FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Evaluación del Mejor Predictor Lineal Insesgado como estimador de un total

Una primera aproximación a la inferencia basada en modelos

Leandro Emiliano González Sosa

Ana Vignolo Cortabarría

Trabajo Final de Grado presentado para la Licenciatura en Estadística, Facultad de Ciencias Económicas y de Administración de la Universidad de la República, como parte de los requisitos necesarios para la obtención del título de Licenciado en Estadística.

Tutores:

Ph.D. Prof. Juan José Goyeneche

M.Sc. Prof. Adj. María Eugenia Riaño

Montevideo – Uruguay

Junio de 2024

González Sosa, Leandro Emiliano; Vignolo Cortabarría, Ana

Evaluación del Mejor Predictor Lineal Insesgado como estimador de un total / Leandro Emiliano González Sosa; Ana Vignolo Cortabarría. - Montevideo: Universidad de la República, Facultad de Ciencias Económicas y de Administración, 2024.

XVI, 125 p. 29.7 cm.

Tutores:

Juan José Goyeneche

María Eugenia Riaño

Trabajo Final de Grado – Universidad de la República, Licenciatura en Estadística, 2024.

Referencias bibliográficas: p. 118 – 121.

1. Muestreo de poblaciones finitas, 2. Inferencia basada en modelos, 3. Modelo de población homogénea, 4. Modelo de población lineal general, 5. Mejor predictor lineal insesgado, 6. Estimador de regresión. I. Goyeneche, Juan José; Riaño, María Eugenia. II. Universidad de la República, Licenciatura en Estadística. III. Título.

INTEGRANTES DEL TRIBUNAL

M.Sc. Prof. Adj. Juan Pablo Ferreira

M.Sc. Prof. Miguel Galmés

Ph.D. Prof. Juan José Goyeneche

Montevideo – Uruguay
Junio de 2024

Agradecimientos

En primer lugar, queremos agradecer a nuestros tutores, Juan José Goyeneche y María Eugenia Riaño, quienes han sido extremadamente generosos tanto con su tiempo como con sus conocimientos. Su calidez y buen humor nos han impulsado a seguir adelante en los momentos de más frustración y dudas. Realmente, este trabajo no hubiera sido posible sin su guía.

Agradecemos también a nuestros compañeros de la Licenciatura y del IESTA, por su interés y palabras de aliento. Ellos nos han mostrado el verdadero significado del compañerismo y el trabajo en equipo.

Leandro:

Quiero agradecer a todos los que de una u otra manera han formado parte de esta etapa.

En primer lugar, agradezco a mi familia y amigos por apoyarme durante todos estos años, tanto en las buenas, compartiendo mi alegría, como en las malas, alentándome a seguir.

También agradezco a todos los docentes y compañeros de la Licenciatura con los que he compartido en algún momento, seguramente de todos me fui nutriendo para seguir avanzando.

Ana:

Muchas gracias a todos los que me han acompañado durante este proceso, este logro les pertenece también a ellos.

Primero, quiero agradecer a los docentes de la Licenciatura, cuya preocupación por sus estudiantes es más que evidente. En especial, muchas gracias a la Prof. Silvia Rodríguez, quien es un referente importantísimo para mí. Su disponibilidad y sus sabios consejos han sido fundamentales.

Muchísimas gracias a mi familia, cuyo apoyo incondicional está a la base de este y todos mis logros. A mi padre, quien siempre tiene una palabra amable,

y a mi madre, mi editora de cabecera. Gracias a mi hermano Andrés, no sólo por toda su ayuda a la hora de compaginar este trabajo, sino también por impulsarme siempre a mejorar tanto a nivel profesional como personal.

Por último, gracias a Maximiliano, cuya serenidad y paciencia todo este tiempo han sido de enorme ayuda. Gracias por acompañarme siempre.

The best books, he perceived, are those that tell you what you know already.

— George Orwell
1984

ABSTRACT

Within survey analysis and sampling theory, there are two main paradigms that rule the estimation process of a population parameter. These are design-based and model-based inference. Because of its unbiasedness under ideal conditions, design-based inference has been favoured over its model-based counterpart. In contrast, the validity of model-based inference hinges on whether the model used to predict non-observed values of the target variable is appropiate or not. Despite its fragility (which is a result of the assumed model), the model-based approach has several advantages. In specific, if models are adequately selected and fitted, the resulting estimations can be much more precise than those obtained through the design-based paradigm. Therefore, evaluating the quality of model-based estimations in various scenarios is of special relevance. This project assesses the performance of two basic models, which were applied to two types of populations. First, a known superpopulation model was used to simulate data. Second, a real-world population was considered. In this case, the model that "generated" each observation was unknown. Furthermore, in order to analyse the effect of using samples that do not "replicate" the distribution of the target variable in the population, different sampling designs were considered. In line with existing literature, the results suggest that although "imperfect" samples can lead to biases, the most important factor in achieving adequate model-based estimations is the appropriate specification of the superpopulation model.

Keywords:
Finite population sampling, Model-based inference, Homogeneous Population Model, General Linear Population Model, Best Unbiased Linear Predictor, Regression estimator.

RESUMEN

Dentro del muestreo de poblaciones finitas, existen dos grandes paradigmas que rigen el proceso de estimación de un parámetro poblacional: la inferencia basada en el diseño y la basada en modelos. Dada su insesgadez en contextos ideales, la inferencia basada en el diseño ha sido mucho más usada que la basada en modelos. En contraposición, la validez de la inferencia basada en modelos está condicionada a que el modelo utilizado para predecir los valores no observados de la variable de interés en la población sea apropiado. A pesar de su fragilidad, producto del modelo asumido, el paradigma basado en modelos tiene varias ventajas. En particular, si el modelo es seleccionado y ajustado adecuadamente, las estimaciones tienen el potencial de ser mucho más precisas que las proporcionadas por el enfoque basado en el diseño. Por lo tanto, es relevante evaluar la calidad de las estimaciones basadas en modelos bajo diferentes condiciones. En este trabajo, se estudió el desempeño de dos modelos sencillos para dos tipos de poblaciones. Por un lado, se trabajó con datos simulados a partir de un modelo superpoblacional conocido y, por otro, se utilizó una población real para la que se desconoce el modelo que la generó. Asimismo, se varió el diseño muestral empleado, de forma de analizar el efecto de usar muestras "que no replican" la distribución de la variable de interés en la población. En línea con la literatura existente, los resultados sugieren que si bien pueden generarse sesgos si se recurre a muestras "imperfectas", el factor más importante para que el enfoque *model-based* arroje buenas estimaciones es la adecuada especificación del modelo superpoblacional.

Palabras clave:
Muestreo de poblaciones finitas, Inferencia basada en modelos, Modelo de población homogénea, Modelo de población lineal general, Mejor predictor lineal insesgado, Estimador de regresión.

Tabla de contenidos

Lista de figuras

Lista de tablas

Capítulo 1

Introducción

1.1. Introducción y motivación inicial

El muestreo puede entenderse como un conjunto de métodos que permite estimar parámetros asociados a una población a partir de la inspección de un subconjunto de sus elementos al que se le denomina "muestra". En la actualidad, su uso es extremadamente extendido tanto en ámbitos académicos como en organismos públicos y privados. En general, el muestreo resulta preferible a la enumeración completa (llamada "censo") de los elementos de la población debido a que es relativamente menos costosa, requiere menos tiempo y puede incluso arrojar estimaciones más precisas (Särndal et al. 1992). Sin embargo, su desarrollo sistemático comenzó recién a principios del siglo XX, cuando la comunidad científica llegó a un consenso sobre la validez de extraer conclusiones para una población a partir de la observación parcial de sus unidades (Brewer y Gregoire, 2009; Chambers y Clark, 2012).

A diferencia de otras ramas de la Estadística, en muestreo el proceso de inferencia puede variar sustancialmente dependiendo de cuál se considere que sea la fuente de aleatoriedad en el proceso de estimación. A grandes rasgos, existen dos tipos de inferencia: basada en el diseño y basada en modelos (Dever y Valliant, 2018).

En la inferencia basada en el diseño, se considera que la aleatoriedad viene dada por el mecanismo de selección de la muestra. En consecuencia, la distribución de los estimadores viene inducida por el diseño muestral. De esta manera, se analiza la probabilidad de que el estimador tome cada posible valor en su recorrido dado el conjunto de muestras que pueden ser seleccionadas

bajo el diseño muestral utilizado. Como se detallará en el siguiente capítulo, la principal ventaja de la inferencia basada en el diseño es el hecho de que, en condiciones ideales en las que no existen errores ajenos al muestreo, arroja estimaciones insesgadas cuya validez no está condicionada al cumplimiento de ciertos supuestos. Esta característica resulta sumamente atractiva y ha contribuido a establecer el predominio del paradigma basado en el diseño por sobre la inferencia basada en modelos.

Un caso especial del enfoque basado en el diseño es la llamada inferencia asistida por modelos. En este caso, si bien se asume un modelo para mejorar la precisión de las estimaciones, la aleatoriedad sigue proviniendo mayoritariamente del diseño, con lo cual se mantiene el insesgamiento aproximado de las mismas. Así, se supone que el modelo "describe" adecuadamente a la población, pero no que la "genera".

Por el contrario, en la inferencia basada en modelos sí se asume la existencia de un modelo superpoblacional desconocido que "genera" los elementos de una población finita. Es decir, que los valores de la variable de interés constituyen realizaciones de un conjunto de variables aleatorias cuya distribución se rige por un cierto modelo estadístico desconocido. En este caso, entonces, la fuente de aleatoriedad es la propia variable de interés. Por lo tanto, para estimar un cierto parámetro poblacional que sea función de la variable de interés, será necesario predecir el valor que toma dicha variable para cada elemento de la población no observado. A su vez, para obtener las predicciones, se deberá estimar el modelo superpoblacional en base a una muestra. Si el modelo especificado posee variables de entrada, esto requerirá contar con información auxiliar a nivel de marco, elemento a elemento, y no bastará con conocer sus totales poblacionales.

Dado que la estimación final será una función de las predicciones arrojadas por el modelo para cada elemento de la población no observado, si existen errores de especificación, se producirá un sesgo potencialmente grande. Por consiguiente, la validez de la inferencia está condicionada a que la estimación del modelo superpoblacional, por definición desconocido, sea adecuada. En este sentido, con frecuencia se dice que la inferencia basda en modelos es "modelo dependiente". Esta cualidad constituye una gran debilidad y es el principal motivo por el cual suele preferirse el enfoque basado en el diseño.

A pesar de este problema, la inferencia basada en modelos presenta varias ventajas. En primer lugar, si el modelo es estimado adecuadamente, las estimaciones resultantes pueden ser mucho extremadamente precisas. Además, no

considera el diseño muestral, con lo cual (al menos en principio) puede ser aplicado a datos cuyo mecanismo de selección es desconocido, y a muestras por conveniencia. Esto determina que, por lo general, utilizar el paradigma basado en modelos sea mucho menos costoso que recurrir a la inferencia basada en el diseño. Sin embargo, es importante señalar que, en este caso, la validez de las estimaciones dependerá de que el modelo estimado a partir de la muestra "coincida" con el que rige para el resto de la población. De lo contrario, es probable que se incurra en sesgos. Para más detalles sobre las características de cada enfoque, así como sobre sus ventajas y desventajas, ver Capítulo 2.

Si bien distintos autores han destacado las ventajas de la inferencia basada en modelos, la mayoría admite que la "modelo-dependencia" es un problema difícil de superar (ver, por ejemplo, Särndal et al. 1978; Smith, 1994; Little, 2004). En este contexto, el objetivo de este trabajo es el de determinar en qué medida se ven afectadas las estimaciones basadas en modelos si existen errores de especificación del modelo y/o las muestras usadas tienen problemas que las hacen "imperfectas". Para ello, se llevó a cabo una serie de aplicaciones prácticas con dos grupos de poblaciones completamente conocidas. En cada caso, se presentan distintos indicadores que intentan cuantificar la "fragilidad" de las estimaciones.

El documento se organiza como sigue. En el presente capítulo, se introduce el problema de estudio, así como los objetivos y las hipótesis consideradas. En el Capítulo 2, se analizan en profundidad los distintos paradigmas de la inferencia en muestreo de poblaciones finitas, se describen sus ventajas y desventajas y se presentan algunos antecedentes de interés. A su vez, en el Capítulo 3, se desarrolla el marco teórico y se demuestran algunos resultados relevantes. Por su parte, en el Capítulo 4, se presentan los datos utilizados en las aplicaciones prácticas, y en el Capítulo 5, se detalla la metodología empleada. En el Capítulo 6, se describen los resultados. En el Capítulo 7 se realizan algunas discusiones e interpretaciones. Finalmente, en el Capítulo 8, se presentan las principales conclusiones, se enumeran las limitaciones del trabajo y se proponen algunas líneas de investigación futuras.

1.2. Presentación del problema

A pesar de que el paradigma basado en modelos permite estimar cualquier parámetro poblacional que sea función del valor de la variable de interés en cada

elemento de la población, este trabajo se limitó a estudiar parámetros sencillos, como ser, un total. Para ello, se consideraron solamente dos modelos: el de población homogénea y el de población lineal general (Chambers y Clark, 2012). Como se verá, el primero no hace uso de ningún tipo de información auxiliar, de manera que la predicción del valor que toma la variable de interés para cada elemento de la población se reduce a la media muestral de los datos considerados. En cambio, el modelo de población lineal general sí permite considerar distintas variables auxiliares mediante un modelo lineal. Para simplificar el análisis, en este trabajo se utilizaron únicamente regresiones lineales simples y se asumió que se cumple el supuesto de homocedasticidad de los errores.

Con respecto a los datos, se usaron dos tipos de poblaciones. En primer lugar, se trabajó con poblaciones simuladas a partir de un modelo superpoblacional completamente conocido. Se trata de un ejemplo teórico en el que no sólo se conoce la realización de la variable de interés en cada elemento de la población, sino que también se sabe cuál fue el mecanismo a través del cual dicha población fue generada. De esta manera, en este primer caso, se supo a ciencia cierta si las estimaciones fueron obtenidas mediante un modelo correctamente especificado o no. Para garantizar la robustez de los resultados, se consideraron 5.000 réplicas del modelo superpoblacional escogido.

En segundo lugar, se trabajó con la población MU281, que recoge diversos indicadores socioeconómicos para 281 municipios suecos. La población MU281 fue originalmente presentada por Särndal et al. (1992) y con frecuencia es utilizado en muestreo como un ejemplo "de laboratorio" para evaluar el desempeño de distintos estimadores y técnicas. Si bien se conoce el valor que toman los indicadores en cada municipio, la población MU281 constituye una única realización de un modelo superpoblacional desconocido. En consecuencia, a diferencia de las poblaciones simuladas, en este caso no es posible determinar si el modelo escogido es adecuado, solamente es posible analizar la validez de sus predicciones. En este sentido, la población MU281 constituye un caso más realista que las poblaciones simuladas.

Para cada población, los dos modelos considerados fueron estimados a partir de muestras obtenidas bajo distintas condiciones. En primera instancia, se consideró un muestreo aleatorio simple sin reposición, también conocido como SI (Särndal et al. 1992). Este diseño asigna a todos los elementos de la población la misma probabilidad de inclusión en la muestra. Por consiguiente, en promedio, es esperable que la muestra refleje la verdadera variabilidad de la variable

de interés. Como se detalla en el Capítulo 2, esto implica que se cumple la denominada condición de "ignorabilidad", según la cual el mecanismo mediante el que se recolectaron los datos no distorsiona la estimación de los parámetros del modelo. Este primer caso constituye, entonces, una situación "ideal", en la que siempre y cuando el modelo esté razonablemente bien especificado, se obtendrán buenas estimaciones.

Sin embargo, en la práctica, las muestras utilizadas son imperfectas y no necesariamente son seleccionadas mediante un diseño ignorable. Por este motivo, se buscó obtener una noción sobre posibles variaciones en la utilidad del modelo bajo distintos tipos de información. Como primera aproximación a estos problemas, se consideraron dos situaciones.

Por un lado, se volvieron a extraer muestras simples, pero se las truncó superiormente en términos de la variable de interés con el fin de replicar el efecto de patrones de no respuesta que dependen de la variable estudiada, a veces conocidos como No Respuesta No Ignorable (Rubin, 1976). En estos casos, la probabilidad de que un individuo responda depende de la variable de interés, con lo cual, si no se la toma en cuenta en la etapa de estimación, los resultados serán sesgados. Se optó por trabajar con este tipo de no respuesta en particular debido a que se da con bastante frecuencia en encuestas socioeconómicas (por ejemplo, los hogares más ricos suelen negarse a declarar sus ingresos en mayor proporción que el resto). Por lo tanto, es importante analizar en qué medida las estimaciones basadas en modelos pueden verse sesgadas en estos casos.

Por otro lado, se analizó el efecto de utilizar un diseño muestral no ignorable para estimar un modelo superpoblacional. Para ello, se recurrió al denominado diseño con probabilidades de inclusión proporcionales al tamaño (Särndal et al. 1992). Éste se nota como πps y se caracteriza por definir, para cada elemento de la población, una probabilidad de inclusión en la muestra proporcional al valor de una variable auxiliar. En particular, en este trabajo, se utilizaron probabilidades proporcionales al inverso de la variable auxiliar considerada (para más detalles, ver Capítulo 5). Es decir, que a mayor valor de la variable original, menor fue la probabilidad de que un elemento perteneciera a la muestra. Así, en promedio, se obtuvieron muestras "distorsionadas" en términos de la variable auxiliar. Dado que es esperable que esta última esté correlacionada con la variable de interés, es probable que las estimaciones *model-based* se vean sesgadas.

Para evaluar el desempeño de las predicciones obtenidas en cada caso, se

consideraron tres indicadores básicos: el error de predicción, la estimación de la varianza del estimador y la cobertura estimada de los intervalos de confianza. A pesar de su simplicidad, estas medidas son sumamente útiles y permiten identificar potenciales sesgos y/o problemas de precisión.

1.3. Objetivos e hipótesis

Como síntesis de lo presentado en la sección anterior, a continuación se enumeran los objetivos del trabajo, así como las hipótesis iniciales.

Objetivo general: Evaluar la calidad de las estimaciones *model-based* de totales bajo distintos diseños muestrales y para dos modelos básicos (homogéneo y lineal general) aplicados a poblaciones reales y simuladas.

Objetivos específicos:

1. Indagar acerca del efecto de incurrir en errores de especificación del modelo utilizado para generar las predicciones.

2. Analizar el efecto de ignorar el mecanismo de selección de la muestra cuando el mismo no es ignorable.

3. Comparar el desempeño de la inferencia basada en modelos con la asistida por modelos cuando ambos enfoques arrojen distintos resultados. Para ello, sólo se considera el estimador de regresión dentro del conjunto de los estimadores *model-assisted*.

Hipótesis:

1. Dado que el diseño SI es ignorable, las estimaciones obtenidas bajo dicho mecanismo serán aproximadamente insesgadas para ambos modelos considerados. En contraste, las muestras truncadas y las obtenidas mediante diseños no ignorables conducirán a estimaciones fuertemente sesgadas.

2. Como el modelo de población lineal general hace uso de la información auxiliar disponible, las predicciones resultantes serán más precisas que en el caso de un modelo homogéneo.

3. Las estimaciones obtenidas para las poblaciones simuladas bajo el modelo de población lineal general serán superiores en términos de precisión e

insesgamiento que para la población MU281. Esto se debe a que sólo en el primer caso será posible garantizar que no existan errores de especificación del modelo superpoblacional.

4. En los casos en que las estimaciones basadas en modelos difieran de las asistidas por modelos, estas últimas presentarán menores sesgos promedio por ser "modelo-independientes". Es decir, que la validez de la inferencia no está condicionada a que el modelo seleccionado sea correcto.

Capítulo 2

Paradigmas de la inferencia en muestreo: caracterización y antecedentes

Como se anticipó en el Capítulo 1, en la etapa de estimación de cantidades descriptivas desconocidas de poblaciones finitas, existen diferentes paradigmas de la inferencia. En otras palabras, existen varias concepciones de la aleatoriedad en el proceso de inferencia. En términos generales, la inferencia puede ser de dos tipos: basada en el diseño o basada en modelos. A su vez, dentro del primer paradigma, se encuentra el caso particular de la inferencia asistida por modelos. Por otra parte, existen algunas técnicas de inferencia "híbridas" (ver, por ejemplo, Ståhl et al. 2016), las cuales combinan aspectos de ambos enfoques.

A continuación, se describen las principales características de estos tres enfoques y se comparan sus ventajas y desventajas. Asimismo, se presentan algunos ejemplos de la aplicación del enfoque basado en modelos en diferentes campos de estudio.

2.1. Predominancia de la inferencia basada en el diseño

En el marco del muestreo de poblaciones finitas, la inferencia basada en el diseño ha sido el enfoque predominante desde que Neyman (1934) publicó su artículo seminal, en el que criticó duramente el uso de muestras balanceadas. En este sentido, múltiples autores consideran que este estudio desincentivó

fuertemente el uso de muestras no aleatorias (ver, por ejemplo, Gregoire, 1998; Brewer y Gregoire, 2009; Ståhl et al. 2016).

La característica distintiva de la inferencia basada en el diseño es que la aleatoriedad viene dada únicamente por el mecanismo de selección de la muestra. De esta manera, autores como Brus y De Gruijter (1997) y Gregoire (1998) consideran que la aleatoriedad es introducida "artificialmente" por el investigador al definir el diseño muestral. Si se conoce la forma en que se seleccionan las observaciones, la probabilidad de extraer cada una de las posibles muestras es conocida *a priori*. Por lo tanto, la "distribución de referencia" bajo este enfoque es la del diseño. A su vez, dicha distribución determina las probabilidades de inclusión de cada elemento de la población en la muestra. Así, la probabilidad de que cada unidad sea seleccionada será igual a la suma de las probabilidades de todas las muestras que la contienen. Cabe señalar que para que un diseño sea aleatorio, todos los elementos de la población deben poseer propiedades de inclusión positivas.

Bajo la inferencia basada en el diseño, el valor que toma la variable de interés en cada elemento de la población se trata como un conjunto de valores fijos (aunque desconocidos) sobre cuya distribución no se realizan supuestos. En consecuencia, cualquier cantidad que sea función de los datos (tales como un total) también será un parámetro fijo que deberá ser estimado.

Estas cantidades pueden ser estimadas insesgadamente mediante funciones de las observaciones muestrales ponderadas por el inverso de sus probabilidades de inclusión. En este sentido, la inferencia basada en el diseño "pura" utiliza los ponderadores para corregir la desproporcionalidad existente en los datos muestrales con respecto a la población, la cual se origina en las probabilidades de inclusión (Pfeffermann, 1993; Natarajan et al. 2008). De esta forma, las estimaciones sólo incorporan información auxiliar a través del diseño muestral (por ejemplo, estratificando la población o utilizando probabilidades de inclusión proporcionales al tamaño de una covariable).

Dado que las probabilidades de inclusión vienen dadas por el diseño muestral por definición aleatorio, los estimadores también lo serán y su valor variará dependiendo de cuál haya sido la muestra extraída. Por consiguiente, sus propiedades estadísticas deberán evaluarse con respecto a la probabilidad de extraer cada muestra. Esta "distribución de referencia" cambiará con el diseño muestral, por lo que bajo esquemas de selección distintos las propiedades de un mismo estimador pueden diferir (Brewer y Gregoire, 2009).

En suma, la característica fundamental del enfoque basado en el diseño es el hecho de que la base de la inferencia está en el diseño muestral y es prácticamente independiente de supuestos sobre la estructura de la población en términos de la variable de interés (en términos estrictos, sólo se asume normalidad a la hora de construir intervalos de confianza, pero esto es razonable para muestras "grandes"). Además, no se requiere ajustar ningún modelo para obtener las estimaciones, con lo cual Brewer y Gregoire (2009) califican a este enfoque como no paramétrico. Por estos motivos, autores como Smith (1994) afirman que se trata de un enfoque robusto "por construcción".

2.2. Resurgimiento de la inferencia basada en modelos

La robustez de la inferencia basada en el diseño constituye su principal atractivo, en especial cuando se requieren resultados confiables como por ejemplo en estadísticas oficiales, para las que el diseño muestral continúa siendo la base de la inferencia que garantiza el insesgamiento de las estimaciones.

Sin embargo, la inferencia basada en el diseño tiene algunas desventajas. En primer lugar, al menos en su versión "pura"(no asistida por modelos), no explota información auxiliar no contenida en el diseño muestral. Esto implica que las estimaciones sean potencialmente menos eficientes de lo que podrían ser, dada la información disponible.

En segundo lugar, Godambe (1955) demostró teóricamente que no existe un estimador basado en el diseño lineal, insesgado y de mínima varianza (BLUE, por su sigla en inglés). Así, si bien existen múltiples estimadores insesgados con respecto al diseño muestral (entre los que se destaca el estimador de Horvitz-Thompson), no es posible obtener un estimador óptimo en términos de su dispersión para todos los diseños muestrales.

En tercer lugar, los estimadores basados en el diseño no permiten predecir el valor que tomará la variable de interés para un cierto individuo no observado en la población, sino que en general se enfocan en estimar parámetros de toda la población como son totales y promedios (Brus y De Gruijter, 1997).

En cuarto lugar, Little (2004) señala que, en general, para que las estimaciones basadas en el diseño sean confiables, la muestra utilizada debe ser relativamente grande. En este sentido, la inferencia basada en el diseño consti-

tuye un enfoque "asintótico" que no estipula claramente cómo proceder frente a muestras pequeñas.

En quinto lugar, cabe destacar que el enfoque basado en el diseño no es aplicable cuando la muestra está "corrompida" por errores no muestrales tales como la no respuesta o errores de medición. En estas situaciones, los ponderadores dados por el inverso de las probabilidades de inclusión no reflejan adecuadamente a cuántos elementos de la población refleja cada unidad muestreada.

En sexto lugar, cabe destacar que el enfoque basado en el diseño se diferencia del resto de la Estadística debido a que trabaja con la distribución que surge de considerar la probabilidad asociada a cada posible muestra bajo un cierto diseño muestral, y no se considera la distribución de la variable de interés en sí misma. Si bien esto no es un problema, ha contribuido a que el muestreo se desarrollara en forma bastante aislada del resto de las ramas de la Estadística (Little, 2004).

2.3. Caracterización de la inferencia basada en modelos

Múltiples autores han buscado sistematizar la literatura existente sobre el enfoque basado en modelos y compararlo con el paradigma basado en el diseño (ver, por ejemplo, Särndal et al. 1978; Royall, 1992; Smith, 1994 y Little, 2004). El rasgo más distintivo de la inferencia basada en modelos es el hecho de que se supone que la aleatoriedad viene dada por la propia variable de interés. En otras palabras, los valores que toma la variable de interés para cada elemento de la población se conciben como realizaciones de variables aleatorias cuya distribución se rige por un cierto modelo. Aquí, la "distribución de referencia" será la de la variable de interés en la población, y por lo tanto las propiedades de los estimadores estarán condicionadas a ella (Gregoire, 1998).

Dado que el valor de la variable de interés en la población se concibe como una serie de variables aleatorias, cualquier cantidad que sea función de ellas como un total también lo será y su valor deberá ser predicho ajustando un modelo a partir de una muestra. Bajo el enfoque basado en modelos, entonces, la inferencia constituye un problema de predicción de elementos no observados de la población (Royall, 1992). Estrictamente hablando, se trata de predicciones y no estimaciones debido a que la cantidad considerada no se entiende como un

parámetro fijo. Sin embargo, por simplicidad, en este trabajo se utilizan ambos términos indistintamente.

El modelo a través del cual se supone que se generan los elementos de la población se denomina "modelo superpoblacional" y puede ser interpretado de dos formas diferentes (Aubry y Francesiaz, 2022). Por un lado, puede entenderse como un proceso estocástico natural. En este caso, es de interés estimar los parámetros superpoblacionales para entender el proceso generador de poblaciones en sí mismo. Así, la población finita bajo estudio constituye una única realización dentro de las infinitas que puede generar el modelo.

Por otra parte, los modelos superpoblacionales pueden entenderse como abstracciones matemáticas útiles para derivar la forma de los estimadores óptimos de una cantidad poblacional. En este sentido, un modelo es cualquier representación de la realidad que incorpore conocimientos previos del comportamiento de la variable de interés en la población. De esta manera, se trata de un concepto sumamente amplio que abarca desde regresiones lineales hasta modelos bayesianos (Little, 2004).

Si bien *a priori* no es necesario que los modelos incorporen variables independientes, gran parte de su atractivo es la capacidad de explotar información auxiliar (Särndal et al. 1978). Así, los modelos logran formalizar el vínculo entre la variable de interés y las covariables (Valliant, 2009). Sin embargo, la validez de las predicciones está condicionada a que el modelo esté adecuadamente especificado y a que efectivamente el comportamiento de la variable de interés pueda ser explicado a partir de la información auxiliar disponible. De lo contrario, pueden producirse sesgos sustanciales. Esto es especialmente relevante si se tiene en cuenta que no existen procedimientos rigurosos de selección de variables en el proceso de identificación de modelos superpoblacionales, más allá del coeficiente de determinación (R^2) para las observaciones de la muestra (Nascimento Silva y Skinner, 1997).

Los modelos utilizados para predicción se estiman a partir de la muestra. Para que la estimación de sus parámetros no se vea sesgada, es necesario utilizar un diseño muestral que refleje el rango de valores que toman las variables auxiliares en la población. Dado que la base de la inferencia está en la aleatoriedad de la variable de interés y no en el diseño, la inferencia basada en modelos no necesariamente requiere el uso de muestras probabilísticas. Bajo ciertos modelos, las muestras no aleatorias y balanceadas arrojan estimaciones más precisas (Royall, 1992; Nedyalkova y Tillé, 2008). Sin embargo, Valliant

(2009) recomienda utilizar esquemas de muestreo aleatorios porque éstos evitan posibles preconceptos errados del investigador y en promedio generan muestras balanceadas. Así, se busca que la muestra se rija por el mismo modelo que la población.

Si bien el diseño puede no ser probabilístico, sí es necesario que sea no informativo, en ocasiones también denominado exógeno (Smith, 1994; Gregoire, 1998; Lumley y Scott, 2017). Es decir, que el mecanismo mediante el cual se eligen los elementos debe ser una función de la información disponible (incluyendo variables auxiliares), y no de la variable de interés desconocida. Analíticamente, esto equivale a que la distribución conjunta de las variables binarias que establecen si un cierto elemento de la población fue incluido en la muestra no varíe al condicionarla a la variable de interés (Little, 2004). A su vez, esto implica que las probabilidades de inclusión no dependerán de dicha variable.

Por otra parte, como en el enfoque basado en modelos no se ponderan las observaciones, es necesario que la muestra refleje la dispersión de la variable de interés de la población. Los diseños que cumplen esto se denominan ignorables (Pfeffermann, 1993). Esto ocurre cuando el mecanismo de selección de la muestra no brinda ninguna información adicional, más allá de la dada por las variables incorporadas en el modelo. Es el caso, por ejemplo, para muestras simples. No obstante, para diseños complejos, suele ser difícil verificar el cumplimiento de las condiciones necesarias para la "ignorabilidad". En el contexto de modelos lineales, por ejemplo, si el diseño contiene información acerca del comportamiento de la variable de interés que no esté recogido en las variables explicativas del modelo, las estimaciones pueden ser sumamente sesgadas (Nordberg, 1989).

En suma, es claro que para que la inferencia basada en modelos funcione bien, es necesario que el modelo sea robusto a errores de especificación y que el diseño muestral permita que las observaciones reflejen el comportamiento de la variable de interés en la población. Si ambas condiciones se cumplen simultáneamente, la estimación de los parámetros superpoblacionales será razonablemente buena, con lo cual la predicción del valor de la variable de interés para cada elemento de la población no muestreado también lo será.

Para predecir la cantidad de interés, se utilizan las predicciones de los individuos no observados obtenidas a partir del modelo estimado. Por este motivo, Brewer y Gregoire (2009) destacan que se trata de un enfoque paramétrico.

Por otra parte, cabe señalar que si se utiliza un modelo que incorpora información auxiliar, para calcular las predicciones individuales se requiere conocer el valor de cada covariable a nivel de marco muestral. Es decir, en contraste con otros enfoques como la inferencia asistida por modelos, no alcanza con tener información auxiliar agregada.

A diferencia de la inferencia basada en el diseño, bajo el enfoque basado en modelos sí resulta posible deducir la forma de los predictores óptimos para cada modelo: lineales, insesgados y de mínima varianza (BLUP por su sigla en inglés). El predictor BLUP será aquel que minimice la dispersión en las estimaciones de los parámetros del modelo, que son los que introducen variabilidad en cada predicción (Lohr, 2009).

Con respecto al cálculo de la varianza, Valliant (2009) destaca que, dado que el total y su predictor constituyen variables aleatorias independientes entre sí, la varianza del error de predicción (por definición, la diferencia entre ambas variables) será igual a la suma de las dos varianzas. Sin embargo, la varianza del total suele ser muy pequeña en comparación a la varianza del predictor. En consecuencia, en la práctica la estimación de la varianza del error de predicción es equivalente a la varianza del predictor.

De esta manera, la optimalidad de los predictores se evalúa con respecto al modelo y no al diseño muestral (Nedyalkova y Tillé, 2008). Sin embargo, de acuerdo con autores como Brus y De Gruijter (1997) y Ståhl et al. (2016), esto no resulta particularmente útil dado que depende de que el modelo haya sido correctamente identificado. En caso contrario, la predicción de la cantidad de interés puede estar sumamente sesgada en comparación con su valor real.

Aubry y Francesiaz (2022) clasifican a los predictores para totales en dos tipos: proyectivos y predictivos. Aunque se refieren a la inferencia basada en modelos, los denominan "estimadores". Así, los estimadores son "proyectivos" cuando se calculan como la suma de las predicciones del modelo para cada elemento de la población. Por su parte, los estimadores "predictivos" son aquellos que se obtienen mediante la suma del total de la variable de interés en la muestra y de las predicciones para los elementos de la población no observados. En general, los estimadores predictivos son preferibles a los proyectivos debido a que son consistentes. Es decir, que cuando se censa a toda la población, la estimación coincidirá con el verdadero total.

En general, los predictores BLUP pertenecen al grupo de los estimadores predictivos (Royall, 1976; Smith, 1994). De todas formas, de acuerdo con Ståhl

et al. (2016), si la población es relativamente grande con respecto a la muestra, no deberían existir grandes diferencias entre ambos.

Queda claro, entonces, que la gran debilidad de la inferencia basada en modelos es el ser "modelo dependiente" (Valliant, 2009). En la medida que el diseño muestral no es tenido en cuenta, la validez de la predicción de una cierta cantidad dependerá de que el modelo superpoblacional haya sido estimado razonablemente bien. Si esta condición no se cumple, es probable que se incurra en grandes sesgos. De acuerdo con Smith (1994), este problema es particularmente grave en el marco de las ciencias sociales, en donde no suelen existir modelos que funcionen universalmente y resulta difícil verificar el cumplimiento de los supuestos realizados.

Es posible mitigar estos riesgos mediante la etapa de diagnóstico del modelo. Sin embargo, los detractores del paradigma basado en modelos argumentan que este tipo de inferencia requiere dedicar mucho tiempo y recursos al desarrollo del modelo. Frente a esta crítica, Little (2004) sostiene que aunque ningún modelo debe ser aplicado a ciegas, existen múltiples modelos "estándar" aplicables a una gran diversidad de poblaciones.

