Universidad Santo Tómas

Facultad de Estadística

Propuesta de un Método de Imputación Múltiple para Variables de Tasas y Proporciones

Trabajo de Grado

Autor:
Mario Avila León

Director:
Mario José Pacheco López

Bogotá D.C., Noviembre 2022

Índice

1. Resumen **2**

2. Abstract **2**

3. Antecedentes **3**

4. Introducción **5**

5. Planteamiento del Problema **7**

6. Objetivos **7**
6.1. Objetivo general 7
6.2. Objetivos específicos 7

7. Justificación **8**

8. Hipótesis **8**

9. Marco Teórico **9**
9.1. Conceptos Básicos 10
9.1.1. Mecanismos de Datos Faltantes 10
9.1.2. Métodos de Imputación de Valores Faltantes 11
9.1.3. Paquetes del Software R 12
9.1.4. Distribución Beta 13
9.2. Aproximación Metodológica 13
9.2.1. Amputación por medio de un mecanismo multivariado . . 14
9.2.2. Imputación por regresión 14
9.2.3. Regresión beta 15

10. Metodología **15**

11. Resultados **18**
11.1. Aplicación para una Base de Datos Completa 18
11.1.1. Información Descriptiva 18
11.1.2. Método de Imputación Regresión Lineal 21
11.1.3. Método de Imputación Regresión Beta 22
11.2. Aplicación para una Base de Datos Incompleta 23
11.2.1. Información Descriptiva 24
11.2.2. Método de Imputación Regresión Lineal 26
11.2.3. Método de Imputación Regresión Beta 26
11.3. Anexos 29

12. Conclusiones **30**

1. Resumen

El análisis de datos faltantes esta orientado a la investigación y aplicación de métodos estadísticos, que permiten trabajar conjuntos de datos, con información faltante. Mediante procesos de estimación estadística, los valores faltantes en una base de datos (o muestra), pueden reemplazarse con valores obtenidos en dicho proceso; y lograr una base de datos con información completa. Algo ideal para un investigador; disponer de la data completa es una garantía para considerar toda la información de la muestra.

Por este motivo, obtener una buena estimación de los valores faltantes, permite mejorar los análisis; por el contrario, reemplazar los valores ausentes con sesgos en su estimación puede generar valoraciones inadecuadas en un análisis.

Los métodos y procesos para la imputación de datos faltantes, mejoran su potencia según el comportamiento de los datos y la información observada. Por tanto, los métodos de imputación pueden aplicarse a variables que cumplan algunas condiciones para asegurar una adecuada estimación. Este trabajo, propone estimar los valores faltantes para variables de proporciones o tasas, con un modelo de regresión Beta.

2. Abstract

Missing data analysis is oriented to the investigation and application of statistical methods, that allow working with data sets, with missing information. By means of statistical estimation processes, the missing values in a database (or sample), can be replaced with values obtained in this process; and achieve a database with complete information. Something ideal for a researcher; having the complete data is a guarantee to consider all the information of the sample.

For this reason, obtaining a good estimate of the missing values, allows to improve the analysis; on the contrary, replacing the missing values with biases in their estimation can generate inadequate valuations in an analysis.

The methods and processes for missing data, imputation improve their power according to the behavior of the data and the information observed. Therefore, imputation methods can be applied to variables that meet some conditions to ensure an adequate estimation. This paper, proposes to estimate missing values for ratio or rate variables, with a Beta regression model.

3. Antecedentes

El análisis estadísto de datos faltantes; se emplea por la necesidad de desarrollar métodos estadísticos, que puedan aplicarse a conjuntos de datos en los cuales, un determinado porcentaje de la muestra presenta ausencia de información en una o todas sus variables. En ocasiones, la información disponible o que se logra recolectar, puede no estar en su totalidad; para solucionar estos casos donde hay información ausente, se utilizan métodos de imputación de datos.

Los primeros estudios para el análisis de datos faltantes, se dieron entre los años 1920 a 1930; desde entonces ya se planteaban soluciones algorítmicas (Geert Molenberghs, 2007). De acuerdo al autor Craig K. Enders, en su libro *Applied Missing Data Analysis*: en la década de los 70', surgieron los avances más significativos en el campo de análisis de datos faltantes y posteriormente se han desarrolado métodos más sofisticados en los procesos de imputación, por supuesto sacando provecho de la evolución y capacidad computacional para el procesamiento de datos.

En la actualidad existen diferentes métodos estadísticos para la imptutación de datos faltantes, y cabe resaltar que es un área de conocimiento que continua ampliando su investigación. Los procesos estadísticos para la imputación de datos; de manera general se pueden distiguir en: Métodos de Imputación Simple y Métodos de Imputación Múltiple (Juan Francisco Muñoz Rojas, 2009), donde la aplicación de cada método dependerá de la estructura de los datos.

Por ejemplo, en una investigación publicada por la Universidad del Rosario de Argentina; emplearon diferentes métodos de imputación para la Encuesta Nacional de Gasto de los Hogores, en particular comprobaron los resultados de imputación para la variable 'Ingreso'. (Mitas, 2018).

Los métodos utilizados y contrastados fueron:

- missForest
- Random Hot Deck
- Vecino más cercano
- Algoritmo EM (Expectation - Maximization)
- Amelia
- MICE

Dentro de los hallazgos del estudio, concluyen que los métodos más precisos fueron missForest y *MICE*, los cuales arrojaron menores coeficientes de Error Cuadrático Medio Normalizado (Mitas, 2018). Sin embargo, el estudio es realizado bajo variables que tienen un comportamiento normal y para los cuales

el método MICE es eficiente. En el mismo, plantean que al tratarse de variables que no cumplen los criterios de normalidad, estos métodos de imputación pueden presentar deficiencias en las estimaciones.

El presente trabajo identifica la posibilidad de explorar las técnicas existentes para los procesos de imputación y basarse en los métodos que tienen aplicabilidad a datos de tipo proporción o tasa, con el fin de proponer un proceso de imputación con un modelo que se adecue a este tipo de variables.

4. Introducción

En la práctica, es habitual encontrarse con información faltante dentro de una base de datos, lo cual puede dificultar y hasta limitar el análisis de la información. Son diversas situaciones que pueden causar la pérdida de información; como una mala codificación o falta de respuestas en un cuestionario (Juan Francisco Muñoz Rojas, 2009). Suponiendo que una muestra contiene datos faltantes y se espera realizar algún tipo de análisis o inferencia; será necesario plantear una manera de trabajar con la información disponible y tal vez, utilizar métodos estadísticos y computacionales, para completar la información de la muestra; estimando de una manera eficiente y eficaz aquellos datos que no están disponibles, como también puede ser una alternativa; eliminar individuos sin ninguna respuesta dentro del cuestionario (Enders, 2010).

Existen diversas técnicas en el análisis estadísticio de datos faltantes, que integran procesos con la capacidad de reemplazar en una muestra o base de datos los valores ausentes. El proceso de imputación de datos tiene impacto en la consistencia de los resultados, o en posibles inferencias y estimaciones para análisis posteriores. Por tanto, cobra relevancia identificar el tipo de variable(s) que se pueda tener en un conjunto de datos y a su vez, conocer la naturaleza y comportamiento de los mismos. También, para el análisis de información faltante, es importante comprender el mecanismo de los datos ausentes; todo esto para lograr una imputación adecuada de la información.(Geert Molenberghs, 2007)

De los diversos conjuntos de datos que pueden presentar información faltante, las variables cuantitativas en particular, cuentan con métodos de imputación eficientes; donde los datos ausentes son reemplazados por valores sustitos, obtenidos a través de algoritmos de simulación. Estos procesos por lo general, se calculan a partir de las variables observadas en una base de datos. Los modelos de regresión por ejemplo, son utilizados para procesos de imputación de datos; estos métodos implican que el modelo aplicado para estimar los valores faltantes; deben tener concordancia con el comportamiento de la variable a imputar, dado que no utilizar un modelo adecuado puede ocasionar errores de estimación, y posiblemente llegar a tener un valor sustituto por fuera del rango de la variable (Fernando Medina, 2007).

