UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Maestría en Estadística Aplicada
Trabajo de grado
Enero 2026

Modelo de regresión beta con valores faltantes en las covariables: una aplicación sobre la tasa de pobreza multidimensional de los municipios de Colombia.

Beta regression model with missing values in the covariates: an application to the multidimensional poverty rate of the municipalities of Colombia.

Rolando Javier Barajas Pérez.[a]
rolandobarajas@usantotomas.edu.co

Mario José Pacheco López.[b]
mariopacheco@usantotomas.edu.co

Resumen

En gran parte de los estudios donde se realiza el modelamiento de variables de interés es habitual encontrarse con la presencia de datos faltantes en las variables explicativas, lo cual constituye un reto en cuanto al correcto tratamiento que deben recibir estos datos, dado que la naturaleza misma de las variables influye en las posibles estrategias que deben implementarse. En este sentido, una de las estrategias más empleadas en el tratamiento de datos faltantes es la imputación. Si bien, existen diversas formas de aplicación de este proceso no siempre se realiza de forma correcta, puesto que en algunos casos no se tiene en consideración que el método estadístico y el método de imputación elegido van de la mano. Por lo anterior, se desarrolló una metodología para el tratamiento de valores faltantes en las variables explicativas dentro de un modelo de regresión beta, permitiendo el modelamiento de variables de tipo proporción o razón, como es el caso de la tasa de pobreza; tema que es de gran interés para la formulación de políticas pública, en especial, para generar acciones y proyectos que se enfoquen en mejorar la calidad de vida de las personas dentro del país.

***Palabras clave*:** Tratamiento de datos faltantes, imputación múltiple, regresión beta, variables de razón, pobreza.

1. Introducción

Desde hace algunos años se han venido aunando esfuerzos, desde diversos sectores, para incentivar mecanismos orientados a mejorar las condiciones de vida de las personas en Colombia (Chavarro et al., 2017). Es por ello, que en el año de 2015, la Organización de Naciones Unidas (ONU), de la cual el estado colombiano es miembro, propone la agenda titulada: "Transformar nuestro mundo: la Agenda 2030 para el Desarrollo Sostenible" (*Transforming our world: the 2030 Agenda for Sustainable Development* en inglés), en la que se definen diecisiete (17) objetivos comunes, los cuales buscan luchar contra la pobreza extrema a partir de la integración y el equilibrio de tres dimensiones esenciales para el desarrollo sostenible (lo económico, lo social y lo ambiental), proporcionando así una valiosa hoja de ruta para articular la formulación de políticas mundiales (Gil, 2018). Es precisamente en este punto que los estudios sobre pobreza en Colombia cobran gran importancia, y más si se tiene en cuenta el comportamiento diferencial

[a] Estudiante
[b] Director

que puede presentarse a nivel de los departamentos, municipios o incluso veredas; situación que es relevante para implementar planes, proyectos y estrategias que estén focalizadas en los factores particulares que presenta la pobreza dentro de cada región.

Así, el Objetivo de Desarrollo Sostenible Nº 1 (ODS – 1), en palabras de Ramos (2016), busca: "reducir al menos a la mitad la proporción de hombres, mujeres y niños(as) de todas las edades que viven en pobreza en todas sus dimensiones con arreglo a las definiciones nacionales". En virtud de este interés común de las naciones, en Colombia se han venido realizando, en los últimos años, diversos estudios encaminados a identificar las formas de pobreza y el método más adecuado para su medición. Es así como el Departamento Administrativo Nacional de Estadística (DANE) ha tomado el enfoque de pobreza multidimensional dada por Alkire y Foster (2007), el cual propone que una persona es pobre cuando sufre múltiples privaciones en la mayoría de las dimensiones humanas; dimensiones que están determinadas (medidas) por distintos indicadores asociados al bienestar social y económico de los individuos. Así, para estos autores, un sujeto cuya variable de medición cae debajo de la línea de corte en la evaluación total de indicadores es definida como pobre.

No obstante, es importante señalar que la mayoría de los estudios estadísticos realizados en Colombia no emplean precisamente esta medida de pobreza, ya que, dependiendo del foco investigativo de cada estudio, se emplean métodos estadísticos distintos. Un ejemplo de ello son los estudios realizados por Narváez et al. (2015) que consideraron diversas variables socioeconómicas para estimar la pobreza desde un enfoque monetario, empleando un modelo Probit que tomaba como información base la encuesta de calidad de vida, para definir así qué hogares tenían mayores probabilidades de ubicarse en situación de pobreza. Este trabajo fue un ajuste del trabajo realizado por Arias et al. (2007), el cual utilizó un análisis de regresión binomial. Otro ejemplo corresponde a los resultados obtenidos dentro de la investigación de Teitelboim (2008), en la que se buscó establecer la relación que tienen una serie de características de los jefes de hogar con los niveles de pobreza, para lo cual se ajustó un modelo logístico que permitiera desagregar dichos niveles respecto a lo regional, municipal y comunal.

Por otra parte, existen casos como el de Madariaga et al. (2013), quien proyectó un modelo de regresión lineal para estudiar los problemas de pobreza en trece de las principales ciudades de Colombia. En este estudio se consideraron los porcentajes de pobreza y extrema pobreza, asociándolos al indicador de desigualdad económica de una población (GINI), para, de este modo, concluir la dependencia lineal entre las variables estudiadas. Sin embargo, también existen distintos antecedentes que toman en consideración la medida de pobreza multidimensional adoptada por el DANE. Al respecto se presenta la investigación desarrollada por Orjuela Montoya (2021), en la que se propone el impacto potencial de la informalidad sobre la pobreza multidimensional. Justamente, este autor presentó un modelo logístico probit que buscaba evaluar el impacto de la política pública en la reducción de la informalidad, para lo cual tomó como indicador de medición el modelo de pobreza multidimensional, siguiendo la propuesta metodológica descrita en párrafos anteriores (Alkire & Foster, 2007).

