ISSN 1680-8789

SERIE

ESTUDIOS ESTADÍSTICOS 97

Desagregación de datos en encuestas de hogares

Metodologías de estimación en áreas pequeñas

Isabel Molina

NACIONES UNIDAS

CEPAL

CEPAL

POR UN DESARROLLO SOSTENIBLE CON IGUALDAD

Gracias por su interés en esta publicación de la CEPAL

Si desea recibir información oportuna sobre nuestros productos editoriales y actividades, le invitamos a registrarse. Podrá definir sus áreas de interés y acceder a nuestros productos en otros formatos.

Deseo registrarme

NACIONES UNIDAS
CEPAL

70 CEPAL
POR UN DESARROLLO SOSTENIBLE CON IGUALDAD

www.cepal.org/es/publications
facebook.com/publicacionesdelacepal
www.cepal.org/apps

SERIE
ESTUDIOS ESTADÍSTICOS
97

Desagregación de datos en encuestas de hogares

Metodologías de estimación en áreas pequeñas

Isabel Molina

NACIONES UNIDAS
CEPAL

Este documento fue preparado por Isabel Molina, Investigadora del Departamento de Estadística de la Universidad Carlos III de Madrid, con la colaboración de Andrés Gutiérrez, Experto Regional, y Álvaro Fuentes, Estadístico, de la Unidad de Estadísticas Sociales, División de Estadísticas de la Comisión Económica para América Latina y el Caribe (CEPAL), en el marco de las actividades del Programa sobre Estadísticas y Datos del décimo tramo de la Cuenta de las Naciones Unidas para el Desarrollo.

Publicación de las Naciones Unidas
ISSN: 1680-8789 (versión electrónica)
ISSN: 1994-7364
LC/TS.2018/82/Rev.1
Distribución: Limitada

Impreso en Naciones Unidas, Santiago
S.19-00419

Esta publicación debe ser citada como: I. Molina. "Desagregación de datos en encuestas de hogares: metodologías de estimación en áreas pequeñas'', *Series Estudios Estadísticos*, Nº 97, (LC/TS.2018/82/Rev.1), Santiago, Comisión Económica para América Latina y el Caribe, (CEPAL), 2019.

Índice

Cuadros

Gráficos

Resumen

Las encuestas de hogares son un instrumento ampliamente utilizado para obtener información sobre la situación socioeconómica y el bienestar de las personas. Sin embargo, la precisión de las estimaciones de las encuestas de hogares decrece sustancialmente cuando se trata de realizar inferencias para grupos poblacionales que representan desagregaciones para las cuales la encuesta no fue diseñada. En este contexto, es posible utilizar procesos de estimación que combinan la información de las encuestas de hogares con información auxiliar existente a nivel poblacional como censos o registros administrativos.

Este documento presenta una guía metodológica de la conjunción de técnicas estadísticas de las encuestas y modelos probabilísticos con el fin de producir desagregaciones para grupos de interés, conocidas como técnicas de estimación de áreas pequeñas (*en inglés, small area estimation o SAE).*

Después de describir el problema de las desagregaciones con datos insuficientes, el texto aborda tres conjuntos de técnicas que permiten lograr el objetivo propuesto. Primeramente, se revisan los estimadores directos (calculados directamente de las encuestas basándose solo en los datos del área correspondiente), que tienen la ventaja de estar libres de sesgo, aunque con baja precisión- cuando se aplican en desagregaciones. Luego, se estudian los estimadoresque toman prestada información de otras áreas para mejorar la precisión, apoyándose en información auxiliar poblacional. Se introducen los modelos probabilísticos para mejorar las propiedades estadísticas de los estimadores de interés. El modelamiento suele realizarse a dos niveles: al nivel de los individuos de interés (hogares o personas) o al nivel de las categorías de desagregación (subgrupos de interés). Se complementa la discusión de la teoría con ilustraciones y ejemplos que se apoyan en el software estadístico R. Por último, se presenta una aplicación práctica de algunos de los métodos revisados y se concluye acerca de la viabilidad de su uso en este problema específico.

Introducción

Las encuestas de hogares son insumo fundamental para la medición de las condiciones de vida de la población de un país y constituyen una herramienta esencial para la definición y seguimiento de las políticas públicas en diversos ámbitos. Esta fuente permite generar información precisa e insesgada a nivel nacional y para las desagregaciones consideradas en el diseño de la encuesta.

Existe una demanda creciente de información para grupos específicos de la población o para áreas geográficas menores. A manera de ejemplo, el marco global de indicadores para el seguimiento de los Objetivos del Desarrollo Sostenible plantea que la información debe ser desagregada no solo de manera geográfica (en subregiones de interés como provincias, municipalidades, comunas o localidades), sino también por grupos de ingreso, sexo, edad, raza, origen étnico, estatus migratorio y condición de discapacidad. No obstante, La confiabilidad en la inferencia de los indicadores disminuye a medida que lo hace el tamaño de muestra, por lo que generalmente no es posible lograr los niveles de desagregación deseados con una precisión adecuada.

Es así como en la última década ha tenido auge el concepto de la desagregación de datos, referida como información numérica que ha sido recolectada de diferentes fuentes, o medida a través de múltiples variables o incluso diferentes unidades de observación y que se compila en forma agregada y resumida. El propósito de esta agregación es presentar a la sociedad estimaciones de interés que gocen de buenas propiedades estadísticas para que sea posible que, a partir de ellas, se pueda extraer información e incluso formular políticas públicas en cada uno de los subgrupos de interés.

Este documento constituye una guía para la desagregación de datos estadísticos relacionados con las condiciones de vida de los individuos, de forma geográfica (a nivel regional) o por subgrupos de la población. El capítulo I comienza describiendo el problema de la desagregación de datos estadísticos (sección I.A); concretamente, se describe exactamente en qué situaciones se da este problema y se definen los términos y conceptos que se utilizan habitualmente y que también aparecerán a lo largo de este documento. A continuación (sección I.B) se pasa a establecer hasta qué nivel es conveniente desagregar los datos estadísticos ya que, a medida que se desagregan las estimaciones directas, debido a la disminución de los tamaños muestrales, aumentan los errores de muestreo que hacen dichas estimaciones demasiado volátiles y por tanto poco fiables. Por ejemplo, consideremos una población dividida sucesivamente a distintos niveles; por ejemplo, España está dividida en comunidades autónomas, que a su

vez se dividen provincias; éstas se dividen en comarcas, y finalmente las comarcas se dividen en municipios. En la Unión Europea en su conjunto, se utiliza la nomenclatura común NUT (Nomenclature des Unités Territoriales Statistiques) y los países (NUT0) se dividen en regiones llamadas NUT1, NUT2, etc. En la sección I.B se proporcionan indicaciones sobre el máximo nivel de desagregación de las estimaciones directas, nivel a partir del cual se pasaría a usar estimaciones indirectas. Éstas últimas son mucho más fiables debido a que toman prestada información de todas las áreas mediante el uso de distintas fuentes de datos. También se comenta en qué medida es recomendable utilizar dichos estimadores indirectos, pues es conveniente contener el posible sesgo de estos estimadores. Por tanto, se dan recomendaciones sobre los casos en los que sería prudente no producir estimación alguna. En cualquier caso, es posible rediseñar la encuesta para que abarque de forma más exhaustiva los dominios para los cuales se desean obtener los datos estadísticos. Hay que tener en cuenta que, a nivel local, la información o conocimiento que poseen las comunidades que habitan la zona podría contradecir los datos proporcionados. Por tanto, establecer hasta qué límite es aconsejable desagregar las estimaciones es esencial para que los datos producidos posean la suficiente calidad y sean realistas, no alejándose en gran medida del conocimiento local. Finalmente, en la sección I.C, se hace un recorrido por las distintas metodologías que proporcionan estimaciones indirectas, las cuales superan los límites de desagregación de las estimaciones directas. Concretamente, se revisan los estimadores indirectos básicos, que incluyen los sintéticos y los compuestos, y los estimadores basados en modelos, que quizás sean los más utilizados a la hora de obtener estimaciones fiables a niveles muy desagregados. Los estimadores "asistidos por modelos", que usan un modelo de trabajo, pero no requieren bondad de ajuste del mismo para mantener su insesgadez, se incluyen en el capítulo III junto con los métodos directos, ya que poseen buenas propiedades teóricas para áreas de tamaño muestral grande.

El capítulo II revisa distintos indicadores sobre la calidad de vida de los individuos; en concreto, medidas de pobreza y desigualdad. Se define de forma más detallada una familia de medidas de pobreza, llamada familia FGT, que se utilizará para ilustrar los distintos procedimientos de los capítulos siguientes. En la descripción de cada procedimiento se detallará cómo se aplicaría a la estimación de indicadores de esta familia y, para algunos de ellos, se realizarán ejemplos mediante el paquete de R `sae` (Molina y Marhuenda, 2015), que el lector puede reproducir.

A continuación, el capítulo III describe con detalle los estimadores directos habituales. Se incluyen estimadores directos básicos como el de Horvitz-Thompson y el de Hájek (sección III.A), así como estimadores asistidos por modelos; concretamente, los estimadores generalizados de regresión y los de calibración (sección III.B), junto con estimadores de sus errores de muestreo. Se ilustra el cálculo de estimadores directos en R mediante dos ejemplos.

El capítulo IV revisa algunos estimadores indirectos básicos como son el estimador post-estratificado sintético (sección IV.A), los estimadores sintéticos de regresión a nivel de área (sección IV.B) y a nivel de individuo (sección IV.C) y los estimadores compuestos (sección IV.D). Estos estimadores se incluyen únicamente debido a que ilustran de forma simple las ideas que subyacen a los métodos más sofisticados que se incluyen en el capítulo V. De nuevo, se incluyen dos ejemplos que ilustran el cálculo de los estimadores post-estratificados sintéticos y compuestos.

Los métodos basados en modelos del capítulo V son sensiblemente más realistas que los métodos indirectos básicos, y son más recomendables para utilizar en aplicaciones reales, pues proporcionan estimaciones potencialmente menos sesgadas. De entre los métodos basados en modelos, se incluyen los estimadores basados en el modelo más popular a nivel de área (sección V.A) y los basados en el modelo básico a nivel de individuo (sección V.B). Se ilustra cómo obtener estos estimadores en R mediante tres ejemplos. Se incluye en esta sección el método ELL, utilizado tradicionalmente por el Banco Mundial para la estimación de indicadores de pobreza y/o desigualdad (sección V.B), ya que en principio este método considera el modelo básico a nivel de individuo. No obstante, como se verá, dicho método es esencialmente sintético, y por tanto quizás debería incluirse en el capítulo IV dedicado a los estimadores sintéticos. Se describe también el método EB (sección V.D), que permite estimar indicadores generales al igual que el método ELL, pero que mejora dicho método al considerar que existe heterogeneidad entre las áreas y, consecuentemente, produce estimaciones más precisas. El procedimiento HB de la sección V.E obtiene estimaciones muy similares al método EB pero con menor coste computacional en el caso de

poblaciones grandes, especialmente a la hora de proporcionar las medidas de error de dichas estimaciones (el error cuadrático medio). Finalmente, la sección V.F describe métodos específicos para estimar indicadores que toman la forma de proporciones o medias de variables binarias. Aunque en principio también es posible utilizar otros métodos para estimar estos indicadores, como los de las secciones V.A o V.B, en general éstos pueden proporcionar estimaciones fuera del espacio natural de una proporción. En ciertos casos, las estimaciones obtenidas por distintos métodos pueden presentar solo ligeras diferencias.

Por otro lado, algunos de los métodos descritos son aplicables solamente a indicadores de tipo lineal; es decir, que sean aditivos en los valores de la variable de interés para las unidades del área, como medias o totales. Otros métodos; concretamente, los métodos basados en modelos a nivel de individuo ELL, EB y HB de las secciones V.C, V.D y V.E, están diseñados para poder estimar indicadores generales definidos como una función de los valores de una variable continua (e.g. los ingresos) en las unidades del área; valores para los cuales se asume un modelo. Los métodos basados en modelos a nivel de área son en principio aplicables a muchos tipos de indicadores, mientras se verifiquen las hipótesis necesarias, pero, en la práctica, es difícil que se verifiquen dichas hipótesis (como la insesgadez de los estimadores directos) para indicadores no lineales. Por tanto, en principio son más adecuados para la estimación de medias o totales en las áreas. En cualquier caso, a continuación de la descripción de cada método, se incluye un resumen detallando a qué indicadores podría ser aplicable, los requerimientos de datos necesarios aparte de las observaciones de la variable de interés obtenidos de una encuesta, y los pros y contras de cada método en comparación con los métodos que serían aplicables al mismo tipo de indicadores.

Es necesario señalar que no es posible describir con detalle todas las metodologías existentes por falta de espacio. Se incluye la descripción detallada de algunos de los métodos más ampliamente estudiados y con buenas propiedades. Éstos ayudarán a la comprensión de métodos más complejos. Sí se mencionan algunas de las extensiones de dichos métodos principales, redirigiendo al lector a la bibliografía correspondiente en caso de que necesite más información. No se incluyen métodos cuyas propiedades teóricas son desconocidas, aunque pudieran ser prometedores. Tampoco se detallan procedimientos que requieren excesiva formulación matemática, como por ejemplo la estimación del error cuadrático medio de los estimadores del capítulo IV y la sección V.B. En ambos casos, se redirige al lector a la bibliografía en la que puede encontrar dicho material. En el caso de los estimadores indirectos básicos del capítulo IV, otra razón por la cual no se incluye dicho material es que no se conocen estimadores fiables del error cuadrático medio de los estimadores que al mismo tiempo sean distintos para cada área. Existen estimadores excesivamente inestables pero distintos para cada área, o estables pero iguales para todas las áreas, pero no ambas cosas al mismo tiempo. Por tanto, este es un problema no resuelto.

Conviene realizar una matización importante sobre el enfoque bajo el cual se evalúa la calidad de un estimador. Existen tres enfoques alternativos para evaluar las propiedades de los estimadores, pero, a menudo, cada tipo de estimador se evalúa solamente usando una medida calculada respecto del enfoque natural para este estimador. Los estimadores directos y los indirectos básicos se evalúan respecto a la distribución del diseño muestral; es decir, respecto de todas las posibles muestras que se pueden extraer de la población mediante el diseño muestral de la encuesta concreta. En este caso, se considera que los valores de la variable de interés en las unidades de la población son valores fijos y únicamente varían (según un procedimiento aleatorio) las unidades que se seleccionan para la muestra. Un buen estimador es entonces aquel que tiene un buen comportamiento promedio para todas las posibles muestras, fijados los valores de la variable en las unidades de la población.

Por otro lado, los métodos de estimación basados en modelos se evalúan respecto de la distribución inducida por el modelo considerado, condicionando a la muestra observada. Es decir, se considera que los valores de la variable de interés en los individuos de la población son aleatorios, y están generados por un modelo (llamado modelo de superpoblación). Según este enfoque, el censo de nuestra variable es una realización posible de un vector aleatorio que sigue un modelo (o distribución de probabilidad). Los estimadores se evalúan respecto a todos los posibles censos que genera el modelo considerado; es decir, un buen estimador sería un estimador que se comporta bien en promedio para los infinitos posibles censos de valores de la variable de interés generados por el modelo, dejando constantes los individuos que aparecen en la muestra (aunque sus valores de la variable de interés varían ya que se extraen de un censo).

Finalmente, los métodos bayesianos, como el método HB de la sección V.E, se evalúan condicionando a las observaciones de la variable de interés en la muestra (distribución posterior). Es decir, un estimador será evaluado respecto de la distribución del indicador condicionada a los datos disponibles, en lugar de promediar para los posibles valores de dichos datos.

No existe un consenso acerca de cuál es el mejor enfoque para evaluar los estimadores de áreas pequeñas. El enfoque "bajo el diseño muestral" es no paramétrico pues no asume ningún modelo. Esto significa que la medida de error que se proporcione bajo este enfoque (habitualmente el error cuadrático medio) recoge el error de estimación a través de las posibles muestras, sin necesidad de que se verifique un modelo. Este es el enfoque preferido por los estadísticos de los organismos oficiales. El enfoque "bajo el modelo" asume un modelo, pero se fija la muestra obtenida, proporcionando el error para la muestra concreta que se tiene en lugar de un promedio para todas las posibles muestras que se podrían extraer. En este enfoque, la medida de error recoge la incertidumbre a lo largo de los posibles censos que el modelo genera; es decir, a través de las posibles realidades que pudieran suceder, variando también los valores observados en la muestra. Por último, el enfoque bayesiano considera que los indicadores en cuestión son realizaciones de variables aleatorias que siguen una distribución, y proporcionan medidas de error en forma de descriptivos de la distribución de dichos indicadores, condicionando a los valores observados de la muestra; es decir, para las observaciones concretas de la muestra que se han obtenido, en lugar de promediar para todos los posibles valores de éstas.

Como ya se ha mencionado, cada método de estimación se evalúa habitualmente atendiendo a su enfoque natural. Es decir, las medidas de error que acompañan a las estimaciones para asesorar su calidad; concretamente, los errores cuadráticos medios, se calculan habitualmente respecto del enfoque bajo el que se obtienen las estimaciones. Esto significa que los errores cuadráticos medios de distintos estimadores, al ser obtenidos bajo enfoques diferentes, no son directamente comparables. Sin embargo, es conocido que, si las hipótesis que asumen los modelos considerados se verifican, dichos errores cuadráticos medios sí son comparables al promediar a lo largo de un gran número de áreas del mismo tamaño muestral. Además, los errores cuadráticos medios bajo el diseño de los estimadores basados en modelos no son fáciles de estimar y no se conocen estimadores aceptables. Por otro lado, realizando previamente un chequeo del modelo para comprobar que se ajusta adecuadamente a los datos disponibles, los estimadores de los errores cuadráticos medios bajo el modelo, que son relativamente estables, pueden compararse con los errores cuadráticos medios bajo el diseño.

I. El problema de desagregación de datos (o estimación en áreas pequeñas)

A. Descripción del problema

Las encuestas oficiales que realizan los Institutos Nacionales de Estadística, así como los Institutos Regionales de Estadística y otros organismos o instituciones a nivel supranacional o internacional, están diseñadas para producir datos estadísticos a un nivel de agregación determinado; es decir, para subdivisiones, bien geográficas, bien socio-económicas, de la población. Por ejemplo, el Módulo de Condiciones Socioeconómicas (MCS) de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) de México está diseñado para proporcionar estimaciones de indicadores de pobreza y desigualdad a nivel nacional y para las 32 entidades federativas (31 estados y la Ciudad de México) desagregando por zonas rurales y urbanas, cada dos años. Sin embargo, en este país existe un mandato para producir estimaciones cada 5 años a nivel de municipio. Esta situación ocurre frecuentemente también en otros países y ámbitos; es decir, una vez efectuada una encuesta, con tamaños muestrales establecidos para producir estimaciones fiables a un nivel de agregación determinado, a posteriori se produce una demanda de datos a un nivel más desagregado. Para ello, se desea poder utilizar los datos de dicha encuesta sin incurrir en costes adicionales debidos a un incremento de la muestra. Sin embargo, en el caso de México, las submuestras de la ENIGH tomadas de cada municipio no permiten la obtención de estimadores directos fiables en todos ellos; de hecho, más de la mitad de los municipios carecen de observaciones. Este es el problema que surge frecuentemente al desear producir datos estadísticos para subdivisiones más pequeñas de las que originariamente estaba planificado producir.

Con anterioridad a la implementación de la encuesta, sería posible mejorar aspectos del diseño muestral que eviten en cierto grado este problema. Por ejemplo, es posible aumentar los tamaños muestrales en las áreas en las que sea necesario (con el correspondiente aumento del coste) o distribuir el tamaño muestral total de la encuesta entre las áreas de una forma más eficiente. A pesar de que existen diversos mecanismos para mejorar el diseño muestral y disponer de un mínimo suficiente de datos en todas las subdivisiones de la población, sin embargo "el cliente siempre demanda más de lo que se ha especificado en la etapa del diseño" (Fuller, 1999).

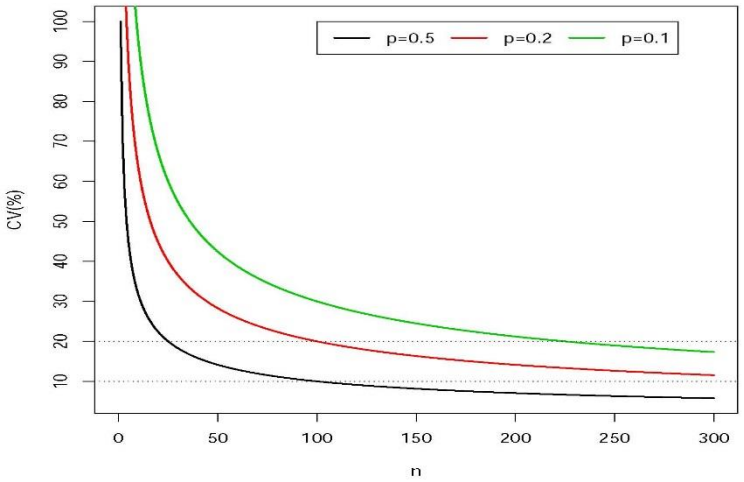
En la literatura, las subdivisiones para las cuales se desea obtener datos estadísticos (o estimaciones) se denominan comúnmente "áreas" o "dominios", con independencia de si son delimitaciones geográficas, o socio-económicas, o un cruce de ambos tipos. A la hora de estimar un indicador concreto en una de estas áreas, llamamos estimador directo a un estimador que usa solamente los datos de la encuesta para esa área. Los estimadores directos habituales son insesgados o prácticamente insesgados respecto a la distribución del diseño muestral, e.g., a lo largo de todas las posibles muestras que se pueden extraer de la población mediante el correspondiente diseño muestral. No obstante, si la encuesta no se había planificado para estimar a un nivel tan desagregado, el tamaño muestral en algunas de las áreas puede ser demasiado pequeño, lo cual se traduce en errores de muestreo excesivamente grandes para los estimadores directos de los indicadores de interés en esas áreas. Las áreas en las que esto ocurre, independientemente de su tamaño poblacional, se conocen en la literatura como "áreas pequeñas". Por tanto, no es el tamaño poblacional del área el que le confiere el adjetivo de "pequeña" pues, en muchos casos, áreas de gran tamaño poblacional (e.g. los estados de EE.UU.) se consideran áreas pequeñas si no se dispone de estimaciones directas de suficiente calidad. Concretamente, el término "área pequeña" hace referencia a áreas, en las cuales el estimador directo del indicador de interés es ineficiente debido al insuficiente número de observaciones obtenidas (o encuestas realizadas) de esa área. Por ejemplo, a la hora de producir estimaciones de indicadores de pobreza y desigualdad basados en el MCS de la ENIGH de México, los municipios se considerarían áreas pequeñas, puesto que la encuesta no está diseñada para obtener estimaciones precisas para éstos.

A menudo, las estimaciones a nivel geográfico muy detallado se representan en forma de cartogramas o mapas que muestran las correspondientes regiones con distintas tonalidades o colores representando los distintos grados en la magnitud del indicador de interés. Por ejemplo, el Banco Mundial produce mapas desagregados de pobreza o desigualdad para un gran número de países, véase e.g. Elbers, Lanjouw y Lanjouw (2003). Estos mapas, así como las estimaciones concretas, son una herramienta esencial para la monitorización de las condiciones de vida en las distintas regiones de un país y se utilizan por los gobiernos y los organismos internacionales para planificar las políticas de desarrollo regional. Es muy conveniente suplementar las estimaciones con medidas de calidad de las mismas (habitualmente los errores de muestreo). Al igual que las estimaciones, éstas también se pueden plasmar en un mapa.

B. Límites de la desagregación de datos estadísticos

Aunque no existe una definición formal, como se ha mencionado, un área se denomina "pequeña" cuando el error de muestreo del estimador directo considerado para el indicador de interés no es aceptable. Sin embargo, no existe un límite superior universal para este error de muestreo por encima del cual el área donde se estima se considere "pequeña". Cada Instituto Nacional de Estadística u organismo internacional establece su propio límite para el error de muestreo relativo o coeficiente de variación (CV), a partir del cual los datos estadísticos se consideran no fiables y por tanto no se publican. En ocasiones, estos datos se publican con alguna indicación de que carecen de la calidad exigida. Tampoco existe un tamaño muestral concreto, por debajo del cual el área se considere pequeña, pues el error de muestreo varía no solo dependiendo del tamaño muestral, sino también del indicador que se estima y del estimador concreto que se utiliza. Por ejemplo, a la hora de estimar la media de una variable continua (e.g. los ingresos medios) con un error de muestreo máximo determinado, a menudo se necesita un tamaño muestral menor que para estimar la proporción de individuos que posee una característica determinada (media de una variable binaria), especialmente si dicha característica es muy poco o muy común; es decir, cuando la proporción real está cerca de cero o de uno. El Gráfico 1 ilustra el tamaño muestral mínimo necesario para obtener un CV máximo concreto para la proporción muestral bajo muestreo aleatorio simple. Se puede ver que el tamaño muestral necesario varía dependiendo del verdadero valor de la proporción que se desea estimar. Concretamente, este gráfico muestra cómo, en el caso de que la verdadera proporción sea $p = 0.5$, un tamaño muestral de aproximadamente $n = 25$ es suficiente para asegurar un CV de la proporción muestral por debajo del 20%, mientras que para $p = 0.2$ se necesita al menos $n = 100$ unidades, y para $p = 0.1$ se necesitan más de $n = 200$ unidades en la muestra. Por tanto, no es posible establecer un tamaño muestral mínimo de las áreas que garantice el nivel de eficiencia deseado para cualquier estimador y/o cualquier indicador objetivo.

Gráfico 1
CV de la proporción muestral $\hat{p}$ en función del tamaño muestral n, para cada valor de la verdadera proporción p.
(En porcentajes)



Fuente: Elaboración propia.

En particular, algunos indicadores de pobreza habituales son proporciones. Por ejemplo, la incidencia de pobreza, también llamada tasa de individuos en riesgo de pobreza es la proporción de individuos con ingresos por debajo del umbral o línea de pobreza. Este umbral es el valor de los ingresos (netos equivalentes) por debajo del cual un individuo se considera en riesgo de pobreza o exclusión. Igualmente, ciertos tipos de carencias se miden como las proporciones de individuos con acceso a determinados servicios básicos (de salud, en la vivienda, alimentación, etc.). Como se ha mencionado, el tamaño muestral necesario para obtener estimaciones directas de estos indicadores con suficiente calidad es habitualmente mayor que el que se necesita para estimar medias o totales de variables cuantitativas.

A pesar de no existir límites superiores universales para los errores de muestreo (ni límites inferiores para los tamaños muestrales) a partir de los cuales se considera que los datos estadísticos no tienen la suficiente calidad, algunos Institutos Nacionales de Estadística coinciden en establecer un dato como "no publicable" cuando su error de muestreo relativo o CV supera un 20%. Por tanto, para estas instituciones, las áreas para las cuales las estimaciones directas de un determinado indicador de interés poseen un CV superior a ese valor se considerarían "pequeñas" para estos indicadores. Por ejemplo, los municipios de México serían áreas pequeñas a la hora de estimar indicadores de pobreza a partir de la ENIGH. En estas áreas, sería necesario aumentar los tamaños muestrales de la encuesta o utilizar métodos "indirectos" para producir datos estadísticos con la calidad suficiente para ser publicables.

Los métodos "indirectos" de estimación no consideran solamente los datos muestrales relativos al dominio o área de interés, sino que además usan los datos muestrales provenientes de otras áreas o dominios. Estos estimadores utilizan la información de otras variables (llamadas variables auxiliares) que están relacionadas con la variable de interés. Esta relación se considera similar para todas las áreas y se representa a través de un modelo que las enlaza por medio de parámetros comunes. Al estimar los parámetros comunes a todas las áreas usando todos los datos de la muestra (en total la muestra suele ser grande), el uso de una mayor cantidad de información proporciona estimadores más eficientes (comparados con los directos). Estos estimadores suelen comprometer ligeramente el sesgo bajo el diseño, a cambio de aumentar en gran medida la eficiencia global del estimador, evaluada en términos de error cuadrático medio.

La ganancia en eficiencia de los estimadores indirectos respecto de los directos es mayor cuanto menor es el tamaño muestral del área. Sin embargo, éstos suelen mejorar en la mayoría de las áreas, incluyendo muchas de tamaño muestral grande. De hecho, algunos estimadores indirectos (véase capítulo V) poseen la buena propiedad de converger a un estimador directo al aumentar el tamaño muestral del área. Por tanto, los estimadores indirectos que poseen esta propiedad se pueden utilizar para todas las áreas, independientemente de si son "pequeñas" o no, con lo cual se reduce la importancia de disponer de una definición más exacta o formal de área pequeña.

No obstante, en la práctica se hace necesario determinar hasta qué nivel de desagregación es conveniente seguir utilizando los estimadores directos convencionales, a partir de qué nivel acudir a los estimadores indirectos, e incluso si conviene producir datos estadísticos para cualquier nivel de desagregación posible, que en el límite sería a nivel de individuo. En virtud de lo comentado anteriormente, conviene utilizar los estimadores directos al nivel para el cual los CVs de estos estimadores no superan el límite establecido para ninguna de las áreas. En caso de superar el límite mencionado para algún área, sería más recomendable utilizar estimadores indirectos para las áreas de ese nivel.

Conviene recalcar que no es recomendable producir estimaciones para cualquier área ya que, si el modelo no se verifica exactamente (prácticamente ningún modelo es exactamente cierto), el sesgo bajo el diseño de los estimadores indirectos aumenta al disminuir el tamaño muestral. Aunque el error cuadrático medio de los estimadores indirectos se mantenga menor que el del estimador directo, no es recomendable comprometer excesivamente el sesgo bajo el diseño. Consecuentemente, tiene sentido establecer un límite superior para el sesgo absoluto relativo de un estimador indirecto y decidir no producir datos para las áreas en las que éste supere dicho límite. Este límite se fijará en función de las exigencias del usuario de los datos (e.g. un 10% o un 5% de sesgo absoluto relativo). Así, se recomienda:

(i) Utilizar los estimadores directos para la población completa y para los niveles de agregación mayores, mientras los estimadores directos para todas las áreas de ese nivel posean un CV por debajo del límite establecido.

(ii) Para niveles más desagregados, usaremos estimadores indirectos en las áreas para las cuales el sesgo absoluto relativo no supere la cantidad máxima prefijada.

(iii) Finalmente, para las áreas donde los estimadores indirectos tengan sesgo absoluto relativo por encima de la cantidad máxima, se aconsejaría no producir estimaciones o modificar el diseño de la encuesta para conseguir disponer de un tamaño muestral mínimo en todas las áreas de interés.

El sesgo bajo el diseño de un estimador no es algo conocido pues depende del verdadero valor del indicador en cuestión. En algún caso, se puede aproximar de forma teórica. Otra opción es obtenerlo de forma empírica a partir de experimentos de simulación. Estos experimentos se pueden realizar imitando en lo posible la realidad, e.g., basados en un censo anterior o utilizando los datos de la encuesta para generar un censo y extrayendo muestras de éste. Estos experimentos tienen una utilidad adicional muy importante, que es la validación de los métodos de estimación en situaciones en las que se conocen los verdaderos valores. En ambos casos, es posible determinar el tamaño muestral mínimo de las áreas que sería necesario para no sobrepasar el límite superior del sesgo absoluto relativo del estimador del indicador en cuestión, véase la sección V.A. Así, a la hora de producir estimaciones con los datos reales, se utilizarían los estimadores indirectos solamente en las áreas para las cuales el tamaño muestral supera a este tamaño muestral mínimo.

C. Metodologías para superar los límites de la desagregación

Como se ha mencionado, si se desea evitar incrementos de la muestra debido a los correspondientes costes, o si la demanda de datos a un nivel más desagregado se ha producido una vez se ha efectuado la encuesta, una forma efectiva en términos de costes de obtener estimadores para todas las áreas de interés más fiables que los directos es acudir a métodos indirectos. Estos métodos no utilizan solamente los datos de la encuesta para el área correspondiente, sino que usan datos de otras áreas asumiendo algún tipo de similitud con el área en cuestión. Esta similitud se suele representar a través de un modelo (que representa un conjunto de hipótesis). Los estimadores indirectos más simples se basan en hipótesis poco realistas y por

tanto pueden poseer un sesgo considerable. Éstos incluyen los estimadores sintéticos, que no recogen la heterogenidad que suele existir entre las áreas. Estimadores sintéticos muy conocidos son el estimador post-estratificado sintético y el estimador sintético de regresión (capítulo III). Otros estimadores indirectos clásicos son los compuestos, que se calculan como una media ponderada entre un estimador directo y otro sintético, e incluyen el conocido estimador dependiente del tamaño muestral o los estimadores compuestos óptimos. El peso que se otorga a cada estimador no depende de la bondad de ajuste del modelo que asume el estimador sintético. Además, en la práctica el peso del estimador directo suele estar cerca de uno y por tanto se toma prestada escasa información.

Estimadores indirectos algo más sofisticados, que consideran la existencia de heterogeneidad entre las áreas, son los basados en modelos de regresión. Existen dos grandes grupos de modelos de regresión utilizados para la estimación en áreas pequeñas: los modelos a nivel de área y los modelos a nivel de individuo, aunque también es posible establecer modelos a niveles de agregación intermedios (e.g. por grupos de sexo-edad dentro de las áreas). Los modelos a nivel de área utilizan solamente datos agregados para las áreas o dominios de estimación. Habitualmente, este tipo de datos se pueden conseguir con menores restricciones, ya que la agregación evita los problemas de confidencialidad. Modelos lineales a nivel de área muy utilizados son los llamados modelos Fay-Herriot (FH), propuestos por Fay y Herriot (1979). Estos modelos tienen una estructura a dos niveles. En el primer nivel, se considera que la relación entre los indicadores de interés para las áreas y las variables auxiliares a nivel de área disponibles es constante para todas las áreas. Por ejemplo, se considera que la disminución en los ingresos medios al pasar de estar trabajando a estar desempleado, manteniendo constantes otras variables, es igual en todas las áreas. Así, se enlazan todas las áreas a través de un modelo de regresión lineal. En el segundo nivel, se asume que, dados los verdaderos valores de los indicadores de interés, los estimadores directos de las áreas son centrados en estos verdaderos valores y con varianzas que se asumen conocidas. Dichas varianzas típicamente varían entre áreas debido a que los tamaños muestrales de las áreas son distintos. Estos modelos han tenido un merecido éxito debido a que los estimadores resultantes para las áreas son una composición o media ponderada entre los estimadores directos y los estimadores sintéticos de regresión. Cuando el modelo sintético no se ajusta bien a los datos (i.e., las variables auxiliares consideradas no explican suficientemente la heterogeneidad del indicador a lo largo de las áreas) o el tamaño muestral de un área es grande, el estimador basado en el modelo FH le da mayor peso al estimador directo, el cual es lo suficientemente preciso. Por el contrario, cuando el modelo sintético se ajusta bien o el tamaño muestral del área es pequeño (estimador directo impreciso), aumenta el peso otorgado al estimador sintético de regresión. En este caso, se aumenta la eficiencia debido a que el estimador sintético posee un coeficiente de regresión común a todas las áreas y por tanto se estima usando los datos de todas las áreas. Además, dado que los estimadores directos son aproximadamente insesgados respecto del diseño muestral, para las áreas con tamaños muestrales mayores, los estimadores obtenidos a partir del modelo Fay-Herriot preservan también un sesgo pequeño bajo el diseño. Una dificultad de estos modelos es determinar los valores de las varianzas de los estimadores directos (o varianzas heteroscedásticas de los términos de error del modelo). Aunque, como se ha mencionado, estas varianzas se asumen conocidas, en la práctica se reemplazan por estimaciones. Dado el pequeño número de datos en algunas de las áreas, las estimaciones de estas varianzas también son muy imprecisas. Existen métodos de suavizado como el método de la función generalizada de varianza (véase Fay y Herriot, 1979) o de estimación no paramétrica de estas varianzas, véase González-Manteiga et al. (2010). La estimación de dichas varianzas añade el problema de incorporar el correspondiente error de estimación en el error del estimador final, aunque algunos estudios indican que dicha contribución es pequeña.

En los modelos a nivel de individuo, como su nombre indica, el modelo se establece para cada individuo de la población (modelo de superpoblación), y por tanto el ajuste de estos modelos requiere disponer de datos individuales de la variable respuesta y las variables auxiliares. El primer modelo de este tipo fue propuesto por Battesse, Harter y Fuller (1988) y se conoce como modelo de errores anidados. Este es un modelo de regresión lineal que incluye, además de los errores individuales del modelo, efectos aleatorios asociados a las áreas, los cuales representan la heterogeneidad entre las áreas no explicada por las variables auxiliares disponibles. Estos modelos son ampliamente utilizados en la actualidad cuando se dispone de los datos necesarios, pues incorporan mucha más información que los modelos a nivel de área y no es necesario conocer las varianzas de los errores del modelo.

El hecho de asumir un modelo estocástico que genera los valores de la variable de interés en los individuos de la población hace que los indicadores de interés sean cantidades aleatorias. Así, es común en la literatura usar el término "predecir", en lugar de "estimar", el valor del indicador de interés y "predictor" en lugar de "estimador". En este documento, se usarán ambos términos como sinónimos. En este contexto, un predictor insesgado de un indicador es aquel cuya esperanza bajo el modelo coincide con la esperanza de dicho indicador. A la hora de estimar indicadores de tipo lineal en los valores de la variable de interés en los individuos de la población, como medias o totales, los modelos básicos a nivel de área o de individuo que se utilizan forman parte de los modelos lineales mixtos que incluyen efectos aleatorios de las áreas de interés. Bajo estos modelos, el estimador indirecto habitual es el mejor predictor lineal insesgado (en inglés, *best linear unbiased predictor*, BLUP), que consiste en la combinación lineal de los valores observados de la variable respuesta en los individuos de la muestra, que es insesgado bajo el modelo y minimiza el error cuadrático medio. El BLUP depende de los parámetros desconocidos del modelo, los cuales representan el comportamiento común entre las áreas. Reemplazando estos parámetros desconocidos por estimadores se obtiene el BLUP empírico (en inglés, *empirical BLUP*, EBLUP). Este es finalmente el estimador (o predictor) habitual basado en un modelo de un indicador lineal en un área pequeña.

El BLUP no requiere ninguna hipótesis de normalidad en el modelo. Por otro lado, para estimar indicadores más generales que los lineales, el mejor predictor (en inglés, *best predictor*) es el que minimiza el error cuadrático medio, sin exigir que sea lineal ni insesgado. Éste es igual a la esperanza bajo el modelo del indicador a estimar, condicionada a los valores observados en la muestra. Bajo normalidad, el mejor predictor de un indicador lineal, una vez estimado el parámetro de regresión mediante mínimos cuadrados ponderados, coincide con el BLUP. Cuando no existe normalidad o cuando el indicador a estimar no es lineal, es posible que la esperanza que define el mejor predictor no se pueda calcular de forma analítica. En ese caso, se recurre a aproximaciones numéricas del mejor predictor. Otros modelos muy utilizados, por ejemplo a la hora de estimar proporciones de variables binarias, son los modelos lineales generalizados con efectos aleatorios (véase el capítulo V).