Las variables de proporción son sensibles a este error de estimación, por la restricción del rango de valores que puede tomar $(0, 1)$. En este contexto, es interesante plantear un proceso de imputación adecuado, como puede ser un modelo de regresión lineal Beta que permita estimar variables de intervalo $(0, 1)$.

El presente trabajo, tiene como objetivo proponer un proceso de imputación de datos faltantes, empleando un modelo de regresión Beta; que permita estimar valores faltantes de una variable de tasa o proporción, con el fin de solucionar los errores de sobreestimación o subestimación de los datos imputados. Aplicando

el proceso de imputación bajo el modelo de regresión Beta, se espera obtener un método apropiado para variables que únicamente tomen valores en el intervalo $(0, 1)$; y por otra parte, se busca comparar los resultados obtenidos a partir de la imputación con el modelo de regresión propuesto, frente a los a métodos estadísticos de la función MICE, disponible en el software R.

5. Planteamiento del Problema

La observación de datos en la estadística es fundamental para realizar diferentes procesos y obtener resultados satisfactorios; que puedan ser la base para el análisis y posterior toma de decisiones. La calidad de los datos y un muestreo adecuado, garantizan que la inferencia de la información sea una aproximación buena a las características de la población. (Villegas, 2005).

Sin embargo, en muchos estudios no es fácil obtener toda la información esperada de la muestra y es imperativo aplicar métodos estadísticos que permitan realizar inferencia de la información, sin sobrestimar el comportamiento de las variables (Moisés, 2014). Una de las dificultades al trabajar con muestras o bases de datos, es la información faltante en distintas variables; que pueden ser de naturaleza cuantitativa y/o cualitativa.

Ahora bien, a pesar de contar con métodos de imputación para diferentes tipos de variables; al momento de estimar los valores ausentes, los resultados pueden arrojar un valor sustituto que no sea adecuado o lógico en la interpretación de una variable en particular. Un caso puntual, son las variables de intervalo $(0, 1)$, que dada su naturaleza; es posible que en un proceso de imputación por medio de métodos de regresión o procesos de simulación, los valores estimados estén fuera del rango de la variable.

Por lo tanto, una posible solución para variables que toman valores entre 0 y 1, es trabajar con un modelo de regresión Beta para la imputación de los datos faltantes; que garantice una estimación conforme al comportamiento de la variable.

6. Objetivos

6.1. Objetivo general

- Plantear un método de imputación para variables de proporción y tasas.

6.2. Objetivos específicos

- Proponer un método de imputación múltiple basado en modelos de regresión Beta.
- Realizar aplicaciones sobre bases de datos reales, que presenten ausencia de información en variables de tasas o proporciones.

7. Justificación

En el planteamiento del problema se identificó la dificultad para imputar variables de tasas y/o proporciones, dadas las caracterśticas particulares de los datos; y a pesar de contar con métodos de imputación eficientes para cierto tipo de variables, como lo señalan Juan Muñoz y Encarnación Álvarez en el artículo *'Métodos de imputación para el tratamiento de datos faltantes'*; no existe un método estadístico específico para variables con un intervalo de $(0, 1)$.

Por tanto, será preciso plantear para variables de tasas o proporciones, un método de imputación adecuado para el comportamiento de los datos y contrastar los resultados obtenidos con las imputaciones del algoritmo *MICE* y bajo la propuesta de un modelo de regresión lineal Beta.

8. Hipótesis

El proceso de imputación por medio de la regresión lineal Beta, es adecuado y permite su aplicación a variables de tasas y/o proporciones.

9. Marco Teórico

El análisis de datos faltantes, cuenta con diferentes autores e investigaciones que proporcionan metodologías estadísticas; para trabajar de manera eficaz con información faltante.

El autor Stef van Buuren, en la segunda edición del libro *'Flexible Imputation of Missing Data'*, detalla los mecanismos de valores faltantes, donde los datos tienen cierta probabilidad de estar ausentes; lo cual debe considerarse en los procesos de imputación, como lo son: *Datos faltantes completamente aleatorios (MCAR por sus siglas en inglés), Datos faltantes aleatorios (MAR) y Datos faltantes no aleatorios (MNAR).*

Posterior a idenitificar la naturaleza de los valores faltantes; se puede ejecutar los procesos de imputación. En el libro *'Statistical Analysis with Missing Data'*, sus autores Rubin y Little; sustentan diversas formas para la imputación de datos faltantes; describiendo métodos de imputación simple, hasta estimaciones por medio de modelos robustos; que serán útiles para aplicar la imputación de datos en el conjunto de datos del presente trabajo.

A su vez, Craig K. Enders en 2010 publicó el libro *'Applied Missing Data Analysis'*, en el cual se encuentran consignados los métodos tradicionales para la imputación de datos y propone las estimaciones de máxima verosimilitud para valores faltantes. Adicional, CEPAL en 2007 publicó un trabajo de investigación de Fernando Molina y Marco Galván titulado: *'Imputación de Datos: Teoría y Práctica'*; el cual compila estudios sobre la distribución de los datos faltantes, como también métodos de imputación simple y múltiple.

En esta misma línea y para ampliar los textos de consulta sobre análisis de datos faltantes; artículos y trabajos de grado preliminares serán herramientas de apoyo. Por mecionar alguno, el proyecto de grado presentado por Leandro Sánchez en enero 2020 a la Universidad Santo Tómas, donde fue aplicado un método de imputación por interpolación.

Por otra parte, la obtención de las bases de datos para el desarrollo del trabajo serán consultadas en la Departamento Administrativo Nacional de Estadística (DANE). También los resultados de las pruebas Saber, publicadas por el Ministerio de Educación; es una posible fuente para obtener las bases de datos. Para complementar con información macroeconómica, otra fuente de consulta para la base de datos es el Banco Mundial.

Por último, en respuesta a la hipótesis planteada en el trabajo; es preciso contar con información referente a la distribución beta y la aplicación del modelo para datos con esta naturaleza. En 2014, González C., Galvis S. y Hurtado T.; publicaron un artículo donde hacen la aplicación de un modelo Beta generalizado; el cual contiene la teoría del modelo. De igual manera, el libro *'Estadística*

Matemática con Aplicaciones'; soporta la información teórica de la distribución Beta.

9.1. Conceptos Básicos

Los datos faltantes se definen como valores no disponibles dentro de un conjunto de datos, en términos generales; datos o valores ausentes. Estos datos no observados en un análisis inferencial, pueden ser determinantes para hallar estimadores insesgados. (Dagnino, 2014)

Al observar una base de datos en particular con valores faltantes, es posible identificar la estructura de los valores observados y ausentes. Esto se define como patrones de datos faltantes (Enders, 2010)

La siguiente tabla representa un conjunto de datos, con presencia de datos faltantes:

	Va1	Va2	Va3
1	✓	✓	
2	✓	✓	✓
3	✓	✓	✓
4		✓	
5	✓	✓	

Cuadro 1: Representación de datos falantes

Donde, se obtienen los siguientes patrones de datos faltantes:

	Va1	Va2	Va3
Patrón 1	✓	✓	
Patrón 2		✓	
Patrón 3	✓	✓	✓

Cuadro 2: Patrón de datos faltantes resultante

9.1.1. Mecanismos de Datos Faltantes

Los mecanismos de datos faltantes, explican la relación que existe entre los datos observados y los valores ausentes (Roderick J. A. Little, 2020), dicha relación se puede definir de la siguiente manera:

Teniendo que $\mathbf{Y}$ es una matriz con y_{ij} datos observados o faltantes; los valores observados se denotan como Y_{obs} y los ausentes como Y_{miss}. Asimismo, se define $\mathbf{M}$ como una matriz indicadora así: cuando $(m_{ij}) = 1$; el dato y_{ij} es un valor faltante y si $m_{ij} = 0$ el valor es observado. Ahora, considerando que las filas y_i y m_i son independientes e idénticamente distribuidas, entonces el mecanismo de ausencia de datos se caracteriza por la distribución condicional m_i dado y_i (Rianne Margaretha Schouten, 2021):

$$f_{M|Y}(m_i|y_i, \phi), \quad \textit{donde } \phi, \textit{ es un parametro desconocido}$$