Si bien es cierto, todos estos estudios han permitido de una u otra forma tener un acercamiento sobre el estado de la medida de pobreza en Colombia, organismos como la Comisión Económica para América Latina y el Caribe (CEPAL) ha hecho hincapié en la necesidad de realizar estudios más rigurosos y bajo un enfoque estadístico que contemple la naturaleza real de las variables asociadas a la tasa de pobreza. Este es el caso de Feres y Mancero (2001b) y Botello (2017), los cuales han descrito los desafíos estadísticos que existen para una adecuada medición del índice de pobreza; toda vez que en palabras de Melo Martínez (2010), en ocasiones: "se cae en el abismo de ajustar fenómenos a los modelos que se tiene a disposición, en lugar de permitir que los datos hablen por sí solos". Lo anterior suele ocurrir con regularidad cuando la intención es modelar covariables con una respuesta, para lo cual se opta por emplear modelos de regresión lineal, lo que, desde el aspecto teórico y el comportamiento de la variable, no es adecuado porque la variable está restringida a un intervalo abierto de cero y uno, como es el caso de la tasa de pobreza.

Por consiguiente, el mismo Melo Martínez (2010) destaca las bondades en el uso de la distribución beta, dado que ella ofrece mayor flexibilidad en los estudios que modelan variables en el intervalo anteriormente mencionado. Para ampliar esta conclusión, el autor cita los trabajos de Ferrari y Cribari-Neto (2004), quienes proponen los modelos de regresión beta para el modelamiento de proporciones y razones, dado que se toman en consideración regresores y parámetros desconocidos. Así, en busca de tomar en consideración lo propuesto por los estudios de la CEPAL, se ajustó un modelo de regresión beta que tomó en cuenta el comportamiento de las covariables que permiten definir la tasa de pobreza en los municipios de Colombia.

En este punto es importante mencionar que la presencia de datos faltantes es una situación a la que se enfrentan, frecuentemente, investigadores y tomadores de decisiones (Medina & Galván, 2007). Es justamente por dicha limitante que resulta de vital importancia el tratamiento que los mismos deben recibir, puesto que la implementación incorrecta de estrategias puede afectar significativamente la etapa inferencial de cualquier estudio estadístico. Para el caso de los estudios de pobreza, esta ausencia de datos es una prioridad dentro de los procesos de medición, más si se considera que existen limitaciones en el acceso y disponibilidad de la información, producto de la brecha, geográfica y tecnológica, que poseen distintas zonas del país que se encuentran ubicadas en la periferia, que no cuenta con acompañamiento estatal o se encuentran en situación de coerción política o conflicto armado; ausencia que afecta la existencia de datos asociados a las variables explicativas que ayudan a comprender el comportamiento de la pobreza en estas regiones.

Por lo anterior, el poder definir un adecuado método de tratamiento de datos faltantes es una tarea fundamental, ya que no basta con tener claridad sobre la naturaleza de la variable o el modelo que mejor se ajuste a su comportamiento, sino que se debe considerar como manejar la información ausente para poder cumplir de forma ideal con las inferencias, resultados y conclusiones que se derivan de la implementación de modelos como los que presentan una distribución beta. En este contexto, Mejía Giraldo y Restrepo Betancur (2019) proponen la imputación como una estrategia idónea en la determinación de valores faltantes, dado que esta técnica es la más conocida y usada para tales fines. Sin embargo, la imputación, en sí misma, tiene sus consideraciones, dado que suele centrarse en planteamientos de estimación de la media y su variación bajo diseños muestrales simples (Muñoz Rosas, Álvarez Verdejo et al., 2009). (Muñoz Rosas, Alvarez Verdejo et al., 2009).

Aunque la imputación suele ser la solución a los problemas de datos faltantes, este no es el único método plausible; existen otras posibilidades que permitan el mejor tratamiento de los datos que pueden considerar métodos heurísticos, técnicas deterministas como el algoritmo EM (Expectation Maximization), técnicas aleatorias o estocásticas incluyendo métodos de imputación múltiple (Viada et al., 2016). Es por todo lo anterior que el tratamiento de valores faltantes dentro de los estudios centrados en la estimación de la pobreza constituye un aspecto fundamental para dar cumplimiento al ODS – 1, toda vez que permitirá generar acciones de política pública que impacten, de forma directa, la proporción de pobreza con adecuación no sólo a la definición nacional, sino coherente y consistente con las realidades particulares de las poblaciones del país.

De este modo, la investigación permitió formular una estrategia de medición (modelo de regresión beta que considere tratamiento de datos faltante) que atienda a las demandas y necesidades de las regiones más vulnerables del país, las cuales son, precisamente, aquellas que carecen de la infraestructura política, tecnológica y comunicacional para contar con información completa e integral sobre la pobreza, favoreciendo así una verdadera transformación en el bienestar y la calidad de vida de sus habitantes.

2. Marco teórico y conceptual

En este apartado se presentan las categorías teóricas que fundamentan la investigación, las cuales corresponden a: pobreza, tratamiento de datos faltantes, reglas de Rubín y los modelos de regresión beta. Así, esta sección realiza una aproximación a los conceptos y autores que definen estas nociones, lo que permite consolidar un mapeo de la información estadística de interés para el objeto de este estudio.

2.1. Pobreza

Comprendiendo la relevancia que tienen los estudios de pobreza en el mundo y la necesidad de implementar políticas públicas que coadyuven al logro de los ODS, es necesario iniciar este capítulo presentando una aproximación a la definición orgánica de la pobreza dentro del país, para lo cual se recurre a los estudios realizados por Herrera (2009), quien cita a Romero (2002) respecto a la importancia que tiene posicionar dentro de los estudios de este tema la visión de la CEPAL. Al respecto, esta entidad señala que: "la noción de pobreza expresa situaciones de carencia de recursos económicos o de condiciones de vida que la sociedad considera básicos de acuerdo con normas sociales de referencia que reflejan derechos sociales mínimos y objetivos públicos".

Por consiguiente, la pobreza no puede considerarse, únicamente, desde un punto de vista monetario, sino que en ella convergen otros factores, los cuales están asociados a la calidad de vida y el bienestar de los individuos dentro de una comunidad. Es justamente por lo anterior que, dentro de las estimaciones de pobreza, se toma como punto de referencia el Índice de Necesidades Básicas Insatisfechas (NBI), el cual se constituye en "un método directo para identificar las carencias que tiene una población y con ello establecer si ésta se encuentra en situación de pobreza" Feres y Mancero (2001a). Este modelo se ha utilizado, dentro del país, como un mecanismo para focalizar la asignación y transferencias de recursos a municipios considerados vulnerables, lo que afecta de forma significativa la manera en que se mide la pobreza dentro de la nación (Baltazar et al., 2007).