Consideremos ahora una población que está dividida en dominios, y a su vez estos dominios están divididos en subdominios, y se desea estimar en uno de los niveles o en los dos. Por ejemplo, México está dividido en 31 Estados más Ciudad de México y, a su vez, cada Estado se divide en un número de municipios. Modelos más apropiados para esta situación incluyen efectos aleatorios a varios niveles (véase e.g. Stukel y Rao, 1999 para la estimación de indicadores lineales o Marhuenda et al., 2018 para la estimación de indicadores generales). Por otro lado, cuando existen diversas variables de interés relacionadas entre sí, se pueden plantear modelos multivariantes (véase Fay, 1987 o Datta, Fay y Ghosh, 1991). Asimismo, cuando existe correlación temporal y/o espacial, se puede recurrir a modelos que incluyen efectos aleatorios que siguen un proceso de series temporales y/o un proceso espacial, véase e.g. Pfeffermann y Burk (1990) o Rao y Yu (1992) para modelos temporales, Molina, Salvati y Pratesi (2008) para un modelo espacial y Marhuenda, Molina y Morales (2013) para un modelo espacio-temporal. Por otro lado, los modelos Bayesianos son una alternativa a los modelos frecuentistas que en muchas ocasiones presentan ventajas computacionales, proporcionando estimaciones prácticamente idénticas a las obtenidas con el correspondiente modelo frecuentista mientras las distribuciones previas consideradas sean no informativas (véase el capítulo IV). La monografía de Rao y Molina (2015) describe con detalle las técnicas más utilizadas en estimación en áreas pequeñas y realiza una revisión concienzuda de la mayoría de los trabajos realizados en este campo hasta la fecha de publicación.

II. Indicadores comunes de pobreza y desigualdad

En la literatura existen infinidad de indicadores de pobreza y desigualdad que resumen distintos aspectos sobre las condiciones de vida de una población. En efecto, a partir de las encuestas oficiales sobre condiciones de vida de los distintos países, los Institutos Nacionales de Estadística suelen producir una gran variedad de indicadores con el objeto de representar las distintas dimensiones de la pobreza o la desigualdad. La forma matemática del indicador concreto que se desea estimar es de gran importancia a la hora de seleccionar las técnicas de estimación en áreas pequeñas apropiadas, pues no todas las técnicas son aplicables a cualquier tipo de estimador.

En este capítulo vamos a revisar muchos de los indicadores que aparecen en la literatura, así como los indicadores que se suelen producir a partir de las encuestas oficiales de condiciones de vida. Aunque no es posible incluir todos los indicadores existentes, algunos de los que se describen en este capítulo se utilizarán para ilustrar las técnicas de estimación en áreas pequeñas más utilizadas para nuestro propósito. Así, en los capítulos que vienen a continuación se va a realizar una revisión de los distintos métodos que se pueden utilizar, con una indicación sobre los tipos de indicadores a los que son aplicables.

Neri, Ballini y Betti (2005) hacen una revisión de indicadores de pobreza y desigualdad. El indicador de pobreza más utilizado es la incidencia o tasa de pobreza, también llamada tasa en riesgo de pobreza, que se calcula como la proporción de individuos con ingresos (netos equivalentes) por debajo del umbral o línea de la pobreza. Otro indicador común es la brecha de la pobreza, que mide la magnitud de la pobreza en lugar de la frecuencia de individuos en riesgo de pobreza. Estos dos indicadores son elementos de una familia de indicadores más amplia definida por Foster, Greer y Thorbecke (1984), que llamaremos familia de indicadores FGT, y que poseen la ventaja de ser aditivos en los individuos. Los métodos de estimación en áreas pequeñas que vamos a describir en capítulos posteriores se van a ilustrar aplicándolos a algunos de los indicadores de esta familia, aunque es importante recalcar que algunos métodos son aplicables a muchos otros indicadores no incluidos en esta familia. En cada capítulo se matizará a qué tipo de indicadores es aplicable cada método.

Llamenos U a la población objetivo (por ejemplo los residentes en un país), de tamaño N, que está dividida en D subpoblaciones, las áreas o dominios donde se desea estimar, de tamaños $N_1, \dots, N_D$. Obsérvese que los tamaños poblacionales de las áreas suelen ser muy grandes pues, como se ha comentado

en el capítulo II, el término "área pequeña" hace referencia al tamaño muestral (más concretamente al error de muestreo del estimador directo utilizado) y no al tamaño poblacional.

Llamamos E_{di} a la medida de poder adquisitivo (e.g. ingresos o gastos) del individuo i en el área $d, d = 1, \dots, D$. Llamamos z al umbral de pobreza utilizado, por debajo del cual un individuo se considera en riesgo de pobreza. La familia de indicadores FGT para el área d se define por :

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

donde $I(E_{di} < z)$ es una función indicadora, que toma el valor 1 si $E_{di} < z$ (individuo i en riesgo de pobreza) o el valor 0 en caso contrario. Tomando $\alpha = 0$, obtenemos la tasa o incidencia de pobreza. La brecha de pobreza es el indicador obtenido tomando $\alpha = 1$.

Un indicador más complejo que utiliza tanto la brecha como la incidencia de pobreza, además del índice de Gini, es el Índice Sen (Sen, 1976). Por otro lado, dentro de los indicadores que no dependen de un umbral de pobreza sino de la situación relativa de los individuos dentro de la ordenación de todos ellos, podemos mencionar los índices Fuzzy monetarios y los Fuzzy suplementarios, véase Betti et al. (2006). Por encima de la dimensión monetaria, a menudo es de interés medir otro tipo de limitaciones o carencias no estrictamente monetarias. Estas carencias se suelen medir como proporciones de individuos que poseen (o que no poseen) acceso a determinados servicios (sanitarios, en el hogar, educativos,...). Por otro lado, indicadores de desigualdad incluyen el Índice de Gini, la entropía generalizada o el Índice de Theil, véase e.g. Neri, Ballini y Betti (2005).

En el Consejo Europeo de diciembre de 2001, como parte de la Estrategia de Lisboa del año 2000 para la coordinación de las políticas sociales de los distintos países, se establecieron un grupo de indicadores de pobreza y exclusión social, llamados indicadores de Laeken. Estos indicadores incluyen la tasa en riesgo de pobreza F_{0d}, la razón entre quintiles de los ingresos (razón entre los ingresos del 20% de la población más rica y los del 20% más pobre; en inglés, *quintile share ratio*), la mediana relativa de la brecha en riesgo de pobreza y el índice de Gini entre otros.

Un ejemplo de medida multidimensional de pobreza es el utilizado por el CONEVAL en México, llamado indicador de pobreza multidimensional, que mide la proporción de individuos con al menos una de entre un conjunto de deprivaciones o carencias establecidas, y cuyos ingresos están por debajo del umbral o línea de bienestar. En los capítulos siguientes se revisarán algunos métodos de estimación en áreas pequeñas que, aunque se ilustren estimando indicadores de la familia FGT, se pueden utilizar de la misma forma para estimar una gran cantidad de indicadores.

III. Métodos directos para la desagregación de datos de pobreza

En este capítulo se describen estimadores directos básicos para la media de una variable en un dominio o área, dada por

$$\bar{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} Y_{di}, \tag{2}$$

donde Y_{di} denota el valor de la variable para el individuo i dentro del área (o dominio) d. Obsérvese que los indicadores FGT dados en (1) también se pueden escribir en forma de medias como en (2) llamando

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

con lo cual obtenemos que $F_{\alpha d}$ es la media de los valores $Y_{di} = F_{\alpha,di}$ para los individuos en el área d, es decir,

$$F_{\alpha d} = N_d^{-1} \sum_{i=1}^{N_d} F_{\alpha,di}. \tag{3}$$

Como ya se ha mencionado, un estimador de un indicador en un área determinada se califica como "directo" si se calcula utilizando solamente datos de esa área, sin hacer uso de datos de ninguna otra área. Estos estimadores son los utilizados por defecto en los Institutos Nacionales de Estadística, debido a sus buenas propiedades respecto del diseño muestral (como la insesgadez) en áreas con suficiente tamaño muestral. Por ejemplo, los estimadores directos han sido usados tradicionalmente para producir estadísticas sobre las condiciones de vida en Chile a nivel nacional, regional y para un conjunto de comunas con muestra representativa, según la Encuesta de Caracterización Socioeconómica Nacional (CASEN). A partir de la CASEN 2015, la metodología para la estimación en las comunas no representativas se realiza usando métodos indirectos basados en modelos; concretamente, el método Fay-Herriot descrito en la introducción, véase el documento sobre Metodología de estimación de pobreza a nivel comunal, con datos de Casen 2015 del Observatorio Social del Ministerio de Desarrollo Social de Chile de 2017.

En este documento, llamaremos s a la muestra extraída de tamaño n de la población U, s_d a la submuestra del área d de tamaño n_d (que puede ser igual a cero) y r_d al conjunto de elementos fuera de la muestra de la misma área, $d = 1, \dots, D$, donde $\sum_{d=1}^{D} n_d = n$. Además, llamaremos π_{di} a la probabilidad de inclusión del individuo i en la muestra del área d, $w_{di} = \pi_{di}^{-1}$ al peso muestral del mismo individuo y $\pi_{d,ij}$ a la probabilidad de inclusión de los individuos i y j en la muestra del área d. A continuación describimos los estimadores directos más conocidos.

A. Estimadores directos básicos

El estimador insesgado con respecto al diseño muestral de la media del área d, $\bar{Y}_d$, es el conocido como estimador de Horvitz-Thompson (HT). Este estimador require conocer el verdadero tamaño del área N_d y los pesos muestrales $w_{di} = \pi_{di}^{-1}$ para los individuos de la muestra en el área d. Asumiendo que éstos son conocidos, el estimador de HT de $\bar{Y}_d$ es

$$\hat{\bar{Y}}_d = N_d^{-1} \sum_{i \in s_d} w_{di}\, Y_{di}. \tag{4}$$

Obsérvese que para el total del área d, $Y_d = \sum_{i=1}^{N_d} Y_{di}$, el estimador de HT es simplemente $\hat{Y}_d = \sum_{i \in s_d} w_{di}\, Y_{di}$ y no requiere conocer el tamaño poblacional del área N_d.

Si $\pi_{di} > 0$ para todo $i = 1, \dots, N_d$, un estimador insesgado de la varianza bajo el diseño del estimador de HT de $\bar{Y}_d$ viene dado por

$$\widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d) = N_d^{-2} \left\{ \sum_{i \in s_d} \frac{Y_{di}^2}{\pi_{di}^2}(1 - \pi_{di}) + 2 \sum_{i \in s_d} \sum_{\substack{j \in s_d \\ j > i}} \frac{Y_{di} Y_{dj}}{\pi_{di}\pi_{dj}} \left(\frac{\pi_{d,ij} - \pi_{di}\pi_{dj}}{\pi_{d,ij}} \right) \right\}. \tag{5}$$

En muchas ocasiones, en la fase de estimación, no se dispone de toda la información sobre el diseño muestral aparte de los pesos muestrales w_{di}. Al no disponer de las probabilidades de inclusión de segundo orden $\pi_{d,ij}$, el estimador (5) no se puede calcular. Sin embargo, para diseños muestrales con probabilidades de inclusión de segundo orden verificando $\pi_{d,ij} \approx \pi_{di}\pi_{dj}$, para $j \neq i$, como por ejemplo en el muestreo de Poisson, donde se da la igualdad, el segundo término de (5) se hace aproximadamente cero. Reemplazando además $w_{di} = \pi_{di}^{-1}$, obtenemos el siguiente estimador de la varianza, que no depende de las probabilidades de inclusión de segundo orden

$$\widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d) = N_d^{-2} \sum_{i \in s_d} w_{di}\,(w_{di} - 1) Y_{di}^2. \tag{6}$$

Este estimador es el que proporciona la función direct() del paquete de R sae, que se utilizará en el Ejemplo 1 para ilustrar estos procedimientos, cuando se incluyen los pesos muestrales. Esta función asume que no se dispone de información sobre el diseño muestral aparte de los pesos muestrales. En el caso de poseer información sobre el diseño muestral, existen paquetes de R más adecuados como survey (Lumley 2017) o sampling (Tillé y Matei 2016). Además, existen aproximaciones alternativas de la varianza dependiendo del diseño muestral y de la información disponible, e.g. el método de los conglomerados últimos o el método de las replicaciones repetidas balanceadas (en inglés, *Balanced Repeated Replications*, BRR) con la corrección de Fay (U.S. Bureau of Labor Statistics y U.S. Census Bureau 2006).

El estimador de HT pondera las observaciones individuales Y_{di} mediante los pesos muestrales o inversos de las probabilidades de inclusión en la muestra, $w_{di} = \pi_{di}^{-1}$. Esto protege de las situaciones en las que la probabilidad de seleccionar a un individuo está relacionada con el valor de la variable de interés (diseño muestral informativo). En efecto, si cierto tipo de individuos (e.g. aquellos con menores ingresos) tiene una mayor probabilidad de aparecer en la muestra, es probable que este tipo de individuos aparezca más frecuentemente en la muestra final, mientras que los individuos con menor probabilidad de aparecer

(e.g. con mayores ingresos) probablemente escaseen en la muestra. Esto significa que, si estimásemos dando el mismo peso a todas las observaciones de la muestra, al igual que en la media muestral básica, tendríamos un sesgo (e.g., se infraestimarían los ingresos medios). Por este motivo, es necesario dar menor peso a las observaciones con mayor probabilidad de aparecer en la muestra, y mayor peso a las que poseen menor probabilidad de aparecer.

Aunque este estimador sea exactamente insesgado respecto al diseño muestral, su varianza bajo el diseño puede ser muy grande cuando el tamaño muestral del área n_d es pequeño. Un estimador ligeramente sesgado para n_d pequeño pero con una varianza algo menor, y que no requiere el conocimiento del tamaño del área N_d para estimar la media $\bar{Y}_d$, es el estimador de Hájek. Este estimador es igual a la media ponderada en las observaciones del área, usando como ponderaciones los pesos muestrales, es decir,

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

Para el total $Y_d = \sum_{i=1}^{N_d} Y_{di}$, el estimador de Hájek es $\hat{Y}_d^{HA} = N_d \hat{\bar{Y}}_d^{HA}$, que sí requiere conocer el tamaño poblacional N_d.

Bajo el diseño muestral, un estimador de la varianza del estimador de Hájek, $\hat{\bar{Y}}_d^{HA}$, se obtiene usando el método de linearización de Taylor. El estimador resultante se obtiene simplemente reemplazando Y_{di} por $\tilde{e}_{di} = Y_{di} - \hat{\bar{Y}}_d^{HA}$ en el estimador de la varianza del estimador de HT del total $\hat{Y}_d$ y dividir por $\widehat{N}_d$; es decir,

$$\widehat{\text{var}}_\pi(\hat{\bar{Y}}_d) = \widehat{N}_d^{-2} \left\{ \sum_{i \in s_d} \frac{(Y_{di} - \hat{\bar{Y}}_d^{HA})^2}{\pi_{di}^2} (1 - \pi_{di}) \right.$$

$$\left. + 2 \sum_{i \in s_d} \sum_{\substack{j \in s_d \\ j > i}} \frac{(Y_{di} - \hat{\bar{Y}}_d^{HA})(Y_{di} - \hat{\bar{Y}}_d^{HA})}{\pi_{di}\pi_{dj}} \left(\frac{\pi_{d,ij} - \pi_{di}\pi_{dj}}{\pi_{d,ij}} \right) \right\}, \qquad (7)$$

suponiendo que $\pi_{di} > 0$, para todo i. Para diseños en los cuales $\pi_{d,ij} \approx \pi_{di}\pi_{dj}$, para $j \neq i$, como en el muestreo de Poisson, esta varianza estimada se reduce a

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

Como se ha mencionado, los indicadores FGT tienen la ventaja de que se pueden escribir como una media para los individuos del área, véase (3). Por tanto, el estimador de Horvitz-Thompson de $F_{\alpha d}$ es entonces

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

Alternativamente, el estimador de Hájek de $F_{\alpha d}$ viene dado por

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

Obsérvese que, al sumar los estimadores directos de HT de los totales Y_d para las áreas de una región más grande, pongamos para la población completa, se obtiene el estimador de HT del total poblacional $\hat{Y} = \sum_{d=1}^{D} \sum_{i \in s_d} w_{di}\, Y_{di}$, es decir,

$$\sum_{d=1}^{D} \hat{Y}_d = \hat{Y}.$$

Dado que a un nivel de agregación mayor como el poblacional, el estimador de HT es eficiente, esta propiedad, llamada propiedad *"benchmarking"* es deseable para los estimadores en las áreas. Sin embargo, otros estimadores, especialmente los indirectos que veremos en los capítulos siguientes, no van a sumar exactamente el estimador directo considerado para el total poblacional (que puede ser distinto del de HT). Para forzar que esto ocurra, se pueden realizar ajustes en los estimadores. Sea $\hat{Y}_d^{EST}$ un estimador que no verifica esta propiedad. Si se desea que éstos sumen el estimador de HT a nivel nacional $\hat{Y}$, un ajuste común es el de tipo razón, dado por

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

Existe una gran cantidad de literatura sobre otros tipos de ajustes, como por ejemplo de diferencia, y sobre métodos específicamente diseñados para que los estimadores calculados verifiquen de forma obligatoria esta propiedad incluso a varios niveles, pero no se incluyen en este documento por brevedad. Para más información, véase e.g. Ghosh y Steorts (2013) y las referencias que ahí se citan.

A continuación, resumimos los tipos de indicadores a los que son aplicables estos estimadores, los datos que son necesarios para su producción además de los datos de la variable de interés obtenidos de una encuesta, y las ventajas e inconvenientes desde un punto de vista eminentemente práctico.

Indicadores objetivo: parámetros aditivos, en el sentido de que son sumas de ciertas variables para cada individuo del área. Estas variables pueden ser funciones de las variables de interés para los individuos (e.g. $F_{\alpha,di}$ es función de la variable considerada para medir el poder adquisitivo para el individuo, E_{di}).

Requerimientos de datos:

- Pesos muestrales w_{di} para los individuos de la muestra en el área d.
- Para el estimador de HT de la media y para el estimador de Hájek del total, tamaño poblacional del área, N_d.

Ventajas:

- El estimador de HT es exactamente insesgado y el de Hájek es aproximadamente insesgado respecto al diseño muestral. Ambos son consistentes respecto al diseño cuando el tamaño muestral del área n_d crece. Por tanto, se comportan bien para áreas con tamaño muestral suficiente bajo diseños muestrales con probabilidades desiguales, incluyendo bajo muestreo informativo, mientras se calculen usando las verdaderas probabilidades de inclusión de los individuos en la muestra del área.
- No necesitan asumir ningún modelo o hipótesis sobre las variables en cuestión Y_{di}; es decir, son completamente no paramétricos.
- Satisfacen la propiedad "*benchmarking*": si sumamos los totales estimados para todas las áreas de una región más grande, obtenemos el total estimado de esa región que se obtiene mediante el mismo método.

Desventajas:

- Son muy ineficientes (i.e. poseen un error de muestreo elevado) para áreas pequeñas debido al pequeño tamaño muestral.
- No se pueden calcular para áreas o dominios no muestreados; es decir, con tamaño muestral n_d igual a cero.

Ejemplo 4.1. Estimadores directos de HT de incidencias de pobreza, con R. Vamos a ilustrar cómo calcular estimadores directos de HT para las incidencias de pobreza, usando datos simulados sobre condiciones de vida en las provincias españolas, incluidos en el fichero de datos de R llamado incomedata del paquete de R sae. Este conjunto de datos incluye, para $n = 17119$ individuos ficticios que residen en las $D = 52$ provincias españolas, el nombre de la provincia donde reside (provlab), el código de la

provincia (prov), el código de la comunidad autónoma (ac), el grupo de edad de 1 a 5 (age), la nacionalidad (nat, 1=si posee la española, 2=si no la posee), el nivel educativo (educ, de 0=menor de 16 años a 3=nivel universitario), la situación laboral (labor, donde 0=menor de 16 años, 1=ocupado, 2=desempleado y 3=inactivo), si está en cada grupo de edad, desde el grupo 2 hasta el 5 (age2 hasta age5), si posee el nivel educativo 1 hasta 3 (educ1 hasta educ3), si posee la nacionalidad española, si está ocupado, parado o inactivo, los ingresos netos equivalentes (income) y el peso muestral (weight). Calculamos los estimadores directos de HT para las incidencias de pobreza en las $D = 52$ provincias españolas.

Después de instalar la librería sae, la cargamos, junto con el conjunto de datos incomedata, que contiene los datos muestrales, y el conjunto de datos sizeprov, que contiene los tamaños poblacionales de las provincias, N_d:

```
library(sae)
data(incomedata)
attach(incomedata)
data(sizeprov)
```

Ahora utilizamos la función direct() para obtener los estimadores directos de HT. En primer lugar, calculamos el tamaño muestral total, el número de provincias, los tamaños muestrales de éstas y extraemos los tamaños poblacionales del fichero sizeprov:

```
n<-dim(incomedata)[1]      # Tamaño muestral total
D<-length(unique(prov))        # Número de provincias (áreas o dominios)
nd<-as.vector(table(prov))      # Tamaños muestrales de las provincias
Nd<-sizeprov$Nd             # Tamaños poblacionales de las provincias
```

Establecemos el umbral de pobreza, que se calcula como 0.6*median(income) con los datos del año anterior, y construímos la variable poor, que es el indicador de tener ingresos por debajo del umbral de pobreza:

```
z<-6557.143
poor<-numeric(n)
poor[income<z]<-1
```

Finalmente, calculamos los estimadores directos de HT de las incidencias de pobreza en las provincias (promedios de la variable poor en las provincias), usando la función direct() incluyendo los pesos muestrales dados por la variable weight:

```
povinc.dir.res<-direct(y=poor,dom=prov,sweight=weight,domsize=sizeprov[,-1])
print(povinc.dir.res,row.names=F)
```

La salida de esta función es:

```
Domain SampSize     Direct         SD        CV
      1       96 0.25503732 0.04846645 19.003670
      2      173 0.14059242 0.03042195 21.638397
      3      539 0.20785096 0.02178689 10.481979
      4      198 0.26763976 0.04090335 15.282986
      5       58 0.05512200 0.02555426 46.359465
      6      494 0.21553890 0.02357906 10.939585
      7      634 0.09999792 0.01536517 15.365488
```

```
 8     1420 0.29812535 0.01618508  5.428952
 9      168 0.21413150 0.04473542 20.891562
10      282 0.27031324 0.03125819 11.563692
11      398 0.14887351 0.02189022 14.703904
12      118 0.17598199 0.03584882 20.370731
13      250 0.20921534 0.03279230 15.673948
14      224 0.29975708 0.03934080 13.124228
15      495 0.25347550 0.02467716  9.735520
16       92 0.26334059 0.05913385 22.455274
17      142 0.18337421 0.03710194 20.232911
18      208 0.31727340 0.04043964 12.745990
19       89 0.17908182 0.04234025 23.642966
20      285 0.23690549 0.03194779 13.485457
21      122 0.12583449 0.03202547 25.450474
22      115 0.24107606 0.04856351 20.144476
23      232 0.31294198 0.04122671 13.173916
24      218 0.18801572 0.03002634 15.970122
25      130 0.15559590 0.03872448 24.887854
26      510 0.25811811 0.02459196  9.527405
27      173 0.37718722 0.05696330 15.102129
28      944 0.18218209 0.01639018  8.996593
29      379 0.22918462 0.02735631 11.936364
30      885 0.17703167 0.01648910  9.314210
31      564 0.16190765 0.01842017 11.376958
32      129 0.22799612 0.04199465 18.419018
33      803 0.26064010 0.02093779  8.033220
34       72 0.30166074 0.07179782 23.800849
35      472 0.16651843 0.02307258 13.855869
36      448 0.18549072 0.02418887 13.040474
37      164 0.16104513 0.02998243 18.617410
38      381 0.18429619 0.02054550 11.148085
39      434 0.34244429 0.03248937  9.487491
40       58 0.22262002 0.05639965 25.334492
41      482 0.20503036 0.02122527 10.352256
42       20 0.02541207 0.02540651 99.978151
43      134 0.32035438 0.04934077 15.401934
44       72 0.27364239 0.06723440 24.570172
45      275 0.12553377 0.02131991 16.983409
46      714 0.21360678 0.02070508  9.693081
47      299 0.19292332 0.03211484 16.646429
48      524 0.21694466 0.02215645 10.212948
49      104 0.30027442 0.06025302 20.065986
50      564 0.10034577 0.01569138 15.637311
51      235 0.19724796 0.03341193 16.939048
52      180 0.19109119 0.03441016 18.007191
```

Finalmente, guardamos los valores estimados en un vector y contamos cuántas provincias tienen CV por encima del 20%:

```
povinc.dir<-povinc.dir.res$Direct
povinc.dir.cv<-povinc.dir.res$CV
sum(povinc.dir.cv>20)
```

Existen 15 provincias cuyos estimadores directos de HT de la incidencia de pobreza poseen un CV superior al 20%. Esas 15 provincias serían áreas pequeñas para dicho indicador. Pero, como veremos, se pueden encontrar estimadores con mejor eficiencia también en otras provincias.

B. Estimadores GREG y de calibración

Un estimador más sofisticado que los estimadores directos básicos dados en el capítulo anterior, en el sentido de que utiliza información auxiliar, es el estimador generalizado de regresión (en inglés, *generalized regression*, GREG). En su versión restringida al área d, este estimador requiere conocer el total $\boldsymbol{X}_d = \sum_{i=1}^{N_d} \boldsymbol{x}_{di}$, o la media $\bar{\boldsymbol{X}}_d = N_d^{-1} \sum_{i=1}^{N_d} \boldsymbol{x}_{di}$, para el área d de un vector $\boldsymbol{x}_{di}$ de valores de p variables auxiliares relacionadas con Y_{di}, para el individuo i dentro del área d. Si $\hat{\bar{\boldsymbol{X}}}_d = N_d^{-1} \sum_{i \in s_d} w_{di}\, \boldsymbol{x}_{di}$ es el estimador de HT de $\bar{\boldsymbol{X}}_d$, el estimador GREG de $\bar{Y}_d$ viene dado por

$$\hat{\bar{Y}}_d^{GREG} = \hat{\bar{Y}}_d + \left(\bar{\boldsymbol{X}}_d - \hat{\bar{\boldsymbol{X}}}_d\right)' \hat{\boldsymbol{B}}_d. \qquad (8)$$

Aquí, $\hat{\boldsymbol{B}}_d = \left(\sum_{i \in s_d} w_{di}\, \boldsymbol{x}_{di} \boldsymbol{x}_{di}'/c_{di}\right)^{-1} \sum_{i \in s_d} w_{di}\, \boldsymbol{x}_{di} Y_{di}/c_{di}$ es el estimador de mínimos cuadrados ponderados (usando los pesos del diseño muestral) del vector de coeficientes de la siguiente regresión lineal asumida para las unidades del área d,

$$Y_{di} = \boldsymbol{x}_{di}' \boldsymbol{\beta}_d + \epsilon_{di}, \quad i = 1, \ldots, N_d, \qquad (9)$$

donde los errores del modelo ϵ_{di} son independientes, con esperanza cero y varianza $\sigma^2 c_{di}$, siendo $c_{di} > 0$ constantes que representan la posible heteroscedasticidad, $i = 1, \ldots, N_d$. Las constantes c_{di} se determinan estudiando los residuos del modelo lineal sin heteroscedasticidad, es decir, con $c_{di} = 1$, $i = 1, \ldots, N_d$. Por ejemplo, observando el diagrama de dispersión de los residuos frente a cada una de las variables auxiliares, podemos observar gráficamente si la varianza de los residuos aumenta con alguna de ellas. En ese caso, se tomaría como constantes c_{di}, los valores de esta variable en las unidades del área o, más generalmente, una función $c_{di} = f(x_{di}) > 0$, de los valores de dicha variable auxiliar.

El estimador GREG de la media del área d, $\bar{Y}_d$, dado en (8), es aproximadamente insesgado bajo el diseño muestral independientemente de si el modelo (9) es correcto o no, ya que el sesgo del estimador del vector de coeficientes de la regresión $\hat{\boldsymbol{B}}_d$, como estimador de su versión poblacional, $\boldsymbol{B}_d = (\sum_{i=1}^{N_d} \boldsymbol{x}_{di} \boldsymbol{x}_{di}'/c_{di})^{-1} \sum_{i=1}^{N_d} \boldsymbol{x}_{di} Y_{di}/c_{di}$, es pequeño. Por este motivo, el modelo (9) se suele llamar "modelo de trabajo" (en inglés, *working model*) y a los estimadores que son insesgados independientemente de si modelo se verifica, al igual que (8), se les llama "asistidos por modelos" (en inglés, *model-assisted*). Por otro lado, el GREG también es insesgado bajo el modelo de regresión (9), condicionando a la muestra s. Aunque el estimador GREG tiende a mejorar la eficiencia del estimador directo $\hat{\bar{Y}}_d$ si las variables auxiliares están linealmente relacionadas con la variable dependiente Y_{di}, este estimador solo usa los datos del área d y por tanto su varianza puede seguir siendo grande para áreas con tamaño muestral n_d pequeño. También es posible definir un estimador GREG basado en un modelo a nivel nacional, donde los coeficientes de regresión son constantes para todas las áreas, pero la eficiencia de este estimador sigue dependiendo del tamaño muestral de área.

Obsérvese que, si deseamos utilizar el estimador GREG para el indicador FGT de orden α, que es igual a la media de las variables $F_{\alpha,di}$ en el área, es decir, $F_{\alpha d} = N_d^{-1} \sum_{i=1}^{N_d} F_{\alpha,di}$, la mejora en eficiencia respecto del estimador directo dependería de la bondad del ajuste del siguiente modelo de regresión:

$$F_{\alpha,di} = \boldsymbol{x}_{di}' \boldsymbol{\beta}_d + \epsilon_{di}, \quad i = 1, \ldots, N_d.$$

Sin embargo, en el caso de los indicadores FGT, las variables $F_{\alpha,di}$ son una función compleja de la variable de interés (la medida del poder adquisitivo E_{di}) dada por $F_{\alpha,di} = \{(z - E_{di})/z\}^{\alpha} I(E_{di} < z)$, $\alpha \geq 0$. No es fácil encontrar variables auxiliares $\boldsymbol{x}_{di}$ que estén linealmente relacionadas con $F_{\alpha,di}$. Por tanto, este modelo es difícil que se verifique en la práctica y, así, para los indicadores FGT, los estimadores GREG tienen una utilidad más reducida que para estimar medias o totales de las variables de interés (e.g., de los ingresos E_{di}).

Los estimadores de calibración son muy utilizados en los Institutos Nacionales de Estadística para estimar medias o totales a nivel nacional y en regiones con suficiente tamaño muestral. Si calibramos a nivel de área, vamos a ver que el estimador resultante está íntimamente relacionado con el estimador GREG. El método de calibración fue propuesto por Deville y Särndal (1992) para estimar el total de una variable de interés usando información auxiliar de p variables relacionadas con ella. Asumiendo que se conocen los totales de las variables auxiliares en el área, $\boldsymbol{X}_d$, y que las variables auxiliares $\boldsymbol{x}_{di}$ están relacionadas linealmente con Y_{di}, el método de calibración consiste en encontrar unos nuevos pesos h_{di}, lo más cercanos posible a los pesos muestrales originales w_{di} en virtud de una distancia $G_{di}(h_{di}, w_{di})$, tales que el total $\boldsymbol{X}_d$ de las variables auxiliares se estima con estos pesos de forma exacta; es decir, sin error. Si la variable de interés está linealmente relacionada con estas variables auxiliares y los totales de éstas se estiman de forma exacta, se espera que los totales de la variable de interés también se estimen con poco error. Formalmente, a la hora de estimar la media $\bar{Y}_d$, se buscan unos nuevos pesos para las unidades de la muestra, h_{di}, $i \in s_d$, que son la solución del problema

$$\min_{\{h_{di}; i \in s_d\}} \quad \sum_{i \in s_d} G_{di}\,(h_{di}, w_{di})$$
$$\text{sujeto a} \quad \sum_{i \in s_d} h_{di}\,\boldsymbol{x}_{di} = \boldsymbol{X}_d,$$

donde $G_{di}(\cdot,\cdot)$ es una pseudo-distancia. Usando la pseudo-distancia ji-cuadrado dada por $G_{di}(h_{di}, w_{di}) = c_{di}(h_{di} - w_{di})^2/w_{di}$, que es probablemente la más popular, y resolviendo el problema mediante el método de multiplicadores de Lagrange, los pesos resultantes son

$$h_{di} = w_{di}\left\{1 + \boldsymbol{x}_{di}{}'\left(\sum_{i \in s_d} w_{di}\,\boldsymbol{x}_{di}\boldsymbol{x}_{di}{}'/c_{di}\right)^{-1}\left(\boldsymbol{X}_d - \sum_{i \in s_d} w_{di}\,\boldsymbol{x}_{di}/c_{di}\right)\right\}, i \in s_d. \qquad (10)$$

Obsérvese que los pesos calibrados h_{di} son el resultado de realizar un ajuste a los pesos originales, $h_{di} = w_{di}g_{di}$, donde el factor de ajuste g_{di} viene dado por el término dentro de las llaves en (10). El estimador de calibración de $\bar{Y}_d$ se obtiene entonces simplemente igual que el estimador de HT, pero usando los pesos calibrados en lugar de los originales, de la forma

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

Es fácil comprobar que, sustituyendo la fórmula obtenida para estos pesos dada en (10) en el estimador de calibración $\hat{\bar{Y}}_d^{CAL}$, obtenemos exactamente el estimador GREG de $\bar{Y}_d$ dado en (8). Deville y Särndal (1992) proponen estimadores de calibración basados en distancias $G_{di}(\cdot,\cdot)$ distintas de la distancia ji-cuadrado. Sin embargo, también demuestran que los estimadores resultantes, bajo ciertas condiciones de regularidad para la distancia $G_{di}(\cdot,\cdot)$, son asintóticamente equivalentes al GREG dado en (8) y por tanto comparten la misma varianza asintótica. Al igual que el estimador GREG, para un tamaño muestral n_d pequeño, la varianza de los estimadores de calibración puede ser grande. Es posible obtener estimadores de calibración usando información auxiliar a otros niveles, como por ejemplo a nivel nacional, pero su varianza sigue dependiendo del tamaño muestral del área.

Un estimador consistente (cuando n_d crece) para la varianza del estimador $\hat{\bar{Y}}_d^{GREG}$ se obtiene usando el método de linearización de Taylor. El estimador resultante se obtiene de reemplazar Y_{di} por $\tilde{e}_{di} = Y_{di} - \boldsymbol{x}_{di}{}'\hat{\boldsymbol{B}}_d$ en la varianza estimada del estimador de HT dada en (5), es decir,

$$\widehat{\text{var}}_\pi(\hat{\bar{Y}}_d^{GREG}) = N_d^{-2}\left\{\sum_{i \in s_d} \frac{\tilde{e}_{di}^2}{\pi_{di}^2}(1 - \pi_{di}) + 2\sum_{i \in s_d}\sum_{\substack{j \in s_d \\ j > i}} \frac{\tilde{e}_{di}\tilde{e}_{di}}{\pi_{di}\pi_{di}}\left(\frac{\pi_{d,ij} - \pi_{di}\pi_{di}}{\pi_{d,ij}}\right)\right\}.$$

Para diseños en los que se verifique $\pi_{d,ij} \approx \pi_{di}\pi_{di}$, para $j \neq i$, como en el muestreo de Poisson, esta varianza estimada, escrita en función de $w_{di} = \pi_{di}^{-1}$, se reduce a

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

Estudios de simulación han mostrado que este estimador puede infraestimar la varianza del GREG. Sin embargo, el estimador que resulta de reemplazar Y_{di} por $g_{di}\tilde{e}_{di}$, siendo g_{di} el factor de ajuste de los pesos w_{di}, en la varianza estimada del estimador de HT, dado por

$$\widehat{\text{var}}_\pi(\hat{\bar{Y}}_d^{GREG}) = N_d^{-2} \left\{ \sum_{i \in s_d} \frac{g_{di}^2 \tilde{e}_{di}^2}{\pi_{di}^2}(1 - \pi_{di}) + 2 \sum_{i \in s_d} \sum_{\substack{j \in s_d \\ j > i}} \frac{g_{di}\tilde{e}_{di} g_{dj}\tilde{e}_{dj}}{\pi_{di}\pi_{di}} \left(\frac{\pi_{d,ij} - \pi_{di}\pi_{di}}{\pi_{d,ij}} \right) \right\}$$

reduce esta infraestimación y sigue siendo consistente cuando n_d crece, véase Fuller (1975) o Estevao, Hidiroglou y Särndal (1995). Además, dicho estimador alternativo de la varianza es aproximadamente insesgado para la varianza del GREG $\hat{\bar{Y}}_d^{GREG}$ bajo el modelo (9) condicionando a la muestra s, para diversos diseños muestrales.

Obsérvese de nuevo que estos estimadores tienen sentido para estimar totales o medias de las variables de interés, y no para otro tipo de parámetros. Por ejemplo, para el indicador FGT de orden α en el área d, $F_{\alpha d} = N_d^{-1} \sum_{i=1}^{N_d} F_{\alpha,di}$, los estimadores GREG o de calibración ganarían eficiencia respecto del estimador directo si las variables auxiliares $\boldsymbol{x}_{di}$ estuvieran relacionadas linealmente con $F_{\alpha,di}$, lo cual es improbable en la práctica.

A continuación se resumen los principales aspectos de estos estimadores:

Indicadores objetivo: medias/totales de las variables de interés.
Requerimientos de datos:

- Pesos muestrales w_{di} para los 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 relacionadas con la variable de interés, obtenidas de la misma encuesta donde se obtienen los datos de la variable de interés.
- Totales $\boldsymbol{X}_d$ o medias $\bar{\boldsymbol{X}}_d$ poblacionales de las p variables auxiliares en el área.

Ventajas:

- Son aproximadamente insesgados (y consistentes cuando n_d crece) respecto al diseño muestral, independientemente de si el modelo se verifica o no. Por tanto, se comportan bien para áreas con tamaño muestral suficiente bajo diseños muestrales con probabilidades desiguales, incluyendo muestreo informativo.
- No requieren que se verifique el modelo considerado para las variables de interés Y_{di}; es decir, son no paramétricos.

Desventajas:

- Aunque pueden mejorar a los estimadores directos básicos si el modelo de regresión se verifica, pueden seguir siendo ineficientes para áreas pequeñas debido al pequeño tamaño muestral.
- No se pueden calcular para áreas o dominios no muestreados; es decir, con tamaño muestral n_d igual a cero.

Ejemplo 4.2. Estimadores GREG de incidencias de pobreza, con R. Continuando el Ejemplo 4.1, ilustramos ahora cómo se podrían calcular los estimadores GREG de las incidencias de pobreza en las provincias con los mismos datos, considerando ahora variables auxiliares; concretamente, consideramos la constante 1, el grupo de edad, el nivel educativo y el estado laboral.