Rubin (1976) clasificó los modelos de pérdida de datos en tres tipos diferentes: MCAR, MAR y MNAR (Moisés, 2014):

- **MCAR** (missing completely at random): La probabilidad que los datos faltantes sean completamente aleatorios, no depende de los valores de las variables observadas y/o ausentes:

$$Pr(M = 1 \mid Y_{obs}, Y_{miss}.\phi) = Pr(M = 1 \mid \phi)$$

- **MAR** (missing at random): La probabilidad que los datos faltantes sean aleatorios, depende de los valores observados en la variable:

$$Pr(M = 1 \mid Y_{obs}, Y_{miss}.\phi) = Pr(M = 1 \mid Y_{obs}, \phi)$$

- **MNAR** (missing not at random): La probabilidad que los datos faltantes no sean aleatorios, es dependiente de los valores ausentes en las variables:

$$Pr(M = 1 \mid Y_{obs}, Y_{miss}.\phi) = Pr(M = 1 \mid Y_{obs}, Y_{miss}, \phi)$$

9.1.2. Métodos de Imputación de Valores Faltantes

La imputación es el proceso de reemplazar los datos faltantes con valores sustitutos. A continuación se presentan algunos métodos tradicionales de imputación (Moisés, 2014):

1. **Análisis de Casos Completos**: Son descartados todos los individuos que presentan uno o más valores faltantes. Este método puede generar una reducción significativa de la muestra y si la pérdida de datos no es MCAR es posible generar estimaciones sesgadas.

2. **Análisis de Casos Disponibles**: Intenta reducir la pérdida de información realizando una estimación con los valores obsevardos variable a variable, donde emplea un subconjunto de datos observados para realizar estimaciones de una variables particular.

3. **Imputación por Media Incondicional**: Los datos faltantes son sustituidos por la media muestral calculada de los datos observados de la variable.

4. **Imputación por Media Condicional (Regresión)**: Los valores faltantes de cada variable son estimados utilizando modelos de regresión, en donde el valor faltante es la variable de respuesta y los demás datos observados de las variables son las explicativas del modelo.

5. **Imputación por Regresión Estocástica**: Emplea modelos de regresión para estimar los valores faltantes de las variables, utilizando los datos observados de las demás variables como explicativas. A diferencia del método anterior, en la regresión estocástica, para cada ecuación de regresión se suma un componente residual aleatorio $\epsilon \sim (0, {\sigma^2}_j)$.

6. **Imputación Hot-Deck**: Utiliza los valores observados de individuos con características similares, para reemplazar los datos faltantes de las variables.

7. **Última Observación Realizada**: Reemplaza un valor faltante con la observación anterior, este método se emplea para datos longitudinales.

8. **Imputación Múltiple**: Se basa en realizar un proceso iterativo de imputaciones de valores faltantes; obteniendo diferentes conjuntos de datos completados, posteriormente los resultados se analizan para la estimación final.

9.1.3. Paquetes del Software R

Algunos paquetes de R, con funciones para el tratamiento de bases de datos, con valores ausentes: (*i*) *MissingDataGUI*, (*ii*) *Amelia II*, (*iii*) *VIM*, (*iv*) *MICE* (Moisés, 2014).

Algoritmo MICE

La función *MICE* (imputación multivariada por ecuaciones encadenandas, por sus siglas en inglés) es utilizada para realizar procesos múltiples de imputación de datos faltantes de naturaleza MAR o MNAR. El algoritmo *MICE* trabaja de forma iterativa con el fin de hacer progresivamente mejores estimaciones de los datos faltantes (Stef Van Buuren, 2011). Además, es una de las funciones más potentes para el procesamiento de datos faltantes.

El algoritmo es capaz de realizar procesos iterativos con modelos de regresión; para reemplazar datos de las variables con valores ausentes. El algoritmo funciona así:

1. Cada dato faltante y_{ij} es reemplazado temporalmente, por la media calculada de los datos observados en la muestra.

2. Luego, para la variable incompleta Y_j, se estima un modelo de regresión con los datos observados, donde Y_j es la variable de respuesta y las demás variables son explicativas del modelo. Entonces los datos imputados en el paso 1 se eliminan y son reemplazados por los valores estimados con el modelo.

3. De manera iterativa, cada variable faltante Y_{j+r} (con $r = 1, 2, 3, ...$) será la variable de respuesta y las demás (observadas o imputadas) serán las

explicativas, obteniendo la estimación del dato faltante para todas las variables.

4. Los pasos 2 y 3 se repiten para cada variable incompleta, hasta completar el proceso con todas las variables y lograr una iteración ($m = 1$). El proceso genera m conjuntos de datos imputados.

(Moisés, 2014).

Adicional, *MICE* contiene en el paquete R una serie de métodos de imputación (argumento *method*), que pueden ser aplicados de acuerdo al tipo de variable que se va a imputar.

9.1.4. Distribución Beta

La distribución Beta es adecuada para modelar datos con un intervalo soporte de las variables aleatorias entre $(0, 1)$ (Dennis D. Wackerly, 2010). La función de densidad Beta está dado por (Silvia L.P. Ferrari, 2004):

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}, \quad donde \;\; 0 < y < 1, \quad a, b > 0 \qquad (1)$$

9.2. Aproximación Metodológica

El análisis estadístico se centra en la observación de datos; contar con toda la información de una población o muestra, permite obtener resultados consistentes; por ejemplo, en estudios clínicos donde las muestras pueden ser limitadas, es importante tener la información completa para cada uno los pacientes (Geert Molenberghs, 2007). De acuerdo al estudio publicado en 2007 por Medina y Galván, en la Comisión Económica para América Latina y el Caribe (CEPAL); la ausencia de información en bases de datos es algo recurrente, para estos casos es necesario plantear la manera adecuada de procesar los datos (Fernando Medina, 2007).

En estadística una forma de trabajar con información faltante, son los métodos de imputación (reemplazar un dato ausente); dicho proceso garantiza que el posterior análisis considere a todos los individuos de la muestra, en donde los valores imputados son estimados de acuerdo a las características de la pobalción y variable de estudio (Mitas, 2018).

El trabajo está enfocado a aplicar un método de imputación, para una base de datos con variables de rango $(0, 1)$ principalmente. Por tal razón, el conjunto de datos objeto de estudio, debe tener valores faltantes y contar con variables de tasa o proporción, entre otras. Es posible trabajar con una base con información completa, en este caso debe realizarse un proceso de simulación para generar valores faltantes.

9.2.1. Amputación por medio de un mecanismo multivariado

La función *ampute*, realiza un proceso de amputación multivariado para generar valores faltantes en una base de datos. La función permite definir uno o varios patrones de datos faltantes; por medio de una matriz indicadora **M** que toma valores de 0 cuando el valor es ausente y 1 cuando se espera que el valor de la variable siga siendo observado; a su vez es posible ingresar un vector con la longitud esperada por cada patrón de datos faltantes. Además, permite que el usuario pueda indicar la proporción de datos faltantes (G. V. Rianne Margaretha Schouten, 2018).

Adicional a otras condiciones que integra la función, es posible especificar el mecanismo para generar datos faltantes: *MCAR, MAR o MNAR*; por medio de una combinación lineal de las variables Y_i y unos pesos ponderados w_i definidos por el investigador. Los pesos asignados a las variables determinan la probabilidad de ausencia para cada dato y_{ij} en la variable Y_j, así (G. V. Rianne Margaretha Schouten, 2018):

$$ws_i = w_1 y_{1i} + w_2 y_{2i} + ... + w_r y_{ri}$$

Cuando los pesos $w_1, w_2, ..., w_r$ toman valores de 0, el mecanismo de datos faltantes es *MCAR*. Para el mecanismo *MAR*, las variables que serán amputadas toma un valor de $w_i = 0$, dado que la probabilidad de ausencia depende de las variables observadas. Por tanto, al asignar $w_i \neq 0$ el mencanismo de datos faltantes es *MNAR*.