Si bien la dependencia del modelo constituye una importante limitación, la inferencia basada en modelos posee múltiples ventajas. En primer lugar, al no requerir muestras aleatorias, es posible utilizar muestras por conveniencia o datos ya disponibles cuyo diseño se desconoce. Por este motivo, puede ser una estrategia mucho menos costosa que la basada en el diseño.

En segundo lugar, la inferencia basada en modelos permite incorporar información no contenida en el diseño muestral. Si el modelo seleccionado es adecuado y sus parámetros son estimados correctamente, las predicciones resultantes son potencialmente mucho más eficientes que las basadas en el diseño, sobre todo teniendo en cuenta que es posible deducir la forma del predictor óptimo para cada modelo. Cabe señalar que si el modelo es correcto, no es necesario ponderar las observaciones. Más aún, utilizar los pesos cuando no es necesario incrementa la variabilidad de las observaciones y reduce la eficiencia de las estimaciones cuanto menor sea el tamaño de la muestra (Pfeffermann, 1993; Solon et al. 2015).

En tercer lugar, el enfoque basado en modelos logra compatibilizar el muestreo con el resto de la Estadística. Hasta su surgimiento, los principios que regían la inferencia clásica no parecían ser aplicables al muestreo. Royall (1992) sostiene que esto se debe al llamado "Principio de aleatorización", el cual

introduce artificialmente una fuente de aleatoriedad simplemente por la forma en que se seleccionan los datos. En cambio, la inferencia basada en modelos recurre a la misma metodología que muchas de las ramas de la Estadística: se propone un modelo en base a la experiencia previa y se lo estima, se verifica el cumplimiento de sus supuestos y en caso de ser necesario se lo modifica (Gregoire, 1998).

En cuarto lugar, la inferencia basada en modelos ofrece alternativas para enfrentar algunos problemas que la inferencia basada en el diseño no ha logrado solucionar como son los errores de medición, la no respuesta y la estimación en áreas pequeñas (ver Subsección 2.5.3). En términos generales, el paradigma basado en modelos permite trabajar con muestras "corrompidas" (Little, 2004).

En quinto lugar, dado que en general se cuenta con un modelo que vincula la variable de interés con la información auxiliar y estimaciones de los parámetros superpoblacionales, el enfoque basado en modelos puede facilitar la interpretación de los resultados desde un punto de vista conceptual (Royall, 1992).

Finalmente, la inferencia basada en modelos es preferible a la basada en el diseño cuando se desea predecir el valor puntual y/o estimar la varianza asociada a la variable de interés para un cierto individuo de la población (Brus y De Gruijter, 1997).

2.4. Surgimiento de otros paradigmas de la inferencia

De acuerdo con Little (2004), en la actualidad, el debate acerca de cuál es el mejor paradigma de la inferencia en muestreo ha perdido intensidad y se ha adoptado una perspectiva más pragmática, según la cual se recurre a uno u otro dependiendo de la situación concreta. En este sentido, varios autores consideran que ningún enfoque ha demostrado ser claramente superior al otro (ver, por ejemplo, Brus y De Gruijter, 1997 y Ståhl et al. 2016).

En este contexto, se desarrolló la inferencia asistida por modelos con el objetivo de aprovechar algunas de las ventajas del enfoque basado en modelos sin perder la robustez característica de la inferencia basada en el diseño (Särndal et al. 1992). Bajo este paradigma, se recurre a modelos que hacen uso de la información auxiliar disponible, lo cual potencialmente puede mejorar la

precisión de las estimaciones. Sin embargo, a diferencia de la inferencia basada en modelos, el enfoque asistido por modelos es "modelo independiente" en el sentido de que la validez de las estimaciones no está condicionada a que el modelo se encuentre correctamente especificado. Esto se debe a que la base de la inferencia sigue estando en el diseño muestral; el modelo se utiliza para describir a la población y no al proceso que la genera (Nedyalkova y Tillé, 2008). Esto implica que, si el modelo está errado, a lo sumo, no se reducirá la varianza de las estimaciones, pero sí se mantendrá su insesgamiento aproximado. Cabe destacar que al igual que bajo el enfoque basado en el diseño, en la inferencia asistida por modelos el insesgamiento se calcula con respecto al diseño muestral y no al modelo. Es decir, que la fuente de aleatoriedad es la muestra y no la variable de interés.

Uno de los estimadores asistidos por modelos más utilizados para el total de una variable de interés es el estimador de regresión (GREG, por su sigla en inglés), el cual utiliza modelos de regresión lineal, aunque se han propuesto algunas extensiones para modelos lineales generalizados (Rondon et al. 2012). Algebraicamente, el estimador de regresión se compone de la suma de dos términos. Por un lado, se considera la suma de las predicciones dadas por el modelo para cada elemento de la población. A dicho valor, se le añade un término de corrección que puede ser entendido como el estimador de Horvitz-Thompson para el error total del modelo a partir de los errores conocidos para la muestra. Este segundo término incorpora el diseño muestral en el estimador y permite que sea aproximadamente insesgado. Además, cuando el modelo tiene un buen ajuste, el término de corrección será relativamente pequeño y la varianza de las estimaciones se reducirá (Ståhl et al. 2016).

Aunque el enfoque asistido por modelos supera el problema de robustez de la inferencia basada en modelos, cabe destacar que frecuentemente ambos paradigmas arrojan resultados muy similares (por ejemplo, Nedyalkova y Tillé, 2008).

Dadas sus ventajas, la inferencia asistida por modelos es el enfoque más utilizado en la actualidad. Sin embargo, se han propuesto otras alternativas. Por ejemplo, Ståhl et al. (2016) proponen la "inferencia híbrida", la cual es útil cuando no se cuenta con variables auxiliares en el marco muestral y su recolección resulta costosa. En este caso, la estimación se da en dos fases. En primer lugar, se utiliza un estimador basado en el diseño para estimar el total de la variable auxiliar deseada. Para ello, se requiere una muestra probabilística.

Una vez estimado el total de la variable auxiliar, se recurre a un modelo para predecir la variable de interés a partir de una segunda muestra independiente de la primera. Como la validez de la predicción dependerá de que no haya errores de especificación del modelo, los autores consideran que la inferencia híbrida es un subtipo de la inferencia basada en modelos.

2.5. Antecedentes de aplicación de la inferencia basada en modelos

2.5.1. Aplicación de la inferencia basada en modelos en estudios científicos

La inferencia basada en modelos ha sido muy utilizada en muchas disciplinas que hacen uso de datos muestrales tales como la geología, la biología y la forestación, entre otras. Sin embargo, con frecuencia esto no es explicitado y no se verifica el cumplimiento de los supuestos del modelo utilizado. En consecuencia, varios autores han comparado los enfoques basados en el diseño y en modelos en el marco de campos de estudio específicos con el objetivo de aclarar estas cuestiones.

Brus y De Gruijter (1997) analizan el uso de los paradigmas de la inferencia en geología. Señalan que muchos estudios afirman erróneamente que al trabajar con muestras de suelos no puede utilizarse el enfoque basado en el diseño ya que los datos se encuentran espacialmente correlacionados. De esta manera, se considera que la independencia de las observaciones viene dada por el fenómeno de interés propiamente dicho, cuando en realidad es el diseño muestral el que por construcción asegura esto. Es decir, que la validez de la inferencia no depende de la distribución espacial de los datos. Por este motivo, en este campo se ha dejado de lado la inferencia basada en el diseño en pos de la basada en modelos. Para los autores, esto es un problema debido a que el paradigma basado en modelos se apoya en algunos supuestos subjetivos cuestionables.

Gregoire (1998) y Ståhl et al. (2016) comparan la aplicación de los distintos paradigmas de la inferencia en estudios forestales. Ambos afirman que muchos artículos científicos no explicitan cuál fue el enfoque utilizado y no trabajan con la suficiente rigurosidad estadística. Por otra parte, Ståhl et al. (2016) destacan que la inferencia basada en el diseño puede no ser factible cuando

existen dificultades en el relevamiento de la información dadas por carreteras en malas condiciones, áreas peligrosas y terrenos privados, entre otras. Esto implica que, en ocasiones, las muestras sean pequeñas y no aleatorias. Por otro lado, con frecuencia se tiene gran cantidad de variables auxiliares vinculadas a diversas variables de interés forestales. En particular, se cuenta con muchos datos satelitales a partir de los cuales es posible estimar distintos modelos. Es decir, que la forestación presenta algunas características particulares que ocasiones vuelven preferible (o necesario) recurrir a la inferencia basada en modelos por sobre la basada en el diseño.

Geuna (2000) evalúa la aplicación de ambos enfoques en estudios de la morfología del sistema nervioso. Dado que no es factible observar todas las células de interés, el muestreo juega un rol fundamental en esta disciplina. Aunque reconoce que ninguno de los dos paradigmas ha probado ser superior al otro, aboga por el enfoque basado en el diseño ya que resulta difícil comprobar la validez de los modelos más comúnmente usados en este campo.

En el marco de la sociología, Shi et al. (2019) presentan la inferencia basada en modelos en el contexto del muestreo *respondent-driven* (RDS por su sigla en inglés). Este método está diseñado para estudiar poblaciones "ocultas" estigmatizadas o que desean mantener su privacidad, para las cuales las técnicas tradicionales suelen ser inefectivas. Para ello, se deja que los propios miembros de la población se recluten entre ellos para participar de una encuesta. Una vez obtenida la muestra, es usual utilizar la inferencia basada en el diseño para estimar las cantidades de interés. Para ello, se le asigna a cada individuo una probabilidad de inclusión proporcional a su "grado" en la población. El concepto de grado se toma del análisis de redes y se define como la cantidad de vínculos que tiene un individuo con el resto de los elementos de una red. Así, a mayor cantidad de contactos con el resto de la población, mayor será la probabilidad de ser reclutado, lo cual parece razonable. Sin embargo, para que las estimaciones basadas en el diseño sean insesgadas, debe cumplirse un conjunto de supuestos de difícil verificación. De acuerdo con los autores, en este caso, una opción es corregir los sesgos generados mediante modelos que tomen en cuenta la estructura de la red y aspectos comportamentales y sociales conocidos. Por este motivo, consideran que el enfoque basado en modelos es superior al basado en el diseño.

Finalmente, Aubry y Francesiaz (2022) estudian la aplicación de los paradigmas basados en modelos y en el diseño en el campo de la ecología como

herramienta para cuantificar la abundancia de una especie en un momento del tiempo y en un cierto espacio geográfico. En el marco de la inferencia basada en modelos, cuestionan la validez del modelo delta-lognormal, muy utilizado para modelar la distribución de las especies en un espacio. Argumentan que antes de utilizarse este modelo, debería realizarse algún contraste de hipótesis para evaluar si los datos muestrales se apartan significativamente de la distribución propuesta o no. Dado que esto puede resultar complejo, prefieren el paradigma basado en el diseño "libre de modelos" por sobre la inferencia basada en modelos.

2.5.2. Ejemplo de aplicación de la inferencia basada en modelos en estadísticas oficiales

En la mayoría de los países, las estadísticas oficiales construidas a partir de muestras utilizan la inferencia basada en el diseño. Una excepción a esto son las estimaciones de la tasa de desempleo por sexo y edad a nivel de localidades LAD (*"Local authority districts"*) en el Reino Unido (National Statistics, 2006). Dado que los tamaños muestrales reducidos no permiten obtener estimaciones basadas en el diseño confiables a nivel de localidad, la Oficina de Estadísticas Nacionales (ONS por su sigla en inglés), desarrolló una metodología basada en modelos. A partir de un modelo logístico, se estimó la proporción anual de las personas en edad de trabajar desempleadas en cada LAD. Se incluyeron distintas variables explicativas obtenidas a partir de registros administrativos y censos. Entre ellas, la más importante es la cantidad de personas que solicitaron un subsidio por desempleo. Además, se agregó un término para modelar efectos aleatorios que recojan diferencias entre localidades no explicadas por las variables seleccionadas. La estimación final para cada LAD se calcula como el promedio ponderado entre la predicción del modelo y la estimación basada en el diseño. A mayor tamaño muestral, mayor es el peso que se le da a la estimación basada en el diseño. En todos los casos, las estimaciones obtenidas mediante esta nueva metodología fueron más precisas que las basadas en el diseño.

2.5.3. Estimación en Áreas Pequeñas

Si bien la inferencia basada en modelos casi no es utilizada en estadísticas oficiales en su versión "pura", crecientemente se ha recurrido a ella en el marco de la Estimación de Áreas Pequeñas (SAE por su sigla en inglés). Esta subdisciplina del muestreo se centra en la estimación de cantidades de interés en áreas geográficas pequeñas y/o dominios no planeados, para los cuales la cantidad de observaciones disponibles dentro de una muestra es demasiado pequeña para arrojar estimaciones precisas (Ghosh y Rao, 1994; Rao y Molina, 2015).

En general las encuestas nacionales arrojan buenos resultados a nivel agregado, pero no funcionan para localidades o grupos muy pequeños. De esta manera, las estimaciones "directas" en estos dominios suelen presentar altos coeficientes de variación y por lo tanto no suelen ser publicadas. Si bien este problema podría resolverse incrementando el tamaño muestral de cada dominio, en general esto resulta ser muy costoso. Además, no es posible anticipar todos los potenciales dominios de interés.

En este contexto, la metodología SAE surge como una posible solución. Para mejorar la calidad de las estimaciones, se combinan las estimaciones "directas" basadas en el diseño con estimaciones "indirectas", basadas en modelos. En este sentido, el estudio presentado en la sección anterior podría ser entendido como un ejemplo de Estimación en Áreas Pequeñas (si bien en el correspondiente informe no se utiliza este término explícitamente).

Así, los métodos SAE recurren a modelos que permiten "pedir prestada" información a otras unidades similares, de forma de lograr una mayor eficiencia en las estimaciones. Es decir, que los modelos permiten vincular dichas unidades, ya sea en forma explícita o implícita (Rao, 2005). Al igual que en la inferencia basada en modelos "pura", esto requiere incorporar información auxiliar tal como censos o registros administrativos.

La metodología SAE ha tenido un gran auge en las últimas décadas, y existen múltiples ejemplos de aplicación. Por ejemplo, en Estados Unidos, se lo utiliza para obtener estimaciones a nivel de condado y/o distrito escolar (Pfeffermann, 2002; Rao, 2005; Ghosh, 2020). En primer lugar, la Oficina del Censo la utiliza para estimar el ingreso medio y los niveles de pobreza por tramo etario y localidad. De igual forma, la Oficina de Estadísticas Laborales estima el desempleo por industria y localidad. Por su parte, el Departamento

de Agricultura recurre a la Estimación en Áreas Pequeñas para mapear la producción de soja y maíz en el territorio nacional. Finalmente, el *Research Triangle Institute* usa este método para estimar la tasa de uso de drogas ilegales por condado y según sexo, edad y raza.

Por otro lado, las estimaciones SAE son muy frecuentemente utilizadas para construir mapas de pobreza que permitan identificar a dónde se deberían destinar más recursos y dirigir las políticas sociales (Corral Rodas et al. 2020). Por ejemplo, en un documento publicado por el Bank (2020), se presentan los casos de Filipinas y Tailandia, cuyas oficinas estadísticas publican estimaciones de pobreza por área desde hace más de 20 años.

2.5.4. Muestras no probabilísticas

La inferencia basada en modelos también puede aplicarse para obtener estimaciones a partir de muestras no probabilísticas (Wu, 2022). Para estas muestras, se desconoce la probabilidad de que un elemento pertenezca a la muestra, conocida como *"propensity score"* o probabilidad de participación. Por lo tanto, no es posible ponderar las observaciones por su probabilidad de inclusión y así corregir el sesgo de la muestra. A pesar de ello, el muestreo no probabilístico es atractivo en la medida que permite relevar información de forma rápida y económica. En consecuencia, es importante desarrollar metodologías que permitan hacer uso de este tipo de datos.

En este contexto, (Wu, 2022) discute distintas alternativas para mejorar la calidez de las estimaciones obtenidas y propone una que hace uso de la inferencia basada en modelos. Para ello, se requiere contar con variables auxiliares presentes tanto en la muestra no probabilística como en una muestra probabilística, conocida como "muestra de referencia".

En primer lugar, se define un modelo, cuyos parámetros se estiman a partir de la muestra no probabilística. Luego, se hace una predicción para cada elemento de la muestra de referencia, y, finalmente, se obtiene el parámetro de interés. A diferencia del enfoque basado en modelos "puro", las covariables no se encuentran disponibles para toda la población de interés, sino solamente para la muestra de referencia. Por consiguiente, las variables auxiliares deberán ponderarse por el inverso de las probabilidades de inclusión dados por la muestra de referencia.

Dado que la variable de interés no se encuentra presente en la muestra

de referencia, puede concebírsela como un conjunto de valores faltantes que deben ser imputados. Por este motivo, los estimadores utilizados bajo esta metodología son conocidos como *"Mass imputation estimators"* (estimadores MI). Para imputar cada valor faltante, puede utilizarse una gran variedad de modelos y técnicas.

Como cualquier método basado en modelos, la principal desventaja los estimadores MI es que su validez depende del cumplimiento de un conjunto de supuestos de difícil verificación. En particular, se asume que la propensión a participar en una muestra no probabilística es independiente del valor que tome la variable de interés, lo cual es cuestionable.

Capítulo 3

Marco teórico

A continuación, se formalizan varias de las nociones tratadas en el Capítulo 2. En primer lugar, se definen algunos conceptos básicos de muestreo y se especifica la notación utilizada. En segundo lugar, en el marco de la inferencia basada en el diseño, se describen algunos diseños muestrales y se introduce el estimador de Horvitz-Thompson. En tercer lugar, se retoman algunas cuestiones asociadas al paradigma basado en modelos y se presentan los dos modelos considerados en este trabajo, a saber, el modelo de población homogénea y el modelo de población lineal general. Posteriormente, se introduce el estimador de regresión como ejemplo de estimador asistido por modelos. Finalmente, se analiza en qué casos la inferencia basada en modelos y la asistida por modelos difieren entre sí, dados los supuestos bajo los cuales se trabajó.

Los conceptos relativos a la inferencia basada en el diseño y a la asistida por modelos fueron tomados de Särndal et al. (1992). Por su parte, los modelos y otros aspectos asociados a la inferencia basada en modelos siguen el desarrollo de Chambers y Clark (2012).

3.1. Notación y conceptos previos

Una población finita se denota como U. Para identificar y acceder a cada una de las unidades que la componen, se requiere de un listado de sus unidades al que se le denomina "marco muestral" y se denota como F o U_F. Dentro del marco, a cada elemento de U se le asocia un entero i perteneciente al conjunto $\{1, 2, \ldots, N\}$. En este trabajo, se asume que se cuenta con un marco completo, de manera que $U = U_F$. Existe, entonces, una relación uno a uno

entre la población y las unidades del marco, y no existen problemas de sub o sobrecobertura.

Por su parte, una muestra s es un subconjunto de U, a la que se identifica a través de un subconjunto de los enteros entre 1 y N. La cantidad de elementos de s se indica con n. Por otra parte, el complemento de s en U, se denota con r, esto es: $r = U - s$ o $U = s \cup r$.

Se le llama variable a cualquier atributo medido en una muestra y/o disponible en el marco. De esta manera, se consideran dos tipos de variables:

- **Variables de interés:** Se denotan como Y y se caracterizan por que sólo se conocen sus valores para los elementos pertenecientes a la muestra s. Es decir, que los valores pertenecientes a r son desconocidos, por lo que deberán ser estimados.

- **Variables auxiliares o covariables:** Se las identifica con las letras X y Z y comprenden las variables para las que se conoce su valor para cada uno de los N elementos de las población. Así, se asume que se cuenta con información a nivel de marco para todas las covariables. Sin embargo, para mejorar las estimaciones, con frecuencia alcanza con conocer los totales poblacionales y los valores únicamente para los n elementos de la muestra.

En este trabajo, las minúsculas se reservan para indicar valores específicos de determinadas unidades. Así, y_i denota el valor de una variable Y asociado al i-ésimo elemento de la población.

Los valores individuales y_i no suelen ser la mayor preocupación en las encuestas por muestreo, sino que el interés suele ser estimar parámetros poblacionales que resuman dichos valores, como, por ejemplo:

- **Total poblacional:** $t_y = \sum_{i=1}^{N} y_i = \sum_{i \in U} y_i = \sum_{U} y_i$

- **Media poblacional:** $\bar{y}_U = {}^{t_y}/_{N}$

3.2. Inferencia basada en el diseño

Como se detalló en el Capítulo 2, en la inferencia basada en el diseño, la aleatoriedad proviene del mecanismo de selección de la muestra y no de una distribución asumida. Por lo tanto, es relevante deducir las probabilidades de

extraer cada muestra posible. Éstas, a su vez, determinarán las probabilidades de inclusión en una muestra de cada elemento de la población.

Sea S el conjunto conocido de las muestras posibles. Como la selección de la muestra es aleatoria, a cada una de ellas le corresponde una cierta probabilidad, $p(s)$, conocida como "diseño muestral". Se trata de una función con dominio en S y recorrido entre 0 y 1: $p(.) \rightarrow [0, 1]$.

Se le denomina I_i a una indicatriz que vale 1 cuando el elemento i pertenece a S y 0 en caso contrario:

$$I_i(S) = \begin{cases} 1 & \text{si } i \in S \\ 0 & \text{en otro caso} \end{cases} \tag{3.1}$$

Por otra parte, la probabilidad de inclusión de primer orden se define como la probabilidad de que un cierto elemento i de la población U pertenezca a la muestra seleccionada mediante un cierto diseño. De esta manera, equivale a la suma de la probabilidad de todas las muestras que contienen a i:

$$\pi_i = P(i \in S) = P(I_i = 1) = \sum_{s \ni i} p(s) \tag{3.2}$$

En consecuencia, I_i es una variable de Bernoulli de parámetro π_i, de manera que $I_i \sim Ber(\pi_i)$.

Análogamente, se conoce como probabilidad de inclusión de segundo orden a la probabilidad de que dos elementos i y j de U pertenezcan simultáneamente a una muestra:

$$\pi_{ij} = P(i, j \in S) = P(I_i = 1, I_j = 1) = P(I_i I_j = 1) = \sum_{s \ni i,j} p(s) \tag{3.3}$$

Cuando todos los elementos de la población tienen una probabilidad de inclusión de primer orden no nula, se dice que el muestreo es aleatorio. En este caso, el estimador de Horvitz-Thompson será insesgado (ver Subsección 3.2.1). Asimismo, si se cumple que las probabilidades de inclusión de segundo orden son no nulas para todo par de elementos, se dice que el diseño es medible. Cuando esto ocurre, el estimador de la varianza del estimador Horvitz-Thompson también será insesgado.

3.2.1. Estimador Horvitz-Thompson

Sea $p(s)$ un diseño aleatorio cualquiera, el estimador de Horvitz-Thompson (π) para t_y se define como la suma de los valores de la variable de interés en la muestra, ponderados por el inverso de sus probabilidades de inclusión:

$$\hat{t}_{y_\pi} = \sum\nolimits_s \frac{y_i}{\pi_i} \tag{3.4}$$

Este estimador posee la propiedad fundamental de ser insesgado respecto al diseño muestral, de forma que:

$$\begin{aligned} E_{p(s)}(\hat{t}_{y_\pi}) &= E\left(\sum\nolimits_S \frac{y_i}{\pi_i}\right) = E\left(\sum\nolimits_U I_i \frac{y_i}{\pi_i}\right) \\ &= \sum\nolimits_U E(I_i)\frac{y_i}{\pi_i} = \sum\nolimits_U y_i = t_y \end{aligned} \tag{3.5}$$

$E_{p(s)}$ indica que la esperanza se calcula con respecto al diseño muestral.

La varianza del estimador de Horvitz-Thompson será, entonces:

$$\begin{aligned} \mathrm{Var}_{p(s)}(\hat{t}_{y_\pi}) &= \mathrm{Var}\left(\sum\nolimits_s \frac{y_i}{\pi_i}\right) = \mathrm{Var}\left(\sum\nolimits_U I_i \frac{y_i}{\pi_i}\right) \\ &= \sum\nolimits_U \mathrm{Var}(I_i)\left(\frac{y_i}{\pi_i}\right)^2 + \sum\nolimits_{i\neq j}\sum\nolimits_U \mathrm{Cov}(I_i, I_j)\left(\frac{y_i}{\pi_i}\right)\left(\frac{y_j}{\pi_j}\right) \\ &= \sum\sum\nolimits_U \mathrm{Cov}(I_i, I_j)\left(\frac{y_i}{\pi_i}\right)\left(\frac{y_j}{\pi_j}\right) \\ &= \sum\sum\nolimits_U \Delta_{ij}\left(\frac{y_i}{\pi_i}\right)\left(\frac{y_j}{\pi_j}\right) \end{aligned} \tag{3.6}$$

donde $\Delta_{ij} = \mathrm{Cov}(I_i I_j) = \pi_{ij} - \pi_i\pi_j$.

Finalmente, la varianza del estimador puede ser estimada ponderando Δ_{ij} por el inverso de las probabilidades de inclusión de segundo orden:

$$\widehat{\mathrm{Var}}_{p(s)}(\hat{t}_{y_\pi}) = \sum\sum\nolimits_s \frac{\Delta_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i}\right)\left(\frac{y_j}{\pi_j}\right) \tag{3.7}$$

3.2.2. Diseño simple

Uno de los diseños muestrales más básicos es el muestreo aleatorio simple sin reposición, en general notado como SI (Särndal et al. 1992, p. 66), y consiste en extraer n elementos de manera independiente y sin reposición de los N presentes en la población U. Bajo estas condiciones, existen C_n^N muestras posibles de

tamaño n, de manera que la probabilidad de extraer cada una ellas será:

$$p(s) = \begin{cases} (C_n^N)^{-1} & \forall\ s \text{ de tamaño } n \\ 0 & \text{en otro caso} \end{cases} \tag{3.8}$$

A partir de lo anterior, puede demostrarse que las probabilidades de inclusión de primer y segundo orden son:

$$\pi_i = \frac{n}{N} \quad \forall i \in U \tag{3.9}$$

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)} \quad \forall i \neq j \in U \tag{3.10}$$

Por lo tanto, el estimador de Horvitz-Thompson para t_y viene dado por:

$$\hat{t}_{y_\pi} = \sum\nolimits_s \frac{y_i}{n/N} = N\bar{y}_s \tag{3.11}$$

donde $\bar{y}_s$ es la media muestral.

Por su parte, se puede demostrar que la varianza de dicho estimador es:

$$\mathrm{Var}(\hat{t}_{y_\pi}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_{y_U}^2}{n} \tag{3.12}$$

donde $s_{y_U}^2 = \frac{1}{N-1} \sum_U (y_i - \bar{y}_U)^2$ es la varianza poblacional de Y.

Aplicando un razonamiento análogo al utilizado para obtener $\hat{t}_\pi$, un estimador insesgado de la varianza del estimador será:

$$\widehat{\mathrm{Var}}(\hat{t}_{y_\pi}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_{y_s}^2}{n} \tag{3.13}$$

siendo $s_{y_s}^2 = \frac{1}{n-1} \sum_s (y_i - \bar{y}_s)^2$ la varianza muestral de Y.

3.2.3. Diseño πps

Bajo un diseño πps (Särndal et al. 1992, p. 90), los elementos de la población tienen probabilidades de ser incluidos en la muestra proporcionales al valor de una cierta variable auxiliar Z:

$$\pi_i \propto z_i \quad \forall i = 1, \ldots, N \tag{3.14}$$

Si el diseño es tal que las probabilidades de inclusión π_i son estrictamente

proporcionales al tamaño de la variable de interés y el diseño es de tamaño fijo, se demuestra que el estimador de Horvitz-Thompson estima sin error t_y, de forma que su varianza es nula. Esto solamente es de interés teórico; en la práctica sólo puede implementarse de forma aproximada siempre y cuando se cuente con una variable auxiliar para la cual la correlación con la variable de interés sea alta. En tal caso, se tendrá:

$$\pi_i = \frac{nz_i}{\sum_U z_i} \quad \forall i = 1, \ldots, N \tag{3.15}$$

donde $\hat{t}_{y_\pi}$ y $\widehat{\text{Var}}(\hat{t}_{y_\pi})$ se calculan de acuerdo a las ecuaciones 3.4 y 3.7, respectivamente.

3.3. Inferencia basada en modelos

Como se analizó en el capítulo anterior, la inferencia basada en modelos supone que los valores de y_i en la población constituyen realizaciones de variables aleatorias (VA's) cuya distribución se rige por un modelo superpoblacional. Es decir, que la aleatoriedad en el proceso de estimación viene dada por la propia variable de interés Y, y no por el mecanismo de selección de la muestra. En este sentido, s se supone constante (Chambers y Clark, 2012, p. 14). Además, t_y es igual a la suma de VA's y por tanto es, en sí misma, otra VA. Análogamente, los predictores $\hat{t}_y$ son funciones de los valores muestrales de Y, con lo cual también son VA's.

En suma, tanto $\hat{t}_y$ como t_y son VA's. Sin embargo, se rigen por dos procesos diferentes:

- **Modelo superpoblacional:** El valor que tome el total t_y dependerá de los valores de Y, los cuales son generados por un proceso aleatorio dado por el modelo superpoblacional.
- **Diseño muestral:** El valor del predictor $\hat{t}_y$ se ve influido por el mecanismo de selección de la muestra, y puede tratarse o no de un proceso aleatorio.

Evidentemente, se desea que $\hat{t}_y$ arroje predicciones lo más cercanas posibles al total de interés, de manera de minimizar el error $\hat{t}_y - t_y$. Aunque el mismo es desconocido para todas las unidades de la población no observadas, es posible derivar sus propiedades estadísticas a partir del modelo superpoblacional

asumido. Esto permite hallar un predictor que genere errores con una esperanza y una varianza pequeñas, condicionales al modelo seleccionado. Así, se obtendrá un $\hat{t}_y$ insesgado y lo más preciso posible, de manera de minimizar su Error Cuadrático Medio (ECM):

$$\text{ECM}(\hat{t}_y) = E(\hat{t}_y - t_y)^2 = \big(E(\hat{t}_y - t_y)\big)^2 + \text{Var}(\hat{t}_y - t_y) \tag{3.16}$$

donde $E(\hat{t}_y - t_y)$ es el sesgo de predicción con respecto al modelo especificado.

Por otra parte, t_y puede descomponerse de la siguiente manera:

$$t_y = t_{y_s} + t_{y_r} \tag{3.17}$$

donde t_{y_s} corresponde al total de la variable Y restringido a las unidades pertenecientes a la muestra s, y t_{y_r} es el total de Y para las unidades pertenecientes al complemento de s, r (unidades no muestreadas). En la práctica, se utilizan muestras muy pequeñas en relación al tamaño de la población, por lo que es posible aproximar el total general mediante el total de los elementos no observados ($t_y \approx t_{y_r}$).

Una vez extraída la muestra s, el total t_{y_s} será conocido. Por lo tanto, el problema se reducirá a predecir t_{y_r}, de forma de minimizar el ECM. Posteriormente, dados el modelo y el predictor, será importante determinar cuál es el mecanismo de selección de la muestra que arroja menores errores.

3.4. Modelo de población homogénea

En el marco de la inferencia basada en modelos, uno de los modelos más básicos que pueden proponerse es el denominado "modelo de población homogénea", frecuentemente aplicado cuando no se cuenta con variables auxiliares o cuando se sabe, *a priori*, que éstas no están correlacionadas con la variable de interés Y. En este contexto, no se cuenta con información previa para los elementos de la población que hagan preferible una muestra por sobre otra. En consecuencia, es razonable utilizar un diseño muestral que dé igual probabilidad de selección a todas las posibles muestras de un cierto tamaño n. Lo más lógico, entonces, es utilizar un diseño SI.

3.4.1. Especificación del modelo

Dado que no se cuenta con información que permita diferenciar las distintas unidades de la población, el modelo de población homogénea asume que todas poseen una misma media y una varianza constante en términos de Y. A su vez, se supone que el valor que toma la variable de interés en cada elemento es independiente de lo que ocurre con el resto de la población. En términos analíticos, estos supuestos se expresan como:

$$E(y_i) = \mu \tag{3.18}$$

$$\text{Var}(y_i) = \sigma^2 \tag{3.19}$$

$$y_i \text{ y } y_j \text{ son independientes cuando } i \neq j \tag{3.20}$$

3.4.2. Mejor predictor empírico

En este trabajo, se considera que un predictor es "óptimo" si se minimiza el ECM. Para hallar su forma, una primera opción es derivar el llamado "mejor predictor empírico" (EB, por su sigla en inglés). Para ello, se parte de un resultado conocido, según el cual el predictor que minimiza la varianza de una VA, W, dado el valor de otra VA, V, es su esperanza condicional, $E(W|V)$. Sea $W = t_y$ y $V = \{y_i, i \in s; z_i, i = 1, \ldots, N\}$, el mejor predictor de t_y resulta ser, entonces:

$$\begin{aligned} t_y^* &= E(t_y | y_i, i \in s; z_i, i = 1, \ldots, N) \\ &= E(t_{y_s} + t_{y_r} | y_i, i \in s; z_i, i = 1, \ldots, N) \\ &= t_{y_s} + E(t_{y_r} | y_i, i \in s; z_i, i = 1, \ldots, N) \end{aligned} \tag{3.21}$$

Puesto que bajo el modelo de población homogénea no se cuenta con variables auxiliares Z, y dado que se supone que las realizaciones de Y son independientes entre sí, la Ecuación 3.21 se convierte en:

$$\begin{aligned} t_y^* &= t_{y_s} + E(t_{y_r} | y_i, i \in s) = t_{y_s} + E\left(\sum\nolimits_r y_i | y_i, i \in s\right) \\ &= t_{y_s} + \sum\nolimits_r E(y_i | y_i, i \in s) = t_{y_s} + \sum\nolimits_r E(y_i) \\ &= t_{y_s} + \sum\nolimits_r \mu = t_{y_s} + (N - n)\mu \end{aligned} \tag{3.22}$$

Como μ es desconocido, es necesario reemplazar este parámetro por una

estimación. Se trabaja, entonces, con un estimador *"plug-in"*. Dado que la media poblacional puede ser estimada de diversas formas, el predictor EB no es único. Sin embargo, en este caso, todos los elementos de la muestra proporcionan la misma información, por lo que parece razonable utilizar la media muestral $\bar{y}_s$. Así, se llega al siguiente predictor, comúnmente conocido como "estimador de expansión":

$$\hat{t}_y^E = t_{y_s} + (N-n)\hat{\mu} = t_{y_s} + (N-n)\bar{y}_s = N\bar{y}_s = \frac{N}{n}t_{y_s} \tag{3.23}$$

Se observa, entonces, que bajo un diseño simple, el estimador de expansión coincide con el estimador de Horvitz-Thompson.

3.4.3. Mejor predictor lineal insesgado

Otra alternativa para predecir t_y es el llamado "mejor predictor lineal insesgado" (BLUP, por su sigla en inglés). Si bien se obtiene mediante un procedimiento más complejo, tiene la ventaja de ser único (a diferencia del predictor EB, cuya forma final dependerá de cómo se estime μ).

Como su nombre lo indica, un predictor BLUP ($\hat{t}_y^{BLUP}$) es lineal, insesgado y de mínima varianza:

- **Predictor lineal:** Puede expresarse como una combinación lineal de los valores muestrales de Y. En general, es deseable que los predictores sean lineales dada la sencillez en su uso. Se tiene, entonces:

$$\hat{t}_y^{BLUP} = \sum\nolimits_s w_i y_i \tag{3.24}$$

  donde w_i es el peso asociado a la observación i.

