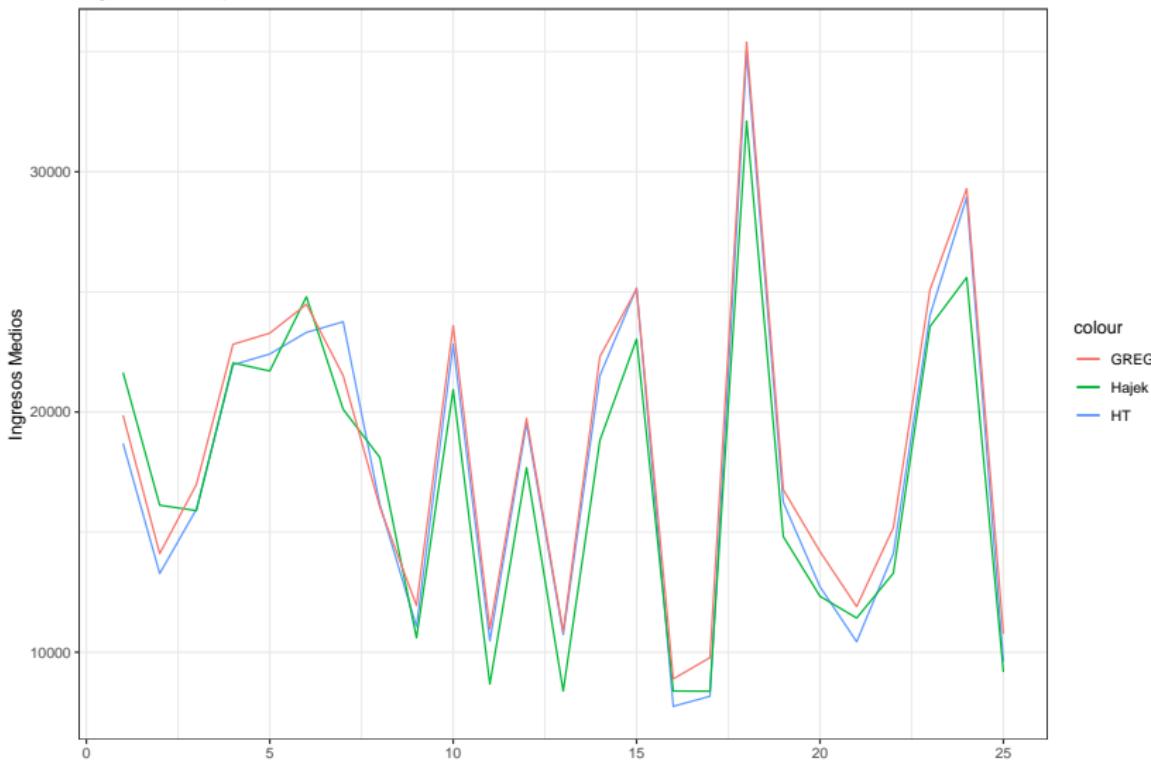


Comparando los estimadores: Mujeres

sec2	ntotal	HT	Hajek	GREG
2	121	13277	16120	14107
1	167	18694	21644	19861
3	186	15951	15896	16981
4	319	21965	22044	22819
6	320	23314	24798	24484
5	495	22414	21706	23282
21	3165	10435	11424	11901
13	3556	10742	8384	10862
18	3950	34943	32103	35391
11	3963	10473	8675	10977
17	4373	8167	8377	9777
10	6302	22823	20929	23589

Comparando los estimadores: Mujeres

Ingresa de mujeres en Montevideo con estimadores directos



¡Gracias!

¡Gracias!

*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

- 1 *Estimador post-estratificado sintético*
- 2 *Estimador sintético de regresión a nivel de área (REG1-SYN)*
- 3 *Estimador sintético de regresión a nivel de individuo (REG2-SYN)*
- 4 *Estimadores compuestos*
- 5 *Resultados: Estimación de ingreso medio en sectores de Montevideo*

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, \dots, U^J con tamaños poblacionales N^1, \dots, N^J

Estimador post-estratificado sintético

- El área d , U_d también es dividida en J grupos, llamados post-estratos, U_d^1, \dots, U_d^J de tamaño N_d^1, \dots, N_d^J .
- Tienen medias $\bar{Y}_d^1, \dots, \bar{Y}_d^J$, donde $\bar{Y}_d^j = \sum_{i \in U_d^j} Y_{di} / N_d^j$, $j = 1, \dots, J$.
- 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,

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

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 \bar{Y}_d así:

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

- 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,

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

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 $\hat{Y}^{j, HA}$ 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

$$F_{\alpha d}^j = F_\alpha^j, \quad j = 1, \dots, J$$

donde F_α^j 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), $N_d^j, j = 1, \dots, J$
 - 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, $\hat{\delta}_d, d = 1, \dots, D$
- El modelo se asume entonces,

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

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 $\text{var}(\hat{\delta}_d)$, $d = 1, \dots, D$.
- 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

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

- Donde

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

Estimador sintético de regresión a nivel de área

- Para α , el sesgo bajo el diseño de $\hat{\delta}_d^{REG1-SYN}$ viene dado por $\mathbf{x}'_d \alpha - \delta_d$.
- 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,
 $\hat{F}_{\alpha d} = \mathbf{x}'_d \boldsymbol{\alpha} + \varepsilon_d, \quad d = 1, \dots, D$ 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 $d = 1, \dots, D$, \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} , $i \in s_d$, $d = 1, \dots, D$.
- Se puede obtener por tanto un modelo lineal a nivel de individuo para Y_{di} .
- Llamamos $\mathbf{y}_d = (Y_{d1}, \dots, Y_{dN_d})$ 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,

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

- 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

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

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

Estimador sintético de regresión a nivel de individuo

- El vector de predicciones para el área d es entonces
 $\hat{\mathbf{y}}_d = (\hat{Y}_{d1}, \dots, \hat{Y}_{dN_d})'$ donde $\hat{Y}_{di} = \mathbf{x}'_{di}\hat{\beta}$, $i = 1, \dots, N_d$
- 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 , $\delta_d = \bar{Y}_d$, si $\bar{\mathbf{X}}_d$ es el vector de medias poblacionales de las p variables auxiliares, $\hat{\bar{Y}}_d^{REG2-SYN}$ sería

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

- Se obtiene el estimador para un área no muestreada de la misma forma.
- Para β conocido, el sesgo bajo el diseño de la media es $\bar{\mathbf{X}}_d' \beta - \bar{Y}_d$, 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

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

- 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, $\bar{\mathbf{X}}_d$, $d = 1, \dots, D$

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”.

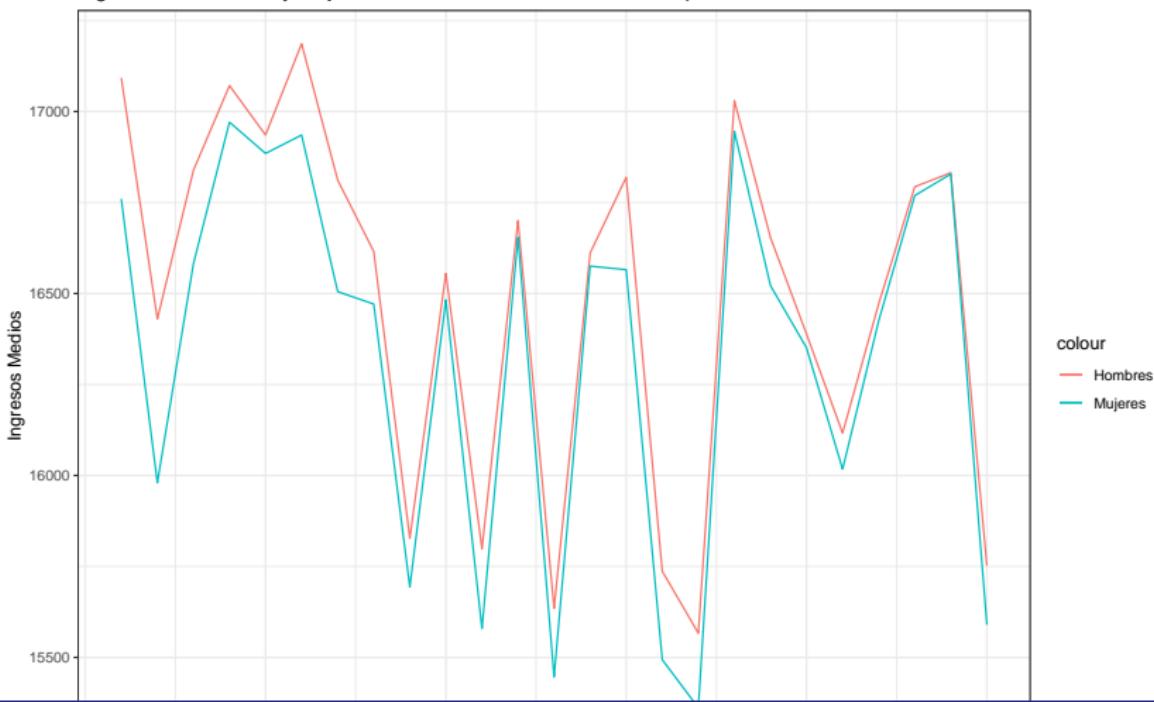
Resultados: Estimación de ingreso medio en sectores de Montevideo

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

Ingresos de hombres y mujeres en Montevideo con el estimador post-estratificado sintético

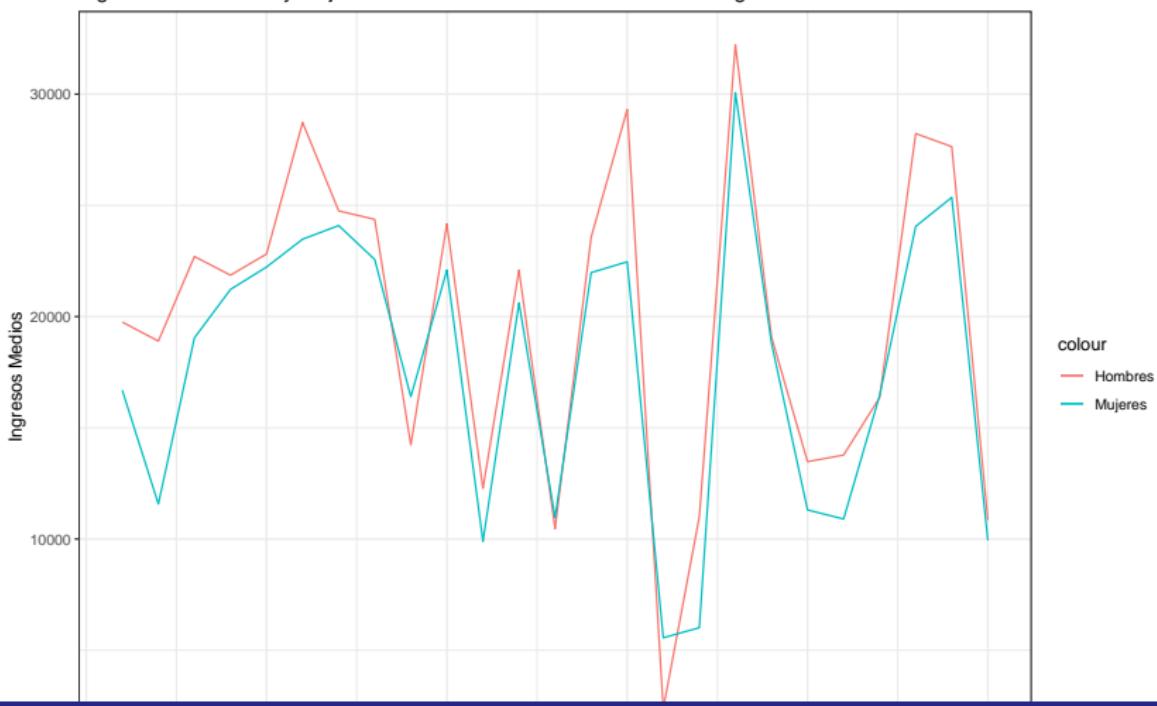


Estimador de regresión sintético a nivel de área: Hombres y Mujeres en Montevideo

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

Ingresos de hombres y mujeres en Montevideo con el estimador de regresión sintético a nivel de área

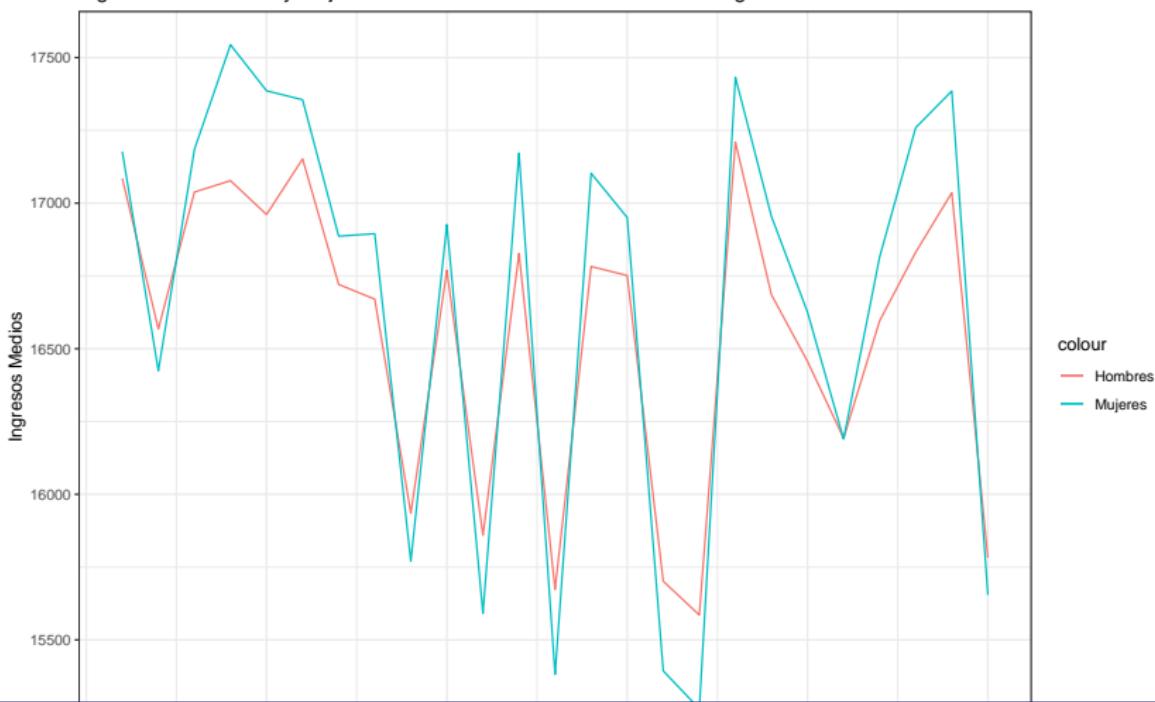


*Estimador de regresión sintético a nivel de individuo:
Hombres y Mujeres en Montevideo*

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

Ingresos de hombres y mujeres en Montevideo con el estimador de regresión sintético a nivel de individuo