En primer lugar, cargamos los ficheros que contienen los datos que nos faltan: los totales de individuos en cada provincia por cada grupo de edad, por cada nivel educativo y por cada estado laboral:

```
data(sizeprovage)
data(sizeprovedu)
data(sizeprovlab)
```

Construimos la matriz con los vectores de proporciones de individuos en cada categoría y provincia. Éstos formarán el vector de medias poblacionales $\bar{\boldsymbol{X}}_d$:

```
Nd<-sizeprov[,3]
Ndage<-as.matrix(sizeprovage[,-c(1,2)])
Ndedu<-as.matrix(sizeprovedu[,-c(1,2)])
Ndlab<-as.matrix(sizeprovlab[,-c(1,2)])

Pdage<-Ndage/Nd
Pdedu<-Ndedu/Nd
Pdlab<-Ndlab/Nd

X<-cbind(const=rep(1,D),Pdage[,3:5],Pdedu[,c(2,4)],Pdlab[,2])
```

Ahora creamos la matriz de diseño para la regresión lineal, con los valores de las variables auxiliares para los individuos de la muestra:

```
Xtot<-model.matrix(poor~age3+age4+age5+educ1+educ3+labor1).
```

Finalmente, calculamos los estimadores GREG para las incidencias de pobreza (medias de la variable poor) en cada provincia:

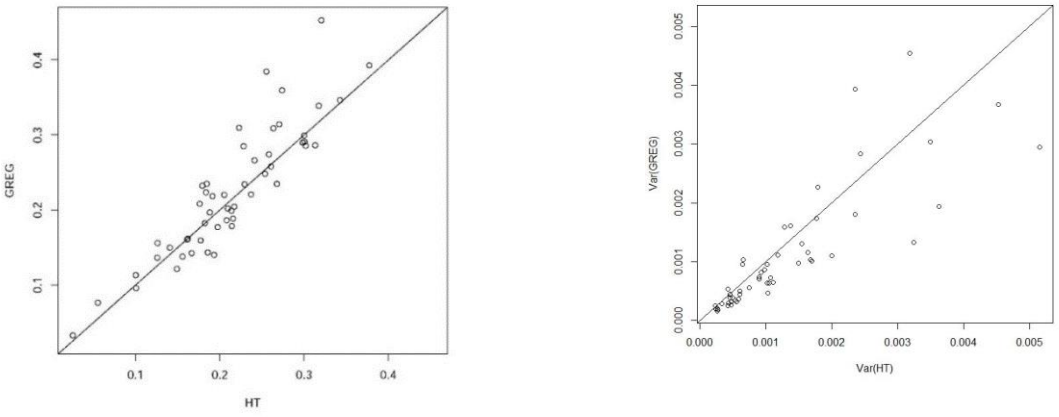
```
provl<-unique(prov)     # Índice de cada provincia
p<-dim(Xtot)[2]         # Número de variables auxiliares

betad<-matrix(0,nr=D,nc=p)   # Matriz con los coeficientes de regresión
                        # para cada provincia (en filas)
Xd.est<-matrix(0,nr=D,nc=p)  # Matriz de estimadores directos de las medias
                        # de las variables auxiliares para cada provincia
povinc.greg<-numeric(D)      # Vector de estimadores GREG en las provincias
povinc.greg.var<-numeric(D)  # Vector con las varianzas estimadas
                        # bajo el diseño de los estimadores GREG

for (d in 1:D){

  Xd<-Xtot[prov==provl[d],]  # Valores de las variables auxiliares
                        # para los indiv. de la provincia
  wd<-weight[prov==provl[d]]    # Pesos muestrales para los individuos
                        # de la provincia
  yd<-poor[prov==provl[d]]   # Valores de la variable de interés
                        # para los individuos de la provincia

  # Ajustamos la regresión para la provincia, con los pesos muestrales
  betad[d,]<-coef(summary(lm(yd~-1+Xd, weights=wd)))[,1]
  # Estimadores directos de las medias de las variables auxiliares en la provincia
```

```
  Xd.est[d,]<-colSums(diag(wd)%*%Xd)/Nd[d]

  # Estimador GREG de la incidencia de pobreza para la provincia
  povinc.greg[d]<-povinc.dir [d]+sum((X[d,]-Xd.est[d,])*betad[d,])

  # Varianza estimada bajo el diseño del estimador
  # GREG de la incidencia de pobreza
  gd<-matrix(1/Nd[d]+
  +(X[d,]-Xd.est[d,])%*%solve(t(Xd)%*%diag(wd)%*%Xd)%*%t(Xd),nr=nd[d])
  ed<-yd-Xd%*%as.matrix(betad[d,],nr=p)
  povinc.greg.var[d]<-sum(wd*(wd-1)*(gd*ed)^2)
}

# CVs de los estimadores GREG
povinc.greg.cv<-100*sqrt(povinc.greg.var)/povinc.greg
```

Representamos los valores de los estimadores GREG frente a los de HT, así como sus varianzas (o errores de muestreo al cuadrado):

```
M<-max(povinc.dir,povinc.greg)
m<-min(povinc.dir,povinc.greg)
plot(povinc.dir,povinc.greg,ylim=c(m,M),xlim=c(m,M),xlab="HT",ylab="GREG")
abline(a=0,b=1)

M<-max(povinc.dir.var,povinc.greg.var)
m<-min(povinc.dir.var,povinc.greg.var)
plot(povinc.dir.var,povinc.greg.var,ylim=c(m,M),xlim=c(m,M),xlab="Var(HT)",
 ylab="Var(GREG)")
abline(a=0,b=1)
```

Gráfico 2
Estimadores GREG de las incidencias de pobreza para las provincias frente a estimadores HT (izquierda), y varianzas estimados de los estimadores GREG frente a las de los estimadores de HT (derecha).
(En proporciones)



Fuente: Elaboración propia.

Podemos ver que los estimadores GREG se parecen a los de HT, pero sus varianzas estimadas son ligeramente menores. Esta ganancia en eficiencia se consigue gracias al uso de información auxiliar.

IV. Métodos indirectos básicos para la desagregación de datos de pobreza

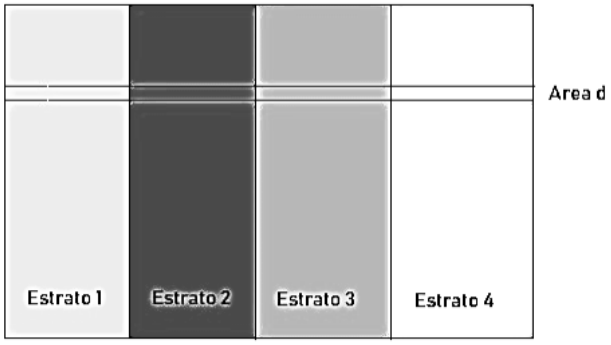
Un estimador indirecto para un indicador en un área concreta es aquel que usa la información de otras áreas gracias al hecho de asumir algún tipo de homogeneidad entre ellas. El empleo de una mayor cantidad de información en el proceso de estimación conlleva a menudo una disminución en el error de muestreo (o aumento de la eficiencia). Vamos a describir en primer lugar estimadores sintéticos. Un estimador sintético es aquel que considera que las áreas son homogéneas en el sentido de poseer parámetros comunes, sin permitir ningún grado de heterogeneidad entre ellas que no sea debido a las variables auxiliares disponibles. Estos estimadores asumen hipótesis fuertes que son poco probables en la práctica y por tanto pueden tener un sesgo grande. A pesar de su potencial sesgo, se incluyen en este documento con el objeto de ilustrar la idea intuitiva subyacente a la estimación en áreas pequeñas, que consiste en tomar prestada información de otras áreas con el fin de mejorar la eficiencia.

A. Estimador post-estratificado sintético

Se recalca de nuevo que este estimador es escasamente utilizado en aplicaciones reales de estimación en áreas pequeñas debido a que se basa en hipótesis poco realistas; sin embargo, se describe en este documento ya que ilustra de una manera sencilla la idea subyacente principal sobre cómo tomar prestada información.

Se dispone de una variable cualitativa relacionada con la variable Y_{di}. Esta variable cualitativa tiene J categorías posibles, las cuales dividen a la población U en J grupos, $U^1, \dots, U^J$ de tamaños $N^1, \dots, N^J$, llamados post-estratos, que se entrecruzan con las áreas. Por tanto, el área U_d de la población queda igualmente dividida en J trozos de post-estratos, $U_d^1, \dots, U_d^J$ de tamaños poblacionales $N_d^1, \dots, N_d^J$ y con medias $\bar{Y}_d^1, \dots, \bar{Y}_d^J$, donde $\bar{Y}_d^j = \sum_{i \in U_d^j} Y_{di} / N_d^j$, $j = 1, \dots, J$, véase el Gráfico 3. En este gráfico y en lo siguiente, por simplicidad, nos referimos a los post-estratos como estratos.

Gráfico 3
Población dividida en 4 post-estratos y área *d*



Fuente: Elaboración propia.

Dado que las medias son indicadores aditivos, podemos descomponerlas en sumas para los J estratos, de la forma

$$\bar{Y}_d = \frac{1}{N_d}\sum_{i=1}^{N_d} Y_{di} = \frac{1}{N_d}\sum_{j=1}^{J} N_d^j \ \bar{Y}_d^j. \tag{11}$$

Se asume que los individuos dentro de cada estrato se comportan de forma homogénea, independientemente del área al que pertenecen; más concretamente, se asume

$$\bar{Y}_d^j = \bar{Y}^j, \quad j = 1, \dots, J, \tag{12}$$

donde $\bar{Y}^j = \sum_{i \in U^j} Y_{di} \, / N^j$ es la media del estrato j. Entonces, podemos aprovechar esta homogeneidad dentro de los estratos para estimar la media de cada área estimando las medias de los estratos (que deben tener tamaños muestrales grandes). Es decir, sustituyendo (12) en (11), resulta

$$\bar{Y}_d = \frac{1}{N_d}\sum_{j=1}^{J} N_d^j \ \bar{Y}^j. \tag{13}$$

El estimador post-estratificado sintético (PS-SYN) de $\bar{Y}_d$ se obtiene estimando las medias de cada estrato en (13) mediante los estimadores de Hájek, es decir,

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

Se considera que el número de estratos J es pequeño, y que éstos tienen suficiente muestra. Por tanto, los estimadores directos $\hat{\bar{Y}}^{j,HA}$ de las medias en los estratos $\bar{Y}^j$ tienen varianza pequeña. Esto significa que, al estimar la media para el área d a través de los estimadores para los estratos, $\hat{\bar{Y}}^{j,HA}$, la varianza también es pequeña. Así, se aprovecha la homogeneidad dentro de cada estrato para mejorar la eficiencia del estimador para el área d usando todos los datos de la muestra. Sin embargo, la hipótesis de homogeneidad dentro de cada uno de los estratos (12) es poco realista y por tanto el estimador post-estratificado sintético puede tener un sesgo considerable.

Dado que el sesgo de estos estimadores no es despreciable, en lugar de su varianza, la cual puede dar una imagen equivocada de la calidad del estimador, es de interés obtener su error cuadrático medio (ECM), que refleja ambas cosas. Para estimadores sintéticos generales, un estimador del ECM bajo el diseño viene dado por

$$\widehat{\mathrm{MSE}}_{\pi}(\hat{\bar{Y}}_d^{SYN}) = (\hat{\bar{Y}}_d^{SYN} - \hat{\bar{Y}}_d^{DIR})^2 - \widehat{\mathrm{var}}_{\pi}(\hat{\bar{Y}}_d^{DIR}),$$

véase Rao y Molina (2015), p.44, eq. (3.2.16). Este estimador es muy inestable pues depende del estimador directo del área correspondiente. Se han propuesto estimadores del ECM más estables basándose en la idea de promediar para todas las áreas, pero los estimadores resultantes no son específicos para cada área; es decir, se daría el mismo valor de ECM para todas las áreas. No se conocen estimadores del ECM para los estimadores sintéticos que sean al mismo tiempo estables y específicos para cada área. Este es un inconveniente de estos estimadores.

Si deseamos utilizar el estimador PS-SYN para un indicador FGT, gracias a la aditividad de estos indicadores, en principio sería posible. Sin embargo, el estimador estaría basado en la hipótesis (poco realista) de que el indicador FGT se mantiene constante dentro de los estratos, es decir

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

siendo F_{α}^{j} el indicador FGT en el estrato j. Por tanto, este estimador sería más útil para estimar medias o totales de una variable continua.

Resumen sobre estos estimadores:

Indicadores objetivo: medias/totales de la variable de interés
Requerimientos de datos:

- Pesos muestrales w_{di} para todos los individuos de la muestra.
- Tamaño poblacional del área, N_d, y tamaños poblacionales de las intersecciones estrato-área, $N_d^j, j = 1, \dots, J$.
- Una variable cualitativa (o una combinación de varias) observada en la misma encuesta que la variable de interés y relacionada con ésta.

Ventajas:

- Si los estratos tienen suficientes observaciones en la muestra, se puede disminuir considerablemente la varianza en comparación con un estimador directo.

Desventajas:

- La hipótesis de homogeneidad considerada para las variables Y_{di} es poco realista en la práctica. Si esta no se verifica, los estimadores resultantes pueden tener un sesgo considerable, y por tanto no reflejar la realidad. Además, al estimar los errores de muestreo se obtendrán valores pequeños. Sin embargo, raramente es posible estimar adecuadamente el sesgo. Por tanto, a falta de estimaciones correctas del sesgo, los estimadores pueden aparentar ser de buena calidad, siendo esto difícilmente cierto.
- No es fácil encontrar estimadores estables del ECM bajo el diseño.

Ejemplo 3. Estimadores post-estratificados sintéticos de incidencias de pobreza con R. Continuando el Ejemplo 2, ilustramos ahora cómo calcular los estimadores post-estratificados sintéticos de las incidencias de pobreza para las provincias usando los niveles educativos (variable `educ`) como post-estratos.

En el Ejemplo 2 habíamos cargado los tamaños poblacionales de las provincias por cada nivel educativo (conjunto de datos `sizeprovedu`). Dichos tamaños deben estar en un objeto de tipo data frame, donde los nombres de las columnas deben coincidir con los códigos utilizados para las categorías de la variable post-estrato (`educ`). Por tanto, añadimos los nombres de las columnas al data frame con los tamaños poblacionales. Después, llamamos a la función `pssynt()`, que calcula los estimadores

post-estratificados para las incidencias de pobreza (medias de la variable poor) usando la variable educ y guardamos los valores estimados:

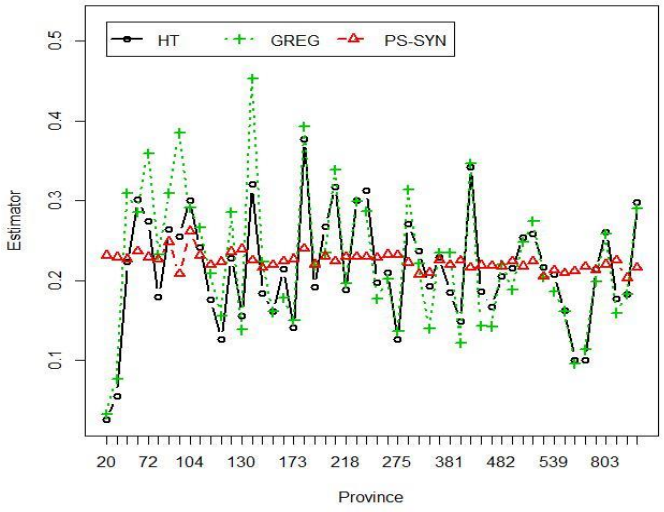
```
colnames(sizeprovedu) <- c("provlab","prov","0","1","2","3")
povinc.psedu.res<-pssynt(y=poor,sweight=weight,ps=educ,domsizebyps=sizeprovedu[,-1])
povinc.psedu<-povinc.psedu.res$PsSynthetic
```

Finalmente, comparamos gráficamente los resultados con los obtenidos mediante los estimadores directos HT y GREG para cada provincia:

```
o<-order(nd)
M<-max(povinc.psedu,povinc.dir,povinc.greg)
m<-min(povinc.psedu,povinc.dir,povinc.greg)
k<-6
plot(1:D,povinc.dir[o],type="n",ylim=c(m,M+(M-m)/k),
 xlab="Province",ylab="Estimator",xaxt="n")
points(1:D,povinc.dir[o],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:D,povinc.greg[o],type="b",col=3,lty=3,pch=3,lwd=2)
points(1:D,povinc.psedu[o],type="b",col=2,lty=2,pch=2,lwd=2)
 axis(1, at=1:D, labels=nd[o])
legend(1,M+(M-m)/k,legend=c("HT","GREG","PS-SYN"),ncol=3,
col=c(1,3,2),lwd=rep(2,3), lty=c(1,3,2),pch=c(1,3,2))
```

Los resultados se muestran en el Gráfico 4. Podemos ver que los estimadores post-estratificados sintéticos son demasiado similares para todas las provincias, ya que asumen homogeneidad para todos los individuos del mismo nivel educativo, independientemente de la provincia a la que pertenecen. Esta hipótesis es difícilmente cierta.

Gráfico 4
Estimaciones HT, GREG y PS-SYN de las incidencias de pobreza para cada provincia
(En proporciones)



Fuente: Elaboración propia.

B. Estimador sintético de regresión a nivel de área

Los estimadores sintéticos de regresión asumen un modelo de regresión lineal que se puede plantear, bien a nivel de área o bien a nivel de individuo, dependiendo de la información auxiliar disponible. Comenzamos considerando que únicamente se posee información auxiliar a nivel de área. Llamamos $\boldsymbol{x}_d$ al vector disponible de p variables auxiliares a nivel de área (e.g., el vector de medias $\bar{\boldsymbol{X}}_d$ de p variables auxiliares). Se asume que el indicador que se desea estimar δ_d (e.g., la media del área, $\delta_d = \bar{Y}_d$) varía respecto de estos datos agregados $\boldsymbol{x}_d$ de forma constante para todas las áreas, según un modelo de regresión lineal. Dado que los verdaderos valores del indicador en las áreas no están disponibles (son los parámetros objetivo), en lugar de éstos, se consideran estimadores directos $\hat{\delta}_d$, $d = 1, \dots, D$. Así, el modelo a nivel de área asume

$$\hat{\delta}_d = \boldsymbol{x}_d{}'\boldsymbol{\alpha} + \varepsilon_d, \quad d = 1, \dots, D, \tag{14}$$

donde los términos de error ε_d se asumen independientes, con esperanza cero y varianza ψ_d conocida, $d = 1, \dots, D$. Obsérvese que, dado que $\boldsymbol{x}_d$ es el valor poblacional y por tanto tiene varianza cero, ψ_d es la varianza del estimador directo $\hat{\delta}_d$, es decir, $\psi_d = \mathrm{var}(\hat{\delta}_d)$. En la práctica, estas varianzas se estiman con los microdatos de la encuesta. El estimador sintético de regresión (REG1-SYN) para el indicador del área d viene entonces dado por la predicción del indicador a través del modelo, es decir, si $\hat{\boldsymbol{\alpha}} = (\sum_{d=1}^{D} \psi_d^{-1} \boldsymbol{x}_d \boldsymbol{x}_d{}')^{-1} \sum_{d=1}^{D} \psi_d^{-1} \boldsymbol{x}_d \hat{\delta}_d$ es el estimador de $\boldsymbol{\alpha}$ obtenido por mínimos cuadrados ponderados, el estimador REG1-SYN de δ_d viene dado por

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

En el modelo (14), ε_d es el error debido a que utilizamos un estimador directo $\hat{\delta}_d$ en lugar del verdadero valor del indicador δ_d, ya que éste es desconocido, y se está asumiendo que el verdadero valor δ_d es exactamente igual al término de la regresión, $\delta_d = \boldsymbol{x}_d{}'\boldsymbol{\alpha}$, sin dejar ningún grado de heterogeneidad a los indicadores de las distintas áreas en cuanto a esta regresión. Este tipo de modelos que, al igual que (14), no incorporan efectos de las áreas representando dicha heterogeneidad, se denominan "modelos sintéticos". De hecho, el sesgo bajo el diseño de $\hat{\delta}_d^{REG1-SYN}$ para $\boldsymbol{\alpha}$ conocido, viene dado por $\boldsymbol{x}_d{}'\boldsymbol{\alpha} - \delta_d$, que no depende del tamaño muestral del área n_d; por tanto, este sesgo no disminuye al aumentar el tamaño muestral del área.

Una ventaja de los estimadores basados en modelos es que permiten estimar en áreas no muestreadas; es decir, con tamaño muestral igual a cero, si se dispone de la correspondiente información auxiliar. Para un área d con $n_d = 0$, si disponemos de $\boldsymbol{x}_d$, entonces el estimador sintético de δ_d es igualmente $\hat{\delta}_d^{REG1-SYN} = \boldsymbol{x}_d{}'\hat{\boldsymbol{\alpha}}$.

Para estimar el indicador de pobreza FGT de orden α, $\delta_d = F_{\alpha d}$, mediante este procedimiento, necesitamos variables auxiliares a nivel de área que verifiquen el modelo a nivel de área

$$\hat{F}_{\alpha d} = \boldsymbol{x}_d{}'\boldsymbol{\alpha} + \varepsilon_d, \quad d = 1, \dots, D, \tag{15}$$

siendo $\psi_d = \mathrm{var}(\hat{F}_{\alpha d})$, $d = 1, \dots, D$ conocidas. Así, el estimador sintético de regresión para el indicador FGT en el área d, $F_{\alpha d}$, viene dado por

$$\hat{F}_{\alpha d}^{REG1-SYN} = \boldsymbol{x}_d{}'\hat{\boldsymbol{\alpha}},$$

donde, en este caso, $\hat{\boldsymbol{\alpha}} = (\sum_{d=1}^{D} \psi_d^{-1} \boldsymbol{x}_d \boldsymbol{x}_d{}')^{-1} \sum_{d=1}^{D} \psi_d^{-1} \boldsymbol{x}_d \hat{F}_{\alpha d}$.

El modelo (14) asumido por el estimador REG1-SYN enlaza todas las áreas a través del parámetro de regresión común $\boldsymbol{\alpha}$. Al estimar este parámetro común con los estimadores directos $\hat{\delta}_d$ de todas las áreas, se obtiene un estimador con varianza mucho más pequeña que para el estimador directo. Sin embargo, este modelo no incorpora heterogeneidad entre las áreas, aparte de la heterogeneidad explicada (o debida) a las variables auxiliares consideradas. En la práctica, difícilmente se cuenta con datos de todas las variables auxiliares que explican completamente la heterogeneidad de los indicadores δ_d en las áreas en las que se

desea estimar. Por tanto, el modelo sintético (14) puede no representar muchos de los casos que aparecen en la práctica, proporcionando estimadores sesgados en estos casos. Además, obsérvese que, en el caso más favorable de conocer el verdadero modelo (y el verdadero valor de $\boldsymbol{\alpha}$), el estimador REG1-SYN sería $\boldsymbol{x}_d{}'\boldsymbol{\alpha}$, con lo cual no se estarían utilizando los datos de la variable de interés para esa área obtenidos de la encuesta. Así, esto podría considerarse un desperdicio para las áreas de tamaño muestral grande. Además, el estimador obtenido puede diferir mucho del estimador directo, que sería fiable para estas áreas. Este es un gran inconveniente de los estimadores (o los modelos) sintéticos. Por otro lado, como se ha comentado en la introducción, al ser estimadores potencialmente sesgados bajo el diseño, su calidad debería ser evaluada en términos de ECM en lugar de varianza (que será pequeña induciendo erróneamente a pensar que el estimador tiene calidad); sin embargo, no se conocen estimadores del ECM bajo el diseño que sean estables y al mismo tiempo distintos para cada área.

Resumen de estos estimadores:

Indicadores objetivo: parámetros generales.
Requerimientos de datos:

- Datos agregados (e.g. medias poblacionales) de las p variables auxiliares consideradas en las áreas, $\boldsymbol{x}_d$, $d = 1, \dots, D$.

Ventajas:

- Puede disminuir considerablemente la varianza en comparación con un estimador directo.
- Se puede estimar en áreas no muestreadas.

Desventajas:

- El modelo de regresión sintético considerado no representa los casos en los que no se dispone de todas las variables auxiliares que expliquen la heterogeneidad entre las áreas. Por tanto, en estos casos, los estimadores resultantes pueden tener un sesgo sustancial.
- Es necesario analizar concienzudamente el modelo (e.g. a través de los residuos), pues el sesgo de estos estimadores depende de la bondad de ajuste de dicho modelo. Especialmente, es de gran importancia contrastar si existe efecto del área, ya que este modelo no lo considera.
- Si se conoce el modelo, no se usarían los datos de la variable de interés para esa área.
- No tiende al estimador directo cuando aumenta el tamaño muestral.
- No se conocen estimadores del ECM bajo el diseño que sean estables y al mismo tiempo distintos para cada área.
- Requieren un reajuste para verificar la propiedad "*benchmarking*" de que la suma de los totales estimados en las áreas de una región mayor coincida con el estimador directo para dicha región.

C. Estimador sintético de regresión a nivel de individuo

Consideramos ahora que se dispone de datos a nivel de individuo (o microdatos) de las p variables auxiliares en la encuesta, $\boldsymbol{x}_{di}$, $i \in s_d$, $d = 1, \dots, D$. En este caso, se puede obtener un estimador sintético de regresión para el indicador de interés asumiendo un modelo de regresión lineal a nivel de individuo para Y_{di}. Llamamos $\boldsymbol{y}_d = (Y_{d1}, \dots, Y_{dN_d})'$ al vector de valores de la variable en cuestión para los individuos del área d. El indicador a estimar en el área d es una función de este vector, es decir, $\delta_d = \delta_d(\boldsymbol{y}_d)$. El modelo de regresión sintético básico considera que las variables Y_{di} para todos los individuos de la población siguen el modelo de regresión lineal

$$Y_{di} = \boldsymbol{x}_{di}{}'\boldsymbol{\beta} + \varepsilon_{di}, \quad i = 1, \dots, N_d, d = 1, \dots, D, \tag{16}$$

donde los errores ε_{di} son independientes, con esperanza cero y varianza $\sigma^2 k_{di}^2$, donde k_{di} son constantes conocidas que representan la posible heteroscedasticidad en el modelo ($k_{di} = 1$ para todo i y d si no existe heteroscedasticidad). Estimando $\boldsymbol{\beta}$ mediante el estimador por mínimos cuadrados ponderados

$\widehat{\boldsymbol{\beta}} = (\sum_{d=1}^{D}\sum_{i\in s_d} a_{di}\,\boldsymbol{x}_{di}\boldsymbol{x}_{di}{}')^{-1}\sum_{d=1}^{D}\sum_{i\in s_d} a_{di}\,\boldsymbol{x}_{di}Y_{di}$, siendo $a_{di} = k_{di}^{-2}$, obtenemos predicciones, a través del modelo, para cada individuo del área, $\hat{Y}_{di} = \boldsymbol{x}_{di}{}'\widehat{\boldsymbol{\beta}}$, $i = 1, \dots, N_d$. El vector de predicciones para el área d es entonces $\widehat{\boldsymbol{y}}_d = (\hat{Y}_{d1}, \dots, \hat{Y}_{dN_d})'$. Usando este vector en lugar de $\boldsymbol{y}_d$ para calcular el indicador, obtenemos el estimador sintético de regresión de δ_d, es decir

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

Por ejemplo, para la media del área d, $\delta_d = \bar{Y}_d$, si $\bar{\boldsymbol{X}}_d$ es el vector de medias poblacionales de las p variables auxiliares consideradas en esa área, el estimador sintético basado en el modelo (16) sería

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

Para un área no muestreada, este estimador se obtiene de la misma forma. Para $\boldsymbol{\beta}$ conocido, el sesgo bajo el diseño del estimador de la media es $\bar{\boldsymbol{X}}_d{}'\boldsymbol{\beta} - \bar{Y}_d$, que no depende del tamaño muestral del área n_d; por tanto, este sesgo no disminuye al aumentar el tamaño muestral.

De nuevo, si quisiéramos estimar los indicadores FGT de pobreza, deberíamos encontrar variables $\boldsymbol{x}_{di}$ linealmente relacionadas con $F_{\alpha,di}$; es decir, que verificasen el modelo

$$F_{\alpha,di} = \boldsymbol{x}_{di}{}'\boldsymbol{\beta} + \varepsilon_{di}, \quad i = 1, \dots, N_d, d = 1, \dots, D. \tag{17}$$

Sin embargo, encontrar variables linealmente relacionadas con $F_{\alpha,di}$ es poco habitual en la práctica. Sería más conveniente asumir el modelo para las variables de interés, es decir, las variables utilizadas para medir el poder adquisitivo, E_{di} o, mejor, para una transformación biyectiva de éstas, $T(E_{di})$, ya que E_{di} suelen tener una distribución muy asimétrica y por tanto un modelo lineal para estas variables no sería muy adecuado. En la práctica, es muy común utilizar la transformación logaritmo; es decir, se tomaría $Y_{di} = \log(E_{di} + c)$ como variable respuesta en el modelo, siendo $c > 0$ una constante positiva que hace la distribución de Y_{di} aproximadamente normal. Esta constante se puede determinar ajustando el modelo para una secuencia de valores de c en el rango de E_{di}, y tomar el valor de c para el cual una medida de asimetría de los residuos del modelo (e.g., el coeficiente de asimetría de Pearson) está lo más cercano posible a cero.

Al igual que en el caso de los anteriores estimadores sintéticos, si no se dispone de todas las variables auxiliares que explican la heterogeneidad de Y_{di} en las áreas; es decir, el modelo sintético que se asume no se verifica, entonces estos estimadores serán sesgados. Sin embargo, su varianza será pequeña ya que el coeficiente de la regresión se estima usando la muestra completa, que suele ser grande. Por tanto, el estimador sintético de regresión tendrá un error de muestreo pequeño. Estos estimadores requieren un estudio de la bondad de ajuste del modelo asumido para evitar grandes sesgos. De nuevo, en el mejor caso, si conociéramos exactamente el modelo, estos estimadores usarían solamente los datos de las variables auxiliares y no los datos de la variable de interés observados en el área en cuestión, y no se acercan a los estimadores directos para las áreas con suficiente tamaño muestral. Además, no se conocen estimadores fiables del ECM bajo el diseño que sean distintos para cada área.

Resumen de estos estimadores:

Indicadores objetivo: parámetros generales.

Requerimientos de datos:

- Observaciones muestrales de las p variables auxiliares relacionadas con la variable de interés, obtenidas de la misma encuesta donde se obtienen los datos de la variable de interés.
- Para indicadores tipo medias/totales de la variable respuesta considerada en el modelo, medias/totales poblacionales de las p variables auxiliares consideradas en las áreas, $\bar{\boldsymbol{X}}_d$, $d = 1, \dots, D$. Para indicadores no lineales en las variables respuesta del modelo, se necesitan los valores de las p variables auxiliares para todos los individuos (microdatos) de esa área, $\{\boldsymbol{x}_{di};\ i = 1, \dots, N_d, d = 1, \dots, D\}$.

Ventajas:

- Puede reducir considerablemente la varianza de los estimadores directos y de los estimadores obtenidos a partir de un modelo a nivel de área.
- Se puede estimar en áreas no muestreadas.

Desventajas:

- El modelo de regresión sintético considerado no representa los casos en los que no se dispone de todas las variables auxiliares que expliquen la heterogeneidad entre las áreas. Por tanto, en estos casos, los estimadores resultantes pueden tener un sesgo sustancial.
- Es necesario analizar concienzudamente el modelo (e.g. a través de los residuos), pues el sesgo de estos estimadores depende de la bondad de ajuste de dicho modelo. Especialmente, es de gran importancia contrastar si existe efecto del área, ya que este modelo no lo considera.
- Si se conociera exactamente el modelo, no usarían los datos de la variable de interés para esa área.
- No tiende al estimador directo cuando aumenta el tamaño muestral.
- No se conocen estimadores del ECM bajo el diseño que sean estables y al mismo tiempo distintos para cada área.
- Requieren un reajuste para verificar la propiedad "*benchmarking*" de que la suma de los totales estimados en las áreas de una región mayor coincida con el estimador directo para dicha región.

D. Estimadores compuestos

Como se ha comentado en capítulos anteriores, los estimadores directos son (al menos aproximadamente) insesgados bajo el diseño muestral, pero pueden poseer una varianza grande para las áreas de tamaños muestrales pequeños. Por otro lado, los estimadores sintéticos tienen varianza pequeña, pero pueden ser considerablemente sesgados bajo el diseño. Los estimadores compuestos nacen con el objeto de disminuir la varianza del estimador directo a cambio de una porción del sesgo de un estimador sintético. Se pretende al mismo tiempo aumentar la eficiencia del estimador directo y reducir el sesgo del estimador sintético. Sea $\hat{\bar{Y}}_d^{DIR}$ un estimador directo genérico de $\bar{Y}_d$ y $\hat{\bar{Y}}_d^{SYN}$ un estimador sintético. Un estimador compuesto de $\bar{Y}_d$ tiene la forma

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

El peso ϕ_d otorgado al estimador directo se puede establecer, bien de forma semi-óptima minimizando una aproximación del error cuadrático medio (ECM) bajo el diseño muestral o fijándolo de forma arbitraria. Drew, Singh y Choudhry (1982) propusieron un peso ϕ_d que depende del tamaño muestral del área, dando lugar al estimador dependiente del tamaño muestral (en inglés, *sample-size dependent*, SSD). Tomando un valor $\delta > 0$ preestablecido (por defecto se puede tomar 1), el peso propuesto tiene la forma

$$\phi_d = \begin{cases} 1, & \text{si} \quad \widehat{N}_d \ge \delta N_d; \\ \widehat{N}_d/(\delta N_d), & \text{si} \quad \widehat{N}_d < \delta N_d, \end{cases}$$

donde $\widehat{N}_d = \sum_{i \in s_d} w_{di}$. Para comprender la idea intuitiva de este estimador, obsérvese que, bajo muestreo aleatorio simple (MAS) en la población (en ese caso el tamaño del área n_d es aleatorio), se obtiene

$$\widehat{N}_d = \sum_{i \in s_d} w_{di} = \sum_{i \in s_d} \frac{N}{n} = N\frac{n_d}{n}$$

y como $\widehat{N}_d$ es insesgado, su esperanza bajo el diseño es igual a $NE_\pi(n_d)/n = N_d$, con lo cual $E_\pi(n_d) = nN_d/N$ y por tanto el peso resulta ser

$$\phi_d = \begin{cases} 1 & \text{si} \quad n_d \geq \delta E_\pi(n_d); \\ n_d/\{\delta E_\pi(n_d)\} & \text{si} \quad n_d < \delta E_\pi(n_d). \end{cases}$$

Si fijamos $\delta = 1$, entonces el estimador SSD le da peso 1 al estimador directo cuando el tamaño muestral del área es mayor o igual al tamaño muestral esperado, y le da un peso menor que 1 en caso contrario. Sin embargo, un área determinada puede tener tamaño muestral n_d pequeño, pero éste puede superar el tamaño esperado, con lo cual se le daría peso 1 al estimador directo y por lo tanto no habría mejora de eficiencia respecto al estimador directo.

El estimador SSD se utilizó en la Encuesta de Población Activa de Canadá para obtener estimadores para las secciones censales tomando $\delta = 2/3$, véase Drew, Singh y Choudhry (1982). Sin embargo, para la mayoría de las áreas consideradas, el peso del estimador directo resultó ser $\phi_d = 1$; para unas pocas, el peso fue $\phi_d = 0.9$, pero en ningún caso se obtuvo un peso menor que 0.8. Por tanto, la ganancia en eficiencia respecto del estimador directo fue muy limitada. Al igual que en esta aplicación, el problema de este estimador es que suele dar un peso cercano a 1 a los estimadores directos aunque el tamaño muestral del área sea pequeño, no habiendo mejora de eficiencia respecto del estimador directo. Además, el peso ϕ_d no tiene en cuenta si las áreas son muy o poco homogéneas en el sentido de satisfacer el modelo considerado por el estimador sintético. Por tanto, es independiente de la calidad del estimador sintético (o de la bondad de ajuste del modelo sintético) para cada área. Por tanto, estos estimadores se pueden considerar demasiado simples como para conllevar una mejora perceptible de eficiencia respecto a los estimadores directos.

Como se ha mencionado, es posible obtener estimadores compuestos aproximadamente óptimos respecto al diseño muestral, tomando el peso ϕ_d que minimiza (aproximadamente) el ECM bajo el diseño del estimador compuesto, $\mathrm{MSE}_\pi(\hat{\bar{Y}}_d^C)$. Considerando que la covarianza entre el estimador directo y el sintético es despreciable, y minimizando

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

se obtiene el peso óptimo

$$\phi_d^* = \mathrm{MSE}_\pi(\hat{\bar{Y}}_d^{SYN})/\{\mathrm{var}_\pi(\hat{\bar{Y}}_d^{DIR}) + \mathrm{MSE}_\pi(\hat{\bar{Y}}_d^{SYN})\}. \tag{18}$$

Un estimador de $\mathrm{MSE}_\pi(\hat{\bar{Y}}_d^{SYN})$ es

$$\widehat{\mathrm{MSE}}_\pi(\hat{\bar{Y}}_d^{SYN}) = (\hat{\bar{Y}}_d^{SYN} - \hat{\bar{Y}}_d^{DIR})^2 - \widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d^{DIR}),$$

véase Rao y Molina (2015, p.44). Reemplazando este estimador en el peso óptimo ϕ_d^* dado en (18), obtenemos un estimador de este peso óptimo, dado por

$$\hat{\phi}_d^* = \widehat{\mathrm{MSE}}_\pi(\hat{\bar{Y}}_d^{SYN})/(\hat{\bar{Y}}_d^{SYN} - \hat{\bar{Y}}_d^{DIR})^2 = 1 - \widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d^{DIR})/(\hat{\bar{Y}}_d^{SYN} - \hat{\bar{Y}}_d^{DIR})^2.$$

Podemos ver que este peso depende del estimador directo $\hat{\bar{Y}}_d^{DIR}$, que es muy volátil. Esto significa que el peso óptimo estimado $\hat{\phi}_d^*$ también es muy volátil. Se puede obtener un peso estimado más estable promediando para todas las áreas, de la forma

$$\begin{aligned} \hat{\phi}^* &= \sum_{\ell=1}^{D} \widehat{\mathrm{MSE}}_\pi(\hat{\bar{Y}}_\ell^{SYN}) / \sum_{\ell=1}^{D} (\hat{\bar{Y}}_\ell^{SYN} - \hat{\bar{Y}}_\ell^{DIR})^2 \\ &= 1 - \left\{ \sum_{\ell=1}^{D} \widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_\ell^{DIR}) / \sum_{\ell=1}^{D} (\hat{\bar{Y}}_\ell^{SYN} - \hat{\bar{Y}}_\ell^{DIR})^2 \right\} \end{aligned}$$

El peso resultante, $\hat{\phi}^*$, es muy estable, pero no depende del área d; es decir, es constante para todas las áreas, sin depender siquiera de su tamaño muestral. Probablemente por estos inconvenientes, los estimadores compuestos óptimos son menos utilizados en la práctica que los "basados en modelos" que veremos en el siguiente capítulo.