  Es importante aclarar que los pesos no tienen por qué coincidir con los ponderadores provenientes del diseño. La única restricción es que no dependan de la variable de interés. Por el contrario, sí es usual que dependan de las unidades pertenecientes a la muestra y de las variables auxiliares (cuando se cuente con ellas).

- **Predictor insesgado:** La esperanza de los errores debe ser nula:

$$E(\hat{t}_y^{BLUP} - t_y) = 0 \tag{3.25}$$

- **Predictor de mínima varianza:** Para cualquier muestra s, debe presentar la menor varianza del error de entre todos los predictores lineales insesgados de t_y, es decir:

$$\text{Var}(\hat{t}_y^{BLUP} - t_y) \leq \text{Var}(\hat{t}_y - t_y) \tag{3.26}$$

  siendo $\hat{t}_y$ cualquier otro predictor lineal insesgado de t_y.

Para derivar el predictor BLUP, es necesario asumir que no existe correlación entre las distintas observaciones, lo cual constituye un supuesto más débil que el de independencia establecido en la Ecuación 3.20.

Por otra parte, se observa que cualquier predictor lineal de t_y puede descomponerse como:

$$\hat{t}_y = \sum\nolimits_s w_i y_i = \sum\nolimits_s y_i + \sum\nolimits_s (w_i - 1) y_i = t_{y_s} + \sum\nolimits_s u_i y_i \tag{3.27}$$

donde $u_i = w_i - 1$.

En consecuencia, el error muestral puede expresarse como:

$$\hat{t}_y - t_y = t_{y_s} + \hat{t}_{y_r} - (t_{y_s} - t_{y_r}) = \sum\nolimits_s u_i y_i - \sum\nolimits_r y_i \tag{3.28}$$

Puede entenderse u_i como el peso de la unidad no observada i. Es decir, es el peso asignado al valor y_i cuando se predice el total no muestral de Y. Así, para obtener el predictor BLUP, alcanza con obtener los u_i (o, en su defecto, los w_i).

Por definición, los pesos deben conducir a un predictor insesgado, de forma que:

$$\begin{aligned} B(\hat{t}_y) &= E(\hat{t}_y - t_y) = E\left(\sum\nolimits_s u_i y_i - \sum\nolimits_r y_i\right) \\ &= \sum\nolimits_s u_i \mu - \sum\nolimits_r \mu = \mu\left(\sum\nolimits_s u_i - (N - n)\right) = 0 \end{aligned} \tag{3.29}$$

La condición anterior se cumple sólo si:

$$\sum\nolimits_s u_i - (N - n) = 0 \tag{3.30}$$

Asimismo, los pesos u_i deben ser tales que el predictor resultante tenga una varianza mínima respecto a los otros predictores de su clase. Para ello, se parte

del siguiente resultado:

$$\text{Var}(\hat{t}_y - t_y) = \text{Var}(\hat{t}_{y_r} - t_{y_r}) = \text{Var}(\hat{t}_{y_r}) - 2\text{Cov}(\hat{t}_{y_r}, t_{y_r}) + Var(t_{y_r}) \quad (3.31)$$

donde:

$$\begin{aligned} \text{Var}(\hat{t}_{y_r}) &= \text{Var}\left(\sum\nolimits_s u_i y_i\right) = \sum\nolimits_s \text{Var}(u_i y_i) \\ &= \sum\nolimits_s u_i^2 \text{Var}(y_i) = \sigma^2 \sum\nolimits_s u_i^2 \end{aligned} \quad (3.32)$$

$$\text{Var}(t_{y_r}) = \text{Var}\left(\sum\nolimits_r y_i\right) = \sum\nolimits_r \text{Var}(y_i) = (N-n)\sigma^2 \quad (3.33)$$

$$\text{Cov}(\hat{t}_{y_r}, t_{y_r}) = 0 \quad (3.34)$$

En la Ecuación 3.34 se hace uso del supuesto de que los valores de y pertenecientes a s están incorrelacionados con los pertenecientes a r.

Dado que $\text{Var}(t_{y_r})$ y $\text{Cov}(\hat{t}_{y_r}, t_{y_r})$ no son funciones de los pesos de predicción u_i, para minimizar $\text{Var}(\hat{t}_y - t_y)$ respecto de los u_i alcanza con minimizar $\text{Var}(\hat{t}_{yr})$. Es decir, los valores óptimos de u_i se obtienen minimizando $\sum_s u_i^2$, sujeto a la restricción de insesgamiento de la Ecuación 3.30. Para ello, se considera el correspondiente Lagrangiano, L (se multiplica λ por dos para simplificar los cálculos):

$$L = \sum\nolimits_s u_i^2 - 2\lambda\left(\sum\nolimits_s u_i - (N-n)\right) \quad (3.35)$$

Derivando L respecto de u_i e igualando a cero, se tiene que $u_i = \lambda$. Sustituyendo esta expresión en la restricción establecida en la Ecuación 3.30, se obtiene que:

$$\lambda = \frac{N-n}{n} \quad (3.36)$$

Esto implica que:

$$u_i = \frac{N-n}{n} \Rightarrow w_i = \frac{N}{n} \quad (3.37)$$

Sustituyendo estos pesos en la Ecuación 3.27, se concluye que:

$$\begin{aligned} \hat{t}_y &= t_{y_s} + \sum\nolimits_s \left(\frac{N-n}{n}\right) y_i = t_{y_s} + \left(\frac{N-n}{n}\right) \sum\nolimits_s y_i \\ &= t_{y_s}\left(1 - \left(\frac{N-n}{n}\right)\right) = \frac{N}{n} t_{y_s} \end{aligned} \quad (3.38)$$

De esta forma, se cumple que el predictor BLUP para el modelo de población homogénea coincide con el predictor de expansión definido en la Ecuación 3.23.

3.4.4. Estimación de la varianza e intervalos de confianza

Al sustituir los pesos hallados en la Ecuación 3.37 en la Ecuación 3.32, se llega a que:

$$\text{Var}(\hat{t}_{yr}^E) = \sigma^2 \sum\nolimits_s \left(\frac{N-n}{n}\right)^2 = \sigma^2 \frac{(N-n)^2}{n} \tag{3.39}$$

Por lo tanto, la varianza del error de predicción es:

$$\begin{aligned} \text{Var}(\hat{t}_y^E - t_y) &= \text{Var}(\hat{t}_{yr}^E) + \text{Var}(t_{yr}) - 2\text{Cov}(\hat{t}_{yr}^E, t_{yr}) \\ &= \sigma^2 \left(\frac{(N-n)^2}{n} + (N-n)\right) \\ &= \sigma^2 (N-n)\frac{N}{n} \\ &= \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sigma^2 \end{aligned} \tag{3.40}$$

Para construir intervalos de confianza para t_y basados en el estimador de expansión, es necesario estimar la expresión anterior. Para ello, un posible estimador insesgado de σ^2 bajo el modelo de población homogénea es la varianza muestral corregida de Y:

$$s_{y_s}^2 = \frac{1}{n-1} \sum\nolimits_s (y_i - \bar{y}_s)^2 \tag{3.41}$$

Así, un estimador insesgado de la varianza de $\widehat{\text{Var}}\left(\hat{t}_y^E\right)$ dada por la Ecuación 3.40 es:

$$\widehat{\text{Var}}\left(\hat{t}_y^E\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) s_{y_s}^2 \tag{3.42}$$

La anterior ecuación se asemeja mucho a la varianza del estimador de Horvitz-Thompson bajo un diseño simple. Sin embargo, la diferencia fundamental entre ambos radica en la concepción que se tiene de σ^2. En el contexto de la inferencia basada en modelos, este parámetro corresponde a la varianza de la variable de interés como variable aleatoria. En cambio, en la inferencia basada en el diseño, $\sigma^2 = s_{y_U}^2$. Es decir, que σ^2 ya no es aleatoria sino un estadístico descriptivo que denota la dispersión de una población.

Sea z el estadístico que surge de estandarizar $\hat{t}_y^E$:

$$z = \frac{(\hat{t}_y^E - t_y)}{\sqrt{\widehat{\text{Var}}(\hat{t}_y^E)}} \tag{3.43}$$

Por el Teorema Central del Límite, z tiene una distribución asintótica normal estándar. En consecuencia, para un tamaño de muestra "grande", un intervalo de confianza aproximado al $100(1-\alpha)\,\%$ para t_y es:

$$\hat{t}_y^E \pm q_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{t}_y^E)} \tag{3.44}$$

donde $q_{1-\alpha/2}$ es el quantil $(1-\alpha/2)$ de una distribución Normal estándar. Dado que t_y es una variable aleatoria y no un parámetro, con frecuencia este intevalo se conoce también como "intervalo de predicción".

3.5. Modelo de población lineal general

En términos generales, el modelo de población lineal general se caracteriza por utilizar una regresión lineal para predecir el comportamiento de Y. Para ello, utiliza como regresores a una o más variables auxiliares Z. Como es usual en los análisis que hacen uso de modelos lineales, se asume que los errores son homocedásticos.

3.5.1. Especificación del modelo

Para derivar la forma del modelo, es necesario recurrir a notación vectorial y matricial. Salvo que se indique lo contrario, se trabaja con vectores columna.

Se denota como $\mathbf{y}_U$ al vector que contiene los N valores poblacionales de la variable de interés Y. En forma análoga, $\mathbf{z}_{1_U}$, $\mathbf{z}_{2_U}, \ldots, \mathbf{z}_{p_U}$ son los vectores conformados por los N valores poblacionales para las p variables auxiliares disponibles. Estos vectores se concatenan para obtener la matriz $\mathbf{Z}_U$. La misma posee N filas y p columnas, donde la k-ésima columna es $\mathbf{z}_{k_U}$. Se nota como $\mathbf{z}_i$ al vector de dimensión p que corresponde a la i-ésima fila de $\mathbf{Z}_U$.

Por otra parte, se dice que el vector $\mathbf{y}_U$ sigue el modelo de población lineal general definido por $\mathbf{Z}_U$ cuando se cumple que:

$$E(y_i|\mathbf{z}_i) = \mathbf{z}_i^{'}\boldsymbol{\beta} \tag{3.45}$$

$$\mathrm{Var}(y_i|\mathbf{z}_i) = \sigma^2 \tag{3.46}$$

$$y_i \text{ y } y_j \text{ son independientes condicionadas a } \mathbf{Z}_U \text{ cuando } i \neq j \tag{3.47}$$

donde $\boldsymbol{\beta}$ es un vector de parámetros desconocidos de regresión de dimensión p.

A diferencia de lo que ocurría con el modelo homogéneo, en este caso σ^2 se define como la varianza de y condicionada a x. Al considerarse la información auxiliar, se reducirá la dispersión de los posibles valores que puede tomar la variable de interés, lo cual, a su vez, también reducirá la incertidumbre en las predicciones. Por lo tanto, la varianza del error de estimación también disminuirá (ver Ecuación 3.54).

3.5.2. Mejor predictor empírico

Al igual que ocurre bajo el modelo de población homogénea, la condición dada por la Ecuación 3.47 es necesaria para obtener un predictor EB. En cambio, para construir el predictor BLUP, es suficiente que se cumpla únicamente el supuesto más débil de covarianzas nulas.

Bajo el modelo lineal general, el "mejor predictor" tendrá la siguiente forma:

$$t_y^* = t_{y_s} + E(t_{y_r}|\mathbf{y}_s, \mathbf{Z}_U) = t_{y_s} + \sum\nolimits_r \mathbf{z}_i^{'}\boldsymbol{\beta} = t_{y_s} + \mathbf{t}_{z_r}^{'}\boldsymbol{\beta} \tag{3.48}$$

donde $\mathbf{y}_s$ refiere al componente muestral de $\mathbf{y}_U$ y $\mathbf{t}_{z_r}^{'}$ es un vector fila de dimensión p, cuyas entradas contienen los totales no muestrales de las variables auxiliares.

Por otra parte, el mejor estimador lineal insesgado de $\boldsymbol{\beta}$ es el estimador de Mínimos Cuadrados Ordinarios (MCO):

$$\hat{\boldsymbol{\beta}}_{MCO} = (\mathbf{Z'}_s\mathbf{Z}_s)^{-1}\mathbf{Z'}_s\mathbf{y}_s = \left(\sum\nolimits_s \mathbf{z}_i\mathbf{z'}_i\right)^{-1}\sum\nolimits_s \mathbf{z}_i y_i \tag{3.49}$$

donde $\mathbf{Z}_s$ es la sub-matriz de dimensiones $n \times p$ de $\mathbf{Z}_U$ definida por los elementos pertenecientes a la muestra.

El predictor EB se obtiene sustituyendo $\hat{\boldsymbol{\beta}}_{MCO}$ por $\boldsymbol{\beta}$ en la Ecuación 3.48.

3.5.3. Mejor predictor lineal insesgado

Bajo el modelo definido en la Subsección 3.5.1, el predictor EB coincide con el predictor BLUP para t_y, de forma que:

$$\hat{t}_y^{BLUP} = t_{ys} + \mathbf{t'}_{zr}\hat{\boldsymbol{\beta}}_{MCO} \tag{3.50}$$

Dado que se trata de un predictor lineal, es posible escribir $\hat{t}_y^{BLUP}$ como una combinación lineal de los valores muestrales y los ponderadores, de manera

que:

$$\hat{t}_y^{BLUP} = \sum\nolimits_s w_i y_i \tag{3.51}$$

donde:

$$w_i = 1 + \mathbf{t}'_{zr} \left(\sum\nolimits_s \mathbf{z}_j \mathbf{z}'_j \right)^{-1} \mathbf{z}_i \tag{3.52}$$

Nuevamente, se cumple que los pesos w_i dependen tanto de los valores muestrales como no muestrales de las variables auxiliares Z, pero no dependen de la variable de interés.

3.5.4. Estimación de la varianza e intervalos de confianza

Para obtener la varianza del error, se utiliza la expresión para la varianza de $\hat{\boldsymbol{\beta}}_{MCO}$:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{MCO}) = \sigma^2 (\mathbf{Z}'_s \mathbf{Z}_s)^{-1} \tag{3.53}$$

De lo anterior, se sigue que la varianza del error de predicción de $\hat{t}_y^{BLUP}$ bajo el modelo lineal general es:

$$\begin{aligned} \mathrm{Var}(\hat{t}_y^{BLUP} - t_y) &= \mathrm{Var}(\mathbf{t}'_{zr} \hat{\boldsymbol{\beta}}_{MCO} - t_{yr}) \\ &= \mathrm{Var}(t_{yr}) + \mathbf{t}'_{zr} \mathrm{Var}(\hat{\boldsymbol{\beta}}_{MCO}) \mathbf{t}_{zr} \\ &= \sigma^2 ((N-n) + \mathbf{t}'_{zr} (\mathbf{Z}'_s \mathbf{Z}_s)^{-1} \mathbf{t}_{zr}) \end{aligned} \tag{3.54}$$

Por otra parte, un posible estimador insesgado para el parámetro σ^2 es:

$$\hat{\sigma}^2 = (n-p)^{-1} \sum\nolimits_s (y_i - \mathbf{z}'_i \hat{\boldsymbol{\beta}}_{MCO})^2 \tag{3.55}$$

Luego, un estimador de la varianza para $\hat{t}_y^{BLUP}$ viene dado por:

$$\widehat{\mathrm{Var}}(\hat{t}_y^{BLUP}) = \hat{\sigma}^2 ((N-n) + \mathbf{t}'_{zr} (\mathbf{Z}'_s \mathbf{Z}_s)^{-1} \mathbf{t}_{zr}) \tag{3.56}$$

A partir de esta estimación, el correspondiente intervalo de confianza al $100(1-\alpha)\,\%$ se obtiene mediante la Ecuación 3.44.

3.6. Modelo de población lineal simple

El modelo de población lineal simple constituye un caso particular del modelo de población lineal general. En este caso, se trabaja con un una regresión

lineal simple. De esta manera, se utiliza una única variable auxiliar Z y, por consiguiente, no resulta necesario recurrir a notación matricial.

3.6.1. Especificación del modelo

Bajo el modelo de población lineal simple, se asume que, para $i = 1, \ldots, N$, se cumple:

$$E(y_i|z_i) = \alpha + \beta z_i \tag{3.57}$$

$$\text{Var}(y_i|z_i) = \sigma^2 \tag{3.58}$$

$$y_i \text{ y } y_j \text{ son independientes cuando } i \neq j \tag{3.59}$$

donde α es un intercepto y β es el coeficiente asociado a Z.

3.6.2. Mejor predictor empírico

Los estimadores MCO para α y β vienen dados, respectivamente, por:

$$a_L = \bar{y}_s - b_L \bar{z}_s \tag{3.60}$$

$$b_L = \frac{\sum_s (y_i - \bar{y}_s)(z_i - \bar{z}_s)}{\sum_s (z_i - \bar{z}_s)^2} \tag{3.61}$$

A partir de estas estimaciones, el predictor EB se obtiene de la misma forma que para el modelo de población lineal general:

$$t_y^* = t_{y_s} + E(t_{yr}|z_i, i \in r) = t_{y_s} + \sum\nolimits_r (\alpha + \beta z_i) \tag{3.62}$$

Si se sustituyen los parámetros desconocidos α y β por sus estimadores MCO, se obtiene el "mejor predictor empírico" para t_y:

$$\hat{t}_y^L = t_{y_s} + \sum\nolimits_r (a_L + b_L z_i) = N\left(\bar{y}_s + b_L(\bar{z}_U - \bar{z}_s)\right) = N(a_L + b_L \bar{z}_U) \tag{3.63}$$

3.6.3. Mejor predictor lineal insesgado

Puede demostrarse que el predictor dado por la Ecuación 3.63 es tanto EB como BLUP. En este último caso, los pesos w_i vienen dados por la expresión:

$$w_i = \frac{N}{n}\left(1 + \frac{(\bar{z}_U - \bar{z}_s)(z_i - \bar{z}_s)}{(1 - n^{-1})s_z^2}\right) \tag{3.64}$$

donde s_z^2 es la varianza muestral corregida de Z, $s_z^2 = (n-1)^{-1}\sum_s (z_i - \bar{z}_s)^2$.

Por otra parte, al igual que para los modelos de población homogénea y lineal general, para derivar la forma del predictor BLUP, se requiere únicamente que $\text{Cov}(y_i, y_j) = 0, \forall i \neq j$. Así, no es necesario que se cumpla el supuesto más fuerte de independencia establecido en la Ecuación 3.59.

3.6.4. Estimación de la varianza e intervalos de confianza

La varianza del error de predicción puede expresarse como:

$$\text{Var}(\hat{t}_y^L - t_y) = \frac{N^2}{n}\sigma^2 \left(\left(1 - \frac{n}{N}\right) + \frac{(\bar{z}_U - \bar{z}_s)^2}{(1 - n^{-1})s_z^2} \right) \tag{3.65}$$

A partir de la Ecuación 3.65, se deduce que la varianza del error se minimiza cuando $\bar{z}_s = \bar{z}_U$. En ese caso, se dice que la muestra es balanceada de primer orden en Z, y constituye una estrategia óptima para predecir t_y. Para obtener una muestra balanceada, es preferible que los valores de Z sean lo más dispersos posibles y no que se asemejen a $\bar{z}_U$, ya que en este último caso la varianza muestral s_z^2 será pequeña y $\text{Var}(\hat{t}_y^L - t_y)$ crecerá.

Se observa que el único parámetro desconocido de la Ecuación 3.65 es σ^2, por lo que será necesario estimarlo como:

$$\hat{\sigma}^2 = (n-2)^{-1}\sum_s (y_i - a_L - b_L z_i)^2 \tag{3.66}$$

Luego, un estimador insesgado de la varianza de $\hat{t}_y^L$ bajo el modelo de población lineal simple es:

$$\widehat{\text{Var}}(\hat{t}_y^L) = \frac{N^2}{n}\hat{\sigma}^2 \left(\left(1 - \frac{n}{N}\right) + \frac{(\bar{z}_U - \bar{z}_s)^2}{(1 - n^{-1})s_z^2} \right) \tag{3.67}$$

Los correspondientes intervalos de confianza al $100(1-\alpha)\,\%$ se calculan en forma análoga a los modelos de población homogénea y lineal general (ver Ecuación 3.44).

3.7. Inferencia asistida por modelos y estimador de regresión

La inferencia asistida por modelos puede concebirse como una "solución de compromiso" entre los paradigmas basados en el diseño y en modelos (Dever y Valliant, 2018). De acuerdo con este enfoque, las poblaciones no se generan mediante un modelo superpoblacional. Sin embargo, los modelos constituyen herramientas útiles para describir a la población y así mejorar la precisión de las estimaciones basadas en el diseño. Así, si el modelo presenta un buen ajuste para una alta proporción de los valores que toma la variable de interés en la población, se reducirá sustancialmente la varianza de las estimaciones. Aun si el modelo es incorrecto, probablemente se logrará un cierto aumento en la eficiencia de las estimaciones.

El estimador de regresión (Särndal et al. 1992) es uno de los estimadores asistidos por modelos más frecuentemente utilizados. El mismo recurre a un modelo superpoblacional ξ para mejorar la precisión del estimador Horviz-Thompson.

3.7.1. Especificación del modelo

Se supone que ξ cumple los siguientes supuestos:

- $y_1, \ldots, y_N$ son realizaciones de las variables aleatorias independientes $Y_1, \ldots, Y_N$
- $E_\xi(Y_i) = \sum_{j=1}^{J} \beta_j x_{ji}$
- $\mathrm{Var}_\xi(Y_i) = \sigma_i^2$

donde E_ξ y Var_ξ denotan la esperanza y la varianza de la variable de interés respecto al modelo.

Cabe señalar que aunque se utiliza una notación ligeramente distinta a la de la Sección 3.5, el modelo utilizado es esencialmente el mismo al empleado para las poblaciones lineales generales. Sin embargo, en este caso, el modelo es aplicado bajo un enfoque asistido y no basado en modelos.

3.7.2. Forma del estimador

El estimador de regresión se define como:

$$\hat{t}_y^{reg} = \hat{t}_{y_\pi} + \sum_{j=1}^{J} \hat{B}_j (t_{x_j} - \hat{t}_{x_{j\pi}}) \tag{3.68}$$

donde $\hat{t}_{y_\pi}$ es el estimador de Horvitz-Thompson para t_y, t_{x_j} es el total de la covariable j y $\hat{t}_{x_{j\pi}}$ es su correspondiente estimador de Horvitz-Thompson. Por su parte, $\hat{B}_1, \ldots, \hat{B}_J$ son los componentes del vector:

$$\hat{\mathbf{B}} = (\hat{B}_1, \ldots, \hat{B}_j)' = \left(\sum\nolimits_s \frac{\mathbf{x}_i \mathbf{x}_i'}{\sigma_i^2 \pi_i} \right)^{-1} \sum\nolimits_s \frac{\mathbf{x}_i y_i}{\sigma_i^2 \pi_i} \tag{3.69}$$

siendo $\mathbf{x}_i$ un vector que contiene el valor de cada una de las covariables para el individuo i. Así, $\mathbf{x}_i$ constituye una fila de la matriz de covariables $\mathbf{X}$.

De esta manera, $\hat{\mathbf{B}}$ es un vector que estima la forma del estimador obtenido mediante el método de Mínimos Cuadrados Generalizados (MCG) de $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$ que se tendría si se conocieran todos los elementos de la población U (por ejemplo, mediante un censo). De esta forma, se tiene que:

$$\mathbf{B} = (B_1, \ldots, B_J)' = \left(\sum\nolimits_U \frac{\mathbf{x}_i \mathbf{x}_i'}{\sigma_i^2} \right)^{-1} \sum\nolimits_U \frac{\mathbf{x}_i y_i}{\sigma_i^2} \tag{3.70}$$

Del análisis de modelos lineales, se sabe que $\mathbf{B}$ es el mejor estimador insesgado de $\boldsymbol{\beta}$. Para estimarlo a partir de una muestra, basta con calcular $\hat{\mathbf{B}}$, el cual utiliza el estimador de Horvitz-Thompson para expandir los resultados a la población. Se tiene, entonces:

$$\mathbf{B} = \mathbf{T}^{-1} \mathbf{t} \tag{3.71}$$

donde $\mathbf{T} = \sum_U \frac{\mathbf{x}_i \mathbf{x}_i'}{\sigma_i^2}$ y $\mathbf{t} = \sum_U \frac{\mathbf{x}_i y_i}{\sigma_i^2}$. Estos parámetros se estiman como:

$$\hat{\mathbf{T}} = \sum\nolimits_s \frac{\mathbf{x}_i x_i'}{\sigma_i^2 \pi_i} \tag{3.72}$$

$$\hat{\mathbf{t}} = \sum\nolimits_s \frac{\mathbf{x}_i y_i}{\sigma_i^2 \pi_i} \tag{3.73}$$

Por lo tanto, $\hat{\mathbf{T}}$ y $\hat{\mathbf{t}}$ son estimadores de Hovitz-Thompson y, por lo tanto, son insesgados. Sin embargo, $\hat{\mathbf{B}}$ constituye una combinación no lineal de ambos

estimadores. En consecuencia, no es un estimador insesgado de $\mathbf{B}$. Sin embargo, usando la linealización de Taylor, se demuestra que sí es aproximadamente insesgado.

3.7.3. Expresiones alternativas del estimador

El estimador de regresión puede expresarse de múltiples formas que permiten visualizar distintos aspectos de interés.

3.7.3.1. Notación matricial

En primer lugar, $\hat{t}_y^{reg}$ puede expresarse en forma matricial, de forma de evitar la sumatoria dada por la Ecuación 3.68. Se tiene, entonces:

$$\hat{t}_y^{reg} = \hat{t}_{y_\pi} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}_\pi})'\hat{\mathbf{B}} \tag{3.74}$$

siendo $\mathbf{t_x} = (t_{x_1}, \ldots, t_{x_J})'$ y $\hat{\mathbf{t}}_{\mathbf{x}_\pi} = (\hat{t}_{x_{1\pi}}, \ldots, \hat{t}_{x_{J\pi}})'$ vectores que contienen los totales de las J variables auxiliares y sus estimadores de Horvitz-Thompson, respectivamente.

Se observa que para calcular el estimador de regresión, solamente es necesario conocer el valor que toman las variables auxiliares para los elementos de la muestra. A nivel poblacional, bastará con conocer el total de cada covariable. Por lo tanto, se trata de un enfoque más flexible que la inferencia basada en modelos.

3.7.3.2. Ajuste de ponderadores

En segundo lugar, partiendo de la expresión matricial de $\hat{t}_y^{reg}$, puede expresarse el estimador de regresión como la suma de las observaciones ponderadas por pesos ajustados. Sustituyendo las ecuaciones 3.72 y 3.73 en la Ecuación 3.74, se obtiene:

$$\begin{aligned}\hat{t}_y^{reg} &= \hat{t}_{y_\pi} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}_\pi})'\hat{\mathbf{T}}^{-1}\hat{\mathbf{t}} = \sum_s \frac{y_i}{\pi_i} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}_\pi})'\hat{\mathbf{T}}^{-1}\sum_s \frac{\mathbf{x}_i y_i}{\sigma_i^2 \pi_i} \\ &= \sum_s \frac{y_i}{\pi_i}\left(1 + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}_\pi})'\hat{\mathbf{T}}^{-1}\frac{\mathbf{x}_i}{\sigma_i^2}\right) = \sum_s g_{is}\frac{y_i}{\pi_i}\end{aligned} \tag{3.75}$$

siendo $g_{is} = 1 + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}_\pi})'\hat{\mathbf{T}}^{-1}\frac{\mathbf{x}_i}{\sigma_i^2}$.

De esta forma, cada elemento de la muestra se ajusta por un término que depende de la información auxiliar disponible y de la muestra seleccionada.

3.7.3.3. Corrección de predicciones

En tercer lugar, el estimador de regresión puede expresarse como la suma de las predicciones obtenidas mediante el modelo para cada elemento de la población, más un término de corrección que protege al estimador en caso de que el modelo asumido no sea correcto. Es decir, que al contrario de lo que ocurre para las demás expresiones, si se utiliza esta fórmula, deberá contarse con información auxiliar a nivel del marco muestral.

Partiendo de la Ecuación 3.74, se tiene:

$$\begin{aligned} \hat{t}_y^{reg} &= \sum\nolimits_s \frac{y_i}{\pi_i} + \left(\sum\nolimits_U \mathbf{x}_i - \sum\nolimits_s \frac{\mathbf{x}_i}{\pi_i}\right)' \hat{\mathbf{B}} \\ &= \sum\nolimits_s \frac{y_i}{\pi_i} + \sum\nolimits_U \mathbf{x}_i' \hat{\mathbf{B}} - \sum\nolimits_s \frac{\mathbf{x}_i' \hat{\mathbf{B}}}{\pi_i} \\ &= \sum\nolimits_U \hat{y}_i + \sum\nolimits_s \frac{y_i - \hat{y}_i}{\pi_i} = \sum\nolimits_U \hat{y}_i + \sum\nolimits_s \frac{e_i}{\pi_i} \end{aligned} \tag{3.76}$$

donde $\hat{y}_i$ es la predicción del modelo para la observación i y $e_i = y_i - \hat{y}_i$ es su residuo muestral asociado.

A partir de la ecuación anterior, se observa que el término de corrección constituye una estimación del error total cometido por el modelo. Para ello, se calcula el estimador de Horvitz-Thompson a partir de los residuos muestrales.

3.7.3.4. Ajuste hipotético

Sean $y_i^0 = \mathbf{x}_i' \mathbf{B}$ el ajuste hipotético de la población al modelo ξ para el elemento i y $E_i = y_i - y_i^0$ su residuo. Mediante estas definiciones, el estimador de regresión puede escribirse como:

$$\begin{aligned} \hat{t}_y^{reg} &= \sum\nolimits_s g_{is} \frac{y_i}{\pi_i} = \sum\nolimits_s g_{is} \frac{y_i^0 + E_i}{\pi_i} = \sum\nolimits_s g_{is} \frac{y_i^0}{\pi_i} + \sum\nolimits_s g_{is} \frac{E_i}{\pi_i} \\ &= \left(\sum\nolimits_s g_{is} \frac{\mathbf{x}_i'}{\pi_i}\right) \mathbf{B} + \sum\nolimits_s g_{is} \frac{E_i}{\pi_i} \end{aligned} \tag{3.77}$$

El primer sumando de la expresión anterior puede reescribirse como:

$$\begin{aligned}\sum_s g_{is}\frac{\mathbf{x}_i'}{\pi_i} &= \sum_s \left(1 + (\mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_{\mathbf{x}_\pi})'\hat{\mathbf{T}}^{-1}\frac{\mathbf{x}_i}{\sigma_i^2}\right)\frac{\mathbf{x}_i'}{\pi_i} \\ &= \sum_s \frac{\mathbf{x}_i'}{\pi_i} + (\mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_{\mathbf{x}_\pi})'\hat{\mathbf{T}}^{-1}\hat{\mathbf{T}} \\ &= \hat{\mathbf{t}}_{\mathbf{x}_\pi} + \mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_{\mathbf{x}_\pi} = \sum_U \mathbf{x}_i'\end{aligned} \tag{3.78}$$

Retomando la Ecuación 3.77, se arriba a la siguiente expresión:

$$\hat{t}_y^{reg} = \sum_U \mathbf{x}_i'\mathbf{B} + \sum_s g_{is}\frac{E_i}{\pi_i} = \sum_U y_i^0 + \sum_s g_{is}\frac{E_i}{\pi_i} \tag{3.79}$$

Se llega, así, a una última expresión para el estimador de regresión, la cual es "hipotética" ya que, salvo que se realice un censo, $\mathbf{B}$ deberá ser estimado mediante $\hat{\mathbf{B}}$.

3.7.4. Estimación de la varianza e intervalos de confianza

Dado que la base de la inferencia está en el diseño, la varianza del estimador de regresión se calcula con respecto al mismo. Como no es posible hacer esto en forma exacta, se recurre a la linealización de Taylor. De esta forma, se demuestra que, cerca de t_y, $\mathbf{t}_\mathbf{x}$, $\hat{\mathbf{T}}$ y $\mathbf{t}$, se cumple que:

$$\hat{t}_{y0}^{reg} = \sum_U \hat{y}_i^0 + \sum_s \frac{E_i}{\pi_i} \tag{3.80}$$

Calculando la varianza de la expresión anterior, se llega a:

$$\begin{aligned}AV(\hat{t}_y^{reg}) = \text{Var}(\hat{t}_{y0}^{reg}) &= \text{Var}\left(\sum_s \frac{E_i}{\pi_i}\right) \\ &= \sum\sum_U \Delta_{ij}\left(\frac{E_i}{\pi_i}\right)\left(\frac{E_j}{\pi_j}\right)\end{aligned} \tag{3.81}$$

donde $\text{Var}\left(\sum_U \hat{y}_i^0\right) = 0$ ya que es una constante bajo el paradigma basado en el diseño.

Finalmente, la expresión anterior puede estimarse como:

$$\widehat{\text{Var}}(\hat{t}_y^{reg}) = \sum\sum_s \frac{\Delta_{ij}}{\pi_{ij}}\left(g_{is}\frac{e_i}{\pi_i}\right)\left(g_{js}\frac{e_j}{\pi_j}\right) \tag{3.82}$$

Una vez estimada la varianza del estimador de regresión, es posible cons-

truir los correspondientes intervalos de confianza de la forma usual (ver Ecuación 3.44).

3.8. Comparación de estimadores basados y asistidos por modelos

A continuación, se demuestra que, bajo ciertas condiciones, se anula el término de corrección en el estimador de regresión, lo cual, para ciertos diseños, implicará que el estimador de regresión coincida con su "versión" basada en modelos, dada por el modelo de población lineal general desarrollado en la Sección 3.5.

3.8.1. Efecto de la estructura de varianza asumida

Sea λ' un vector constante de dimensión J. Si se cumple que:

$$\sigma_i^2 = \lambda' \mathbf{x}_i \quad \forall i = 1, \ldots, N \tag{3.83}$$

se anula el término de ajuste por los errores de predicción para cualquier muestra s (Särndal et al. 1992, p. 232). Se cumple, entonces, que $\sum_s \frac{e_i}{\pi_i} = 0$. La demostración de este resultado puede consultarse en el Apéndice 1.

Esto implica que cuando se trabaja con un modelo en el que la varianza de la variable de interés es proporcional a alguna de las covariables, el estimador de regresión resulta ser igual a la suma de las predicciones para cada elemento de la población.

En modelos de regresión lineal, en general se supone que la matriz de diseño está dada y no es aleatoria, de manera que $\mathbf{X}$ es "fija". Asimismo, suele asumirse que el término de error es homocedástico, es decir que $\text{Var}(\varepsilon_i) = \sigma^2$. Ambos supuestos implican que la varianza de la variable dependiente $\mathbf{Y}$ respecto al modelo también será constante. Para cada elemento i, se tiene:

$$\text{Var}(y_i|\mathbf{x}_i) = \text{Var}(\mathbf{x'}_i\boldsymbol{\beta} + \varepsilon_i|\mathbf{x}_i) = \text{Var}(\varepsilon_i) = \sigma^2 \quad \forall i = 1, \ldots, N \tag{3.84}$$

Si se trabaja con una regresión lineal con intercepto bajo las anteriores

condiciones, se cumplirá la condición establecida en la Ecuación 3.83:

$$\text{Var}(y_i|\mathbf{x}_i) = \sigma^2 = \begin{bmatrix}\sigma^2 & 0 & \cdots & 0\end{bmatrix} \begin{bmatrix}1\\ x_{1i}\\ x_{2i}\\ \vdots\\ x_{Ji}\end{bmatrix} = \lambda'\mathbf{x}_i \tag{3.85}$$

donde J la cantidad de variables auxiliares. El primer elemento de $\mathbf{x}_i$ vale uno debido a que el modelo tiene constante.