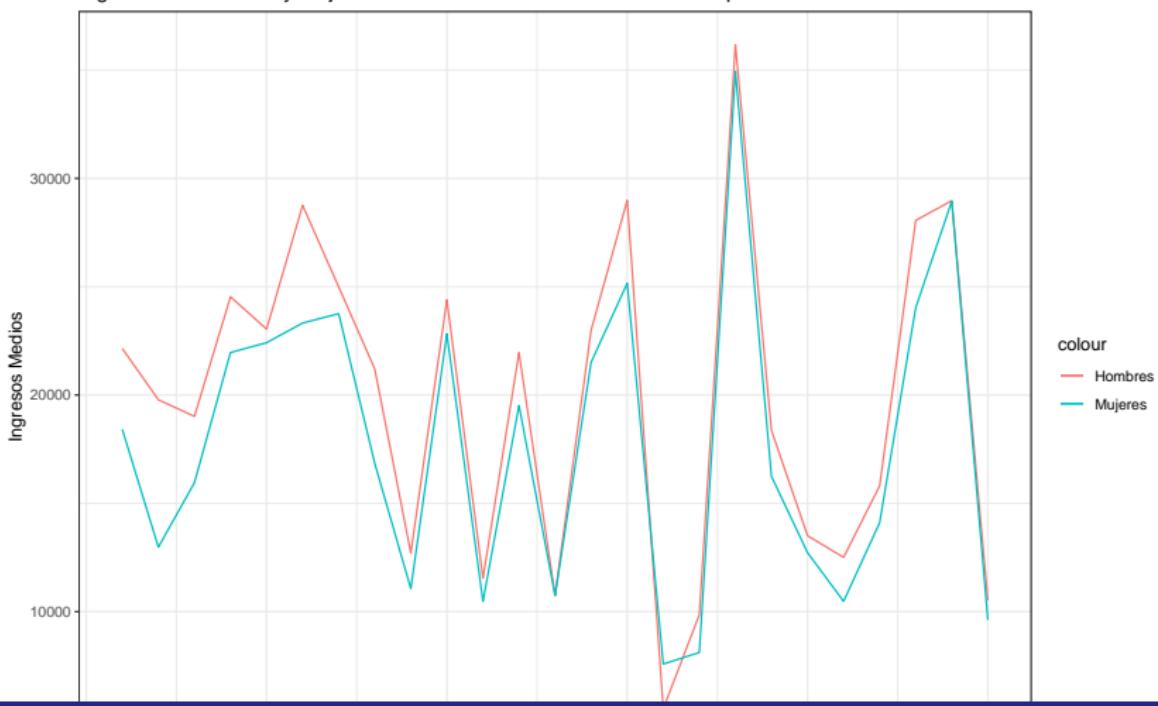


Estimador compuesto: Hombres y Mujeres en Montevideo

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

Ingresos de hombres y mujeres en Montevideo con el estimador compuesto

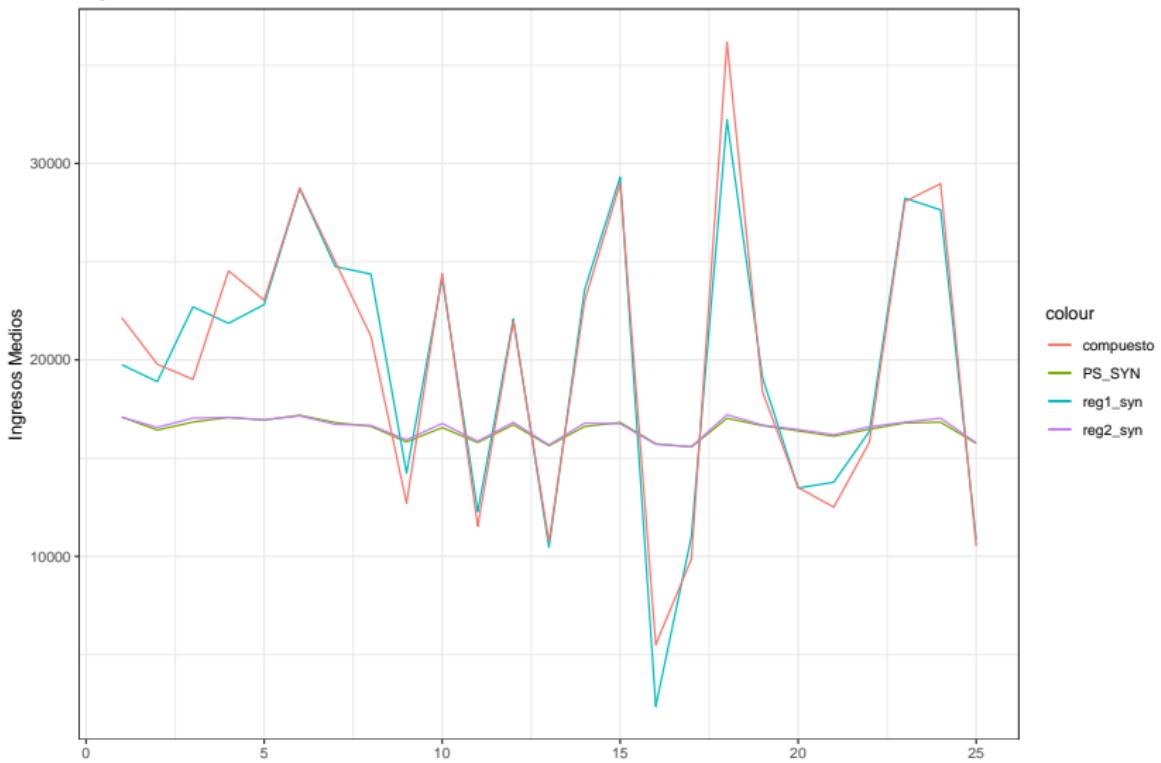


Comparando los estimadores: Hombres

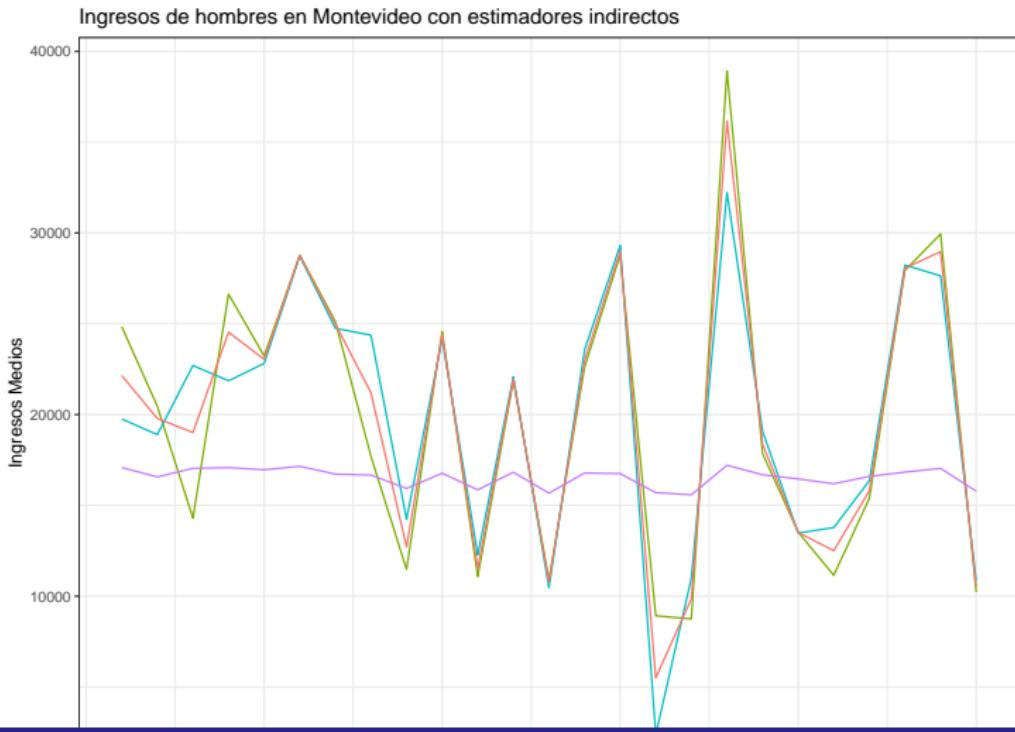
sec2	ntotal	HT	PS_SYN	reg1_syn	reg2_syn	compuesto
2	121	20461	16430	18896	16568	19781
1	167	24837	17093	19756	17085	22149
3	186	14299	16838	22701	17038	19015
4	319	26635	17071	21862	17077	24534
6	320	28784	17186	28726	17152	28757
5	495	23223	16935	22816	16961	23042
21	3165	11148	16117	13776	16193	12506
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

Ingresa de hombres en Montevideo con estimadores indirectos



Comparando los estimadores: Hombres, usando HT para referencia

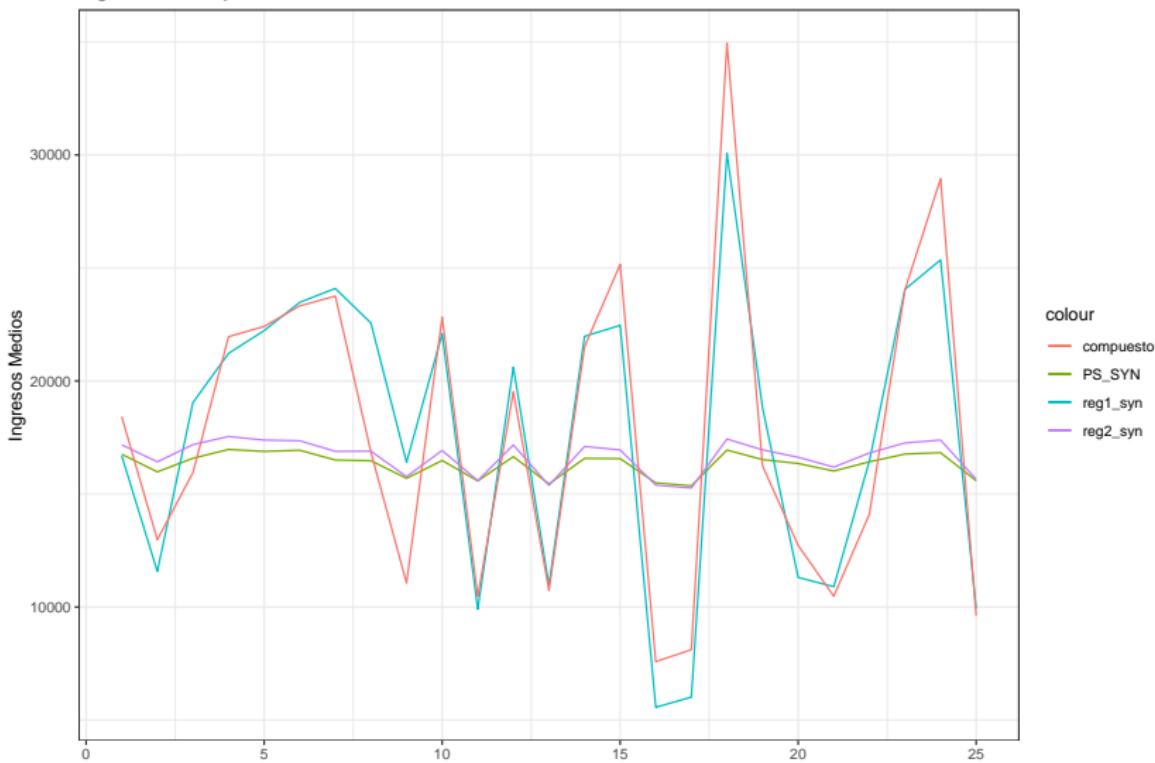


Comparando los estimadores: Mujeres

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

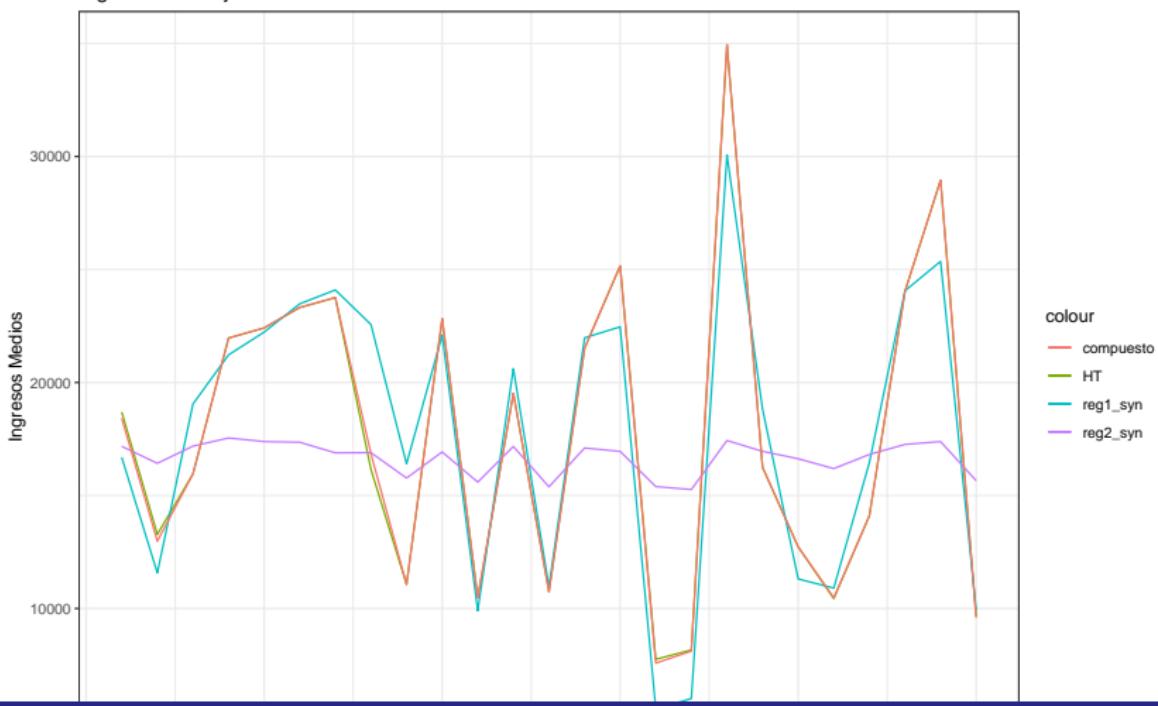
Comparando los estimadores: Mujeres

Ingresa de mujeres en Montevideo con estimadores indirectos



Comparando los estimadores: Mujeres, usando HT para referencia

Ingresos de mujeres en Montevideo con estimadores indirectos



¡Gracias!

¡Gracias!

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

- 1 BLUP/EBLUP basado en el modelo Fay-Herriot
- 2 Resultados: Estimación de ingreso medio en sectores de Montevideo

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 , $d = 1, \dots, D$, asumiendo que varían respecto a un vector de p covariables, \mathbf{x}_d , de forma constante.
- Viene dado por

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

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

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

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 muestreo.
- $\hat{\delta}_d^{DIR}$ todavía se considera insesgado bajo el diseño muestral.

BLUP/EBLUP basado en el modelo Fay-Herriot

- Podemos definir, entonces,

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

donde e_d es el error debido al muestreo, $e_d \stackrel{ind}{\sim} (0, \psi_d)$.

- Dichas varianzas $\psi_d = \text{var}_{\pi}(\hat{\delta}_d^{DIR} | \delta_d)$, $d = 1, \dots, D$, se estiman con los microdatos de la encuesta.
- Por tanto, el modelo se hace,

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

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,

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

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

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

BLUP/EBLUP basado en el modelo Fay-Herriot

- Note que

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

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 ,

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

donde

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

y

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

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 β 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:

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

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

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

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

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

y

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

Resumen del estimador FH

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

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”.