En virtud de lo anterior, se hace necesario definir y analizar las posibles variables teóricas que están asociadas a las estimaciones de la pobreza. Inicialmente, dentro de esta investigación se recoge, lo señalado por la CEPAL, a través de las formulaciones de Feres y Mancero (2001b), así como de Núñez Méndez y Ramírez Jaramillo (2002), quienes establecen, siguiendo la metodología del NBI, que, la disminución o aumento de la pobreza está asociado con "los niveles de ingreso de los hogares, en los cuales habitan tres personas por miembro ocupado y en donde el jefe tenga, como máximo, dos años de educación primaria aprobados" (Fresneda Bautista, 2007). Así, siguiendo la investigación, los indicadores que determinan o definen a un sujeto pobre se sitúan en cuatro áreas de necesidades básicas que tienen los individuos: vivienda, servicios sanitarios, educación e ingreso. Justamente, algunos estudios muestran, siguiendo esta línea, que en Colombia la pobreza medida con el NBI ha venido disminuyendo sostenidamente como consecuencia de la mejoría en el acceso a los servicios de educación, salud, vivienda y agua potable, pero no sucede lo mismo cuando se mide la pobreza por el nivel de ingresos de los hogares.

Seguidamente se la definición, con sus variables, tomada y ajustada por el DANE sobre pobreza multidimensional, la cual considera un doble umbral de percepción: el primero identifica a las personas u hogares que presentan privaciones en determinados aspectos y, el segundo, que distingue multimodalmente la cantidad de aspectos que generan una "condición" en individuos u hogares para no superar determinado valor referencial (Alkire & Foster, 2007). Así pues, a partir del anterior concepto surge el Índice de Pobreza Multidimensional (IPM) de Alkire y Foster, el cual permite obtener información acerca de su incidencia, profundidad y severidad. Este índice podrá, siguiendo la base de su formulación, ser adaptable a las dinámica y realidades de cada país, permitiendo monitorear, en un marco global pero desde el contexto local, su propio nivel de pobreza y reducirla según las prioridades regionales pero desde una agenda internacional (Alkire & Foster, 2007).

De este modo, el DANE desagrega este IPM dentro de la Medida de Pobreza Multidimensional Municipal, la cual es una fuente de tipo censal, que se encuentra constituida por las siguientes dimensiones humanas

y sociales: condiciones educativas del hogar, condiciones de la niñez y la juventud, salud, trabajo y situaciones propias de la vivienda, así como acceso a servicios públicos domiciliarios. A su vez, estas dimensiones a su vez están asociadas a quince (15) indicadores, todos ellos amparados en la encuesta de calidad de vida (ECV), según la información establecida por el mismo DANE dentro de su página oficial.

2.2. Tratamiento de datos faltantes

Dentro de este marco, se hace necesario realizar las precisiones asociadas al componente estadístico en los procesos de estimación de la pobreza, el cual permitirá comprender de mejor forma la metodología que se busca proponer para el tratamiento de valores faltantes en las covariables de un modelo de regresión beta; lo anterior pensado en una aplicación de la estimación los indicadores descritos en los municipios de Colombia. De este modo, a continuación, se enlista una serie de conceptos propios de los estudios estadísticos, la mayoría de ellos desde la perspectiva de Rubin y Little (2019):

- **Datos faltantes:** son valores no se encuentran disponibles o cuya informa sería útil o significativa para el análisis de los resultados o el estudio mismo. Existen varios tipos de datos faltantes y aun más razones por las cuales pueden presentarse; sin embargo, se destacan dos factores que son decisivos al enfrentar la ausencia de datos en el momento de analizar los resultados, donde lo principal es decidir si la pérdida es aleatoria, es decir, afecta por igual a todos los individuos, o bien puede ser debida a una razón o razones específicas que pueden introducir sesgos que invaliden los resultados.

- **Mecanismos de datos faltantes:** describen posibles relaciones entre variables medidas y la probabilidad de datos faltantes. Nótese que un patrón de datos faltantes sólo describe la localización de los agujeros en los datos, y no explica por qué los datos están faltantes. Entre los mecanismos existe:

  - MCAR (*Missing Completely At Random*): la probabilidad de que una respuesta a una variable sea dato faltante es independiente tanto del valor de esta variable como del valor de otras variables del conjunto de datos.
  - MAR (*Missing At Random*): la probabilidad de que una respuesta sea dato faltante es independiente de los valores de la misma variable, pero es dependiente de los otros valores de variables del conjunto de datos.
  - NMAR (*Not Missing At Random*): la probabilidad de que una respuesta a una variable sea dato faltante es dependiente de los valores de la variable.

- **Tratamiento de datos faltantes:** constituye el adecuado manejo de la información no disponible en el estudio; para lo cual existen alternativas para lidiar con los datos faltantes, la primera de ellas consiste en omitir variables con datos faltantes y la segunda en omitir individuos en quienes hay datos faltantes. Aunque, estos métodos son los más usados, probablemente con mayor frecuencia, tiene el agraviante que producen una pérdida de información y rigurosidad del estudio; además, no modifican el riesgo de sesgos, por lo que la mayoría de los autores recomiendo no ser usados. Así pues, existe una tercera alternativa que consiste en estimar (imputar) los datos faltantes, donde estos son reemplazados con valores predichos desde los datos presentes.

- **Imputación:** consiste en el proceso de conversión de los datos faltantes en datos simulados, de modo que se pueda intentar no interferir o minimizar de la mejor forma la incidencia en el resultado esperado. Este proceso genera una o más versiones de los datos imputados ("simulados") basado en la cantidad de iteraciones especificadas. Cada iteración genera una versión de los datos que deben ser analizadas y estimadas según el interés sobre los datos. Al finalizar, la totalidad de estimaciones iteradas son agrupados en un solo conjunto que permita la validación de ese dato imputado con relación al conjunto de datos totales.