Tanto un modelo homogéneo como un modelo de población lineal general del tipo presentado en las secciones 3.4 y 3.5 constituyen modelos de regresión lineal con intercepto y homocedásticos. En consecuencia, verifican la condición dada por la Ecuación 3.83, independientemente del diseño muestral utilizado.

A priori, aunque se anule el término de error, el estimador de regresión asistido por modelos no tiene por qué ser igual a su análogo basado en modelos. Mientras que el estimador de regresión para un total será igual a la suma de las predicciones para cada individuo, los estimadores basados en modelos se compondrán de la suma de los verdaderos valores de la variable de interés para los elementos de la muestra más las predicciones para los elementos desconocidos. En cada caso, se tendrá respectivamente:

$$\begin{aligned}\hat{t}_y^{reg} &= \sum\nolimits_U \hat{y}_i\\ \hat{t}_y^{mod} &= \sum\nolimits_s y_i + \sum_r \hat{y}_i\end{aligned} \tag{3.86}$$

donde $\hat{t}_y^{mod}$ es un estimador basado en modelos y r es el conjunto de elementos de la población que no fueron seleccionados en la muestra.

Sin embargo, para diseños muestrales autoponderados, el estimador de regresión resulta ser igual a ciertos estimadores basados en modelos. Como se muestra a continuación, el diseño simple es un caso claro de esto. En cambio, para muestras extraídas a través de un diseño πps, la igualdad no se cumple.

3.8.2. Diseño simple

3.8.2.1. Modelo de población homogénea

Utilizar un modelo homogéneo equivale a trabajar con una regresión lineal sin variables auxiliares en la que el intercepto es igual a la media muestral de la variable de interés, $\bar{y}_s$. Recordando el supuesto de homocedasticidad y que la muestra fue extraída mediante un diseño simple, el parámetro a estimar para obtener el estimador de regresión toma la siguiente forma:

$$\begin{aligned}\hat{\mathbf{B}} &= \left(\sum_s \frac{\mathbf{x}_i \mathbf{x}_i'}{\sigma_i^2 \pi_i}\right)^{-1} \sum_s \frac{\mathbf{x}_i y_i}{\sigma_i^2 \pi_i} \\ &= \left(\sum_s \frac{1}{\sigma^2 (n/N)}\right)^{-1} \sum_s \frac{y_i}{\sigma^2 (n/N)} = \frac{1}{n} \sum_s y_i = \bar{y}_s\end{aligned} \tag{3.87}$$

siendo $\hat{\mathbf{B}}$ la estimación a partir de una muestra del único parámetro del modelo. Dado que no se cuenta con variables auxiliares, para cada individuo se tiene que $\mathbf{x}_i = 1$.

A partir de $\hat{\mathbf{B}}$, la estimación del total de la variable de interés es:

$$\hat{t}_y^{reg} = \sum_U \hat{y}_i = \left(\sum_U \mathbf{x}_i\right)' \hat{\mathbf{B}} = N\hat{\mathbf{B}} = N\bar{y}_s \tag{3.88}$$

que resulta ser igual al estimador basado en modelos bajo un modelo homogéneo, $\hat{t}_y^E$ (ver Ecuación 3.23). Es decir, que si se cuenta con una muestra simple y se recurre a un modelo homogéneo que se supone homocedástico, es indistinto trabajar con el paradigma basado en modelos o con el asistido por modelos.

3.8.2.2. Modelo de población lineal general

En este trabajo, se utilizó el modelo de población lineal general en su versión simple (con homocedasticidad). En este caso, el correspondiente estimador de regresión está asistido por un modelo con dos parámetros: un intercepto, $\hat{B}_0$, y un coeficiente $\hat{B}_1$ asociado a una única variable auxiliar x. Sin embargo, los resultados presentados a continuación son extendibles a un modelo de regresión lineal múltiple.

Bajo un diseño simple, el vector de parámetros estimados $\hat{\mathbf{B}}$ con las estima-

ciones MCO obtenidas a partir de una muestra:

$$\begin{aligned}\hat{\mathbf{B}} = \begin{bmatrix}\hat{B}_0 \\ \hat{B}_1\end{bmatrix} &= \left(\sum\nolimits_s \frac{\mathbf{x}_i\mathbf{x}_i'}{\sigma^2(n/N)}\right)^{-1} \sum\nolimits_s \frac{\mathbf{x}_i y_i}{\sigma^2(n/N)} \\ &= \left(\sum\nolimits_s \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum\nolimits_s \mathbf{x}_i y_i\end{aligned} \tag{3.89}$$

Por lo tanto, la estimación de los parámetros coincidirá con la obtenida bajo la inferencia basada en modelos. A su vez, esto implica que el estimador de regresión será igual al estimador de población lineal general, $\hat{t}_y^L$, obtenido en la Ecuación 3.63:

$$\hat{t}_y^{reg} = \sum_U \hat{y}_i = \left(\sum_U \mathbf{x}_i\right)' \hat{\mathbf{B}} = \sum_U (\hat{B}_0 + \hat{B}_1 x_i) = N(\hat{B}_0 + \hat{B}_1 \bar{x}_U) \tag{3.90}$$

En suma, al igual que ocurre con el modelo homogéneo, bajo un muestreo simple y el supuesto de varianza constante para la variable de interés, el estimador de regresión coincide con el estimador basado en un modelo de población lineal general.

3.8.3. Diseño πps

Como se vio en la Subsección 3.2.3, el muestreo πps se caracteriza por el hecho de que cada elemento de la población tiene una probabilidad de ser seleccionado en una muestra sin remplazo proporcional a su valor en términos de una variable auxiliar z_i conocida para toda la población. Por lo tanto, para poder extraer una muestra bajo este diseño, es necesario contar con al menos una covariable a nivel del marco muestral. De lo contrario, no será posible dar más peso a una observación que a otra y se retornará a un diseño simple:

$$\pi_i = \frac{1}{\sum_U 1} n = \frac{n}{N} \quad \forall i = 1, \ldots, N \tag{3.91}$$

Para evitar este resultado trivial, en lo que sigue se asume que en la etapa de selección de la muestra se utilizó una cierta covariable, aun si en la etapa de estimación se aplica un modelo que no haga uso de ella.

3.8.3.1. Modelo de población homogénea

Al igual que ocurre para un diseño simple, bajo el modelo homogéneo el vector de covariables disponibles para asistir la estimación se transforma en un escalar igual a uno. Sin embargo, al cambiar las probabilidades de inclusión, la forma de $\hat{\mathbf{B}}$ se verá modificada. Si se mantiene el supuesto de homocedasticidad, se tiene que:

$$\hat{\mathbf{B}} = \left(\sum_s \frac{1}{\sigma^2 \frac{z_i}{\sum_U z_i} n} \right)^{-1} \sum_s \frac{y_i}{\sigma^2 \frac{z_i}{\sum_U z_i} n} = \left(\sum_s \frac{1}{z_i} \right)^{-1} \sum_s \frac{y_i}{z_i} \tag{3.92}$$

donde $\sum_U z_i$ es una constante que puede ser eliminada de la expresión.

A partir de un razonamiento análogo al realizado en la Ecuación 3.88, se obtiene el estimador de regresión para una muestra πps:

$$\hat{t}_y^{reg} = \sum_U \hat{y}_i = N\hat{\mathbf{B}} = N \left(\sum_s \frac{1}{z_i} \right)^{-1} \sum_s \frac{y_i}{z_i} \tag{3.93}$$

Esta expresión no coincide con la forma del estimador basado en modelos para un modelo homogéneo. Por ende, tiene sentido comparar el desempeño de ambos.

3.8.3.2. Modelo de población lineal general

Nuevamente, si bien los resultados que se presentan a continuación son válidos para un modelo de regresión múltiple, para facilitar los cálculos, sólo se detalla su versión simple. Si se utiliza un modelo de población lineal general con varianza de la variable de interés constante, la estimación del vector $\hat{\mathbf{B}}$ también resulta ser diferente a la hallada para un diseño simple:

$$\begin{aligned} \hat{\mathbf{B}} = \begin{bmatrix} \hat{B}_0 \\ \hat{B}_1 \end{bmatrix} &= \left(\sum_s \frac{\mathbf{x}_i \mathbf{x}_i'}{\sigma^2 \frac{z_i}{\sum_U z_i} n} \right)^{-1} \sum_s \frac{\mathbf{x}_i y_i}{\sigma^2 \frac{z_i}{\sum_U z_i} n} \\ &= \left(\sum_s \frac{\mathbf{x}_i \mathbf{x}_i'}{z_i} \right)^{-1} \sum_s \frac{\mathbf{x}_i y_i}{z_i} \end{aligned} \tag{3.94}$$

En la medida que los parámetros del modelo $\hat{B}_0$ y $\hat{B}_1$ no coincidirán por el método MCO, también el estimador $\hat{t}_y^{reg} = N(\hat{B}_0 + \hat{B}_1 \bar{x}_U)$ será diferente al hallado en la Ecuación 3.63. Al igual que para el modelo homogéneo, los estimadores basados y asistidos por modelos bajo un diseño πps serán diferentes

y podrán ser comparados entre sí.

Capítulo 4

Descripción de poblaciones

Como fue detallado en el Capítulo 1, se trabajó con dos tipos de poblaciones. En primer lugar, a partir de simulaciones, se replicó una situación teórica "ideal" en la que el modelo superpoblacional es conocido y, por lo tanto, no existen problemas de especificación del mismo. A lo largo de este trabajo, a este tipo de poblaciones se las llama indistintamente "ficticias" o "simuladas".

En segundo lugar, se utilizó una población real conocida como MU281. Aunque en este caso se conoce el valor que toma la variable de interés para cada uno de los elementos de la población, la diferencia fundamental con la población simulada es que el modelo superpoblacional que los generó es desconocido. En consecuencia, la validez de las predicciones estará condicionada a la selección de un modelo de estimación robusto.

A continuación, se presentan ambos tipos de poblaciones, los modelos utilizados para la predicción de totales y la forma de los estimadores que se desprenden de dichos modelos.

4.1. Poblaciones simuladas

4.1.1. Modelo superpoblacional y procedimiento de simulación

Para analizar el comportamiento de los estimadores basados en modelos cuando el modelo superpoblacional es conocido, se simularon 5.000 poblaciones de tamaño 300 a partir de una regresión lineal simple. Es importante destacar que tanto la cantidad de réplicas como el tamaño de la población fueron

elegidas de forma arbitraria. Se buscó trabajar con una cantidad de simulaciones relativamente grande. En lo que respecta al tamaño de la población, a fin de no generar distorsiones en la comparación de los resultados, se optó por generar una cantidad de elementos similar a la de la población MU281.

El modelo superpoblacional se definió como sigue. Sea y la variable de interés cuyo total se desea estimar, x una variable auxiliar conocida para toda la población y ε un término de ruido, se tiene entonces:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \forall i = 1, \ldots, 300 \tag{4.1}$$

En esta situación "ideal" no sólo se cumple que el modelo superpoblacional es completamente conocido, sino que puede ser estimado sin errores de especificación a partir de una muestra en la etapa de predicción debido a que la variable que la explica, x, es justamente una covariable conocida para toda la población.

A partir de la Ecuación 4.1, queda claro que, para generar cada población, es necesario obtener 300 realizaciones de las variables aleatorias X y ε [1]. Además, se debe definir el valor de los parámetros β_0 y β_1. Se definió que:

- $X_1, X_2, \ldots, X_{300}$ son independientes e idénticamente distribuidas (iid). Se consideró una distribución normal con media 10 y varianza 1, respectivamente:
$$X_i \sim \text{Normal}(10, 1) \tag{4.2}$$

- $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{300}$ es un término de error desconocido que hace que la relación entre la variable auxiliar x y la variable de interés y no sea determinística. Sus realizaciones fueron obtenidas a partir de una distribución normal estándar:
$$\varepsilon_i \sim \text{Normal}(0, 1) \tag{4.3}$$

- Los valores de los parámetros de la regresión fueron fijados en $\beta_0 = 1$ y $\beta_1 = 2$.

Una vez generado el conjunto de variables auxiliares, se las dejó de tratar como variables auxiliares y se las supuso fijas. Es decir, que la distribución normal de media igual a 10 y varianza igual a 1 fue utilizada sólo como

[1] La notación con mayúsculas hace referencia a que se trata de variables aleatorias y no de su realización.

una herramienta para obtener el regresor fijo x en la regresión lineal simple superpoblacional.

El valor concreto de estos parámetros fue elegido de forma arbitraria, con el único objetivo de obtener conjuntos de datos generados a través de un modelo conocido. En la medida en que estas características inevitablemente afectarán los resultados obtenidos, es importante tenerlos en cuenta en su interpretación. En este trabajo, no se utilizaron variaciones del modelo superpoblacional ya que el foco fue puesto en el efecto de conocer o no el modelo superpoblacional sobre la inferencia basada en modelos, y no en la forma puntual de dichos modelos.

4.1.2. Distribución de la variable de interés y de su total

La distribución de Y variará dependiendo de si se cuenta o no con información acerca de las realizaciones de X. A continuación, se presenta, para ambos casos, la distribución de la variable de interés, así como de su total poblacional.

4.1.2.1. Distribución sin información auxiliar

Cuando no se cuenta con información auxiliar, Y constituye una combinación lineal de variables aleatorias normales e independientes, X y ε, por lo que también será normal. Los parámetros que la caracterizan se calculan como sigue:

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 x_i + \varepsilon_i) = 21 \quad \forall i = 1, \ldots, 300 \\ \text{Var}(Y_i) &= \text{Var}(\beta_0 + \beta_1 X_i + \varepsilon_i) = 5 \quad \forall i = 1, \ldots, 300 \end{aligned} \tag{4.4}$$

Así, mientras X no sea incorporada en el cálculo de los momentos de Y, esta variable tendrá una distribución normal con media de 21 y varianza igual a 5:

$$Y \sim \text{Normal}(21, 5) \tag{4.5}$$

A modo de ejemplo, en la Figura 2.0.1 del Apéndice 2, se presentan los resultados para una de las 5.000 poblaciones generadas. Dicha figura permite visualizar la normalidad de X y de Y, así como la correlación existente entre ambas variables.

Por otra parte, el total poblacional de la variable de interés es una variable aleatoria construida mediante la suma de 300 realizaciones independientes de

Y. Esto implica que, si no se incorpora información auxiliar, su distribución corresponderá a una normal de media 6.300 y varianza 1.500:

$$\begin{aligned} E(t_Y) &= E\left(\sum_{i=1}^{300} Yi\right) = 6.300 \\ \text{Var}(t_Y) &= \text{Var}\left(\sum_{i=1}^{300} Yi\right) = 1.500 \end{aligned} \tag{4.6}$$

Se concluye, entonces, que:

$$t_Y \sim \text{Normal}(6.300, 1.500) \tag{4.7}$$

El cálculo teórico de los momentos de t_Y no condicionado al valor particular de x es útil ya que brinda una primera idea de cómo se comportarán las 5.000 poblaciones simuladas. Si bien los totales para la variable de interés serán diferentes para cada réplica, es esperable que exhiban una distribución simétrica centrada en 6.300 aproximadamente.

4.1.2.2. Distribución con información auxiliar

Cuando se cuenta con información auxiliar, el conjunto de $x_1, x_2, \ldots, x_{300}$ se supone fijo, lo que hace posible utilizarlo como una variable independiente en el marco de una regresión lineal clásica. Se tiene, entonces:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = 1 + 2x_i + \varepsilon_i \quad \varepsilon_i \sim \text{Normal}(0, 1) \tag{4.8}$$

En este caso, x se entiende como información que permite mejorar las predicciones del valor que tomará la variable dependiente y para cada elemento de la población. Por este motivo, tanto la esperanza como la varianza de la variable de interés deben calcularse condicionadas a x:

$$\begin{aligned} E(y_i|x_i) &= E(\beta_0 + \beta_1 x_i + \varepsilon_i|x_i) = 1 + 2x_i \quad \forall i = 1, \ldots, 300 \\ \text{Var}(y_i|x_i) &= \text{Var}(\beta_0 + \beta_1 x_i + \varepsilon_i|x_i) = \text{Var}(\varepsilon_i) = 1 \quad \forall i = 1, \ldots, 300 \end{aligned} \tag{4.9}$$

Al igual que cuando no se cuenta con información auxiliar, y tendrá una distribución normal debido a que constituye una combinación lineal de ε, pero

esta vez estará caracterizada por los parámetros calculados en la Ecuación 4.9:

$$y_i|x_i \sim \text{Normal}(1 + 2x_i, 1) \tag{4.10}$$

Por lo tanto, si bien las 300 realizaciones de $y_i|x_i$ serán independientes, no serán idénticamente distribuidas, sino que su esperanza dependerá del valor que tome x_i para cada elemento de la población.

Puesto que se trata de una combinación lineal de normales independientes, el total de la variable de interés en presencia de información auxiliar, t_y, tendrá una distribución normal. Su esperanza y varianza serán:

$$\begin{aligned} E(t_y|x_i) &= E\left(\sum_{i=1}^{300} y_i|x_i\right) = \sum_{i=1}^{300} 1 + 2x_i = 300 + 2t_x = 300(1 + 2\bar{x}) \\ \text{Var}(t_y|x_i) &= \text{Var}\left(\sum_{i=1}^{300} y_i|x_i\right) = \sum_{i=1}^{300} \text{Var}(y_i|x_i) = 300 \end{aligned} \tag{4.11}$$

En resumen, la esperanza de la distribución de t_y condicional a la variable auxiliar x tendrá una distribución normal del tipo:

$$t_y|x_i \sim \text{Normal}(300(1 + 2\bar{x}), 300) \tag{4.12}$$

4.1.3. Forma de los estimadores

A continuación, se presenta la forma concreta que tomarán los estimadores en las poblaciones simuladas para el total de una variable de interés y para su varianza bajo los dos modelos considerados, a saber, el modelo homogéneo y el modelo de población lineal general. Para ello, se utilizan las fórmulas y expresiones presentadas en el Capítulo 3.

4.1.3.1. Modelo de población homogénea

Como fue dicho, el modelo homogéneo es con frecuencia aplicado cuando no se posee información auxiliar desagregada a nivel de cada elemento de la población, y equivale a trabajar con una regresión sin variables independientes. Bajo estas condiciones, se cuenta únicamente con una constante estimada mediante la media muestral. Entonces, el estimador para las poblaciones simuladas bajo

el modelo homogéneo será de la forma:

$$\hat{t}_y^E = \frac{N}{n} t_{y_s} = N\bar{y}_s = 300\bar{y}_s \tag{4.13}$$

donde $\hat{t}_y^E$ es el estimador del total para el modelo homogéneo, t_{y_s} es el total de la variable de interés dentro de la muestra extraída e $\bar{y}_s$ es el promedio muestral. Por su parte, N y n son los tamaños de la población y de la muestra respectivamente.

Para las poblaciones simuladas, es claro que la forma del modelo superpoblacional es diferente al modelo homogéneo estimado a partir de una muestra, con lo cual inevitablemente se incurrirá en un error de especificación. Sin embargo, como se verá en el Capítulo 6, al trabajarse con muestras "buenas" que reflejan la verdadera variabilidad de la población, las predicciones obtenidas son aceptables.

Por su parte, la varianza del error de estimación tendrá la siguiente forma:

$$\begin{aligned} \text{Var}(\hat{t}_y^E - t_y) &= \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sigma^2 = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\text{Var}(yi) \\ &= \frac{300^2}{30}\left(1 - \frac{30}{300}\right)(5) = 13.500 \end{aligned} \tag{4.14}$$

A diferencia de lo que ocurre en el modelo de población lineal general, en este caso no se utiliza información relativa a la variable auxiliar x que brinde información acerca del comportamiento de y. Por este motivo, en el desarrollo del modelo homogéneo, Chambers y Clark (2012) definen σ^2 como la varianza incondicional de las y_i (ver Ecuación 3.19). En el marco de las poblaciones simuladas, esto implica que σ^2 sea igual a 5. Esto es razonable debido a que, si bien la población fue generada mediante un modelo superpoblacional que hace uso de x, esta variable no fue incorporada como información auxiliar en el cálculo de las predicciones. Por consiguiente, es esperable que la variabilidad de las estimaciones sea mayor que si sí se utilizara dicha información.

Para concluir, si se estima σ^2 mediante la varianza muestral de y, s_y^2, la varianza del estimador $\hat{t}_y^E$ puede ser aproximada a través de la estimación de la varianza del error:

$$\widehat{\text{Var}}(\hat{t}_y^E) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)s_y^2 = \frac{300^2}{30}\left(1 - \frac{30}{300}\right)s_y^2 = 2.700 s_y^2 \tag{4.15}$$

4.1.3.2. Modelo de población lineal general

En el caso del modelo de población lineal general, se hace uso de las covariables disponibles incluyéndolas en una regresión lineal estimada a través de una muestra. Si se considera un modelo lineal con x como variable independiente, la forma del modelo superpoblacional coincidirá con la del modelo utilizado para predecir los valores de y de los elementos de la población no pertenecientes a la muestra. Por lo tanto, no habrá problemas de especificación del modelo en el marco de las poblaciones simuladas.

Para cada población simulada, se estima el siguiente modelo lineal mediante el método de Mínimos Cuadrados Ordinarios (MCO):

$$\begin{aligned} &y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \\ &\varepsilon_i \sim Normal(0, \sigma^2) \quad \forall i = 1, \ldots, 300 \end{aligned} \tag{4.16}$$

A partir de las estimaciones de β_0 y β_1 para cada simulación, se estima t_y siguiendo lo establecido en la Ecuación 3.63:

$$\hat{t}_y^L = N(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_U) = 300(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_U) \tag{4.17}$$

donde $\hat{t}_y^L$ es el estimador del total bajo el modelo de población lineal general y $\bar{x}_U$ es la media poblacional para la variable x, conocida para cada elemento de la población.

Sea s_x^2 la varianza muestral de la variable auxiliar, la varianza del error de estimación viene dada por la Ecuación 3.65:

$$\begin{aligned} \mathrm{Var}(\hat{t}_y^L - t_y) &= \frac{N^2}{n}\sigma^2 \left(\left(1 - \frac{n}{N}\right) + \frac{\bar{x}_U - \bar{x}_s}{(1 - n^{-1})s_x^2} \right) \\ &= \frac{N^2}{n}\mathrm{Var}(y_i|x_i) \left(\left(1 - \frac{n}{N}\right) + \frac{\bar{x}_U - \bar{x}_s}{(1 - n^{-1})s_x^2} \right) \\ &= \frac{300^2}{30}(1) \left(\left(1 - \frac{30}{300}\right) + \frac{\bar{x}_U - \bar{x}_s}{(1 - 1/30)s_x^2} \right) \end{aligned} \tag{4.18}$$

Para el modelo de población lineal general, la varianza del error de estimación varía en función de la muestra que se extraiga. Como se discutió en el marco teórico (ver Subsección 3.6.4), cuanto más se asemeje la media muestral a la poblacional, menor será la varianza. En el caso extremo, si se eligiera una muestra perfectamente balanceada, se anularía el término $\bar{x}_U - \bar{x}_s$, con lo cual

la varianza del error de estimación sería de 2.700, muy inferior al valor de 13.500 hallado para el modelo homogéneo. Sin embargo, en este trabajo se consideraron solamente estimadores basados en modelos para muestras ya dadas.

Al igual que para el modelo homogéneo, la varianza del estimador $\hat{t}_y^L$ puede estimarse a partir de la expresión para la varianza del error. Para ello, se estima σ^2 mediante MCO:

$$\widehat{\text{Var}}(y_i|x_i) = (n-2)^{-1} \sum_s (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{4.19}$$

Siguiendo la Ecuación 3.67, la varianza del estimador del total, $\widehat{\text{Var}}(\hat{t}_y^L)$, será, entonces:

$$\begin{aligned} \widehat{\text{Var}}(\hat{t}_y^L) &= \frac{N^2}{n} \widehat{\text{Var}}(y_i|x_i) \left(\left(1 - \frac{n}{N}\right) + \frac{\bar{x}_U - \bar{x}_s}{(1-n^{-1})s_x^2} \right) \\ &= \frac{300^2}{30} \widehat{\text{Var}}(y_i|x_i) \left(\left(1 - \frac{30}{300}\right) + \frac{\bar{x}_U - \bar{x}_s}{(1-1/30)s_x^2} \right) \end{aligned} \tag{4.20}$$

4.2. Población MU281

4.2.1. Descripción de la población

Para analizar el desempeño de los estimadores de totales basados en modelos para una población real, se eligió el conjunto de datos "MU284", presentado a modo de ejemplo en Särndal et al. (1992) y en el paquete *sampling* del software estadístico R (Tillé y Matei, 2021). El mismo contiene indicadores relevados entre 1975 y 1984 para los 284 municipios de Suecia tales como ingresos fiscales, población y cantidad de empleados, entre otros.

Se optó por trabajar con la población MU284 por varios motivos:

- Es una población muy conocida y utilizada múltiples veces como ejemplo por Särndal et al. (1992) y en la documentación de la librería *sampling*.
- Tiene un tamaño relativamente pequeño y una cantidad de indicadores manejable. Sin embargo, se cuenta con información suficiente como para definir una variable de interés y múltiples variables auxiliares.
- Se cuenta con varias variables cuantitativas. Esto se debe a que, si bien es posible incorporar variables cualitativas como regresores en modelos

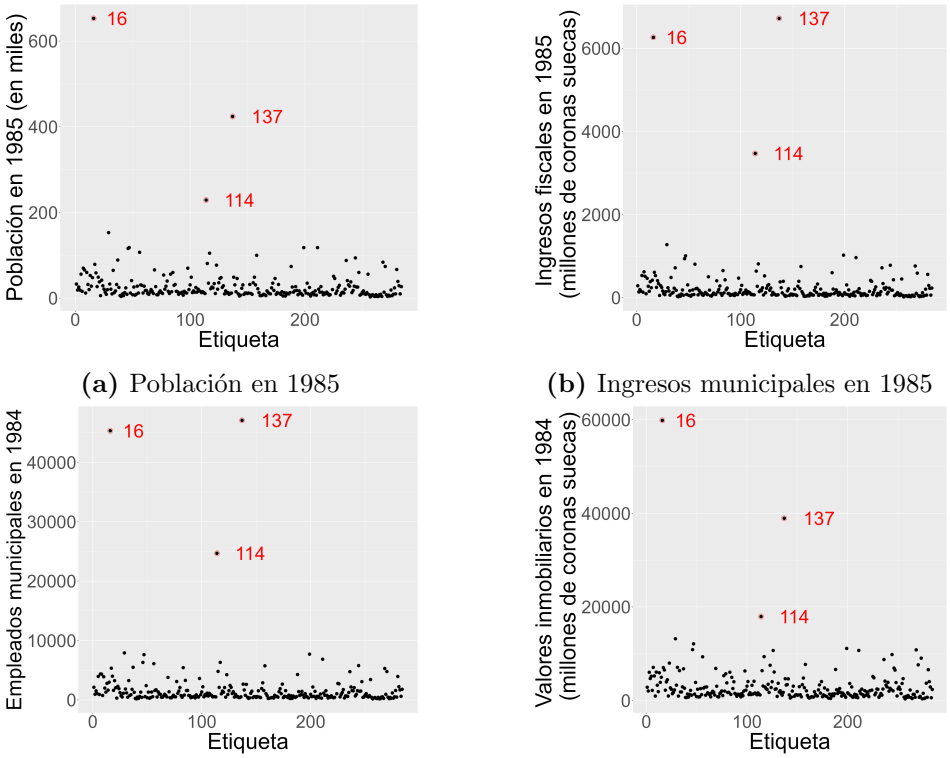
lineales, se trata de una metodología diseñada esencialmente para explicar y/o predecir variables cuantitativas.

En la Tabla 4.2.1, se presenta el diccionario de variables para la población MU284. La base se compone de 284 filas correspondientes a los 284 elementos de la población y 11 columnas: ocho indicadores, un número de identificación para los 284 elementos de la población y códigos de región y *cluster*. Dado que no se cuenta con información respecto a cómo fueron construidas estas dos últimas variables, se las excluyó del análisis. Por otro lado, no se conoce a qué municipalidad corresponde cada número de identificación.

Variable	**Descripción**
id	Nro. de identificación
pob85	Población de 1985 (miles de habitantes)
pob75	Población de 1975 (miles de habitantes)
ing_muni_85	Ingresos fiscales municipales de 1985 (millones de coronas)
esca_cons_82	Nro. de escaños conservadores en el Consejo Municipal
esca_soc_82	Nro. de escaños social-democrátas en el Consejo Municipal
esca_tot_82	Nro. total de escaños en el Consejo Municipal
emp_mun_84	Nro. de empleados municipales en 1984
val_inmob_84	Valores inmobiliarios de 1984 (millones de coronas)
region	Código de región geográfica
cluster	Cluster

Tabla 4.2.1 – Variables disponibles en la base MU284.

Como se muestra en la Figura 4.2.1, un primer análisis descriptivo de la población MU284 muestra que existen tres observaciones que exhiben valores extremos para varias de las variables disponibles. Dichos municipios son Estocolmo, Malmö y Gotemburgo, identificados con las etiquetas 16, 114 y 137 (no necesariamente en ese orden). En la medida en que estos tres municipios comprenden tres de las ciudades más grandes de Suecia, es normal que tengan características diferentes al resto y pueden ser clasificados como valores atípicos.



(a) Población en 1985

(b) Ingresos municipales en 1985

(c) Empleados municipales en 1984

(d) Valores inmobiliarios en 1984

Figura 4.2.1 – Distribución de variables seleccionadas para la población MU284. Las observaciones se ordenan según su número de identificación.

Dado que estos municipios pueden distorsionar el análisis, en varias de sus aplicaciones, Särndal et al. (1992) opta por excluirlos y denomina a la población restante "MU281". Teniendo en cuenta que las observaciones atípicas pueden ser influyentes y modificar fuertemente la estimación de los coeficientes en modelos lineales, en este trabajo también se decidió remover estos tres elementos de la base.

Si bien en la práctica no es posible identificar observaciones atípicas en términos de la variable de interés debido a que la misma es desconocida, sí es posible y aconsejable estudiar la distribución de las variables auxiliares. En caso de hallarse valores extremos, deberá analizarse si considerar muestras que contengan dichos elementos a la hora de estimar el modelo seleccionado.

Por último, se eligió como variable de interés a los ingresos fiscales de los municipios (variable *ing_muni_85*). Por su parte, se supuso que los restantes siete indicadores (variables *pob85*, *pob75*, *esca_cons_82*, *esca_soc_82*, *emp_mun_84* y *val_inmob_84*) constituyen información auxiliar conocida para cada elemento de la población. Aunque esta elección fue esencialmente arbitraria, tiene la

ventaja de que el total recaudado tiene un sentido económico real que podría ser relevante estimar en la práctica.

Si bien se analizará más a fondo la distribución de estas variables en la etapa de selección de la información auxiliar bajo el modelo de población lineal general, es útil brindar un primer panorama acerca de las características de los datos con los que se trabajará. Con este propósito, a continuación se presentan algunos estadísticos de resumen para las ocho variables consideradas en la población MU281 (ver Tabla 4.2.2).

Variable	**Mínimo**	**Q1**	**Mediana**	**Media**	**Q3**	**Máximo**
pob85	3	10	16	25	30	153
pob75	4	10	15	24	28	138
ing_muni_85	21	67	113	189	226	1277
esca_cons_82	1	6	8	9	11	24
esca_soc_82	8	17	21	22	26	46
esca_tot_82	31	41	45	47	49	85
emp_mun_84	173	485	784	1381	1579	7910
val_inmob_84	347	1136	1833	2695	3298	13 205

Tabla 4.2.2 – Estadísticos de resumen para la población MU281.

En primer lugar, las fuertes diferencias entre la media y la mediana dan cuenta de una clara asimetría hacia la derecha para varios de los indicadores disponibles como la población en 1975 y 1985, los ingresos fiscales, la cantidad de empleados municipales y los valores inmobiliarios del municipio. Por su parte, las variables referidas a la cantidad de escaños de los distintos partidos en el Consejo Municipal exhiben un comportamiento un poco más simétrico, con una media bastante similar a la mediana. Probablemente, esto se deba a que, a diferencia de las demás, estas variables no hacen referencia al tamaño del municipio y tienen un límite superior. Es decir, que en la medida en que existen unos pocos municipios mucho más grandes que otros en términos de población e ingresos, es razonable que las variables vinculadas a estos aspectos sean asimétricas.

Para la variable de interés, *ing_muni_85*, la media fue de 189, bastante superior a la mediana de 113. Lo que es más, el máximo se ubicó en 1.277, extremadamente alejado del valor del tercer cuartil, de 226. Si bien se eliminaron las tres observaciones más alejadas del resto, cuya recaudación en todos los casos superaba los 3.000 millones de coronas suecas (ver Figura 4.2.1), siguen

existiendo municipios con ingresos muy por encima del resto. Por lo tanto, para que los modelos arrojen buenas predicciones, se deberá trabajar con muestras que efectivamente capten esta variabilidad. Este aspecto será retomado en profundidad en el análisis de muestras truncadas, las cuales buscan replicar el efecto de la no respuesta no ignorable sobre los estimadores basados en modelos.

4.2.2. Forma de los estimadores

En el marco de la inferencia basada en modelos, la población MU281 puede ser entendida como una única realización disponible de un modelo superpoblacional desconocido. A su vez, la población generada determina el valor realizado del total poblacional. De esta manera, si bien se conoce cada elemento de la población, se posee sustancialmente menos información que en el caso ideal de las poblaciones simuladas. Por consiguiente, es probable que el modelo propuesto sea incorrecto, y por ello es fundamental ajustar un modelo robusto a errores de especificación.

4.2.2.1. Modelo de población homogénea

Al igual que para las poblaciones simuladas, bajo el modelo homogéneo, el total de la variable de interés se estimó como el producto del tamaño de la población, $N = 281$, por la media muestral:

$$\hat{t}^{E}_{\text{ingresos}} = \frac{281}{n} t_{\text{ingresos}_s} \tag{4.21}$$

donde t_{ingresos_s} es el total de los ingresos municipales en la muestra seleccionada.

Por su parte, la varianza del error de estimación tendrá la siguiente forma:

$$\text{Var}(\hat{t}^{E}_{\text{ingresos}} - t_{\text{ingresos}}) = \frac{281^2}{n} \left(1 - \frac{n}{281}\right) \sigma^2 \tag{4.22}$$

Es relevante señalar que, a pesar de que se conocen todos los elementos de la población MU281, la distribución de la variable aleatoria que la generó es desconocida. En consecuencia, la varianza de la variable de interés, σ^2, también lo será. No es posible obtener, entonces, el valor real de la varianza del error de estimación a partir de una población, la cual se compone de un conjunto de realizaciones de un modelo superpoblacional.

Finalmente, se estima la varianza del estimador del total de ingresos fiscales

a partir de la varianza del error de estimación. Para ello, se estima la varianza del ingreso mediante su varianza muestral corregida, s^2_{ingresos}:

$$\widehat{\text{Var}}(\hat{t}^E_{\text{ingresos}}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) s^2_{\text{ingresos}} \tag{4.23}$$

4.2.2.2. Modelo de población lineal general

Al igual que para las poblaciones simuladas, para la población MU281 se trabajó con un modelo de regresión lineal simple. De esta manera, fue posible obtener una primera aproximación acerca de cuánto pueden mejorarse las estimaciones del modelo homogéneo a partir de relativamente poca información auxiliar cuando el modelo superpoblacional es desconocido.