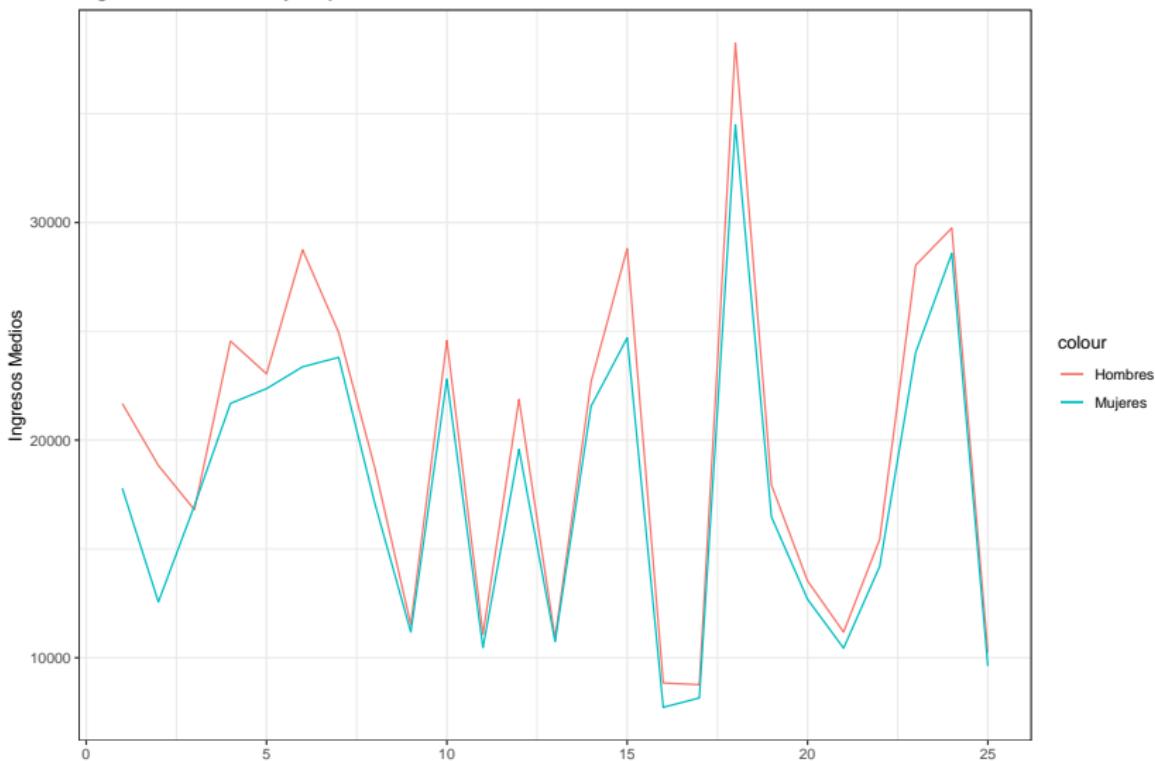
Resultados: Estimación de ingreso medio en sectores de Montevideo

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

Ingresaos de hombres y mujeres en Montevideo con el estimador FH

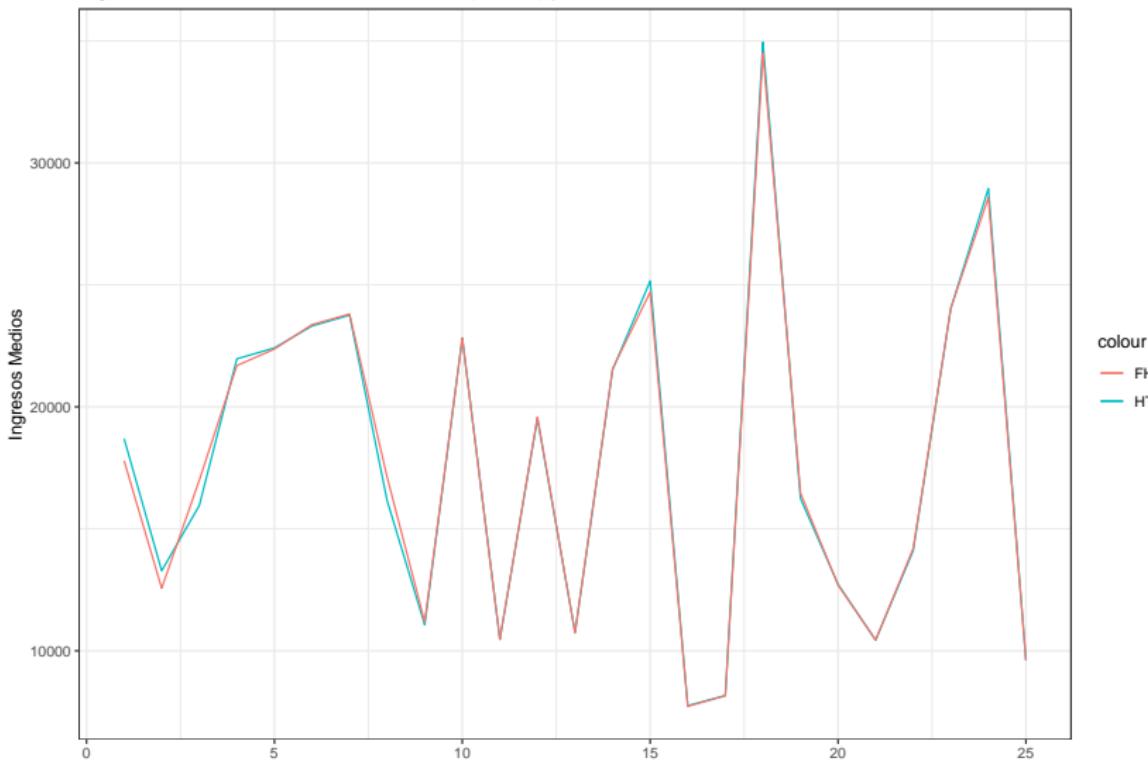


Comparando los estimadores: Hombres

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

Ingresaos de hombres en Montevideo: HT (directo) y FH

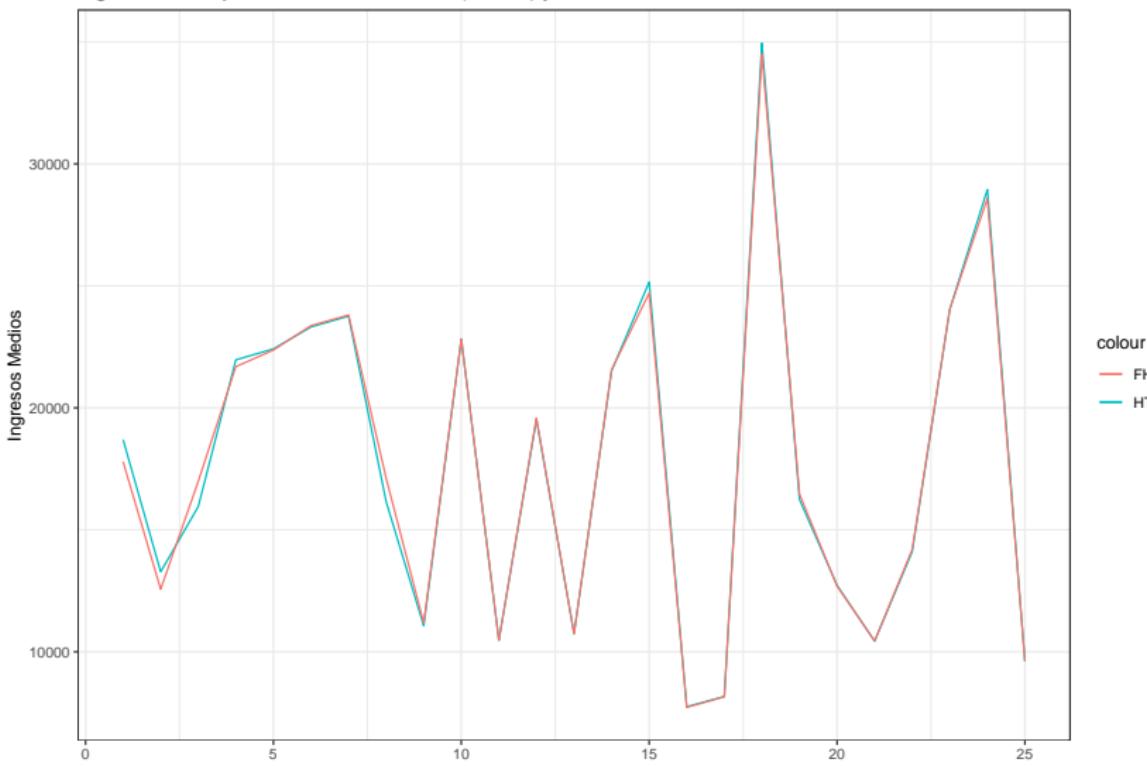


Comparando los estimadores: Mujeres

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

Ingresos de mujeres en Montevideo: HT (directo) y FH



¡Gracias!

¡Gracias!

Curso Internacional de Desagregación de Estimaciones en Áreas Pequeñas usando R

*Métodos indirectos con modelos de unidad: EBLUP basado en
el modelo BHF y el método ELL*

División de Estadísticas
Comisión Económica para América Latina y el Caribe

2020

- 1 BLUP/EBLUP basado en el modelos con errores anidados (BHF)
- 2 Método ELL
- 3 Resultados: Estimación de ingreso medio en sectores de Montevideo

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

- 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.
- 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

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

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

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

y

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

- 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:

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

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

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

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

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

- $\bar{y}_{da} = a_{d\cdot}^{-1} \sum_{i \in s_d} a_{di} Y_{di}$ 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 $\mathbf{y}_d = (Y_{d1}, \dots, Y_{dN_d})'$ un vector de variables respuestas para el área d y $\mathbf{X}_d = (\mathbf{x}_{d1}, \dots, \mathbf{x}_{dN_d})'$ la matriz de covariables en el área d .
- Bajo el modelo de errores anidados, $\mathbf{y}_d \stackrel{ind}{\sim} N(\mathbf{X}_d\boldsymbol{\beta}, \mathbf{V}_d)$, $d = 1, \dots, D$, donde

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

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

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.

- 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 \bar{Y}_d se puede escribir como

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

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

- 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

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

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, $\theta = (\sigma_u^2, \sigma_e^2)'$ con $\hat{\theta} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$, obtenemos el estimador EBIUP:

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

donde

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

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 θ por $\tilde{\theta}$ en $\tilde{\beta}$.
- Donde

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

y

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

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 $\bar{\mathbf{X}}'_d \hat{\beta}$.
- 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{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 $u_d^{*(b)} \stackrel{iid}{\sim} N(0, \hat{\sigma}_u^2)$, $d = 1, \dots, D$

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

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

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

Nótese que este cálculo no requiere los valores individuales de \mathbf{x}_{di} 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

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

- 6) Para la muestra original $s = s_1 \cup \dots \cup s_D$, sea $\mathbf{y}_s^{*(b)} = ((\mathbf{y}_{1s}^{*(b)})', \dots, (\mathbf{y}_{Ds}^{*(b)})')'$ 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 $\hat{Y}_d^{EBLUP*(b)}$, $d = 1, \dots, D$.

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

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

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

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
- 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, $\bar{\mathbf{X}}_d$.

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:

- 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 muestradas.

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:
 - 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.
- 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

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

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

y

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

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 $d = 1, \dots, D$, y errores e_{di}^* , para cada individuo $i = 1, \dots, N_d$, $d = 1, \dots, D$
- 2) Se generan valores bootstrap de la variable respuesta

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

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.
- Obtenemos el indicador de $F_{\alpha d}$ con los valores Y_{di}^* generados para cada censo a de la forma

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

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

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

Método ELL

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

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

- A lo largo de las réplicas bootstrap,

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

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

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

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

Resumen del método ELL

- Indicadores objetivos: Parámetros generales
- 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:

- 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:
 - 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.

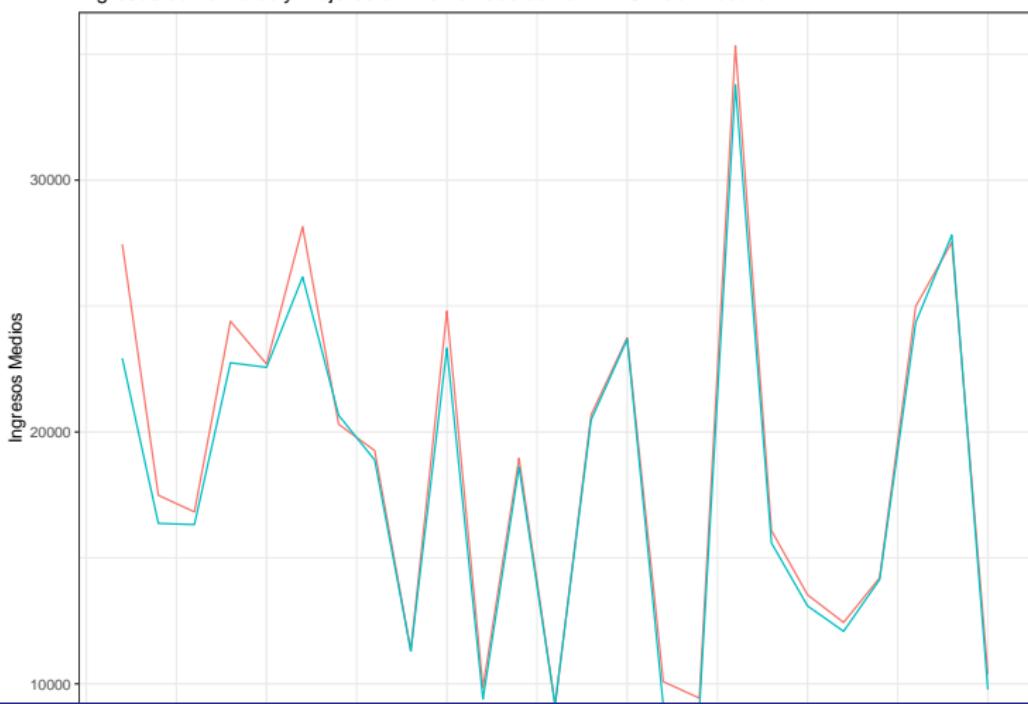
Resultados: Estimación de ingreso medio en sectores de Montevideo