9.2.2. Imputación por regresión

El método de imputación por regresión lineal, establece para cada uno de los patrones faltantes un modelo de regresión múltiple, para estimar el dato faltante y_{ij}. Entonces, cada valor ausente será la variable de respuesta y se ajusta una ecuación de regresión donde las variables observadas son las explicativas del modelo; con el fin de predecir y reemplazar el dato y_{ij}, como una media condicionada de las covariables (Fernando Medina, 2007).

Por ejemplo, la siguiente ecuación, representa el modelo de regresión lineal para el patrón 1 de datos faltantes en el cuadro 2:

$$\hat{Y}_{var3} = \hat{\beta}_0 + \hat{\beta}_1 Y_{var1} + \hat{\beta}_2 Y_{var2} \tag{2}$$

La imputación por regresión estocástica, trabaja con las ecuaciones de regresión y este método integra un término residual aleatorio $u \sim N(0, \sigma_j^2)$:

$$\hat{Y}_{var3} = \hat{\beta}_0 + \hat{\beta}_1 Y_{var1} + \hat{\beta}_2 Y_{var2} + u \tag{3}$$

9.2.3. Regresión beta

El modelo de regresión Beta, es utilizado para modelar variables continuas que asumen valores en el intervalo $(0, 1)$. Este modelo se basa en una parametrización de la distribución Beta en términos de la media y el parámetro de precisión (Francisco Cribari-Neto, 2010).

Donde, para la ecuación 1; Ferrari y Cribari Neto en su artículo publicado en 2004; propone realizar la siguiente parmetrización: $\mu = a/(a+b)$ y $\phi = a+b$, entonces la ecuación resultante es:

$$f(x) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad donde \ 0 < y < 1, \quad 0 < \mu < 1, \quad \phi > 0 \tag{4}$$

Por tanto, $y \sim B(\mu, \phi)$, con $E(y) = \mu$ y $Var(y) = \frac{\mu(1-\mu)}{1+\phi}$ y ϕ es el parámetro de precisión (Francisco Cribari-Neto, 2010).

Ahora, sea $y_1, y_2, ..., y_n$ una muestra aleatoria para la cual $y_i \sim B(\mu, \phi)$, con $i = 1, 2, ..., n$; el modelo de regresión beta está definido por:

$$g(\mu_i) = x_i^\top \beta = \eta_i \tag{5}$$

Donde, $\beta = (\beta_1, ..., \beta_k)^\top$ es un vector de parámetros de regresión desconocidos k x 1, con $k < n$; además $x_i = (x_{i1}, ..., x_{ik})^\top$ es un vector de k coeficientes de regresión; $\eta_i = \beta_1 x_{i1} + ... + \beta_k x_{ik}$ es un predictor lineal y $g(\cdot)$ es la función de enlace que por defecto es la función logística (Francisco Cribari-Neto, 2010).

10. Metodología

Aplicar un método de imputación de datos, implica conocer el tipo de variables con información faltante, un posible escenario es tener un conjunto de datos de variables cuantitativas y cualitativas; aquí no es posible aplicar un único método para imputar los valores de todas las variables. Por tanto, los métodos para imputar deben estar en concordancia con el comportamiento de la variable y así lograr una imputación adecuada (G. V. Rianne Margaretha Schouten, 2018). Las variables de tasas y proporciones tienen una restricción de rango $(0, 1)$, por tanto es posible que algunos procesos de imputación, arrojen estimaciones fuera del intervalo mencionado.

Una manera eficiente para evaluar los procesos de imputación, es trabajar con una base de datos completa; y por medio de de simulación generar valores faltantes, luego a partir de la base de datos con información faltante aplicar algún método de imputación. Este proceso permitirá comparar los resultados estimados frente a la base de datos completa (G. V. Rianne Margaretha Schouten, 2018). El software R integra diferentes algoritmos para la imputación de datos y además, es posible realizar un proceso de amputación por medio del

paquete *MICE* con la función *ampute*. Esta función es conveniente para el desarrollo del presente trabajo, dado que será utilizado para la base de datos objeto de estiudio.

Imputación con Modelo de Regresión Lineal

La imputación por regresión bajo la función *MICE*, realiza un proceso iterativo; haciendo una primer estimación para los datos ausentes con la media muestral; calculada con los valores observados de la variable. Para la segunda iteración, es aplicada una regresión lineal, donde para cada columna Y_j se elimina el dato generado en el primer paso y es estimado a partir de los datos observados en las otras columnas; cada columna incompleta Y_j será la variable dependiente y las demás columnas son las predictoras del modelo. Entonces, el proceso iterativo continua estimando los valores faltantes utilizando los valores imputados del paso inmediatamente anterior. La función *MICE* en R, permite definir la cantidad de iteraciones con el argumento *maxit* y también es posible generar m conjuntos de datos imputados (Stef Van Buuren, 2011).

Imputación Propuesta con Modelo de Regresión Beta

El objetivo del trabajo es proponer un método de imputación adecuado para las variables de tasas y proporciones. Utilizando el software R, se plantea una función para imputar datos con la siguiente metodología:

El usuario debe ingresar en la función; la base de datos con información faltante, el número de iteraciones que espera del proceso de imputación y las variables de intervalo $(0, 1)$ (definidas con un valor o vector que indica las columnas de variable que cumplen con el criterio mencionado).

1. La función en primer instancia realiza una imputación para toda la base de datos, empleando el algoritmo *MICE*, bajo el método *sample*; que reemplaza los valores faltantes de forma aleatoria de acuerdo a los valores observados. Obteniendo una base de datos sin valores faltantes.

   Adicional, es definida la matriz indicadora de datos faltantes ($\mathbf{M}$), que será útil para generar en los pasos 2 y 3 los valores faltantes de la base de datos ingresada.

2. Con los resultados del paso 1 -y para variables que no sean de tasa o proporción-, se calcula una imputación por medio del método *norm.predict* del paquete *MICE*; este método reempalza los datos faltantes, a partir de las estimaciones por medio de un modelo lineal.

3. Finalmente, con los resultados obtenidos en el paso anterior, se aplica un proceso de imputación con el modelo lineal Beta -únicamente para variables de rango $(0, 1)$-. Entonces, se ejecuta un proceso iterativo que reemplaza los valores ausentes de cada variable, mediante la regresión

lineal Beta. El modelo estima los valores faltantes para cada una de las variables, así:

Con la matriz $\mathbf{M}$ se generan nuevamente los datos faltantes originales para la variable de respuesta ($\mathbf{Y_i}$) y las demás variables serán las explicativas ($\mathbf{X_{ij}}$). Los valores estimados de $\mathbf{Y_i}$ son almacenados; logrando una nueva base de datos completa, con valores imputados por el modelo Beta.

El modelo de regresión Beta es calculado con la función *betareg*.

11. Resultados

11.1. Aplicación para una Base de Datos Completa

El Ministerio de Educación por medio del portal datos abiertos, en julio de 2022 actualizó la base de datos; que contiene los principales indicadores para la educación preescolar, básica y media; la cual recoge información del año 2011 a 2021 (este último es un dato preliminar). Esta base de datos cuenta con 322[1] registros históricos por departamento y hay un total de 37 variables.

11.1.1. Información Descriptiva

La información contenida en la base, inicia con variables descriptivas que son: el año de la información, el departamento y el tamaño de la pobalción. Luego se encuentran indicadores con mediciones en porcentajes y tasas. Esta información cuantitativa cumple la característica fundamental para el prósito del trabajo; en cuánto a proponer un método de imputación para datos correspondientes a tasas y proporciones. Las variables utilizadas en el desarrollo del trabajo corresponden a los indicadores totalizados de preescolar, básica y media, por cada departamento: (i) tasa de matriculación, (ii) cobertura de educación neta, (iii) la tasa de deserción, (iv) tasa de aprobación estudiantil, (v) tasa de reprobación y (vi) la tasa de repitencia.

La siguiente tabla muestra estadísticas generales de la base de datos completa:

	Mínimo	Máximo	Promedios	Desv. Estándar
Tasa matriculación	0.523	0.994	0.842	0.102
Cobertura	0.508	0.993	0.848	0.104
Deserción	0.012	0.91	0.046	0.065
Aprobación	0.739	0.991	0.909	0.047
Reprobación	0.001	0.99	0.102	0.166
Repitencia	0.01	0.98	0.103	0.223

Cuadro 3: Datos de la base completa

Ahora, la gráfica de cajas, da una visualización del comportamiento de las variables de la base de datos:

[1] Los datos del año 2021 no son considerados.