Los estimadores compuestos son interesantes debido al compromiso que se pretende conseguir entre sesgo y varianza. Sin embargo, veremos en los capítulos siguientes que se pueden obtener estimadores compuestos con mayor eficiencia que los estimadores directos a partir de modelos de regresión que tienen en cuenta heterogeneidad entre las áreas. Estos estimadores compuestos serán óptimos respecto a la distribución de probabilidad inducida por el modelo asumido, y por este motivo se denominan estimadores "basados en modelos". En estos estimadores, los pesos dependen del tamaño muestral del área y de la bondad de ajuste del modelo sintético, dando mayor peso a los estimadores directos cuando el modelo sintético es pobre (variables auxiliares poco informativas o áreas muy heterogéneas) o cuando el tamaño muestral del área es grande, y dando mayor peso al estimador sintético a medida que disminuye el tamaño muestral o que el modelo tenga más potencia predictiva. Por tanto, los estimadores basados en modelos dominan a estos estimadores compuestos simples.

A continuación resumimos las características del estimador SSD, como el representante más común de los estimadores compuestos:

Indicadores objetivo: parámetros aditivos.
Requerimientos de datos:

- Pesos muestrales w_{di} para los individuos de la muestra en el área d.
- Tamaño poblacional del área, N_d, si se usa el estimador de HT de la media o el estimador de Hájek del total.

Ventajas:

- Están diseñados para reducir al mismo tiempo el sesgo del estimador sintético y la varianza del estimador directo. No pueden tener peor eficiencia que el estimador directo ni mayor sesgo que el estimador sintético.

Desventajas:

- Para un área de tamaño muestral pequeño, mientras este tamaño no sea inferior al tamaño muestral esperado, no se toma prestada información de las demás áreas a través del estimador sintético. Por tanto, no se ganará eficiencia respecto del estimador directo considerado.
- El peso que se da al estimador sintético no depende de lo bien explicada que esté la variable de interés por las variables auxiliares; es decir, no depende de la bondad de ajuste del modelo.
- No se pueden calcular para áreas o dominios no muestreados; es decir, con tamaño muestral n_d igual a cero.
- No se conocen estimadores estables del ECM bajo el diseño que al mismo tiempo sean distintos para cada área.
- Requieren un reajuste para verificar la propiedad "*benchmarking*": que la suma de los totales estimados en las áreas de una región mayor coincida con el estimador directo para dicha región.

Ejemplo 4. Estimadores compuestos de incidencias de pobreza, con R. Continuando los ejemplos anteriores, ilustramos ahora cómo obtener estimadores compuestos SSD de las incidencias de pobreza para las provincias, usando los estimadores directos HT obtenidos en el Ejemplo 1 y los post-estratificados sintéticos obtenidos en el Ejemplo 3. Para ello, llamamos a la función `ssd()` usando el valor por defecto del parámetro `delta` (`delta=1`) y guardamos los resultados:

```
povinc.ssd.res<-ssd(dom=prov,sweight=weight,domsize=sizeprov[,c(2,3)],
 direct=povinc.dir.res[,c("Domain","Direct")],synthetic=povinc.psedu.res)
povinc.ssd<-povinc.ssd.res$ssd
```

Analizamos el peso que otorga el estimador SSD al estimador directo para cada provincia mediante un resumen descriptivo de estos pesos:

```
summary(povinc.ssd.res$CompWeight)
```

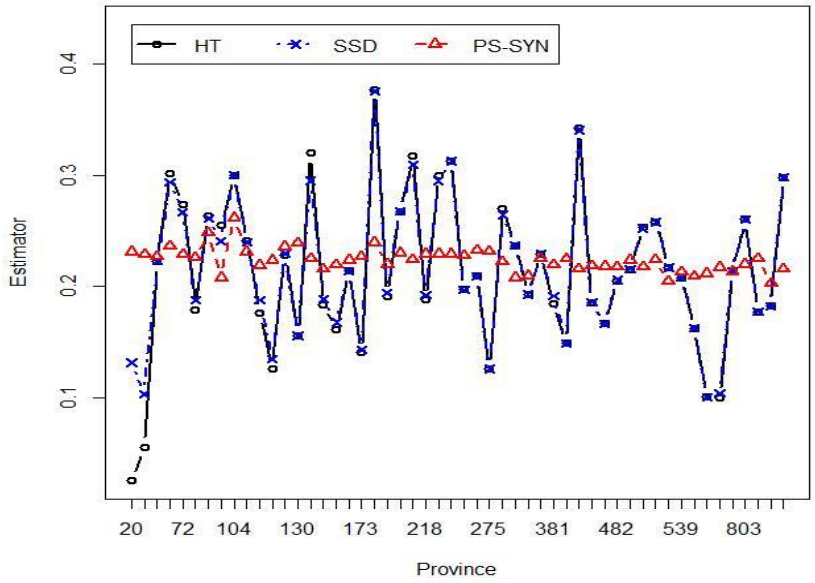
El resultado es:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4846  0.8800  0.9779  0.9224  1.0000  1.0000
```

Podemos ver que se le da peso igual a uno a los estimadores directos para al menos un cuarto de las provincias. En esas provincias concretas, no se está tomando información prestada de las demás. Por otro lado, en este estimador dicho peso no depende de la variable de interés. Si estimamos, por ejemplo, los ingresos medios, obtenemos exactamente los mismos pesos. Efectivamente, si comparamos gráficamente las estimaciones SSD con las directas HT y las post-estratificadas sintéticas (Gráfico 5), podemos ver que se parecen en gran medida a las estimaciones directas HT. En este gráfico, las provincias (en el eje) se ordenan de menor a mayor tamaño muestral, y se indican sus tamaños muestrales en las etiquetas del eje:

```
o<-order(nd)
k<-2
M<-max(povinc.psedu,povinc.dir,povinc.ssd)
m<-min(povinc.psedu,povinc.dir,povinc.ssd)
plot(1:D,povinc.dir[o],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",
 ylab="Estimator",xaxt="n")
points(1:D,povinc.dir[o],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:D,povinc.greg[o],type="b",col=3,lty=3,pch=3,lwd=2)
points(1:D,povinc.ssd[o],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:D,povinc.psedu[o],type="b",col=2,lty=2,pch=2,lwd=2)
axis(1, at=1:D, labels=nd[o])
legend(1,M+(M-m)/k,legend=c("HT","GREG","SSD","PS-SYN"),ncol=4,col=c(1,3,4,2),
 lwd=rep(2,3),lty=c(1,3,4,2),pch=c(1,3,4,2))
```

Gráfico 5
Estimaciones HT, PS-SYN y SSD de las incidencias de pobreza para cada provincia
(En proporciones)



Fuente: Elaboración propia.

V. Métodos indirectos basados en modelos

Los estimadores en áreas pequeñas basados en modelos entran dentro del grupo de los estimadores indirectos ya que toman prestada información de otras áreas. No obstante, son algo más sofisticados que los estimadores indirectos básicos estudiados en el capítulo IV, en el sentido de que incorporan heterogeneidad entre áreas no explicada por las variables auxiliares consideradas. Esto se realiza incorporando en el modelo de regresión considerado efectos aleatorios aditivos de las áreas. Veremos que estos efectos aleatorios proporcionan una muy buena propiedad a los estimadores basados en modelos lineales, y es que se pueden escribir como estimadores compuestos que tienden a un estimador directo en las áreas con suficiente tamaño muestral. Disponer de todas las variables que expliquen completamente la heterogeneidad entre áreas de nuestra variable de interés va a ocurrir en escasas ocasiones. Por tanto, estos modelos son sensiblemente más realistas que los modelos sintéticos, lo cual se traduce en estimadores con menor sesgo bajo el diseño muestral.

A. EBLUP basado en el modelo Fay-Herriot

El modelo Fay-Herriot (FH) es un modelo a nivel de área muy popular que fue introducido por Fay y Herriot (1979) para estimar los ingresos per cápita en áreas pequeñas de EE. UU. Este modelo se utiliza actualmente por la Oficina del Censo de EE. UU. (U.S. Census Bureau) dentro del programa "Small Area Income and Poverty Estimates (SAIPE)", para estimar proporciones de niños pobres en edad escolar en condados y distritos escolares, para más detalles véase Bell (1997) o http://www.census.gov/hhes/www/saipe. Este modelo también ha sido utilizado en Chile para estimar las tasas de incidencia de pobreza en las comunas chilenas, véase Casas-Cordero Valencia, Encina y Lahiri (2015), y en España para estimar la incidencia y la brecha de pobreza en las provincias por género, véase Molina y Morales (2009).

Este modelo enlaza los indicadores de interés para todas las áreas δ_d, $d = 1, \dots, D$, asumiendo que éstos varían respecto de un vector con valores de p variables auxiliares $\boldsymbol{x}_d$ de forma constante para todas las áreas, siguiento el modelo de regresión lineal

$$\delta_d = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d, \quad d = 1, \dots, D, \tag{19}$$

donde $\boldsymbol{\beta}$ es el vector de coeficientes común a todas las áreas y u_d es el término de error de la regresión, diferente para cada área, también conocido como efecto aleatorio del área d. Estos efectos aleatorios u_d

representan la heterogeneidad de los indicadores δ_d a través de las áreas, no debida a (o no explicada por) las variables auxiliares consideradas. En el modelo más simple, se asume que dichos efectos aleatorios u_d son independientes e idénticamente distribuidos (iid), con varianza común σ_u^2 desconocida; denotamos esto por $u_d \sim^{iid}(0, \sigma_u^2)$.

Dado que los verdaderos valores de los indicadores δ_d no son observables, el modelo (19) no se puede ajustar. Usando un estimador directo $\hat{\delta}_d^{DIR}$ de δ_d, debemos tener en cuenta que este estimador tiene un error debido al muestreo. El modelo FH considera que este estimador directo $\hat{\delta}_d^{DIR}$ es insesgado bajo el diseño. En este caso, podemos representar el error debido al muestreo de este estimador mediante el modelo:

$$\hat{\delta}_d^{DIR} = \delta_d + e_d, \quad d = 1, \dots, D, \tag{20}$$

donde e_d es el error de muestreo en el área d. Se asume que los errores de muestreo e_d son independientes entre sí y también son independientes de los efectos aleatorios de las áreas, u_d, tienen media cero y varianzas conocidas ψ_d; es decir, $e_d \sim^{ind}(0, \psi_d)$. En la práctica, 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. Combinando los modelos (19) y (20), se obtiene el modelo lineal mixto dado por

$$\hat{\delta}_d^{DIR} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \dots, D. \tag{21}$$

Calculando por medio del método de multiplicadores de Lagrange el estimador lineal en los datos $\hat{\delta}_d^{DIR}$, $d = 1, \dots, D$, que es insesgado bajo el modelo (21) y que minimiza el ECM bajo el modelo, obtenemos el mejor predictor lineal insesgado (en inglés, *best linear unbiased predictor*, BLUP) de $\delta_d = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d$. El estimador resultante se obtiene simplemente ajustando el modelo mixto (21); es decir, el BLUP bajo el modelo FH de δ_d viene dado por

$$\tilde{\delta}_d^{FH} = \boldsymbol{x}_d{}'\tilde{\boldsymbol{\beta}} + \tilde{u}_d, \tag{22}$$

donde $\tilde{u}_d = \gamma_d(\hat{\delta}_d^{DIR} - \boldsymbol{x}_d{}'\tilde{\boldsymbol{\beta}})$ es el BLUP de u_d, siendo $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d)$ y donde $\tilde{\boldsymbol{\beta}}$ es el estimador de mínimos cuadrados ponderados de $\boldsymbol{\beta}$ bajo el modelo (21), dado por

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

Obsérvese que, sustituyendo $\tilde{u}_d = \gamma_d(\hat{\delta}_d^{DIR} - \boldsymbol{x}_d{}'\tilde{\boldsymbol{\beta}})$ en el BLUP bajo el modelo FH dado en (22), podemos expresar el BLUP como una combinación lineal convexa del estimador directo y del estimador sintético de regresión, es decir,

$$\tilde{\delta}_d^{FH} = \gamma_d \hat{\delta}_d^{DIR} + (1 - \gamma_d)\boldsymbol{x}_d{}'\tilde{\boldsymbol{\beta}}, \tag{23}$$

con un peso para el estimador directo dado por $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d) \in (0,1)$. Este peso depende del tamaño muestral del área a través de la varianza ψ_d del estimador directo y de la bondad de ajuste del modelo sintético medido por σ_u^2 (en otras palabras, la heterogeneidad no explicada entre las áreas). Por tanto, para un área d en la que el estimador directo $\hat{\delta}_d^{DIR}$ sea eficiente debido a disponer de suficiente tamaño muestral; es decir, con una varianza muestral ψ_d pequeña comparada con la heterogeneidad no explicada σ_u^2, $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d)$ es cercano a uno y por tanto $\tilde{\delta}_d^{FH}$ le da más peso al estimador directo. Por otro lado, en las áreas d en las que el estimador directo carece de calidad debido al pequeño tamaño muestral, donde su varianza muestral ψ_d sea mayor que la heterogeneidad no explicada σ_u^2, entonces γ_d se acerca a cero y por tanto se le da más peso al estimador sintético de regresión $\boldsymbol{x}_d{}'\tilde{\boldsymbol{\beta}}$, el cual utiliza los datos de todas las áreas para estimar el parámetro común $\boldsymbol{\beta}$. Es decir, este estimador toma prestada información de las otras áreas a través del estimador sintético de regresión $\boldsymbol{x}_d{}'\tilde{\boldsymbol{\beta}}$ en la medida en que es necesario, dependiendo de la eficiencia del estimador directo.

Además, el hecho de que el BLUP $\tilde{\delta}_d^{FH}$ se acerque al estimador directo cuando el tamaño muestral del área es grande (ψ_d pequeña) es una propiedad muy deseable, pues no necesitamos saber cuándo un área es lo suficientemente "pequeña" para utilizar este estimador en lugar del estimador directo, ya que tiende al estimador directo cuando el tamaño muestral crece, y al mismo tiempo mejora al estimador

directo en las áreas con tamaño muestral pequeño. Por tanto, en principio este estimador se puede utilizar para todas las áreas mientras exista alguna "pequeña" (si no existiera ninguna, no sería necesario utilizarlo).

El BLUP de δ_d depende del verdadero valor de la varianza σ_u^2 de los efectos aleatorios u_d. En la práctica, esta varianza es desconocida y debemos estimarla. Métodos habituales de estimación son máxima verosimilitud (en inglés, *máximum likelihood*, ML) y máxima verosimilitud restringida o residual (en inglés, *restricted/residual* ML, REML). El método REML corrige el estimador ML de la varianza σ_u^2 por los grados de libertad debidos a estimar los coeficientes de la regresión $\boldsymbol{\beta}$ y proporciona por tanto un estimador menos sesgado para tamaño muestral finito n. Un método de ajuste basado en momentos, que no necesita una distribución paramétrica para obtener la verosimilitud, es el método propuesto por Fay y Herriot (1979), que llamamos método FH. Sea $\hat{\sigma}_u^2$ un estimador consistente de σ_u^2, como los obtenidos por dichos métodos. Reemplazando σ_u^2 por $\hat{\sigma}_u^2$ en (22), obtenemos el BLUP empírico (en inglés, *empirical BLUP*, EBLUP) de δ_d,

$$\hat{\delta}_d^{FH} = \hat{\gamma}_d \hat{\delta}_d^{DIR} + (1 - \hat{\gamma}_d) \boldsymbol{x}_d{}' \widehat{\boldsymbol{\beta}}, \tag{24}$$

donde $\hat{\gamma}_d = \hat{\sigma}_u^2 / (\hat{\sigma}_u^2 + \psi_d)$ y $\widehat{\boldsymbol{\beta}} = (\sum_{d=1}^{D} \hat{\gamma}_d \, \boldsymbol{x}_d \boldsymbol{x}_d{}')^{-1} \sum_{d=1}^{D} \hat{\gamma}_d \, \boldsymbol{x}_d \hat{\delta}_d^{DIR}$. En este documento, para resumir, llamaremos estimador FH al EBLUP basado en el modelo FH dado en (24).

Si los parámetros del modelo $\boldsymbol{\beta}$ y σ_u^2 son conocidos, el ECM del BLUP, $\tilde{\delta}_d^{FH}$, basado en el modelo (21) viene dado por

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

Por tanto, dado el verdadero valor del indicador δ_d, si σ_u^2 y $\boldsymbol{\beta}$ son conocidos, el BLUP bajo el modelo FH, $\tilde{\delta}_d^{FH}$, no puede ser menos eficiente que el estimador directo. En la práctica, σ_u^2 y $\boldsymbol{\beta}$ son estimados y el error debido a la estimación de estos dos parámetros se añade al ECM del estimador FH. Sin embargo, estos dos términos que se añaden al ECM tienden a cero cuando el número de áreas *D* tiende a infinito. Por tanto, para un número de áreas *D* suficiente, es probable que el estimador FH todavía mejore al estimador directo en cuanto a ECM. Por este motivo, estos estimadores suelen mejorar en la mayoría de las áreas mientras exista un número de áreas suficiente. Sin embargo, las ganancias en eficiencia pueden ser pequeñas si el número de áreas no es lo suficientemente grande. Los modelos a nivel de unidad, basados en el tamaño muestral total n, pueden ganar mucha más eficiencia que los modelos a nivel de área, mientras existan variables auxiliares a nivel de individuo que sean suficientemente informativas sobre la variable respuesta. Sin embargo, una ventaja del estimador FH dado en (24) es que usa los pesos del diseño muestral a través del estimador directo y es consistente bajo el diseño cuando el tamaño muestral del área n_d crece, mientras el peso del estimador directo es $\gamma_d > 0$. Además, su sesgo absoluto bajo el diseño viene dado por

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

con lo cual, será menos sesgado que el estimador sintético de regresión basado en el mismo vector de coeficientes $\boldsymbol{\beta}$ mientras $\gamma_d > 0$.

Para un área no muestreada; i.e., con tamaño muestral $n_d = 0$, la varianza del estimador directo ψ_d tendería a infinito y γ_d tendería a cero. Tomando el valor límite $\gamma_d = 0$, se obtiene el estimador sintético de regresión,

$$\hat{\delta}_d^{FH} = \boldsymbol{x}_d{}' \widehat{\boldsymbol{\beta}}.$$

Bajo normalidad de u_d y e_d, Prasad y Rao (1990) obtuvieron una aproximación de segundo orden (es decir, con error $o(D^{-1})$ cuando el número de áreas *D* es grande) para el ECM del estimador FH, dada por

$$\mathrm{MSE}(\hat{\delta}_d^{FH}) = g_{d1}(\sigma_u^2) + g_{d2}(\sigma_u^2) + g_{d3}(\sigma_u^2),$$

donde

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

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

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

Aquí, $\overline{\text{var}}(\hat{\sigma}_u^2)$ es la varianza asintótica del estimador $\hat{\sigma}_u^2$ de σ_u^2, que depende del método de estimación utilizado, $g_{1d}(\sigma_u^2)$ es el error debido a la predicción del efecto aleatorio del área u_d, de orden $O(1)$ cuando D crece (es decir, no tiende a cero), $g_{2d}(\sigma_u^2)$ es el error debido a la estimación del vector de coeficientes de regresión $\boldsymbol{\beta}$ y $g_{3d}(\sigma_u^2)$ es el error debido a la estimación de la varianza σ_u^2, donde los dos últimos términos tienden a cero cuando D crece con orden $O(D^{-1})$; es decir, a la misma velocidad que D^{-1}. Esto significa que $g_{2d}(\sigma_u^2)$ y $g_{3d}(\sigma_u^2)$ desaparecen para D lo suficientemente grande, mientras que $g_{1d}(\sigma_u^2)$ no desaparece, pero para D moderado es necesario tener en cuenta los tres términos para evitar infraestimación del ECM.

Si $\hat{\sigma}_u^2$ es el estimador REML, la varianza asintótica se obtiene como el inverso de la información de Fisher $\mathcal{I}(\sigma_u^2)$, y viene dada por

$$\overline{\text{var}}(\hat{\sigma}_u^2) = \mathcal{I}^{-1}(\sigma_u^2) = 2\left\{\sum_{d=1}^{D}(\sigma_u^2+\psi_d)^{-2}\right\}^{-1}. \quad (25)$$

En este caso, $g_{d2}(\hat{\sigma}_u^2)$ y $g_{d3}(\hat{\sigma}_u^2)$ son estimadores respectivos de $g_{2d}(\sigma_u^2)$ y $g_{3d}(\sigma_u^2)$ insesgados de segundo orden. Esto significa que su sesgo es $o(D^{-1})$, es decir, tiende a cero más rápido que D^{-1} cuando D crece. Sin embargo, $g_{d1}(\hat{\sigma}_u^2)$ tiene un sesgo no despreciable como estimador de $g_{d1}(\sigma_u^2)$ que resulta ser igual a $-g_{3d}(\sigma_u^2)+o(D^{-1})$. Por tanto, para corregir el sesgo de $g_{1d}(\hat{\sigma}_u^2)$, debemos sumar dos veces $g_{3d}(\hat{\sigma}_u^2)$. Así, un estimador insesgado de segundo orden del ECM del estimador FH, llamado aquí estimador Prasad-Rao, viene entonces dado por

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

Si $\hat{\sigma}_u^2$ es el estimador ML, su varianza asintótica es la misma que para el estimador REML, dada en (25). Sin embargo, este estimador tiene un sesgo que viene dado por

$$b(\sigma_u^2) = -\{2\mathcal{I}(\sigma_u^2)\}^{-1}\text{traza}\left[\left\{\sum_{d=1}^{D}(\sigma_u^2+\psi_d)^{-1}\boldsymbol{x}_d\boldsymbol{x}_d{}'\right\}^{-1}\sum_{d=1}^{D}(\sigma_u^2+\psi_d)^{-2}\boldsymbol{x}_d\boldsymbol{x}_d{}'\right].$$

En este caso, el sesgo del estimador ML añade un término al sesgo de $g_{d1}(\hat{\sigma}_u^2)$ como estimador de $g_{d1}(\sigma_u^2)$. Este sesgo es igual a $b(\sigma_u^2)\nabla g_{1d}(\sigma_u^2) - g_{3d}(\sigma_u^2)$, donde

$$\nabla g_{1d}(\sigma_u^2) = (1-\gamma_d)^2.$$

Como $b(\hat{\sigma}_u^2)\nabla g_{1d}(\hat{\sigma}_u^2)$ sí es un estimador insesgado de segundo orden de $b(\sigma_u^2)\nabla g_{1d}(\sigma_u^2)$, podemos corregir el sesgo de $g_{1d}(\hat{\sigma}_u^2)$ restando este término. De esta forma, obtenemos el siguiente estimador insesgado de segundo orden del ECM del estimador FH,

$$\text{mse}_{PR}(\hat{\delta}_d^{FH}) = g_{d1}(\hat{\sigma}_u^2) - b(\hat{\sigma}_u^2)\nabla g_{1d}(\hat{\sigma}_u^2) + g_{d2}(\hat{\sigma}_u^2) + 2g_{d3}(\hat{\sigma}_u^2). \quad (26)$$

Si $\hat{\sigma}_u^2$ es el estimador obtenido por el método FH basado en momentos, el estimador insesgado de segundo orden del ECM tiene la misma forma que (26), pero el sesgo del estimador FH de σ_u^2 y la varianza asintótica cambian, y vienen dados por

$$\overline{\text{var}}(\hat{\sigma}_u^2) = 2D\left\{\sum_{d=1}^{D}(\sigma_u^2+\psi_d)^{-1}\right\}^{-2}, \quad (27)$$

$$b(\sigma_u^2) = \frac{2[D\sum_{d=1}^{D}(\sigma_u^2+\psi_d)^{-2} - \{\sum_{d=1}^{D}(\sigma_u^2+\psi_d)^{-1}\}^2]}{\{\sum_{d=1}^{D}(\sigma_u^2+\psi_d)^{-1}\}^3}.$$

A la hora de estimar mediante el modelo FH el indicador FGT de orden α, $\delta_d = F_{\alpha d}$, se deben encontrar variables auxiliares $\boldsymbol{x}_d$ que verifiquen el modelo

$$F_{\alpha d} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d, \quad d = 1, \dots, D \tag{28}$$

y se asume que el estimador directo $\hat{F}_{\alpha d}^{DIR}$ de $F_{\alpha d}$ satisface

$$\hat{F}_{\alpha d}^{DIR} = F_{\alpha d} + e_d, \quad d = 1, \dots, D. \tag{29}$$

El modelo lineal mixto que se obtiene combinando (28) y (29) viene dado por

$$\hat{F}_{\alpha d}^{DIR} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \dots, D. \tag{30}$$

Ajustando este modelo, el BLUP de $F_{\alpha d} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d$ sería

$$\tilde{F}_{\alpha d}^{FH} = \boldsymbol{x}_d{}'\tilde{\boldsymbol{\beta}} + \tilde{u}_d, \tag{31}$$

donde, en este caso, $\tilde{u}_d = \gamma_d(\hat{F}_{\alpha d}^{DIR} - \boldsymbol{x}_d{}'\tilde{\boldsymbol{\beta}})$ es el BLUP de u_d y $\tilde{\boldsymbol{\beta}}$ se calcula de la forma

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

El estimador FH final de $F_{\alpha d}$ se obtiene simplemente reemplazando la varianza σ_u^2 por un estimador consistente $\hat{\sigma}_u^2$ en el BLUP (31).

Resumen de las características del estimador FH:

Indicadores objetivo: parámetros generales.
Requerimientos de datos:

- Datos agregados (e.g. medias poblacionales) de las p variables auxiliares consideradas en las áreas, $\boldsymbol{x}_d$, $d = 1, \dots, D$.

Ventajas:

- Suele mejorar la eficiencia del estimador directo.
- El modelo de regresión considerado incorpora heterogeneidad no explicada entre las áreas.
- Es un estimador compuesto que automáticamente toma prestada información del resto de áreas (dando mayor peso al estimador sintético de regresión) en la medida en que es necesario (cuando el estimador directo tiene mayor varianza, o menor tamaño muestral). Tiende al estimador directo cuando el tamaño del área crece (ya que ψ_d se hace pequeño).
- Si para un área d, el peso dado al estimador directo es estrictamente positivo ($\gamma_d > 0$), se usan los pesos muestrales w_{di} a través del estimador directo $\hat{\delta}_d^{DIR}$; es decir, se tiene en cuenta el diseño muestral. Como consecuencia, es consistente bajo el diseño (al igual que el estimador directo). Esto significa que se verá menos afectado por diseños informativos (diseños con probabilidades de selección de los individuos dependiendo de la variable de interés), considerando que los pesos muestrales son los verdaderos.
- Debido a que se utilizan datos agregados, el estimador FH no se encuentra excesivamente afectado por datos atípicos aislados (en este caso estimadores directos atípicos para algún área).
- Al usar solo información auxiliar agregada, evita los problemas de confidencialidad de los microdatos obtenidos de un censo o registro administrativo.
- 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.
- Se puede estimar en áreas no muestreadas.
- El estimador Prasad-Rao del ECM es estable (o eficiente) y es insesgado bajo el diseño cuando se promedia a lo largo de muchas áreas.

Desventajas:

- Los estimadores se basan en un modelo; por tanto, es necesario analizar dicho modelo (e.g. a través de los residuos). Para parámetros no lineales, podemos tener problemas de linealidad.
- Las varianzas muestrales de los estimadores directos ψ_d se asumen conocidas, aunque en la práctica es necesario estimarlas, lo cual sufre del mismo problema (la falta de datos en un área). Incorporar el error de estimación de estas varianzas en el ECM del estimador FH no es automático y a menudo el ECM estimado no incorpora este error.
- El número de observaciones usadas para ajustar el modelo es el número de áreas muestreadas, el cual suele ser mucho menor que el tamaño muestral total n usado para ajustar los modelos a nivel de individuo. Por tanto, los parámetros del modelo se estiman con menor eficiencia y las ganancias en eficiencia respecto a los estimadores directos serán menores que con los modelos a nivel de individuo (esta eficiencia aumenta con el número de áreas).
- A la hora de estimar varios indicadores que dependen de una variable común (e.g., $F_{\alpha d}$ para distintos valores de α), al contrario que los métodos basados en modelos a nivel de unidad, se requiere modelización y búsqueda de variables auxiliares útiles para cada uno de los indicadores por separado.
- El estimador del ECM bajo el modelo de Prasad-Rao es correcto bajo el modelo con normalidad de u_d y e_d, y no es insesgado bajo el diseño para el ECM bajo el diseño para un área concreta.
- Una vez se ha ajustado el modelo a nivel de área, los estimadores $\hat{\delta}_d^{FH}$ no se pueden desagregar para subdominios o subáreas dentro de las áreas a menos que se encuentre un nuevo modelo modelo adecuado para ese nuevo nivel o, alternativamente, se ajuste un modelo con efectos aleatorios a varios niveles.
- Requieren un reajuste para verificar la propiedad "*benchmarking*": que la suma de los totales estimados en las áreas de una región mayor coincida con el estimador directo para dicha región.

Ejemplo 5. Estimadores FH de incidencias de pobreza, con R. Continuando con los ejemplos anteriores, ilustramos cómo obtener en R estimadores FH de las incidencias de pobreza para las provincias.

En primer lugar, para comprobar si la hipótesis de normalidad del modelo se verifica, podemos analizar gráficamente la distribución de los estimadores directos de las incidencias de pobreza a través del histograma:

```
hist(povinc.dir,prob=TRUE,main="",xlab="HT estimators pov.
incidence")
```

La forma de este histograma (no se incluye por brevedad) es algo asimétrica pero no se aleja demasiado de una densidad normal, lo cual es esperable puesto que se aplica el Teorema Central del Límite a los estimadores directos de las áreas.

A continuación, cargamos los conjuntos de datos con los tamaños poblacionales de las provincias y los mismos por grupos de nacionalidad, edad y estado laboral (algunos ya estaban cargados en los ejemplos anteriores):

```
data(sizeprov)
data(sizeprovnat)
data(sizeprovage)
data(sizeprovedu)
data(sizeprovlab)
```

Usamos estos tamaños poblacionales para calcular las proporciones de individuos en cada categoría dentro de cada provincia. Éstos serán nuestras variables explicativas en un modelo Fay-Herriot:

```
Nd<-sizeprov[,3]
Ndnat<-as.matrix(sizeprovnat[,-c(1,2)])
Ndage<-as.matrix(sizeprovage[,-c(1,2)])
Ndedu<-as.matrix(sizeprovedu[,-c(1,2)])
Ndlab<-as.matrix(sizeprovlab[,-c(1,2)])

Pdnat<-Ndnat/Nd
Pdage<-Ndage/Nd
Pdedu<-Ndedu/Nd
Pdlab<-Ndlab/Nd

# Matriz de diseño para modelo FH
X<-cbind(const=rep(1,D),nat1=Pdnat[,1],Pdage[,3:5],Pdedu[,c(1,3)],Pdlab[,c(2,3)])
```

Llamamos a la función que calcula los estimadores FH de las incidencias de pobreza para las provincias, usando los estimadores directos HT obtenidos en el Ejemplo 1 y sus correspondientes varianzas muestrales:

```
povinc.FH.res<-eblupFH(povinc.dir~X-1,vardir=povinc.dir.res$SD^2)
povinc.FH<-povinc.FH.res$eblup
```

Usando los coeficientes de regresión estimados obtenidos del ajuste del modelo Fay-Herriot, podemos calcular también estimadores sintéticos de regresión basados en el modelo a nivel de área:

```
povinc.rsyn1<-X%*%povinc.FH.res$fit$estcoef[,1]
```

Aunque estos estimadores están basados en el estimador de los coeficientes de regresión obtenidos del ajuste del modelo Fay-Herriot y no del modelo sintético, también son estimadores sintéticos pues no consideran heterogeneidad entre áreas no explicada por las variables auxiliares consideradas. Además, los estimadores de los coeficientes de regresión obtenidos bajo ambos modelos, usando las mismas variables auxiliares, son asintóticamente equivalentes. Por tanto, para un número de áreas grande, ambos serán muy similares.

Como los estimadores FH son estimadores compuestos entre los directos y los sintéticos de regresión, calculamos los pesos que se dan a los estimadores directos en la composición y mostramos un resumen descriptivo de éstos:

```
gammad<-povinc.FH.res$fit$refvar/(povinc.FH.res$fit$refvar+povinc.dir.res$SD^2)
summary(gammad)
```

Resultado:

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.4537  0.7182  0.8108  0.7906  0.8977  0.9477
```

Vemos que, a diferencia de los estimadores SSC, en este caso el peso que se le otorga al estimador directo no es igual a uno para ninguna provincia, aunque sí toma valores cercanos a uno para algunas provincias.

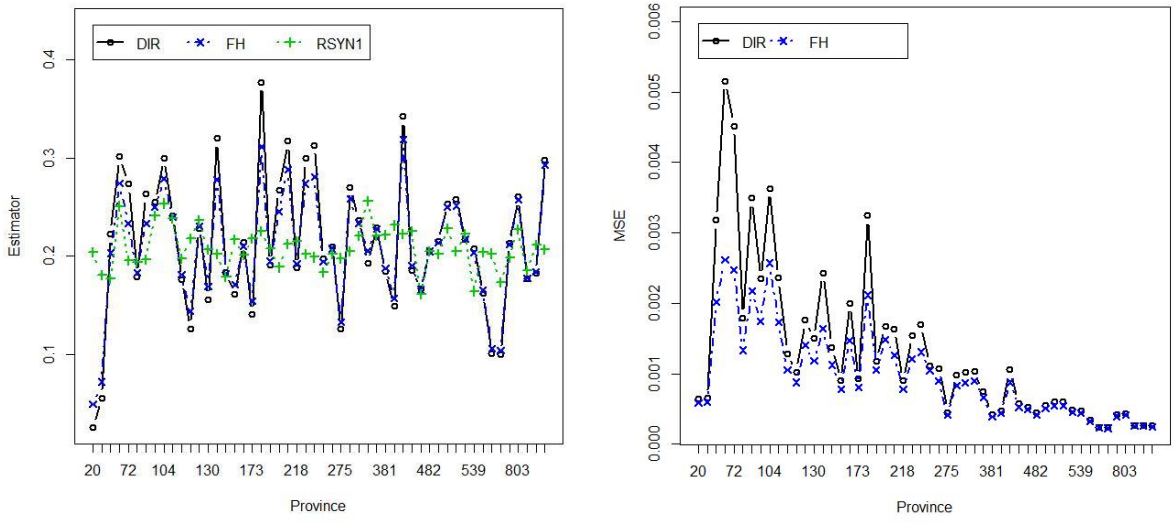
Ahora comparamos gráficamente las estimaciones FH con las directas HT y sintéticas (llamadas RSYN1) para cada provincia. Las provincias (en el eje) se ordenan de menor a mayor tamaño muestral, e indicamos los tamaños muestrales de éstas en el eje:

```
o<-order(nd)
k<-6
M<-max(povinc.dir,povinc.FH,povinc.rsyn1)
m<-min(povinc.dir,povinc.FH,povinc.rsyn1)
plot(1:D,povinc.dir[o],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="Estimator",
 xaxt="n")
points(1:D,povinc.dir[o],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:D,povinc.FH[o],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:D,povinc.rsyn1[o],type="b",col=3,lty=3,pch=3,lwd=2)
axis(1, at=1:D, labels=nd[o])
legend(1,M+(M-m)/k,legend=c("DIR","FH","RSYN1"),ncol=3,col=c(1,4,3),lwd=rep(2,3),
 lty=c(1,4,3),pch=c(1,4,3))
```

Finalmente, estimamos el ECM de los estimadores FH, llamando a la función `mseFH()`, calculamos los CVs estimados y graficamos los ECMs junto a las varianzas de los estimadores directos:

```
povinc.FH.mse.res<-mseFH(povinc.dir~X-1,vardir=povinc.dir.res$SD^2)

povinc.FH.mse<-povinc.FH.mse.res$mse
povinc.FH.cv<-100*sqrt(povinc.FH.mse)/povinc.FH
M<-max(povinc.dir.var,povinc.FH.mse)
m<-min(povinc.dir.var,povinc.FH.mse)
plot(1:D,povinc.dir.cv[o],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="CV",xaxt="n")
points(1:D,povinc.dir.var[o],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:D,povinc.FH.mse[o],type="b",col=4,lty=4,pch=4,lwd=2)
axis(1, at=1:D, labels=nd[o])
legend(1,M+(M-m)/k,legend=c("DIR","FH"),ncol=3,col=c(1,4),lwd=rep(2,2),lty=c(1,4),pch=c(1,4))
```

Gráfico 6
Estimaciones FH, directas HT y RSYN1 de las incidencias de pobreza para las provincias (izquierda), y ECMs estimados de los estimadores FH y directos HT (derecha)
(En proporciones)



Fuente: Elaboración propia.

De nuevo, podemos ver en el Gráfico 6 (izquierda) que los estimadores sintéticos de regresión toman valores similares para todas las provincias, a diferencia de los estimadores directos, que varían más a lo largo de éstas. Los estimadores FH se acercan a los directos, pero al mismo tiempo toman información prestada de las demás provincias a través de los estimadores sintéticos, especialmente para las provincias de menor tamaño muestral (parte izquierda del gráfico). Aunque en este ejemplo las variables auxiliares consideradas no sean muy potentes, el Gráfico 6 (derecha) sugiere que los estimadores FH son más eficientes que los directos.

Finalmente, vamos a comparar los CVs estimados para los estimadores HT, GREG y FH para las 5 provincias con menores tamaños muestrales:

```
compardirFH<-data.frame(povinc.dir.cv,povinc.greg.cv,povinc.FH.cv)

selprov<-o[1:5]
compardirFH[selprov,]
```

Resultados:

```
   povinc.dir.CV povinc.greg.cv povinc.FH.cv
42      99.97815       94.72703     49.34572
5       46.35946       42.04802     33.74811
40      25.33449       21.77035     21.64444
34      23.80085       19.02477     18.27171
44      24.57017       16.86049     20.47468
```

Podemos ver la reducción en los CVs que consiguen los estimadores FH en comparación con los estimadores directos HT. También ganan eficiencia respecto de los estimadores GREG para las cuatro provincias de menores tamaños muestrales, y las ganancias son considerables para las dos provincias de tamaños muestrales más pequeños.

B. EBLUP basado en el modelo con errores anidados

El modelo con errores anidados fue propuesto por Battese, Harter y Fuller (1977) para estimar la producción de maíz y soja en condados de EE. UU. Este modelo relaciona de forma lineal los valores de una variable de interés Y_{di} para el individuo i dentro del área d, con los valores de p variables auxiliares para ese mismo individuo, de la forma

$$Y_{di} = \boldsymbol{x}_{di}{}'\boldsymbol{\beta} + u_d + e_{di}, \quad i = 1, \dots, N_d,\ d = 1, \dots, D, \tag{32}$$

donde $\boldsymbol{\beta}$ es el vector de coeficientes de las variables auxiliares, común para todas las áreas, u_d es el efecto aleatorio del área y e_{di} es el error a nivel de individuo. Los efectos aleatorios representan la heterogeneidad no explicada de los valores Y_{di} a través de las áreas. Los efectos aleatorios se consideran independientes de los errores, con $u_d \sim^{iid}(0, \sigma_u^2)$ y $e_{di} \sim^{ind}(0, \sigma_e^2 k_{di}^2)$, siendo k_{di} constantes conocidas que representan la posible heteroscedasticidad.