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

Ingresos de hombres y mujeres en Montevideo con el EBLUP del modelo BHF

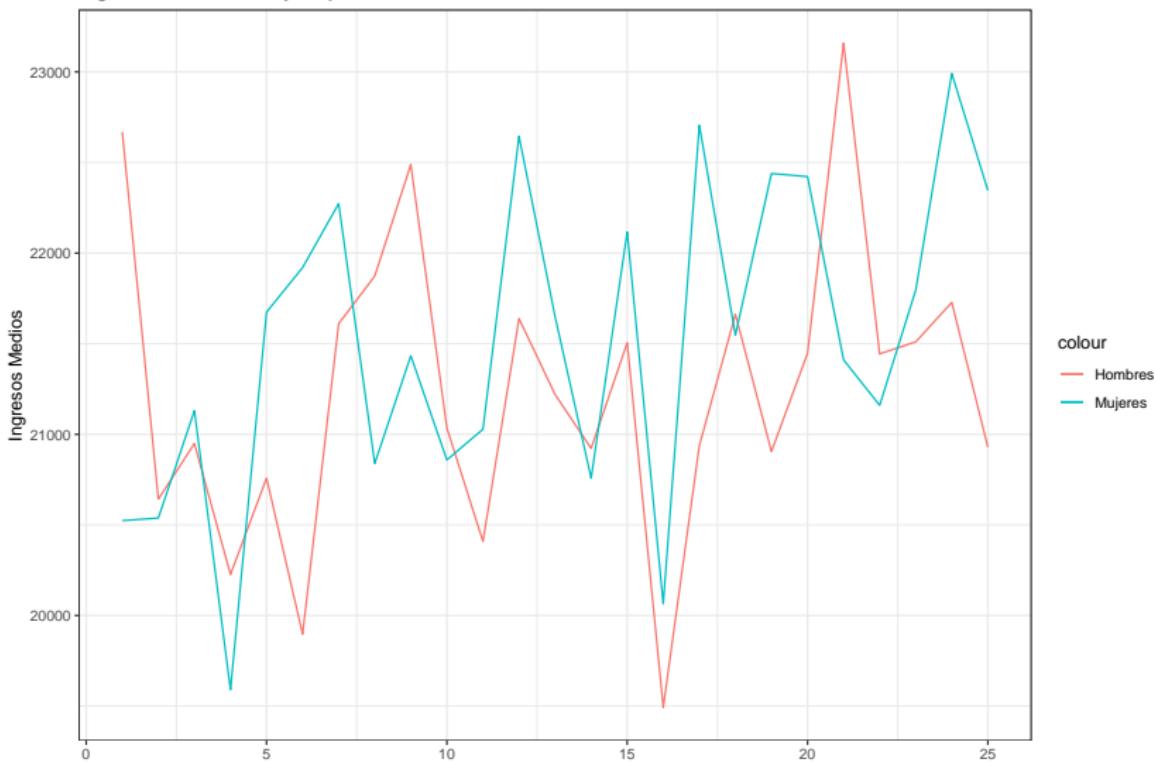


Método ELL: Hombres y Mujeres en Montevideo

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

Ingresos de hombres y mujeres en Montevideo con el método ELL

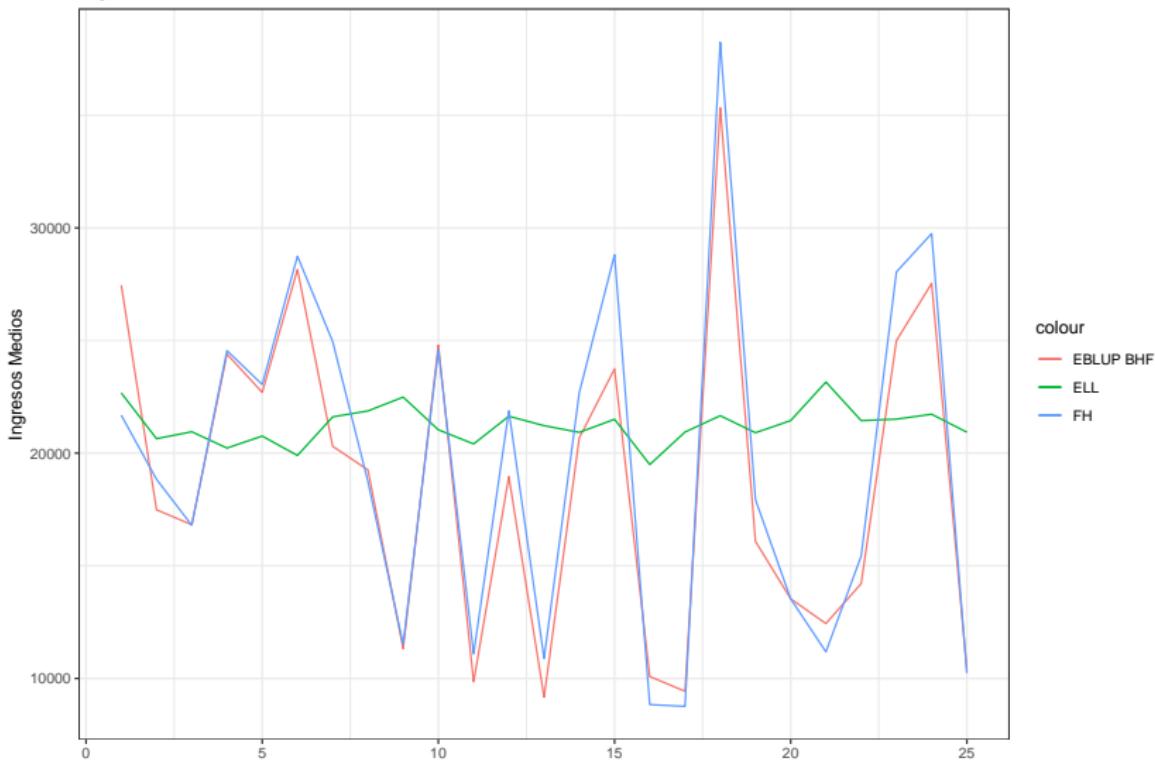


Comparando los estimadores: Hombres

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

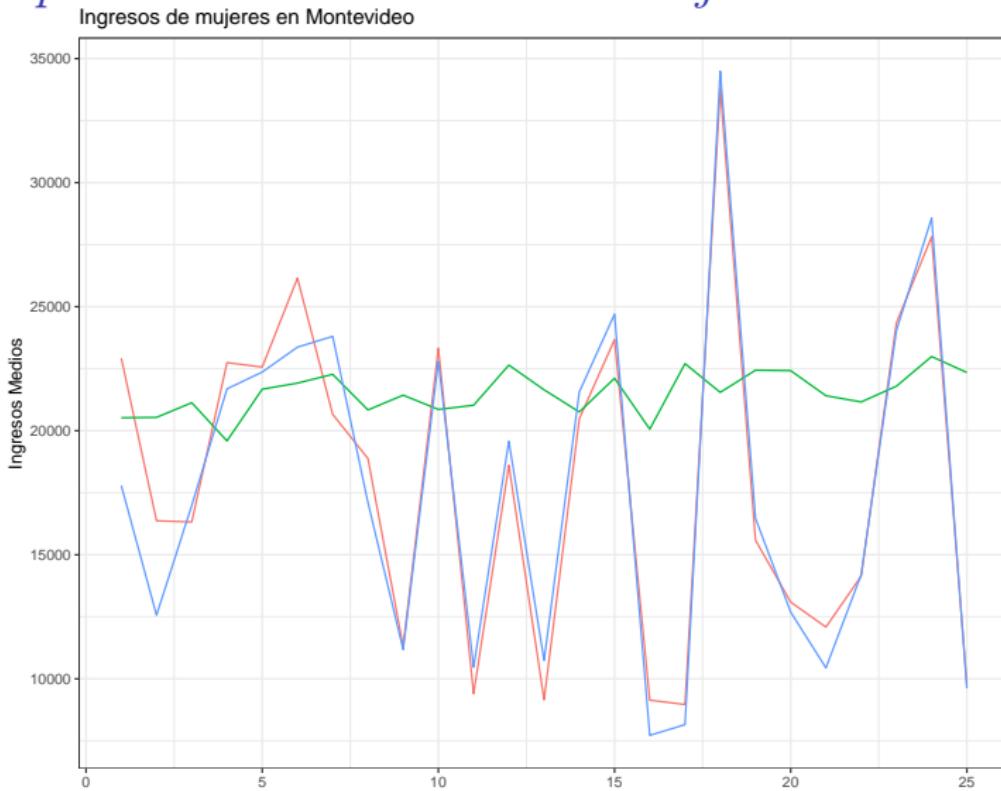
Ingresaos de hombres en Montevideo



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!

¡Gracias!

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

- 1 Mejor predictor empírico (EB, empirical best) bajo el modelo con errores anidados
- 2 Resultados: Estimación de ingreso medio en sectores de Montevideo

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, 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 colecciónadas
- 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

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

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 $\mathbf{y}_d = (Y_{d1}, \dots, Y_{dN_d})'$, $d = 1, \dots, D$, son independientes y verifican

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

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

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

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

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

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

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 $i \in r_d$:

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

donde la media y la varianza condicionadas vienen dadas por

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

y

$$\sigma_{di|s}^2 = \sigma_u^2 (1 - \gamma_d) + \sigma_e^2 k_{di}^2$$

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,

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

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

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

- Aquí,

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

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

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

y

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

- 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 $\hat{\theta} = (\hat{\beta}', \hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ para los verdaderos parámetros ajustando del modelo de errores anidados a los datos $(\mathbf{y}_s, \mathbf{X}_s)$
- 2) Generar $a = 1, \dots, A$ 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 $\mathbf{y}_d^{(a)} = (\mathbf{y}'_{ds}, (\mathbf{y}'_{dr})')'$ 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

$\mathbf{y}_s = (\mathbf{y}'_{1s}, \dots, \mathbf{y}'_{Ds})'$ para obtener estimaciones de los parámetros del modelo, $\hat{\beta}$, $\hat{\sigma}_u^2$ y $\hat{\sigma}_e^2$

- 2) Generar efectos bootstrap para las áreas con

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

- 3) Generar errores bootstrap para los individuos con

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

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

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

- 5) Calcular los indicadores de interés, $\delta_d^{*(b)} = \delta_d(\mathbf{y}_d^{*(b)})$,
 $d = 1, \dots, D$, donde $\mathbf{y}_d^{*(b)} = (Y_{d1}^{*(b)}, \dots, Y_{dN_d}^{*(b)})'$
- 6) Sea $\mathbf{y}_s^{*(b)} = ((\mathbf{y}_{1s}^{*(b)})', \dots, (\mathbf{y}_{Ds}^{*(b)})')'$ 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 $\hat{\delta}_d^{EB*(b)}$, $d = 1, \dots, D$

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:

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

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 $\{(E_{di}, \mathbf{x}_{di}); i \in s_d, d = 1, \dots, D\}$.
- Porque se necesita datos de las mismas variables auxiliares para todas las áreas, 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 $\tilde{F}_{\alpha,di}^B(\theta) = F_{\alpha,di}^B(\theta) = E[F_{\alpha,di}|\mathbf{y}_{ds}; \theta]$, también para los individuos de la muestra como si no se observaran.
- el predictor Census best de $F_{\alpha d}$ viene dado por

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

- 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_{di}
- 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)

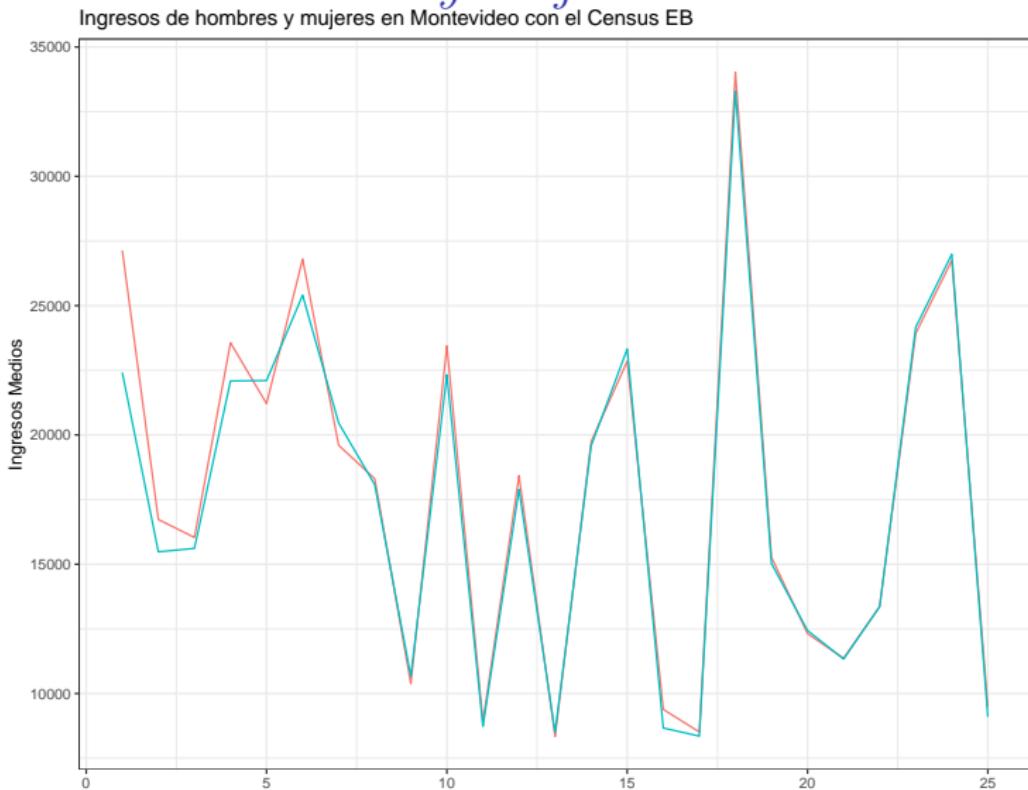
- Desventajas:
 - Estimadores de ECM usando el bootstrap son computacionalmente intensivos:
 - Para la aproximación Monte-Carlo del estimador EB, se necesita A censos $y^{(a)}$ de grande tamaño
 - Para el estimar el ECM a través del proceso bootstrap requiere que se repita la aproximación Monte Carlo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

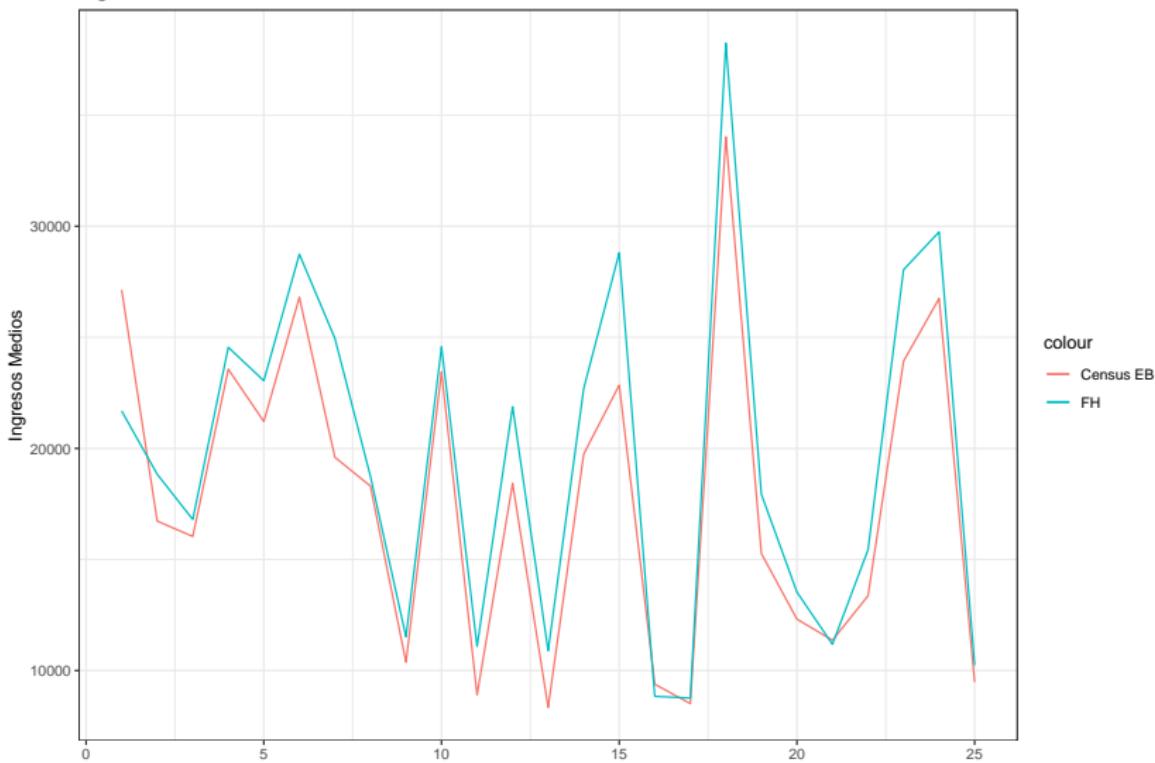


Comparando los estimadores: Hombres

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

Ingresaos de hombres en Montevideo

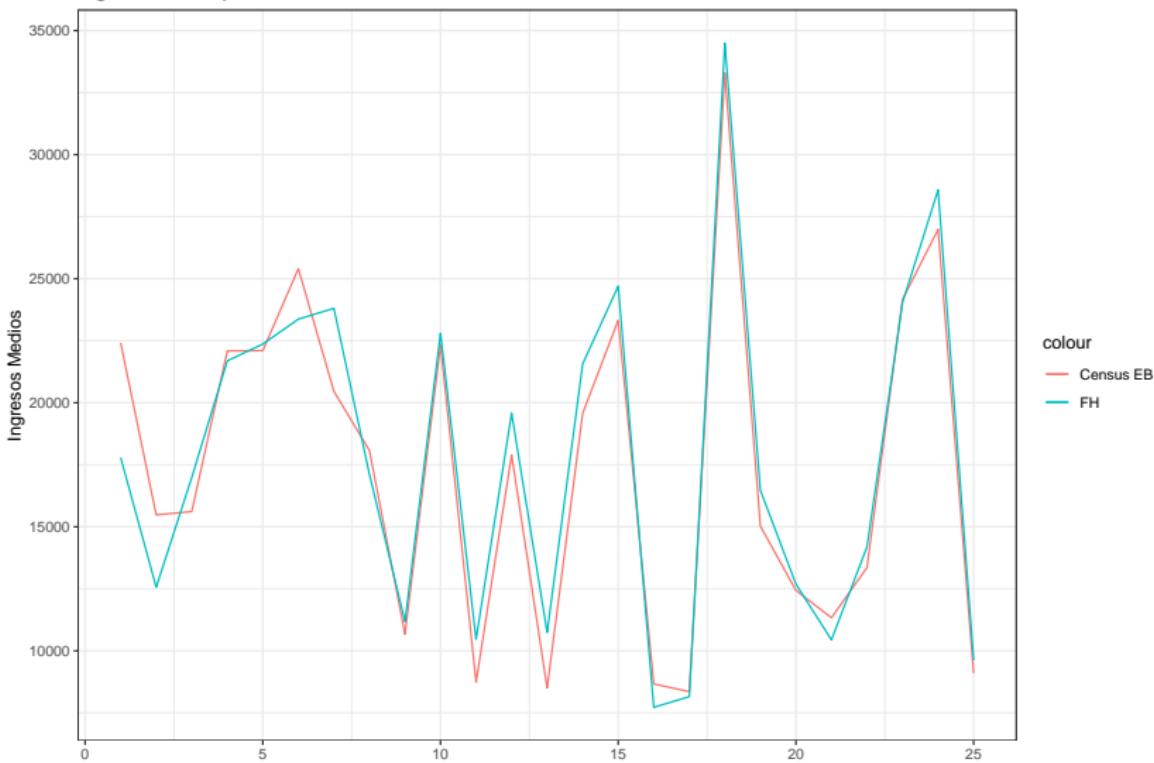


Comparando los estimadores: Mujeres

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

Ingresos de mujeres en Montevideo



¡Gracias!