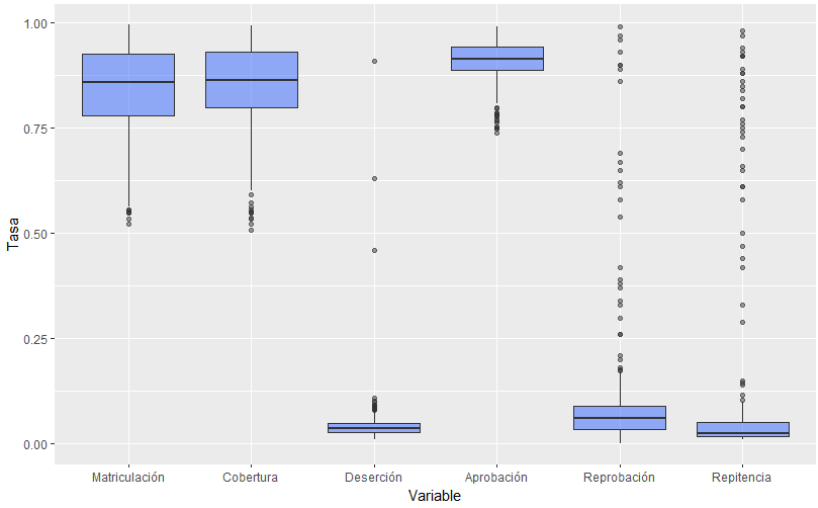

Figura 1: Boxplot

A partir del gráfico se observa que los porcentajes de la Deserción escolar, presentan la menor variabilidad, sus tasas están principalmente agrupadas entorno a la media de 4.5 %; para esta variable se evidencia fácilmente 3 valores atípicos (46, 63, 91 %). Asimismo, la desviación estándar de las variables Aprobación es estadísticamente baja, las dos últimas variables (Reprobación escolar y Repitencia) presentan la mayor cantidad de datos atípicos y su media es cercana al 10 %. Otras variables que tienen un comportamiento similiar entorno a su media es la Tasa de Matriculación y Cobertura; que a pesar de tener presencia de datos atípicos no están alejados del límite inferior.

En los años 2015 a 2020, los departamentos de Guaviare, Vaupés y Vichada registran las menores tasas de matriculación y cobertura en educación, tasas entre 50 *y* 55 %; mientras que los departamentos de la región caribe y central del país principalmente, reportan las tasas más altas (por encima del 95 %) destacando al departamento de Tolima con tasas superiores al 99 % en el 2020. En el mismo año, Bogotá registró la tasa de desereción estudiantil más baja historícamente; mientras que para 2020 los departamentos con mayor Repitencia son Vichada (14,4 %), Guainá (10,5 %) y Amazonas (10 %).

Amputación de datos

Las variables mencionadas no presentan ausencia de datos, por tanto por medio del paquete *MICE* del software R; se utiliza la función *ampute*, para generar valores faltantes en la base de datos. Este proceso de amputación es un método multivariado; donde se genera una proporción de datos faltantes de 30, 40 *y* 50 % con el mecanismo de valores faltantes *MAR*, para garantizar que la probabilidad de ausencia está dada por los datos observados. El proceso de

amputación, simula los siguientes datos faltantes:

Variable	Datos amputados, según %					
	No. NAs (30 %)	% NAs (30 %)	No. NAs (40 %)	% NAs (40 %)	No. NAs (50 %)	% NAs (50 %)
Matriculación	16	4.97	24	7.45	25	7.76
Cobertura	16	4.97	21	6.52	18	5.6
Deserción	11	3.42	21	6.52	23	7.14
Aprobación	18	5.6	23	7.14	36	11.18
Reprobación	16	4.97	12	3.73	25	7.76
Repitencia	20	6.21	22	6.83	32	9.94

Cuadro 4: Datos base amputada

Patrones de datos faltantes

Luego del proceso de amputación, para cada variable se generaron valores faltantes, como resultado del proceso de amputación, en la siguiente imagen, se identifican los patrones de datos faltantes:

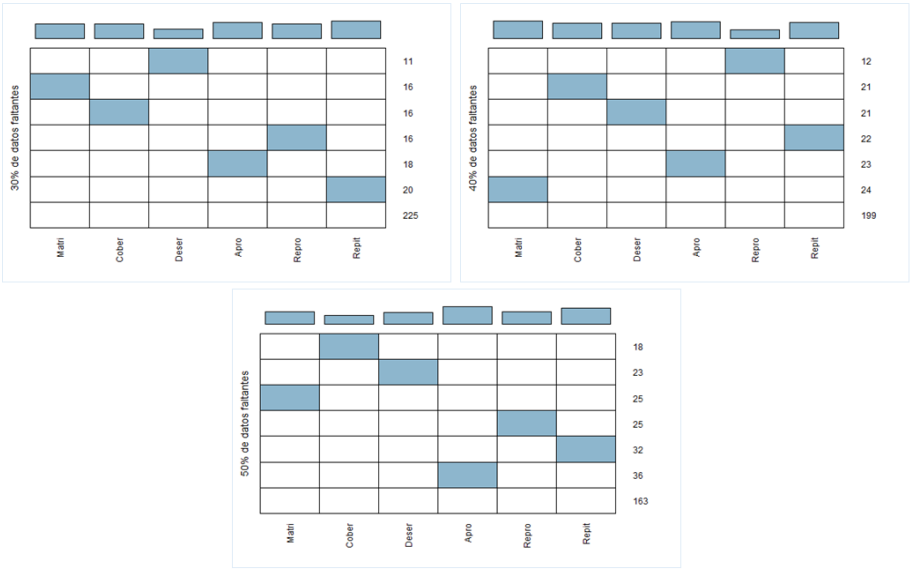

Figura 2: Patrones de datos faltantes

En total hay 7 patrones de datos faltantes. Los números de la derecha representan la cantidad de datos ausentes por cada patrón, el color azul indica que para esa variable el valor es faltante. En total, se generaron 97, 123 y 159 datos

faltantes en el proceso de amputación del 30, 40 y 50 %, respectivamente y El patrón en blanco (último renglón de las gráficas); corresponde a la cantidad de registros que cuentan con información en todas sus variables.

11.1.2. Método de Imputación Regresión Lineal

Inicialmente el método aplicado para reemplazar los valores faltantes, es a paritr de la imputación por regresión lineal predictiva. Por tanto, dentro de la función *MICE*, se define el método de imputación *norm.predict*. Los datos imputados para la variable *Repitencia*, se pueden visualizar en rojo en la siguiente gráfica:

Nota: Repitencia estudiantil, presenta la mayor variabilidad en su información; por esta razón los resultados que se muestran en el trabajo. se basarán en esta variable.

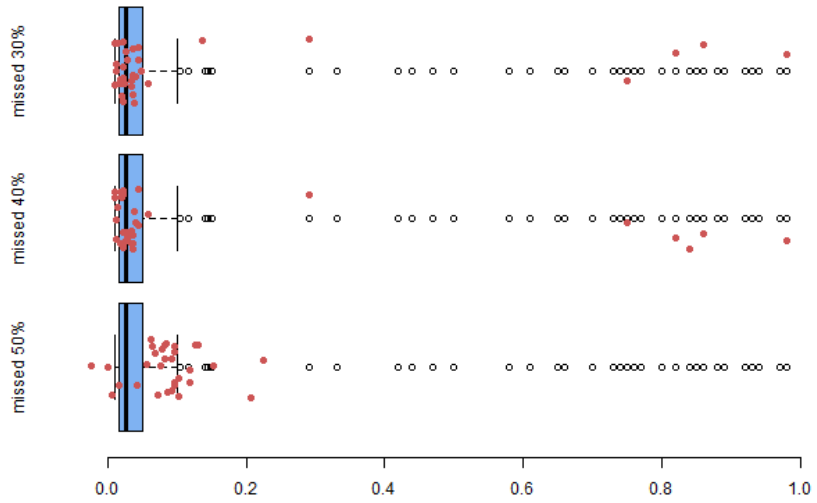

Figura 3: Datos imputados por regresión lineal

Bajo el modelo lineal, los valores imputados para la variable *Repitencia* están resaltados en rojo. Para los casos de pérdida de información del 30 % *y* 40 % el modelo logra captar parte de la variabilidad y se puede observar que los valores imputados están dentro de los intervalos de la variable.