- **Métodos de imputación:** son las diferentes técnicas que permiten llegar acabo un adecuado análisis de los datos mediante la obtención de la base completa que inicialmente no se disponía. Si una imputación se lleva a cabo de manera adecuada, podría disminuir el sesgo, en caso de existir; pero el uso de imputación también puede llevar a afectar las distribuciones conjuntas, o incluso, distribuciones marginales de las variables (Viada et al., 2016). Sin embargo, existe diversos métodos de imputación mencionados por Viada et al. (2016), citando el libro de Rubin y Little (2019) entre los que se tiene:

  - Métodos Heurísticos, consiste en una solución a priori cuando la cantidad de información perdida (faltante) es pequeña; se emplea el análisis de casos completos o casos disponibles; para lo cual, en el primer caso se descarta los casos faltantes y solo se aplica el análisis sobre aquellos valores observados de los cuales se disponga información de todas las variables. Respeto al otro, es una alternativa que busca utilizar para el cálculo toda la información disponible en la muestra, para así poder determinar medidas como de tendencia central o de variabilidad.
  - Métodos fundamentados en información externa, están basados en parámetros relacionados con una fuente de información (encuesta) pertinente a otras bases de datos; presentan subtécnicas como los métodos deductivos (los dados faltantes se deducen con cierto grado de incertidumbre) y tablas *Look – Up* (Tablas de información relacionada para imputar datos).
  - Métodos deterministas, donde se repite la imputación en los registros, de modo que al mantener las mismas condiciones se obtenga las mismas respuestas. Esta técnica presenta las siguientes diversificaciones: imputación de la media o moda, imputación de media de clases, imputación por regresión, emparejamiento media, imputación por el vecino más cercano, algoritmo EM (*Expectation Maximization*); y en este último se encuentra las mejoras como lo son las redes neuronales y modelos de serie de tiempo.
  - Métodos aleatorios o estocásticos, a diferencia de los métodos deterministas, estos buscan repetir la imputación bajo las mismas condiciones para cada registro, pero buscando obtener resultados distintos. Entre los tipos de técnicas estocásticas se encuentran: la imputación aleatoria de un caso seleccionado, imputación aleatoria de un caso seleccionado entre clases, imputación secuencial *Hot – Deek*, imputación jerárquica *Hot – Deek*, imputación por regresión aleatoria y la imputación por regresión logística.
  - Métodos de imputación simple, buscan solucionar el tema de valores faltantes reemplazando los mismos por datos estimados a partir de la información suministrada por la misma muestra; por lo cual, se construye una matriz completa sobre la cual se realiza los análisis requeridos por el estudio.
  - Método KNN, es una técnica de imputación empleada por medio computacional a través de la librería vim (Kowarik and Templ 2016), la cual genera un modelo que establece el valor faltante sobre la media de todos los valores situados a k-distancias del registro que contiene la información faltante; es decir, utiliza una distancia euclideana para determinar el valor faltante.
  - Predictive mean matching (pmm), es un procedimiento de imputación múltiple *Hot - Deck* de última generación, donde la calidad de sus resultados depende, entre otras cosas, de la disponibilidad de casos de disponibles adecuados.

- **Algoritmo MICE**: otro de los métodos utilizados es la imputación multivariada por medio de ecuaciones encadenadas (MICE, por sus siglas en inglés), llamada también "especificación totalmente condicional" o "imputación múltiple secuencial de regresión", que crea múltiples predicciones para cada valor faltante; sin embargo, si la calidad de los valores observados para el valor faltante es baja, se tiene como resultado una imputación de baja calidad. Los pasos generales para la imputación son:

  1. Se crea un valor temporal, que corresponde a la media de los datos observados para cada y_{ij}.

2. En la variable incompleta Y_{ij}, se estima un modelo de regresión con los datos observados, donde Y_j es la variable de respuesta y las demás son usadas como explicativas y estos "nuevos" valores reemplazan los valores generados en el paso anterior.
3. Los datos generados en los pasos anteriores, son repetidos para cada variable que tiene valores faltantes. El ciclo en cada variable consiste en una iteración o ciclo. Al final de cada ciclo todos los valores faltantes son reemplazados con predicciones que reflejan las relaciones observadas en los datos.
4. Los pasos 2 al 4 se repiten durante varios ciclos, actualizandose las imputaciones en cada ciclo.

Además, Mice (*Multivariate Imputation by Chained Equations*), es una librería (van Buuren and Groothuis-Oudshoorn 2011) que ofrece una serie de funciones orientadas a la visualización e imputación de datos perdidos; busca crear múltiples imputaciones para datos multivariados, donde cada variable puede ser imputada con un modelo estadístico particular.

- **Evaluación de imputación**: es bastante simple aplicar algunas de las metodologías disponibles de imputación; sin embargo, evaluar cuál es la más eficaz en cada caso es ya más complejo. El proceso completo de imputación implica realizar la imputación, estimar los parámetros y el agrupamiento ("*pooling*") de los resultados. En tal sentido, el proceso de estimación de los parámetros implicar una serie de etapas: búsqueda del modelo de imputación adecuado, optimización, validación, predicción y finalmente evaluación de la calidad del ajuste del modelo.

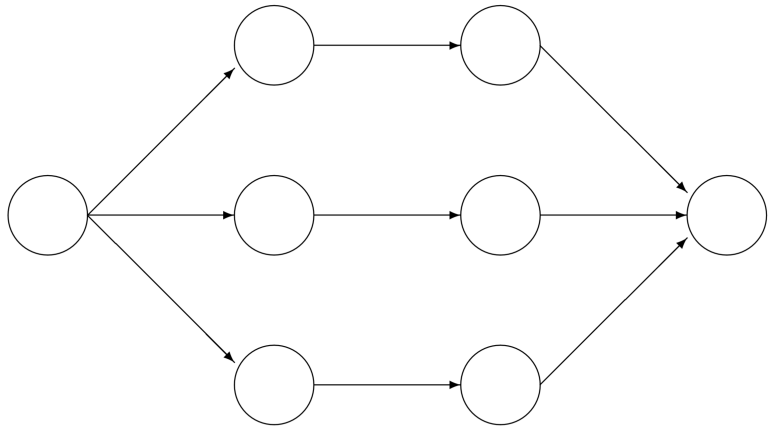

Figura 1. Pasos principales en el proceso de imputación. (Buuren 2018)

2.3. Reglas de Rubín

Rubín en el año de 1987 considera que la incertidumbre generada por los valores imputados es de vital importancia, por lo cual propone el método de Imputación Múltiple (MI), el cual consiste en un proceso estocástico que selecciona los posibles valores para la información faltante y utiliza dichos valores para recoger el componente aleatorio del dato imputado. Este proceso es realizado ya que se requiere que la validez del proceso de imputación este basado en el supuesto de que la información faltante ha sido omitida en forma completamente aleatoria, es decir que la probabilidad de no reportar la información no depende del valor de la variable.

No obstante, el proceso de Imputación Múltiple no soluciona el problema de la información faltante, sino que lo ajusta desde el enfoque estadístico. En este sentido, se puede contar con información

completa, pero es necesario disponer de múltiples bases de datos imputadas donde cada una de ellas tiene un valor posible para la observación faltante. Así pues, se debe desarrollar en análisis en cada una de las m bases de datos completas y luego combinar los resultados a fin de obtener las conclusiones; la combinación de los resultados sigue las reglas propuestas por Rubín.