¡Gracias!

Curso Internacional de Desagregación de Estimaciones en Áreas Pequeñas usando R

*Método jerárquico de Bayes y método basado en modelos
generalizados lineales mixtos*

División de Estadísticas
Comisión Económica para América Latina y el Caribe

2020

1 Método jerárquico de Bayes

2 Métodos basados en modelos lineales generalizados mixtos

3 Aplicación

References

- (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

- De nuevo, 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 colecciónadas
- Esto se realiza incorporando efectos aleatorios de las áreas en los modelos de interés

Método jerárquico de Bayes

Método jerárquico Bayes bajo el modelo BHF

- Molina, Nandrum, y Rao (2014) propusieron el método jerárquico Bayes (*hierarchical Bayes, HB*), que no requiere el uso del bootstrap
- El método reparametriza el modelo de errores anidados en términos del coeficiente de correlación intraclase
$$\rho = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$$
- Considera distribuciones *a priori* para los parámetros $(\beta, \rho, \sigma_e^2)$

Método jerárquico Bayes bajo el modelo BHF

El modelo HB viene dado por:

1)

$$Y_{di} | u_d, \beta, \sigma_e^2 \stackrel{ind}{\sim} N(\mathbf{x}'_{di}\beta + u_d, \sigma_e^2 k_{di}^2), \quad i = 1, \dots, N_d,$$

2)

$$u_d | \rho, \sigma_e^2 \stackrel{iid}{\sim} N\left(0, \frac{\rho}{1-\rho} \sigma_e^2\right), \quad d = 1, \dots, D,$$

3)

$$\pi(\beta, \rho, \sigma_e^2) \propto \frac{1}{\sigma_e^2}, \quad \epsilon \leq \rho \leq 1 - \epsilon, \sigma_e^2 > 0, \beta \in \mathbb{R}^p$$

donde $\epsilon > 0$ refleja la falta de información

Método jerárquico Bayes bajo el modelo BHF

- La distribución a priori de los parámetros del modelo se puede calcular en función de las distribuciones condicionadas usando la regla de cadena
- La densidad conjunta del parámetros $\theta = (\mathbf{u}', \beta', \sigma_e^2, \rho)'$ viene dada por
- $$\pi(\mathbf{u}, \beta, \sigma_e^2, \rho | \mathbf{y}_s) = \pi_1(\mathbf{u} | \beta, \sigma_e^2, \rho, \mathbf{y}_s) \pi_2(\beta | \sigma_e^2, \rho, \mathbf{y}_s) \pi_3(\sigma_e^2 | \rho, \mathbf{y}_s) \pi_4(\rho | \mathbf{y}_s)$$
- Todas las distribuciones tienen forma conocida excepto π_4 , y generamos esos valores usando un método de rejilla

Método jerárquico Bayes bajo el modelo BHF

- Así, se pueden generar muestras de $\boldsymbol{\theta} = (\mathbf{u}', \boldsymbol{\beta}', \sigma_e^2, \rho)'$ directamente de la distribución a posteriori
- Dado $\boldsymbol{\theta}$, las variables Y_{di} para todos los individuos verifican

$$Y_{di} | \boldsymbol{\theta} \stackrel{ind}{\sim} N(\mathbf{x}'_{di}\boldsymbol{\beta} + u_d, \sigma_e^2 k_{di}^2), \quad i = 1, \dots, N_d, d = 1, \dots, D$$

- La densidad predictiva de \mathbf{y}_{dr} viene dada por

$$f(\mathbf{y}_{dr} | \mathbf{y}_s) = \int \prod_{i \in r_d} f(Y_{di} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta} | \mathbf{y}_s) d\boldsymbol{\theta},$$

Método jerárquico Bayes bajo el modelo BHF

- Finalmente, el estimador HB viene dado por

$$\hat{\delta}_d^{HB} = E_{\mathbf{y}_{dr}}(\delta_d | \mathbf{y}_s) = \int \delta_d(\mathbf{y}_d) f(\mathbf{y}_{dr} | \mathbf{y}_s) d\mathbf{y}_{dr}$$

lo que estimamos usando una simulación Monte Carlo

- Generamos muestras de la distribución a posteriori $\pi(\theta | \mathbf{y}_s)$

Método jerárquico Bayes bajo el modelo BHF: proceso Monte Carlo

- Primero, generamos un valor $\rho^{(a)}$ de $\pi_4(\rho|\mathbf{y}_s)$ con un método de Rejilla
- Despues, generamos $\sigma_e^{2(a)}$ de $\pi_3(\sigma_e^2|\rho^{(a)}, \mathbf{y}_s)$, $\beta^{(a)}$ de $\pi_2(\beta|\sigma_e^{2(a)}, \rho^{(a)}, \mathbf{y}_s)$ y $\mathbf{u}^{(a)}$ de $\pi_1(\mathbf{u}|\beta^{(a)}, \sigma_e^{2(a)}, \rho^{(a)}, \mathbf{y}_s)$
- Para cada de los A valores del vector θ , generamos los valores de los individuos afuera de la muestra $\mathbf{y}_{dr}^{(a)}$, y creamos el vector censal $\mathbf{y}_d^{(a)} = (\mathbf{y}'_{ds}, (\mathbf{y}_{dr}^{(a)})')'$

Método jerárquico Bayes bajo el modelo BHF:

- Para cada vector censal, producimos $\delta_d^{(a)} = \delta_d(\mathbf{y}_d^{(a)})$
- Se aproxima El estimador HB por

$$\hat{\delta}_d^{HB} = E_{\mathbf{y}_{dr}}(\delta_d | \mathbf{y}_s) \approx \frac{1}{A} \sum_{a=1}^A \delta_d^{(a)}$$

- La varianza se aproxima por

$$V(\delta_d | \mathbf{y}_s) \approx \frac{1}{A} \sum_{a=1}^A \left(\delta_d^{(a)} - \hat{\delta}_d^{HB} \right)^2$$

- Para un indicador FGT, el estimador HB se aproxima en la forma:

$$\hat{F}_{\alpha d}^{HB} \approx \frac{1}{A} \sum_{a=1}^A F_{\alpha d}^{(a)}$$

Resumen del estimador del método HB:

- Indicadores objetivos: Parámetros generales
- Requerimientos de datos:
 - Microdatos de las p variables auxiliares de la misma de la variable de interés
 - Área de interés obtenida de la misma encuesta
 - Microdatos de las p covariables a partir de un censo o registro administrativo

Resumen del estimador del método HB:

- Ventajas:
 - Basado en datos a nivel de individuo, lo que provee información más detallada
 - Se puede estimar cualquier indicador que es una función de la variable Y_{di}
 - Es insesgado bajo el modelo si los parámetros son conocidos
 - Minimiza la varianza a posteriori

Resumen del estimador del método HB:

- Ventajas:

- Resulta prácticamente igual al estimador EB
- Una vez se ajusta el modelo, se puede estimar en subáreas sin reajustar el modelo
- Una vez se ajusta el modelo, se puede estimar cualquier indicador sin reajustar el modelo

Resumen del estimador del método HB:

- Ventajas:

- No se usa el procedimiento *Markov Chain Monte Carlo*, MCMC, al contrario de muchos procesos bayesianos
- No requiere el uso de métodos bootstrap para estimar ECM, lo que disminuye el tiempo computacional
- El cálculo de intervalos creíbles o cualquier otro resumen de la distribución es automático

Resumen del estimador del método HB:

- Desventajas:
 - Es basado en un modelo, por tanto es necesario comprobar dicho modelo
 - No tiene en cuenta el diseño muestral
 - Puede ser seriamente afectado por atípicos aislados
 - El método no se puede extender a un modelo más complejos sin perder algunas ventajas mencionadas, como el evitar los métodos MCMC por ejemplo

Métodos basados en modelos lineales generalizados mixtos

Métodos basados en modelos lineales generalizados mixtos

- Los modelos mixtos hasta ahora no dan predicciones entre [0, 1]
- Para estimar proporciones, sería útil usar un modelo que proporciona valores en ese rango
- Esto incluye la incidencia de pobreza F_{0d} , pero no la brecha de pobreza, F_{1d}
- Para hacer eso, es habitual usar un modelo lineal generalizado mixto (*generalized linear mixed models, GLMM*)

Métodos basados en modelos lineales generalizados mixtos

- Asumimos que

$$Y_{di} | v_d \sim \text{Bern}(p_{di}), \quad g(p_{di}) = \mathbf{x}'_{di} \boldsymbol{\alpha} + v_d$$

y

$$v_d \stackrel{iid}{\sim} N(0, \sigma_v^2), \quad i = 1, \dots, N_d, \quad d = 1, \dots, D$$

- v_d es el efecto de área d y $\boldsymbol{\alpha}$ es el vector de coeficientes de la regresión
- $g : (0, 1) \rightarrow \mathbb{R}$ es la función link (biyectiva, con derivada continua)
- El link logístico, $g(p) = \log(p/(1-p))$, el más utilizado

Métodos basados en modelos lineales generalizados mixtos: Método mejor predictor

- El mejor predictor, el que minimiza el ECM bajo el modelo, viene dado por

$$\tilde{P}_d^B(\theta) = E(P_d | \mathbf{y}_{ds}; \theta) = \frac{1}{N_d} \left\{ \sum_{i \in s_d} Y_{di} + \sum_{i \in r_d} E(Y_{di} | \mathbf{y}_{ds}; \theta) \right\}$$

- En la práctica, obtenemos el predictor EB reemplazando θ por estimaciones consistentes, es decir

$$\hat{P}_d^{EB} = \tilde{P}_d^B(\hat{\theta})$$

donde $\hat{\theta}$ se encuentra ajustando el modelo GLMM a los datos muestrales $\mathbf{y}_s = (\mathbf{y}'_{1s}, \dots, \mathbf{y}'_{Ds})'$

Métodos basados en modelos lineales generalizados mixtos: Método mejor predictor

- Una manera de estimar $E(Y_{di}|\mathbf{y}_{ds}; \boldsymbol{\theta})$ sería utilizar el Teorema de Bayes y que las variables Y_{di} son independientes dado v_d
- En este caso,

$$E(Y_{di}|\mathbf{y}_{ds}; \hat{\boldsymbol{\theta}}) = \frac{E\{h(\mathbf{x}'_{di}\boldsymbol{\alpha} + v_d)f(\mathbf{y}_{ds}|v_d); \hat{\boldsymbol{\theta}}\}}{E\{f(\mathbf{y}_{ds}|v_d); \hat{\boldsymbol{\theta}}\}}, \quad i \in r_d,$$

donde $h = g^{-1}$ es el link inverso

$h(\mathbf{x}'_{di}\boldsymbol{\alpha} + v_d) = \exp(\mathbf{x}'_{di}\boldsymbol{\alpha} + v_d) / \{1 + \exp(\mathbf{x}'_{di}\boldsymbol{\alpha} + v_d)\}$ y

$$f(\mathbf{y}_{ds}|v_d) = \prod_{i \in s_d} p_{di}^{Y_{di}} (1 - p_{di})^{(1 - Y_{di})}$$

Métodos basados en modelos lineales generalizados mixtos: Método mejor predictor

- Podemos usar un proceso Monte Carlo para generar $v_d^{(r)} \sim N(0, \hat{\sigma}_v^2)$, $r = 1, \dots, R$
- Después, calculamos

$$E(Y_{di} | \mathbf{y}_{ds}; \hat{\theta}) \approx \frac{R^{-1} \sum_{r=1}^R h(\mathbf{x}'_{di} \hat{\alpha} + v_d^{(r)}) \hat{f}(\mathbf{y}_{ds} | v_d^{(r)})}{R^{-1} \sum_{r=1}^R \hat{f}(\mathbf{y}_{ds} | v_d^{(r)})}, \quad i \in r_d,$$

donde \hat{f} es la distribución condicionada $f(\mathbf{y}_{ds} | v_d)$ en $\hat{\theta}$

Métodos basados en modelos lineales generalizados mixtos: Método mejor predictor



$$\tilde{P}_d^B(\boldsymbol{\theta}) = E(P_d | \mathbf{y}_{ds}; \boldsymbol{\theta}) = \frac{1}{N_d} \left\{ \sum_{i \in s_d} Y_{di} + \sum_{i \in r_d} E(Y_{di} | \mathbf{y}_{ds}; \boldsymbol{\theta}) \right\}$$

tiene ECM mínimo y es insesgado bajo el modelo linear generalizado mixto

- Sin embargo, el proceso que se ha descrito es computacionalmente intensivo debido a las réplicas Monte Carlo

Métodos basados en modelos lineales generalizados mixtos: método plug-in

- Existen estimadores que se obtienen directamente de la salida del software que estima el GLMM
- Cuando se hace la regresión, el software estima $\hat{\alpha}$ y \hat{v}_d
- Se puede crear un estimador por el método de la analogía (*plug-in estimator*) que viene dado por:

$$\hat{P}_d^{PI} = \frac{1}{N_d} \left(\sum_{i \in s_d} Y_{di} + \sum_{i \in r_d} \hat{p}_{di} \right)$$

donde

$$\hat{p}_{di} = h(\mathbf{x}'_{di} \hat{\alpha} + \hat{v}_d)$$

- El estimador plug-in no es insesgado a menos que la función link es lineal

Métodos basados en modelos lineales generalizados mixtos: método plug-in

- Aunque es más fácil calcular, el estimador plug-in no es insesgado a menos que la función link es lineal
- Sin embargo, el link logístico $g(p) = \log(p/(1-p))$ es aproximadamente lineal para $p \in (02, 08)$
- Debido a esta propiedad, se puede comprobar que el EB y plug-in de la proporción $P_d = \bar{Y}_d$ se parecen al EBLUP, $\hat{P}_d^{EBLUP} = \hat{\bar{Y}}_d^{EBLUP}$, basado en el modelo con errores anidados (BHF)

Métodos basados en modelos lineales generalizados mixtos: ECM bootstrap

El ECM del estimador EB o plug-in se puede estimar con un procedimiento bootstrap 1) Ajustar el modelo GLMM

$Y_{di} | v_d \sim \text{Bern}(p_{di})$, $g(p_{di}) = \mathbf{x}'_{di}\boldsymbol{\alpha} + v_d$ a los datos de la muestra para obtener los estimadores $\hat{\sigma}_v^2$ y $\hat{\boldsymbol{\alpha}}$ 2) Generar efectos aleatorios bootstrap

$$v_d^{*(b)} \stackrel{iid}{\sim} N(0, \hat{\sigma}_v^2), \quad d = 1, \dots, D$$

Métodos basados en modelos lineales generalizados mixtos: ECM bootstrap

- 3) Generar el censo bootstrap $\mathbf{y}_d^{*(b)} = (Y_{d1}, \dots, Y_{dN_d})'$ en la siguiente forma:

$$Y_{di}^{*(b)} \stackrel{\text{ind}}{\sim} \text{Bern}(p_{di}^{*(b)}),$$

y

$$p_{di}^{*(b)} = h(\mathbf{x}'_{di}\hat{\boldsymbol{\alpha}} + v_d^{*(b)}), \quad i = 1, \dots, N_d, \quad d = 1, \dots, D,$$

Calcular los verdaderos valores de los indicadores para el censo $P_d^{*(b)} = \bar{Y}_d^{*(b)}$, $d = 1, \dots, D$.

