TÉCNICAS DE IMPUTACIÓN, UNA APLICACIÓN PARA MEDIR INGRESO

DARÍO PADULA, LETICIA DEBERA

Intendencia de Montevideo, Uruguay.

estadisticas@piso456.imm.gub.uy

RESUMEN

En una investigación es frecuente que individuos encuestados no respondan a una o más preguntas del cuestionario o que no se consiga contactar al individuo seleccionado. Cuando esto ocurre se dice que se tienen datos ausentes o missing y estamos bajo un problema de no-respuesta. La no-respuesta puede introducir sesgo en la estimación e incrementar la varianza muestral debido a la reducción del tamaño de la muestra. En el marco del proyecto "Iniciativa Pobreza y Medio Ambiente" del PNUD y PNUMA, se estudia la población de hogares clasificadores de residuos sólidos urbanos con el objetivo de mostrar la existencia de una relación entre pobreza y medio ambiente. Se encuentra una gran cantidad de datos faltantes en las variables relacionadas con el ingreso, lo cual motiva el estudio de distintas técnicas de imputación de datos. La imputación es un método por medio del cual se completan los valores faltantes, ya sea a una o más variables de interés, por lo general utilizando información auxiliar. A partir de ello, se obtiene un conjunto de datos completo y consistente. En el presente trabajo se imputa la variable *ingresos recibidos por trabajo en relación de dependencia* utilizando distintos métodos de imputación y se comparan los resultados mediante simulaciones.

PALABRAS CLAVE: Datos faltantes, Métodos de Imputación

1 INTRODUCCIÓN

En una investigación es frecuente que individuos encuestados no respondan a una o más preguntas del cuestionario o que no se consiga contactar al individuo seleccionado. Cuando esto ocurre se dice que se tienen datos ausentes o missing y estamos bajo un problema de no-respuesta. La no-respuesta puede introducir sesgo en la estimación e incrementar la varianza muestral debido a la reducción del tamaño muestral.
La imputación de datos corresponde a la etapa final del proceso de depuración de datos, tras el proceso de edición, en el cual los valores missing o que han fallado alguna regla de edición del conjunto de datos son reemplazados por valores aceptables conocidos. La principal razón por la cual se realiza la imputación es obtener un conjunto de datos completo y consistente al cual se pueda aplicar las técnicas de estadística clásicas.
Las razones para utilizar estos procedimientos en el análisis de datos son:

- Reducir el sesgo de las estimaciones (sesgo debido a la no-respuesta).
- Facilitar procesos posteriores de análisis de los datos.
- Facilitar la consistencia de los resultados entre distintos tipos de análisis.
- Mantener la estructura de asociación entre las variables.
- Obtener intervalos de confianza más robustos.

De forma general, existen dos grandes grupos de no-respuesta: aquella asociada a unidades, o sea, registros que tienen todos los campos faltantes y aquella asociada a ítems, es decir, registros que tienen ciertos campos con valor faltante. Los métodos más usados para el tratamiento de la no respuesta asociada a unidades son métodos de calibración (se tratan todas las variables de forma simultánea), mientras que para tratar la no respuesta a los ítems se sugiere la imputación de datos. Esta investigación se centrará en la no-respuesta asociada a ítems.

Dentro de las técnicas más utilizada se encuentra la eliminación de los registros de las variables de interés que presenten algún missing. Sin embargo, esto implica una pérdida de información

considerable sobre todo cuando se cuenta con pocos registros.
La calidad de las imputación dependerá de la disponibilidad de información auxiliar relevante, del mecanismo que genera los datos faltantes y de los métodos utilizados para imputar. Los valores imputados pueden ser tomados de otras unidades o pueden ser producidos a través de un modelo paramétrico ajustado a los datos completos.
Existen dos tipos de métodos de imputación: simple y múltiple. La imputación simple consiste en generar un solo valor y asignar ese valor a la celda vacía, por ejemplo utilizando la media, haciendo una regresión, etc. En cambio, la imputación múltiple consiste en generar m ($m \geq 2$) valores simulados para una única celda obteniendo m matrices completas [5].