Obsérvese que la media del área d se puede descomponer en la suma de los valores observados en la muestra y los no muestreados, de la forma

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

No es necesario predecir los valores observados en la muestra pues nos vienen dados. El BLUP de $\bar{Y}_d$ bajo el modelo con errores anidados (32) se obtiene simplemente ajustando el modelo a los datos de la muestra y prediciendo los valores de las variables Y_{di} fuera de la muestra del área d, es decir,

$$\tilde{\bar{Y}}_d^{BLUP} = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} \tilde{Y}_{di}^{BLUP}\right), \tag{33}$$

donde, tomando el estimador de mínimos cuadrados ponderados $\tilde{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ bajo el modelo (32), los valores predichos son

$$\tilde{Y}_{di}^{BLUP} = \boldsymbol{x}_{di}{}'\tilde{\boldsymbol{\beta}} + \tilde{u}_d,$$
$$\tilde{u}_d = \gamma_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}{}'\tilde{\boldsymbol{\beta}}), \gamma_d = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/a_{d\cdot}),$$

siendo $\bar{y}_{da} = a_{d\cdot}^{-1}\sum_{i\in s_d} a_{di}\, Y_{di}$ y $\bar{\boldsymbol{x}}_{da} = a_{d\cdot}^{-1}\sum_{i\in s_d} a_{di}\,\boldsymbol{x}_{di}$ las medias muestrales ponderadas de la variable respuesta y las variables auxiliares, respectivamente, con pesos $a_{di} = k_{di}^{-2}$, y donde $a_{d\cdot} = \sum_{i\in s_d} a_{di}$. De nuevo $\tilde{u}_d$ es el BLUP de u_d y los valores predichos $\tilde{Y}_{di}^{BLUP}$ son los BLUPs de las variables Y_{di}, $i \in r_d$, bajo el modelo (32).

Construimos el vector de variables respuesta para el área d, $\boldsymbol{y}_d = (Y_{d1}, \dots, Y_{dN_d})'$ y la matriz correspondiente de variables auxiliares, $\boldsymbol{X}_d = (\boldsymbol{x}_{d1}, \dots, \boldsymbol{x}_{dN_d})'$. Bajo el modelo de errores anidados (32), $\boldsymbol{y}_d \sim^{ind} N(\boldsymbol{X}_d\boldsymbol{\beta}, \boldsymbol{V}_d)$, $d = 1, \dots, D$, donde

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

donde $\boldsymbol{A}_d = \mathrm{diag}(k_{di}^2; i = 1, \dots, N_d)$. Ahora descomponemos el vector $\boldsymbol{y}_d$ del área d en los subvectores para las unidades de la muestra y para las unidades fuera de la muestra de la forma $\boldsymbol{y}_d = (\boldsymbol{y}_{ds}{}', \boldsymbol{y}_{dr}{}')'$, y las matrices $\boldsymbol{X}_d$ y $\boldsymbol{V}_d$ de la misma forma,

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

Con esta notación, el estimador de mínimos cuadrados ponderados de $\boldsymbol{\beta}$ viene dado por

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \boldsymbol{X}_{ds}\boldsymbol{V}_{ds}^{-1}\boldsymbol{X}_{ds}{}'\right)^{-1}\sum_{d=1}^{D}\boldsymbol{X}_{ds}\boldsymbol{V}_{ds}^{-1}\boldsymbol{y}_{ds}. \tag{34}$$

Para áreas con fracción muestral despreciable, es decir, donde $n_d/N_d \approx 0$, el BLUP de la media $\bar{Y}_d$ se puede escribir de la forma

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

Como $\gamma_d \in (0,1)$, el BLUP es una media ponderada entre el estimador $\bar{y}_{da} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_{da})'\tilde{\boldsymbol{\beta}}$, conocido como estimador "*survey regression*" y el estimador sintético de regresión, $\bar{\boldsymbol{X}}_d{}'\tilde{\boldsymbol{\beta}}$. El estimador "*survey regression*" se obtiene de ajustar el mismo modelo (32), pero tomando los efectos de las áreas u_d como fijos en lugar de aleatorios. Además, obsérvese que esta media ponderada es similar a la obtenida mediante el estimador FH dado en (24), pero donde el estimador "survey-regression" $\bar{y}_{da} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_{da})'\tilde{\boldsymbol{\beta}}$ hace el papel de estimador directo. En efecto, este estimador se puede considerar como directo, pues su varianza es $O(n_d^{-1})$; es decir, su varianza aumenta cuando el tamaño muestral del área n_d se hace pequeño.

Para interpretar este estimador, por simplicidad consideremos un modelo homoscedástico; es decir, con $k_{di} = 1$ para todo i y d. En este caso, se tiene $\gamma_d = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/n_d)$. Para un área con tamaño muestral n_d pequeño, γ_d es cercano a cero y el BLUP se acerca al estimador sintético de regresión, el cual toma información prestada de las demás áreas. Sin embargo, para un área con tamaño muestral n_d grande, γ_d se acerca a uno y el BLUP se acerca al estimador "*survey regression*". Además, γ_d también depende de la heterogeneidad entre áreas medida por σ_u^2. Si las áreas son muy heterogéneas (σ_u^2 grande comparada con σ_e^2/n_d), o equivalentemente, si las variables auxiliares consideradas no explican gran parte de la variabilidad, entonces γ_d se acerca a uno y se le da mayor peso al estimador "*survey regression*", que es

similar a un estimador directo. En caso contrario, si las áreas son homogéneas o, en otras palabras, las variables auxiliares son predictores potentes, entonces se le da mayor peso al estimador sintético obtenido mediante la regresión con estas variables auxiliares.

De nuevo, el BLUP dado en (33) depende de los verdaderos valores de las componentes de la varianza del modelo (32), $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)'$. Sustituyendo el verdadero $\boldsymbol{\theta}$ por un estimador consistente $\widehat{\boldsymbol{\theta}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ en el BLUP (33), obtenemos el EBLUP, dado por

$$\hat{\bar{Y}}_d^{EBLUP} = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} \hat{Y}_{di}^{EBLUP}\right), \tag{35}$$

donde, llamando $\widehat{\boldsymbol{\beta}}$ al resultado de sustituir $\boldsymbol{\theta}$ por el estimador $\widehat{\boldsymbol{\theta}}$ en $\widetilde{\boldsymbol{\beta}}$ dado en (34), los valores predichos son ahora

$$\hat{Y}_{di}^{EBLUP} = \boldsymbol{x}_{di}{}'\widehat{\boldsymbol{\beta}} + \hat{u}_d,$$
$$\hat{u}_d = \hat{\gamma}_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}{}'\widehat{\boldsymbol{\beta}}), \hat{\gamma}_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_e^2/a_{d\cdot}),$$

El BLUP es insesgado bajo el modelo (32) y es óptimo, en el sentido de minimizar el ECM, entre los estimadores lineales en la muestra e insesgados. Al sustituir $\boldsymbol{\theta}$ por el estimador $\widehat{\boldsymbol{\theta}}$, el EBLUP sigue siendo insesgado bajo el modelo (32), bajo ciertas condiciones sobre el estimador $\widehat{\boldsymbol{\theta}}$. Los métodos de estimación habituales, concretamente ML, REML y el método de Henderson III, satisfacen estas condiciones. Sin embargo, ni el BLUP ni el EBLUP son insesgados bajo el diseño muestral. De hecho, no tienen en cuenta el diseño muestral y por tanto en principio están diseñados para muestreo aleatorio simple (MAS). En cualquier caso, los EBLUPs aumentan considerablemente la eficiencia respecto de los estimadores directos e incluso respecto de los estimadores FH, ya que utilizan información mucho más detallada y de forma más eficiente (sin reducir los datos a medias). Bajo diseños muestrales con probabilidades desiguales, pueden tener un sesgo bajo el diseño no despreciable. You y Rao (2002) propusieron una variación denominada pseudo EBLUP que incluye los pesos muestrales y es consistente bajo el diseño cuando el tamaño del área n_d crece.

Para un área no muestreada, es decir, con tamaño muestral $n_d = 0$, tomando $\gamma_d = 0$, obtenemos el estimador sintético de regresión $\bar{\boldsymbol{X}}_d{}'\widehat{\boldsymbol{\beta}}$.

Bajo MAS y tomando $k_{di} = 1$, para todo i y d, dado que el estimador "*survey regression*" es aproximadamente insesgado bajo el diseño, el sesgo bajo el diseño del BLUP cuando $n_d/N_d \approx 0$ es $-(1-\gamma_d)(\bar{Y}_d - \bar{\boldsymbol{X}}_d{}'\boldsymbol{\beta})$. Por tanto, el sesgo absoluto relativo (SAR) bajo el diseño es igual a

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

es decir, es menor que el sesgo absoluto relativo bajo el diseño del estimador sintético de regresión $\bar{\boldsymbol{X}}_d{}'\boldsymbol{\beta}$ para el mismo vector de coeficientes $\boldsymbol{\beta}$, $|(\bar{Y}_d - \bar{\boldsymbol{X}}_d{}'\boldsymbol{\beta})/\bar{Y}_d|$, mientras $\gamma_d > 0$. Si establecemos un límite superior B para el sesgo absoluto relativo (e.g. $B = 0.20$ o $B = 0.10$), en el caso de que se supere este límite B para alguna de las áreas, podemos reemplazar el sesgo absoluto relativo del estimador sintético por una cantidad constante para cada área, como por ejemplo el máximo; es decir, consideramos

$$M = \max_{1\leq d\leq D}\left|\frac{\bar{Y}_d - \bar{\boldsymbol{X}}_d{}'\boldsymbol{\beta}}{\bar{Y}_d}\right|.$$

La cantidad $(1-\gamma_d)M$ decrece de forma monótona con el tamaño muestral del área n_d, a través de γ_d. Podemos encontrar el tamaño muestral n_d^* a partir del cual $(1-\gamma_d)M$ supera B. Si $M > B$ (en caso contrario el SAR no supera B para ninguna provincia), el tamaño muestral resultante es

$$n_d^* = \frac{\sigma_e^2}{\sigma_u^2}\left(\frac{M}{B} - 1\right).$$

Así, para las áreas con tamaño muestral $n_d < n_d^*$, el sesgo absoluto relativo podría superar el límite superior B y podemos decidir no producir estimaciones para dichas áreas. Sin embargo, n_d^* depende de ciertas cantidades desconocidas. Por tanto, en la práctica estimamos dichas cantidades desconocidas y obtenemos un valor estimado de n_d^*. Un estimador sería

$$\hat{n}_d^* = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_u^2}\left(\frac{\widehat{M}}{B} - 1\right),$$

donde

$$\widehat{M} = \max_{1 \le d \le D} \left| \frac{\hat{\bar{Y}}_d^{EBLUP} - \bar{\boldsymbol{X}}_d{}' \widehat{\boldsymbol{\beta}}}{\hat{\bar{Y}}_d^{EBLUP}} \right|,$$

suponiendo que $\widehat{M} > B$.

El ECM del EBLUP $\hat{\bar{Y}}_d^{EBLUP}$ de $\bar{Y}_d$, así como un estimador de segundo orden de este ECM, se pueden aproximar mediante una fórmula analítica correcta de segundo orden para D grande de forma muy similar a la fórmula de Prasad-Rao que se describe en la introducción para el estimador FH. Otra opción que no requiere un número de áreas D grande, aunque computacionalmente más costosa, es recurrir a procedimientos de remuestreo. Aquí describimos un procedimiento *Bootstrap* paramétrico para poblaciones finitas propuesto por González-Manteiga et al. (2008), particularizado aquí para la estimación de las medias de las áreas, $\bar{Y}_d$. El procedimiento *Bootstrap* es el siguiente:

1. Ajustar el modelo (32) a los datos de la muestra $\boldsymbol{y}_s = (\boldsymbol{y}_{1s}{}', \dots, \boldsymbol{y}_{Ds}{}')'$ y obtener los estimadores de los parámetros del modelo $\widehat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ y $\hat{\sigma}_e^2$.
2. Generar los efectos de las áreas de la forma $u_d^{*(b)} \sim^{iid} N(0, \hat{\sigma}_u^2)$, $d = 1, \dots, D$.
3. Generar, independientemente de los efectos de las áreas $u_d^{*(b)}$, errores *Bootstrap* para las unidades de la muestra en el área, $e_{di}^{*(b)} \sim^{iid} N(0, \hat{\sigma}_e^2)$, $i \in s_d$. Generar también las medias poblacionales de los errores en las áreas, $\bar{E}_d^{*(b)} \sim^{iid} N(0, \hat{\sigma}_e^2/N_d)$, $d = 1, \dots, D$.
4. Calcular las verdaderas medias *Bootstrap* de las áreas,
$$\bar{Y}_d^{*(b)} = \bar{\boldsymbol{X}}_d{}' \widehat{\boldsymbol{\beta}} + u_d^{*(b)} + \bar{E}_d^{*(b)}, \quad d = 1, \dots, D.$$
Obsérvese que el cómputo de la media $\bar{Y}_d^{*(b)}$ no requiere los valores individuales $\boldsymbol{x}_{di}$, para cada unidad fuera de la muestra del área $i \in r_d$.
5. Usando los vectores de valores de las variables auxiliares para las unidades de la muestra $\boldsymbol{x}_{di}$, $i \in s_d$, generar las variables respuesta para las unidades de la muestra a partir del modelo
$$Y_{di}^{*(b)} = \boldsymbol{x}_{di}{}' \widehat{\boldsymbol{\beta}} + u_d^{*(b)} + e_{di}^{*(b)}, \quad i \in s_d, \quad d = 1, \dots, D. \qquad (36)$$
6. Para la muestra original $s = s_1 \cup \cdots \cup s_D$, sea $\boldsymbol{y}_s^{*(b)} = ((\boldsymbol{y}_{1s}^{*(b)})', \dots, (\boldsymbol{y}_{Ds}^{*(b)})')'$ el vector *Bootstrap* de valores en la muestra. Ajustar el modelo (32) a los datos *Bootstrap* $\boldsymbol{y}_s^{*(b)}$ y calcular los EBLUPs *Bootstrap* $\hat{\bar{Y}}_d^{EBLUP*(b)}$, $d = 1, \dots, D$.
7. Repetir los pasos 2)–6), para $b = 1, \dots, B$ y obtenemos las medias verdaderas $\bar{Y}_d^{*(b)}$ y los correspondientes EBLUPs $\hat{\bar{Y}}_d^{EBLUP*(b)}$ para la repetición *Bootstrap* b. Los estimadores "*naive Bootstrap*" del ECM de los EBLUPs $\hat{\bar{Y}}_d^{EBLUP}$, obtenidos mediante el *Bootstrap* paramétrico son
$$mse_B\left(\hat{\bar{Y}}_d^{EBLUP}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\bar{Y}}_d^{EBLUP*(b)} - \bar{Y}_d^{*(b)}\right)^2, \qquad (37)$$
$$d = 1, \dots, D.$$

El estimator *Bootstrap* (37) 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. Existen distintas correcciones

de sesgo en la literatura pero, o bien producen estimadores que pueden tomar indeseados valores negativos, o bien son estrictamente positivos pero no insesgados de segundo orden. Además, estas correcciones aumentan la varianza del estimador del ECM. Por tanto, el estimador "*naive Bootstrap*" que no realiza corrección de sesgo es una opción aceptable dentro de los estimadores no analíticos.

Resumen de características del EBLUP basado en el modelo con errores anidados:
Indicadores objetivo: medias/totales de la variable de interés.
Requerimientos de datos:

- Microdatos de las p variables auxiliares consideradas, de la misma encuesta donde se observa la variable de interés.
- Área de interés obtenida de la misma encuesta donde se observa la variable de interés.
- Medias poblacionales de las p variables auxiliares consideradas en las áreas, $\bar{\boldsymbol{X}}_d$, $d = 1, \dots, D$.

Ventajas:

- El número de observaciones usadas para ajustar el modelo es el tamaño muestral total n, mucho mayor que el número de observaciones (igual al número de áreas) en los modelos FH. Por tanto, los parámetros del modelo se estiman con mayor eficiencia y las ganancias en eficiencia respecto a los estimadores directos tienden a ser mayores que con los modelos FH.
- El modelo de regresión considerado incorpora heterogeneidad no explicada entre las áreas.
- Es un estimador compuesto, que automáticamente toma prestada información del resto de áreas (dando mayor peso al estimador sintético de regresión) en la medida en que es necesario (cuando el tamaño muestral es pequeño). Tiende al estimador "survey regression" cuando el tamaño del área crece.
- Al contrario que en el modelo FH, no se necesita conocer ninguna varianza.
- El estimador del ECM bajo el modelo es un estimador estable del ECM bajo el diseño y es insesgado bajo el diseño cuando se promedia a lo largo de muchas áreas.
- Se pueden desagregar las estimaciones para cualquier subdominio o subárea deseada dentro de las áreas; incluso a nivel de individuo.
- Se puede estimar en áreas no muestreadas.

Desventajas:

- Los estimadores se basan en un modelo; por tanto, es necesario analizar dicho modelo (e.g. a través de los residuos).
- No tiene en cuenta el diseño muestral. Por tanto, no es insesgado bajo el diseño y es más apropiado para el muestreo aleatorio simple. Se verá afectado por diseños muestrales informativos.
- Se ve afectado de observaciones atípicas aisladas o por la falta de normalidad.
- Los microdatos suelen ser obtenidos de un censo o registro administrativo, y suelen existir problemas de confidencialidad que limiten el uso de este tipo de datos.
- El estimador del ECM bajo el modelo de Prasad-Rao es correcto bajo el modelo con normalidad, y no es insesgado bajo el diseño para el ECM bajo el diseño para un área concreta.
- Requieren un reajuste para verificar la propiedad "*benchmarking*": que la suma de los totales estimados en las áreas de una región mayor coincida con el estimador directo para dicha región.

Ejemplo 6. EBLUPs basados en el modelo con errores anidados de las incidencias de pobreza, con R.
Continuando con los ejemplos anteriores, ilustramos cómo obtener en R los EBLUPs de las incidencias de pobreza basados en un modelo con errores anidados. En un conjunto de datos predefinido en R, se dispone de los valores de las variables auxiliares fuera de la muestra para las cinco provincias con menores

tamaños muestrales. A partir de estos datos y la muestra, se pueden calcular las medias poblacionales de dichas variables para estas provincias, pero no disponemos de las verdaderas medias para las demás provincias. Por tanto, se ilustra la obtención de los EBLUPs sólo para dichas provincias, aunque el modelo se ajuste a la muestra con todas las provincias.

En primer lugar, cargamos el conjunto de datos que contiene los valores de las variables auxiliares fuera de la muestra para las provincias seleccionadas y calculamos las medias poblacionales de dichas variables en las provincias. Para ello, usamos los valores en la muestra (conjunto de datos incomedata) y los valores fuera de la muestra (Xoutsamp). Además, incluímos los códigos de las provincias en la primera columna de la matriz de medias:

```
data(Xoutsamp)

I<-length(selprov)              # Número de provincias seleccionadas
p<-dim(Xoutsamp)[2]-1           # Número de variables auxiliares

auxvar<-names(Xoutsamp)[-1]   # Nombres de var. aux. en Xoutsamp
meanXpop<-matrix(0,nr=I,nc=p) # Matriz con medias de var. aux.
Ni<-numeric(I)                # Tamaño poblacional de las provincias

for (i in 1:I){               # Bucle para las provincias seleccionadas
 d<-selprov[i]
 Xsd<-incomedata[prov==d,auxvar]        # Valores muestrales var. aux.
 Xrd<-Xoutsamp[Xoutsamp$domain==d,-1]   # Valores no muestrales
 Ni[i]<-dim(Xrd)[1]+dim(Xsd)[1]# Tamaño poblacional de la prov.
 for (k in 1:p){
  meanXpop[i,k]<-(sum(Xrd[,k])+sum(Xsd[,k]))/Ni[i]
 }
}
Xmean<-data.frame(selprov,meanXpop)
```

Ahora llamamos a la función que calcula los EBLUPs de las incidencias de pobreza para las provincias seleccionadas, basados en el modelo con errores anidados ajustado a los datos de la muestra (para todas las provincias). Guardamos las estimaciones obtenidas en un vector:

```
povinc.BHF.res<-eblupBHF(poor ~ age2+age3+age4+age5+nat1+educ1+educ3+labor1+labor2,
 dom=prov,selectdom=selprov,meanxpop=Xmean,popnsize=sizeprov[,-1])

povinc.BHF<-numeric(D)
povinc.BHF[selprov]<-povinc.BHF.res$eblup$eblup
```

Consultamos los resultados del ajuste del modelo con errores anidados y calculamos el estimador sintético de regresión basado en el modelo a nivel de individuo:

```
betaest<-povinc.BHF.res$fit$fixed  # Coeficientes de regresión
upred<-povinc.BHF.res$fit$random     # Efectos predichos de prov.
sigmae2est<-povinc.BHF.res$fit$errorvar  # Var. estimada del error
sigmau2est<-povinc.BHF.res$fit$refvar     # Varianza estimada de los efectos de las provincias

povinc.rsyn2<-numeric(D)
povinc.rsyn2[selprov]<-cbind(1,meanXpop)%*%betaest
```

Analizamos cuánto peso le da el EBLUP al estimador "*survey regression*":

```
gammad.BHF<-sigmau2est/(sigmau2est+sigmae2est/nd)
summary(gammad.BHF)
```

Resultado:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3458  0.7743  0.8606  0.8352  0.9276  0.9741
```

Cuando más cerca de cero esté el resultado de gammad.BHF para un área, más información se está tomando prestada del estimador sintético de regresión a nivel de individuo. En este caso, existe una provincia para la cual se está tomando prestada mucha información, dado que el mínimo valor de gammad.BHF es relativamente pequeño.

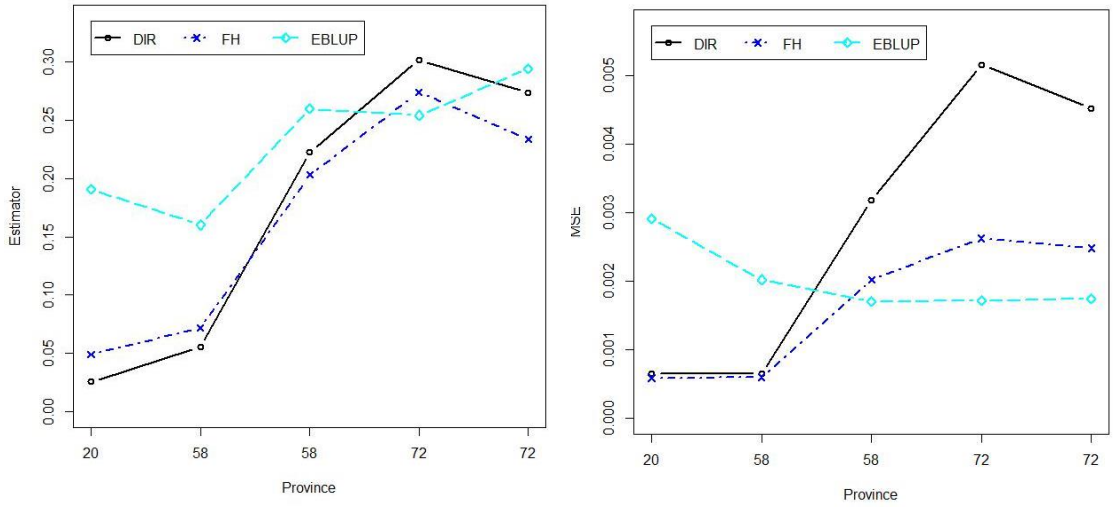
Ahora calculamos los estimadores del ECM de los EBLUPs mediante el *Bootstrap* paramétrico descrito anteriormente. Para ello, llamamos a la función pbmseBHF() usando B=200 repeticiones *Bootstrap*. Esta función también devuelve los EBLUPs y resultados del ajuste exactamente igual que la función eblupBHF().

```
povinc.mse.res<-pbmseBHF(poor~age3+age4+age5+nat1+educ1+educ3+labor1+labor2,
 dom=prov,selectdom=selprov,meanxpop=Xmean,popnsize=sizeprov[,-1],B=200)
```

Finalmente, comparamos los EBLUPs basados en el modelo con errores anidados con los estimadores directos HT y FH, graficando las estimaciones puntuales obtenidas y sus ECMs estimados para las cinco provincias seleccionadas:

```
M<-max(povinc.dir[selprov],povinc.FH[selprov],povinc.BHF[selprov])
m<-min(povinc.dir[selprov],povinc.FH[selprov],povinc.BHF[selprov])
plot(1:5,povinc.dir[selprov],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="Estimator",
 xaxt="n")
points(1:5,povinc.dir[selprov],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:5,povinc.FH[selprov],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:5,povinc.BHF[selprov],type="b",col=5,lty=5,pch=5,lwd=2)
axis(1, at=1:5, labels=nd[selprov])
legend(1,M+(M-m)/k,legend=c("DIR","FH","EBLUP"),ncol=3,col=c(1,4,5),lwd=rep(2,3),
 lty=c(1,4,5),pch=c(1,4,5))