Van Buuren (2018) establece que: sea θ un parámetro poblacional unidimensional de interés y $\widehat{\theta}$ su estimación puntual si no hubiera datos faltantes, denotando con $\widehat{v}(\widehat{\theta})$ la varianza estimada del estimador puntual. Luego de analizar datos imputados se tiene m estimaciones $\widehat{\theta}_1$, $\widehat{\theta}_2$, ... ,$\widehat{\theta}_m$ con varianzas estimadas asociadas $\widehat{v}(\widehat{\theta}_1)$, $\widehat{v}(\widehat{\theta}_2)$, ... ,$\widehat{v}(\widehat{\theta}_m)$. Se derivan las siguientes reglas de Rubín para combinar las m estimaciones; donde la estimación puntual para θ basada en imputación múltiple, $\overline{\theta}$, y su varianza asociada, $\widehat{v}(\overline{\theta})$, están dadas por:

$$\overline{\theta} = \frac{1}{m}\sum_{j=1}^{m}\widehat{\theta}_j \tag{1}$$

$$\widehat{v}(\overline{\theta}) = \frac{1}{m}\sum_{j=1}^{m}\widehat{v}(\widehat{\theta}_j) + (1+\frac{1}{m}) + B \tag{2}$$

donde

$$B = \frac{1}{m-1}\sum_{j=1}^{m}(\widehat{\theta}_j - \overline{\theta})^2 \tag{3}$$

En este sentido, $\frac{1}{m}\sum_{j=1}^{m}\widehat{v}(\widehat{\theta}_j)$ representa la variabilidad intra-imputaciones y B la variabilidad entre-imputaciones.

Por lo anterior, como medidas de severidad producto del proceso de imputación se pueden calcular:

- **Proporción de la varianza debida a los datos faltantes**:

$$\lambda = \frac{(1+\frac{1}{m})B}{\widehat{v}(\overline{\theta})} \tag{4}$$

- **Incremento relativo de la varianza**:

$$r = \frac{(1+\frac{1}{m})B}{\overline{v}(\theta)} \tag{5}$$

- **Fracción de información faltante**:

$$\gamma = (1+r)^{-1}\left[r + \frac{2}{v+3}\right] \tag{6}$$

siendo $v = \frac{n(n-1)}{(n+2)}(1-\lambda)$

Extensiones Multivariadas para $q > 1$

Severidad del problema de datos faltantes ($q > 1$)

- **Proporción de la varianza debida a los datos faltantes:**

$$\lambda = \left(1 + m^{-1}\right) q^{-1} \text{tr}\left(\mathbf{B}\left[\hat{V}(\bar{\boldsymbol{\theta}})\right]^{-1}\right) \tag{7}$$

- **Incremento relativo de la varianza:**

$$r = \left(1 + m^{-1}\right) q^{-1} \text{tr}\left(\mathbf{B}\bar{\boldsymbol{\Sigma}}^{-1}\right) \tag{8}$$

- **Fracción de información faltante sobre $\boldsymbol{\theta}$:**

$$\gamma = (1 + r)^{-1}\left[r + 2(\nu + 3)^{-1}\right] \tag{9}$$

con $\nu = \frac{n-q+1}{n-q+3}(n - q)(1 - \lambda)$

2.4. Modelos de Regresión Beta

En la teoría de probabilidad y estadística, representa una familia de distribuciones de probabilidad continuas con soporte en el intervalo (0,1); cuya densidad beta es caracterizada por dos parámetros positivos, indicados generalmente por α y β, los cuales son parámetros de localización y de escala. La distribución beta ha sido aplicada para modelar el comportamiento de variables aleatorias limitadas a intervalos de amplitud finita, en una gran variedad de áreas.

En tal sentido, una variable Y tiene distribución beta si su función de densidad está dada por:

$$f(y|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1 - y)^{\beta-1} I(0, 1)(y) \tag{10}$$

donde $\alpha > 0$, $\beta > 0$, $\Gamma(.)$ denota función gamma y $I(0, 1)(y)$ la función de indicador en el intervalo abierto (0, 1). Adicionalmente, la media y la varianza de Y, $\mu = E(Y)$ y $\sigma^2 = Var(Y)$, están dadas por:

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{11}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{12}$$

Por otra parte, los parámetros de media y dispersión se pueden modelar en función de variables explicativas, por ejemplo, como se propuso en Cepeda Cuervo y Garrido Lopera (2011), dado que los cambios en el parámetro de precisión pueden ser explicados por variables explicativas. La distribución beta dada en (1) también se puede repararmetrizar como una función de la media y la varianza, con:

$$\alpha = \frac{(1 - \mu)\mu^2 - \mu\sigma^2}{\sigma^2} \tag{13}$$

$$\beta = \frac{(1 - \mu)[\mu - \mu^2 - \sigma^2]}{\sigma^2} \tag{14}$$

Aunque escribiendo (1) en función de μ y σ^2 puede resultar en una expresión compleja, el modelado conjunto de la media y la varianza se puede lograr fácilmente aplicando la metodología bayesiana propuesta en Cepeda Cuervo y Garrido Lopera (2011); asi pues, el modelado conjunto de la media y

la varianza podría ser más apropiado que el modelado conjunto de la media y el llamado parámetro de precisión, dado que los parámetros de la regresión en los modelos serían más fáciles de interpretar.

La regresión beta es una aproximación bajo modelos lineales generalizado (GLM); esta regresión modela variables dependientes distribuidas con la distribución beta; es decir, datos con la distribución beta incluyen proporciones y razones, donde los valores x se encuentran entre 0 y 1 pero no inclusivo. Ahora bien, atendiendo a la recopilación realizada por Real Pérez (2015) se presenta las hipótesis de los modelos de regresión beta:

- Hipótesis distribucional, el modelo se basa en asumir que la variable objetivo y se distribuye según la distribución beta, la cual es apropiada para modelar proporciones. Por lo cual, se realiza la parametrización $\mu = \frac{\alpha}{\alpha+\beta}$ y $\phi = \alpha + \beta$, se tiene que $\alpha = \mu\phi$ y $\beta = (1-\mu)\phi$; en consecuencia, la esperanza y la varianza de la variable objetivo se pueden escribir en función de estos nuevos parámetros de la siguiente forma:

$$E[y] = \mu \tag{15}$$

$$Var[y] = \frac{\mu(1-\mu)}{1+\phi} \tag{16}$$

  donde μ es la media de la variable respuesta y ϕ puede ser interpretado como el parámetro de precisión, en el sentido de, fijado μ, a mayor valor de ϕ, menor es la varianza de la variable. La función de densidad puede ser escrita con la nueva parametrización como:

$$f(y,\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, 0<y<1 \tag{17}$$

  donde $0 < \mu < 1$ y $\phi > 0$ El modelo propuesto es también útil para situaciones donde la variable respuesta se restringe a intervalos genéricos (a,b), donde a y b son escalares a ¡b. En este caso se modela con $\frac{y-a}{b-a}$ en vez de modelar con y directamente.