- 4) Para cada área, extraer del censo los elementos de la muestra Y_{di} , $i \in s_d^{*(b)}$ para construir el vector $\mathbf{y}_{ds}^{*(b)}$ y, después, $\mathbf{y}_s^{*(b)} = ((\mathbf{y}_{1s}^{*(b)})', \dots, (\mathbf{y}_{Ds}^{*(b)})')'$ el vector con los valores en la muestra de todas las áreas

Métodos basados en modelos lineales generalizados mixtos: ECM bootstrap

- 5) Ajustar el modelo GLMM a los datos bootstrap $\mathbf{y}_s^{*(b)}$ y calcular $\hat{P}_d^{EB*(b)}$, $d = 1, \dots, D$
- 6) Repetir pasos 2-5 B veces. El estimador bootstrap de ECM viene dado por

$$mse_B(\hat{P}_d^{EB}) = B^{-1} \sum_{b=1}^B (\hat{P}_d^{EB*(b)} - P_d^{*(b)})^2$$

Resumen del predictor EB/plug-in basado en GLMM

- Indicadores objetivos: Proporciones o totales de una variable binaria (e.g.carencia o no de determinado bien o servicio)
- Requerimientos de datos:
 - Microdatos de las p covariables obtenidas de la misma encuesta de la variable de interés
 - Áreas de interés obtenidas de la misma encuesta
 - Microdatos de las p covariables de un censo o registro. Esto es necesario para calcular la esperanza $E(Y_{di}|\mathbf{y}_{ds}; \theta)$

Resumen del predictor EB/plug-in basado en GLMM

- Ventajas:

- El número de observaciones usadas es el tamaño muestral, mucho mayor que el número de áreas
- El modelo GLMM incorpora heterogeneidad no explicada entre las áreas
- No se necesita conocer ninguna varianza, al contrario que para el modelo FH

Resumen del predictor EB/plug-in basado en GLMM

- Ventajas:

- El estimador del ECM bajo el modelo es un estimador estable y insesgado bajo el diseño (y bajo el diseño cuando el número de áreas es grande)
- Se puede estimar en cualquier subárea sin reajustar el modelo
- Se puede estimar en áreas no muestradas

Resumen del predictor EB/plug-in basado en GLMM

- Desventajas:
 - Es basado en un modelo y por tanto es necesario analizar el modelo (a través de los residuos, por ejemplo)
 - No tiene en cuenta el diseño muestral y, por eso, no es insesgado bajo el diseño
 - El uso de microdatos de un censo puede conllevar problemas de confidencialidad
 - El estimador ECM bootstrap *no* es insesgado bajo el modelo para el ECM bajo el modelo para un área concreta

Resumen del predictor EB/plug-in basado en GLMM

- Desventajas:
 - El predictor EB (no el plug-in) es computacionalmente intensivo
 - El ECM del estimador EB usando un proceso bootstrap es excesivamente intensivo, pero se puede cortar usando el plug-in
 - Requiere un reajuste para verificar la propiedad “benchmarking”

Aplicación

Antecedentes

- La ENDES 2018 permite la estimación a nivel nacional, departamental y por zona (urbano/rural) de indicadores demográficos relacionados con salud, natalidad, planificación familiar, etc.
- Las estimaciones están limitadas al alcance del diseño de muestreo de la encuesta, por lo que la precisión de dichas estimaciones disminuye con el número de desagregaciones.
- Perú cuenta con 25 departamentos y 196 provincias donde se incluye la provincia constitucional del Callao.
- El proyecto fue financiado conjuntamente por UNFPA y CEPAL.

Indicadores de planificación familiar

- La información sobre prácticas de prevención y autocuidado se enmarcan en la noción de que la salud sexual y reproductiva constituye un derecho de los hombres y las mujeres a lo largo de todo su ciclo vital.
- Las parejas y los individuos tienen derecho a decidir de manera libre y responsable sobre el número de hijos que desean tener, el espaciamiento de los nacimientos, etc., además de disponer de la información y los medios necesarios para ello.

Indicadores de planificación familiar

Tomando como base la propuesta de indicadores del Consenso de Montevideo (Cepal, 2018), se propone la estimación de los siguientes indicadores:

- Proporción de mujeres que hace uso de métodos anticonceptivos.
- Proporción de mujeres que cubren sus necesidades de planificación familiar con métodos modernos.
- Proporción de mujeres en edad reproductiva (o con pareja) cuyas necesidades de planificación familiar no están cubiertas.

Estimaciones SAE con modelos de unidad

Cada respuesta dentro de la encuesta se transforma en una variable binaria. Por ejemplo, en el caso del uso de métodos anticonceptivos modernos en mujeres en edad fértil para el área d , se tiene que y_{di} es la respuesta binaria de la i -ésima mujer en el área d medida como

$$y_{di} = \begin{cases} 1, & \text{si el individuo usa anticonceptivos moderno} \\ 0, & \text{si el individuo no usa anticonceptivos modernos} \end{cases}$$

Modelo lineal generalizado

Como $y_{di} \in \{0, 1\}$ es posible definir un modelo lineal generalizado mixto para cada área como

$$y_{di} | \nu_d \sim \text{Ber}(\theta_{di}), \quad d = 1, \dots, D, i = 1, \dots, N_d$$

En donde $g(\theta_{di}) = \mathbf{x}_{di}^T \boldsymbol{\beta} + \nu_d$, siendo $\boldsymbol{\beta}$ los coeficientes de regresión del modelo, $g : (0, 1) \rightarrow \mathbb{R}$ el enlace logit dado por $g(\theta) = \log(\frac{\theta}{1-\theta})$ y $\nu_1, \dots, \nu_D \stackrel{\text{i.i.d}}{\sim} N(0, \sigma_\nu^2)$ representan los efectos de área.

Predicción

Una vez ajustado el modelo lineal generalizado mixto, se predicen los valores para todo el censo a través del modelo, de la siguiente manera:

$$\hat{\theta}_d^{PI} = \frac{1}{N_d} \sum_{i \in U_d} \hat{\theta}_{di}$$

En donde

$$\hat{\theta}_{di} = g^{-1}(\mathbf{x}_{di}^T \hat{\beta} + \hat{\nu}_d) = \frac{\exp(\mathbf{x}_{di}^T \hat{\beta} + \hat{\nu}_d)}{1 + \exp(\mathbf{x}_{di}^T \hat{\beta} + \hat{\nu}_d)}$$

es el valor predicho de las observaciones obtenidas del censo, siendo $\hat{\beta}$ y $\hat{\nu}_d$ las estimaciones correspondientes de β y ν_d .

Predicción en áreas no muestreadas

Para el caso en que existan áreas no muestreadas, y por tanto no sea posible obtener el efecto aleatorio $\hat{\nu}_d$, la estimación correspondiente del valor predicho para el individuo i del área está dada por

$$\hat{\theta}_{di} = g^{-1}(\mathbf{x}_{di}^T \boldsymbol{\beta})$$

De manera análoga fue posible calcular los estimadores respectivos para los otros indicadores de interés.

Estimación del ECM

El ECM de los predictores se estima mediante un procedimiento bootstrap a partir del modelo propuesto. Para ello es necesario seguir el siguiente procedimiento (I. Molina 2019):

- ① Ajustar el modelo

$y_{di} | \nu_d \sim \text{Ber}(\theta_{di}), \quad d = 1, \dots, D, i = 1, \dots, n_d$ a los datos de la encuesta Endes para obtener los estimadores $\hat{\sigma}_\nu$ y $\hat{\beta}$.

- ② generar $\hat{\nu}_d^{*(b)} \stackrel{\text{i.i.d}}{\sim} N(0, \hat{\sigma}_\nu)$ para $d = 1, \dots, D$

- ③ Generar para cada área d un censo

$\mathbf{Y}_d^{*(b)} = (y_{d1}^{*(b)}, \dots, y_{dN_d}^{*(b)})^T$ de la forma

$$y_{di}^{*(b)} \stackrel{\text{i.i.d}}{\sim} \text{Ber}(\theta_{di}^{*(b)}), \quad \theta_{di}^{*(b)} = \frac{\exp(x_{di}^T \hat{\beta} + \hat{\nu}_d^{*(b)})}{1 + \exp(x_{di}^T \hat{\beta} + \hat{\nu}_d^{*(b)})}$$

y con estos valores calcular $\theta_d^{*(b)} = \frac{1}{N_d} \sum_{i=1}^{N_d} y_{di}^{*(b)}$ para $d = 1, \dots, D$.

Estimación del ECM

- 4) Para cada área muestreada d , extraer una muestra aleatoria de tamaño n_d mediante un muestreo aleatorio estratificado proporcional al tamaño N_d de cada provincia.
- 5) Ajustar el modelo de (1) a la muestra obtenida y calcular los predictores *Bootstrap* $\hat{\theta}_d^{EB*(b)}$ para $d = 1, , D$.
- 6) Repetir los pasos 2)-5) para $b = 1, , B$. Finalmente, el estimador *Bootstrap* del ECM para la estimación de $\hat{\theta}_d$ está dada por

$$\text{ECM}_B(\hat{\theta}_d) = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_d^{EB*(b)} - \hat{\theta}_d^{*(b)})^2$$

Estimación

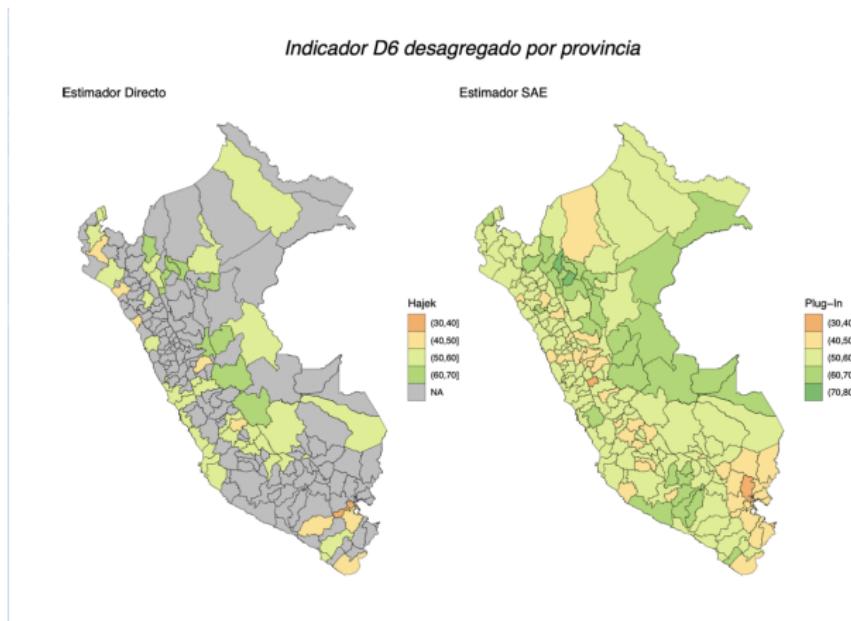


Figura 1: Uso de métodos de planificación familiar

Estimación

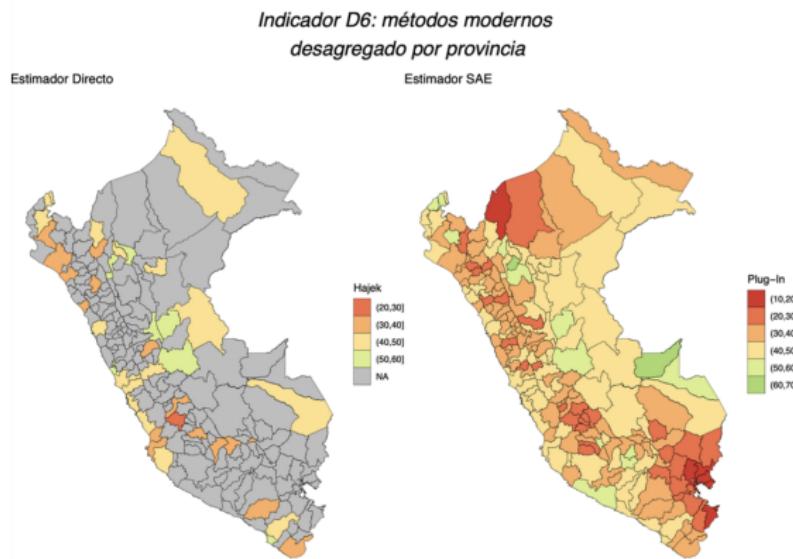


Figura 2: Uso de métodos modernos de planificación familiar

Estimación

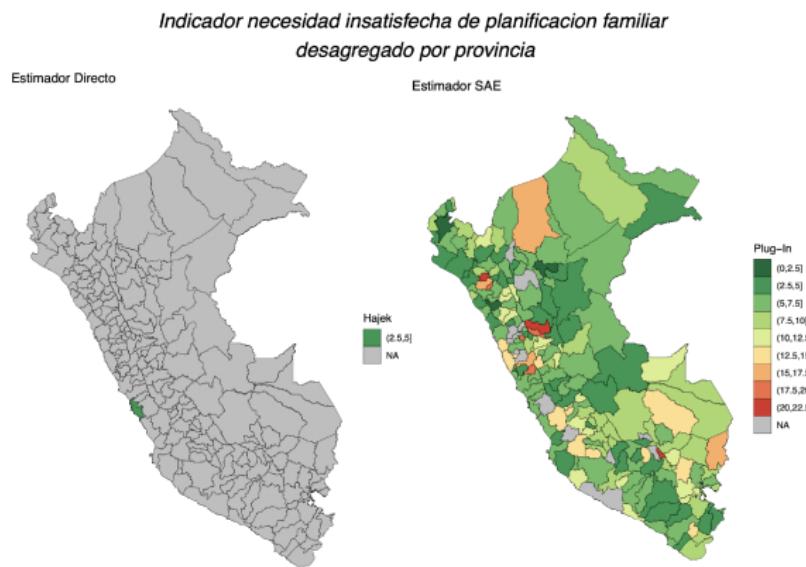


Figura3: Indicador de necesidades insatisfechas de planificación familiar

¡Gracias!

¡Gracias!

Curso Internacional de Desagregación de Estimaciones en Áreas Pequeñas usando R

*Realización de mapas para indicadores desagregados
geográficamente usando R*

División de Estadísticas
Comisión Económica para América Latina y el Caribe

2020

1 Datos cartográficos

2 Mapas con ggplot2

3 Mapas con tmap

4 Mapas en leaflet

Datos cartográficos

Introducción

- Los mapas son una herramienta gráfica poderosa para la visualización de datos.

Introducción

- Los mapas son una herramienta gráfica poderosa para la visualización de datos.
- Para indicadores sociales-demográficos estos son una gran referencia visual para desagregaciones a nivel País, región, departamento, provincia, distrito, municipio, comuna, etc.

Introducción

- Los mapas son una herramienta gráfica poderosa para la visualización de datos.
- Para indicadores sociales-demográficos estos son una gran referencia visual para desagregaciones a nivel País, región, departamento, provincia, distrito, municipio, comuna, etc.
- El software de código libre utilizado para análisis estadístico R posee un sin fin de métodos de programación para representar dichos mapas.

Datos cartográficos

- Para graficar mapas es necesario contar con información geoespacial, datos que contienen las coordenadas o delimitaciones geográficas de determinado país o región.

Datos cartográficos

- Para graficar mapas es necesario contar con información geoespacial, datos que contienen las coordenadas o delimitaciones geográficas de determinado país o región.
- Sitios web como <http://www.diva-gis.org/gdata> ofrecen de manera gratuita bases de datos o shapes que contienen los vectores asociados a las geografías correspondientes.

Datos cartográficos

- Dichos conjuntos de datos poseen observaciones sobre la longitud y latitud lo cuál permite graficar en R un conjunto de puntos cuya unión en el gráfico formarán las formas los polígonos que dan forma a las áreas geográficas.

Datos cartográficos

- Dichos conjuntos de datos poseen observaciones sobre la longitud y latitud lo cuál permite graficar en R un conjunto de puntos cuya unión en el gráfico formarán las formas los polígonos que dan forma a las áreas geográficas.
- Como ejemplo, se utilizarán los resultados obtenidos mediante la metodología SAE en Perú a nivel provincia para el indicador ODS 3.7.1 o D.7 del consenso de Montevideo (Mujeres unidas en edad fértil que utilizan métodos modernos).