Para seleccionar la variable auxiliar a incluir en el modelo, se consideraron las correlaciones de Pearson de los ingresos fiscales con los restantes siete indicadores. Para ello, se tuvieron en cuenta los 281 elementos de la población. Es evidente que, en la práctica, la variable de interés será desconocida y por lo tanto será imposible realizar este ejercicio. En dicho caso, una alternativa sería considerar las correlaciones a nivel de muestra. Si se cuenta con una muestra "buena" que refleje la verdadera variabilidad de la población en términos de la variable de interés, las correlaciones muestrales y poblacionales no deberían ser demasiado diferentes.

En la Figura 4.2.2, se presentan los coeficientes de correlación lineal para variables seleccionadas de la base MU281 junto con sus correspondientes diagramas de dispersión. También se incluyen las densidades estimadas de cada variable. A partir de éstas últimas, se observan las asimetrías anticipadas en la sección anterior (ver Tabla 4.2.2).

Asimismo, se advierte una alta correlación del ingreso con todas las variables presentadas, a saber, la población, la cantidad total de escaños en el Consejo Municipal, la cantidad de empleados municipales y los valores inmobiliarios. Esto tiene sentido en la medida en que todas estas variables se vinculan al tamaño del municipio. Para la cantidad de habitantes y el número de empleados municipales, las correlaciones con los ingresos fiscales superan el 99 %.

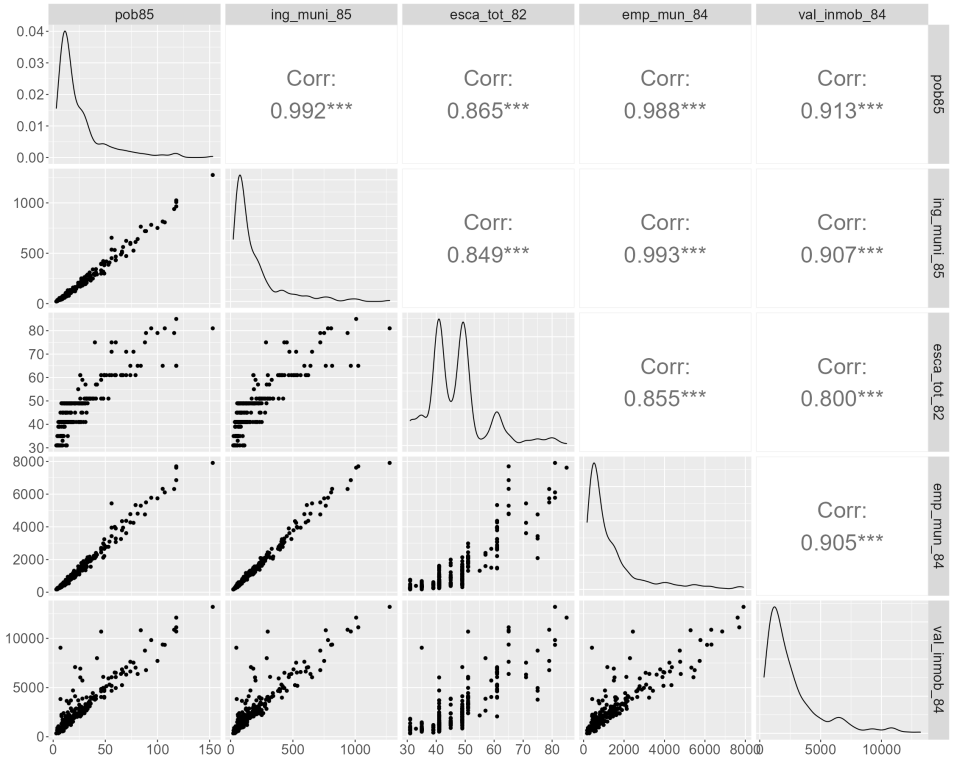


Figura 4.2.2 – Correlaciones lineales, diagramas de dispersión y densidades estimadas para variables seleccionadas.

Por su parte, la correlación de los ingresos fiscales con los valores inmobiliarios es algo menos trivial desde un punto de vista conceptual, y da cuenta de que a mayor valor de las propiedades inmobiliarias y/o mayor cantidad de ellas, mayor será la recaudación de impuestos. Así, la coyuntura del mercado inmobiliario podría ser utilizado como un predictor de la recaudación de los municipios. Claramente, los impuestos a las propiedades no son la única fuente de recaudación de los municipios, con lo cual es razonable que la correlación lineal entre los ingresos fiscales y los valores inmobiliarios sea algo más baja que para las otras variables (90,7 %).

Más allá de su relevancia teórica, la variable de valores inmobiliarios es interesante justamente por el hecho de no exhibir una correlación lineal extremadamente alta con los ingresos fiscales. Si se la utiliza como variable auxiliar, el modelo estimado a partir de cada muestra no estará perfectamente especificado. Bajo estas condiciones, si las predicciones del modelo resultan ser razonablemente buenas, se tendrá un primer elemento para afirmar que los estimadores dados por el modelo de población lineal general son robustos a errores de especificación.

Por estos motivos, se seleccionó a los valores inmobiliarios como variable independiente y se definió el siguiente modelo de regresión lineal:

$$\begin{aligned} \text{ingresos}_i &= \beta_0 + \beta_1 \text{valores}_i + \varepsilon_i \quad \forall i = 1, \ldots, 281 \\ \varepsilon_i &\sim \text{Normal}(0, \sigma^2) \end{aligned} \tag{4.24}$$

Al igual que para la población simulada, los coeficientes β_0 y β_1 se estimaron mediante el método MCO. A partir de dichas estimaciones, se obtuvo la predicción para cada elemento de la población y se estimó el total de ingresos fiscales:

$$\hat{t}^L_{\text{ingresos}} = N(\hat{\beta}_0 + \hat{\beta}_1 \overline{\text{valores}}_U) = 281(\hat{\beta}_0 + \hat{\beta}_1 \overline{\text{valores}}_U) \tag{4.25}$$

donde $\hat{t}^L_{\text{ingresos}}$ es el estimador del total de ingresos fiscales municipales bajo el modelo de población lineal general y $\overline{\text{valores}}_U$ es la media poblacional de los valores inmobiliarios.

En la medida en que la distribución de la variable aleatoria teórica que genera las realizaciones del ingreso fiscal en cada municipio es desconocida, también lo será su varianza condicional a los valores inmobiliarios. De esta forma, σ^2 es un parámetro desconocido y no es posible obtener un valor para la varianza del error de predicción:

$$\text{Var}(\hat{t}^L_{\text{ingresos}} - t_{\text{ingresos}}) = \frac{281^2}{n}\sigma^2 \left(\left(1 - \frac{n}{281}\right) + \frac{\overline{\text{valores}}_U - \overline{\text{valores}}_s}{(1 - n^{-1}) s^2_{\text{valores}}} \right) \tag{4.26}$$

donde s^2_{valores} es la varianza muestral corregida de los valores inmobiliarios y $\overline{\text{valores}}_U$ es su media poblacional.

La varianza del estimador $\hat{t}^L_{\text{ingresos}}$ se estima a partir del error de estimación. Para ello, deberá estimarse σ^2 como el promedio de los residuos del modelo al cuadrado:

$$\widehat{\text{Var}}(\text{ingresos}_i | \text{valores}_i) = (n-1)^{-1} \sum_s (\text{ingresos}_i - \hat{\beta}_0 - \hat{\beta}_1 \text{valores}_i)^2 \tag{4.27}$$

Capítulo 5

Metodología

En este capítulo, se describen los pasos seguidos para abordar el problema introducido en la Sección 1.2. Para las poblaciones ficticias, se detalla el procedimiento de simulación a partir del modelo superpoblacional presentado en el Capítulo 4. Asimismo, se explica el mecanismo de selección de las muestras bajo los distintos diseños utilizados, así como los estimadores y medidas de resumen obtenidas. Para la población MU281, la metodología utilizada fue similar. Sin embargo, en este caso se trabajó con un conjunto de datos reales, de forma que las muestras seleccionadas provinieron siempre de una única realización de un modelo superpoblacional desconocido.

5.1. Poblaciones simuladas

5.1.1. Generación de poblaciones

Para evaluar el desempeño de distintos estimadores basados en modelos y asistidos por modelos, bajo un modelo superpoblacional conocido, se simularon 5.000 poblaciones a partir de dicho modelo. En cada réplica, se llevó a cabo el siguiente procedimiento:

1. Se simularon 300 realizaciones independientes de una variable $X \sim \text{Normal}(10, 1)$ y de un término de error $\varepsilon \sim \text{Normal}(0, 1)$ (ver Subsección 4.1.1).

2. A partir de la regresión lineal simple de parámetros $\beta_0 = 1$ y $\beta_1 = 2$ especificada en la Ecuación 4.1, se simularon 300 realizaciones de Y. Así,

quedaron definidos los 300 elementos de la población, cada uno con su valor de la variable de interés y y de la covariable x.

5.1.2. Extracción de muestras

Para cada una de las 5.000 poblaciones obtenidas, se extrajeron muestras de diferentes tamaños bajo los tres diseños introducidos en el Capítulo 1:

1. **Diseño simple:** En cada réplica, se seleccionaron cinco muestras de 30, 60, 90, 120 y 150 elementos bajo un diseño SI. De esta forma, la probabilidad de selección de todos los individuos fue la misma:

$$\pi_i = {}^{n}/_{N} = {}^{n}/_{300} \quad \forall i = 1, \ldots, 300 \tag{5.1}$$

2. **Diseño simple con muestras truncadas:** Para obtener muestras truncadas superiormente de tamaño $n = 30, 60, 90, 120, 150$, se extrajeron cinco muestras SI de tamaño $n + 5$ y se eliminaron los cinco elementos de mayor valor en términos de la variable de interés y. La cantidad de elementos a ser retirados fue elegida arbitrariamente, con el único fin de simular la no respuesta de individuos atípicos con valores de y extremadamente altos.

   Se optó por tomar muestras de $n + 5$ elementos y no de n para que la muestra final truncada fuera de tamaño n y no $n - 5$. En caso contrario, en la medida en que n aparece en la fórmula de la estimación de la varianza de los estimadores, se habrían generado distorsiones que no habrían permitido comparar los resultados con los obtenidos en muestras no truncadas.

3. **Diseño con probabilidades proporcionales al tamaño:** Al igual que en los dos casos anteriores, para cada población se trabajó con cinco muestras de 30, 60, 90, 120 y 150 individuos, pero seleccionadas mediante un diseño πps. Se eligieron elementos sin reposición con probabilidades de inclusión proporcionales a la inversa de la covariable x. Se tiene entonces:

$$\pi_i = \frac{\frac{1}{x_i}}{\sum_U \frac{1}{x_i}} n \quad \forall i = 1, \ldots, 300 \tag{5.2}$$

   En comparación al diseño SI, bajo este esquema fue mayor la probabilidad

de sacar elementos con valores "pequeños" de x. Si se ignora este aspecto a la hora de obtener las estimaciones de interés, es probable que se incurra en sesgos. En consecuencia, es esperable que un estimador asistido por modelos, que toma en cuenta el diseño muestral, funcione mejor que un estimador basado en modelos, que no lo hace.

5.1.3. Cálculo de estimadores y medidas de resumen

Una vez extraídas las cinco muestras de tamaño variable bajo los tres esquemas de muestreo para cada una de las 5.000 poblaciones simuladas, se obtuvieron diferentes estimadores basados y/o asistidos por modelos.

5.1.3.1. Muestras simples

Tanto para las muestras simples truncadas como para las no truncadas, se siguieron los siguientes pasos:

1. Para las 5.000 réplicas, se estimó el total de la variable de interés, t_y, a partir de los modelos homogéneo y lineal general. Para ello, se emplearon las fórmulas para $\hat{t}_y^E$ y $\hat{t}_y^L$ presentadas en las secciones 3.4 y 3.6. Dado que se asumió una estructura de varianza homocedástica, ambos estimadores coincidieron con su versión asistida por modelos.

2. Para cada réplica, se calculó el error de predicción bajo cada modelo:

$$e\left(\hat{t}_y\right) = t_y - \hat{t}_y \tag{5.3}$$

3. Para cada población, se estimó la varianza de ambos estimadores a partir de lo establecido en las ecuaciones 3.42 y 3.67. Así, se obtuvieron 5.000 realizaciones de $\widehat{\text{Var}}(\hat{t}_y^E)$ y $\widehat{\text{Var}}(\hat{t}_y^L)$.

4. Se construyeron histogramas para las estimaciones y los errores calculados en los puntos 1, 2 y 3 para muestras de tamaño 30.

5. Para los distintos tamaños de muestra, se calculó el porcentaje de cobertura de los intervalos de confianza al 95 % para el estimador del total bajo los modelos homogéneo y de población lineal general.

5.1.3.2. Muestras con probabilidades proporcionales al tamaño

Como se detalló en la Sección 3.8, si se trabaja con un diseño πps, los estimadores basados en modelos y asistidos por modelos, no necesariamente serán iguales. Por consiguiente, en este caso fue posible comparar el desempeño de ambos paradigmas. Para ello, se siguió el procedimiento que se presenta a continuación:

1. Para las 5.000 poblaciones simuladas y sus respectivas muestras πps, se repitieron los pasos 1 a 5 utilizados para las muestras simples.

2. Adicionalmente, se calcularon los estimadores de regresión asistidos por los modelos homogéneo y lineal general, de acuerdo con las fórmulas presentadas en la Subsección 3.7.3. En primera instancia, se calibraron los ponderadores únicamente por el tamaño de la población ($N = 300$), y luego, por el total de la covariable x en cada simulación. Es decir, se obtuvo la "versión" asistida por modelos de los estimadores basados en modelos previamente calculados.

3. Se construyeron los correspondientes histogramas para las estimaciones y sus errores con muestras de 30 elementos.

4. Para estimar la varianza del estimador de regresión y así poder construir los correspondientes intervalos de confianza, se recurrió a las funciones *calibrate()* y *svytotal()* del paquete *survey* (Lumley, 2010). Dicho paquete utiliza la siguiente expresión, diferente de la presentada en el marco teórico (ver Subsección 3.7.4):

$$\widehat{\text{Var}}(\hat{t}_y^{reg}) = \frac{1}{n(n-1)} \sum\nolimits_s \left(e_i wc_i n - \sum\nolimits_s e_i wc_i \right)^2 \tag{5.4}$$

   siendo $e_i = y_i - \hat{y}_i$ el error de predicción de y para el elemento i bajo su respectivo modelo y wc_i el ponderador calibrado.

5. A partir de los intervalos de confianza al 95 % para las muestras de tamaño 30, 50, 60, 120 y 150 para cada una de las 5.000 réplicas, se obtuvo su porcentaje de cobertura del verdadero total de y.

5.2. Población MU281

La población MU281 se concibe como una única realización de un modelo superpoblacional desconocido y todas las muestras seleccionadas son tomadas del mismo conjunto de 281 municipios. Es decir, que lo único que varía en cada simulación es el conjunto de elementos seleccionados en la muestra y no la población en sí. En consecuencia, la metodología aplicada para la población MU281 fue ligeramente diferente a la descrita en la sección anterior.

En este caso, la variable de interés fue *ing_muni_85*, que corresponde a los ingresos de cada municipio en 1985. Por su parte, la única covariable utilizada fue *val_inmob_84*, la cual recoge los valores inmobiliarios en 1984 de cada municipio. Ambas variables se miden en millones de coronas suecas.

5.2.1. Extracción de muestras

Para extraer las muestras, se utilizaron los mismos tres diseños que para el caso de las poblaciones simuladas. En cada caso, se seleccionaron 5.000 muestras con $n = 30, 60, 90, 120, 150$ con el objetivo de conocer el total de ingresos de 281 municipios suecos, t_{ingresos}. Se tiene entonces:

1. **Diseño simple:** Al utilizarse un muestreo aleatorio simple sin reemplazo, la probabilidad de inclusión de cada elemento resulta ser constante:

$$\pi_i = {}^{n}/_{N} = {}^{n}/_{281} \quad \forall i = 1, \ldots, 281 \tag{5.5}$$

2. **Diseño simple con muestras truncadas:** Al igual que para las poblaciones simuladas, se extrajeron muestras de tamaño $n + 5$ y se eliminaron las cinco observaciones con mayor valor en términos de la variable de interés *ing_muni_85*.

3. **Diseño con probabilidades proporcionales al tamaño:** Para extraer muestras πps, se consideraron los valores inmobiliarios de cada municipio. De esta manera, se recurrió a la misma variable auxiliar utilizada para estimar el modelo de población lineal general en el caso basado en modelos. Las resultantes probabilidades de inclusión de primer orden fueron:

$$\pi_i = \frac{\frac{1}{\text{valores}_i}}{\sum_U \frac{1}{\text{valores}_i}} n \quad \forall i = 1, \ldots, 281 \tag{5.6}$$

5.2.2. Cálculo de estimadores y medidas de resumen

El procedimiento seguido para obtener las estimaciones a partir de las distintas muestras fue prácticamente igual al detallado para las poblaciones simuladas. La única diferencia radica en que cada muestra proviene de una misma población y no de realizaciones diferentes del modelo poblacional. En consecuencia, el verdadero total de la variable de interés y la variable auxiliar se mantuvo inalterado, de forma que $t_{\text{ingresos}} = 53.151$ y $t_{\text{valores}} = 757.246$.

Para las muestras simples, se repitieron los pasos 1 a 5 presentados en la Subsubsección 5.1.3.1. Así, se analizó el desempeño de los estimadores basados en los modelos homogéneo y lineal general. Al igual que ocurre con las poblaciones ficticias, el supuesto de homocedasticidad implica que dichos estimadores coincidan con sus análogos asistidos por modelos, de manera que no fue posible compararlos.

En cambio, para las muestras con probabilidades de inclusión proporcionales al tamaño de la variable auxiliar, los estimadores basados en modelos y asistidos por modelos difieren y por lo tanto se pudo comparar su desempeño. Dentro de los estimadores asistidos por modelos, nuevamente se utilizó el estimador de regresión y se consideraron dos situaciones: una en la que sólo se calibró los ponderadores por el tamaño de la población, $N = 281$, y otro en el que se incluyó también el total de valores inmobiliarios, t_{valores}. Para ello, se siguieron los pasos 1-5 presentados en la Subsubsección 5.1.3.2.

Capítulo 6

Resultados

En este capítulo, se detallan los resultados obtenidos bajo los tres esquemas de muestreo considerados para las poblaciones simuladas y la población MU281. En cada caso, se presentan las estimaciones basadas en modelos para el total de la variable de interés junto con las correspondientes estimaciones de su varianza para una muestra de 30 elementos. Asimismo, se exhibe el porcentaje de cobertura del verdadero total de los intervalos de confianza al 95 % para las 5.000 réplicas para distintos tamaños muestrales. En todas las figuras, se utilizan las mismas escalas en los ejes de abscisas y de ordenadas para facilitar la comparación entre modelos.

6.1. Poblaciones simuladas

Como se detalló en capítulos anteriores, para las poblaciones simuladas, se generó la variable de interés y a partir de una regresión lineal con una variable independiente x. En cada una de las 5.000 réplicas, se generó una nueva población de 300 elementos y se buscó estimar el total de y, t_y. De esta manera, los valores simulados de t_y pueden ser entendidos como realizaciones de una variable aleatoria cuya distribución se rige por un cierto modelo.

Es importante destacar que en este caso el modelo superpoblacional a partir del cual se generaron las distintas poblaciones es completamente conocido. Así, para cada modelo, el error de especificación (en caso de existir) será completamente conocido. Esto implica que la calidad de las estimaciones dependerá básicamente de la muestra utilizada para estimar los parámetros superpoblacionales. Como se detalló en el Capítulo 4, el modelo de población

lineal general que incluye a x como único regresor estará perfectamente definido, con lo cual en ese caso el error de especificación será nulo. En cambio, el modelo de población homogénea sí incurrirá en problemas de especificación por no incorporar la información auxiliar relevante.

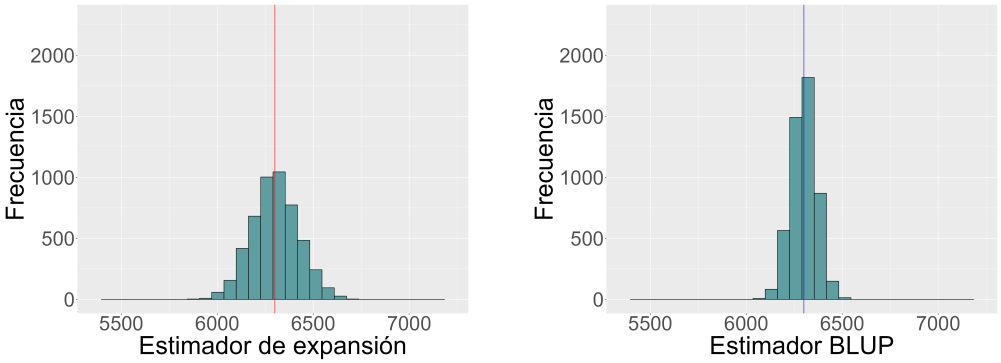
6.1.1. Muestras simples

Tal como se presentó en los capítulos 1 y 2, el diseño simple es ignorable en el sentido de que conduce a la selección de muestras que replican, en promedio, la forma de la distribución de la variable de interés en la población. En consecuencia, no sesgan la estimación de los parámetros superpoblacionales. Si no se incurre en errores de especificación del modelo, las estimaciones deberán ser, entonces, insesgadas.

Por otra parte, para muestras obtenidas bajo un diseño SI y una variable de interés con varianza constante con respecto al modelo utilizado, el estimador de regresión coincidirá con el basado en el modelo de población lineal general. Por lo tanto, en este caso, los estimadores basados en modelos coincidirán con los asistidos por modelos y no será posible compararlos. Para más detalles, ver Capítulo 3.

6.1.1.1. Estimaciones del total

En la Figura 6.1.1, se presentan los histogramas con las 5.000 estimaciones basadas en modelos de t_y. En el panel de la izquierda, se muestran las estimaciones dadas por el modelo homogéneo y en el de la derecha, las que surgen del modelo de población lineal general.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.1.1 – Distribución de los estimadores de t_y basados en los modelos homogéneo y lineal general para poblaciones simuladas y muestras seleccionadas mediante un diseño simple de 30 elementos. La línea roja corresponde a la esperanza del total poblacional bajo el modelo homogéneo. La línea azul es la media de t_y para las 5.000 poblaciones.

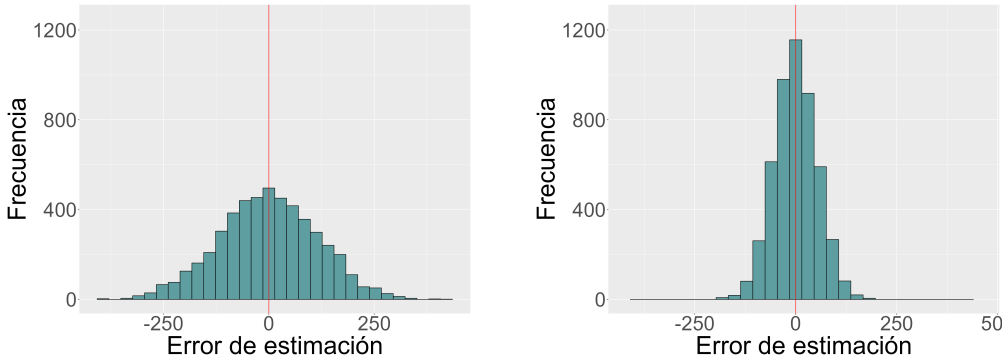
En primer lugar, para el modelo homogéneo, se aprecia una distribución simétrica de las estimaciones del total ($\hat{t}_y^E$) en torno al 6.300, indicado por la línea roja. En ausencia de información auxiliar, t_y se comporta como una variable aleatoria normal con media 6.300 y varianza 1.500 (ver Capítulo 4). Por lo tanto, el hecho de que las estimaciones se centren en dicho valor puede verse como un indicio de que las estimaciones no incurren en grandes sesgos. En este sentido, el promedio de $\hat{t}_y^E$ para las 5.000 simulaciones fue de 6.301,4 (ver Tabla 6.1.1 en la Subsección 6.1.4).

En segundo lugar, también la distribución de las estimaciones de t_y bajo el modelo lineal general ($\hat{t}_y^L$) resultó ser simétrica. Sin embargo, como fue explicado en la Subsubsección 4.1.2.2, en este caso la distribución de t_y deberá condicionarse a la información auxiliar dada por x. Dado que el valor que tome dicha variable en cada elemento cambiará para cada simulación, la esperanza de $t_y|x_i$ también variará. En este sentido, como muestra la Ecuación 4.12, $t_y|x_i$ posee una distribución normal con una media igual a $300(1 + 2\bar{x})$ y varianza de 300. Es decir, que la esperanza del total no tomará un valor constante sino que dependerá de las realizaciones de X en cada simulación. Sin embargo, en la medida que $\bar{x}$ es un estimador insesgado de $E(X) = 10$, es esperable que el promedio de t_y en las 5.000 simulaciones se acerque a 6.300:

$$E(300(1 + 2\bar{x})) = 300(1 + 2E(X)) = 300(1 + 2(10)) = 6.300 \qquad (6.1)$$

En el gráfico, la media de t_y para todas las simulaciones se indica con la línea azul. Efectivamente, dicho valor resultó ser 6.300,2, prácticamente igual a 6.300. Por lo tanto, el hecho de que las estimaciones se encuentren centradas en este valor (como muestra la Tabla 6.1.1, el promedio de las estimaciones fue de 6.300,7) sugiere que las estimaciones dadas por el modelo de población lineal general son aproximadamente insesgadas.

A partir del análisis de los histogramas de $\hat{t}_y$ bajo ambos modelos, queda claro que no resulta fácil analizar directamente su sesgo debido a que el valor de t_y cambiará en cada simulación. Por lo tanto, no existe un único valor con el que comparar todas las estimaciones. Para resolver este problema, en la Figura 6.1.2 se presenta, para cada modelo, un histograma de los errores de estimación absolutos. Ambas distribuciones resultaron ser aproximadamente simétricas en torno a cero. Para el modelo homogéneo, el error medio fue de -1,3 (ver Tabla 6.1.1), y el error relativo promedio fue de -0,02 %. Por su parte, el error promedio bajo el modelo lineal general fue de -0,5, equivalente a un error relativo del -0,01 %. De esta manera, se concluye que ambos modelos arrojan estimaciones aproximadamente insesgadas.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.1.2 – Distribución de los errores de estimación de $\hat{t}_y$ bajo los modelos homogéneo y lineal general para poblaciones simuladas y muestras seleccionadas mediante un diseño simple de 30 elementos.

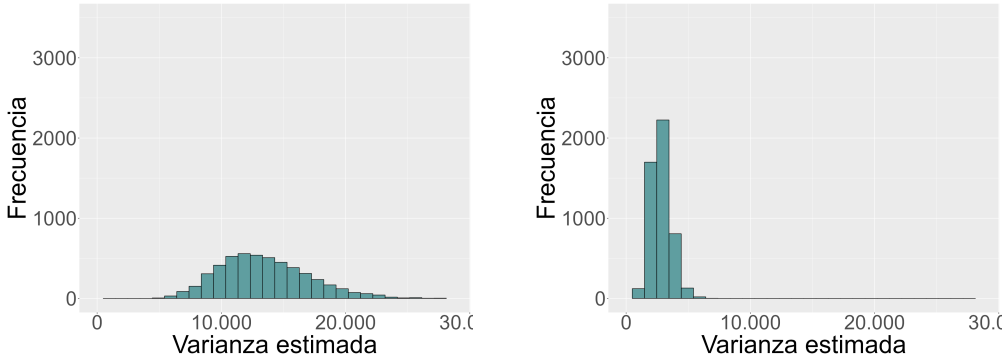
Las figuras 6.1.1a y 6.1.2a muestran que si bien el modelo homogéneo no tiene la misma forma que el modelo superpoblacional utilizado para generar las poblaciones, su desempeño es aceptable. Es decir, que bajo un diseño simple, es robusto a errores de especificación. Esto es esperable ya que, en estas condiciones, el estimador basado en modelos coincide con el estimador basado en el diseño de Horvitz-Thompson, por definición insesgado (ver Capítulo 3).

Por otro lado, también el modelo lineal general permitió arribar a buenas estimaciones. Esto resulta razonable debido a que se utilizó un modelo perfectamente especificado, y estimado a partir de datos que reflejan el verdadero comportamiento de la población tanto en términos de y como de x.

Finalmente, al comparar las figuras 6.1.1a y 6.1.1b, se advierte que aunque ambas distribuciones son aproximadamente simétricas, la dispersión de los estimadores basados en el modelo de población lineal general fue mucho menor que la de los estimadores basados en el modelo de población homogénea. Esto es razonable en la medida que el primer modelo hace uso de la información auxiliar disponible y el segundo no. Dicho de otra forma, dado que el modelo superpoblacional establece que y depende de x, es esperable que incorporar esta variable en el modelo de estimación redunde en un incremento en la eficiencia de los estimadores. A su vez, esto implica que los errores de estimación bajo el modelo lineal general también posean una menor variabilidad que en el caso del modelo homogéneo, tal como muestra la Figura 6.1.2.

6.1.1.2. Estimaciones de la varianza de los estimadores

En la Figura 6.1.3, se exhiben los histogramas de las estimaciones de las varianzas de $\hat{t}_y^E$ y $\hat{t}_y^L$. Nuevamente, la figura de la izquierda muestra los resultados relativos al modelo de población homogénea y la de la derecha los correspondientes al modelo de población lineal general.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.1.3 – Distribución de la varianza estimada de $\hat{t}_y$ bajo los modelos homogéneo y lineal general para poblaciones simuladas y muestras seleccionadas mediante un diseño simple de 30 elementos.

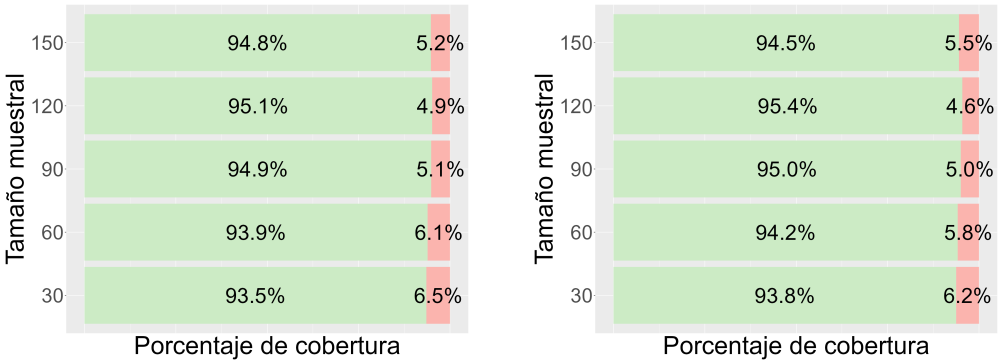
En línea con lo hallado en la Subsubsección 6.1.1.1, la estimación de la varianza resultó ser mucho menor para el modelo lineal general que para el

modelo homogéneo, y su dispersión también fue más pequeña. De esta manera, la distribución de la Figura 6.1.3b se concentra en valores mucho menores que la de la Figura 6.1.3a. Para el modelo homogéneo, el rango de la estimación de la varianza se ubicó entre 5.000 y 25.000 aproximadamente, con una leve asimetría hacia la derecha y una media de 13.451 (ver Tabla 6.1.1). En cambio, para el modelo de población lineal general, la mayor parte de las estimaciones registraron valores menores a 5.000, con una media de 2.800.

6.1.1.3. Cobertura de los intervalos de confianza

En cada modelo, los intervalos de confianza para t_y al 95 % fueron construidos siguiendo lo establecido en la Ecuación 3.44. Para ello, se utilizó tanto $\hat{t}_y$ como la estimación de su varianza. Por consiguiente, la cobertura de los intervalos del verdadero valor de t_y se ve afectada por dos componentes. Por un lado, cuanto menor sea el sesgo de estimación, más probable será que el intervalo capte el verdadero total de y. Adicionalmente, cuanto mayor sea el valor de la varianza estimada, mayor será la amplitud del intervalo, con lo cual la cobertura crecerá. Es decir, que aunque es deseable que los estimadores sean lo más precisos posible, una varianza relativamente grande puede redundar en un porcentaje de cobertura alto. Por lo tanto, al analizar la calidad de las estimaciones, es importante considerar la cobertura de los intervalos en conjunción con el error y la varianza.

En la Figura 6.1.4, se muestra el porcentaje de cobertura de los intervalos de confianza al 95 % para distintos tamaños muestrales bajo ambos modelos. Así, para poblaciones de 300 elementos, se consideran muestras de tamaño 30, 60, 90, 120 y 150.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.1.4 – Cobertura de los intervalos de confianza al 95 % bajo los modelos homogéneo y lineal general para poblaciones simuladas y muestras seleccionadas mediante un diseño simple.

En todos los casos, se observa que la proporción de los 5.000 intervalos que contienen t_y se acerca a su valor deseado de 95 %. Para muestras de tamaño 30, dicho valor fue de 93,5 % bajo el modelo homogéneo y 93,8 % bajo el modelo lineal general. Esto es lógico debido a que ambos modelos arrojaron estimaciones relativamente buenas en términos de insesgamiento y eficiencia. En este sentido, cabe destacar que no existen indicios de que la alta cobertura se origine en una excesiva amplitud de los intervalos. Como referencia, en promedio, el desvío estándar estimado de $\hat{t}_y$ fue de 116 para el modelo población homogénea y 53 para el modelo de población lineal general (ver varianzas en la Tabla 6.1.1). Teniendo en cuenta que las estimaciones fueron del orden de 6.300, es claro que los intervalos de confianza resultantes son bastante precisos.

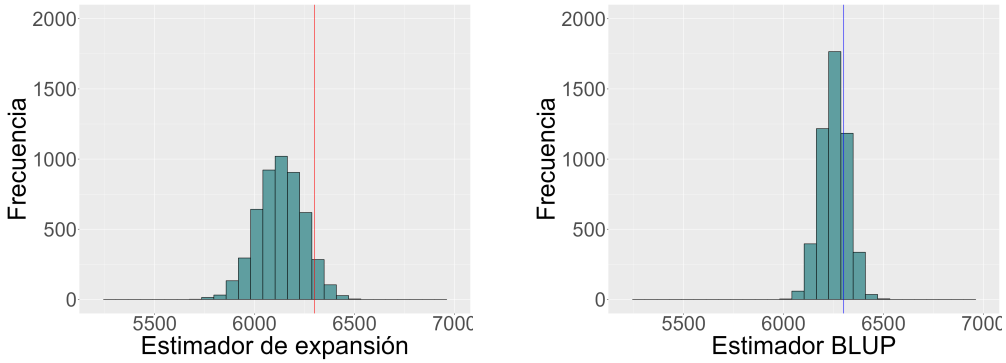
Para ambos modelos, la cobertura aumenta levemente con el tamaño muestral hasta estabilizarse en muestras de 120 elementos. Luego, se observa un pequeño descenso en las muestras de tamaño 150. Además, para la mayoría de los tamaños de muestra, la cobertura fue apenas mayor en el modelo de población lineal general que en el modelo homogéneo, aun cuando aquel presenta una menor estimación de la varianza. Esto sugiere que las estimaciones dadas por el modelo lineal general son muy buenas, lo cual es esperable porque dicho modelo está completamente libre de errores de especificación y fue estimado a partir de una muestra "perfecta".