Sin embargo; a pesar de que los valores se agrupan un poco más entorno a la media, con una pérdida de información del 50 %; se generaron 3 imputaciones con valores negativos -representados debajo del límite inferior de la última gráfica- $(-0{,}0004, -0{,}023, -0{,}035)$. Esta imputación en particular, no arrojo resultados

adecuados para el tipo de variable y una propuesta para solucionar este error en la imputación, es trabajar con un modelo Beta.

11.1.3. Método de Imputación Regresión Beta

Ahora, por medio del algoritmo *MICE* y la función *betareg*; se propone un método de imputación para las variables de tasas o proporciones con el modelo de regresión Beta. El algoritmo aplicado en este trabajo, es una función iterativa que integra inicialmente la funcionalidad de *MICE*, y a partir de las imputaciones obtenidas; es aplicada la regresión Beta:

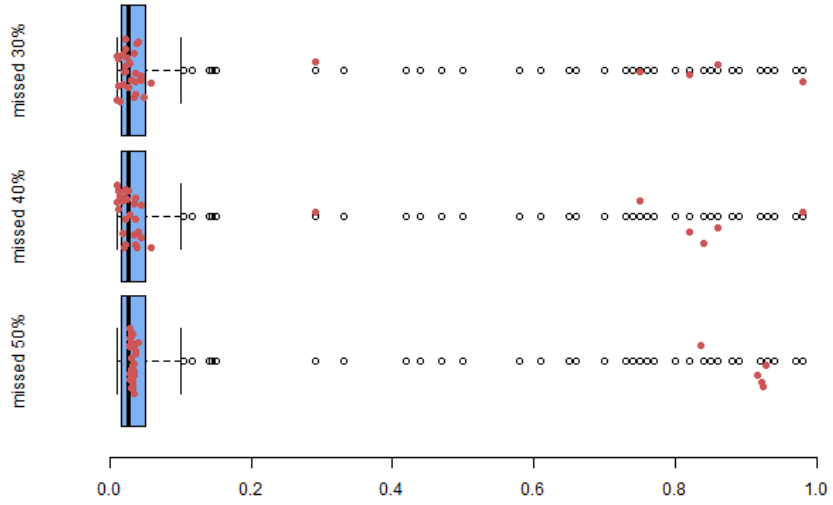

Figura 4: Datos imputados por regresión Beta

Para el 50 % de información faltante; las imputaciones del modelo de regresión Beta, se empiezan aproximar más hacia la media. En general, este proceso de impiutación permitió corregir el error de subestimación del modelo lineal.

Comparación de las Imputaciones

La siguiente tabla, resume las medias de las variables de los datos originales y de las imputaciones realizadas para la base que presentaba 50 % de información faltante:

Método	Matriculación	Cobertura	Deserción	Aprobación	Reprobación	Repitencia
Reales	0.842	0.848	0.046	0.909	0.102	0.103
Regresión Lineal	0.842	0.848	0.046	0.91	0.105	0.095
Regresión Beta	0.838	0.85	0.045	0.911	0.101	0.104

Cuadro 5: Promedios de la base de datos original y las imputadas

No se observan diferencias significativas en los promedios. Para comprobar este supuesto, se realizó una prueba de medias *t Student*, la cual permite contrastar si existen diferencias significativas en la media de las variables de la base de datos original; frente a las base de datos con valores imputados (Cristian Tellez, 2014). De esta manera es posible afirmar que las medias no tienen presentan diferencias estadísticamente significativas.

Finalmente se visualiza una gráfica de los datos imputados con el modelo lineal y modelo Beta, para la variable *Repitencia*.

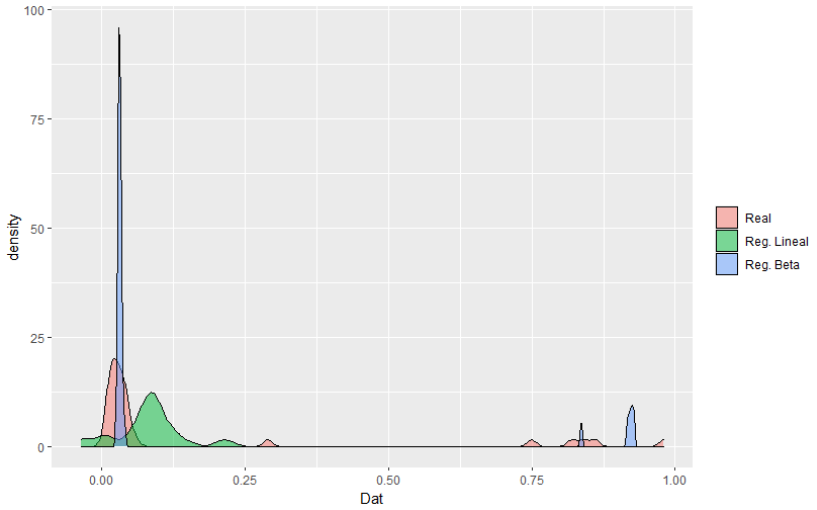

Figura 5: Comparación datos reales vs. imputados

Nuevamente se puede evidenciar que bajo pérdida de información del 50 %, las imputaciones por medio de la regresión Beta se agrupan hacia la media.

11.2. Aplicación para una Base de Datos Incompleta

El segundo ejemplo de aplicación se hará para una base de datos obtenida por medio del Banco Mundial. La base de datos reúne información de 217 países, con estadísticas ecónomicas de la agricultura. A continuación un detalle de las variables cuantitativas:

- **Fertilizantes**: Consumo de fertilizantes, representado en kilogramos por hectárea cultivable.
- **Cultivos por área**: Tierra cultivable, calculado como un porcentaje del área de la tierra.
- **Cosecha**: Índice de variación de la producción agrícola con respecto a los periodos 2004 y 2006.

- **Alimentos**: Índice de variación de la producción de alimentos con respecto a los periodos 2004 y 2006.
- **Animal**: Índice de la producción animal con relación a los periodos 2004 y 2006.
- **%PIB Agricultura**: Porcentaje de la agrícultura en el PIB del país.
- **%Población Rural**: Porcentaje de población rural, con respecto al total de la población.

11.2.1. Información Descriptiva

La base de datos contiene información de 7 variables cuantitativas y 3 de ellas son variables continuas que toman valores entre 0 y 1 (*Cultivos por área*, *%PIB agricultura* y *%Población rural*). En total se presentan 144 valores ausentes.

La siguiente tabla muestra estadísticas generales de la base de datos original:

	Mín	Máx	Media	Desviación	NAs	% NAs
Fertilizantes	0	3573.9	198.2	407.3	53	24.42
Cosecha	64.4	156.1	101.8	12.1	17	7.83
Alimentos	74.2	152.2	103.1	9.9	15	6.91
Animal	62.9	154.7	103.7	9.9	16	7.37
Cultivables	0.001	0.6	0.14	0.14	12	5.53
%PIB Agricultura	0	0.59	0.1	0.1	28	12.9
%Población Rural	0	0.87	0.39	0.24	3	1.38

Cuadro 6: Datos de la base incompleta

En 2018 los países con mayor consumo de ferlizantes son Hong Kong y Malasia (3573 kg, 2106 kg, respectivamente); el país de Singapur no reporta consumo de felitizantes durante este año. También cabe resaltar que Dinamarca, Bangladesh, Ucrania, India y Moldavia cuentan con la mayor extensión de tierras productivas (áreas aprovechables mayores que 51 %), mientras que Yibuti es el país con menor porcentaje de tierras cultivables (8,6 %). Los países de Mongolia y Granada muestran las variaciones más significativas de producción agrícola, de alimentos y animal, con respecto al periodo 2004-2006. Otro punto a resaltar es Sierra Leona, como la agricultura que más aporta al PIB nacional (58,9 %), en contraste con San Marino donde su agricultura aporta un 1,6 % del PIB. Por último es importante destacar que en el 2018, lapara 69 países; la población rural superaba la población urbana.