- Hipótesis estructural, Sean $y_1, ..., y_n$ variables aleatorias independientes, donde para cada y_t, $t = 1, ..., n$, se asume la hipótesis distribucional anterior con media μ_t y parámetro de precisión ϕ desconocido. El modelo se obtiene asumiendo que la media de cada y_t, μ_t, es vinculado al predictor lineal, η_t, como sigue:

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = x_t^\intercal\beta = \eta_t \tag{18}$$

  donde $\beta = (\beta_1, ..., \beta_k)^\intercal$ es un vector desconocido de parámetros de regresión ($\beta \epsilon \mathbb{R}$) y $(x_{t1}, ..., x_{tk})$ es un vector de k componentes ($k < n$). Finalmente, $g : (0,1) \mapsto \mathbb{R}$ es una función vínculo estrictamente creciente y dos veces diferenciable. La varianza de y_t es una función de μ_t y, en consecuencia, de los valores covariantes x_t. Por tanto, la varianza de la variable respuesta puede ser incorporada al modelo. Algunas posibles elecciones para la función vínculo $g()$ son:

  - **Función logit:** $g(\mu) = log\{\frac{\mu}{1-\mu}\}$
  - **Función probit:** $g(\mu) = \Phi^{-1}(\mu)$, donde $\Phi(\cdot)$ es la función de distribución de una variable aleatoria normal estándar.
  - **Función log-log:** $g(\mu) = -log\{-log(1-\mu)\}$
  - **Función log-log complementaria:** $g(\mu) = log\{-log(1-\mu)\}$

  En particular, la función vínculo logit se puede escribir:

$$x_t^\intercal\beta = log\{\frac{\mu}{1-\mu}\} \Leftrightarrow \mu_t = \frac{e^{x_t^\intercal\beta}}{1+e^{x_t^\intercal\beta}} \tag{19}$$

  donde $x_t^\intercal = (x_{t1}, ..., x_{tk})$, $t = 1, ..., n$. Aquí, los parámetros de regresión tienen una interpretación interesante. Supongase que el valor de la i-ésima variable, x_{ti}, es incrementado en c

unidades, manteniendo las demás variables con su valor original, y sea μ^* la media de y bajo los nuevos valores de las variables explicativas y μ la media de y con los valores originales. Entonces, es fácil comprobar que:

$$e^{c\beta_i} = \frac{\mu^*/(1-\mu^*)}{\mu/(1-\mu)} \tag{20}$$

de manera que $e^{c\beta_i}$ es igual a la odds ratio.

2.5. Análisis de residuales

En el marco de la validación del modelo resulta fundamental realizar un análisis efectivo de los residuales, dado que el diagnóstico para cualquier modelo de regresión beta, constituye una herramienta estadística esencial cuando la variable de respuesta consiste en proporciones o tasas. Por lo anterior, en la distribución beta, tal y como la describen los estudios de Ferrari y Cribari-Neto (2004), se maneja la naturaleza intrínsecamente acotada y heterocedástica de los datos, el proceso no está completo sin una inspección rigurosa de los errores del modelo. La necesidad de este análisis surge precisamente porque los supuestos de normalidad y homocedasticidad de la regresión lineal tradicional quedan invalidados, por lo que el foco se desplaza hacia la evaluación de los residuales transformados

2.5.1. Residuales de Pearson

El diagnóstico de los modelos de Regresión Beta, ajustados a partir de quince conjuntos de datos generados mediante imputación múltiple, se centró en la evaluación de los **Residuales de Pearson** ($r_{P,i}$), definidos como:

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}(\hat{\mu}_i)}} \tag{21}$$

Este análisis resulta fundamental para verificar los supuestos de la modelización de tasas y proporciones. El examen de los gráficos de Residuales de *Pearson vs. Valores Ajustados* mostró un patrón de dispersión homogénea y aleatoria alrededor de la línea cero en las quince imputaciones. La ausencia de curvaturas o de patrones tipo "embudo" (indicativos de heterocedasticidad) confirma que los supuestos de **linealidad de la media** (μ_i) y de **homocedasticidad** (varianza constante o adecuadamente modelada) se cumplen de manera consistente. Este hallazgo respalda la estabilidad del parámetro de precisión (ϕ) y la adecuación funcional de la media estimada en cada modelo.

El análisis de los **Histogramas de Residuales de Pearson** refuerza esta consistencia. En todos los paneles, la distribución de los residuales presenta una forma aproximadamente simétrica y acampanada, con la media centrada en cero. Si bien la teoría de la Regresión Beta establece que solo los **residuales cuantil** ($r_{Q,i} = \Phi^{-1}(F(y_i; \hat{\mu}_i, \hat{\phi}))$) deberían seguir una distribución **Normal Estándar** $N(0,1)$, la similitud observada en los residuales de Pearson representa un resultado alentador. La uniformidad de las distribuciones entre imputaciones sugiere una estructura de errores estable y predecible, lo que confirma que el proceso de imputación generó réplicas de datos coherentes con la calidad diagnóstica del modelo original.

El **Gráfico Normal Q-Q** aplicado a los residuales agregados evidenció una excelente correspondencia entre los cuantiles muestrales y los teóricos en el rango central. Esta alineación indica una aproximación razonable a la normalidad, requisito clave para la validez de los residuales de devianza ($r_{D,i}$) o cuantil. No obstante, se observó una ligera desviación en las colas inferiores (negativas), lo cual sugiere la presencia de valores atípicos o una leve asimetría negativa en los errores extremos. A pesar de esta pequeña desviación, los supuestos de linealidad y homocedasticidad se mantienen sólidos.