M<-max(povinc.dir.var[selprov],povinc.FH.mse[selprov],povinc.BHF.mse[selprov])
m<-min(povinc.dir.var[selprov],povinc.FH.mse[selprov],povinc.BHF.mse[selprov])
plot(1:5,povinc.dir.cv[selprov],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="CV",
 xaxt="n")
points(1:5,povinc.dir.var[selprov],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:5,povinc.FH.mse[selprov],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:5,povinc.BHF.mse[selprov],type="b",col=5,lty=5,pch=5,lwd=2)
axis(1, at=1:5, labels=nd[selprov])
legend(1,M+(M-m)/k,legend=c("DIR","FH","EBLUP"),ncol=3,col=c(1,4,5),lwd=rep(2,3),
 lty=c(1,4,5),pch=c(1,4,5))
```

Gráfico 7
EBLUPs basados en el modelo con errores anidados de las incidencias de pobreza para las provincias junto a estimaciones directas HT y FH (izquierda), y ECMs estimados de los tres estimadores (derecha)
(En porporciones)



Fuente: Elaboración propia.

Según el Gráfico 7 (izquierda), podemos ver cómo, para las cinco provincias de menor tamaño muestral, los estimadores FH toman valores similares a los estimadores directos, pero son ligeramente más estables para las 5 provincias seleccionadas que los directos y FH. Los EBLUPs son claramente más estables para las 5 provincias seleccionadas que los directos y FH. Además, según podemos observar en el Gráfico 7 (derecha), los ECMs estimados de los EBLUPs son menores para las provincias de la derecha, debido a que toman prestada de las demás provincias una mayor cantidad de información, ya que el modelo de errores anidados se ajusta con todos los individuos de la muestra (de las D=52 provincias). Por otro lado, estos ECMs crecen de forma suave al disminuir el tamaño muestral, lo cual tiene sentido. Por el contrario, los ECMs estimados de los estimadores directos y FH toman valores excesivamente pequeños para las provincias de menor tamaño muestral (lo cual es poco creíble). En el caso de los estimadores directos, sus varianzas están estimadas con los escasos datos de cada provincia y por tanto dichas varianzas estimadas (iguales a los ECMs) no son fiables. Los BLUPs basados en el modelo FH con parámetros conocidos tienen un ECM que no puede superar a la varianza de los estimadores directos; si estas varianzas están estimadas incorrectamente, entonces el ECM del estimador FH también lo está en ese caso.

C. Método ELL

El método de Elbers, Lanjouw y Lanjouw (2003), que llamaremos método ELL, es el método tradicionalmente utilizado por el Banco Mundial para construir mapas de pobreza o desigualdad. Este método fue el primero que apareció en la literatura que permite estimar indicadores más complejos que medias o totales, mientras sean función de una variable que mide el poder adquisitivo de los individuos (habitualmente ingresos netos disponibles o gastos). Este método asume el modelo con errores anidados (32) para la transformación logaritmo de esta variable, donde los efectos aleatorios son para las unidades de primera etapa del diseño muestral (conglomerados) en lugar de para las áreas de interés. Sin embargo, para facilitar la comparabilidad con el resto de los métodos presentados en este documento y al mismo tiempo simplificar la notación, consideraremos que las unidades de primera etapa son iguales a las áreas. En este caso, si E_{di} es la variable que mide el poder adquisitivo del individuo i en el área d, tomando $Y_{di} = \log(E_{di} + c)$, donde $c > 0$ es una constante (tradicionalmente este método tomaba $c = 0$), el modelo ELL es

$$Y_{di} = \boldsymbol{x}_{di}'\boldsymbol{\beta} + u_d + e_{di}, \quad i = 1, \dots, N_d,\ d = 1, \dots, D, \tag{38}$$

donde $u_d \sim^{iid}(0, \sigma_u^2)$ y $e_{di} \sim^{ind}(0, \sigma_e^2 k_{di}^2)$, siendo u_d y e_{di} independientes, y k_{di} constantes conocidas que representan la posible heteroscedasticidad.

El estimador ELL de un parámetro general $\delta_d = \delta_d(\boldsymbol{y}_d)$ bajo el modelo (38) se obtiene mediante un procedimiento *Bootstrap*. Este procedimiento *Bootstrap* proporciona una aproximación numérica del estimador ELL teórico, que viene dado por la esperanza marginal $\hat{\delta}_d^{ELL} = E[\delta_d]$, a diferencia del predictor EB considerado en el capítulo V.B, que condiciona a la muestra $\boldsymbol{y}_s$. El mismo procedimiento *Bootstrap* se utiliza para obtener una estimación del ECM del estimador ELL.

El procedimiento *Bootstrap* funciona de la forma siguiente. En primer lugar, a partir de los residuos del modelo (38) ajustado a los datos, se generan efectos aleatorios u_d^* para cada área $d = 1, \dots, D$, y errores e_{di}^*, para cada individuo $i = 1, \dots, N_d$, $d = 1, \dots, D$. A partir de ellos, del estimador $\widehat{\boldsymbol{\beta}}$ del parámetro de regresión $\boldsymbol{\beta}$, y usando los valores de las variables auxiliares para los individuos dentro y fuera de la muestra, se generan valores *Bootstrap* de la variable respuesta para todos los individuos de la población, de la forma

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

Esto nos proporciona un censo de la variable respuesta, con el que se puede calcular cualquier tipo de indicador. Este proceso de generación se repite para $a = 1, \dots, A$, obteniendo A censos completos. Para cada censo a, calculamos el indicador de interés $\delta_d^{*(a)} = \delta_d(\boldsymbol{y}_d^{*(a)})$, donde $\boldsymbol{y}_d^{*(a)} = (Y_{d1}^{*(a)}, \dots, Y_{dN_d}^{*(a)})'$ son los valores de la variable respuesta en el área d en el censo *Bootstrap* a. Finalmente, el estimador ELL se obtiene promediando para los A censos,

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

Además, en este método, el ECM se estima de la forma

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

Para estimar el indicador FGT de orden α mediante este método, en primer lugar escribimos este indicador en función de las variables respuesta del modelo $Y_{di} = \log(E_{di} + c)$. Sustituyendo $E_{di} = \exp(Y_{di}) - c$ en la fórmula del indicador FGT dada en (1), obtenemos

$$F_{\alpha d} = \frac{1}{N_d}\sum_{i=1}^{N_d} \left(\frac{z + c - exp\,(Y_{di})}{z}\right)^{\alpha} I(\exp(Y_{di}) < z + c). \tag{39}$$

Así, calculamos este indicador con los valores Y_{di}^* generados para cada censo a, de la forma

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

y el estimador ELL de $F_{\alpha d}$ se aproxima entonces promediando estos indicadores para los A censos generados, es decir,

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

Finalmente, el ECM del estimador $\hat{F}_{\alpha d}^{ELL}$ se estima de la forma

$$\text{mse}_{ELL}(\hat{F}_{\alpha d}^{ELL}) = \frac{1}{A}\sum_{a=1}^{A}(F_{\alpha d}^{*(a)} - \hat{F}_{\alpha d}^{ELL})^2.$$

Es fácil comprobar que, para áreas de tamaño poblacional N_d grande (suele ser el caso en aplicaciones reales), si estimamos mediante este método la media del área d, $\bar{Y}_d$, al promediar $\bar{Y}_d^{*(a)} \approx \bar{\boldsymbol{X}}_d{}'\hat{\boldsymbol{\beta}} + u_d^{*(a)}$ a lo largo de los A censos, el promedio de los efectos aleatorios *Bootstrap* $u_d^{*(a)}$, a lo largo de las repeticiones *Bootstrap*, es $A^{-1}\sum_{a=1}^{A} u_d^{*(a)} \approx E(u_d) = 0$. Por tanto, el estimador ELL, $\hat{\bar{Y}}_d^{ELL} = E[\bar{Y}_d]$, resulta ser el estimador sintético de regresión,

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

Esto es debido a que la media marginal $E[\delta_d]$, sin condicionar a los datos disponibles de Y_{di} en la muestra, no utiliza dichas observaciones muestrales y por tanto se ciñe a la predicción obtenida a través del modelo, sin tener en cuenta los efectos aleatorios de las áreas ya que desaparecen. Por tanto, el estimador ELL tiene los mismos problemas que el estimador sintético de regresión; concretamente, puede ser muy sesgado si el modelo de regresión sin los efectos aleatorios no se verifica; es decir, si las variables auxiliares consideradas no explican toda la heterogeneidad de la variable respuesta a lo largo de las áreas.

Además, en el método *Bootstrap* utilizado, al contrario que en los métodos *Bootstrap* habituales, no se vuelve a ajustar el modelo y estimar con las muestras *Bootstrap* (que se deberían extraer de los censos *Bootstrap*). Por tanto, no se está replicando el proceso del mundo real en el mundo *Bootstrap*. Como consecuencia, el ECM estimado según este método no reproduce correctamente el error que se incurre en la estimación en el mundo real. Finalmente, en el método ELL original, los efectos aleatorios incluidos en el modelo son para los conglomerados o unidades de primera etapa del muestreo y no para las áreas de interés. Si se considera este modelo, pero las variables auxiliares disponibles no explican toda la heterogeneidad entre áreas, se puede infraestimar seriamente el error del estimador ELL.

Resumen de las características del estimador ELL:

Indicadores objetivo: parámetros generales.

Requerimientos de datos:

- Microdatos de las p variables auxiliares consideradas, de la misma encuesta donde se observa la variable de interés.
- Área de interés obtenida de la misma encuesta donde se observa la variable de interés.
- Microdatos de las p variables auxiliares consideradas en las áreas a partir de un censo o un registro administrativo (medidas de la misma forma que en la encuesta).

Ventajas:

- Basado en datos a nivel de individuo, que proporcionan información más detallada que los datos a nivel de área. Además, el tamaño muestral suele ser mucho mayor (n comparado con D).
- Permite estimar indicadores cualesquiera, mientras estén definidos como una función de las variables respuesta Y_{di}.
- Son insesgados bajo el modelo si los parámetros del modelo son conocidos.
- Una vez se ajusta el modelo, se puede estimar para cualquier subarea o subdominio. Incluso se puede estimar a nivel de individuo.
- Una vez se ajusta el modelo, se pueden estimar al mismo tiempo todos los indicadores (que sean función de Y_{di}) que se desee, sin necesidad de ajustar un modelo distinto para cada indicador.

Desventajas:

- Los estimadores ELL pueden presentar un alto ECM bajo el modelo, pudiendo incluso comportarse peor que los estimadores directos si la heterogeneidad entre áreas no explicada es significativa, véase Molina y Rao (2010). Para la estimación de medias, los estimadores ELL son los estimadores sintéticos de regresión, que asumen un modelo sin efectos aleatorios de las áreas.
- Están basados en un modelo. Por tanto, es necesario comprobar que el modelo se ajusta correctamente a los datos.
- No son insesgados bajo el diseño y pueden tener un sesgo considerable bajo diseño informativo.
- Pueden verse seriamente afectados por atípicos aislados.
- Si el modelo incluye efectos de los conglomerados en lugar de efectos de las áreas de interés, pero existe heterogeneidad entre las áreas, los estimadores ELL infraestiman el verdadero ECM. Incluso si los efectos del área se incluyen en el modelo, los estimadores ELL del ECM no estiman correctamente el verdadero ECM de los estimadores ELL para cada área.

D. Mejor predictor empírico bajo el modelo con errores anidados

El mejor predictor (en inglés, *best/Bayes predictor*, BP) basado en el modelo con errores anidados fue propuesto por Molina y Rao (2010) para estimar indicadores no lineales generales. Estos autores lo han utilizado para estimar la incidencia y la brecha de pobreza en las provincias españolas por género. También ha sido utilizado por el Consejo Nacional para la Evaluación de la Política de Desarrollo Social (CONEVAL) de México en estudios comparativos con otros métodos, como el ELL, para la estimación de indicadores de pobreza y desigualdad en los municipios mexicanos. Este método asume que las variables $Y_{di} = \log(E_{di} + c)$ siguen el modelo (32) con normalidad para los efectos aleatorios de las áreas u_d y para los errores e_{di}. Bajo este modelo, los vectores de variables para cada área, $\boldsymbol{y}_d = (Y_{d1}, \dots, Y_{dN_d})'$, $d = 1, \dots, D$, son independientes y verifican $\boldsymbol{y}_d \sim^{ind} N(\boldsymbol{\mu}_d, \boldsymbol{V}_d)$, con vector de medias $\boldsymbol{\mu}_d = \boldsymbol{X}_d\boldsymbol{\beta}$, siendo $\boldsymbol{X}_d = (\boldsymbol{x}_{d1}, \dots, \boldsymbol{x}_{dN_d})'$ y matriz de covarianzas $\boldsymbol{V}_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}_{N_d}{}' + \sigma_e^2 \boldsymbol{A}_d$, donde $\boldsymbol{A}_d = \text{diag}(k_{di}^2; i = 1, \dots, N_d)$. Para un indicador general definido como una función de $\boldsymbol{y}_d$, es decir, $\delta_d = \delta_d(\boldsymbol{y}_d)$, el mejor predictor es aquel que minimiza el ECM y viene dado por

$$\tilde{\delta}_d^B(\boldsymbol{\theta}) = E_{\boldsymbol{y}_{dr}}[\delta_d(\boldsymbol{y}_d)|\boldsymbol{y}_{ds}; \boldsymbol{\theta}], \tag{40}$$

donde la esperanza se toma respecto de la distribución del vector de valores fuera de la muestra $\boldsymbol{y}_{dr}$ del dominio d dados los valores en la muestra $\boldsymbol{y}_{ds}$. Esta distribución condicionada depende del verdadero valor de los parámetros del modelo para $\boldsymbol{\theta}$. Reemplazando $\boldsymbol{\theta}$ por un estimador consistente $\widehat{\boldsymbol{\theta}}$ en el mejor predictor (40), obtenemos el llamado mejor predictor empírico (en inglés, *empirical best/Bayes*, EB), $\hat{\delta}_d^{EB} = \tilde{\delta}_d^B(\widehat{\boldsymbol{\theta}})$. De nuevo, los métodos habituales de estimación, que proporcionan estimadores consistentes incluso si no existe normalidad, son ML y REML, ambos bajo la verosimilitud normal, y el método de Henderson III.

Bajo el modelo de errores anidados (32), la distribución de $\boldsymbol{y}_{dr}|\boldsymbol{y}_{ds}$, necesaria para calcular el mejor predictor (40), se obtiene de la forma siguiente. En primer lugar, descomponemos las matrices $\boldsymbol{X}_d$ y $\boldsymbol{V}_d$ en la parte muestral y fuera de la muestra de forma similar a como hemos descompuesto $\boldsymbol{y}_d$, es decir,

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

Dado que $\boldsymbol{y}_d$ sigue una distribución normal, entonces las condicionadas también tienen distribución normal, es decir,

$$\boldsymbol{y}_{dr}|\boldsymbol{y}_{ds} \overset{ind}{\sim} N(\boldsymbol{\mu}_{dr|s}, \boldsymbol{V}_{dr|s}), \quad d = 1, \dots, D, \tag{41}$$

donde el vector de medias condicionadas y la correspondiente matriz de covarianzas toman la forma

$$\boldsymbol{\mu}_{dr|s} = \boldsymbol{X}_{dr}\boldsymbol{\beta} + \gamma_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}^T\boldsymbol{\beta})\mathbf{1}_{N_d-n_d}, \tag{42}$$

$$\boldsymbol{V}_{dr|s} = \sigma_u^2(1-\gamma_d)\mathbf{1}_{N_d-n_d}\mathbf{1}_{N_d-n_d}^T + \sigma_e^2\,\mathrm{diag}_{i\in r_d}(k_{di}^2), \tag{43}$$

siendo $\mathbf{1}_k$ un vector de unos de tamaño k. Concretamente, para el individuo $i \in r_d$, se tiene

$$Y_{di}|\boldsymbol{y}_{ds} \sim N(\mu_{di|s}, \sigma_{di|s}^2), \tag{44}$$

donde la media y la varianza condicionadas vienen dadas por

$$\mu_{di|s} = \boldsymbol{x}_{di}'\boldsymbol{\beta} + \gamma_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}^T\boldsymbol{\beta}), \tag{45}$$

$$\sigma_{di|s}^2 = \sigma_u^2(1-\gamma_d) + \sigma_e^2 k_{di}^2. \tag{46}$$

Si ahora deseamos estimar el indicador de pobreza FGT de orden α, $\delta_d = F_{\alpha d}$, en primer lugar asumimos que $Y_{di} = \log(E_{di} + c)$, para $c > 0$, verifica el modelo con errores anidados. Reescribimos el indicador FGT en cuestión como una función de las variables respuesta en el modelo Y_{di}, es decir, como en (39), y calculamos la esperanza que define el mejor predictor $\tilde{F}_{\alpha d}^B = E_{\boldsymbol{y}_{dr}}[F_{\alpha d}|\boldsymbol{y}_{ds};\boldsymbol{\theta}]$. Para ello, separamos la suma que define el indicador FGT dado en (1) en la parte muestral y la parte fuera de la muestra e, introduciendo la esperanza dentro de la suma, obtenemos

$$\tilde{F}_{\alpha d}^B(\boldsymbol{\theta}) = \frac{1}{N_d}\left(\sum_{i\in s_d} F_{\alpha,di} + \sum_{i\in r_d} \tilde{F}_{\alpha,di}^B(\boldsymbol{\theta})\right), \tag{47}$$

donde $\tilde{F}_{\alpha,di}^B(\boldsymbol{\theta}) = E[F_{\alpha,di}|\boldsymbol{y}_{ds};\boldsymbol{\theta}]$ y la esperanza se toma respecto de la distribución de $Y_{di}|\boldsymbol{y}_{ds}$, $i \in r_d$, dada en (44)–(46). Para $\alpha = 0{,}1$, las esperanzas son fáciles de calcular, y vienen dadas respectivamente por

$$\tilde{F}_{0,di}^B(\boldsymbol{\theta}) = \Phi(\alpha_{di}), \tag{48}$$

$$\tilde{F}_{1,di}^B(\boldsymbol{\theta}) = \Phi(\alpha_{di})\left\{1 - \frac{1}{z}\left[\exp\left(\mu_{di|s} + \frac{\sigma_{di|s}^2}{2}\right)\frac{\Phi(\alpha_{di} - \sigma_{di|s})}{\Phi(\alpha_{di})} - c\right]\right\}, \tag{49}$$

donde $\Phi(\cdot)$ es la función de distribución de una variable aleatoria Normal estándar y $\alpha_{di} = [\log(z + c) - \mu_{di|s}]/\sigma_{di|s}$, con $\mu_{di|s}$ y $\sigma_{di|s}^2$ dados en (45)–(46).

Para indicadores $\delta_d = \delta_d(\boldsymbol{y}_d)$ más complejos; por ejemplo, los indicadores FGT para $\alpha \neq 0{,}1$, a menudo la esperanza que define el mejor predictor no se puede calcular de forma analítica. En estos casos, el mejor predictor se puede aproximar de forma empírica usando simulación Monte Carlo. El proceso sería el siguiente:

1. Obtener un estimador $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}', \hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ del vector de parámetros $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)'$ ajustando del modelo (32) a los datos $(\boldsymbol{y}_s, \boldsymbol{X}_s)$.
2. Generar, para $a = 1, \ldots, A$, vectores de variables respuesta para los individuos fuera de la muestra del área d, $\boldsymbol{y}_{dr}^{(a)}$, a partir de la distribución de $\boldsymbol{y}_{dr}|\boldsymbol{y}_{ds}$ dada en (41)–(43), con $\boldsymbol{\theta}$ reemplazado por su estimador $\widehat{\boldsymbol{\theta}}$ obtenido en (a).
3. Aumentar el vector generado $\boldsymbol{y}_{dr}^{(a)}$ con los datos de la muestra $\boldsymbol{y}_{ds}$ para formar un vector censal para el área d, $\boldsymbol{y}_d^{(a)} = (\boldsymbol{y}_{ds}', (\boldsymbol{y}_{dr}^{(a)})')'$. Usando $\boldsymbol{y}_d^{(a)}$, calcular el indicador de interés $\delta_d^{(a)} = \delta_d(\boldsymbol{y}_d^{(a)})$ y repetir para $a = 1, \ldots, A$. La aproximación Monte Carlo del predictor EB del indicador δ_d se obtiene promediando los indicadores para los A censos simulados, es decir,

$$\hat{\delta}_d^{EB} = \frac{1}{A}\sum_{a=1}^{A} \delta_d^{(a)}. \tag{50}$$

En el paso 2, es necesario simular A veces un vector $\boldsymbol{y}_{dr}^{(a)}$ con distribución Normal multivariante de tamaño $N_d - n_d$, que puede ser realmente grande (e.g. del tamaño de una provincia), lo cual puede ser computacionalmente muy costoso o incluso imposible debido al gran tamaño del vector multivariante a generar. Esto se puede evitar observando que la matriz de covarianzas de este vector, $\boldsymbol{V}_{dr|s}$, dada en (43), corresponde a la matriz de covarianzas de un vector aleatorio $\boldsymbol{y}_{dr}^{(a)}$ generado a partir del modelo

$$\boldsymbol{y}_{dr}^{(a)} = \boldsymbol{\mu}_{dr|s} + v_d^{(a)}\mathbf{1}_{N_d-n_d} + \boldsymbol{\epsilon}_{dr}^{(a)}, \tag{51}$$

donde $v_d^{(a)}$ y $\boldsymbol{\epsilon}_{dr}^{(a)}$ son independientes, y verifican respectivamente

$$v_d^{(a)} \sim N(0, \sigma_u^2(1-\gamma_d)), \quad \boldsymbol{\epsilon}_{dr}^{(a)} \sim N(\mathbf{0}_{N_d-n_d}, \sigma_e^2 \mathrm{diag}_{i\in r_d}(k_{di}^2)); \tag{52}$$

véase Molina y Rao (2010). Usando el modelo (51)–(52), en lugar de generar un vector $\boldsymbol{y}_{dr}^{(a)}$ Normal multivariante de tamaño $N_d - n_d$, tan solo es necesario generar las $1 + N_d - n_d$ variables Normales independientes $v_d^{(a)} \sim^{ind} N(0, \sigma_u^2(1-\gamma_d))$ y $\epsilon_{di}^{(a)} \sim^{ind} N(0, \sigma_e^2 k_{di}^2)$, para $i \in r_d$. Usando el vector $\boldsymbol{y}_{dr}^{(a)}$ generado a partir del modelo (51), en el paso (c) construimos el vector censal $\boldsymbol{y}_d^{(a)} = (\boldsymbol{y}_{ds}', (\boldsymbol{y}_{dr}^{(a)})')'$ y calculamos el indicador de interés $\delta_d^{(a)} = \delta_d(\boldsymbol{y}_d^{(a)})$.

Para un área d no muestreada (i.e., con $n_d = 0$), generamos $\boldsymbol{y}_{dr}^{(a)}$ del modelo (51) tomando $\gamma_d = 0$ y, al no existir parte muestral en este caso, el vector censal del área d es igual al vector generado $\boldsymbol{y}_d^{(a)} = \boldsymbol{y}_{dr}^{(a)}$.

En el caso de indicadores complejos, calcular aproximaciones analíticas para el ECM de los correspondientes predictores EB es complicado. Molina y Rao (2010) describen un método *Bootstrap* paramétrico para estimar el ECM basado en el método *Bootstrap* para poblaciones finitas de González-Manteiga et al. (2008). Este método consiste en realizar los siguientes pasos:

1. Ajustar el modelo (32) a los datos de la muestra $\boldsymbol{y}_s = (\boldsymbol{y}_{1s}', \dots, \boldsymbol{y}_{Ds}')'$, obteniendo estimaciones de los parámetros del modelo, $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ y $\hat{\sigma}_e^2$.
2. Generar efectos *Bootstrap* de las áreas de la forma
$$u_d^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2), \quad d = 1, \dots, D.$$
3. Generar, independientemente de $u_1^{*(b)}, \dots, u_D^{*(b)}$, errores *Bootstrap*
$$e_{di}^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2), \quad i = 1, \dots, N_d, d = 1, \dots, D$$
4. Generar una población (o censo) *Bootstrap* de valores de la variable respuesta a través del modelo,
$$Y_{di}^{*(b)} = \boldsymbol{x}_{di}'\hat{\boldsymbol{\beta}} + u_d^{*(b)} + e_{di}^{*(b)}, \quad i = 1, \dots, N_d, d = 1, \dots, D.$$
5. Definimos el vector censal de variables respuesta del área d, dado por $\boldsymbol{y}_d^{*(b)} = (Y_{d1}^{*(b)}, \dots, Y_{dN_d}^{*(b)})'$. Calcular los indicadores de interés a partir del censo *Bootstrap* $\delta_d^{*(b)} = \delta_d(\boldsymbol{y}_d^{*(b)})$, $d = 1, \dots, D$.
6. Para la muestra original $s = s_1 \cup \cdots \cup s_D$, sea $\boldsymbol{y}_s^{*(b)} = ((\boldsymbol{y}_{1s}^{*(b)})', \dots, (\boldsymbol{y}_{Ds}^{*(b)})')'$ el vector que contiene las observaciones *Bootstrap* cuyos índices están en la muestra, es decir, que contiene a las variables $Y_{di}^{*(b)}$, $i \in s_d$, $d = 1, \dots, D$. Ajustar de nuevo el modelo (32) a los datos *Bootstrap* $\boldsymbol{y}_s^{*(b)}$ y obtener los predictores EB *Bootstrap* de los indicadores de interés, $\hat{\delta}_d^{EB*(b)}$, $d = 1, \dots, D$.
7. Repetir los pasos 2)–6) para $b = 1, \dots, B$, y obtenemos los verdaderos valores, $\delta_d^{*(b)}$, y los correspondientes predictores EB, $\hat{\delta}_d^{EB*(b)}$, para cada área $d = 1, \dots, D$, y para cada réplica *Bootstrap*, $b = 1, \dots, B$.

8. Los estimadores "naive *Bootstrap*" del ECM de los predictores EB, $\hat{\delta}_d^{EB}$, vienen dados por

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

Obsérvese que, para estimar indicadores complejos, tanto el método ELL descrito en el capítulo anterior como el EB presentado en este capítulo, requieren datos de una encuesta con las observaciones de la variable de interés y de las variables auxiliares para todas las áreas, $\{(y_{di}, \boldsymbol{x}_{di}); i \in s_d, d = 1, \dots, D\}$, así como un censo con los valores de las mismas variables auxiliares para todas las unidades de la población, $\{\boldsymbol{x}_{di}; i = 1, \dots, N_d, d = 1, \dots, D\}$. En principio, el método EB requiere adicionalmente identificar en el censo las unidades que además están en la muestra dentro de cada área s_d. Vincular los datos de la encuesta y el censo no siempre es posible en la práctica. Sin embargo, el tamaño muestral del área, n_d, es típicamente muy pequeño comparado con el tamaño poblacional del área, N_d. Entonces podemos usar el predictor "*Census best*" propuesto por Correa, Molina y Rao (2012), que se obtiene calculando las esperanzas condicionadas $\tilde{F}_{\alpha,di}^{B}(\boldsymbol{\theta})$, también para los individuos de la muestra como si no se observaran, es decir, el predictor *Census best* de $F_{\alpha d}$ viene dado por

$$\tilde{F}_{\alpha d}^{CB}(\boldsymbol{\theta}) = \frac{1}{N_d}\sum_{i=1}^{N_d}\tilde{F}_{\alpha,di}^{B}(\boldsymbol{\theta}). \tag{53}$$

De la misma forma que el predictor EB, definimos el predictor *Census EB* de $F_{\alpha d}$, reemplazando en (53) un estimador consistente de $\boldsymbol{\theta}$. Si la esperanza que define $\tilde{F}_{\alpha,di}^{B}(\boldsymbol{\theta})$ no se puede calcular de forma analítica, como ocurre cuando el indicador tiene una forma complicada, en cada repetición del procedimiento Monte Carlo descrito en (1)–(3), se genera el vector censal completo $\boldsymbol{y}_d$ en lugar de solamente el vector de observaciones fuera de la muestra $\boldsymbol{y}_{dr}$; es decir, aplicamos la aproximación Monte Carlo (50) generando $\boldsymbol{y}_d^{(a)} = \boldsymbol{\mu}_{d|s} + v_d^{(a)}\mathbf{1}_{N_d-n_d} + \boldsymbol{\epsilon}_d^{(a)}$, donde $\boldsymbol{\mu}_{d|s} = \boldsymbol{X}_d\boldsymbol{\beta} + \gamma_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}^T\boldsymbol{\beta})\mathbf{1}_{N_d}$ y $\boldsymbol{\epsilon}_d^{(a)} \sim N(\mathbf{0}_{N_d}, \sigma_e^2\,\text{diag}_{i=1,\dots,N_d}(k_{di}^2))$. Si la fracción muestral n_d/N_d es despreciable, como suele ser en la mayoría de casos en la realidad, el estimador *Census EB* de $\delta_d = F_{\alpha d}$ será prácticamente igual al estimador EB original.

Para indicadores cuyo cálculo tiene un alto coste computacional, como por ejemplo aquellos que requieren ordenar los individuos de la población en función de su poder adquisitivo como los indicadores Fuzzy monetarios y Fuzzy suplementarios, el tiempo computacional del procedimiento total, incluyendo el método *Bootstrap* para el cálculo del ECM, se dispara. En este caso, Ferretti y Molina (2012) propusieron una variación del predictor EB, llamada *fast EB*, mucho más rápida computacionalmente. En el procedimiento Monte Carlo 1-3 para la aproximación del predictor EB, este procedimiento reemplaza la generación del censo en el paso 2 por la generación de una muestra (distinta en cada repetición Monte Carlo) y el cálculo de los verdaderos valores de los indicadores en el paso 3 por el cálculo de estimadores basados en el diseño, que solo necesitan una muestra en lugar del censo completo.

Propiedades del predictor EB (aproximadas para el *Census EB* si n_d/N_d es despreciable):

Indicadores objetivo: parámetros generales.

Requerimientos de datos:

- Microdatos de las p variables auxiliares consideradas, de la misma encuesta donde se observa la variable de interés.
- Área de interés obtenida de la misma encuesta donde se observa la variable de interés.
- Microdatos de las p variables auxiliares consideradas a partir de un censo o un registro administrativo (medidas de la misma forma que en la encuesta).

Ventajas:

- Basado en datos a nivel de individuo, que proporcionan información más detallada que los datos a nivel de área (también es posible incorporar variables a nivel de área). Además, el tamaño muestral suele ser mucho mayor (n comparado con D).
- Permite estimar indicadores cualesquiera, mientras estén definidos como una función de las variables respuesta Y_{di}.
- Son insesgados bajo el modelo si los parámetros del modelo son conocidos.
- Son óptimos en el sentido de minimizar el ECM bajo el modelo, para valores conocidos de los parámetros.
- Se comportan sustancialmente mejor que los estimadores ELL en términos de ECM bajo el modelo (32) cuando la heterogeneidad no explicada entre áreas es significativa. Para áreas no muestreadas (con $n_d = 0$), los estimadores EB y ELL son prácticamente iguales. También serán prácticamente iguales, en este caso para todas las áreas, si toda la heterogeneidad entre áreas está explicada por las variables auxiliares ($\sigma_u^2 = 0$).
- Una vez se ajusta el modelo, se puede estimar para cualquier subarea o subdominio. Incluso se puede estimar a nivel de individuo.
- Una vez se ajusta el modelo, se pueden estimar al mismo tiempo todos los indicadores (que sean función de Y_{di}) que se desee, sin necesidad de ajustar un modelo distinto para cada indicador.

Desventajas:

- Están basados en un modelo. Por tanto, es necesario comprobar que el modelo se ajusta correctamente (e.g., a través de los residuos).
- No tienen en cuenta el diseño muestral. No son insesgados bajo el diseño y pueden tener un sesgo considerable bajo diseño informativo.
- Pueden verse seriamente afectados por atípicos aislados o por la falta de normalidad.
- Los estimadores del ECM obtenidos mediante el método *Bootstrap* paramétrico son computacionalmente intensivos.

Ejemplo 7. Estimadores EB de las incidencias de pobreza, con R. Continuando con los ejemplos anteriores, ilustramos cómo obtener en R los estimadores EB de las incidencias de pobreza, basados en un modelo con errores anidados para el logaritmo de los ingresos (trasladados con una constante). El umbral de pobreza se ha calculado de antemano como el 60% de la mediana de los ingresos, y resulta ser $z = 6557.143$. Usando este umbral, necesitamos definir la función que nos da la incidencia de pobreza:

```
povertyincidence <- function(y) {
 result <- mean(y < 6557.143)
 return (result)
}
```

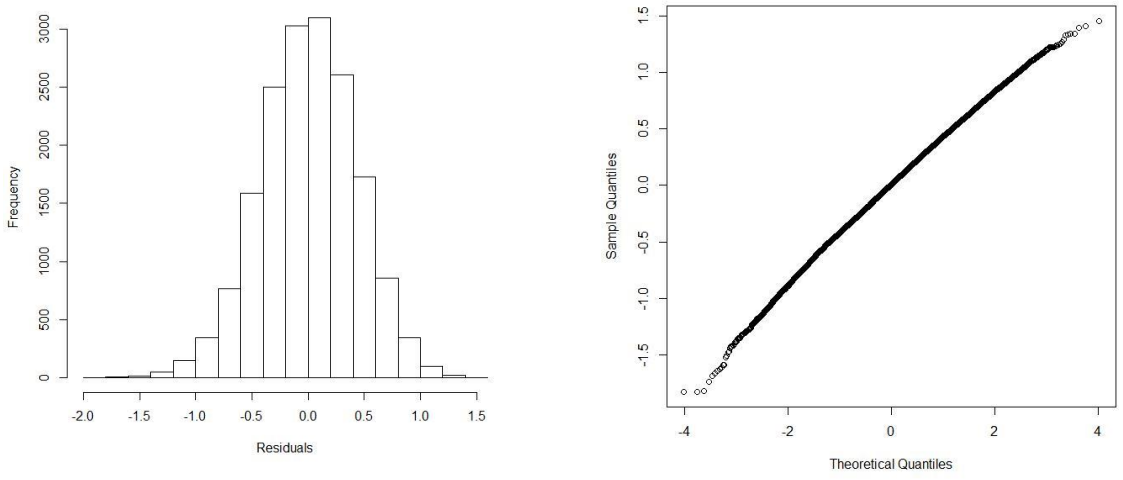
Ahora llamamos a la función que calcula los estimadores EB seleccionando como indicador dicha función `povertyincidence`, tomando transformación logaritmo (por defecto) y añadiendo la constante `constant=3500` a los ingresos antes de dicha transformación, y utilizando repeticiones para la aproximación de Monte Carlo de los estimadores EB. La constante mencionada se selecciona de manera que los residuos del ajuste muestren una distribución aproximadamente simétrica, ya que el método EB descrito se basa en la distribución normal. Antes de llamar a la función, fijamos la semilla de los generadores de números aleatorios para que la función nos proporcione las mismas estimaciones en el caso de repetir la llamada a esta función, e inicializamos el vector que contendrá a los estimadores EB. :

```
povinc.EB<-numeric(D)

set.seed(123)      # Fijamos la semilla para números aleatorios
res.EB<-ebBHF(income~age2+age3+age4+age5+nat1+educ1+educ3+labor1+labor2,dom=prov,
selectdom=selprov,Xnonsample=Xoutsamp,MC=50,constant=3500,indicator=povertyincidence)
povinc.EB[selprov]<-res.EB$eb$eb
```

Para cualquier modelo, conviene analizar los residuos para comprobar que los datos no presenten evidencias claras en contra del modelo asumido. Dado que el método EB requiere normalidad, representamos un histograma y un gráfico q-q de normalidad de los residuos:

```
resid.EB<-res.EB$fit$residuals
hist(resid.EB,main="",xlab="Residuals")
qqnorm(resid.EB,main="")
```

Gráfico 8
Histograma (izquierda) y q-q plot de normalidad (derecha) de los residuos del ajuste del modelo con errores anidados al logaritmo de los ingresos
(En unidades)



Fuente: Elaboración propia.

Ambos gráficos (gráfico 8) muestran que la distribución de los residuos es aproximadamente normal. Por el contrario, si ajustamos el modelo a los ingresos sin la transformación logaritmo, tanto el histograma como el gráfico q-q de normalidad (no se incluyen por brevedad) muestran una distribución marcadamente asimétrica a la derecha. Por tanto, dicha transformación es necesaria para no alejarnos de la hipótesis de normalidad.

Finalmente, calculamos los estimadores *Bootstrap* del ECM de los estimadores EB con B=200 iteraciones *Bootstrap* y MC=50 iteraciones para la aproximación Monte Carlo de los estimadores EB.

```
set.seed(123)
povinc.mse.res<-
pbmseebBHF(income~age2+age3+age4+age5+nat1+educ1+educ3+labor1+labor2,dom=prov,select
dom=selprov,Xnonsample=Xoutsamp,B=200,MC=50,constant=3500,
```

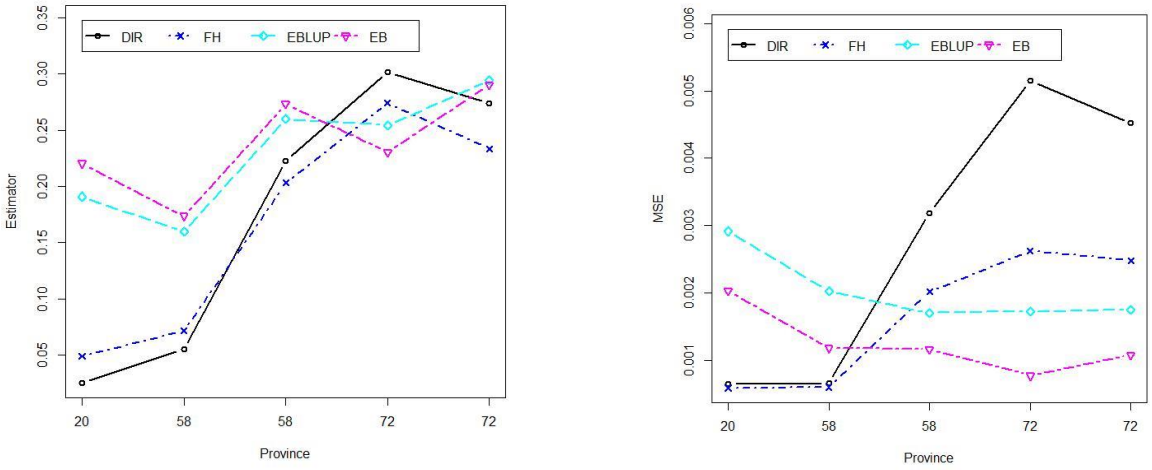
```
 indicator=povertyincidence)

povinc.eb.mse<-numeric(D)
povinc.eb.mse[selprov]<-povinc.mse.res$mse$mse
```

Finalmente, comparamos gráficamente los estimadores EB con los directos HT, FH y EBLUPs basados en el modelo con errores anidados de las incidencias de pobreza para las provincias seleccionadas:

```
k<-6
M<-max(povinc.dir[selprov],povinc.FH[selprov],povinc.BHF[selprov],povinc.EB[selprov])
m<-min(povinc.dir[selprov],povinc.FH[selprov],povinc.BHF[selprov],povinc.EB[selprov])
plot(1:5,povinc.dir[selprov],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="Estimator",
 xaxt="n")
points(1:5,povinc.dir[selprov],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:5,povinc.FH[selprov],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:5,povinc.BHF[selprov],type="b",col=5,lty=5,pch=5,lwd=2)
points(1:5,povinc.EB[selprov],type="b",col=6,lty=6,pch=6,lwd=2)
axis(1, at=1:5, labels=nd[selprov])
legend(1,M+(M-m)/k,legend=c("DIR","FH","EBLUP","EB"),ncol=4,col=c(1,4,5,6),lwd=rep(2,4),
 lty=c(1,4,5,6),pch=c(1,4,5,6))