Patrones de datos faltantes

La siguiente imagen, identifica los patrones de datos faltantes en la base de datos:

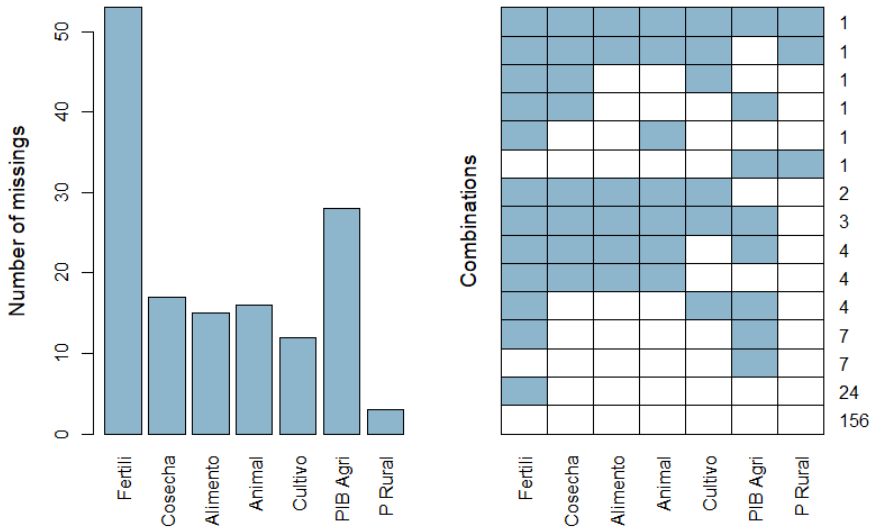

Figura 6: Patrones de datos faltantes

En total hay 15 patrones de datos faltantes; que incluye los 156 países que cuentan con información en todas sus variables. Además se puede observar que un país (Isla de San Martín) no cuenta con información en ninguna de sus variables; por tanto este registro no será considerado para los procesos de imputación, considerando que los procesos de imputación con regresión estiman los valores con las variables observadas.

11.2.2. Método de Imputación Regresión Lineal

Los datos imputados para la variable *%PIB Agricultura*, se pueden visualizar en rojo en la siguiente gráfica:

Nota: El porcentaje del PIB de agricultura, es la variable con mayor cantidad de datos ausentes; por lo tanto los resultados presentados corresponden a esta variable.

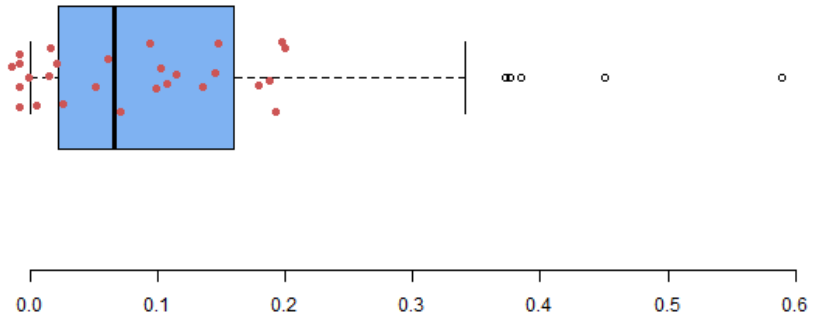

Figura 7: Datos imputados por regresión lineal

En este caso, el proceso de imputación por medio del modelo lineal generó 6 valores negativos. Ahora, se realizará la imputación con un modelo Beta, buscando solucionar las imputaciones negativas para esta variable en particular.

11.2.3. Método de Imputación Regresión Beta

La siguiente gráfica, representa las imputaciones generadas con un modelo lineal Beta:

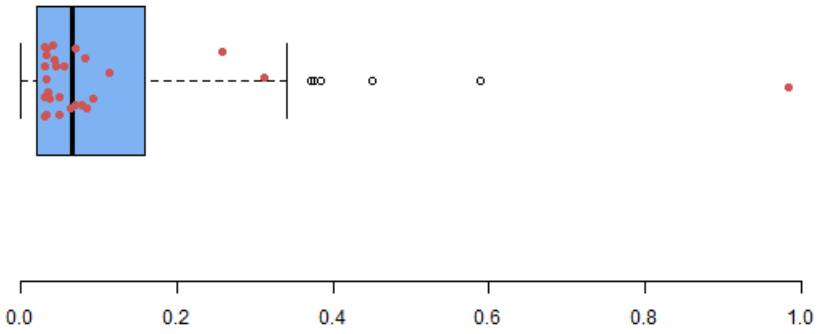

Figura 8: Datos imputados por regresión Beta

Las imputaciones obtenidas a partir de la regresión Beta, generaron un valor atípico para Tuvalu (0,98) este país tampoco cuenta con información en las variables *Fertilizantes* y *Cultivables*. Sin embargo, las demás imputaciones se modelan de mejor manera con respecto a la media.

Comparación de las Imputaciones

En la siguiente tabla, se muestran los promedios de las variables con la base de datos original, en comparación con los promedios calculados a partir de las imputaciones realizadas por cada método:

Método	Fertilizantes	Cosecha	Alimentos	Animal	Cultivables	% PIB Agri.	%P. Rural
Reales	198.2	101.9	103.1	103.7	0.141	0.101	0.39
Regresión Lineal	201	101.7	103.1	103.7	0.14	0.098	0.39
Regresión Beta	190.9	101.9	102.9	103.5	0.141	0.101	0.39

Cuadro 7: Promedios de la base de datos original y las imputadas

No se observan diferencias significativas en los promedios.

Para comparar los resultados, se muestra la gráfica de densidad para los datos originales omitiendo los valores faltantes (figura a la derecha), frente a la densidad de los valores imputados por los 2 métodos utilizados, para la variable *% PIB Agricultura*.

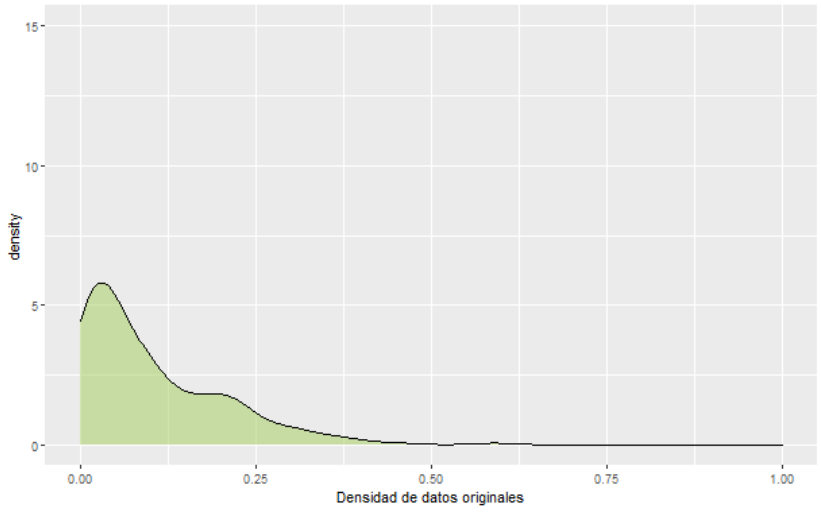

Figura 9: Datos reales, omitiendo NAs

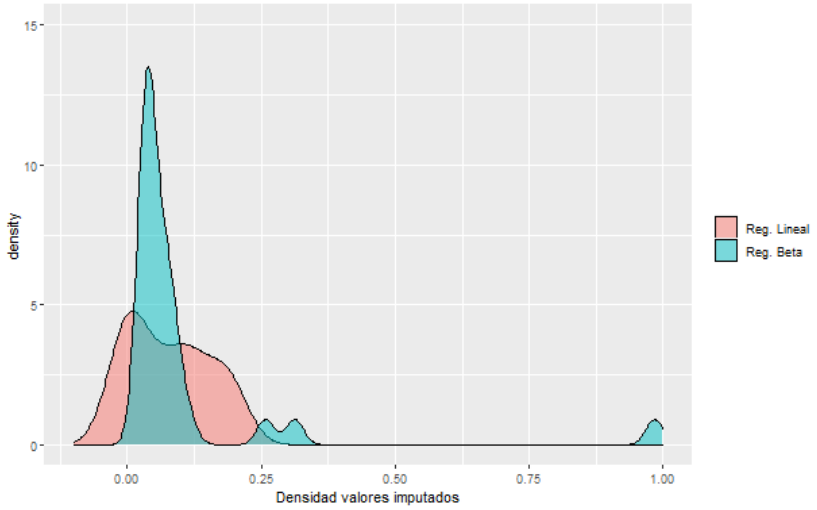

Figura 10: Datos imputados imputados

Con el gráfico de densidad es posible observar que los valores imputados del modelo lineal tienen un comportamiento más semejante a la distribución de la variable *% PIB Agricultura*, sin embargo; para análisis posteriores es necesario corregir los valores negativos. Las imputaciones del modelo Beta están captando mayor variabilidad.