2.5.2. Residuales Cuantil ($r_{Q,i}$)

El análisis de los Residuales Cuantil es una extensión fundamental en el diagnóstico de modelos de Regresión Beta, que permite evaluar de forma más rigurosa la adecuación del modelo a los datos observados. A diferencia de los residuales de Pearson, los residuales cuantil se construyen transformando las probabilidades acumuladas del modelo ajustado a la escala de una distribución Normal estándar, lo que facilita la evaluación de normalidad.

El residual cuantil para la observación i se define como:

$$r_{Q,i} = \Phi^{-1}(F(y_i; \hat{\mu}_i, \hat{\phi})) \tag{22}$$

donde:

- $\Phi^{-1}(\cdot)$ es la función inversa de la distribución Normal estándar.
- $F(y_i; \hat{\mu}_i, \hat{\phi})$ corresponde a la función de distribución acumulada (CDF) de la distribución Beta evaluada en y_i, con parámetros $\hat{\mu}_i$ (media estimada) y $\hat{\phi}$ (precisión).

En un modelo bien especificado, los $r_{Q,i}$ deberían seguir aproximadamente una distribución Normal estándar $N(0, 1)$. Esto implica que:

- La media de los residuales debe ser cercana a 0.
- La dispersión de los residuales debe estar en torno a 1.
- En el gráfico Q-Q, los puntos deben alinearse con la línea de referencia.

2.5.3. Análisis de Influencia (Distancia de Cook)

El Análisis de Influencia, utilizando la Distancia de Cook, es una herramienta esencial para evaluar la estabilidad y robustez de los modelos de Regresión Beta ajustados a partir de la imputación múltiple. La Distancia de Cook mide el impacto agregado de cada observación sobre el conjunto de estimaciones de los coeficientes de regresión. Un valor bajo indica que la eliminación de esa observación particular apenas alteraría el modelo, mientras que un valor alto señala una observación influyente que podría estar distorsionando los resultados. Este diagnóstico es crucial para asegurar que la inferencia no dependa de un grupo de puntos de datos atípicos.

3. Objetivos

3.1. Objetivo general

Determinar una metodología para el tratamiento de valores faltantes en las covariables cuantitativas de un modelo de regresión beta que permita el modelamiento de pobreza multidimensional en los municipios de Colombia en el año 2018.

3.2. Objetivos específicos

1. Implementar un método de imputación para variables de proporción a partir de las cuales se construye la medida de pobreza multidimensional.

2. Ajustar un modelo de regresión beta para la tasa de pobreza multidimensional de los municipios de Colombia, empleando covariables de tipo exógenas que presente mayor frecuencia de reporte.
3. Calcular las medidas de severidad asociadas a la estimación de los parámetros del modelo debidas a los datos faltantes.
4. Realizar el diagnostico del modelo de regresión ajustado, mediante el error cuadrático medio (ECM), el coeficiente de determinación (R^2) y el análisis de residuales y de datos influyentes.

4. Metodología

A continuación, se realiza la descripción del enfoque propuesto junto con las posibles variables de interés que consolidarán la base de datos a trabajar y que permitirán realizar el modelamiento de la pobreza multidimensional. De igual forma se expondrán las etapas que fundamentan el cumplimiento de los objetivos propuestos con los resultados esperados, así como el cronograma de actividades.

4.1. Tipo de estudio

En consideración de la naturaleza del estudio se realiza un enfoque explicativo y correlacional, donde se busca modelar la medida de pobreza multidimensional en los municipios de Colombia en el año 2018 a partir de diversas variables cuantitativas, de origen exógeno, que presenten mayor frecuencia de reporte y cuyo acceso a la información sea más asequible para el establecimiento de políticas publicas a nivel local. Así mismo, es fundamental el establecimiento de un tratamiento para cuando las variables seleccionadas presentan valores faltantes, de modo que se pueda brindar un adecuado manejo que reconozca el comportamiento de cada covariable elegida.

4.2. Variables de interés en el estudio

Desde 2018, el DANE elabora el Índice de Pobreza Multidimensional (IPM) que ofrece una visión amplia de la pobreza, abordando múltiples facetas y complementando las evaluaciones basadas en ingresos. En lugar de basarse únicamente en el poder adquisitivo, se enfoca en diversas dimensiones que reflejan la calidad de vida y el bienestar de las personas. Es por ello que se toma el IPM (de fuente del Censo Nacional de Población y Vivienda) como variable respuesta para el modelo propuesto. El valor del IPM refleja el grado de privación que experimenta un hogar en diversas dimensiones de bienestar. Se calcula en función de las privaciones en las diferentes variables y dimensiones que componen el índice.

Al tratarse de una aproximación de la medición de la pobreza multidimensional en Colombia a nivel municipal, la medida cuenta con cinco dimensiones (condiciones educativas del hogar, condiciones de la niñez y juventud, salud, trabajo, acceso a servicios públicos domiciliarios y condiciones de la vivienda) y 15 indicadores. Cada dimensión tiene un peso del 20 por ciento y los indicadores cuentan con el mismo peso dentro de su dimensión respectiva. En este caso, se consideran como pobres a los hogares que tengan privación en por lo menos el 33,3 por ciento de los indicadores. Por lo anterior, es posible desglosar el puntaje del IPM para entender en qué dimensiones específicas se encuentran las privaciones de un hogar, lo que permite identificar áreas prioritarias de intervención. Al comparar los puntajes del IPM a lo largo de los años, se pueden identificar tendencias y evaluar el impacto de políticas públicas en la reducción de la pobreza multidimensional.

Respecto a las covariables cuantitativas consideradas en el estudio, inicialmente se han considerado variables exógena asociadas a las Necesidades Básicas Insatisfechas (NBI) en cuanto a dimensiones de coberturas de servicios básicos, educación, salud y seguridad; de igual forma, se considera el impacto del déficit habitacional, el valor agregado municipal y las densidades poblaciones a nivel urbano y rural. A continuación, se brinda un definición de las covariables preliminarmente consideradas:

Densidad poblacional: medida que expresa la cantidad de personas que en promedio habitan por unidad de superficie, usualmente por kilómetros cuadrados. Para su cálculo se divide el número total de población de una unidad geográfica o administrativa entre la cantidad total de superficie, usualmente expresada en Kilómetros cuadrados. La fuente de información esta dada por el Censo Nacional de Población y publicada por el Instituto Nacional de Estadística (INEC) para el año 2018.