2 METODOLOGÍA

Tipos de datos faltantes
Cuando se va a realizar una imputación de datos se debe indagar cuál es el mecanismo que genera los datos faltantes.
Hay tres tipos de mecanismos:

1. *MCAR*, Missing completely at Random (Completamente aleatorio). Cuando la probabilidad de observar missing es completamente al azar.

2. *MAR*, Missing at Random (Aleatorio): La probabilidad de observar missing no depende de la variable pero sí de alguna otra variable observada.

3. *NMAR*, No missing at Random. Se produce este tipo de mecanismo cuando la probabilidad de observar missing depende de la misma variable.

Técnicas de imputación
En esta investigación se describen de forma sucinta algunas de las técnicas de imputación existentes que luego serán utilizadas para comparar los resultados por medio de simulaciones.

Regresión
Consiste en asignar, a los campos a imputar, valores en función del modelo $y_k = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$ donde y_k es la variable dependiente a imputar, x_j con $j = 1, 2, ..., n$

variables independientes. En este tipo de modelos se asume que los datos son MAR o MCAR.

Imputación por la media (MEDIA)
Consiste en un caso particular del método de regresión en el cual se considera solamente una constante en el modelo. De esta manera, se le asigna el valor medio de la variable a todos los valores faltantes de la población. Se asume que la estructura de los datos faltantes es MCAR.

CART(Classification And Regression Tree)
Se parte de un nodo inicial con n observaciones $H = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, donde $x_i \in R^d$. Se realizan sucesivas particiones binarias en el espacio de las variables X, de forma tal que en cada partición se formen grupos homogéneos con respecto a la variable Y, hasta que un criterio de parada es conformado. Cuando la variable Y es categórica se denomina árbol de clasificación y se asigna la moda de las categorías como predicción; mientras que si es continua se denomina árbol de regresión y se asigna la media como predicción. [7],[8]

MOB (Model-based Recursive Partitioning)
Método similar al CART en el sentido de que también se genera un árbol. Las sucesivas particiones en este caso se realizan ajustando un modelo paramétrico al conjunto de datos, luego se busca dentro del set de variables X cuál proporciona la mejor partición. La mejor partición será aquella que genere la mayor "inestabilidad" en las estimaciones de los parámetros del modelo dentro de cada hoja y además reduzca la suma de los residuos al cuadrado. El proceso se repite hasta que se cumpla algún criterio de parada. [3]

Random Forest (RF)
Se generan N árboles de clasificación en forma aleatoria. La aleatoriedad se logra construyendo cada árbol a partir de una muestra boostrap y eligiendo al azar $q < q_0$ variables. [1],[2]

Imputación Múltiple (IM)
Se imputan m valores para cada una de las celdas vacías en la matriz de datos generando

m conjuntos de datos "completos". Los datos faltantes son imputados mediante simulación. Para cada uno de los m conjuntos de datos completos se realiza el análisis que se desee, para luego combinarlos en un sólo resultado. [4],[5],[6]

Imputación Múltiple Adaptada (IMA)
Se realiza una modificación al método de Imputación Múltiple donde se generan n conjuntos de datos completos para luego imputar la media de los n valores. De esta forma se genera una única base de datos.

3 RESULTADOS

Para comparar los distintos métodos de imputación se decide trabajar con la encuesta continua de hogares (ECH) siendo la variable a imputar el *ingreso recibido por trabajo en relación de dependencia*, y las variables explicativas características de los individuos, educación y variables relacionadas al trabajo. Se consideran solamente las personas ocupadas en empleos públicos y privados residentes en Montevideo. De esta población se seleccionan N personas y se elimina, a n de ellas, el registro en la variable que indica el salario. La eliminación de estos registros se realiza de tal forma que representa los distintos mecanismos de generación de datos faltantes. Se realizaron dos simulaciones de la estructura de los datos faltantes:

- MCAR: Se realizan 10 iteraciones para distintos tamaños de muestra (700, 500, 400 y 300) variando el porcentaje de datos faltantes en cada muestra (10%, 15%, 20%, 30%, 40% y 50%)