Datos cartográficos

- La base de datos contiene el nombre del departamento, el código ubigeo para las circunscripciones territoriales (a nivel provincia) y la estimación puntual del indicador para dicha provincia.

id	departamento	ubigeo	D7
0	Amazonas	0101	0.7674031
1	Amazonas	0102	0.7320016
2	Amazonas	0103	0.7494394
3	Amazonas	0104	0.3430710
4	Amazonas	0105	0.6352737
5	Amazonas	0106	0.8246836

Datos cartográficos

- La librería rgdal de R nos permite leer la información geoespacial en formato .shp contenida en la carpeta llamada PER_adm.

Datos cartográficos

- La librería rgdal de R nos permite leer la información geoespacial en formato .shp contenida en la carpeta llamada PER_adm.
- ohsPER2 contiene la información geoespacial necesaria para mapear los polígonos a nivel provincia en Perú.

Datos cartográficos

- Perú se divide administrativamente en Departamentos, Provincias y distritos. La información anterior está construida para la segunda desagregación geográfica (196 provincias).

```
library(rgdal)
ohsPER2 <- readOGR("../PER_adm/provincias/PROVINCIAS.shp")
```

Datos cartográficos

Name	Type	Value
ohsPER2	S4 [196 x 6] (sp::SpatialPolygons)	S4 object of class SpatialPolygonsDataFrame
data	list [196 x 6] (S3: data.frame)	A data.frame with 196 rows and 6 columns
polygons	list [196]	List of length 196
plotOrder	integer [196]	138 140 142 144 145 194 ...
bbox	double [2 x 2]	-81.3282 -18.3509 -68.6523 -0.0386
proj4string	S4 (sp::CRS)	S4 object of class CRS

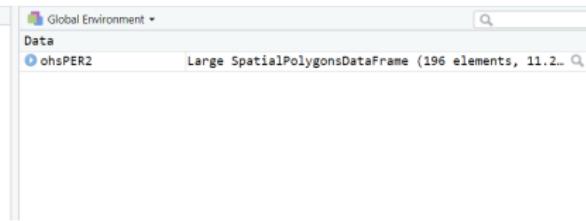


Figure 1: archivo .shp

Mapas con ggplot2

Librería ggplot

- ggplot2 es una potente librería gráfica que permite crear mediante códigos computacionales distintos entornos gráficos en R.

Librería ggplot

- ggplot2 es una potente librería gráfica que permite crear mediante códigos computacionales distintos entornos gráficos en R.
- Mediante `geom_polygon()` es posible utilizar un conjunto de vectores de un archivo .shp para graficar formas poligonales que generan el mapa de una determinada división administrativas de un país.

Librerías requeridas

Las librerías necesarias se muestran en la siguiente sintaxis.

```
library(dplyr)
library(tidyverse)
library(magrittr)
library(sp)
library(RColorBrewer)
library(ggplot2)
library(maptools)
library(scales)
library(gridExtra)
library(grid)
library(sf)
library(rgdal)
```

Librerías requeridas

- Las librerías dplyr, tidyverse y magritr se utilizan para realizar manejo de bases de datos.

Librerías requeridas

- Las librerías dplyr, tidyverse y magritr se utilizan para realizar manejo de bases de datos.
- RColorBrewer permite utilizar funciones para explorar las grillas de colores y paletas que dispone R.

Librerías requeridas

- Las librerías dplyr, tidyverse y magritr se utilizan para realizar manejo de bases de datos.
- RColorBrewer permite utilizar funciones para explorar las grillas de colores y paletas que dispone R.
- sp se utiliza para crear las data.frame a partir de la información geoespacial y poder realizar los emparejamientos necesarios con los datos que se desean graficar.

Librerías requeridas

- Las librerías dplyr, tidyverse y magritr se utilizan para realizar manejo de bases de datos.
- RColorBrewer permite utilizar funciones para explorar las grillas de colores y paletas que dispone R.
- sp se utiliza para crear las data.frame a partir de la información geoespacial y poder realizar los emparejamientos necesarios con los datos que se desean graficar.
- Las librerías grid y gridExtra permiten la unificación de distintos mapas en una sola grilla.

Librerías requeridas

- Una vez cargado el archivo .shp en el entorno de R, se utiliza la función `fortify()` para convertir la lista de información geoespacial en una `data.frame` con la cual se realiza el emparejamiento mediante la variable `id`.

Librerías requeridas

- Una vez cargado el archivo .shp en el entorno de R, se utiliza la función `fortify()` para convertir la lista de información geoespacial en una `data.frame` con la cual se realiza el emparejamiento mediante la variable `id`.
- Se debe tener en cuenta como está formado el `id` y la concordancia que tiene con la información que se dispone.

Librerías requeridas

- Una vez cargado el archivo .shp en el entorno de R, se utiliza la función `fortify()` para convertir la lista de información geoespacial en una `data.frame` con la cual se realiza el emparejamiento mediante la variable `id`.
- Se debe tener en cuenta como está formado el `id` y la concordancia que tiene con la información que se dispone.
- Si la información .shp proviene de fuentes oficiales no se debería tener problemas al momento de la unión, información no oficial puede estar desactualizada u ordenada de otra forma.

Librerías requeridas

- En el caso de Perú, las provincias están ordenadas según la circunscripciones territoriales del INEI Ubigeo. Las provincias se ordenan por orden alfabético de departamento seguido de la capital del departamento, luego, se ordenan alfabéticamente las provincias correspondientes del departamento.

Librerías requeridas

- En el caso de Perú, las provincias están ordenadas según la circunscripciones territoriales del INEI Ubigeo. Las provincias se ordenan por orden alfabético de departamento seguido de la capital del departamento, luego, se ordenan alfabéticamente las provincias correspondientes del departamento.
- En este caso `id = 0` corresponde a Chacapoyas capital de Amazonas el primer departamento ordenado en orden alfabético. Los siguientes 6 `id` corresponden a las provincias que se encuentran en el departamento de Amazonas ordenados de manera alfabética.

Creando mapas en R con ggplot2

- Es necesario leer los *layers* del mapa. Perú se divide en 24 departamentos y una provincia constitucional que a su vez están constituidas por 196 provincias. Esta forma viene dada por el archivo PER_adm2.shp.

```
ohsPERI2 <- fortify(ohsPER2)
shapes <- ohsPERI2 %>% merge(D7, by = "id")
```

Creando mapas en R con ggplot2

- Se debe combinar estos datos con la variable de interés.

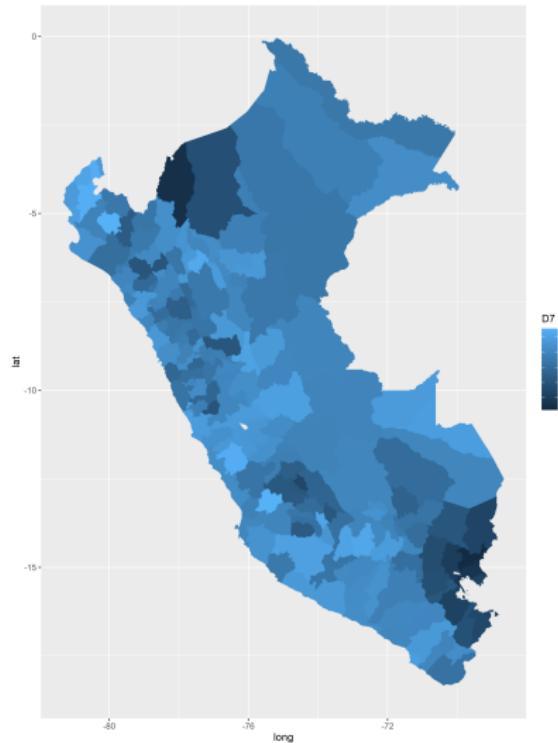
Creando mapas en R con ggplot2

- Se debe combinar estos datos con la variable de interés.
- Se crea el mapa utilizando ggplot2 y se puede personalizar utilizando las distintas herramientas que entrega la librería ggplot2.

Creando mapas en R con ggplot2

```
Mapa <- ggplot() +  
  geom_polygon(data=shapes, aes(long, lat, group=group, fill=D7),  
               colour ="black", size = 0.005)  
Mapa
```

Creando mapas en R con ggplot2



Creando mapas en R con ggplot2

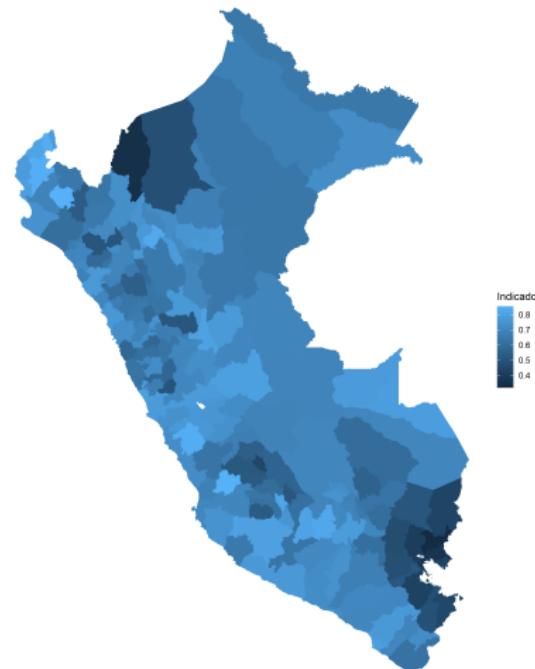
Podemos personalizar el mapa anterior mediante líneas de códigos de ggplot2.

```
Mapa + coord_quickmap() +
  theme_void() +
  labs(fill = "Indicador", title="Estimador SAE")

# ggsave(mapa2, file = "Mapa2.png",
#         width = 8.5, height = 11, type =
#         "cairo-png")
```

Creando mapas en R con ggplot2

Estimador SAE



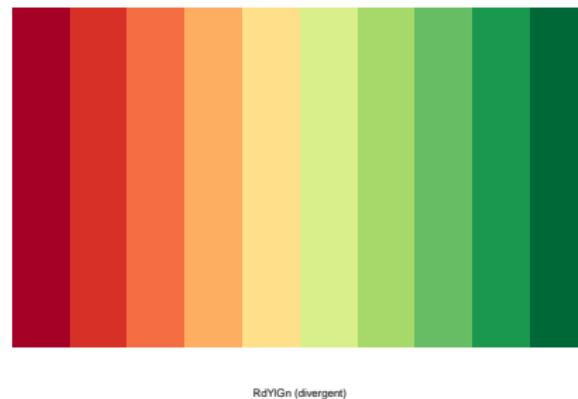
Discretización de la variable

Es posible discretizar la variable de interés para generar gráficos más limpios con intervalos predeterminados. La librería RColorBrewer nos permite acceder a grillas de colores junto a su codificación respectiva.

```
brewer.pal(n = 10, name = 'RdYlGn')  
## [1] "#A50026" "#D73027" "#F46D43" "#FDAE61" "#FEE08B"  
## [8] "#66BD63" "#1A9850" "#006837"
```

Discretización de la variable

```
display.brewer.pal(n = 10, name = 'RdYlGn')
```



Discretización de la variable

Los primeros 3 intervalos y el último de la variable discreteada no poseen datos. Utilizamos por tanto, los códigos de colores que corresponden a cada intervalo. Se tienen 10 códigos de colores para los 10 intervalos generados.

```
shapes$discrete_value = cut(100*shapes$D7,
                            breaks=seq(from=0,to=100, length.out=11))
```

```
table(shapes$discrete_value)
```

0-10	10-20	20-30	30-40	40-50
0	0	0	14267	55069

50-60	60-70	70-80	80-90	90-100
82274	276329	236335	30392	0

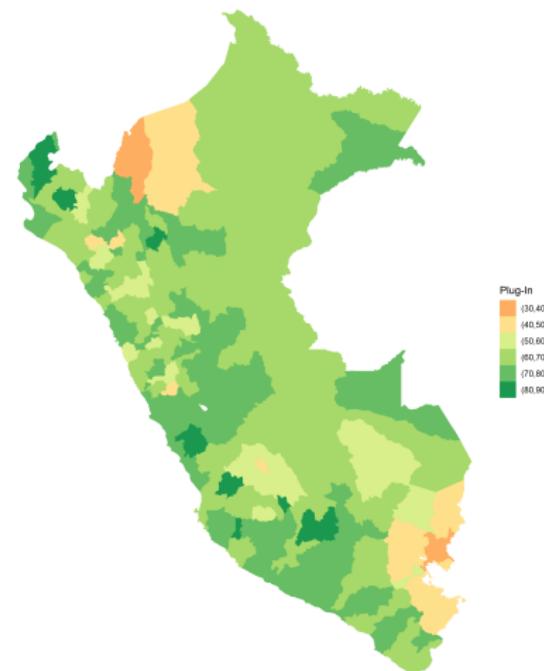
Discretización de la variable

Generamos el mapa y añadimos los elementos deseados para el mapa final.

```
ggplot() +  
  geom_polygon(data=shapes, aes(long, lat, group=group, fill=discrete_value),  
               colour ="black", size = 0.005) +  
  scale_fill_manual(  
    values = c("#FDAE61", "#FEE08B", "#D9EF8B", "#A6D96A", "#66BD63", "#1A9850"),  
    na.value="grey") +  
  coord_quickmap() + theme_void() + labs(fill = "Plug-In", title="Estimador SAE")
```

Discretización de la variable

Estimador SAE



Departamentos de Colombia

- Es posible obtener la información geoespacial de Colombia en <http://www.diva-gis.org/gdata>.

Departamentos de Colombia

- Es posible obtener la información geoespacial de Colombia en <http://www.diva-gis.org/gdata>.
- Colombia se divide en 32 departamentos y un distrito capital Bogotá.

Departamentos de Colombia

- Es posible obtener la información geoespacial de Colombia en <http://www.diva-gis.org/gdata>.
- Colombia se divide en 32 departamentos y un distrito capital Bogotá.
- El Shape para dicha división viene dado por el archivo `COL_adm1.shp`.

Departamentos de Colombia

- Es posible obtener la información geoespacial de Colombia en <http://www.diva-gis.org/gdata>.
- Colombia se divide en 32 departamentos y un distrito capital Bogotá.
- El Shape para dicha división viene dado por el archivo `COL_adm1.shp`.
- Con el archivo cargado debemos unir la data con la variable de interés, en este caso, utilizaremos números aleatorios

Departamentos en Colombia

```
ohsCol2 <- readOGR("COL_adm/COL_adm1.shp")
ohsColI2 <- fortify(ohsCol2)
grupo2 <- data.frame(id = unique(ohsColI2[ , c("id")]))
grupo2[ , "Porcentaje"] <- runif(nrow(grupo2), 0, 1)
ohsColI2 <- merge(ohsColI2, grupo2, by = "id")
save(ohsColI2, file = "ColData.RData")
```

id	long	lat	order	hole	piece	group	Porcentaje
0	-69.43138	-1.078474	1	FALSE	1	0.1	0.2320404
0	-69.42591	-1.096313	2	FALSE	1	0.1	0.2320404
0	-69.42345	-1.104404	3	FALSE	1	0.1	0.2320404
0	-69.41992	-1.111588	4	FALSE	1	0.1	0.2320404
0	-69.41006	-1.131676	5	FALSE	1	0.1	0.2320404
0	-69.39285	-1.148357	6	FALSE	1	0.1	0.2320404

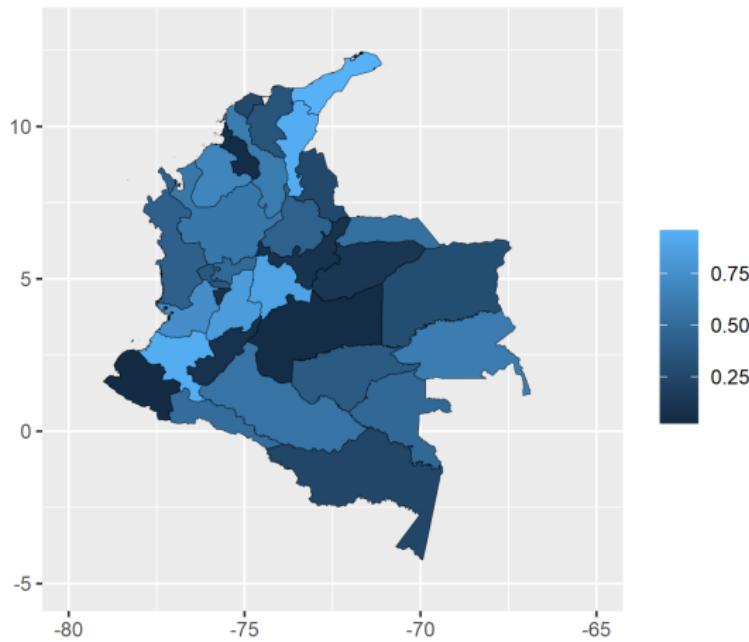
Departamentos de Colombia

Usando la librería ggplot generamos el mapa.

```
mapColDep <- ggplot() +  
  geom_polygon(data=ohsColI2, aes(x=long, y=lat, group = group,  
                                    fill = Porcentaje), colour ="black", size = 0.1) +  
  labs(title = "Colombia", fill = "") +  
  labs(x="",y="",title="Colombia") +  
  scale_x_continuous(limits=c(-80,-65))+  
  scale_y_continuous(limits=c(-5,13))
```

Departamentos en Colombia

Colombia



Departamentos en Colombia

Para guardar el mapa utilizamos la función ggsave.

```
ggsave(mapColDep, file = "mapColDep.png",
       width = 5, height = 4.5, type =
"cairo-png")
```

Mapas con tmap

Mapas en tmap

- La librería tmap funciona de manera similar a ggplot2.