6.1.2. Muestras simples truncadas

Evidentemente, el caso de las muestras simples constituye una situación "ideal" en la que los datos utilizados para estimar el modelo superpoblacional replican la distribución de y en la población. Sin embargo, en la práctica, suelen existir problemas que harán que las muestras sean "imperfectas". Esto puede generar sesgos en las estimaciones debido a que los parámetros superpoblacionales serán estimados a partir de datos que no reflejan el comportamiento real de la variable de interés. En este contexto, se trabajó con muestras truncadas superiormente con el fin de replicar patrones de no respuesta en los que se tiende a perder valores "altos" en términos de la variable de interés.

6.1.2.1. Estimaciones del total

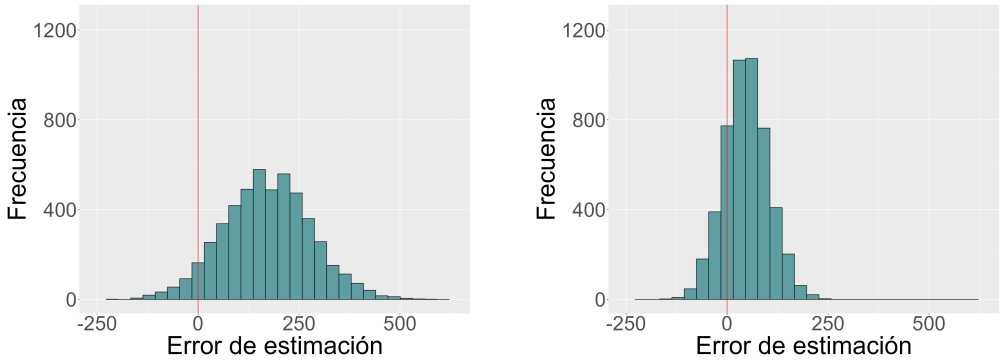
En la Figura 6.1.5, se presentan los histogramas de las estimaciones de t_y para los modelos homogéneo y lineal general. Nuevamente, la línea roja indica la esperanza de t_y y vale 6.300, y la azul denota el promedio de t_y para las 5.000 réplicas generadas, de 6.300,2. Como es esperable, al eliminar las 5 observaciones más grandes en términos de y de cada muestra, se tiende a subestimar su verdadero total. En este sentido, se advierte un desplazamiento de ambas distribuciones hacia la izquierda. Para el modelo homogéneo, esto se dio de manera mucho más acentuada: mientras que para el modelo lineal general la estimación promedio de t_y fue de 6.253,3, bajo el modelo homogéneo dicho valor fue de apenas 6.128,9 (ver Tabla 6.1.1). Esto parece razonable en la medida que el modelo de población lineal general permite explotar la información dada por x y así "compensar" la distorsión generada al truncar cada muestra. En contraste, el modelo de población homogénea incurre en un error de especificación que agrava los efectos de utilizar datos "imperfectos".



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.1.5 – Distribución de los estimadores de t_y basados en los modelos homogéneo y lineal general para poblaciones simuladas y muestras simples truncadas de tamaño 30. La línea roja corresponde a la esperanza del total poblacional bajo el modelo homogéneo. La línea azul es la media de t_y para las 5.000 poblaciones.

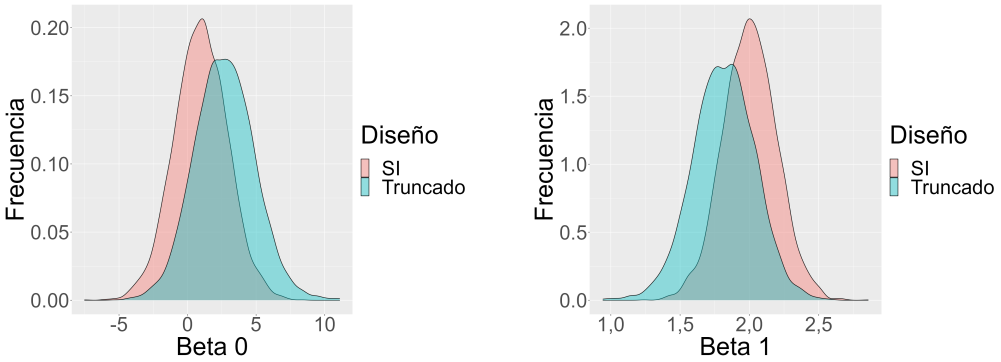
En línea con lo anterior, los histogramas de los errores de estimación evidencian un menor sesgo bajo el modelo lineal general que bajo el modelo homogéneo (ver Figura 6.1.6). En ambos casos, dado que las estimaciones subestiman t_y, los errores tendieron a ser positivos. En consecuencia, su distribución se trasladó hacia la derecha, lo cual da cuenta de la existencia de un sesgo sistemático. Sin embargo, mientras que el error promedio para el modelo homogéneo fue de 171,3, para el modelo lineal general dicho valor fue mucho menor (46,9). Como se detalla en la Tabla 6.1.1, los errores relativos promedio resultaron ser de 2,7 % y 0,7 %, respectivamente.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.1.6 – Distribución de los errores de estimación de $\hat{t}_y$ bajo los modelos homogéneo y lineal general para poblaciones simuladas y muestras simples truncadas de tamaño 30.

Por otra parte, es importante destacar que estos errores se deben principalmente a que los dos modelos considerados fueron ajustados a partir de muestras no ignorables. Al utilizar datos que no reflejan la dispersión de la variable de interés en la población, se generarán distorsiones en el proceso de estimación de los parámetros superpoblacionales. Por ejemplo, en la Figura 6.1.7, se exhibe la densidad estimada de $\hat{\beta}_0$ y $\hat{\beta}_1$ para el modelo de población lineal general. Dado que el modelo de población homogénea equivale a utilizar una regresión lineal sin variables independientes, todas las predicciones serán iguales a un intercepto estimado mediante la media muestral. Por consiguiente, resulta ser un caso trivial y, para facilitar la exposición, se optó por omitirlo.



(a) Estimación de β_0 **(b)** Estimación de β_1

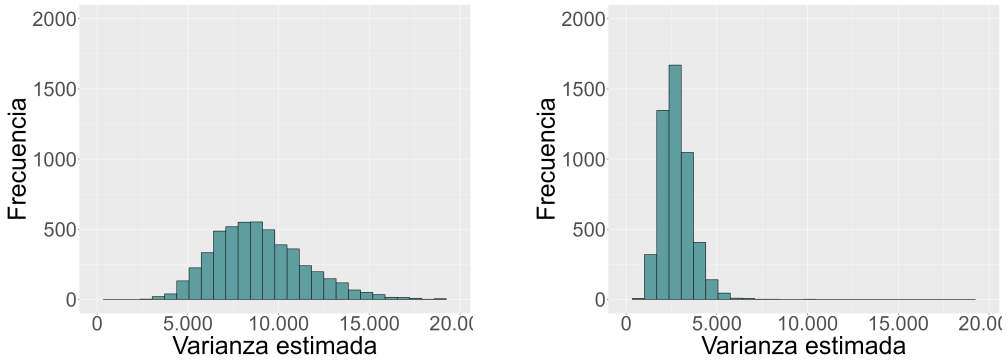
Figura 6.1.7 – Densidades de $\hat{\beta}_0$ y $\hat{\beta}_1$ en el marco del modelo de población lineal general. Se comparan los resultados obtenidos para poblaciones simuladas mediante muestras simples de 30 elementos truncadas y no truncadas.

Como muestra la Figura 6.1.7, las estimaciones de ambos parámetros se vieron sustancialmente afectadas por el hecho de utilizar muestras truncadas. La distribución de $\hat{\beta}_0$ para las muestras truncadas se ubicó en valores más grandes del recorrido que la distribución de las muestras simples. Es decir, que la densidad se trasladó hacia la derecha y dejó de estar centrada en su valor real de 1. Esto implica que la recta de regresión estimada, en promedio, se desplazó hacia arriba. Por lo tanto, se pasó a sobreestimar su ordenada en el origen. Por su parte, al truncar las muestras, la distribución de $\hat{\beta}_1$ se desplazó hacia la izquierda (aunque en ningún momento llegó a tomar valores negativos). Así, la recta de regresión estimada tendió a disminuir su pendiente, cuyo verdadero valor era 2. Es decir, que el usar muestras truncadas generó un sesgo negativo en las estimaciones del parámetro $\hat{\beta}_1$ y, en consecuencia, se tendió a subestimar su valor. Esto determina que el incremento esperado en y dado un cierto aumento

en x es menor si se utilizan muestras truncadas superiormente que bajo un muestreo simple. Al eliminarse los elementos de la muestra más grandes en términos de y, se subestimará el efecto medio de x sobre dicha variable.

6.1.2.2. Estimaciones de la varianza de los estimadores

La Figura 6.1.8 muestra los histogramas de las estimaciones de la varianza de $\hat{t}_y^E$ y $\hat{t}_y^L$. Ambas distribuciones presentan una leve asimetría a la derecha. Sin embargo, se aprecia una dispersión mucho mayor bajo el modelo homogéneo que bajo el modelo de población lineal general. Además, la distribución de la Figura 6.1.8a se concentra en valores más grandes del recorrido que la de la Figura 6.1.8b. En consecuencia, las estimaciones promedio resultaron ser 8.931,1 y 2.777,3, respectivamente (ver Tabla 6.1.1).



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

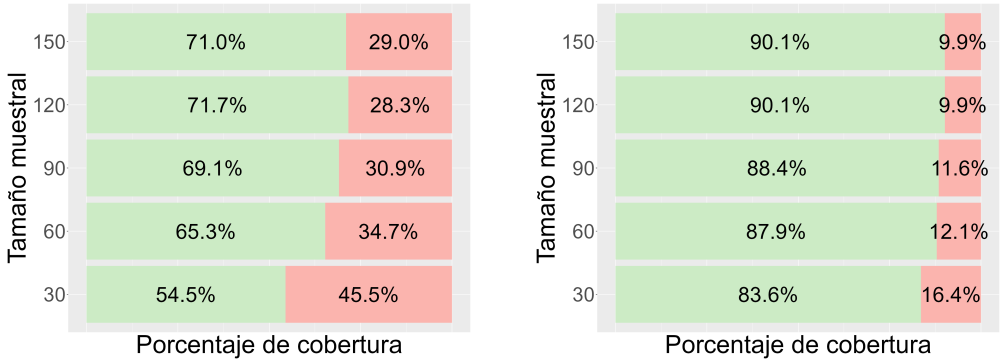
Figura 6.1.8 – Distribución de la varianza estimada de $\hat{t}_y$ bajo los modelos homogéneo y lineal general para poblaciones simuladas y muestras truncadas de tamaño 30.

De esta forma, la evidencia sugiere que el modelo de población lineal general arroja estimaciones más eficientes que el modelo homogéneo aun cuando las muestras utilizadas hayan sido truncadas. Es probable que esto se deba a que el modelo de población general no incurre en errores de especificación y hace uso de información auxiliar significativa.

6.1.2.3. Cobertura de los intervalos de confianza

En la Figura 6.1.9, se detallan los porcentajes de cobertura de los intervalos de confianza al 95 % para muestras de diferente tamaño bajo los dos modelos. En ambos casos, esta proporción aumenta con el tamaño muestral hasta alcanzar

su máximo en $n = 120$. Sin embargo, para el modelo homogéneo, dicho valor fue muy inferior a su nivel deseado para todos los tamaños muestrales: de un mínimo de 54,5 % para $n = 30$, se llega a 71,7 % en $n = 120$. Aunque este problema fue menos grave para el modelo de población lineal general, tampoco en este caso se alcanzó el valor preestablecido de 95 %. En este segundo modelo, se pasa de una cobertura de 83,6 % cuando $n = 30$ a un 90,1 % para $n = 120$ y $n = 150$.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.1.9 – Cobertura de los intervalos de confianza al 95 % bajo los modelos homogéneo y lineal general para poblaciones simuladas y muestras truncadas.

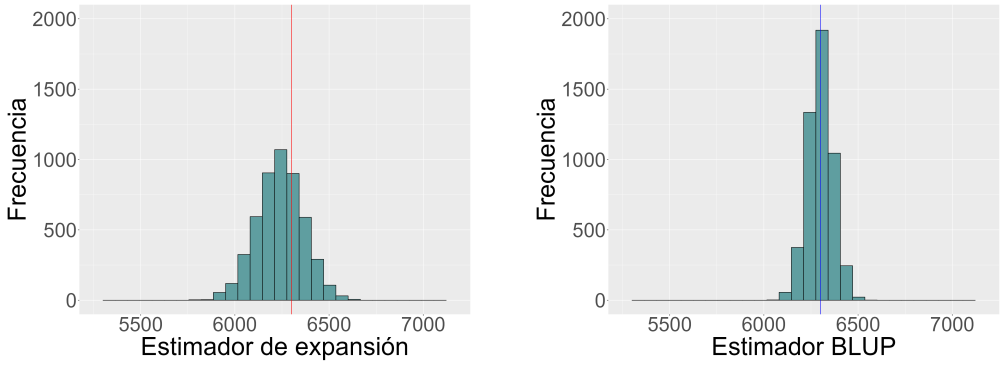
Todos estos resultados indican que el utilizar muestras truncadas genera sesgos en las estimaciones de t_y y distorsiona las estimaciones de su varianza y, en consecuencia, la cobertura de los correspondientes intervalos de confianza se reduce. Este problema resulta ser particularmente grave en el caso del modelo homogéneo, en el que al problema de utilizar muestras "imperfectas" se le suma el error de especificación del modelo superpoblacional.

6.1.3. Muestras πps

Como segundo caso de muestras "imperfectas", se trabajó con muestras con probabilidades de inclusión proporcionales al inverso de x. De esta manera, los elementos de la población tendrán mayor probabilidad de ser seleccionados cuanto menor sea su valor de x y, por lo tanto, será necesario considerar los ponderadores de cada uno a la hora de "replicar" el comportamiento de y en la población. Es decir, que el diseño no es ignorable debido a que la muestra sin ponderadores no refleja el verdadero comportamiento de la variable de interés.

6.1.3.1. Estimaciones del total

La Figura 6.1.10 muestra los histogramas de $\hat{t}_y^E$ y $\hat{t}_y^L$ obtenidos a partir de muestras πps. Al igual que en el caso de las muestras truncadas, se tiende a subestimar el verdadero valor de t_y debido a que las muestras tienden a incluir relativamente menos observaciones "grandes" en términos de y que la población general. De esta forma, el no tener en cuenta un diseño "no ignorable" conlleva sesgos en la estimación de los parámetros superpoblacionales. Sin embargo, se advierte que este problema fue más acentuado para el modelo homogéneo que para el modelo lineal general. En efecto, mientras que la distribución de la Figura 6.1.10a se encuentra levemente a la izquierda del valor de referencia de 6.300, el histograma de la Figura 6.1.10b se encuentra aproximadamente centrado en dicho valor. Como se muestra en la Tabla 6.1.1, el promedio de las estimaciones fue de 6.239,7 para el modelo homogéneo y 6.300,1 para el lineal general.



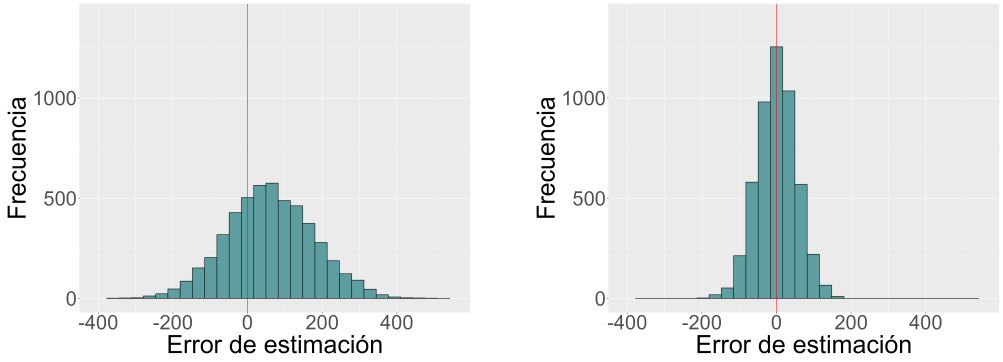
(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.1.10 – Distribución de los estimadores de t_y basados en los modelos homogéneo y lineal general para las poblaciones simuladas y muestras πps de tamaño 30. La línea roja corresponde a la esperanza del total poblacional bajo el modelo homogéneo. La línea azul es la media de t_y para las 5.000 poblaciones.

Por otro lado, al igual que ocurría tanto con las muestras simples como con las truncadas, se aprecia una mayor concentración de la distribución de las estimaciones en la Figura 6.1.10b que en la Figura 6.1.10a. Se concluye, entonces, que las estimaciones obtenidas bajo el modelo de población general fueron más precisas que las dadas por el modelo homogéneo.

En línea con lo anterior, a partir de la siguiente figura, se observa que el modelo homogéneo tiende a subestimar t_y y, por lo tanto, la distribución de los errores se concentra en valores mayores a cero (ver Figura 6.1.11). En

contraposición, la distribución correspondiente al modelo de población lineal general se encuentra centrado en cero y no se advierten errores sistemáticos.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

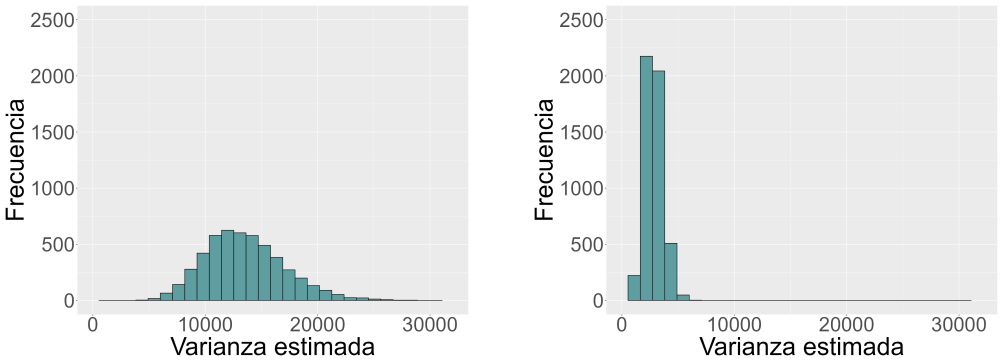
Figura 6.1.11 – Distribución de los errores de estimación de $\hat{t}_y$ bajo los modelos homogéneo y lineal general para poblaciones simuladas y muestras πps de tamaño 30.

Los errores de estimación medios fueron de 60,5 unidades para el modelo homogéneo y de 0,1 para el modelo de población lineal general. En términos relativos, esto equivale a un error del 1,0 % y prácticamente 0 %, respectivamente. Es decir, que para el segundo modelo prácticamente no existe un sesgo de estimación, y aun para el primero el error promedio es comparable al obtenido al utilizar muestras simples. Es posible que las estimaciones no se hayan visto tan afectadas como con las muestras truncadas debido a que, al contrario de lo que ocurría en aquel caso, sí es posible incluir observaciones "grandes" de y. Así, si bien la probabilidad de seleccionar dichas observaciones es más baja que bajo un diseño simple (dada la fuerte correlación de y con x), ésta no es nula, con lo cual algunas de las réplicas consideradas sí reflejarán el verdadero comportamiento de la variable de interés en la población.

6.1.3.2. Estimaciones de la varianza de los estimadores

En la Figura 6.1.12, se presentan las estimaciones de la varianza de los estimadores del total bajo los dos modelos. Los resultados no difieren sustancialmente de los obtenidos mediante muestras simples. En este caso, las estimaciones de la varianza promedio fueron de 13.498,9 para el modelo homogéneo y de 2.837,8 para el lineal general. A su vez, nuevamente se observa que las estimaciones del modelo homogéneo presentan una mayor variabilidad que las arrojadas por el modelo de población lineal general. En este último caso,

el no incurrir en errores de especificación del modelo superpoblacional parece contribuir fuertemente a la robustez de las estimaciones frente a muestras "imperfectas".

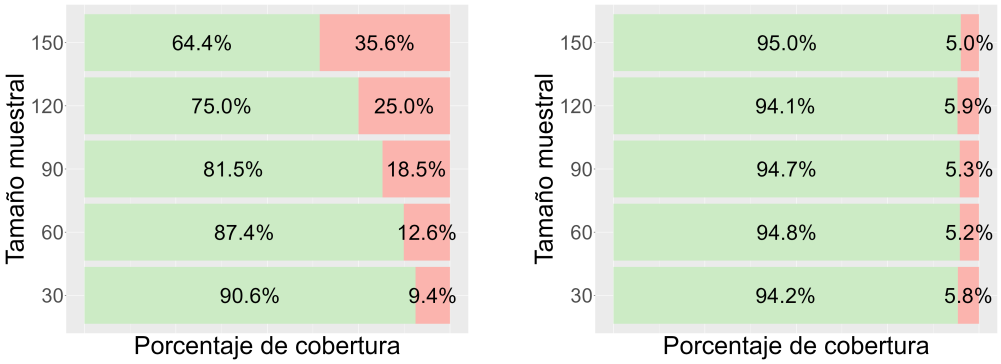


(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.1.12 – Distribución de la varianza estimada de $\hat{t}_y$ bajo los modelos homogéneo y lineal general para poblaciones simuladas y muestras πps de tamaño 30.

6.1.3.3. Cobertura de los intervalos de confianza

A continuación, se presenta la cobertura de los intervalos de confianza al 95 % para los distintos tamaños de muestra (ver Figura 6.1.13). Para el modelo homogéneo, resulta llamativo el hecho de que dicha cobertura parece reducirse con n. De esta manera, la cobertura descendió desde 90,6 % cuando $n = 30$ a 64,4 % para $n = 150$. Esto puede deberse a que la estimación de la varianza del estimador de un total se reduce con el tamaño muestral (ver Ecuación 3.42). Por lo tanto, si las mismas están sesgadas, es razonable que empeore la cobertura de los intervalos de confianza a medida que aumenta n.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

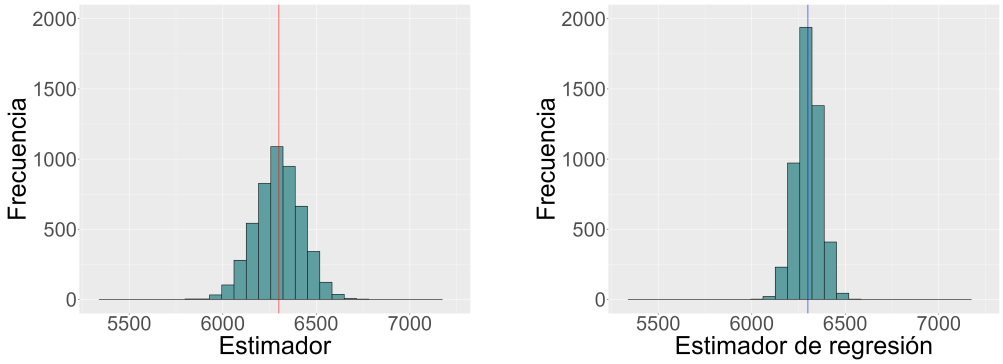
Figura 6.1.13 – Cobertura de los intervalos de confianza al 95 % bajo los modelos homogéneo y lineal general para poblaciones simuladas y muestras πps.

Por el contrario, para el modelo de población general, las estimaciones son insesgadas. En consecuencia, el porcentaje se mantiene cercano al 95 % para todos los tamaños muestrales y los resultados no difieren demasiado de lo hallado para el caso de las muestras simples.

6.1.3.4. Comparación con la inferencia asistida por modelos

Como se demostró en la Subsección 3.8.3, a diferencia de lo que ocurre con muestras simples, para un diseño πps el estimador de regresión (asistido por modelos) no coincide con su estimador equivalente bajo el paradigma basado en modelos. Por este motivo, en este caso resulta interesante analizar el impacto de incorporar la información dada por un diseño muestral no ignorable en las estimaciones.

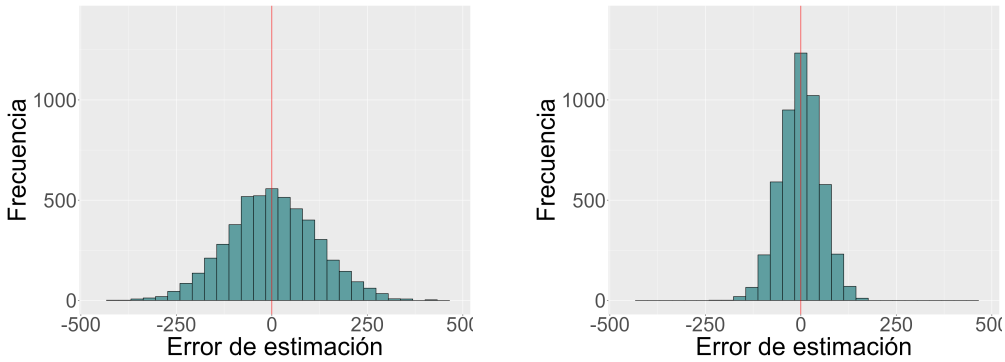
En la Figura 6.1.14, se exhibe la distribución de las estimaciones de t_y dadas por el estimador de regresión $\hat{t}_y^{reg}$. Como se detalló en la metodología (Capítulo 5), se consideraron dos casos posibles. Primero, se calibraron los ponderadores únicamente por el tamaño de la población y, segundo, se consideró también a la variable auxiliar x. De esta forma, se obtuvieron los estimadores asistidos por modelos análogos a los correspondientes basados en los modelos homogéneo y lineal general.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.1.14 – Distribución de los estimadores de t_y asistidos por los modelos homogéneo y lineal general para las poblaciones simuladas y muestras πps de tamaño 30. La línea roja corresponde a la esperanza del total poblacional bajo el modelo homogéneo. La línea azul es la media de t_y para las 5.000 poblaciones.

Si se compara la Figura 6.1.14a con la Figura 6.1.10a, se advierte que el calibrar los ponderadores por el tamaño de la población permite corregir el sesgo negativo de las estimaciones basadas en el modelo homogéneo. Es decir, que alcanza con utilizar únicamente al tamaño de la población como información auxiliar para mitigar dicho sesgo. Así, las estimaciones nuevamente pasaron a estar centradas en su valor de referencia de 6.300, y la estimación promedio fue de 6.298,4 (ver Tabla 6.1.1). Análogamente, la distribución de los errores de estimación volvió a alcanzar su máximo en cero (ver figuras 6.1.11a y 6.1.15a). En este caso, el error medio fue de 1,8 unidades, lo cual equivale a apenas 0,03 %.

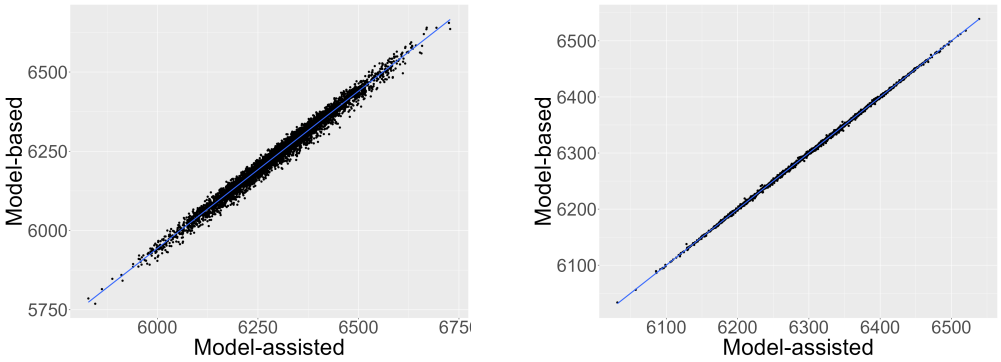


(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.1.15 – Distribución de los errores de estimación de $\hat{t}_y^{reg}$. La inferencia es asistida por los modelos homogéneo y lineal general para poblaciones simuladas y muestras πps de tamaño 30.

Por su parte, no se observan grandes diferencias entre las estimaciones basadas en modelos y asistidas por modelos, para el modelo de población lineal general (ver figuras 6.1.10b y 6.1.14b). En ambos casos, la estimación promedio dada por el estimador de regresión fue de 6.300,1. El error de estimación medio también se mantuvo constante en 0,1 unidades. Esto resulta lógico si se tiene en cuenta que no se advierten sesgos en la Figura 6.1.11b, con lo cual la potencial mejora asociada a incorporar el diseño muestral es limitada.

Cabe señalar que a pesar de que las estimaciones son sumamente parecidas si se considera una regresión lineal simple, ambos enfoques difieren. Como se observa en la Figura 6.1.16b, cada una de las 5.000 réplicas arroja estimaciones levemente distintas para cada paradigma, con lo cual los puntos del correspondiente gráfico de dispersión no se encuentran perfectamente alineados. Teóricamente, esto se explica por el hecho de que las estimaciones basadas en modelos se componen de la suma del total de y en la muestra y de la suma de las predicciones para cada elemento no observado. Por el contrario, en este caso se cumplen ciertas condiciones que aseguran que el estimador de regresión equivale a la suma de las predicciones del modelo para toda la población (ver Ecuación 3.86). Dado que se utiliza una muestra pequeña en comparación al tamaño de la población, es de esperar que ambas estimaciones se asemejen.

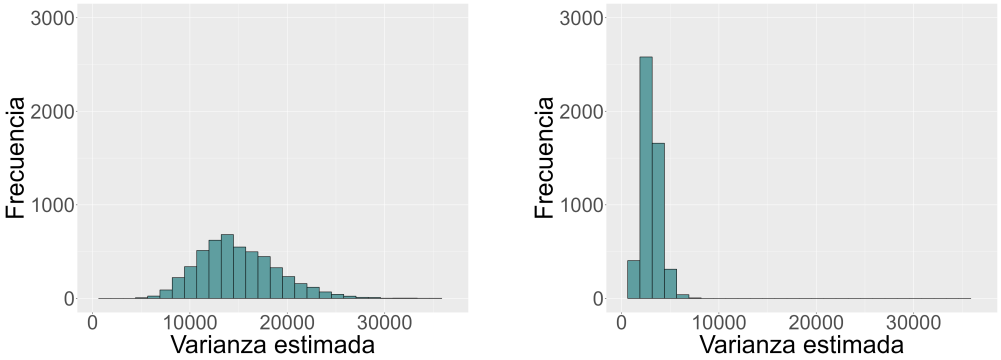


(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.1.16 – Gráfico de dispersión de los estimadores de t_y basados y asistidos por los modelos homogéneo y lineal general para las poblaciones simuladas y muestras πps de tamaño 30.

En la Figura 6.1.17, se presentan las distribuciones de las estimaciones de la varianza del estimador de regresión. Bajo ambos modelos, los resultados se asemejan mucho a los obtenidos mediante el paradigma basado en modelos, y en todos los casos se advierte una leve asimetría a la derecha (ver Figura 6.1.12). De

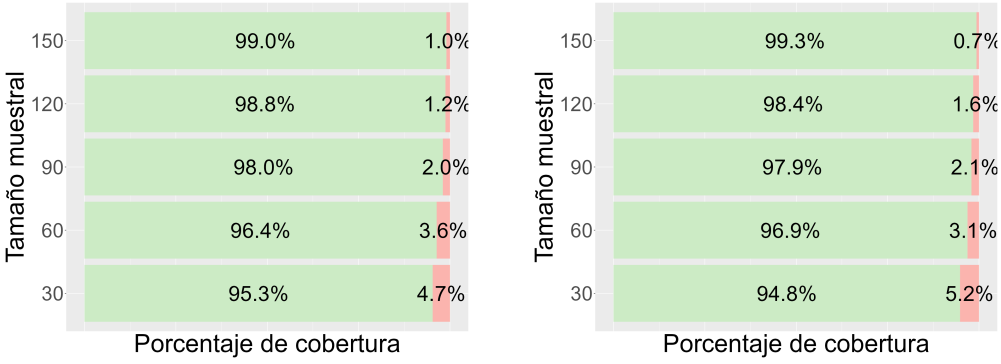
esta manera, el utilizar un estimador asistido por modelos no parece impactar mayormente en la varianza de las estimaciones.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.1.17 – Distribución de la varianza estimada de $\hat{t}_y^{reg}$. La inferencia es asistida por los modelos homogéneo y lineal general para poblaciones simuladas y muestras πps de tamaño 30.

Finalmente, a partir de la Figura 6.1.18, se concluye que el incorporar el diseño muestral en las estimaciones dadas por el modelo homogéneo permite corregir el problema de subcobertura de los intervalos de confianza. Así, incluso para muestras relativamente pequeñas ($n = 30$), se logra una cobertura del 95,3 %, superior al nivel deseado. Dicho porcentaje crece con el tamaño muestral hasta alcanzar un 99 % cuando $n = 150$, de forma que se revierte el comportamiento atípico observado bajo el paradigma basado en modelos, según el cual la cobertura disminuía al crecer la muestra.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.1.18 – Cobertura de los intervalos de confianza al 95 % para estimaciones asistidas por los modelos homogéneo y lineal general para poblaciones simuladas y muestras πps.

En cambio, para el modelo de población lineal general, la mejora en la

cobertura asociada a utilizar un enfoque asistido por modelos en lugar del basado en modelos es algo más limitada. Nuevamente, se observa que el porcentaje de cobertura crece con el tamaño de la muestra, y se pasa de un 94,8 % para $n = 30$ a un 99,3 % para $n = 150$. Esto implica que el hecho de tomar en cuenta el diseño muestral conlleva una sobrecobertura de los intervalos de confianza.

6.1.4. Estadísticos de resumen

En la Tabla 6.1.1 se presentan algunos indicadores que resumen el comportamiento de las estimaciones para cada modelo y diseño muestral utilizado. En particular, se considera la media de $\hat{t}_y$ y de $\widehat{\text{Var}}\left(\hat{t}_y\right)$. Asimismo, se considera el error de estimación medio. En cada réplica, el error se calcula como $e\left(\hat{t}_y\right) = t_y - \hat{t}_y$. Por último, el error de estimación relativo se calcula como el error absoluto sobre el verdadero total de y: $p\left(\hat{t}_y\right) = {}^{e\left(\hat{t}_y\right)}/_{t_y}$.

Modelo homogéneo				
Diseño	$\hat{\mathbf{t}}_\mathbf{y}$	$\widehat{\mathbf{Var}}\left(\hat{\mathbf{t}}_\mathbf{y}\right)$	$\mathbf{e}\left(\hat{\mathbf{t}}_\mathbf{y}\right)$	$\mathbf{p}\left(\hat{\mathbf{t}}_\mathbf{y}\right)$ [%]
Simple	6301,5	13 451,0	−1,3	−0,02
Truncado	6128,9	8931,2	171,3	2,72
πps (MB)	6239,7	13 498,9	60,5	0,96
πps (MA)	6298,4	14 989,2	1,8	0,03
Modelo lineal general				
Diseño	$\hat{\mathbf{t}}_\mathbf{y}$	$\widehat{\mathbf{Var}}\left(\hat{\mathbf{t}}_\mathbf{y}\right)$	$\mathbf{e}\left(\hat{\mathbf{t}}_\mathbf{y}\right)$	$\mathbf{p}\left(\hat{\mathbf{t}}_\mathbf{y}\right)$ [%]
Simple	6300,7	2800,3	−0,5	−0,01
Truncado	6253,3	2777,3	46,9	0,74
πps (MB)	6300,1	2837,7	0,1	0,00
πps (MA)	6300,1	3021,0	0,1	0,00

Tabla 6.1.1 – Promedios de las estimaciones y error medio (absoluto y relativo) bajo los modelos homogéneo y lineal general para las 5.000 réplicas de las poblaciones simuladas. Las estimaciones basadas en modelos y asistidas por modelos se notan como “MB” y “MA” respectivamente.