Por último, para la variable *% PIB Agricultura*, se presenta un gráfico de cajas resaltando los valores imputados de los 2 procesos ejecutados en la base de datos:

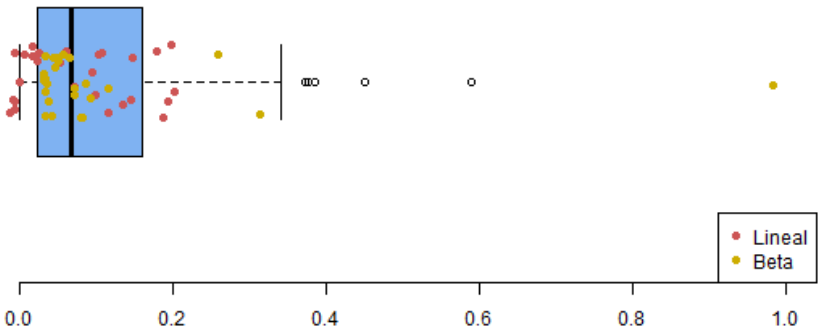

Figura 11: Comparación datos imputados en la variable

Los valores resaltados en rojo representan las imputaciones bajo un modelo lineal y los datos en amarillo corresponden a las imputaciones del modelo Beta.

11.3. Anexos

1. Código R, para la base de datos del Ministerio de Educación.
2. Base de datos, Ministerio de Educación, en Excel.
3. Código R, para la base de datos del Banco Mundial.
4. Base de datos, Banco Mundial, en Excel.
5. Carta para el tratamiento de datos.
6. Presentación.

12. Conclusiones

El presente trabajo, se planteó como objetivo; proponer un método adecuado para imputar datos, para las variables con valores en el intervalo $(0, 1)$. La distribución Beta modela este tipo de variables, por tanto el método de imputación propuesto se basa en el modelo de regresión lineal Beta.

La función elaborada en el software R, aplicada a las bases de datos del Ministerio de Educación (indicadores estudiantiles del país) y del Banco Mundial (datos macroeconómicos de agricultura); cuenta con la capacidad de imputar las variables de tasas y proporciones, -con ayuda del algoritmo *MICE*-; además de otro tipo de variables cuantitativas. La función *MICE* en el paquete R, es utilizada para una imputación inicial de datos, como también para la imputación de variables que no cumplan la condición de rango $(0, 1)$. Por último, y aprovechando las imputaciones obtenidas con el algoritmo *MICE*, aplica la función *betareg* para imputar las variables de tasas o proporciones en las bases de datos. La función procesó la información correctamente y logró completar el proceso de imputación.

Para evaluar la eficiencia del método propuesto, se aplicó a las bases de datos un proceso de imputación múltiple por regresión lienal, con el algoritmo *MICE*. Los procesos de imputación aplicados en las variables de intervalo $(0, 1)$, en algunos casos arrojaron valores negativos. Esto podría generar limitaciones de interpretación en análisis posteriores, además de perder confiabilidad en las estimaciones observadas por este método de imputación.

Los procesos de impiutación se aplicaron en dos bases de datos con características diferentes. La información en la base de datos del Ministerio de Educación estaba completa, mientras que la data del Banco Mundial presentaba valores faltantes. Entonces, para la primer base se ejecutó un proceso de amputación con la funcionalidad de *MICE*, generando pérdida de información de 30, 40 y 50 %; para esta última fue donde se obtuvieron imputaciones negativas con el modelo de regresión lineal. Las 6 variables seleccionadas de la base del Min. Educación, presentaban la restricción del rango $(0, 1)$ y fue evaluado el método de imputación propuesto con la función *betareg*.

La base del Banco Mundial presentaba información faltante en todas sus variables (7 cuantitativas), esta base a diferencia de la primera, contaba con valores fuera del intervalo $(0, 1)$. Las imputaciones con el algoritmo *MICE*, generaron para una de sus variables $(0, 1)$ imputaciones negativas. Para las variables diferentes al intervalo mencionado, la función generó las imputaciones con regresión lineal por medio del algoritmo *MICE* (integrado en la función) y las variables de intervalo $(0, 1)$ son imputadas con el modelo regresión Beta.

Los valores imputados con el modelo propuesto, garantizan estimaciones que siempre se encuentren en el rango $(0, 1)$. Aunque la función fue evaluada

con otras simulaciones, cada base de datos tiene particularidades; por tanto esta función podría tener un mayor alcance y mejora a medida que pueda ser ejecutado con diferentes bases de datos.

Se espera que la función propuesta y desarrollada en el software R; bajo un modelo Beta, pueda tener utilidad para posteriores investigaciones sobre el tema; teniendo en cuenta que los métodos de imputación disponibles en R no cuentan con una metodología de imputación especiífica para variables con intervalo $(0, 1)$.

Referencias

Cristian Tellez, S. G., Diego Lemus. (2014). *Estadística Inferencial con Aplicaciones en R*. Institución Universitaria Los Libertadores.

Dagnino, J. (2014). Datos Faltantes (Missing Values). *Revista Chilena de Anestesía*, (43), 332-334.

Dennis D. Wackerly, R. L. S., William Mendenhall III. (2010). *Statistical Analysis with Missing Data* (7st). Cengage Learning Editores, SA de CV.

Enders, C. K. (2010). *Applied Missing Data Analysis*. The Guilford Press.

Fernando Medina, M. G. (2007). *Imputación de Datos: Teoría y Práctica* (inf. téc.). CEPAL.

Francisco Cribari-Neto, A. Z. (2010). Beta Regression in R. *Journal of Statistical Software*, *34*.

Geert Molenberghs, M. G. K. (2007). *Missing Data in Clinical Studies*. John Wiley & Sons Ltd.

Juan Francisco Muñoz Rojas, E. Á. V. (2009). Métodos de imputación para el tratamiento de datos faltantes: aplicación mediante R/Splus. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, (3), 3-30.

Mitas, J. B. L. H. G. M. F. M. G. (2018). *Visualización y Métodos de Imputación de Datos Faltantes en la Encuesta de Gasto de los Hogares* (inf. téc.). Universidad del Rosario de Argentina.

Moisés, C. C. (2014). *Imputación de Datos Faltantes en un Modelo de Tiempo de Fallo Acelerado* (Tesis de maestría). Universidad da Coruña.

Rianne Margaretha Schouten, G. V. (2021). The Dance of the Mechanisms: How Observed Information Influences the Validity of Missingness Assumptions. *Sociological Methods & Research*, *50*, 1243-1258.

Rianne Margaretha Schouten, G. V., Peter Lugtig. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, *88*(15), 2909-2930.

Roderick J. A. Little, D. B. R. (2020). *Statistical Analysis with Missing Data* (3rd). John Wiley & Sons, Inc.

Silvia L.P. Ferrari, F. C.-N. (2004). *Beta Regression for Modelling Rates and Proportions*.

Stef Van Buuren, K. G.-O. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*, 1-67.

Villegas, M. Á. G. (2005). *Inferencia Estadística*. Díaz de Santos.