Mapas en tmap

- La librería tmap funciona de manera similar a ggplot2.
- En primer lugar se define el objeto espacial a dibujar utilizando la función `tm_shape()`.

Mapas en tmap

- La librería `tmap` funciona de manera similar a `ggplot2`.
- En primer lugar se define el objeto espacial a dibujar utilizando la función `tm_shape()`.
- Para graficar el mapa mediante polígonos utilizando el indicador correspondiente, utilizamos la función `tm_polygons()`.

Mapas en tmap

- La librería tmap funciona de manera similar a ggplot2.
- En primer lugar se define el objeto espacial a dibujar utilizando la función `tm_shape()`.
- Para graficar el mapa mediante polígonos utilizando el indicador correspondiente, utilizamos la función `tm_polygons()`.
- Unimos el indicador a los datos contenidos en `ohsPER2@data` de la información espacial.

Mapas en tmap

```
library(tmap)
library(maptools)
ohsPER2 <- readOGR("../PER_adm/provincias/PROVINCIAS.shp")
D7 <- readRDS("../D7.rds")

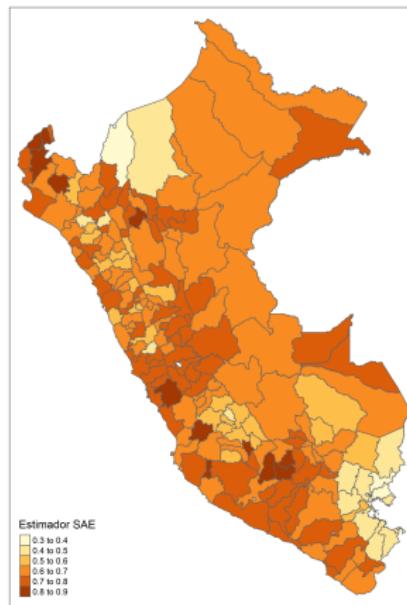
ohsPER2@data <- ohsPER2@data %>% left_join(D7 %>% select(ubigeo,D7),
                                               by = c("IDPROV" = "ubigeo"))
```

Mapas en tmap

Con lo anterior, ya es posible graficar.

```
tm_shape(ohsPER2) + tm_polygons("D7", title = "Estimador SAE")
```

Mapas en tmap



Mapas en tmap

- Utilizando la opción palette podemos escoger la paleta de colores para el indicador

Mapas en tmap

- Utilizando la opción palette podemos escoger la paleta de colores para el indicador
- anteponiendo un signo – es posible invertir el orden de dicha paleta.

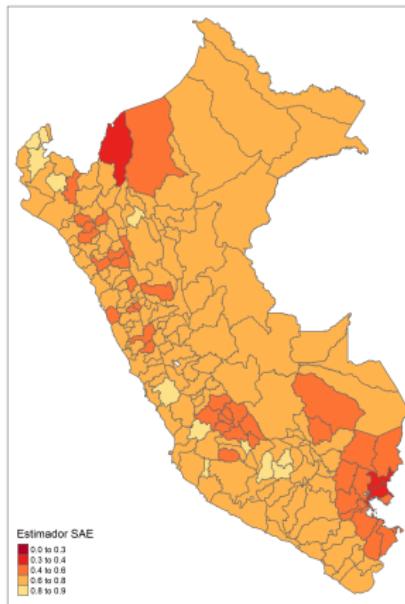
Mapas en tmap

- Utilizando la opción palette podemos escoger la paleta de colores para el indicador
- anteponiendo un signo – es posible invertir el orden de dicha paleta.
- Podemos escoger los intervalos utilizando la opción breaks.

Mapas en tmap

```
tm_shape(ohsPER2) + tm_polygons("D7", title = "Estimador SAE",  
                                palette = "-YlOrRd",  
                                breaks = c(0,.4,.5,.6,.7,.8,.9))
```

Mapas en tmap



Mapas en tmap

Para guardar el mapa generado utilizamos la función `tmap_save`

```
tmap_save(mapa, file = "Pics/MapaPeru.png",
          width = 11, height = 8.5, units = "in")
```

Mapa de Chile

- Es posible descargar el mapa vectorial de las comunas de Chile en https://www.bcn.cl/siit/mapas_vectoriales

Mapa de Chile

- Es posible descargar el mapa vectorial de las comunas de Chile en https://www.bcn.cl/siit/mapas_vectoriales
- Utilizando la función `st_read()` leemos la información geoespacial de las 346 comunas de Chile.

Mapa de Chile

```
ChileSP <- st_read("comunas/comunas.shp")
```

```
Simple feature collection with 6 features and 11 fields
geometry type:  MULTIPOLYGON
dimension:      XY
bbox:           xmin: -8133264 ymin: -4748322 xmax: -7828105 ymax: -4017907
projected CRS: WGS 84 / Pseudo-Mercator
  objectid shape_leng dis_elec cir_sena cod_comuna codregion st_area_sh st_length_
1        48 170038.62     16      8    6204       6 968577420 206184.27
2        29 125730.10     15      8    6102       6 415744636 151911.58
3        30  63026.08     15      8    6103       6 144856484  76355.33
4        31  89840.90     15      8    6104       6 325657168 108874.62
5        78 122626.49     23     11    9121       9 699072708 156680.41
6        79 279936.00     23     11    9103       9 3127304688 360052.12
                                             Region   Comuna Provincia          geometry
1 Región del Libertador Bernardo O'Higgins Marchigüe Cardenal Caro MULTIPOLYGON (((-7992819 -4...
2 Región del Libertador Bernardo O'Higgins Codegua Cachapoal MULTIPOLYGON (((-7831652 -4...
3 Región del Libertador Bernardo O'Higgins Coinco Cachapoal MULTIPOLYGON (((-7892616 -4...
4 Región del Libertador Bernardo O'Higgins Coltauco Cachapoal MULTIPOLYGON (((-7906458 -4...
5             Región de La Araucanía Cholchol Cautín MULTIPOLYGON (((-8121756 -4...
6             Región de La Araucanía Cunco Cautín MULTIPOLYGON (((-7992287 -4...
```

Mapa de Chile

- Generamos un indicador aleatorio para cada comuna de Chile o realizar un emparejamiento del indicador sobre las comunas.

```
ChileSP <- ChileSP %>%
  mutate(indicador = runif(346, 0, 1))
```

Mapa de Chile

- Generamos un indicador aleatorio para cada comuna de Chile o realizar un emparejamiento del indicador sobre las comunas.
- Es posible graficar el país completo o escoger una Región determinada mediante la función filter() de dplyr

```
ChileSP <- ChileSP %>%
  mutate(indicador = runif(346, 0, 1))
```

Mapa de Chile

```
tm_shape(ChileSP) + tm_polygons("indicador")
```

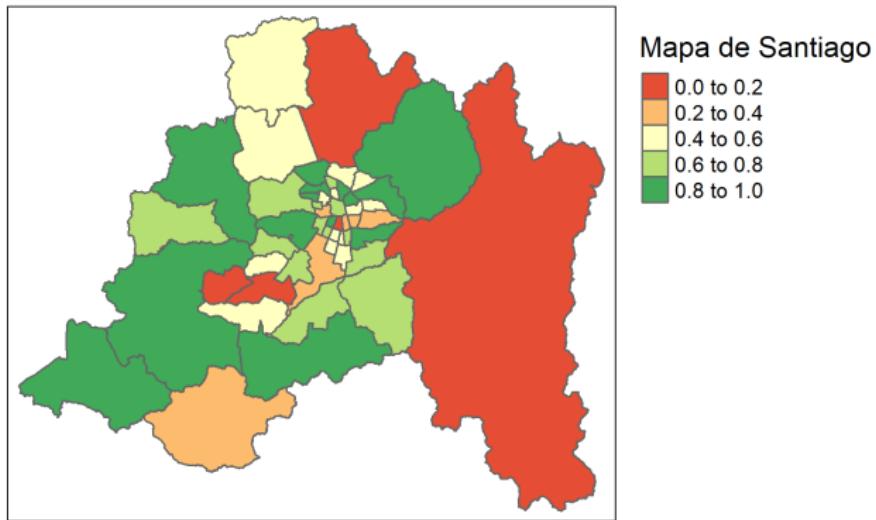
Mapa de Chile



Mapa de Chile: Santiago

```
chile2 <- tm_shape(ChileSP %>% filter(codregion == 13)) +  
  tm_polygons("indicador", palette = "RdYlGn",  
              title = "Mapa de Santiago") +  
  tm_legend(legend.outside = TRUE,  
            legend.outside.position="right")
```

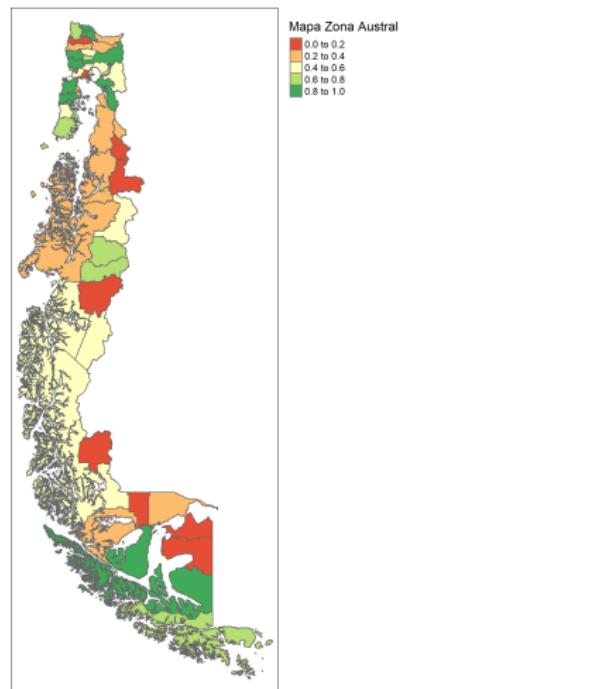
Mapa de Chile: Santiago



Mapa de Chile: Zona Austral

```
tm_shape(ChileSP %>%
          filter(codregion%in% c(10,11,12))) +
  tm_polygons("indicador", palette = "RdYlGn",
              title = "Mapa Zona Austral") +
  tm_legend(legend.outside = TRUE,
            legend.outside.position="right")
```

Mapa de Chile: Zona Austral



Mapas en leaflet

Mapas en leaflet

- Leaflet es una de las librerías de código abierto de Javascript mas utilizada para generar mapas.

Mapas en leaflet

- Leaflet es una de las librerías de código abierto de Javascript mas utilizada para generar mapas.
- A diferencia de lo hecho con ggplot permite generar mapas interactivos a través de populares GIS como CartoDB, OpenStreetMaps y Mapboox.

Mapas en leaflet

Las librerías necesarias para la creación de estos mapas se muestra en la siguiente sintaxis:

```
library(dplyr)
library(leaflet)
library(leaflet.extras)
library(leaflet.providers)
library(rgeos)
library(rgdal)
```

Mapa en leaflet

Para empezar a generar mapas interactivos, la función `Leaflet()` genera el entorno para crear el mapa. Podemos generar el mapa mundial con la función `addTiles()`.

```
leaflet() %>% addTiles()
```

Mapa en leaflet



Mapa en leaflet

- Una vez cargada la información geoespacial, se debe indexar el indicador dentro de la data que contiene la lista.

```
ohsPER2 <- readOGR(".../PER_adm/provincias/PROVINCIAS.shp")  
  
ohsPER2@data <- D7 %>% select(ubigeo,D7) %>%  
  right_join(ohsPER2@data, by = c("ubigeo"="IDPROV"))
```

Mapa en leaflet

- Una vez cargada la información geoespacial, se debe indexar el indicador dentro de la data que contiene la lista.
- En este caso, las provincias de la información geoespacial están escritas en otro formato, se procede a indexarlas en una nueva `data.frame()` para añadir el indicador a la lista. Esta información se encuentra disponible en `ohsPER@data`.

```
ohsPER2 <- readOGR(".../PER_adm/provincias/PROVINCIAS.shp")
```

```
ohsPER2@data <- D7 %>% select(ubigeo,D7) %>%
  right_join(ohsPER2@data, by = c("ubigeo"="IDPROV"))
```

Mapas en leaflet

- Con la información geoespacial actualizada con el indicador, se añade al mapa los polígonos mediante la función `addPolygons()`.

```
d7_pal <- colorNumeric("RdYlGn", domain = ohsPER2@data$D7)

ohsPER2 %>% leaflet() %>% addTiles() %>%
  addPolygons(weight = 1, color = ~d7_pal(D7), fillOpacity = .5,
              label = ~paste0(PROVINCIA, ":", round(D7, 3)))
```

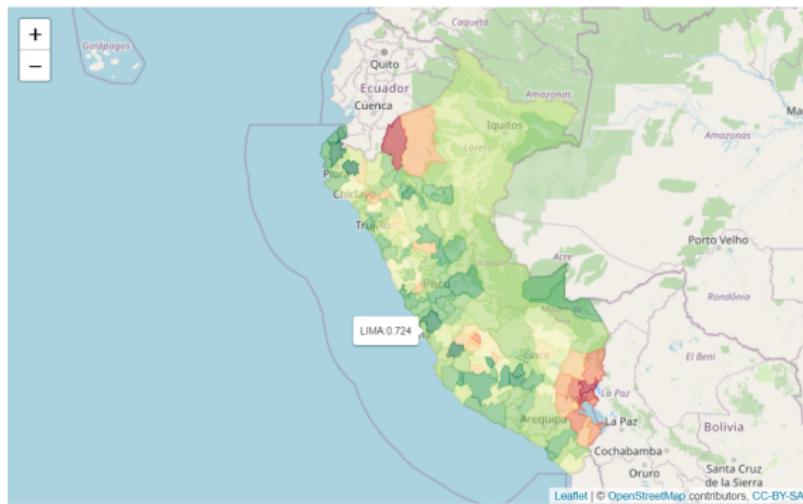
Mapas en leaflet

- Con la información geoespacial actualizada con el indicador, se añade al mapa los polígonos mediante la función `addPolygons()`.
- Una forma de escoger los colores deseados para mapear el indicador es creando un objeto mediante la función `ColorNumeric()`.

```
d7_pal <- colorNumeric("RdYlGn", domain = ohsPER2@data$D7)

ohsPER2 %>% leaflet() %>% addTiles() %>%
  addPolygons(weight = 1, color = ~d7_pal(D7), fillOpacity = .5,
              label = ~paste0(PROVINCIA, ":", round(D7, 3)))
```

Mapas en leaflet



Mapas en leaflet

Podemos guardar nuestro mapa en un archivo .html utilizando el código que se muestra a continuación:

```
saveWidget(MAPA,  
          file = "PeruTargetGroups.html")
```

¡Gracias!

¡Gracias!