6.2. Población MU281

A diferencia del caso de las poblaciones simuladas, la población MU281 puede ser entendida como una única realización de un modelo superpoblacional desconocido. De esta manera, en cada una de las 5.000 réplicas, la población se

mantiene fija y sólo varía la muestra seleccionada bajo cada diseño. A su vez, no se conoce el mecanismo a través del cual se generó la población y, por lo tanto, no se sabe a ciencia cierta si existen errores de especificación del modelo. En consecuencia, se agrega una nueva fuente de error en las estimaciones, la cual no estaba presente en el caso de las poblaciones simuladas.

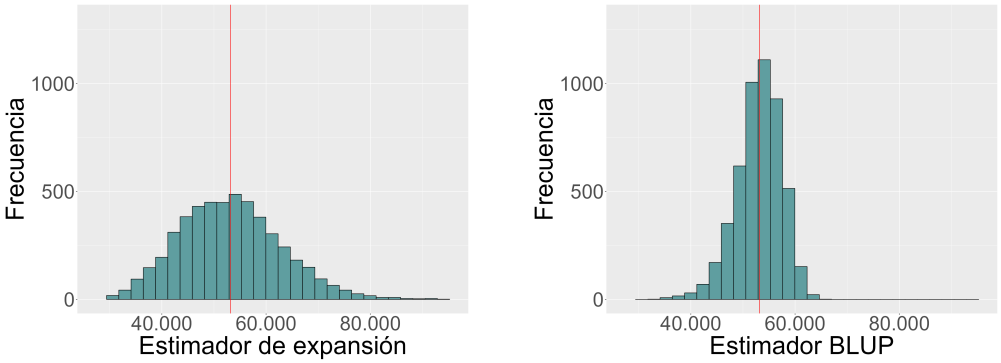
Como se detalló en el Capítulo 4, se tomó como variable de interés al indicador *ing_muni_85*, el cual refiere a los ingresos fiscales de 1985 de cada municipio sueco. Así, se buscó estimar el total de ingresos ($\hat{t}_{\text{ingresos}}$) de los 281 municipios considerados, el cual es de 53.151 millones de coronas suecas. Además, se supuso que la única variable auxiliar disponible fue *val_inmob_84*, que recoge los valores inmobiliarios de cada municipio en 1984.

6.2.1. Muestras simples

Al igual que ocurría con las poblaciones simuladas, las muestras tomadas bajo un diseño simple constituyen una situación "ideal", en la cual la forma en que se selecciona cada elemento no introduce distorsiones en la distribución de la variable de interés en la población. Por lo tanto, las muestras son ignorables y, si el modelo especificado es razonablemente bueno, los parámetros superpoblacionales tenderán a ser insesgados.

6.2.1.1. Estimaciones del total

En la Figura 6.2.1, se presentan los histogramas de $\hat{t}_{\text{ingresos}}$ para los modelos homogéneo y lineal general. Se observa que ambos modelos arrojan distribuciones levemente asimétricas, aunque en direcciones opuestas: hacia la derecha el modelo homogéneo y hacia la izquierda el lineal general. Sin embargo, ambas se encuentran aproximadamente centradas en el verdadero valor del total, indicado mediante la línea roja. En este sentido, tal como se detalla en la Tabla 6.2.1, el promedio de las estimaciones fue de 53.159 para el modelo homogéneo y de 53.051 para el modelo lineal general, lo que sugiere que ambos conducen a estimaciones aproximadamente insesgadas.

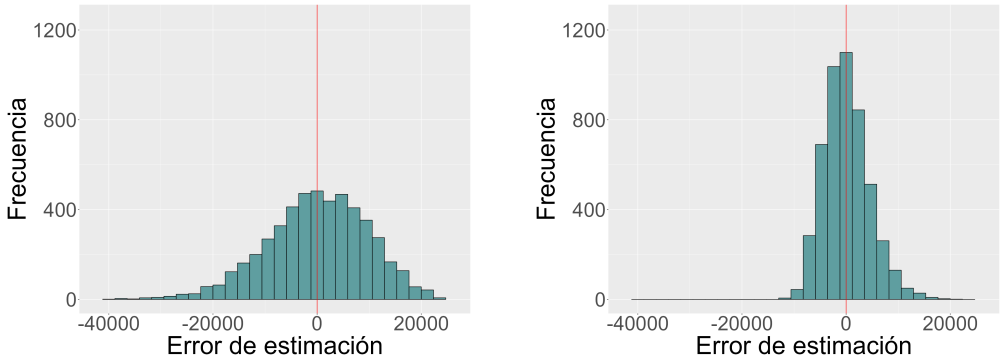


(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.2.1 – Distribución de los estimadores de t_{ingresos} basados en los modelos homogéneo y lineal general para la población MU281 y muestras seleccionadas mediante un diseño simple de 30 elementos. La línea roja corresponde al total de ingresos municipales.

Asimismo, cabe señalar que el histograma de la Figura 6.2.1b presenta una menor dispersión que el de la Figura 6.2.1a. Nuevamente, entonces, el modelo de población lineal general permite explotar la información auxiliar y obtener estimaciones más precisas que el modelo de población homogénea.

En la siguiente figura, se muestran las distribuciones de los errores de estimación para cada modelo (ver Figura 6.2.2). Cabe señalar que en la medida que t_{ingresos} no varía a lo largo de las 5.000 réplicas (a diferencia de lo que ocurría con las poblaciones simuladas), se cuenta con un único valor de referencia y por lo tanto resulta sencillo determinar si las estimaciones son sesgadas o no a partir de la Figura 6.2.1. En consecuencia, en este caso no es tan necesario trabajar con los errores. De todas formas, la Figura 6.2.2 es útil en la medida que permite visualizar más fácilmente el hecho de que los errores se ubican en torno a cero bajo ambos modelos. Para el modelo homogéneo, se aprecia una leve asimetría a la izquierda, lo cual implica que las estimaciones sobrestiman ligeramente t_{ingresos}. Por el contrario, los errores de estimación bajo el modelo lineal general exhiben una asimetría a la derecha, de manera que las estimaciones subestiman el verdadero total de interés. De esta manera, mientras que el error promedio bajo el modelo homogéneo fue de -8,0, para el modelo lineal general dicho valor fue de 99,9, con lo cual presentan signos opuestos. En términos relativos, dichos errores equivalen a -0,01 % y 0,2 %, relativamente (ver Tabla 6.2.1).



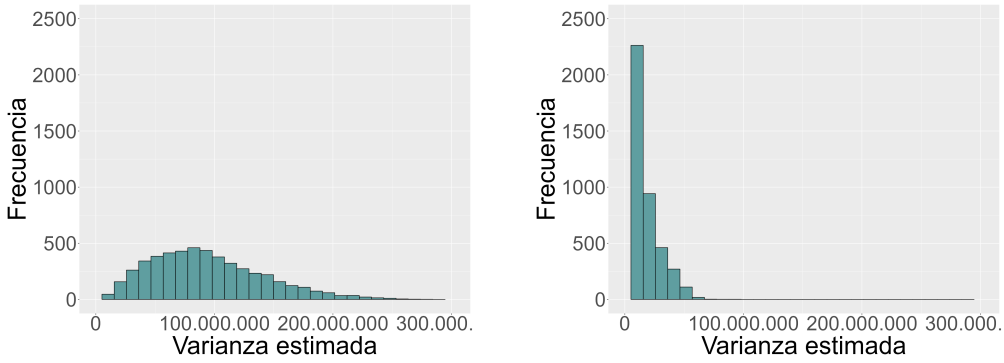
(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.2.2 – Distribución de los errores de estimación de $\hat{t}_{\text{ingresos}}$ bajo los modelos homogéneo y lineal general para la población MU281 y muestras seleccionadas mediante un diseño simple de 30 elementos.

Si bien ambos modelos arrojan resultados comparables, es interesante notar que el modelo homogéneo presenta un error de estimación medio un poco menor que el modelo lineal general. Probablemente, esto se deba a que, bajo un diseño SI, el estimador dado por el modelo homogéneo coincide con el estimador de Horvitz-Thompson, por definición insesgado. Por consiguiente, bajo estas condiciones "ideales", el modelo homogéneo tiende a ser más robusto a errores de especificación que el lineal general.

6.2.1.2. Estimaciones de la varianza de los estimadores

En la Figura 6.2.3, se presentan los histogramas de las estimaciones de la varianza de los estimadores bajo los dos modelos considerados ($\widehat{\text{Var}}\left(\hat{t}_{\text{ingresos}}\right)$). En primer lugar, se advierte una mucho mayor dispersión para el modelo homogéneo que para el modelo lineal general. Además, las estimaciones de $\text{Var}(\hat{t}_{\text{ingresos}})$ se concentran en valores más bajos, lo cual es consistente con la mayor precisión observada en la Figura 6.2.1b. De esta manera, en promedio, la estimación de la varianza de los estimadores fue de alrededor de 94 millones bajo el modelo homogéneo y de 15 millones para el modelo lineal general (ver Tabla 6.2.1). Estos valores equivalen a desvíos estándar de 9.702 y 3.894.

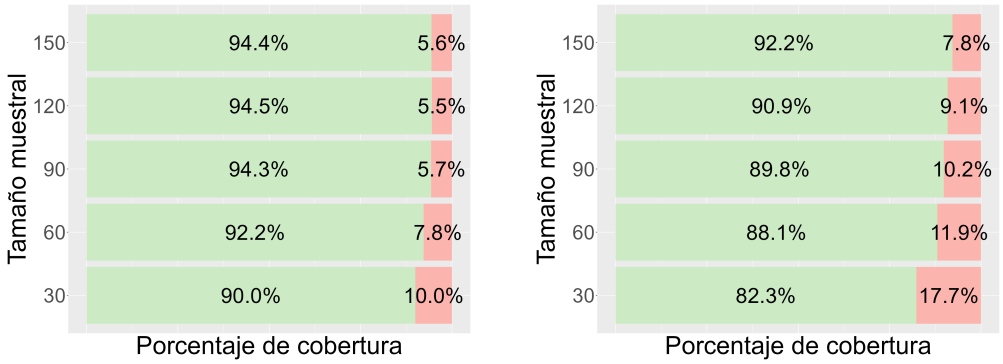


(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.2.3 – Distribución de la varianza estimada de $\hat{t}_{\text{ingresos}}$ bajo los modelos homogéneo y lineal general para la población MU281 y muestras seleccionadas mediante un diseño simple de 30 elementos.

6.2.1.3. Cobertura de los intervalos de confianza

La Figura 6.2.4 detalla el porcentaje de cobertura de los intervalos de confianza al 95 % para t_{ingresos}. Al igual que ocurría con las poblaciones simuladas, en términos generales dicho porcentaje aumenta con el tamaño muestral en ambos modelos (ver Figura 6.1.4). Sin embargo, a diferencia de lo observado para datos ficticios, en este caso el modelo homogéneo se acerca más a la cobertura deseada que el modelo lineal general para todos los tamaños de muestra considerados. De esta manera, para el modelo homogéneo, la cobertura crece desde un 90 % cuando $n = 30$ a un 94,5 % para $n = 120$, valor en el que se mantiene aproximadamente estable. En cambio, bajo el modelo lineal general se pasa de un 82,3 % para muestras de tamaño 30 a un 92,2 % en muestras de 150 elementos. Nuevamente, es probable que esto se deba a que el modelo de población homogénea es más robusto a errores de especificación que el modelo de población lineal general.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

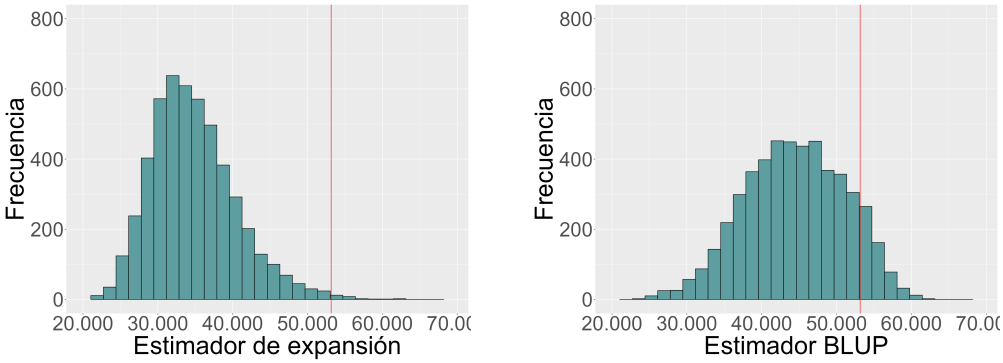
Figura 6.2.4 – Cobertura de los intervalos de confianza al 95 % bajo los modelos homogéneo y lineal general para la población MU281 y muestras seleccionadas mediante un diseño simple.

6.2.2. Muestras simples truncadas

Al remover de la muestra los 5 municipios más grandes en términos de ingresos, la dispersión de la variable *ing_muni_85* en la muestra tenderá a diferir de la observada para la población, lo cual afectará las estimaciones de los parámetros superpoblacionales. Es decir, que al hecho de no conocer la forma del modelo, se le agregará una fuente de error como es el hecho de tener muestras "imperfectas". Además, es interesante destacar que este problema es más "realista" que el trabajado en la Sección 6.1 en la medida que replica un patrón de no respuesta bastante frecuente, según el cual los individuos con valores "altos" en términos de la variable de interés son menos propensos a revelarlo.

6.2.2.1. Estimaciones del total

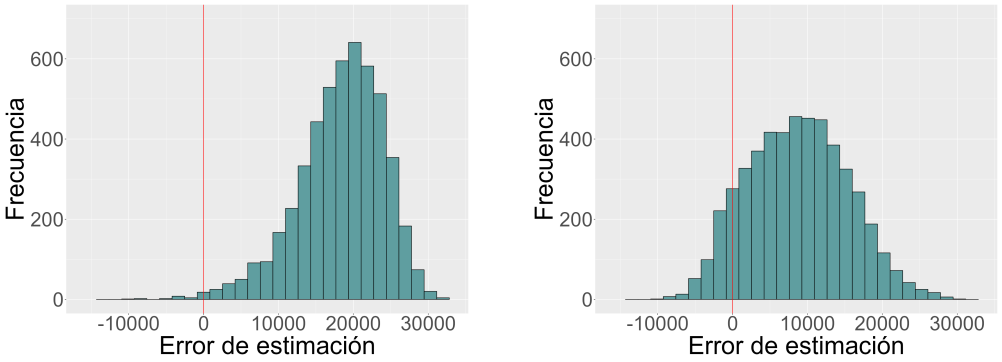
La Figura 6.2.5 contiene los histogramas de las estimaciones del total de ingresos municipales obtenidas a partir de muestras truncadas. Al igual que en el caso de las poblaciones simuladas, se observa que las estimaciones subestiman en forma sistemática t_{ingresos} (ver Figura 6.1.5). Como se muestra en la Tabla 6.2.1, en promedio, los modelos homogéneo y lineal general arrojaron estimaciones de 34.814,9 y 44.449,0, respectivamente.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.2.5 – Distribución de los estimadores de t_{ingresos} basados en los modelos homogéneo y lineal general para la población MU281 y muestras simples truncadas de tamaño 30. La línea roja corresponde al total de ingresos municipales.

Sin embargo, este problema se vio bastante más acentuado en la población MU281, sobre todo en lo que respecta al modelo lineal general. Como muestra la Figura 6.2.6, la distribución de los errores se encuentra bastante por encima del cero, con lo cual se concluye que existe un sustancial sesgo a la derecha. En términos absolutos, el error medio medio fue de 18.336 en el caso del modelo homogéneo y 8.702 en el del modelo lineal general, lo cual equivale a errores relativos de 34,5 % y 16,4 % (ver Tabla 6.2.1).



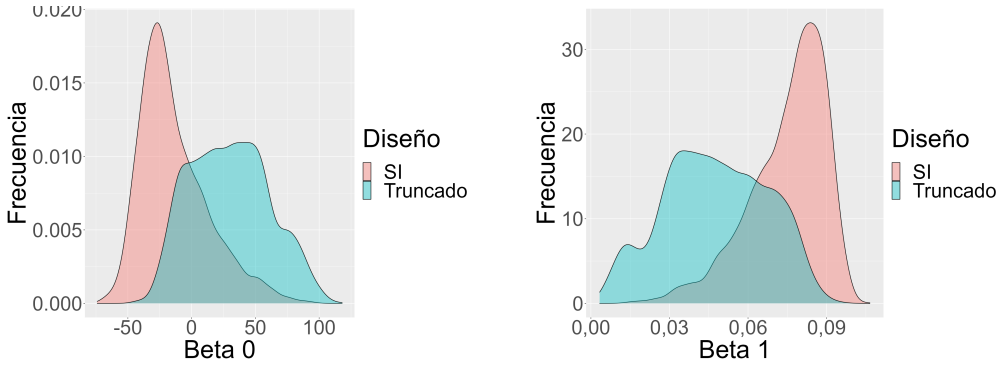
(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.2.6 – Distribución de los errores de estimación de $\hat{t}_{\text{ingresos}}$ bajo los modelos homogéneo y lineal general para la población MU281 y muestras simples truncadas de tamaño 30.

Al contrario de lo que ocurría con las muestras simples, en este caso el modelo homogéneo exhibe un peor desempeño que el modelo lineal. Es probable que existan problemas de especificación con ambos modelos; no obstante,

posiblemente el modelo lineal general funciona mejor debido a que hace mejor uso de la información auxiliar disponible.

Si bien el modelo lineal general exhibe un mejor desempeño que el modelo homogéneo, la estimación de sus parámetros superpoblacionales se vio fuertemente afectada al truncar las muestras. Como muestra la Figura 6.2.7, al eliminar los 5 municipios de mayores ingresos de cada muestra, la distribución de β_0 se desplazó hacia la derecha y la de β_1 hacia la izquierda. Si se toman como marco de referencia las distribuciones obtenidas bajo muestras simples, esto implica que se pasó a subestimar el efecto esperado sobre los ingresos asociado a incrementar en un millón de coronas suecas los valores inmobiliarios de un cierto municipio. Esto es coherente con el sesgo positivo de las estimaciones evidente en las figuras 6.2.5b y 6.2.5a.



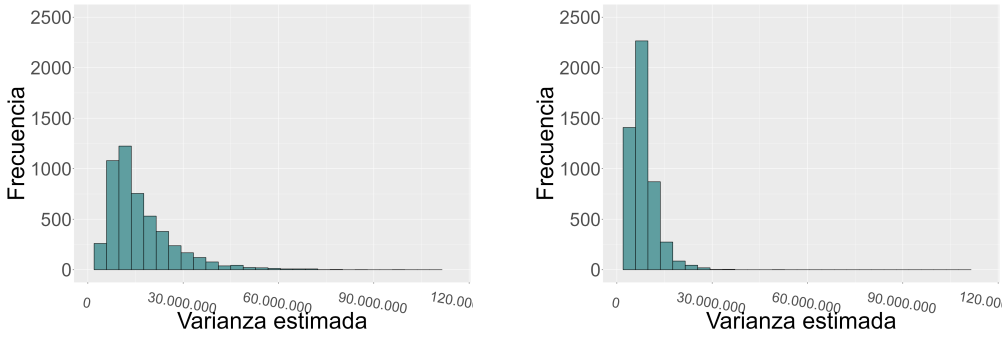
(a) Estimación de β_0 **(b)** Estimación de β_1

Figura 6.2.7 – Densidades de $\hat{\beta}_0$ y $\hat{\beta}_1$ en el marco del modelo de población lineal general. Se comparan los resultados obtenidos para la población MU281 mediante muestras simples de 30 elementos truncadas y no truncadas.

6.2.2.2. Estimaciones de la varianza de los estimadores

La Figura 6.2.8 muestra la distribución de las estimaciones de la varianza de $\hat{t}_{\text{ingresos}}$ al truncar las muestras. Aunque se advierte una mayor dispersión de las estimaciones bajo el modelo homogéneo que bajo el modelo lineal general, las diferencias entre ambos modelos no fueron tan importantes como al utilizar muestras simples no truncadas. Así, la estimación media de $\widehat{\text{Var}}\left(\hat{t}_{\text{ingresos}}\right)$ fue de aproximadamente 16,6 millones bajo el modelo homogéneo y 8 millones para el modelo lineal, lo cual equivale a un desvío estándar de 4.071 y 2.870 (ver Tabla 6.2.1). Ambos valores fueron mucho menores a los obtenidos a partir de muestras SI, lo cual sugiere que el truncar las muestras llevó a subestimar

la variabilidad de la población en términos de los ingresos municipales y, por consiguiente, también la varianza de las estimaciones.

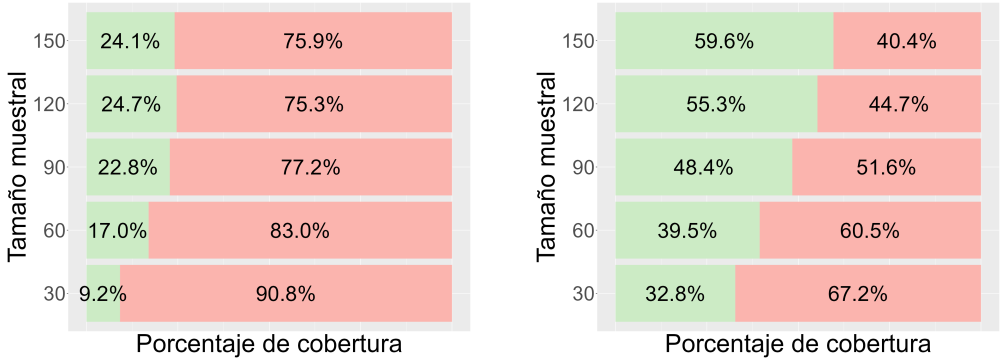


(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.2.8 – Distribución de la varianza estimada de $\hat{t}_{\text{ingesos}}$ bajo los modelos homogéneo y lineal general para la población MU281 y muestras truncadas de tamaño 30.

6.2.2.3. Cobertura de los intervalos de confianza

Como muestra la Figura 6.2.9, ambos modelos conducen a bajísimas tasas de cobertura de los intervalos de confianza al 95 % para todos los tamaños de muestra considerados, muy inferiores a las obtenidas para las poblaciones simuladas (ver Figura 6.1.9). Esto se debe tanto a que las estimaciones son sesgadas como a que la estimación de su varianza es excesivamente baja. De esta forma, los intervalos no sólo se ubicaron relativamente lejos de t_{ingresos}, sino que resultaron ser excesivamente angostos. En consecuencia, la probabilidad de captar el verdadero valor del parámetro de interés se vio fuertemente reducida.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.2.9 – Cobertura de los intervalos de confianza al 95 % bajo los modelos homogéneo y lineal general para poblaciones simuladas y muestras truncadas.

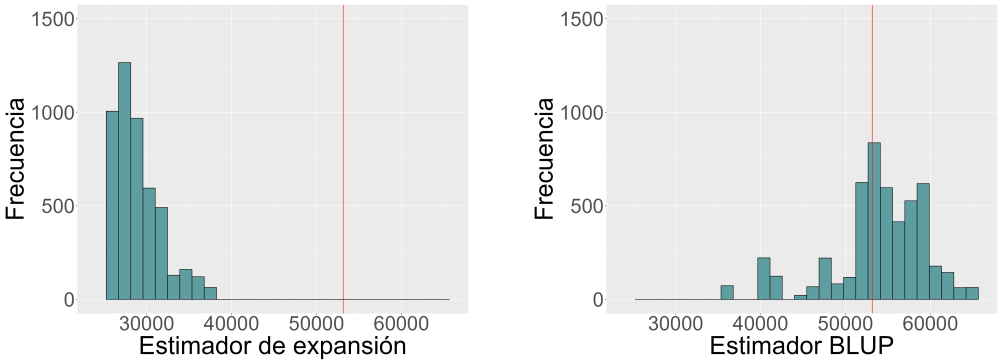
A pesar de que en todos los casos la cobertura se encontró muy por debajo de su nivel deseado, el modelo lineal exhibe un desempeño algo mejor que el modelo homogéneo para cada valor de n. Asimismo, el porcentaje crece con el tamaño muestral. Por un lado, la cobertura del modelo homogéneo creció desde 9,2 % cuando $n = 30$ hasta 24,1 % cuando $n = 150$. Por su parte, para el modelo lineal, la cobertura pasó de 32,8 % para $n = 30$ a 59,6 % para $n = 150$.

6.2.3. Muestras πps

El tercer esquema de muestreo utilizado fue un diseño πps. Para ello, se establecieron probabilidades de inclusión proporcionales al inverso de la covariable *val_inmob_84*, de forma que los municipios más bajos en términos de dicha variable tuvieron una mayor probabilidad de ser seleccionados. En la medida que los valores inmobiliarios de cada municipio poseen una correlación positiva con sus ingresos (ver Figura 4.2.2), es esperable que esto genere un sesgo negativo en las estimaciones.

6.2.3.1. Estimaciones del total

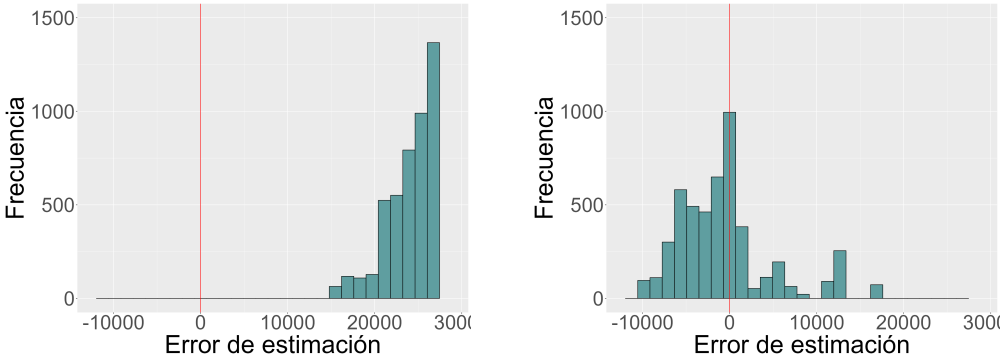
Como fue anticipado, a partir de la Figura 6.2.10a, se concluye que el modelo homogéneo subestima fuertemente t_{ingresos}. En efecto, toda la distribución estimada a partir de 5.000 réplicas para $\hat{t}_{\text{ingresos}}$ se encuentra sumamente alejada del verdadero total y la estimación media fue de apenas 28.806 (ver Tabla 6.2.1). En contraste, aunque la distribución de las estimaciones de la Figura 6.2.10b no es simétrica, se encuentra centrada en 53.942. Es decir, que aún bajo un diseño muestral no ignorable, el modelo lineal general arrojó resultados razonablemente buenos.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.2.10 – Distribución de los estimadores de t_{ingresos} basados en los modelos homogéneo y lineal general para la población MU281 y muestras πps de tamaño 30. La línea roja corresponde al total de ingresos municipales.

En la Figura 6.2.11, se presentan los correspondientes histogramas de los errores de estimación. Nuevamente, se aprecia un sesgo importante para el modelo homogéneo. Como se detalla en la Tabla 6.2.1, en promedio, se subestimó el verdadero total de ingresos en más de 24.000 millones de coronas suecas, lo cual implica que el error relativo medio fue de 45,8 %. Por su parte, el error medio bajo el modelo lineal fue de -791,5 millones de coronas suecas, o equivalentemente, de -1,5 %. De esta manera, el sesgo fue mucho más pequeño que bajo el modelo homogéneo, aun cuando es probable que existan errores de especificación.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

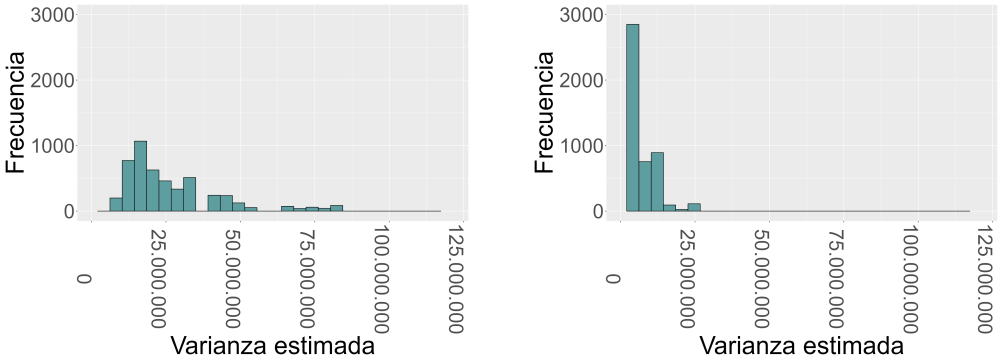
Figura 6.2.11 – Distribución de los errores de estimación de $\hat{t}_{\text{ingresos}}$ bajo los modelos homogéneo y lineal general para la población MU281 y muestras πps de tamaño 30.

Si bien el modelo lineal general arrojó mejores resultados que el modelo homogéneo, los mismos fueron bastante peores que los obtenidos para las

poblaciones simuladas (ver Tabla 6.1.1). Es probable que esto se deba a que el modelo superpoblacional es desconocido, con lo cual los problemas asociados a utilizar un diseño muestral no ignorable se ven agravados por potenciales errores de especificación.

6.2.3.2. Estimaciones de la varianza de los estimadores

A continuación, se presentan los histogramas de $\widehat{\text{Var}}\left(\hat{t}^{E}_{\text{ingresos}}\right)$ y $\widehat{\text{Var}}\left(\hat{t}^{L}_{\text{ingresos}}\right)$ para las muestras extraídas bajo un diseño πps. Al comparar las figuras 6.2.12a y 6.2.12b, se advierte que la distribución de las estimaciones de la varianza bajo el modelo lineal se concentra en valores más bajos que las obtenidas a partir del modelo homogéneo. Mientras que la estimación promedio de $\widehat{\text{Var}}\left(\hat{t}^{L}_{\text{ingresos}}\right)$ fue de alrededor de 6,5 millones, la media de $\widehat{\text{Var}}\left(\hat{t}^{E}_{\text{ingresos}}\right)$ fue de más de 27,8 millones (ver Tabla 6.2.1). De esta forma, los desvíos estándar medios fueron de 2.551,3 y 5.277,2, respectivamente. Dado que ambos valores son inferiores a los obtenidos para un diseño simple, es probable que se esté subestimando la verdadera variabilidad del estimador del total.



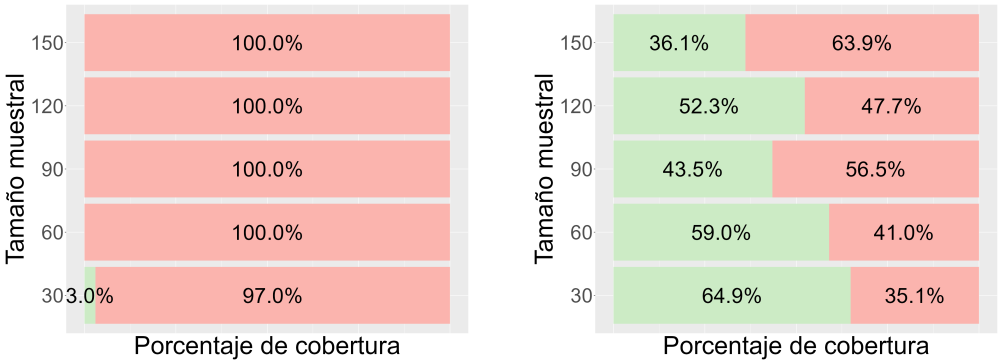
(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.2.12 – Distribución de la varianza estimada de $\hat{t}_{\text{ingresos}}$ bajo los modelos homogéneo y lineal general para la población MU281 y muestras πps de tamaño 30.

6.2.3.3. Cobertura de los intervalos de confianza

En la Figura 6.1.13, se indican los porcentajes de cobertura de los intervalos de confianza al 95 % para cada modelo y tamaño de muestra. En primer lugar, se observa que, bajo el modelo homogéneo, la cobertura es prácticamente nula para todos los tamaños considerados. Para explorar la naturaleza de esta reducción, se recurrió a la Figura 6.2.14, la cual muestra el porcentaje de

cobertura asociado a muestras de entre 10 y 50 elementos. Se observa que esta tasa cae con gran rapidez a medida que crece la cantidad de municipios seleccionados. Así, se pasa de una cobertura de alrededor de 22 % para cuando $n = 10$ hasta prácticamente 0 % cuando $n = 40$. Esto parece deberse tanto al sesgo de las estimaciones como al hecho de que las estimaciones de la varianza fueron excesivamente bajas, de forma que los intervalos resultantes fueron demasiado angostos.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.2.13 – Cobertura de los intervalos de confianza al 95 % bajo los modelos homogéneo y lineal general para la población MU281 y muestras πps.

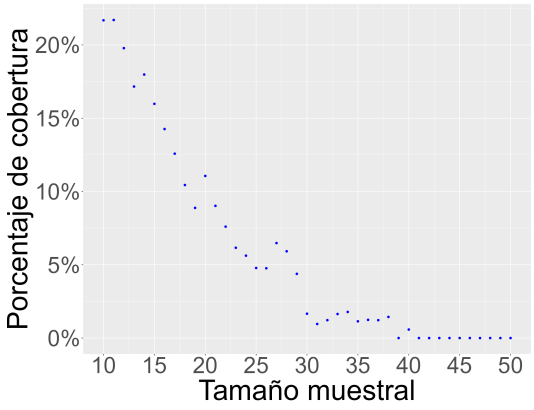


Figura 6.2.14 – Cobertura de los intervalos de confianza al 95 % para el modelo de población homogénea para muestras de entre 10 y 50 elementos.

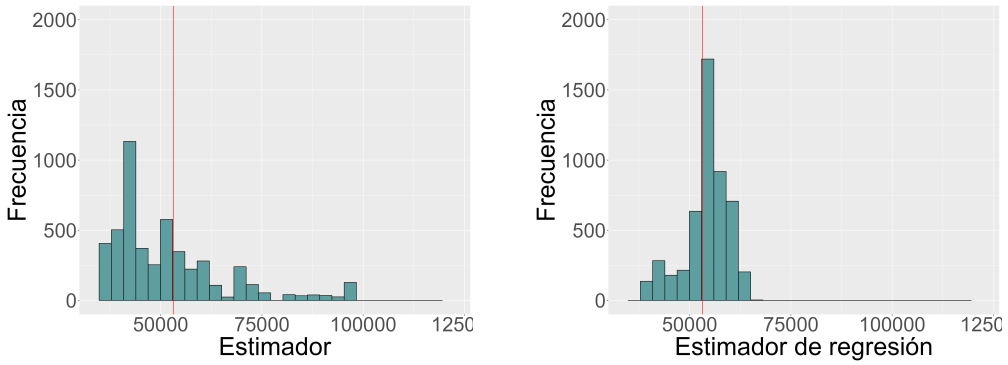
En segundo lugar, si bien se advierte un desempeño algo mejor para el modelo lineal general, también en este caso el porcentaje de cobertura estuvo muy por debajo de su valor deseado. A su vez, se observa que dicho porcentaje decrece a medida que el tamaño de muestra se incrementa. Así, se pasó de una cobertura del 64,9 % para $n = 30$ hasta 36,1 % para $n = 150$. Nuevamente, esto

probablemente se deba a que las estimaciones de la varianza son excesivamente pequeñas.

