Uso de R en la estimación de las tasas de pobreza comunal en Chile: Basado en un modelo de estimación de áreas pequeñas.

Fabiola Aguilar , Angela Arriagada , Alejandra Tapia , Marcelo Rodríguez

Abstract This paper proposes to use the R programming language to estimate poverty rates by commune in Chile. A process of Extraction, Transformation and Loading (ETL) of data is used, together with the Fay-Herriot model for the estimation of small areas (communes). The main objective is to automate the generation of bulletins informing about the communal poverty index in the country. The implementation of this methodology is done through the development of a code in R, making use of several libraries and functions of specific packages. The National Socioeconomic Characterization Survey (CASEN) is used as the official data source for poverty statistics in Chile and the communal estimation MDS (2015) that contains the historical and multidimensional poverty for each commune in Chile.

Palabras clave: Estimación SAE - ETL - R - CASEN - Pobreza comunal

Introducción

No existe una definición única y universal de pobreza, en realidad, no hay una única forma de pobreza, sino que diferentes formas en que este fenómeno se presenta. Pese a lo anterior, un elemento común a todas las definiciones es la privación de los elementos necesarios para la vida humana dentro de la sociedad. Es por eso que durante anos la erradicacion de la pobreza ha sido centro de diferentes politicas en Chile y ha provocado que la politica publica realice diferentes orientaciones para su reduccion.

Desde la decada de 1990, se ha usado una metodologia de medicion de pobreza por ingresos o indirecta. Esta metodología permite distinguir entre quienes se encuentran en situacion de pobreza o pobreza extrema, y aquellos que no lo estan, sobre la base de un determinado nivel de ingresos. Este nivel de ingresos corresponde a un indicador respecto a la capacidad de satisfacer las necesidades basicas, a partir del costo de una canasta basica de alimentos por persona. La pobreza es medida actualmente a nivel oficial a través de dos metodologias: la medicion de pobreza por ingresos y la medicion de pobreza multidimensional. Las estimaciones de las tasas de pobreza para las comunas extremadamente pequeñas con poca muestra no estan disponibles para diferentes tipos de tiempo, ya que la encuesta se enfoca en territorios con grandes areas geograficas, lo que provoca que las estimaciones de encuesta directas de las tasas de pobrezas no posean la precision deseada para estas comunas pequenas, por lo que, la evaluacion politica no es sencilla. Sin embargo, es de gran interes mejorar la metodologia para estimar las tasas de pobreza, para asi, monitorear con mayor exito las tendencias, identificar factores influyentes y desarrollar mejores politicas publicas para la erradicacion de la pobreza.

El Ministerio de Desarrollo Social de Chile decidio crear una metodologia que pudiera aprovechar de mejor manera los datos de CASEN y los registros administrativos disponibles en Chile. En el ano 2011 las estimaciones de las tasas de pobreza a nivel comunal, se obtuvieron utilizando el metodo de estimacion de areas pequenas (SAE) para todas las comunas con muestra en CASEN 2009 y el metodo sintetico Ministerio-PNUD para comunas sin muestra en CASEN 2009.

Usando como base esta metodologia SAE implementada por el Ministerio de Desarrollo Social de Chile, se busca definir un indice de pobreza comunal para Chile, basado en un modelo a nivel de area y automatizar la creacion de los boletines que reportan estos resultados. Todo esto se realiza en el software R utilizando diversos paquetes, entre ellos están el paquete `{haven}` creado por Wickham, Miller, and Smith (2023), `{dplyr}` creado por Wickham, François, et al. (2023), `{tidyverse}` por Wickham (2023), `{sae}` por Molina and Marhuenda (2022), `{rmarkdown}` por Xie (2023), `{ggplot2}` por Wickham, Chang, et al. (2023) y `{sf}`.

Metodología

En este trabajo se propone una metodología de estimación de áreas pequeñas basado en el modelo de área Fay and Herriot (1979), la cual mejora la precisión del estimador directo utilizado en el diseño de muestreo para inferir la verdadera media de área pequeña para cada una de las áreas (comunas) que se quieren analizar, con el objetivo de lograr obtener el índice de pobreza para las comunas de Chile. Además, se desarrolla a través de R una automatización de boletines para reportar estas tasas de pobreza anteriormente calculadas. Estas metodologías y propuestas son implementadas en R y se desarrollan en base al algoritmo descrito a continuación:

- Paso 1: Extracción, Transformación y Carga (ETL) de los datos de CASEN 2017 y la estimación comunal MSD (2015), utilizando principalmente las funciones de los paquetes `{haven}`, `{dplyr}` y `{tidyverse}`.
- Paso 2: Elegir las variables independientes que serán consideradas en el modelo para realizar las estimaciones y fucionar las bases de datos agrupando por comunas, de esta forma se tranforma la data a nivel de unidad a una que se podrá trabajar a nivel de área para el modelo propuesto.
- Paso 3: Implementar la metodología SAE basada en el modelo a nivel de área Fay-Herriot. Para esto se calculan las estimaciones directas y posteriormente se obtiene el estimador EBLUP (Empirical Best Linear Unbiased Prediction) utilizando las funciones `direct` y `eblup` respectivamente, del paquete `{sae}`.
- Paso 4: Se crea un algoritmo para automatizar boletines que reporten las estimaciones obtenidas con la función `render` del paquete `{rmarkdown}`.
- Paso 5: Se crean gráficos o mapas para visualizar e identificar las comunas con su respetiva estimación de tasa de pobreza. Esto con el fin de sea utilizado en los boletines. Se utilizan los paquetes `{ggplot2}` y `{sf}`.

Aplicación

En el estudio se utilizaron datos reales obtenidos de CASEN 2017 de Chile, siendo el objetivo realizar estimaciones de las tasas de pobreza a nivel comunal de Chile mediante el modelo Fay-Herriot. Para esto, se realizó una elección de aquellas variables independientes que son utilizadas como covariables, puesto que, nuestra variable de respuesta es aquella que define a la población en situación o no de pobreza. Para que la data utilizada sea a nivel de área y podamos aplicar el modelo ya mencionado, se debe realizar un ETL en el cual construimos una nueva data que tendrá tantas observaciones como comunas. Puesto que los datos extraídos son a nivel de unidad o de individuo, para hacer esta transformación utilizamos diversos paquetes que nos sirven tanto para importar los datos de la CASEN 2017, hasta para calcular proporciones y simplificar códigos. La estimación de este índice se realizó basado en el modelo a nivel de área Fay-Herriot, del cual se obtiene el estimador EBLUP (Empirical Best Linear Unbiased Prediction) que se define como

$$\theta_i^{\text{EBLUP}} = (1 - \widehat{B}_i)Y_i + \widehat{B}_i x_i^\tau \tilde{\beta} \quad \text{para } i = 1, \ldots, m.$$

donde $\widehat{B}_i$ corresponde a la estimación de $B_i = D_i/A + D_i$ cuando A es remplazado por un estimador $\widehat{A}$. Donde A es la varianza desconocida del efecto aleatorio del área (comuna) y D_i la varianza del error de muestreo (diferente para cada área o en este caso comuna). Además, $\tilde{\beta}$, que corresponde a los parámetros de regresión desconocidos esta dado por

$$\tilde{\beta} = \frac{\sum_{i=1}^m x_i Y_i/(\widehat{A} + D_i)}{\sum_{i=1}^m x_i x_i^\tau/(\widehat{A} + D_i)}$$

donde $x_i = (1, x_{1i}, \ldots, x_{(p-1)i})$ son los valores de $p-1$ covariables para el área o comuna i.

Por otro lado, θ_i^{EBLUP} correspondera a nuestra estimación del índice de pobreza para cada comuna e Y_i corresponde a la estimación directa basada en el diseño de muestreo.

Con este estimador logramos obtener los índices de pobreza para cada comuna de Chile y posterior a esto logramos la automatización de boletines en formato PDF que reporten estos resultados, utilizando como apoyo visual gráficos y mapas. Lo cual se logró utilizando paquetes de R como `{rmarkdown}`, `{ggplot2}`, `{sf}`, entre otros.

Referencias

10 Fay, R. E., and R. A. Herriot. 1979. "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data." *Journal of the American Statistical Association* 74: 269–77.

Molina, I., and Y. Marhuenda. 2022. *sae: Small Area Estimation*. https://CRAN.R-project.org/package=sae.

Wickham, H. 2023. *tidyverse: Easily Install and Load 'Tidyverse' Packages*. https://CRAN.R-project.org/package=tidyverse.

Wickham, H., W. Chang, L. Henry, T. Lin Pedersen, K. Kohske Takahashi, C. Wilke, K. Woo, H. Yutani, and D. Dunnington. 2023. *ggplot2: Elegant Graphics for Data Analysis*. https://CRAN.R-project.org/package=ggplot2.

Wickham, H., R. François, L. Henry, K. Müller, and D. Vaughan. 2023. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, H., E. Miller, and D. Smith. 2023. *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. https://CRAN.R-project.org/package=haven.

Xie, Y. 2023. *rmarkdown: Dynamic Documents for r*. https://CRAN.R-project.org/package=rmarkdown.