M<-max(povinc.dir.var[selprov],povinc.FH.mse[selprov],povinc.BHF.mse[selprov],
 povinc.eb.mse[selprov])
m<-min(povinc.dir.var[selprov],povinc.FH.mse[selprov],povinc.BHF.mse[selprov],
 povinc.eb.mse[selprov])
plot(1:5,povinc.dir.var[selprov],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="CV",
 xaxt="n")
points(1:5,povinc.dir.var[selprov],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:5,povinc.FH.mse[selprov],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:5,povinc.BHF.mse[selprov],type="b",col=5,lty=5,pch=5,lwd=2)
points(1:5,povinc.eb.mse[selprov],type="b",col=6,lty=6,pch=6,lwd=2)
axis(1, at=1:5, labels=nd[selprov])
legend(1,M+(M-m)/k,legend=c("DIR","FH","EBLUP","EB"),ncol=4,col=c(1,4,5,6),lwd=rep(2,4),
 lty=c(1,4,5,6),pch=c(1,4,5,6))
```

Según el Gráfico 9 (izquierda), los estimadores EB se parecen en gran medida a los EBLUPs. Esto es razonable ya que ambos se basan en un modelo a nivel de individuo, aunque los estimadores EB ajustan el modelo para el logaritmo de los ingresos, mientras que los EBLUP ajustan el modelo al indicador binario de tener o no ingresos por debajo del umbral (variable poor). Teóricamente, el modelo asumido por los EBLUPs no es cierto, puesto que la variable respuesta es binaria y los predictores pueden proporcionar valores fuera del intervalo. Además, a pesar de la similitud de las estimaciones EB y EBLUP, el Gráfico 9 (derecha) indica que los estimadores EB son más eficientes que los EBLUPs.

Gráfico 9
Estimaciones EB, EBLUP de la inicidencia de pobreza basados en el modelo por errores anidados, FH y directas HT (izquierda), y ECMs de dichos estimadores (derecha) para las provincias seleccionadas
(En proporciones)



Fuente: Elaboración propia.

E. Método jerárquico Bayes bajo el modelo con errores anidados

El cálculo de los estimadores EB (o *Census EB*) junto con sus ECMs estimados requiere computación intensiva y puede ser inviable para poblaciones muy grandes o para indicadores muy complejos (por ejemplo, los que requieren ordenaciones). Obsérvese que para obtener la aproximación Monte Carlo del estimador EB, es necesario construir A censos $\boldsymbol{y}^{(a)}$, $a = 1, \dots, A$, que pueden ser de gran tamaño. Además, para estimar el ECM mediante *Bootstrap*, es necesario repetir la aproximación Monte Carlo para cada réplica *Bootstrap*. Con el objeto de desarrollar un método computacionalmente más eficiente, Molina, Nandram y Rao (2014) propusieron el método jerárquico Bayes (en inglés, *hierarchical Bayes*, HB) para la estimación de indicadores generales. Este procedimiento no requiere el uso de métodos *Bootstrap* para la estimación del ECM ya que proporciona muestras de la distribución posterior del indicador de interés, a partir de las cuales se pueden obtener fácilmente varianzas posteriores que juegan el papel de ECM, o cualquier otra medida resumen.

El método HB se basa en reparametrizar el modelo con errores anidados (32) en términos del coeficiente de correlación intraclase $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$ y considerando distribuciones previas para los parámetros del modelo $(\boldsymbol{\beta}, \rho, \sigma_e^2)$ que reflejan la falta de información previa sobre ellos. Concretamente, consideramos el siguiente modelo HB:

$$\text{(i) } Y_{di}|u_d, \boldsymbol{\beta}, \sigma_e^2 \overset{ind}{\sim} N\left(\boldsymbol{x}_{di}'\boldsymbol{\beta} + u_d, \sigma_e^2 k_{di}^2\right), \quad i = 1, \dots, N_d,$$
$$\text{(ii) } u_d|\rho, \sigma_e^2 \overset{iid}{\sim} N\left(0, \frac{\rho}{1-\rho}\sigma_e^2\right), \quad d = 1, \dots, D,$$
$$\text{(iii) } \pi(\boldsymbol{\beta}, \rho, \sigma_e^2) \propto \frac{1}{\sigma_e^2}, \quad \epsilon \leq \rho \leq 1 - \epsilon,\ \sigma_e^2 > 0, \boldsymbol{\beta} \in R^p,$$

donde $\epsilon > 0$ se selecciona muy pequeño para reflejar falta de información previa. Véase la aplicación realizada por Molina, Nandram y Rao (2014), donde la inferencia no es sensitiva a pequeños cambios de ϵ.

La distribución posterior de los parámetros del modelo se puede calcular en función de las distribuciones condicionadas usando las reglas de la cadena de la forma siguiente. En primer lugar, obsérvese que, bajo el método HB, los efectos aleatorios $\boldsymbol{u} = (u_1, \dots, u_D)'$ se consideran parámetros adicionales. Entonces, la densidad conjunta del vector de parámetros $\boldsymbol{\theta} = (\boldsymbol{u}', \boldsymbol{\beta}', \sigma_e^2, \rho)'$ dadas las observaciones de la muestra $\boldsymbol{y}_s$ viene dada por

$$\pi(\boldsymbol{u}, \boldsymbol{\beta}, \sigma_e^2, \rho | \boldsymbol{y}_s) = \pi_1(\boldsymbol{u}|\boldsymbol{\beta}, \sigma_e^2, \rho, \boldsymbol{y}_s)\pi_2(\boldsymbol{\beta}|\sigma_e^2, \rho, \boldsymbol{y}_s)\pi_3(\sigma_e^2|\rho, \boldsymbol{y}_s)\pi_4(\rho|\boldsymbol{y}_s), \tag{54}$$

donde todas las densidades condicionadas excepto π_4 tienen formas conocidas. Como ρ está definido en un intervalo cerrado dentro de (0,1), podemos generar valores de π_4 usando un método de rejilla, para más detalles véase Molina, Nandram y Rao (2014). Así, se pueden generar muestras de $\boldsymbol{\theta} = (\boldsymbol{u}', \boldsymbol{\beta}', \sigma_e^2, \rho)'$ directamente de la distribución posterior dada en (54), sin necesidad de utilizar métodos Monte Carlo de cadenas de Markov (en inglés, *Markov Chain Monte Carlo*, MCMC). Bajo condiciones generales, se puede garantizar una distribución posterior propia.

Dado $\boldsymbol{\theta}$, bajo el modelo HB (i)–(iii), las variables Y_{di} para todos los individuos de la población son independientes y verifican

$$Y_{di}|\boldsymbol{\theta} \overset{ind}{\sim} N(\boldsymbol{x}_{di}'\boldsymbol{\beta} + u_d, \sigma_e^2 k_{di}^2), \quad i = 1, \dots, N_d,\ d = 1, \dots, D. \tag{55}$$

La densidad predictiva de $\boldsymbol{y}_{dr}$ viene dada por

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

donde $\pi(\boldsymbol{\theta}|\boldsymbol{y}_s)$ viene dada en (54). Finalmente, el estimador HB del parámetro $\delta_d = \delta_d(\boldsymbol{y}_d)$ es

$$\hat{\delta}_d^{HB} = E_{\boldsymbol{y}_{dr}}(\delta_d|\boldsymbol{y}_s) = \int \delta_d(\boldsymbol{y}_d) f(\boldsymbol{y}_{dr}|\boldsymbol{y}_s) d\boldsymbol{y}_{dr}. \tag{56}$$

Este estimador se puede aproximar mediante simulación Monte Carlo. Para ello, generamos muestras de la distribución posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y}_s)$ de la forma siguiente. En primer lugar, generamos un valor $\rho^{(a)}$ de $\pi_4(\rho|\boldsymbol{y}_s)$ usando un método de rejilla (véase Molina, Nandram y Rao, 2014); después, generamos $\sigma_e^{2(a)}$ de $\pi_3(\sigma_e^2|\rho^{(a)}, \boldsymbol{y}_s)$; a continuación, $\boldsymbol{\beta}^{(a)}$ se genera de $\pi_2(\boldsymbol{\beta}|\sigma_e^{2(a)}, \rho^{(a)}, \boldsymbol{y}_s)$ y, finalmente, se genera $\boldsymbol{u}^{(a)}$ de $\pi_1(\boldsymbol{u}|\boldsymbol{\beta}^{(a)}, \sigma_e^{2(a)}, \rho^{(a)}, \boldsymbol{y}_s)$. Este proceso se repite un número grande A de veces, para obtener una muestra aleatoria $\boldsymbol{\theta}^{(a)}$, $a = 1, \dots, A$, de $\pi(\boldsymbol{\theta}|\boldsymbol{y}_s)$. Por cada valor generado $\boldsymbol{\theta}^{(a)}$ de $\pi(\boldsymbol{\theta}|\boldsymbol{y}_s)$, generamos los valores fuera de la muestra $\{Y_{di}^{(a)}, i \in r_d\}$ de la distribución dada en (55) obteniendo, para cada área d, el vector de variables fuera de la muestra $\boldsymbol{y}_{dr}^{(a)}$. Uniéndolo al vector de datos en la muestra $\boldsymbol{y}_{ds}$, construimos el vector censal $\boldsymbol{y}_d^{(a)} = (\boldsymbol{y}_{ds}', (\boldsymbol{y}_{dr}^{(a)})')'$. Usando ahora $\boldsymbol{y}_d^{(a)}$, calculamos el indicador en cuestión $\delta_d^{(a)} = \delta_d(\boldsymbol{y}_d^{(a)})$, y repetimos para $a = 1, \dots, A$. Finalmente, el estimador HB de δ_d es la media posterior, que se aproxima de la forma

$$\hat{\delta}_d^{HB} = E_{\boldsymbol{y}_{dr}}(\delta_d|\boldsymbol{y}_s) \approx \frac{1}{A}\sum_{a=1}^{A} \delta_d^{(a)}. \tag{57}$$

Para áreas no muestreadas ($n_d = 0$), ya que no existen observaciones muestrales, se tiene $\boldsymbol{y}_{dr}^{(a)} = \boldsymbol{y}_d^{(a)}$, así que se genera el vector censal completo $\boldsymbol{y}_d^{(a)} = (Y_{d1}^{(a)}, \dots, Y_{dN_d}^{(a)})'$ de la distribución (55).

Como medida de error de estimación del estimador HB, $\hat{\delta}_d^{HB}$, se proporciona la varianza posterior aproximada de forma similar,

$$V(\delta_d|\boldsymbol{y}_s) \approx \frac{1}{A}\sum_{a=1}^{A} \left(\delta_d^{(a)} - \hat{\delta}_d^{HB}\right)^2. \tag{58}$$

Para el caso particular del indicador FGT de orden α, $\delta_d = F_{\alpha d}$, en la repetición Monte Carlo a, calculamos $F_{\alpha d}^{(a)}$ usando $\boldsymbol{y}_d^{(a)}$ aplicando (39) y el estimador HB es

$$\hat{F}_{\alpha d}^{HB} \approx \frac{1}{A}\sum_{a=1}^{A} F_{\alpha d}^{(a)}. \tag{59}$$

Al igual que los métodos ELL y EB, si se desea estimar algún indicador no lineal, este método requiere tener disponible, además de los datos de la encuesta de donde se extraen los valores de la variable de interés, un censo o registro administrativo del cual obtener los microdatos de las variables auxiliares. Si no es posible identificar los individuos de la encuesta en el censo o registro, es posible calcular un estimador *Census HB* de forma similar al *Census EB*. En este estimador, aun disponiendo de valores en la muestra $\boldsymbol{y}_{ds}$, éstos se ignorarían y se generaría el vector censal completo $\boldsymbol{y}_d^{(a)}$, generando cada valor $Y_{di}^{(a)}$ de (55) y el procedimiento sería el mismo que si el área no estuviera muestreada.

Resumen del estimador HB basado en el modelo con errores anidados:

Indicadores objetivo: parámetros generales.

Requerimientos de datos:

- Microdatos de las p variables auxiliares consideradas, de la misma encuesta donde se observa la variable de interés.
- Área de interés obtenida de la misma encuesta donde se observa la variable de interés.
- Microdatos de las p variables auxiliares consideradas a partir de un censo o un registro administrativo (medidas de la misma forma que en la encuesta).

Ventajas:

- Basado en datos a nivel de individuo, que proporcionan información más detallada que los datos a nivel de área (también es posible incorporar variables a nivel de área). Además, el tamaño muestral suele ser mucho mayor (n comparado con D).
- Permite estimar indicadores cualesquiera, mientras estén definidos como una función de las variables respuesta Y_{di}.
- Son insesgados bajo el modelo si los parámetros del modelo son conocidos.
- Son óptimos en el sentido de minimizar la varianza posterior.
- En nuestros estudios de simulación, resultan ser prácticamente iguales a los estimadores EB.
- Una vez se ajusta el modelo, se puede estimar para cualquier subarea o subdominio. Incluso se puede estimar a nivel de individuo.
- Una vez se ajusta el modelo, se pueden estimar al mismo tiempo todos los indicadores (que sean función de Y_{di}) que se desee, sin necesidad de ajustar un modelo distinto para cada indicador.
- Al contrario que la mayoría de los procedimientos bayesianos, el método HB propuesto no requiere el uso de métodos MCMC y por tanto no requiere la monitorización de la convergencia de las cadenas de Markov.
- No requiere la utilización de métodos *Bootstrap* para la estimación del ECM. Por lo tanto, el tiempo computacional total puede ser considerablemente menor que en el método EB+*Bootstrap*.
- El cálculo de intervalos creíbles o cualquier otro resumen de la distribución a posterior es automático.

Desventajas:

- Están basados en un modelo. Por tanto, es necesario comprobar que el modelo se ajusta correctamente (e.g., a través de los residuos predictivos o de validación cruzada, véase Molina, Nandram y Rao, 2014).
- No tienen en cuenta el diseño muestral. No son insesgados bajo el diseño y pueden tener un sesgo considerable bajo diseño informativo.
- Pueden verse seriamente afectados por datos atípicos aislados o por la falta de normalidad.
- El método HB no se puede extender directamente a modelos más complejos sin perder alguna de las ventajas mencionadas, como por ejemplo el evitar la aplicación de métodos MCMC.

F. Métodos basados en modelos lineales generalizados mixtos

El acceso a determinados servicios educativos o sanitarios, o la disponibilidad de ciertas comodidades en la vivienda se miden habitualmente en un área concreta mediante las proporciones de personas en esa área que tienen acceso o no al servicio o comodidad en cuestión. Los modelos lineales mixtos considerados hasta ahora no proporcionan predicciones en el espacio natural $[0,1]$ en el que están las proporciones. Para obtener predicciones en dicho espacio, es habitual considerar modelos lineales generalizados mixtos (en inglés, *generalized linear mixed models*, GLMM). Si $Y_{di} \in \{0,1\}$ es la variable binaria que mide la carencia o no del servicio o comodidad en cuestión, el modelo de estimación en áreas pequeñas más habitual es el GLMM con efectos aleatorios de las áreas, dado por

$$Y_{di}|v_d \sim \text{Bern}(p_{di}), g(p_{di}) = \boldsymbol{x}_{di}{}'\boldsymbol{\alpha} + v_d, v_d \overset{iid}{\sim} N(0, \sigma_v^2), i = 1, \dots, N_d, d = 1, \dots, D, \tag{60}$$

donde v_d es el efecto del área d, $\boldsymbol{\alpha}$ es el vector de coeficientes de la regresión y g: $(0,1) \to R$ es la función vínculo (biyectiva, con derivada continua). En particular, el vínculo logístico dado por $g(p) = \log(p/(1-p))$ es probablemente el más utilizado en la práctica.

Como se ha comentado anteriormente, el mejor predictor bajo el modelo (el cual minimiza el ECM bajo el modelo) de la proporción $P_d = \bar{Y}_d$, viene dado por

$$\tilde{P}_d^B(\boldsymbol{\theta}) = E(P_d|\boldsymbol{y}_{ds}; \boldsymbol{\theta}) = \frac{1}{N_d}\left\{\sum_{i \in s_d} Y_{di} + \sum_{i \in r_d} E(Y_{di}|\boldsymbol{y}_{ds}; \boldsymbol{\theta})\right\}. \tag{61}$$

La distribución de $Y_{di}|\boldsymbol{y}_{ds}$ depende del vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \sigma_v^2)'$ de parámetros del modelo. En la práctica, obtenemos el predictor EB reemplazando $\boldsymbol{\theta}$ por un estimador consistente $\widehat{\boldsymbol{\theta}}$ en el mejor predictor, es decir, $\hat{P}_d^{EB} = \tilde{P}_d^B(\widehat{\boldsymbol{\theta}})$.

El estimador $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\alpha}}', \hat{\sigma}_v^2)'$ de $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \sigma_v^2)'$ se obtiene ajustando el modelo GLMM dado en (60) a los datos muestrales $\boldsymbol{y}_s = (\boldsymbol{y}_{1s}{}', \dots, \boldsymbol{y}_{Ds}{}')'$. Si se desea ajustar el modelo mediante el método de máxima verosimilitud, es necesario maximizar la verosimilitud dada por $f(\boldsymbol{y}_s) = \int_{R^D} f(\boldsymbol{y}_s|\boldsymbol{v}) f(\boldsymbol{v}) d\boldsymbol{v}$, donde $\boldsymbol{v} = (v_1, \dots, v_D)'$. Bajo el GLMM mencionado, dicha verosimilitud no tiene forma explícita. Por tanto, para este método de ajuste es necesario utilizar aproximaciones de la integral (e.g. numéricas) junto con técnicas de maximización numérica. Una vez se ha ajustado el modelo, necesitamos calcular las esperanzas $E(Y_{di}|\boldsymbol{y}_{ds}; \widehat{\boldsymbol{\theta}})$ que definen el predictor EB. Una forma de aproximar esta esperanza sería utilizando el Teorema de Bayes y el hecho de que, dado v_d, las variables $\{Y_{di}; i = 1, \dots, N_d\}$ son todas independientes. En este caso, dicha esperanza se puede expresar de la forma

$$E(Y_{di}|\boldsymbol{y}_{ds}; \widehat{\boldsymbol{\theta}}) = \frac{E\{h(\boldsymbol{x}_{di}{}'\boldsymbol{\alpha} + v_d) f(\boldsymbol{y}_{ds}|v_d); \widehat{\boldsymbol{\theta}}\}}{E\{f(\boldsymbol{y}_{ds}|v_d); \widehat{\boldsymbol{\theta}}\}}, \quad i \in r_d, \tag{62}$$

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

$$f(\boldsymbol{y}_{ds}|v_d) = \prod_{i\in s_d} p_{di}^{Y_{di}}\,(1-p_{di})^{(1-Y_{di})} = \prod_{i\in s_d} h\,(\boldsymbol{x}_{di}{}'\boldsymbol{\alpha}+v_d)^{Y_{di}}\{1-h(\boldsymbol{x}_{di}{}'\boldsymbol{\alpha}+v_d)\}^{(1-Y_{di})}. \tag{63}$$

Para el vínculo logístico, el vínculo inverso es $h(\boldsymbol{x}_{di}{}'\boldsymbol{\alpha}+v_d) = \exp(\boldsymbol{x}_{di}{}'\boldsymbol{\alpha}+v_d)/\{1+\exp(\boldsymbol{x}_{di}{}'\boldsymbol{\alpha}+v_d)\}$. Utilizando (63), podemos aproximar las dos esperanzas que aparencen en (62) mediante simulación Monte Carlo, generando $v_d^{(r)} \sim N(0,\hat{\sigma}_v^2)$, $r=1,\dots,R$, y después calculando

$$E(Y_{di}|\boldsymbol{y}_{ds};\widehat{\boldsymbol{\theta}}) \approx \frac{R^{-1}\sum_{r=1}^{R} h\,(\boldsymbol{x}_{di}{}'\widehat{\boldsymbol{\alpha}}+v_d^{(r)})\hat{f}(\boldsymbol{y}_{ds}|v_d^{(r)})}{R^{-1}\sum_{r=1}^{R}\hat{f}\,(\boldsymbol{y}_{ds}|v_d^{(r)})}, \quad i\in r_d, \tag{64}$$

donde $\hat{f}$ es la densidad condicionada $f(\boldsymbol{y}_{ds}|v_d)$, con $\boldsymbol{\alpha}$ reemplazado por $\widehat{\boldsymbol{\alpha}}$.

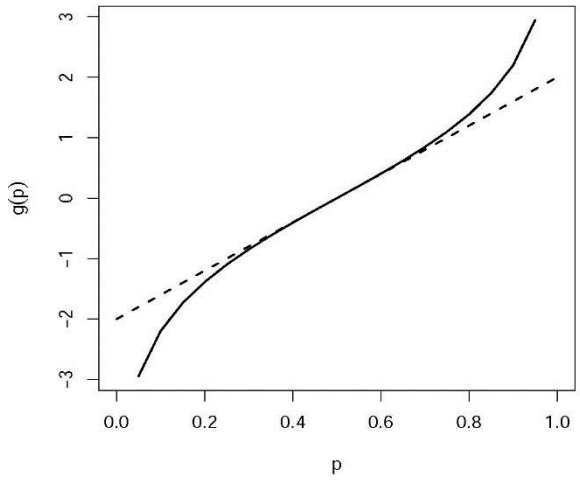
El mejor predictor (61) tiene ECM mínimo y es insesgado bajo el modelo (60). Sin embargo, ajustar el GLMM y calcular la aproximación Monte Carlo de $\hat{P}_d^{EB}$ como se ha descrito, requiere un tiempo computacional considerable. La estimación del ECM de los predictores EB mediante un procedimiento de remuestreo incrementa considerablemente el tiempo computacional, haciéndolo inviable para poblaciones muy grandes. Además, al estimar los parámetros del modelo $\boldsymbol{\theta}$ y reemplazar los estimadores para obtener la versión empírica del mejor predictor (EB), perdemos la insesgadez.

Existen estimadores simples que, a pesar de no ser óptimos, se asimilan enormemente a los óptimos bajo ciertas condiciones y se pueden obtener directamente de la salida del software habitual para el ajuste de GLMMs. Cuando se estima una proporción, si $\widehat{\boldsymbol{\alpha}}$ y $\hat{v}_d$ son los estimadores de $\boldsymbol{\alpha}$ y v_d que devuelve el software, se puede calcular un estimador por el método de la analogía (en inglés, *plug-in*) simplemente prediciendo los valores fuera de la muestra a través del modelo, es decir, tomando

$$\hat{P}_d^{PI} = \frac{1}{N_d}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d}\hat{p}_{di}\right), \tag{65}$$

donde $\hat{p}_{di} = h(\boldsymbol{x}_{di}{}'\widehat{\boldsymbol{\alpha}}+\hat{v}_d)$ es el valor predicho de la observación fuera de la muestra Y_{di}, $i\in r_d$. Para $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \sigma_v^2)'$ conocido, el estimador *plug-in*, $\hat{P}_d^{PI}$, no puede tener menor ECM que el mejor predictor $\hat{P}_d^{B}$. De hecho, al contrario que el mejor predictor, el estimador *plug-in* no es insesgado a menos que la función vínculo sea lineal. Sin embargo, el estimador *plug-in* es mucho más fácil de calcular. Los dos estimadores coinciden cuando la función vínculo $g(\cdot)$ es lineal. En el caso del vínculo logístico $g(p) = \log(p/(1-p))$, éste es aproximadamente lineal para $p\in(0.2,0.8)$ como se muestra en el Gráfico 10. Esta linealidad aproximada de $g(p)$ para valores centrales de p nos hace pensar que el estimador *plug-in* (65) basado en el modelo con vínculo logístico debe ser muy similar al predictor EB, $\hat{P}_d^{EB}$, en términos de ECM, al menos para valores no muy extremos de p. Además, esta linealidad aproximada para valores centrales de p también hace que ambos estimadores, EB y *plug-in* de la proporción $P_d = \bar{Y}_d$, se parezcan al EBLUP, $\hat{P}_d^{EBLUP} = \hat{\bar{Y}}_d^{EBLUP}$, basado en el modelo con errores anidados descrito en el capítulo I. Esto significa que, para estimar proporciones de individuos con cualidades ni demasiado poco ni extremadamente frecuentes, también tiene sentido utilizar dicho EBLUP.

Gráfico 10
Vínculo logístico



Fuente: Elaboración propia.

Ambos métodos, EB y plug-in, basados en modelos no lineales como el GLMM dado en (60), incluso estimando medias $\bar{Y}_d$, requieren disponer de los valores de las variables auxiliares para todos los individuos (microdatos), obtenidos de un censo o un registro administrativo. Esto es necesario para calcular la esperanza $E(Y_{di}|\boldsymbol{y}_{ds};\boldsymbol{\theta})$ en el caso del predictor EB, o para predecir la probabilidad $\hat{p}_{di}$ en el caso del estimador *plug-in*. Sin embargo, el EBLUP de $\bar{Y}_d$ solo requiere, además de los datos de la encuesta, las medias poblacionales de dichas variables en las áreas. Dichos datos agregados suelen estar disponbibles sin restricciones de confidencialidad.

En principio, se podría utilizar el GLMM dado en (60) para estimar la incidencia de pobreza (indicador FGT de orden $\alpha = 0$), F_{0d}. Para la brecha de pobreza (indicador FGT con $\alpha = 1$), F_{1d}, no tendría sentido utilizarlo pues no son proporciones, ya que los valores individuales $F_{1,di}$ no son variables binarias. En el caso de la incidencia de pobreza, tomando como variable respuesta binaria $Y_{di} = I(E_{di} < z)$, obtenemos $P_d = F_{\alpha d}$. El mejor predictor resultante toma la expresión (47) de la Sección V.D, pero la esperanza que aparece en el segundo término sería respecto de la distribución condicionada bajo el modelo (60) y habría que aproximarla de forma numérica; e.g., como en (64) ya que, en este caso, las distribuciones condicionadas $Y_{di}|\boldsymbol{y}_{ds}$ no tienen una forma conocida. Como se ha mencionado, el estimador *plug-in* (65) sería menos costoso computacionalmente. De nuevo, en el caso de no poder identificar las unidades de la encuesta en el censo o registro, es posible utilizar un estimador *Census EB* sustituyendo en el predictor (61) las observaciones de la muestra Y_{di}, $i \in s_d$, con predicciones obtenidas como en (62), o usando como predicción $\hat{p}_{di}$ en el caso del estimador *Census plug-in*.

El ECM del correspondiente predictor (EB o *plug-in*) se puede estimar mediante un procedimiento *Bootstrap* de la forma siguiente (véase González-Manteiga et al., 2007):

1) Ajustar el GLMM dado en (60) a los datos de la muestra s, obteniendo estimadores $\hat{\sigma}_v^2$ y $\widehat{\boldsymbol{\alpha}}$ de los parámetros del modelo.
2) Generar efectos aleatorios *Bootstrap*
$$v_d^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_v^2), \quad d = 1, \dots, D.$$
3) Generar un censo *Bootstrap* $\boldsymbol{y}_d^{*(b)} = (Y_{d1}, \dots, Y_{dN_d})'$, de la forma
$$Y_{di}^{*(b)} \overset{ind}{\sim} \text{Bern}(p_{di}^{*(b)}), p_{di}^{*(b)} = h(\boldsymbol{x}_{di}'\widehat{\boldsymbol{\alpha}} + v_d^{*(b)}), i = 1, \dots, N_d, d = 1, \dots, D, \qquad (66)$$
y calcular los verdaderos valores de los indicadores $P_d^{*(b)} = \bar{Y}_d^{*(b)}$, $d = 1, \dots, D$.

4) Para cada área $d = 1, \dots, D$, extraer del censo *Bootstrap* $\boldsymbol{y}_d^{*(b)}$ los elementos de la muestra de esa área, Y_{di}, $i \in s_d^{)}$, construyendo el vector $\boldsymbol{y}_{ds}^{*(b)}$. Sea $\boldsymbol{y}_s^{*(b)} = ((\boldsymbol{y}_{1s}^{*(b)})', \dots, (\boldsymbol{y}_{Ds}^{*(b)})')'$ el vector con los valores en la muestra de todas las áreas, siendo $s = s_1 \cup \cdots \cup s_D$ la muestra original.
5) Ajustar el modelo (60) a los datos *Bootstrap* $\boldsymbol{y}_s^{*(b)}$ y calcular los predictores *Bootstrap* $\hat{P}_d^{EB*(b)}$, $d = 1, \dots, D$.
6) Repetir los pasos 2)–5), para $b = 1, \dots, B$. El estimador *Bootstrap* del ECM del predictor $\hat{P}_d^{EB}$ 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 de características del predictor EB/*plug-in* basado en el GLMM, en comparación con los métodos aplicables a la estimación de medias:

Indicadores objetivo: Proporciones o totales de una variable binaria (e.g. carencia o no de determinado bien o servicio).

Requerimientos de datos:

- Microdatos de las p variables auxiliares consideradas, de la misma encuesta donde se observa la variable de interés.
- Área de interés obtenida de la misma encuesta donde se observa la variable de interés.
- Microdatos de las p variables auxiliares consideradas a partir de un censo o un registro administrativo (medidas de la misma forma que en la encuesta).

Ventajas:

- El número de observaciones usadas para ajustar el modelo es el tamaño muestral total n, mucho mayor que el número de áreas en los modelos FH. Por tanto, los parámetros del modelo se estiman con mucha eficiencia y las ganancias en eficiencia respecto a los estimadores directos tienden a ser mayores que con los modelos FH.
- El modelo de regresión considerado incorpora heterogeneidad no explicada entre las áreas.
- Al contrario que en el modelo FH, no se necesita conocer ninguna varianza.
- Se pueden desagregar las estimaciones para cualquier subdominio o subárea deseada dentro de las áreas; incluso a nivel de individuo.
- Se puede estimar en áreas no muestreadas.

Desventajas:

- Se basan en un modelo; por tanto, es necesario analizar dicho modelo (e.g. a través de los residuos).
- No tiene en cuenta el diseño muestral. Por tanto, no es insesgado bajo el diseño y es más apropiado para el muestreo aleatorio simple. Se verá afectado por diseños muestrales informativos.
- Los microdatos suelen ser obtenidos de un censo o registro administrativo, y suelen existir problemas de confidencialidad que limiten el uso de este tipo de datos.
- El estimador del ECM bajo el modelo obtenido e.g. por procedimientos *Bootstrap* es correcto bajo el modelo considerado, y no es insesgado bajo el diseño para el ECM bajo el diseño para un área concreta.
- El predictor EB (a diferencia del estimador *plug-in*) posee un elevado coste computacional.
- El ECM del predictor EB obtenido e.g. mediante un procedimiento *Bootstrap* tiene un coste computacional excesivamente alto y puede ser inviable para poblaciones muy grandes. Para el predictor *plug-in*, el coste es considerablemente menor.
- Requieren un reajuste para verificar la propiedad "*benchmarking*": que la suma de los totales estimados en las áreas de una región mayor coincida con el estimador directo para dicha área.

VI. Aplicación: estimación de ingresos medios y tasas de pobreza en Montevideo

En este capítulo vamos a utilizar algunas de las técnicas descritas anteriormente para la estimación de los ingresos medios y las incidencias de pobreza no extrema para las secciones censales y para los dos géneros en Montevideo, Uruguay. Para ello, usaremos datos de la Encuesta Continua de Hogares (ECH) y del Censo de Población, ambos de 2011. El objeto de esta aplicación es meramente ilustrativo y, probablemente, mejorable mediante la realización de una búsqueda más intensiva de información auxiliar. Por tanto, los resultados obtenidos en esta aplicación no se deben considerar como estimaciones definitivas.

Puesto que los parámetros de los modelos a considerar pueden depender del género, para cada tipo de estimador ajustaremos modelos separados para cada género. Concretamente, vamos a calcular estimaciones directas usando los microdatos provenientes de la ECH para cada sección y género, estimaciones FH basadas en el modelo básico a nivel de área (21), usando como información auxiliar ciertos totales de población para cada sección y género, extraídos del censo y, finalmente, estimaciones *Census EB* basadas en el modelo básico a nivel de individuo (38) para el logaritmo de los ingresos, usando microdatos del censo de algunas variables medidas también en la ECH. Obsérvese que, incluso si solo se estimase la media de los ingresos, que es un parámetro lineal en los ingresos de los individuos en el área, al realizar una transformación no lineal (logarítmica) de la variable respuesta en el modelo con errores anidados (38), el parámetro objetivo, escrito en función de los valores de la variable respuesta del modelo, es un parámetro no lineal en los valores de la variable respuesta del modelo. Por tanto, en este caso, el EBLUP no tiene sentido ya que es un estimador lineal en los valores de la variable respuesta del modelo en la muestra y necesitamos recurrir a la metodología EB. Adicionalmente, dado que los individuos de la ECH no están identificados en el censo, consideramos el estimador *Census EB*. Además de los estimadores puntuales, se obtendrán estimaciones de los ECMs de cada estimador. Se han realizado los cálculos usando los paquetes de R `sae` (Molina y Marhuenda, 2015) y `lme4` (Bates et al. 2015).

Los tamaños poblacionales según el cuestionario censal completo (para moradores presentes en viviendas particulares) del Censo de Población de 2011 en Montevideo son $\mathrm{N} = 656{,}162$ para mujeres y $\mathrm{N} = 566{,}698$ para hombres. Los tamaños muestrales de la ECH, después de descartar datos faltantes, son $n = 26{,}233$ para mujeres y $n = 22{,}464$ para hombres. Para las $D = 25$ secciones que aparecen en el censo, los tamaños muestrales varían entre 56 y 3482 para mujeres y entre 65 y 2820 para hombres.

Aunque no son tamaños muestrales excesivamente pequeños, vamos a ver que, aun así, las técnicas de estimación en áreas pequeñas pueden proporcionar estimaciones más precisas, midiendo dicha precisión en términos de error cuadrático medio. También hay que tener en cuenta que, según los datos disponibles, las incidencias de pobreza en Montevideo son relativamente bajas y, para estimar estas cantidades con precisión usando estimadores directos, los tamaños muestrales necesarios por sección y género deben ser mayores que para estimar proporciones cercanas a 0.5 o medias de variables continuas, como los ingresos medios. De hecho, incluso si el tamaño muestral no es excesivamente pequeño, el estimador directo puede resultar igual a cero debido a que no se obtenga ningún individuo con ingresos por debajo del umbral de pobreza. El umbral de pobreza no extrema para áreas urbanas en el año 2011 es de 3,182 pesos uruguayos.

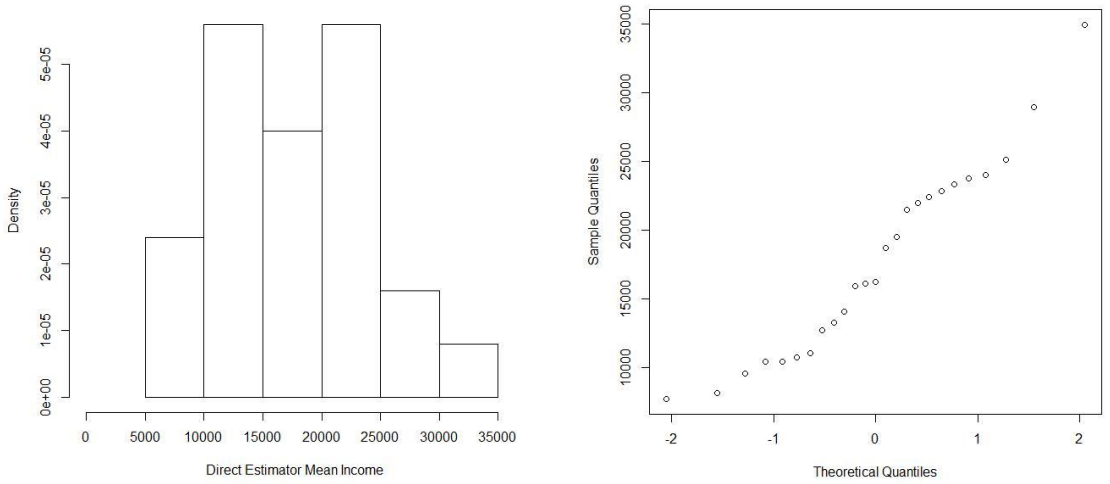
Tanto para los ingresos medios $\bar{E}_d = N_d^{-1} \sum_{i=1}^{N_d} E_{di}$ como para las incidencias de pobreza $F_{0d} = N_d^{-1} \sum_{i=1}^{N_d} I\,(E_{di} < z)$ para cada sección censal y género, los estimadores directos correspondientes, $\hat{\bar{E}}_d^{DIR}$ y $\hat{F}_{0d}^{DIR}$, y sus varianzas muestrales estimadas $\widehat{\text{var}}_\pi(\hat{\bar{E}}_d^{DIR})$ y $\widehat{\text{var}}_\pi(\hat{F}_{0d}^{DIR})$ se obtienen usando los microdatos de la ECH en las fórmulas (4)–(6). Esto es lo que nos proporciona la función direct() del paquete sae, introduciendo los pesos muestrales de la ECH. Como tamaños poblacionales de las secciones censales, N_d, utilizamos los tamaños obtenidos del Censo.

Los estimadores FH y sus errores cuadráticos medios estimados se obtienen a partir del modelo (21) para $\delta_d = \bar{E}_d$ o $\delta_d = F_{0d}$. En el caso de los ingresos medios, $\delta_d = \bar{E}_d$, para ambos géneros, consideramos como variables auxiliares agregadas a nivel de sección censal (componentes de $\boldsymbol{x}_d$ en el modelo), las proporciones censales de individuos alfabetizados, de individuos inactivos distintos de jubilados, la edad media y los años de estudio medios. Para las incidencias de pobreza, $\delta_d = F_{0d}$, solo resultan significativas las proporciones de individuos alfabetizados y las de individuos cesantes. Los estimadores FH se pueden obtener mediante la función eblupFH() del paquete sae que implementa la fórmula dada en (24). Como vector de variables respuesta del modelo, se establece el vector de estimaciones directas obtenidas anteriormente $\hat{\bar{E}}_d^{DIR}$ o $\hat{F}_{0d}^{DIR}$ según sea el caso, y, como varianzas ψ_d, las estimaciones de las varianzas muestrales $\widehat{\text{var}}_\pi(\hat{\bar{E}}_d^{DIR})$ o $\widehat{\text{var}}_\pi(\hat{F}_{0d}^{DIR})$. Los ECMs estimados, $\text{mse}_{PR}(\hat{\delta}_d^{FH})$, se obtienen mediante las fórmulas analíticas de la Sección V.A, para el ajuste mediante el método REML, y en R se obtienen usando la función mseFH() del paquete anterior.

Los estimadores FH de la incidencia de pobreza pueden tomar el valor cero (al igual que los estimadores directos) en los dominios en los que no existan individuos con ingresos por debajo del umbral de la pobreza. Además, los ECMs estimados mediante la fórmula analítica mencionada también toman el valor cero. En estos casos, consideramos que dichas estimaciones FH no son fiables y, en su lugar, calculamos los estimadores sintéticos $\hat{\delta}_d^{FH} = \boldsymbol{x}_d{}'\widehat{\boldsymbol{\beta}}$. Sus ECMs se obtienen mediante la fórmula (6.2.14) de Rao y Molina (2015), reemplazando el estimador REML de la varianza de los efectos del dominio.

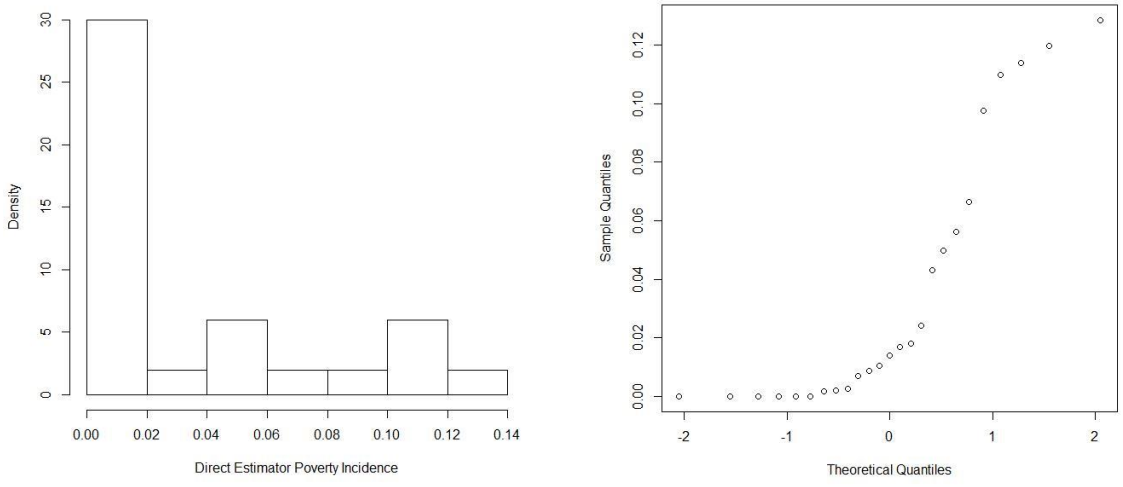
A pesar de que el EBLUP basado en el modelo Fay-Herriot no requiere normalidad, la aproximación analítica del ECM obtenido de esta forma sí requiere normalidad. Como podemos ver en el histograma y el gráfico q-q de normalidad para mujeres (Gráfico 11), la distribución de los estimadores directos de los ingresos medios para las D=25 secciones censales no se ajustan a una distribución normal, pero tampoco se aleja excesivamente, hay que tener en cuenta que el número de observaciones usadas para construir el histograma (D=25) es pequeño. Para hombres, los gráficos son similares. Este no es el caso de los estimadores directos de las incidencias de pobreza no extrema, véase el Gráfico 12. Por tanto, hay que tener en cuenta que los ECMs estimados de dichas incidencias de pobreza pueden no corresponder a la realidad.

Gráfico 11
Histograma (izquierda) y gráfico q-q de normalidad (derecha) de los estimadores directos de los ingresos medios para las $D = 25$ secciones censales de Montevideo, en el caso de Mujeres
(En pesos uruguayos, año 2011)



Fuente: Elaboración propia.

Gráfico 12
Histograma (izquierda) y gráfico q-q de normalidad (derecha) de los estimadores directos de las incidencias de pobreza no extrema para las $D = 25$ secciones censales de Montevideo, en el caso de Mujeres
(En proporciones)

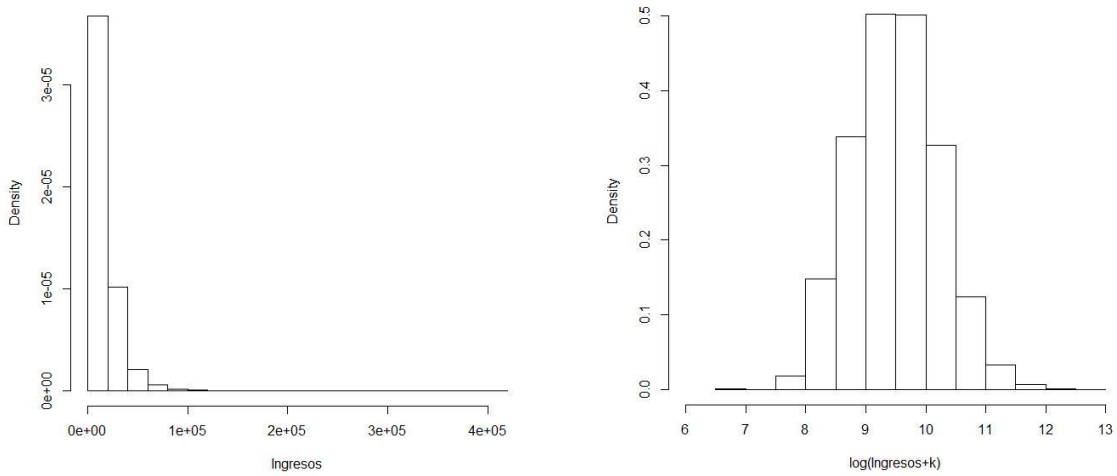


Fuente: Elaboración propia.

Finalmente, obtenemos los estimadores *Census EB* basados en el modelo a nivel de individuo (38), usando como variable respuesta $\log(\text{ingresos} + 1000)$, donde la constante 1000 añadida a los ingresos se ha determinado de manera que el histograma de los residuos del modelo ajustado sea aproximadamente simétrico, véase el histograma de los ingresos originales y de los ingresos transformados de esta forma en el Gráfico 13. Como variables auxiliares a nivel de individuo en $\boldsymbol{x}_{di}$, consideramos los indicadores de las