6.2.3.4. Comparación con la inferencia asistida por modelos

Al igual que para las poblaciones simuladas, se comparó el desempeño de las anteriores estimaciones basadas en modelos con su "versión" asistida por modelos, dada por el estimador de regresión. Nuevamente, sólo existen diferencias entre ambos enfoques cuando se utiliza un diseño muestral πps, en el que las probabilidades de inclusión variables generan diferencias entre ambos enfoques.

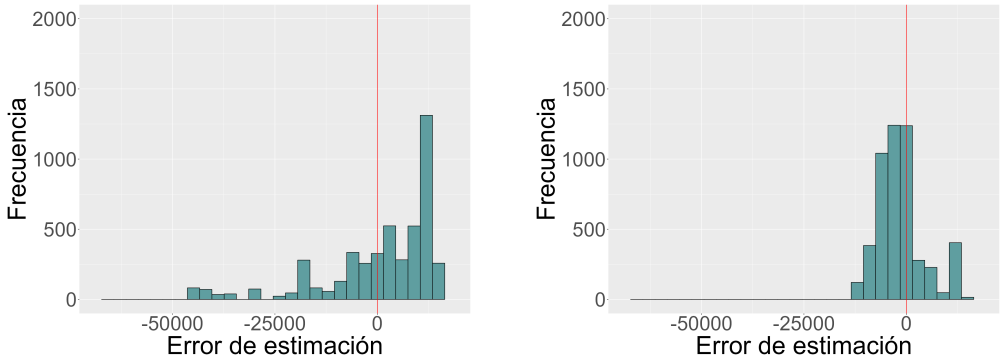
En la Figura 6.2.15, se exhibe la distribución de los dos estimadores de regresión considerados. En primer lugar, se advierte que al calibrar los resultados únicamente por el tamaño de la población, se logró corregir el sesgo negativo que se observó en la Figura 6.2.10a, correspondiente al modelo homogéneo. Como se detalla en la Tabla 6.2.1, bajo el paradigma asistido por modelos, la distribución de las estimaciones bajo este modelo tiene su centro cerca del total de ingresos real, con una estimación media de 52.428,7 millones de coronas.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.2.15 – Distribución de los estimadores de t_{ingresos} asistidos por los modelos homogéneo y lineal general para la población MU281 y muestras πps de tamaño 30. La línea roja corresponde al total de ingresos municipales.

Análogamente, como muestra la Figura 6.2.16, los errores de estimación pasaron a distribuirse entorno a cero aproximadamente. En términos absolutos, el error promedio fue de 722,3 millones, lo cual equivale a un error relativo de 1,4 % (ver Tabla 6.2.1). Estos resultados sugieren que basta con utilizar información auxiliar mínima para mejorar sensiblemente la calidad de las estimaciones.

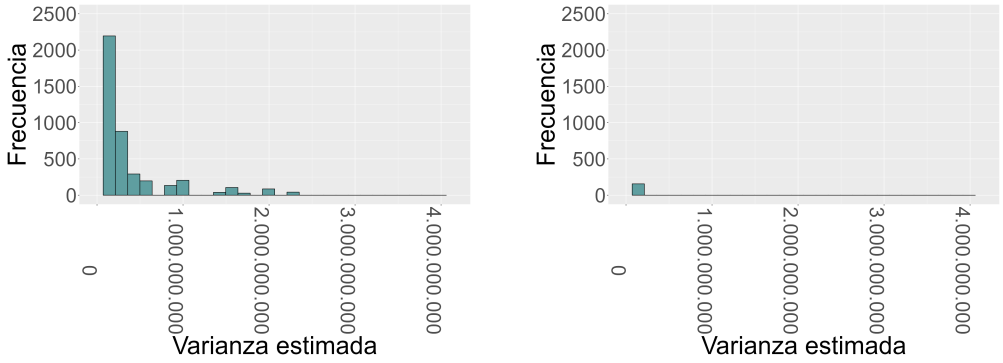


(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.2.16 – Distribución de los errores de estimación de $\hat{t}_{\text{ingresos}}$. La inferencia es asistida por los modelos homogéneo y lineal general para la población MU281 y muestras πps de tamaño 30.

Por otro lado, al comparar las figuras 6.2.10a y 6.2.15a, se concluye que los resultados bajo el modelo lineal general fueron muy similares bajo ambos enfoques. Esto es razonable en la medida que las estimaciones basadas en modelos fueron aproximadamente insesgadas y, por lo tanto, no debieron ser corregidas calibrando los ponderadores mediante el total de valores inmobiliarios. Como se indica en la Tabla 6.2.1, en promedio, se obtuvo una estimación del total de ingresos de 54.255,6 millones de coronas suecas, lo cual implica que el error medio fue de -1.104,6. Por su parte, el error relativo fue de -2,1 %, superior al obtenido sin calibrar.

En la Figura 6.2.17, se presentan las estimaciones de la varianza del estimador de regresión. Para los dos modelos considerados, se observa que las estimaciones de la varianza crecieron fuertemente con respecto a su versión basada en modelos (ver Figura 6.2.12). Esto resulta positivo debido a que, como se vio, dichas estimaciones subestiman la variabilidad del estimador por el mero hecho de que el diseño πps asigna probabilidades de inclusión relativamente bajas a los municipios de ingresos más altos.

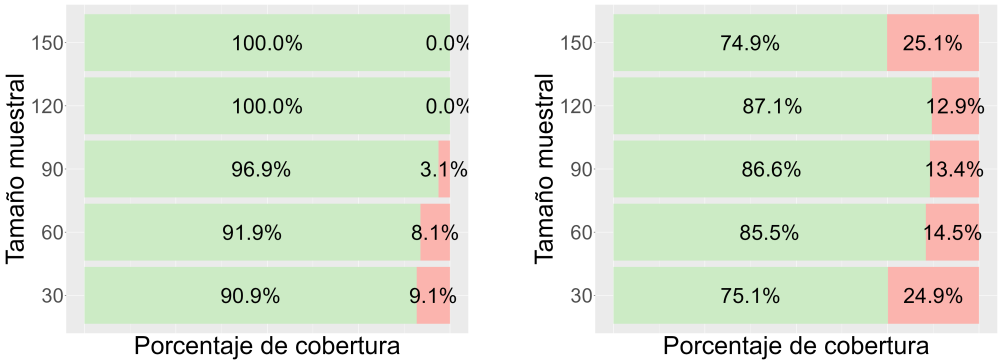


(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.2.17 – Distribución de la varianza estimada de $\hat{t}_{\text{ingresos}}$. La inferencia es asistida por los modelos homogéneo y lineal general para la población MU281 y muestras πps de tamaño 30.

Asimismo, la Figura 6.2.17 muestra que el modelo lineal arroja estimaciones mucho más precisas que el modelo homogéneo, lo cual resulta esperable en la medida que hace uso de más información auxiliar. En promedio, se estimó una varianza de 383,7 millones y de 20,7 millones para los modelos homogéneo y lineal, respectivamente (ver Tabla 6.2.1). Los correspondientes desvíos estándar fueron, entonces, de 19.589,4 y 4.546,4 millones de coronas.

Finalmente, en la Figura 6.2.18 se observa que la cobertura de los intervalos de confianza al 95 % mejoraron para ambos modelos al calibrar los resultados. Para el modelo homogéneo, dichos porcentajes superaron incluso el nivel preestablecido. Teniendo en cuenta que la estimación de la varianza aumentó fuertemente, es probable que esta mejora en la cobertura se deba a que los intervalos de confianza fueron excesivamente amplios. Es decir, que se incrementó la cobertura de los intervalos al costo de una menor precisión, con lo cual los intervalos resultaron ser menos informativos.



(a) Modelo de población homogénea **(b)** Modelo de población lineal general

Figura 6.2.18 – Cobertura de los intervalos de confianza al 95 % para estimaciones asistidas por los modelos homogéneo y lineal general para la población MU281 y muestras πps.

En lo que respecta al modelo lineal, también se aprecia una mejora en la cobertura al utilizar un enfoque asistido por modelos, producto de la mejora en la estimación de la varianza. Sin embargo, para ningún tamaño muestral se llegó al nivel deseado de 95 %.

6.2.4. Estadísticos de resumen

La Tabla 6.2.1 contiene el promedio de las estimaciones del total ($\hat{t}_{\text{ingresos}}$) y la varianza del estimador ($\widehat{\text{Var}}\left(\hat{t}_{\text{ingresos}}\right)$), así como el error medio en términos absolutos ($e\left(\hat{t}_{\text{ingresos}}\right)$) y relativos $p\left(\hat{t}_{\text{ingresos}}\right)$).

Modelo homogéneo				
Diseño	$\hat{\mathbf{t}}_{\text{ingresos}}$	$\widehat{\mathbf{Var}}\left(\hat{\mathbf{t}}_{\text{ingresos}}\right)$	$\mathbf{e}\left(\hat{\mathbf{t}}_{\text{ingresos}}\right)$	$\mathbf{p}\left(\hat{\mathbf{t}}_{\text{ingresos}}\right)$ [%]
Simple	53 159,0	94 135 923,7	−8,0	−0,01
Truncado	34 814,9	16 573 690,3	18 336,1	34,50
πps (MB)	28 806,1	27 848 973,9	24 344,9	45,80
πps (MA)	52 428,7	383 748 129,1	722,3	1,36
Modelo lineal general				
Diseño	$\hat{\mathbf{t}}_{\text{ingresos}}$	$\widehat{\mathbf{Var}}\left(\hat{\mathbf{t}}_{\text{ingresos}}\right)$	$\mathbf{e}\left(\hat{\mathbf{t}}_{\text{ingresos}}\right)$	$\mathbf{p}\left(\hat{\mathbf{t}}_{\text{ingresos}}\right)$ [%]
Simple	53 051,1	15 166 902,2	99,9	0,19
Truncado	44 449,0	8 238 507,3	8702,0	16,37
πps (MB)	53 942,5	6 508 936,4	−791,5	−1,49
πps (MA)	54 255,6	20 668 725,3	−1104,6	−2,08

Tabla 6.2.1 – Promedios de las estimaciones y error medio (absoluto y relativo) bajo los modelos homogéneo y lineal general para las 5.000 muestras de la población MU281.

Capítulo 7

Discusión

A continuación, se discuten los resultados presentados en el capítulo Capítulo 6 y se realizan comparaciones tanto entre poblaciones como entre esquemas de muestreo. Esto permitirá permitirá arribar a las conclusiones de este trabajo y confirmar o refutar las hipótesis planteadas en el Capítulo 1.

7.1. Muestras simples

En primer lugar, se trabajó con un diseño SI, el cual puede ser concebido como un caso "ideal". Esto se debe a que, en promedio, el mismo tiende a replicar la distribución de la variable de interés en la población, con lo cual los parámetros superpoblacionales podrán ser estimados adecuadamente. Por lo tanto, si el modelo fue correctamente especificado, las predicciones del valor que toma la variable de interés en cada elemento de la población serán razonablemente buenas.

Bajo estas condiciones "óptimas", los resultados para las poblaciones simuladas fueron satisfactorios. Así, las estimaciones fueron insesgadas tanto para el modelo homogéneo como para el modelo lineal general, aunque bajo este último se logró un mayor nivel de precisión. Esto permitió alcanzar tasas de cobertura cercanas al nivel preestablecido para ambos modelos.

También se obtuvieron estimaciones aproximadamente insesgadas para la población MU281 bajo los dos modelos propuestos. Esto es particularmente relevante si se considera el hecho de que en este segundo caso se desconoce cuál es el modelo superpoblacional que generó la población, lo cual implica que los errores de especificación no fueron completamente controlables. En este sentido,

es interesante notar que la covariable seleccionada para el modelo lineal general, a saber, los valores inmobiliarios, no exhibe una correlación lineal demasiado alta con los ingresos de cada municipio (ver Figura 4.2.2). Por lo tanto, es probable que el modelo superpoblacional esté errado, y aún así, el desempeño de los estimadores basados en modelos fue aceptable (si bien la cobertura de los intervalos de confianza fue algo más baja que para las poblaciones simuladas).

7.2. Muestras truncadas

En segundo lugar, se trabajó con muestras truncadas superiormente en términos de la variable de interés, de forma de tener un ejemplo concreto de los efectos de trabajar con una muestra "imperfecta". Como era previsible, esto impactó en la estimación de los parámetros superpoblacionales en ambas poblaciones, lo cual a su vez afectó las predicciones para cada elemento y por lo tanto sesgó la estimación del total de interés.

En el caso de las poblaciones simuladas, el modelo homogéneo tendió a subestimar fuertemente tanto a t_y como a $\text{Var}(\hat{t}_y^E)$, lo cual se tradujo en porcentajes de cobertura deficientes. El modelo lineal general también exhibió un sesgo negativo. No obstante, la magnitud de dicho sesgo fue mucho menor que bajo el modelo homogéneo debido a que la distorsión en las estimaciones de los parámetros β_0 y β_1 fue relativamente pequeña. Además, la estimación de la varianza de $\hat{t}_y^E$ se asemejó bastante a la obtenida bajo un diseño simple, con lo cual los intervalos de confianza resultantes no fueron excesivamente angostos y, por consiguiente, la cobertura se acercó más a los niveles preestablecidos.

En contraste, para la población MU281, los efectos de utilizar muestras truncadas se vieron amplificados bajo ambos modelos. Bajo el modelo homogéneo, en particular, los sesgos fueron tan grandes que las tasas de cobertura de los intervalos de confianza fueron prácticamente nulas. Por su parte, el modelo lineal general logró mitigar parcialmente estos problemas haciendo uso de la información auxiliar disponible. Sin embargo, se observaron sesgos en la estimación de los parámetros superpoblacionales β_0 y β_1. En consecuencia, se continuó subestimando sustancialmente el total de la variable de interés, lo cual, en conjunción con estimaciones de la varianza del estimador excesivamente bajas, redundó en porcentajes de cobertura de los intervalos muy por debajo de su valor deseado.

Estos resultados muestran que los problemas asociados a usar muestras

"imperfectas" se ven fuertemente exacerbados cuando se incurre en errores de especificación del modelo superpoblacional. En este caso, se "superponen" distintas fuentes de error en las predicciones, lo cual conduce a grandes sesgos en las estimaciones.

7.3. Muestras πps

En tercer lugar, se evaluó el desempeño de las predicciones basadas en modelos cuando se extraen muestras mediante un diseño no "ignorable" como el πps, según el cual algunos elementos de la población tienen mayor probabilidad de ser seleccionados que otros. Al igual que al truncar las muestras, esto implicó que la distribución de la variable de interés en la muestra difiriera de la de la población.

Para las poblaciones simuladas, se asignaron probabilidades de inclusión proporcionales al inverso de x. Dado que esta variable presenta una correlación lineal positiva con y, se tendió a incluir en la muestra elementos "pequeños" en términos de la variable de interés. En estas condiciones, el modelo homogéneo subestimó sistemáticamente t_y. Esto es razonable ya que dicho modelo utiliza a la media muestral como predictor, la cual tendió a reducirse dadas las características del diseño implementado. Por el contrario, el modelo lineal general arrojó estimaciones aproximadamente insesgadas. Es posible que esto se deba a que, a diferencia de las muestras truncadas, el diseño πps no impidió que se observaran valores "grandes" de y (si bien lo volvió menos probable). Por lo tanto, la muestra fue capaz de captar la correlación existente entre x y y, y no se introdujeron sesgos en la estimación de los parámetros superpoblacionales.

Cabe destacar que, bajo ambos modelos, la estimación promedio de la varianza del estimador se asemejó a la obtenida bajo un diseño simple. Por este motivo, se presume que, a diferencia de lo observado para las muestras truncadas, en este caso no se subestimó la varianza de $\hat{t}_y^E$ ni de $\hat{t}_y^L$. En la medida que el modelo lineal general arrojó estimaciones insesgadas, esto implicó que la cobertura de los intervalos se acercara a su nivel preestablecido. Por su parte, si bien los intervalos de confianza bajo el modelo homogéneo tuvieron la amplitud adecuada debido a que la varianza $\mathrm{Var}(\hat{t}_y^E)$ fue correctamente estimada, el sesgo en las estimaciones de t_y condujo a tasas de cobertura inferiores a lo deseado.

En suma, el caso de las poblaciones simuladas sugiere que cuando no existen errores de especificación, el utilizar un diseño πps no afectó sustancialmente la

calidad de las estimaciones. En consecuencia, los resultados bajo el modelo lineal general fueron aceptables y no se advirtieron mejoras al utilizar el estimador de regresión. En contraste, las estimaciones bajo el modelo homogéneo exhibieron un sesgo considerable. Sin embargo, el mismo pudo ser corregido calibrando las estimaciones por el tamaño de la población. Es decir, que alcanzó con incorporar información auxiliar mínima para mejorar sensiblemente los resultados, si bien no se logró el nivel de precisión del modelo lineal.

Al repetir el ejercicio para la población MU281, los problemas advertidos para las poblaciones simuladas bajo el modelo homogéneo se vieron fuertemente agudizados. Probablemente, esto se debió a que al hecho de no contar con un diseño "ignorable", se le sumó el utilizar modelos superpoblacionales no perfectamente especificados. Así, se subestimó tanto el total de ingresos municipales como la varianza de las estimaciones. Por consiguiente, la cobertura de los correspondientes intervalos de confianza fue prácticamente nula. Los resultados para el modelo lineal general también empeoraron con respecto a lo observado para las poblaciones simuladas; no obstante, se incurrió en sesgos mucho menores que bajo el modelo homogéneo, y los porcentajes de cobertura fueron bastante más altos.

Al igual que para las poblaciones simuladas, el estimador de regresión permitió corregir el sesgo de las estimaciones obtenidas por el modelo homogéneo. Sin embargo, no se pudo reducir el error promedio bajo el modelo lineal general, aunque las estimaciones de $\mathrm{Var}(\hat{t}^L_{\text{ingresos}})$ tendieron a ser mayores, lo cual redundó en intervalos de confianza más amplios y mayores niveles de cobertura.

Capítulo 8

Conclusiones

8.1. Resultados principales

En el marco del muestreo de poblaciones finitas, este trabajo fue una primera aproximación a la inferencia basada en modelos como paradigma alternativo a la inferencia basada en el diseño, más frecuentemente utilizada. En particular, se evaluó la calidad de las estimaciones basadas en modelos bajo una variedad de situaciones que buscaron dar cuenta de algunos de los factores que inciden (y en ocasiones, distorsionan) el proceso de inferencia.

En primer lugar, se trabajó con dos tipos de poblaciones finitas: las ficticias, en las que cada conjunto de elementos fue simulado a partir de un modelo preestablecido, y la MU281, para la que se conoce cada elemento pero no el modelo superpoblacional que la generó. Como se detalló en el Capítulo 7, los resultados fueron más satisfactorios para las poblaciones simuladas que para la población MU281, lo cual pone de manifiesto las dificultades asociadas a desconocer el modelo superpoblacional y los riesgos de utilizar un modelo incorrecto.

En segundo lugar, se consideraron dos modelos sencillos: el de población homogénea, que no hace uso de información auxiliar, y el de población lineal general, que explota la información dada por variables auxiliares mediante una regresión lineal. Esto permitió visualizar la utilidad de incorporar información adicional a las predicciones con el fin de mejorar la precisión de las estimaciones.

Finalmente, se analizó el efecto de utilizar tres tipos de muestras: un diseño simple, un diseño simple con muestras truncadas superiormente y un diseño πps. El primer caso fue "ideal" en la medida que se utilizó un diseño "ignorable"

sin datos faltantes ni problemas de no respuesta. Es decir, que el mecanismo de selección aseguró que, en promedio, las muestras replicaran la distribución de la variable de interés en la población. Por ende, no se generaron distorsiones en la etapa de estimación de los parámetros superpoblacionales. Por su parte, las muestras truncadas simularon patrones de no respuesta no aleatorios, y las muestras πps brindaron una noción acerca del efecto de usar un diseño "no ignorable". Como era de esperar, en estos dos casos, las estimaciones exhibieron sesgos sistemáticos. Como fue dicho, los errores tendieron a ser mayores para la población MU281, para la cual es probable que existieran problemas de especificación del modelo. Estos resultados ponen de relieve el hecho de que los sesgos que surgen de utilizar muestras "imperfectas" se ven agravados si se combinan con modelos incorrectos.

8.2. Hipótesis

A partir de lo anterior, es posible dar respuesta a las hipótesis planteadas en el Capítulo 1:

1. Bajo un diseño muestral simple, tanto el modelo homogéneo como el lineal general mostraron un buen desempeño para las dos poblaciones consideradas. En cambio, al utilizar muestras truncadas u obtenidas mediante un diseño πps, las estimaciones pasaron a ser sesgadas. De esta manera, se confirma la primera hipótesis formulada.

2. Las estimaciones tendieron a ser más precisas para el modelo lineal general que para el modelo homogéneo debido a que el primero explota la información auxiliar disponible. Esto es consistente con lo propuesto en la segunda hipótesis. Sin embargo, cabe destacar que el uso de muestras "imperfectas" puede generar sesgos. En este caso, una mayor precisión puede traducirse en menores porcentajes de cobertura de los intervalos de confianza. Es decir, que en condiciones no "ideales", la menor variabilidad en las estimaciones puede ser perjudicial.

3. Tal como se planteó en la tercera hipótesis, los resultados fueron mejores para las poblaciones simuladas que para la población MU281. Probablemente esto se deba a que solamente en el primer caso fue posible eliminar completamente los errores de especificación.

4. En el marco del diseño πps, la calibración de las estimaciones por el tamaño de la población permitió corregir el sesgo de las estimaciones basadas en el modelo homogéneo. Por el contrario, el estimador de regresión no mejoró sustancialmente las estimaciones basadas en modelos. Esto contradice lo postulado en la cuarta hipótesis propuesta, según la cual el estimador de regresión siempre arroja mejores resultados que los estimadores basados en modelos. Sin embargo, es importante señalar que para el modelo lineal general, la media de la estimación de la varianza del estimador fue mayor y se acercó más a lo hallado bajo un diseño simple. De esta manera, el enfoque asistido por modelos permitió alcanzar mayores niveles de cobertura que la inferencia basada en modelos.

8.3. Limitaciones y trabajos a futuro

Si bien estos resultados son útiles y dan la pauta de algunos de los aspectos que afectan las estimaciones basadas en modelos, el alcance de este trabajo es acotado y tiene varias limitaciones. Por un lado, las poblaciones simuladas fueron generadas en base a un conjunto de parámetros definidos arbitrariamente. Además, en todos los casos se trabajó con poblaciones relativamente pequeñas. Por consiguiente, a futuro, sería interesante analizar el efecto variar el tamaño poblacional y el valor de los parámetros superpoblacionales.

Asimismo, sólo se trabajó con dos modelos extremadamente sencillos que asumen que la varianza de la variable de interés es constante. En trabajos futuros, podría estudiarse si los resultados se mantienen para otro tipo de modelos más sofisticados.

8.4. Consideraciones finales

Los resultados obtenidos ilustran la fragilidad del enfoque basado en modelos. En línea con lo planteado por la literatura (ver Capítulo 2), se observa que, si bien las estimaciones pueden llegar a ser muy eficientes cuando el modelo utilizado es apropiado, ello no es fácilmente verificable. Lo que es más, en caso de existir errores de especificación importantes, se puede llegar a incurrir en sesgos sustanciales, sobre todo si las muestras utilizadas son “imperfectas”.

Frente a esta situación, la alternativa frecuentemente utilizada es recurrir

a la inferencia asistida por modelos. Este paradigma usa los modelos para describir la población, pero la inferencia sigue estando basada en el diseño. De esta manera, logra combinar el insesgamiento propio de los estimadores basados en el diseño con la eficiencia del enfoque basado en modelos. Sin embargo, los resultados muestran que, si las muestras no tienen grandes problemas y el modelo es razonable, los beneficios de utilizar el enfoque asistido por modelos son limitados. En estos casos, puede ser preferible recurrir al enfoque basado en modelos, en general menos costoso.

De esta forma, a pesar de sus desventajas, el enfoque basado en modelos puede arrojar buenas estimaciones siempre y cuando se lo utilice con precaución. Por ende, es importante conocer sus potencialidades y limitaciones en comparación con el paradigma basado en el diseño y el asistido por modelos, de forma de poder determinar cuál de ellos se adecúa más a cada situación concreta. En palabras de Brewer y Gregoire (2009)[1]:

> Las tres visiones siguen vigentes. La visión predominante es que la inferencia basada en el diseño y la asistida por modelos ha funcionado bien por varias décadas, y las razones para cambiarlas son insuficientes. El enfoque basado en modelos continúa siendo presentado por otros como el único que puede ser utilizado consistentemente por un estadístico bien formado. Y algunos se preguntan "¿Por qué no usar ambos [enfoques]?" Sólo el tiempo y la experiencia resolverán el tema, pero mientras tanto, las tres visiones tienen que ser claramente comprendidas.

[1]Traducción propia del inglés.

Referencias

Aubry, P., y Francesiaz, C. (2022). On comparing design-based estimation versus model-based prediction to assess the abundance of biological populations. *Ecological Indicators*, *144*.

Bank, A. D. (2020). *Introduction to Small Area Estimation Techniques: A Practical Guide for National Statistics Offices*.

Brewer, K., y Gregoire, T. G. (2009). Chapter 1 - Introduction to Survey Sampling. En C. Rao (Ed.), *Handbook of Statistics* (pp. 9-37, Vol. 29). Elsevier. https://doi.org/https://doi.org/10.1016/S0169-7161(08)00001-1

Brus, D., y De Gruijter, J. (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, *80*(1-2), 1-44.

Chambers, R., y Clark, R. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford University Press.

Corral Rodas, P. A., Molina, I., Cojocaru, A., y Segovia Juarez, S. C. (2020). *Guidelines to small area estimation for poverty mapping*. World Bank Group. http://documents.worldbank.org/curated/en/099115306242236696/P1694340364c9803d0b7df097798bc42eac

Dever, J. A., y Valliant, R. (2018). Survey Weights: A Step-by-step Guide to Calculation.

Geuna, S. (2000). Appreciating the difference between design-based and model-based sampling strategies in quantitative morphology of the nervous system. *Journal of Comparative Neurology*, *427*(3), 333-339.

Ghosh, M., y Rao, J. N. K. (1994). Small Area Estimation: An Appraisal. *Statistical Science*, *9*(1), 55-76. https://doi.org/10.1214/ss/1177010647

Ghosh, M. (2020). Small area estimation: its evolution in five decades. *Statistics in Transition New Series*, *21*(4), 1-22. https://doi.org/10.21307/stattrans-2020-022

Godambe, V. P. (1955). A Unified Theory of Sampling From Finite Populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *17*(2), 269-278.

Gregoire, T. G. (1998). Design-based and model-based inference in survey sampling: appreciating the difference. *Canadian Journal of Forest Research*, *28*(10), 1429-1447.

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, *99*(466), 546-556.

Lohr, S. L. (2009). Introduction to Part 1. En C. Rao (Ed.), *Handbook of Statistics* (pp. 3-8, Vol. 29). Elsevier. https://doi.org/https://doi.org/10.1016/S0169-7161(09)70006-9

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R*. John Wiley; Sons.

Lumley, T., y Scott, A. (2017). Fitting regression models to survey data. *Statistical Science*, 265-278.

Nascimento Silva, P., y Skinner, C. J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, *23*(1), 23-32.

Natarajan, S., Lipsitz, S. R., Fitzmaurice, G., Moore, C. G., y Gonin, R. (2008). Variance estimation in complex survey sampling for generalized linear models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *57*(1), 75-87.

National Statistics. (2006). Model-Based Estimates of ILO Unemployment for LAD/UAs in Great Britain. Guide for Users.

Nedyalkova, D., y Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika*, *95*(3), 521-537.

Nordberg, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics*, *5*(3), 223-239.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, 317-337.

Pfeffermann, D. (2002). Small Area Estimation: New Developments and Directions. *International Statistical Review / Revue Internationale de Statistique*, *70*(1), 125-143. Consultado el 21 de agosto de 2024, desde http://www.jstor.org/stable/1403729

Rao, J. (2005). *Small Area Estimation*. Wiley.

Rao, J., y Molina, I. (2015). *Small Area Estimation*. Wiley.

Rondon, L. M., Vanegas, L. H., y Ferraz, C. (2012). Finite population estimation under generalized linear model assistance. *Computational Statistics & Data Analysis*, *56*(3), 680-697.

Royall, R. M. (1976). The Linear Least-Squares Prediction Approach to Two-Stage Sampling. *Journal of the American Statistical Association*, *71*(355), 657-664. https://doi.org/10.1080/01621459.1976.10481542

Royall, R. M. (1992). The model based (prediction) approach to finite population sampling theory. *Lecture Notes-Monograph Series*, *17*, 225-240.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Särndal, C.-E., Swensson, B., y Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer New York.

Särndal, C.-E., Thomsen, I., Hoem, J. M., Lindley, D. V., Barndorff-Nielsen, O., y Dalenius, T. (1978). Design-Based and Model-Based Inference in Survey Sampling [with Discussion and Reply]. *Scandinavian Journal of Statistics*, *5*(1), 27-52.

Shi, Y., Cameron, C. J., y Heckathorn, D. D. (2019). Model-based and design-based inference: reducing bias due to differential recruitment in respondent-driven sampling. *Sociological Methods & Research*, *48*(1), 3-33.

Smith, T. (1994). Sample surveys 1975-1990; an age of reconciliation? *International Statistical Review/Revue Internationale de Statistique*, *62*(1), 5-19.

Solon, G., Haider, S. J., y Wooldridge, J. M. (2015). What are we weighting for? *Journal of Human resources*, *50*(2), 301-316.

Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S. P., Patterson, P. L., Magnussen, S., Næsset, E., McRoberts, R. E., et al. (2016). Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. *Forest Ecosystems*, *3*(1), 1-11.

Tillé, Y., y Matei, A. (2021). *sampling: Survey Sampling* [R package version 2.9]. https://CRAN.R-project.org/package=sampling

Valliant, R. (2009). Chapter 23 - Model-Based Prediction of Finite Population Totals. En C. Rao (Ed.), *Handbook of Statistics* (pp. 11-31, Vol. 29). Elsevier. https://doi.org/https://doi.org/10.1016/S0169-7161(09)00223-5

Wu, C. (2022). Statistical inference with non-probability survey samples. *Statistics Canada*, *48*(2). https://www150.statcan.gc.ca/n1/pub/12-001-x/2022002/article/00002-eng.htm

APÉNDICES

Apéndice 1

Anulación de los errores en el estimador de regresión

Como se afirma en el Capítulo 1, cuando la varianza de la variable de interés para cada observación es proporcional al valor de alguna covariable (ver Ecuación 3.83), se anula el término de error del estimador de regresión, t_y^{reg}. Esto se demuestra como sigue:

$$\begin{aligned}
\sum_s \frac{e_i}{\pi_i} &= \sum_s \frac{(y_i - \hat{y}_i)}{\pi_i} \\
&= \sum_s \frac{y_i}{\pi_i} - \sum_s \frac{\mathbf{x}_i'\hat{\mathbf{B}}}{\pi_i} = \sum_s \frac{y_i}{\pi_i} - \sum_s \frac{\sigma_i^2 \mathbf{x}_i'\hat{\mathbf{B}}}{\sigma_i^2\pi_i} \\
&= \sum_s \frac{y_i}{\pi_i} - \lambda' \sum_s \frac{\mathbf{x}_i\mathbf{x}_i'}{\sigma_i^2\pi_i}\hat{\mathbf{B}} \\
&= \sum_s \frac{y_i}{\pi_i} - \lambda' \sum_s \frac{\mathbf{x}_i\mathbf{x}_i'}{\sigma_i^2\pi_i} \left(\sum_s \frac{\mathbf{x}_i\mathbf{x}_i'}{\sigma_i^2\pi_i}\right)^{-1} \sum_s \frac{\mathbf{x}_i y_i}{\sigma_i^2\pi_i} \\
&= \sum_s \frac{y_i}{\pi_i} - \lambda' \sum_s \frac{\mathbf{x}_i y_i}{\sigma_i^2\pi_i} \\
&= \sum_s \frac{y_i}{\pi_i} - \sum_s \frac{\sigma_i^2 y_i}{\sigma_i^2\pi_i} = \sum_s \frac{y_i}{\pi_i} - \sum_s \frac{y_i}{\pi_i} = 0
\end{aligned} \tag{1.1}$$

Apéndice 2

Ejemplo de una población simulada

La Figura 2.0.1 muestra las realizaciones de X y Y para una de las 5.000 poblaciones simuladas. En primer lugar, en el cuadrante superior izquierdo, se estima la densidad de la variable X para los 300 elementos obtenidos. Se advierte una forma acampanada y simétrica alrededor del $x = 10$, con todas las observaciones comprendidas entre 7 y 13 aproximadamente. En la medida en que los datos fueron generados a partir de una normal de media 10 y varianza 1, estos resultados son razonables.

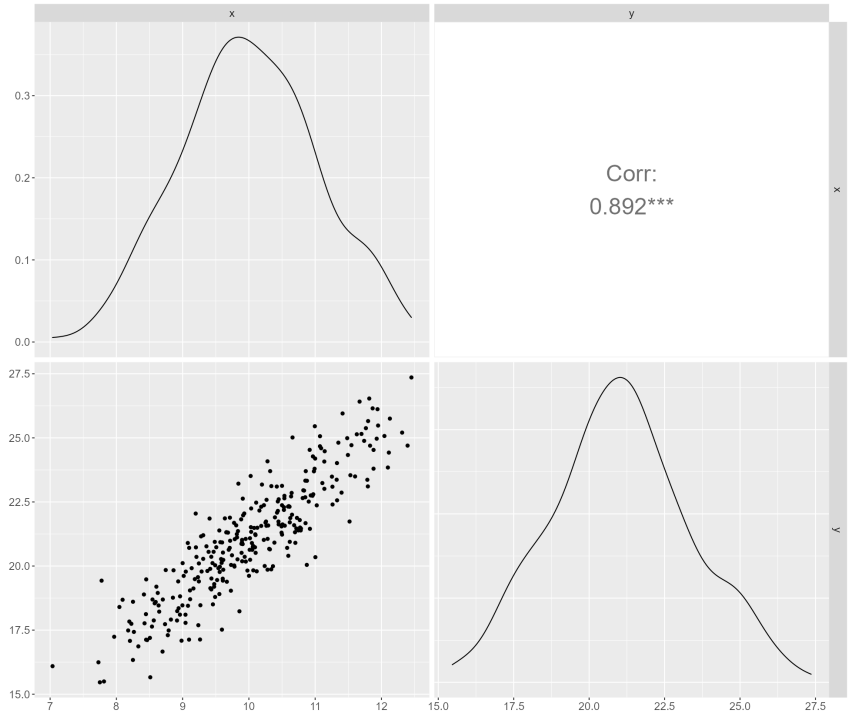


Figura 2.0.1 – Ejemplo de una población simulada. Se presentan las densidades estimadas de x e y, el gráfico de dispersión de ambas y su coeficiente de correlación de Pearson.

En segundo lugar, en el cuadrante inferior derecho de la figura, se muestra la densidad estimada para la variable Y. Los resultados son consistentes con la distribución obtenida en la Ecuación 4.5 de la Subsubsección 4.1.2.1: se aprecia una distribución aproximadamente simétrica y centrada en 21.

Finalmente, los dos paneles restantes de la figura dan cuenta de la relación lineal existente (por construcción) entre X e Y. En esta simulación particular, el coeficiente de correlación lineal de Pearson para x e y fue de 0,89. En el cuadrante inferior izquierdo, el gráfico de dispersión muestra una asociación fuerte entre ambas variables.