- NMAR: Se realizan 10 iteraciones para distintos tamaños de muestra (2000, 1800, 1600, 1400, 1200, 1000 y 500) con las siguientes estructuras de correlación (escenarios),

Cuadro 1: Porcentaje de extracción por percentil

	0-20	**20-50**	**50-70**	**70-90**	**90-100**
Escenario 1	5	7	15	20	30
Escenario 2	5	5	5	30	40
Escenario 3	30	20	5	5	5

A modo de ejemplo, se presentan los resultados comparativos para una simulación donde se consideran 500 personas con 21% de missing para una estructura de datos faltantes NMAR. El Gráfico 1 muestra las estimaciones puntuales de la media muestral y los intervalos de confianza al 95% para los distintos métodos junto con la media estimada para la muestra completa ("Verdadero"). Se observa que todas las estimaciones subestiman el verdadero valor. Esto se debe a que el mecanismo utilizado para la generación de missing asigna mayor probabilidad a mayor ingreso. Cabe destacar que las peores estimaciones resultaron las obtenidas por los métodos de eliminación de registro ("MISS") y de imputación por la media que a su vez proporciona un intervalo de confianza más acotado. Por otro lado, la performance de las otras técnicas registran en esta simulación resultados muy parecidos.

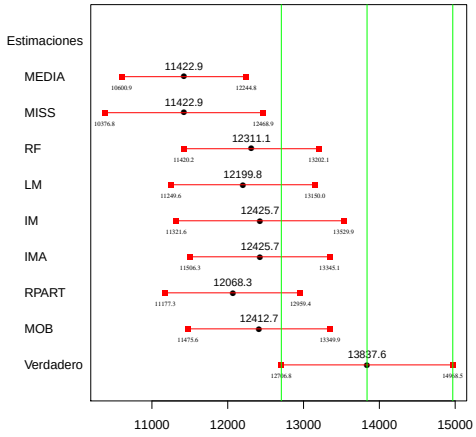

Gráfico 1: Estimaciones de la media muestral e intervalos de confianza al 95% para los distintos métodos

En el Gráfico 2 se muestran los diagramas de caja de la variable *Ipc* de la muestra completa considerando los valores imputados y no imputados para cada uno de los métodos utilizados. Se desea que las distribuciones sean similares a la distribución sin datos faltantes ("Reales"). Por este motivo, se destaca que el método de Random Forest conjuntamente con el de Imputación Múltiple son los que mejor ajustan la distribución original y por el contrario, se destaca el mal ajuste del método de la media.

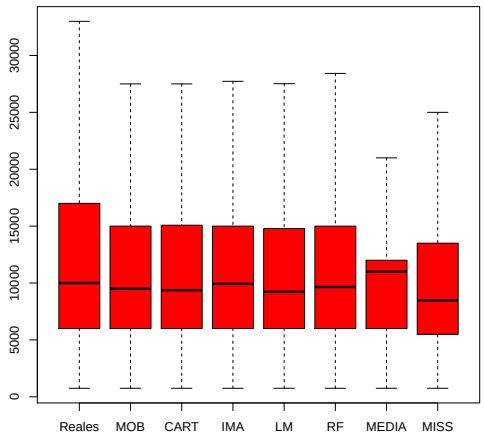

Gráfico 2: Diagramas de caja considerando la muestra completa con los valores imputados

Referencias

[1] G. Biau, L. Devroye y G. Lugosi, *Consistency of random forests and other averaging classifiers*, 2008.

[2] L. Devroye, *Random forests.* Machine Learning 2001.

[3] A. Zeileis, T. Hothorn y K. Hornik, *party with the mob: Model-based Recursive Partitioning in R.*

[4] J. Honaker, G. King y M. Blackwell, *AMELIA II: A Program for Missing Data*, 2010.

[5] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, 1987.

[6] D. B. Rubin, *Multiple Imputation after 18+ years*, 2005.

[7] R. Timofeev, *Classification and Regression Trees, (CART), Theory and Applications*, Master thesis , 2004.

[8] Wei-Yin Loh, *Classification and regression trees*, 2011.