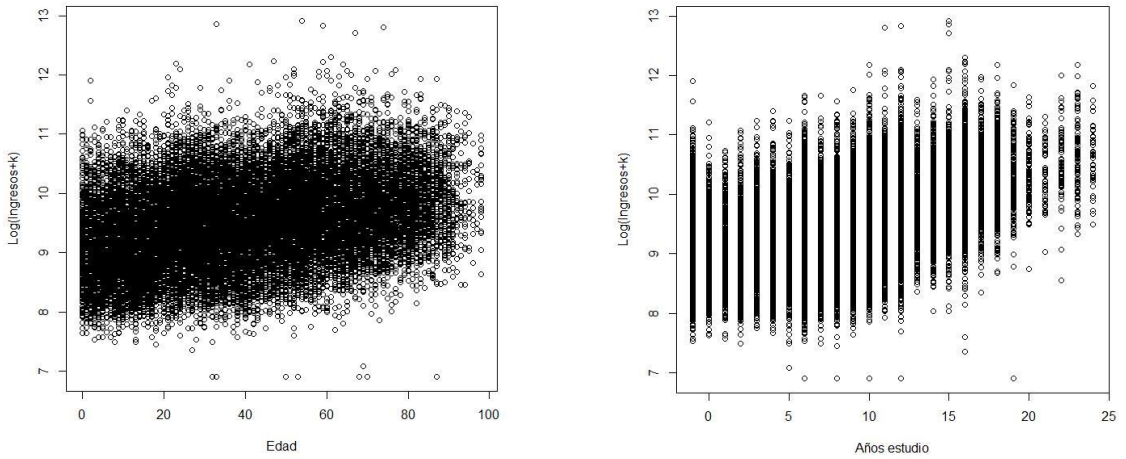
categorías de la condición de actividad, la edad y los años de estudio. El Gráfico14 para mujeres muestra una relación aproximadamente lineal creciente entre los ingresos transformados y la edad o los años de estudio. El gráfico para hombres es similar. Dado que la transformación de los ingresos es monótona, esta relación indica que, a medida que aumentan la edad o los años de estudio, también aumentan los ingresos.

Gráfico 13
Histograma de los ingresos sin transformar (izquierda) y transformados de la forma log(ingresos + 1000) (derecha), en el caso de Mujeres
(En pesos uruguayos, año 2011)



Fuente: Elaboración propia.

Gráfico 14
Ingresos transformados frente a la edad (izquierda) y frente a los años de estudio (derecha), en el caso de Mujeres
(En pesos uruguayos- transformación logarítmica-, año 2011)



Fuente: Elaboración propia.

Los estimadores *Census EB* de las incidencias de pobreza $F_{\alpha d}$ basados en el modelo con errores anidados para los ingresos transformados se calculan utilizando las fórmulas (53) y (48), reemplazando $\boldsymbol{\theta}$ por el estimador $\widehat{\boldsymbol{\theta}}$; en este caso, hemos usado el estimador REML. Aunque, como se ha visto en el Ejemplo 7, la función `ebBHF()` del paquete `sae` proporciona los estimadores EB pero no los estimadores *Census EB*, si las fracciones muestrales de las áreas son pequeñas, podemos utilizar la misma función para obtener valores aproximados de los estimadores *Census EB*, fijando el atributo `Xnonsample` de dicha función (matriz de valores de las variables auxiliares para la parte de la población fuera de la muestra) igual a la matriz con los microdatos censales de dichas variables para todos los individuos de las secciones consideradas. En este caso, se puede comprobar que las estimaciones *Census EB* y las obtenidas de esta forma presentan diferencias realmente pequeñas.

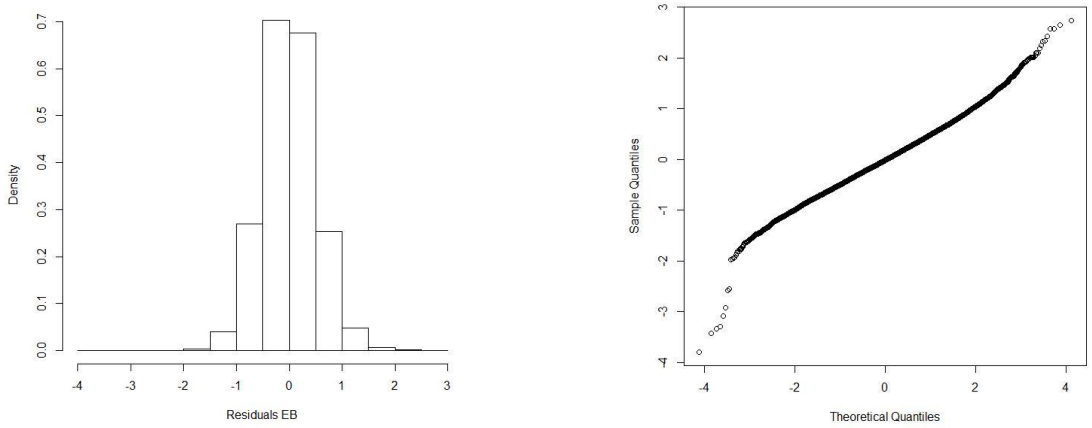
El mismo modelo ajustado permite obtener los estimadores *Census EB* de los ingresos medios. Los estimadores *Census EB* de los ingresos medios $\delta_d = \bar{E}_d$ basados en dicho modelo se obtienen de forma similar. Concretamente, se obtienen de la forma $\hat{\bar{E}}_d^{CEB} = N_d^{-1} \sum_{i=1}^{N_d} \hat{E}_{di}^{CEB}$ donde, teniendo en cuenta que los ingresos E_{di} se obtienen en función de las variables respuesta en el modelo Y_{di} de la forma $E_{di} = \exp(Y_{di}) + 1000$, entonces $\hat{E}_{di}^{CEB} = E[\exp(Y_{di})|\boldsymbol{y}_s; \widehat{\boldsymbol{\theta}}] + 1000$. Dicha esperanza se puede obtener usando la aproximación Monte Carlo (50) e implementada en la función `ebBHF()`, o mediante la fórmula analítica dada en Molina y Martín (2018). En este caso, se ha utilizado dicha fórmula analítica pues prácticamente no tiene coste computacional.

Los ECMs estimados de los estimadores *Census EB* se obtienen mediante una ligera modificación del procedimiento *Bootstrap* descrito en el capítulo V (diseñado en principio para los estimadores EB), usando $B = 500$ repeticiones *Bootstrap*. La diferencia entre los estimadores EB y *Census EB* radica en el hecho de que no se pueden identificar las unidades de la ECH en el censo. Por tanto, en cada repetición *Bootstrap* no podemos generar vectores censales $\boldsymbol{y}_d^{*(b)}$, $d = 1, \dots, D$, y de ellos tomar la parte de la muestra $\boldsymbol{y}_s^{*(b)}$. En el caso de los estimadores *Census EB*, generamos los censos *Bootstrap* $\boldsymbol{y}_d^{*(b)}$, $d = 1, \dots, D$, usando los valores de las variables auxiliares del censo y, por otro lado, se genera el vector de la muestra *Bootstrap* $\boldsymbol{y}_s^{*(b)}$ usando los valores de las mismas variables auxiliares, pero tomados de la ECH. Los verdaderos valores de los parámetros *Bootstrap* se obtienen a partir de los censos *Bootstrap* generados, $\delta_d^{(b)} = \delta_d(\boldsymbol{y}_d^{*(b)}), d = 1, \dots, D$.

El método EB (o *Census EB*) y el procedimiento *Bootstrap* utilizados se basan en la hipótesis de normalidad; por tanto, en este caso es crucial estudiar si esta hipótesis se verifica, al menos aproximadamente. En el Gráfico 15 se muestran el histograma y el gráfico q-q de normalidad de los residuos del ajuste del modelo para los ingresos transformados para mujeres. Aunque los datos reales difícilmente se ajustan a un modelo de forma exacta, y cualquier test va a rechazar la hipótesis nula de normalidad si el tamaño muestral es grande como en este caso, podemos ver en estos gráficos que la distribución no se aleja excesivamente de la normal. Si se desea utilizar una distribución que se ajuste algo mejor a los ingresos, se puede utilizar el método EB basado en un modelo multivariante GB2 como el propuesto por Graf, Marín y Molina (2018).

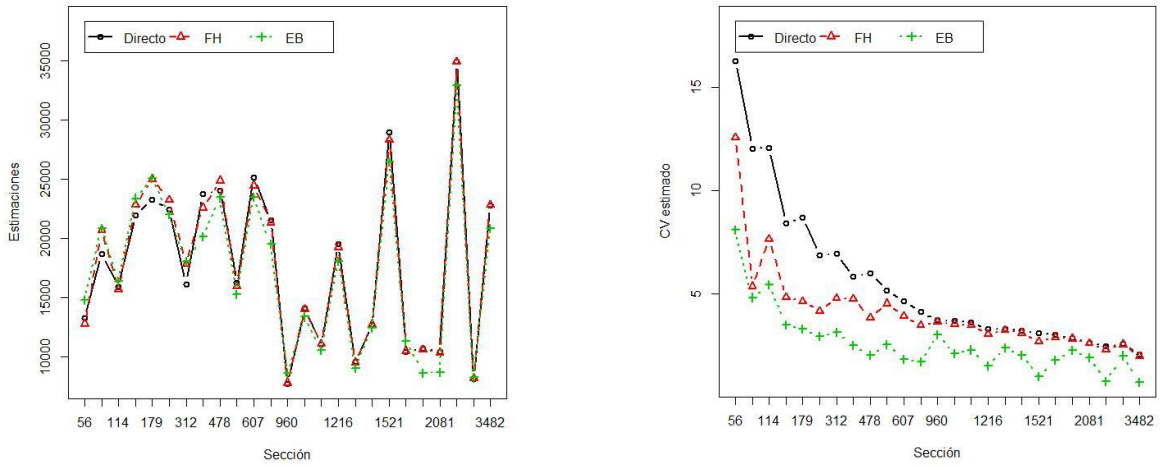
Los resultados numéricos detallados para cada sección censal se muestran en los Cuadros 1–4 del Anexo. A continuación, analizamos gráficamente dichos resultados y comentamos los resultados obtenidos para los distintos estimadores. El Gráfico16 muestra los valores obtenidos de los estimadores directos, FH y *Census EB* de los ingresos medios (izquierda), y los CVs estimados de dichos estimadores (derecha) para cada sección censal, para Mujeres. Las secciones censales (eje x) están ordenadas de menor a mayor tamaño muestral y se han indicado sus tamaños muestrales en las etiquetas del eje x. Se puede observar cómo los tres estimadores toman valores similares, aunque los estimadores directos y FH obtienen prácticamente los mismos valores en este caso. Esto es debido a que, a la hora de estimar los ingresos medios, los tamaños muestrales de las secciones no son excesivamente pequeños y el peso que otorgan los estimadores FH a los correspondientes estimadores directos se acercan a uno. Esta es una ventaja de los estimadores basados en modelos con efectos aleatorios. Sin embargo, y aunque los tamaños muestrales sean moderados, como podemos observar en el gráfico de la derecha, los estimadores *Census EB* son claramente más eficientes que los directos y FH para todas las secciones censales. Esto es debido a que utilizan una mayor cantidad de información, los microdatos del censo. Para hombres (Gráfico 17), podemos extraer conclusiones similares.

Gráfico 15
Histograma (izquierda) y gráfico q-q de normalidad (derecha) de los residuos del modelo con errores anidados para los ingresos transformados, en el caso de Mujeres
(En proporciones)



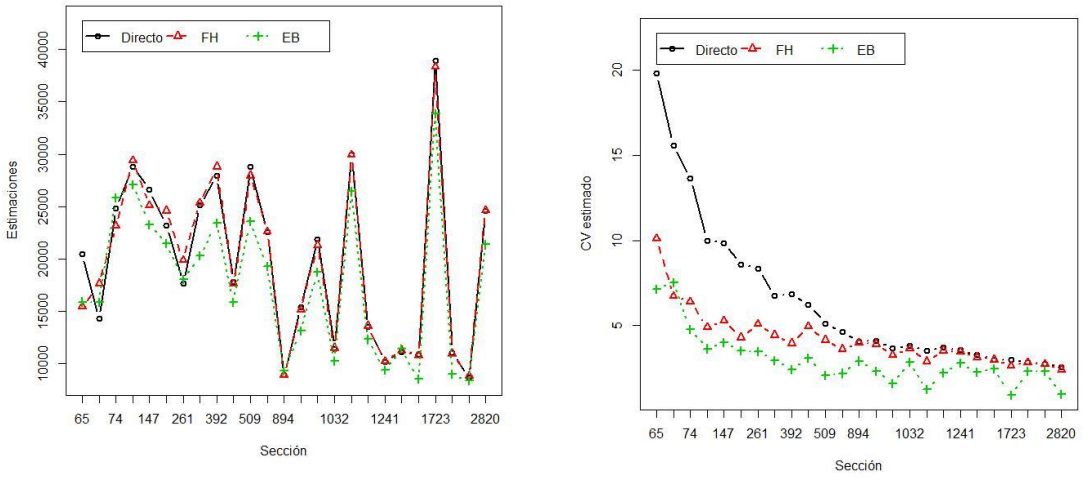
Fuente: Elaboración propia.

Gráfico 16
Estimaciones directas, FH y *Census EB* (izquierda) de los ingresos medios, y CVs de los estimadores (derecha) para las $D = 25$ secciones censales de Montevideo, en el caso de Mujeres. Secciones censales (eje x) ordenadas de menor a mayor tamaño muestral, con tamaños muestrales indicados en el eje
(En pesos uruguayos, año 2011)



Fuente: Elaboración propia.

Gráfico 17
Estimaciones directas, FH y *Census EB* (izquierda) de los ingresos medios, y CVs de los estimadores (derecha) para las $D = 25$ secciones censales de Montevideo, en el caso de Hombres. Secciones censales (eje x) ordenadas de menor a mayor tamaño muestral, con tamaños muestrales indicados en el eje
(En pesos uruguayos, año 2011)



Fuente: Elaboración propia.

Para las incidencias de pobreza, las estimaciones y los errores cuadráticos medios para Mujeres y Hombres se muestran respectivamente en las Gráficos18 y 19. En este caso mostramos ECMs en lugar de CVs debido a que, en el caso de proporciones, para un tamaño muestral fijo, los CVs aumentan a medida que la proporción disminuye; por tanto, los CVs tienen menos sentido como medidas de error de estimación, sobre todo cuando las proporciones estimadas toman valores muy pequeños, como es este caso. De nuevo, los valores de los tres estimadores se parecen para todas las secciones censales excepto para las de menor tamaño muestral. De hecho, en estas secciones, los estimadores directos para el caso de Mujeres toman el valor (poco creíble) cero debido a que no existen individuos muestreados con ingresos por debajo del umbral. De hecho, las varianzas estimadas de los estimadores directos también toman el valor cero para éstas secciones. Hay que tener en cuenta que las varianzas estimadas de los estimadores directos también están basadas en las pocas observaciones muestreadas de cada sección y género. Si consideramos los estimadores directos poco fiables, sus varianzas estimadas también lo son. Para los dominios con estimadores directos iguales a cero, los estimadores FH y sus ECMs también son teóricamente cero. En estos casos, como se ha comentado anteriormente, se han utilizado los estimadores sintéticos obtenidos del mismo modelo. Podemos observar en los gráficos de la derecha cómo los ECMs de los estimadores directos y FH presentan grandes altibajos. Obsérvese que los ECMs de los estimadores FH son especialmente grandes para los dominios en los que se han utilizado los estimadores sintéticos. Por el contrario, los ECMs de los estimadores EB aumentan de forma suave en relación al tamaño muestral de la sección censal. Además de tomar valores más razonables, los ECMs estimados de los estimadores EB permanecen por debajo de los ECMs de los otros dos estimadores para la mayoría de las secciones censales.

Hay que recalcar que los estimadores basados en modelos pueden proporcionar estimaciones incluso para las áreas no muestreadas, aunque esto no es recomendable ya que para estas áreas no es posible estudiar la bondad de ajuste del modelo. Hablando de bondad de ajuste, como ya se ha mencionado, para la incidencia de pobreza no extrema no se verifica la hipótesis de normalidad en el modelo Fay-Herriot. Esto ocurre debido a que los tamaños muestrales son pequeños para algunas de las secciones y las incidencias de pobreza verdaderas parecen ser bastante pequeñas, con lo cual los estimadores directos tienen una distribución marcadamente asimétrica y no se verifica el Teorema Central del Límite. Aunque la normalidad no es necesaria para obtener el estimador FH, sí se asume para la estimación del ECM

mediante las fórmulas analíticas consideradas en la Sección V.A y proporcionadas por la función mseFH() del paquete sae. De hecho, un inconveniente adicional de los estimadores obtenidos a partir de este modelo Fay-Herriot es que pueden resultar valores negativos o mayores que uno, que en el caso de proporciones no es deseable. Una solución simple es truncar las estimaciones al valor cero cuando sean negativas y uno cuando superen este valor. Otra posibilidad es considerar el modelo de regresión (19) para una transformación biyectiva de las incidencias de pobreza, $g(F_{0d})$, que traslade valores del espacio $[0,1]$ a valores reales. Sin embargo, la misma transformación del estimador directo, $g(\hat{F}_{0d}^{DIR})$, no tiene por qué ser un estimador insesgado para $g(F_{0d})$ y, por tanto, el modelo (20) no se verifica para $g(\hat{F}_{0d}^{DIR})$. En este caso, el modelo FH para $g(\hat{F}_{0d}^{DIR})$ tendría un sesgo adicional, a menos que se considere el modelo (20) para $\hat{F}_{0d}^{DIR}$ junto con el modelo de regresión anterior para $g(F_{0d})$. En este caso, los dos modelos considerados no se pueden resumir en un modelo lineal mixto como el dado en (21), es decir, son modelos desparejados (en inglés, *unmatched models*). Estimadores basados en modelos desparejados de este tipo se han obtenido por You y Rao (2002b) basándose en Inferencia Bayesiana.

Como se ha visto, en el caso de poseer información auxiliar a nivel de individuo, las ganancias en eficiencia de los estimadores que usan esta información suelen ser mayores. Sin embargo, en esta aplicación se han utilizado las dos fuentes de datos del mismo año. En años en los que no se dispone de un censo actualizado, los estimadores basados en modelos a nivel de individuo pueden proporcionar estimaciones algo sesgadas. Por tanto, en estos casos, conviene buscar otras fuentes de datos actuales, como registros administrativos. En el caso de no disponer de fuentes actualizadas de datos a nivel de individuo, se recomienda ceñirse a modelos a nivel de área. En algunos casos, se pueden encontrar fuentes de datos agregados a un nivel menor que el área. En ese caso, se podrían utilizar modelos para datos agregados a ese nivel, que incluyan efectos aleatorios anidados a dos niveles (en inglés, *two-fold subarea level models*), véase Torabi y Rao (2014).

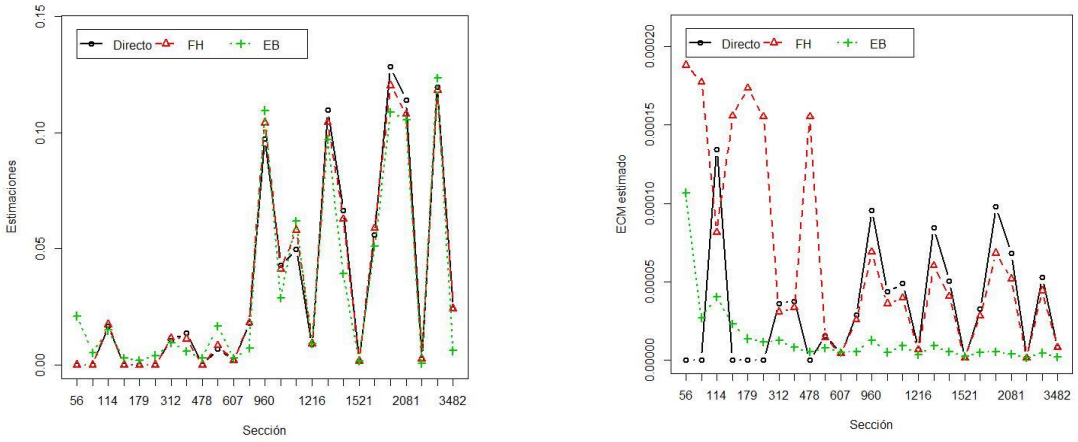
VII. Conclusiones

En este documento se ha descrito el problema de la desagregación de las estimaciones estadísticas en áreas o subgrupos de la población. Se dan recomendaciones acerca de los límites de la desagregación de las estimaciones directas y se describen los métodos indirectos básicos, así como algunos más sofisticados, que permiten superar estos límites. Como se ha visto a lo largo de este documento, los métodos a utilizar en cada aplicación concreta dependen principalmente de la forma del indicador en cuestión y del tipo de información auxiliar disponible, no existiendo métodos universales que se puedan utilizar para cualquier tipo de indicador o de información disponible. Por tanto, en cada caso, es necesario estudiar qué métodos son potencialmente aplicables en función de los requerimientos de datos y de las hipótesis que asume cada método. En aplicaciones que permitan la utilización de diversos métodos, la precisión de los estimadores finales va a depender de en qué medida las variables auxiliares disponibles sean buenas predictoras de la variable que se modelice en cada caso y de en qué medida se verifiquen las hipótesis correspondientes.

No se puede olvidar que, al mismo tiempo que se demandan estimaciones lo más precisas posible, sus medidas de error (habitualmente, los errores cuadráticos medios) deben también estar estimadas con la mayor exactitud posible o, como mínimo, que no exista una infraestimación de éstas, de modo que no se proporcione una visión erróneamente optimista de las estimaciones obtenidas. Como se ha mencionado, al producir estimaciones a nivel local, las comunidades que habitan cada zona a menudo poseen información (aunque sea subjetiva) sobre los valores plausibles de los indicadores en cuestión, y las estimaciones proporcionadas podrían contradecir este conocimiento local. Por tanto, siempre es necesario recordar a los usuarios de los datos estadísticos que dichos datos tienen un cierto grado de error, y las medidas de error que acompañen a estos datos deben reflejar los verdaderos errores cometidos para cada área.

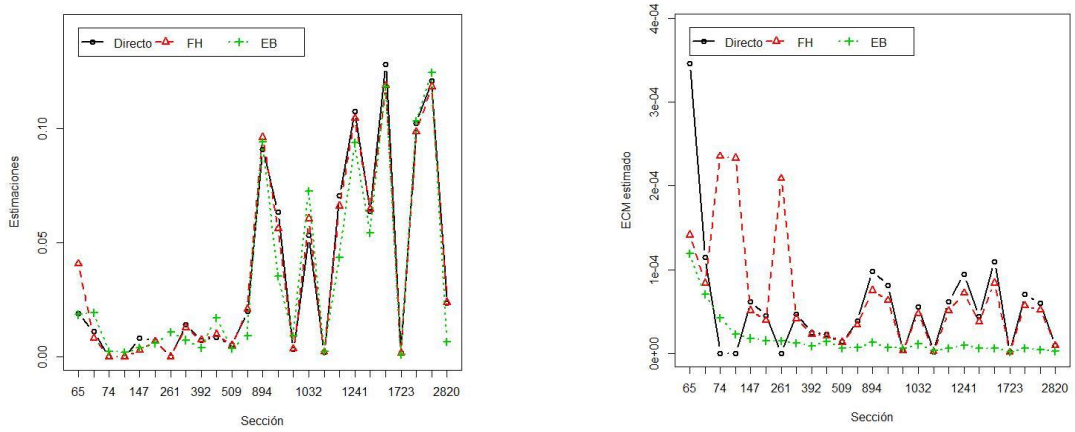
En este documento se han incluido también métodos bien extendidos para la estimación de los errores cuadráticos medios de los correspondientes estimadores indirectos. Sin embargo, no se hace referencia en este documento a medidas de error que incorporen errores ajenos al muestreo, como por ejemplo los errores de cobertura, de no respuesta, errores en los datos, debidos a la sustitución de datos faltantes, etc. Estos temas requieren de un mayor estudio dentro de la estimación en áreas pequeñas.

Gráfico 18
Estimaciones directas, FH y *Census EB* (izquierda) de la incidencia de pobreza, y ECMs de los estimadores (derecha) para las $D = 25$ secciones censales de Montevideo, en el caso de Mujeres. Secciones censales (eje x) ordenadas de menor a mayor tamaño muestral, con tamaños muestrales indicados en el eje
(En proporciones)



Fuente: Elaboración propia.

Gráfico 19
Estimaciones directas, FH y *Census EB* (izquierda) de la incidencia de pobreza, y ECMs de los estimadores (derecha) para las $D = 25$ secciones censales de Montevideo, en el caso de Hombres. Secciones censales (eje x) ordenadas de menor a mayor tamaño muestral, con tamaños muestrales indicados en el eje
(En proporciones)



Fuente: Elaboración propia.

Tampoco se debe considerar este documento como un compendio exhaustivo de métodos para la desagregación (o para la estimación del error), pues existen multitud de métodos no descritos por falta de espacio, véase Rao y Molina (2015) para una descripción más completa de la mayoría de métodos publicados anteriormente. En este documento se ha tratado de realizar una introducción al tema en cuestión, incluyendo los métodos fundamentales, en el sentido de que conforman la base para el estudio de métodos más avanzados, incluyendo solo algunos métodos más avanzados que están diseñados para la estimación de indicadores sobre las condiciones de vida.

Bibliografía

Bates, D., Maechler, M., Bolker, B. y Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67, 1-48.

Battese, G.E., Harter, R.M. y Fuller, W.A. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data, Journal of the American Statistical Association, 83, 28-36.

Bell, W. (1997). Models for county and state poverty estimates. Preprint, Statistical Research Division, U. S. Census Bureau.

Betti, G., Cheli, B., Lemmi, A. y Verma, V. (2006). Multidimensional and Longitudinal Poverty: an Integrated Fuzzy Approach, in Lemmi, A., Betti, G. (eds.) Fuzzy Set Approach to Multidimensional Poverty Measurement, 111-137, Springer, New York.

Casas-Cordero Valencia, C., Encina, J. y Lahiri, P. (2015). Poverty Mapping for the Chilean Comunas, In M. Pratesi (Ed.), Analysis of Poverty Data by Small Area Estimation: Methods for poverty mapping, New York: Wiley.

Correa, L., Molina, I. y Rao, J.N.K., (2012). Comparison of methods for estimation of poverty indicators in small areas. Unpublished report.

Datta, G.S., Fay, R.E. y Ghosh, M. (1991). Hierarchical and Empirical Bayes Multivariate Analysis in Small Area Estimation, in Proceedings of Bureau of the Census 1991 Annual Research Conference, U.S. Bureau of the Census, Washington, DC, 63–79.

Deville, J.C. y Särndal, C.E. (1992). Calibration estimation in Survey Sampling, Journal of the American Statistical Association, 87, 376–382.

Drew, D., Singh, M.P. y Choudhry, G.H. (1982). Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey, Survey Methodology, 8, 17–47.

Elbers, C., Lanjouw, J.O. y Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. Econometrica, 71, 355–364.

Estevao, V., Hidiroglou, M.A. y Särndal, C.E. (1995), Methodological Principles for a Generalized Estimation Systems at Statistics Canada, Journal of Official Statistics, 11, 181–204.

Fay, R.E. (1987). Application of Multivariate Regression to Small Domain Estimation, in R. Platek, J.N.K. Rao, C.E. Särndal y M.P. Singh (Eds.), Small Area Statistics, New York: Wiley, 91–102.

Fay, R.E. y Herriot, R.A. (1979). Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data, Journal of the American Statistical Association, 74, 269–277.

Ferretti, C. y Molina, I. (2012). Fast EB Method for Estimating Complex Poverty Indicators in Large Populations. Journal of the Indian Society of Agricultural Statistics, 66, 105–120.

Foster, J., Greer, J. y Thorbecke, E. (1984). A class of decomposable poverty measures, Econometrica, 52, 761–766.

Fuller, W.A. (1975), Regression Analysis for Sample Surveys, Sankhyā, Series C, 37, 117–132.
_____(1999). Environmental Surveys Over Time. Journal of Agricultural, Biological and Environmental Statistics, 4, 331–345.
Graf, M., Marín, J.M. y Molina, I. (2018). A generalized mixed model for skewed distributions applied to small area estimation. Manuscrito no publicado.
Ghosh, M. y Steorts, R.C. (2013). Two-stage benchmarking as applied to small area estimation, Test, 22, 670–687.
González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. y Santamaría, L. (2008). Bootstrap Mean Squared Error of a Small-Area EBLUP, Journal of Statistical Computation and Simulation, 75, 443–462.
_____(2007), Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model, Computational Statistics and Data Analysis, 51, 2720–2733.
González-Manteiga, W. (2010), Small area estimation under Fay-Herriot Models with nonparametric estimation of heteroscedasticity, Statistical Modelling, 10, 215–239.
Lumley, T. (2017). Survey: analysis of complex survey samples. R package version 3.32.
Prasad, N.G.N. y Rao, J.N.K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators, Journal of the American Statistical Association, 85, 163–171.
Molina, I. y Marhuenda, Y. (2015), sae: An R Package for Small Area Estimation, The R Journal, 7, 81–98.
Marhuenda, Y., Molina, I. y Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. Computational Statistics and Data Analysis, 58, 308–325.
Marhuenda, Y., Molina, I., Morales, D. y Rao, J.N.K. (2018). Poverty mapping in small areas under a two-fold nested error regression model. Journal of the Royal Statistical Society, Series A, en prensa.
Molina, I. y Martín, N. (2018). Empirical best prediction under a nested error model with log transformation, Annals of Statistics, en prensa.
Molina, I. y Morales, D. (2009). Small area estimation of poverty indicators. Boletín de Estadística e Investigación Operativa, 25, 218–225.
Molina, I., Nandram, B. y Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach. Annals of Applied Statistics, 8, 852–885.
Molina, I. y Rao, J.N.K. (2010). Small Area Estimation of Poverty Indicators. Canadian Journal of Statistics, 38, 369–385.
Molina, I., Salvati, N. y Pratesi, M. (2009). Bootstrap for estimating the MSE of the Spatial EBLUP. Computational Statistics, 24, 441–458.
Neri, L., Ballini, F. y Betti, G. (2005). Poverty and inequality in transition countries. Statistics in Transition, 7, 135–157.
Observatorio Social, Ministerio de Desarrollo Social de Chile (2017). Metodología de estimación de pobreza a nivel comunal, con datos de Casen 2015. Aplicación de metodologías de estimación directa, de estimación para áreas pequeñas (SAE) e imputación de medias por conglomerados (IMC). Serie Documentos Metodológicos Casen 3428.
Pfeffermann, D. y Burk, L. (1990). Robust small area estimation combining time series and cross-sectional data. Survey Methodology, 16, 217–237.
Rao, J.N.K. y Molina (2015). Small area estimation, Second Ed., Hoboken, NJ: Wiley.
Rao, J.N.K. y Yu, M. (1992). Small area estimation by combining time series and cross-sectional data, Proceedings of the Section on Survey Research Method, American Statistical Association, 1–9.
Sen A. (1976), Poverty: An Ordinal Approach to Measurement. Econometrica, 44, 219–231.
Stukel, D. y Rao, J.N.K. (1999). On small-area estimation under two-fold nested error regression models. Journal of Statistical Planning and Inference, 78, 131–147.
Tillé, Y. y Matei, A. (2016). sampling: Survey Sampling. R package version 2.8.
U.S. Bureau of Labor Statistics y U.S. Census Bureau. (2006). Design and Methodology: Current Population Survey, Technical Paper 66. Available at https://www.cen–sus.gov/prod/2006pubs/tp-66.pdf
Torabi, M. y Rao, J.N.K. (2014). On small area estimation under a sub-area level model. Journal of Multivariate Analysis, 127, 36–55.
You, Y. y Rao, J.N.K. (2002a). A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights. Canadian Journal of Statistics, 30, 431–439.
You, Y. y Rao, J.N.K. (2002b). Small area estimation using unmatched sampling and linking models, Canadian Journal of Statistics, 30, 3–15.

Anexo

Resultados de la estimación de ingresos medios y de las tasas de pobreza en Montevideo

Cuadro A.1
Estimaciones directas, FH y *Census EB* de los ingresos medios, errores cuadráticos medios y coeficientes de variación estimados de cada estimador, para cada sección censal de Montevideo, para mujeres

Sec	n_d	Directo			FH			*Census EB*		
		Est	var	cv	Est	mse	cv	Est	mse	cv
1	93	18 693,71	5 057 851,88	12,03	20714,21	1 240 278,11	5,38	21 095,71	1 042 681,38	4,84
2	56	13 277,12	4 664 014,87	16,27	12804,17	2 589 382,73	12,57	14 721,23	1 423 595,65	8,10
3	114	15 950,53	3 705 060,97	12,07	15709,41	1 449 903,59	7,66	16 405,42	804 002,00	5,47
4	172	21 964,73	3 420 289,14	8,42	22818,88	1 222 004,22	4,84	23 513,94	685803,09	3,52
5	277	22 414,35	2 388 487,30	6,90	23 267,98	946 571,10	4,18	22 041,95	423 246,74	2,95
6	179	23 313,57	4 128 976,00	8,72	25 010,83	1 352 338,30	4,65	25 195,60	703 110,23	3,33
7	421	23 755,31	1 924 432,84	5,84	22 592,03	1 163 899,22	4,78	20 071,35	256 888,38	2,53
8	312	16 154,17	1 263 151,79	6,96	17 851,16	736 152,51	4,81	18 028,97	321 051,60	3,14
9	1 113	11 063,69	161 639,01	3,63	11 127,05	151 476,54	3,50	10 598,71	59 822,24	2,31
10	3 482	22 823,33	230 630,51	2,10	22 817,42	209 158,99	2,00	20 764,66	24 042,91	0,75
11	2 081	10 473,08	76 660,70	2,64	10 390,16	74 127,46	2,62	8 758,45	29 006,70	1,94
12	1 216	19 519,30	419 289,74	3,32	19 251.93	347 714,08	3,06	18 123,74	75 945,76	1,52
13	1 844	10 741,54	95 042,46	2,87	10 634,85	91 443,15	2,84	8 652,12	38 841,43	2,28
14	792	21 514,74	790 097,23	4,13	21 340,52	560 681,72	3,51	19 518,73	113 098,81	1,72
15	607	25 157,71	1 369 375,39	4,65	24 472,89	931 112,13	3,94	23 471,75	187 914,34	1,85
16	960	7 748,40	84 359,99	3,75	7 817,03	81 443,96	3,65	8 592,14	68 306,77	3,04
17	2 278	8 167,08	44 950,84	2,60	8 217,67	44 341,90	2,56	8 286,72	28 268,28	2,03
18	2 227	34 942,88	746 573,51	2,47	34 893,09	656 698,10	2.32	33 015.11	64 575,27	0,77
19	504	16 244,32	709 726,52	5,19	15 953,47	526 515,95	4,55	15 340,75	156 341,38	2,58
20	1 402	12 724.70	168 612,47	3,23	12 758,27	158 968,95	3,13	12 436,10	65 960,74	2,07
21	1 667	10 435,48	99 478,80	3.02	10 526,60	95 005,58	2,93	11 354,96	43 098,30	1,83
22	1 073	14 104,97	272 183,12	3,70	14 011.56	248 767,70	3,56	13 370,10	82 132,18	2,14
23	478	24 032,62	2084022,21	6,01	24 886,94	920 424,29	3,85	23 547,61	230 604,23	2,04
24	1 521	28 948,32	822 681,41	3,13	28 302,65	588 417,88	2,71	26 395,05	74 187,48	1,03
99	1 364	9 614,82	101 119,90	3,31	9 548,70	96674,58	3,26	9 081,88	47 799,87	2,41

Cuadro A.2
Estimaciones directas, FH y *Census EB* de la pobreza no extrema (en %), errores cuadráticos medios y coeficientes de variación estimados de cada estimador, para cada sección censal de Montevideo, para mujeres

Sec	n_d	Directo		FH		*Census EB*	
		Est	var	Est	mse	Est	mse
1	93	0,00	0,0000	0,94	1,7750	0,50	0,2707
2	56	0,00	0,0000	6,48	1,8817	2,24	1,0674
3	114	1,68	1,3444	1,74	0,8148	1,51	0,4029
4	172	0,00	0,0000	0,00	1,5588	0,30	0,2303
5	277	0,00	0,0000	0,00	1,5516	0,42	0,1158
6	179	0,00	0,0000	0,00	1,7350	0,20	0,1362
7	421	1,39	0,3732	1,12	0,3327	0,58	0,0810
8	312	1,06	0,3617	1,16	0,3075	0,98	0,1230
9	1 113	4,99	0,4874	5,80	0,3948	6,21	0,0888
10	3 482	2,41	0.0813	2,43	0,0786	0,55	0.0184
11	2 081	11,40	0,6827	10,79	0,5181	10,55	0,0393
12	1 216	0,88	0,0681	0,92	0,0662	0,90	0,0341
13	1 844	12,85	0,9809	12,03	0,6817	10,97	0,0521
14	792	1,80	0,2872	1,82	0,2580	0,70	0,0532
15	607	0,20	0,0406	0,20	0,0403	0,29	0,0480
16	960	9,75	0,9567	10,43	0,6890	11,09	0,1267
17	2 278	11,97	0,5287	11,84	0,4425	12,28	0,0449
18	2 227	0,27	0,0128	0,26	0,0127	0,06	0,0148
19	504	0,69	0,1549	0,83	0,1450	1,72	0,0783
20	1 402	6,64	0,5047	6,27	0,4047	3,83	0,0518
21	1 667	5.61	0,3258	5,89	0,2824	5,08	0,0487
22	1 073	4,30	0,4345	4,11	0,3600	2,89	0,0497
23	478	0,00	0,0000	0,00	1,5526	0,29	0,0510
24	1 521	0,17	0,0136	0,17	0,0135	0,17	0,0233
99	1 364	10,98	0,8465	10,43	0,6014	9,65	0,0898

Cuadro A.3
Estimaciones directas, FH y *Census EB* de los ingresos medios, errores cuadráticos medios y coeficientes de variación estimados de cada estimador, para cada sección censal de Montevideo, para hombres

Sec	n_d	Directo			FH			*Census EB*		
		Est	var	cv	Est	mse	cv	Est	mse	cv
1	74	24 836,71	11 478 551,50	13,64	23 153,44	2 192 630,31	6,40	25 762,01	1 516 931,82	4,78
2	65	20 460,87	16 472 932,99	19,84	15 443,37	2 435 603,57	10,11	15 765,25	1 269 312,32	7,15
3	72	14 298,89	49 73 475,40	15,60	17 664,10	1 405 480,47	6,71	16 069,72	1 460 721,97	7,52
4	147	26 634,78	6 878 201,92	9,85	25 124,25	1 749 062,81	5,26	23 452,91	883 029,38	4,01
5	218	23 222,64	3 977 409,63	8,59	24 597,89	1 109 370,86	4,28	21 535,17	568 700,05	3,50
6	141	28 783,87	8 278 641,51	10,00	29 394,58	2 072 316,24	4,90	26 974,58	956 465,19	3,63
7	343	25 127,32	2 855 763,61	6,73	25 322,16	1 257 216,13	4,43	20 211,77	353 216,26	2,94
8	261	17 701,92	2 187 547,36	8,36	19 904,65	1 018 640,94	5,07	17 975,03	384 896,60	3,45
9	1032	11 475,41	190 257,67	3,80	11 517,36	174 746,99	3,63	10 253,44	84 239,24	2,83
10	2820	24 575,73	396 689,08	2,56	24 658,69	348 365,27	2,39	21 279,59	39 334,03	0,93
11	1882	11 079,86	99 418,16	2,85	10 975,09	95 447,04	2,81	9 008,44	42 747,93	2,30
12	1009	21 868,24	638 076,52	3,65	21 342,01	481 650,83	3,25	18 731,09	88 898,30	1,59
13	1712	10 897,03	106 703,48	3,00	10 766,13	103 066,30	2,98	8 566,47	44 008,29	2,45
14	641	22 605,70	1 090 803,69	4,62	22 636,34	664 159,72	3,60	19 332,12	177 272,29	2,18
15	509	28 797,75	2 175 732,47	5,12	27 945,27	1 339 582,93	4,14	23 540,89	232 540,09	2,05
16	894	8920,20	130 401,00	4,05	8 893,51	125 439,87	3,98	9 342,98	73 446,28	2,90
17	2095	8749,60	58 161,78	2,76	8 830,39	57 265,48	2,71	8 402,07	37 470,56	2,30
18	1723	38 931,89	1 332 962,25	2,97	38 347,54	1 019 441,37	2,63	33 874,56	93 681,13	0,90
19	417	17 855,77	1 230 524,64	6,21	17 640,61	755 965,09	4,93	15 893,35	237 964,08	3,07
20	1179	13 531,48	248 104,13	3,68	13 611,71	229 000,88	3,52	12 318,07	72 710,30	2,19
21	1498	11 147,80	132 574,18	3,27	11 290,17	125 020,79	3,13	11 380,39	65 707,49	2,25
22	929	15 394,11	397 876,11	4,10	15 137,23	349 742,04	3,91	13 156,00	92 731,14	2,31
23	392	27 941,71	3 640 071,50	6,83	28 804,90	1 294 705,53	3,95	23 483,98	321 582,84	2,41
24	1170	29 940,63	1 097 458,14	3,50	29 943,72	745 397,44	2,88	26 410,63	108 953,03	1,25
99	1241	10 230,59	130 610,91	3,53	10 227,46	124 563,28	3,45	9 379,01	67 878,14	2,78

Cuadro A.4
Estimaciones directas, FH y *Census EB* de la pobreza no extrema (en %), errores cuadráticos medios y coeficientes de variación estimados de cada estimador, para cada sección censal de Montevideo, para hombres

Sec	n_d	Directo		FH		*Census EB*	
		Est	var	Est	mse	Est	mse
1	74	0,00	0,0000	0,00	2,3526	0,24	0,4195
2	65	1,89	3,4599	4,06	1,4112	1,94	1,1930
3	72	1,10	1,1480	0,82	0,8380	1,90	0,7071
4	147	0,80	0,6193	0,30	0,5133	0,37	0,1798
5	218	0,68	0,4522	0,66	0,3936	0,57	0,1535
6	141	0,00	0,0000	0,00	2,3300	0,19	0,2299
7	343	1,39	0,4707	1,27	0,4145	0,73	0,1259
8	261	0,00	0,0000	1,89	2,0897	1,15	0,1482
9	1 032	5,32	0,5585	6,04	0,4740	7,40	0,1099
10	2 820	2,34	0,0959	2,38	0,0928	0,61	0,0233
11	1 882	10,23	0,7013	9,86	0,5710	10,38	0,0589
12	1 009	0,31	0,0312	0,34	0,0309	0,93	0,0605
13	1 712	12,81	1,0956	11,88	0,8397	11,89	0,0636
14	641	1,99	0,3835	2,09	0,3381	0,88	0,0724
15	509	0,53	0,1388	0,50	0,1337	0,36	0,0644
16	894	9,07	0,9783	9,60	0,7497	9,33	0,1292
17	2 095	12,09	0,5981	11,83	0,5141	12,47	0,0419
18	1 723	0,17	0,0134	0,16	0,0134	0,08	0,0199
19	417	0,83	0,2247	0,98	0,2099	1,72	0,1449
20	1 179	7,04	0,6212	6,60	0,5067	4,28	0,0596
21	1 498	6,36	0,4374	6,47	0,3789	5,45	0,0606
22	929	6,35	0,8152	5,60	0,6296	3,39	0,0734
23	392	0,71	0,2482	0,73	0,2305	0,39	0,0902
24	1 170	0,21	0,0209	0,22	0,0208	0,24	0,0355
99	1 241	10,76	0,9442	10,44	0,7215	9,38	0,0935

NACIONES UNIDAS

CEPAL

Serie Estudios Estadísticos

Números publicados

Un listado completo así como los archivos pdf están disponibles en

www.cepal.org/publicaciones

97. Desagregación de datos en encuestas de hogares: metodologías de estimación en áreas pequeñas, Isabel Molina, (LC/TS.2018/82/Rev.1), 2019.
96. ¿Cuál es el alcance de las transferencias no contributivas en América Latina?: discrepancias entre encuestas y registros, Pablo Villatoro, Simone Cecchini, (LC/TS.2018/46), junio de 2018.
95. Avances y desafíos de las cuentas económico-ambientales en América Latina y el Caribe, Franco Carvajal, (LC/TS.2017/148), enero de 2018.
94. La situación de las estadísticas, indicadores y cuentas ambientales en América Latina y el Caribe, (LC/TS.2017/135), diciembre 2017.
93. Indicadores no monetarios de carencias en las encuestas de los países de América Latina: disponibilidad, comparabilidad y pertinencia, Pablo Villatoro, (LC/TS.2017/130), diciembre de 2017.
92. Un índice de pobreza multidimensional para América Latina, María Emma Santos, Pablo Villatoro, Xavier Mancero Pascual Gerstenfeld, (LC/L.4129), diciembre de 2015.
91. Ajuste de los ingresos de las encuestas a las Cuentas Nacionales. Una revisión de la literatura, Pablo Villatoro, (LC/L.4002), abril de 2015.
90. La evolución del ingreso de los hogares en América Latina durante el período 1990-2008 ¿Ha sido favorable a los pobres?, Fernando Medina y Marco Galván, (LC/L.3975) marzo de 2015.
89. ¿Qué es el crecimiento propobre?, Fundamentos teóricos y metodologías para su medición, Fernando Medina y Marco Galván, (LC/L.3883), agosto de 2014.
88. Cuentas satélite y cuentas de salud: un análisis comparativo, Federico Dorin, Salvador Marconi y Rafael Urriola (LC/L.3865), julio de 2014.
87. Sensibilidad de los índices de pobreza a los cambios en el ingreso y la desigualdad: lecciones para el diseño de políticas en América Latina, 1997-2008, Fernando Medina y Marco Galván, (LC/L.3823), julio de 2014.
86. Una propuesta regional de estrategia de implementación del Sistema de Cuentas Ambientales Económicas (SCAE) 2012 en América Latina (LC/L.3786), diciembre de 2013.
85. América Latina y el Caribe: estimación de las series del PIB y del consumo de los hogares en PPA. Un ejercicio preliminar para el período 2000-2011, Hernán Epstein y Salvador Marconi, (LC/L.3781), enero de 2014.
84. El Sistema de Cuentas Ambientales y Económicas (SCAE) 2012: fundamentos conceptuales para su implementación (LC/L.3752), noviembre 2013.
83. Consumo efectivo de los hogares en salud: resultado de estudios piloto en seis países de América Latina, David Debrott Sánchez, (LC/L.3751), abril de 2014.
82. Crecimiento económico, pobreza y distribución del ingreso: fundamentos teóricos y evidencia empírica para América Latina 1997-207 (LC/L.3689), Fernando Medina, Marco Galván, marzo de 2014.

ESTUDIOS ESTADÍSTICOS

Números publicados:

97. Desagregación de datos en encuestas de hogares.
Metodologías de estimación en áreas pequeñas
Isabel Molina

96. ¿Cuál es el alcance de las transferencias no contributivas en América Latina?
Discrepancias entre encuestas y registros
Pablo Villatoro
Simone Cecchini

95. Avances y desafíos de las cuentas económico-ambientales en América Latina y el Caribe
Franco Carvajal

94. La situación de las estadísticas, indicadores y cuentas ambientales en América Latina y el Caribe

CEPAL
Comisión Económica para América Latina y el Caribe (CEPAL)
Economic Commission for Latin America and the Caribbean (ECLAC)
www.cepal.org