Déficit habitacional (DEFVIV): este indicador determina el total y la proporción de hogares con carencias habitacionales según los criterios definidos en paredes, pisos, tipo de vivienda y número de hogares en vivienda. Para ello se estiman los déficits cuantitativo, que surge de la comparación entre el número de hogares y viviendas apropiadas existentes, y cualitativo, las viviendas con deficiencias en pisos, espacio (hacinamiento) y acceso a servicios públicos domiciliarios. La fuente de estos datos es emanada por el DANE para el año 2018.

Medida de desempeño municipal (MDM): esta medición permite orientar la toma de decisiones en torno a las políticas públicas tanto a nivel nacional como territorial y las decisiones de inversión y gestión en las Entidades Territoriales. Este indicador tiene como objetivo medir el desempeño de las entidades territoriales entendido como: la capacidad de gestión y de generación de resultados de desarrollo, teniendo en cuenta las condiciones iniciales de los municipios, como instrumento para el fortalecimiento de las capacidades territoriales, y la inversión orientada a resultados. En este sentido, se centrará en el componente de gestión está compuesto por cuatro dimensiones y 12 indicadores, los cuales miden la capacidad de las entidades territoriales para: 1) generar recursos propios que se traduzcan en inversión (movilización de recursos propios); 2) ejecutar los recursos de las fuentes de financiamiento de acuerdo con su presupuesto, planeación o asignación inicial (ejecución de recursos); 3) atender al ciudadano y presentar la rendición de cuenta de cuentas de las administraciones locales (gobierno abierto y transparencia) y 4) la utilización de los instrumentos de ordenamiento territorial para el recaudo local y la efectiva organización de la información (gestión de instrumentos de ordenamiento territorial). La fuente de estos datos es emanada por el Departamento Nacional de Planeación (DNP) para el año 2018.

Necesidades Básicas Insatisfechas (NBI): busca determinar, con ayuda de algunos indicadores simples, si las necesidades básicas de la población se encuentran cubiertas. Los grupos que no alcancen un umbral mínimo fijado, son clasificados como pobres. Los indicadores simples seleccionados, son: viviendas inadecuadas, Viviendas con hacinamiento crítico, Viviendas con servicios inadecuados, Viviendas con alta dependencia económica, Viviendas con niños en edad escolar que no asisten a la escuela. La fuente de estos datos es tomada del Información del Censo nacional de población y vivienda 2018 y publicada por DANE.

Tasa de cobertura de servicios básicos: este indicador muestra el porcentaje de hogares con disponibilidad de servicios básicos (agua potable, saneamiento, electricidad, gas natural, aseo) en la vivienda. Cuanto mayor sea el número de hogares que cuente con acceso a estos servicios mayor será su calidad de vida y menor la exposición a riesgos sanitarios por el consumo de agua tratada, así como por la eliminación de las excretas. La fuente de estos datos son los registros del año 2018 encontrados en el aplicativo TerriData creado por el Departamento Nacional de Planeación (DNP).

Tasa de cobertura en internet: variable que se encuentran en unidad de medida porcentaje denotando el total de hogares que cuenta con conexión de internet vía cable o satelital. Información publicada por el DANE a partir del Censo del año 2018.

Tasa de cobertura en seguridad: variable que se encuentran en unidad de medida porcentaje que denota percepción seguridad en el municipio entorno a los componentes de hurto, homicidios y violencia intrafamiliar. Es tomada con base en información publicada por el DNP de la subdirección de Planeación Territorial (SPT) para el año 2018.

Tasa de cobertura en salud: variable que se encuentran en unidad de medida porcentaje denotando el total de personas vinculadas al sistema de salud. Es tomada con base en información publicada por el DNP para el año 2018.

Tasa de cobertura en educación: variable que se encuentran en unidad de medida porcentaje denotando el total de personas neto que se vinculan al sistema de educación. Es tomada con base en información publicada por el DNP para el año 2018.

Valor agregado municipal: medida expresada en unidad de porcentaje y refleja la participación y el excedente económico de producción que aporta cada municipio al departamento, pero no refleja el Producto Interno Bruto municipal. Es tomada con base en información publicada por el DANE para el año 2018.

4.3. Procedimiento

Para cumplir con los objetivos propuestos (general y específicos), este proyecto abordará la siguiente secuencia de etapas:

Etapa 1. Análisis exploratorio de la base de datos. Desde el punto de vista exploratorio se busca iniciar con una revisión detallada de la estructura inherente de la base de datos, tipificando cada variable según su naturaleza. Esta fase permite determinar el panorama metodológico ante la presencia de observaciones incompletas, identificando el patrón de comportamiento de los datos faltantes y el mecanismo que produjo la de pérdida de información.

Etapa 2. Estudio y tratamiento de datos faltantes. Seguidamente, el proyecto se enfocará en realizar la revisión y estudio de las diversas técnicas de tratamientos faltantes. Así, se procederá a probar diversas técnicas para este fin, centrando particular atención en los métodos basados en Random Forest, KNN, Mice (*Multivariate Imputation by Chained Equations*), *Predictive Mean Matching* (pmm) y regresión.

Etapa 3. Ajuste y evaluación del Modelo Beta. Una vez contrastados y probados los métodos de tratamiento de datos faltantes, el proyecto se propone realizar la modelación de la tasa de pobreza con cada una de las bases imputadas; esta modelación culminará con la comprobación de la calidad en las estimaciones arrojadas a partir del coeficiente de determinación, análisis de residuales y análisis de datos influyentes.

Etapa 4. Definición y justificación de la metodología. Finalmente, y con base en los resultados obtenidos en la etapa inmediatamente anterior, se procederá a definir la metodología más idónea para la imputación de datos faltantes en modelos de regresión beta con aplicación a la estimación de la tasa de pobreza en municipios de Colombia.

5. Hallazgos

En este apartado se busca exponer y analizar los resultados empíricos obtenidos de la aplicación de la metodología desarrollada para el tratamiento de la información faltante en las covariables. El proceso inicia con una etapa exploratoria de los datos, para luego analizar el comportamiento de los datos faltantes e implementar estrategias de tratamiento de dichos datos, para luego definir el modelamiento con la confirmación de la coherencia entre la estrategia de imputación y el Modelo de Regresión Beta, para posteriormente efectuar la validación de la solidez de los datos mediante el análisis de sus residuales.

El valor agregado del presente estudio esta orientado en la interpretación de cómo las variables explicativas inciden sobre la Tasa de Pobreza Multidimensional de los municipios de Colombia en 2018. Estos hallazgos permiten generar conocimiento aplicado de alto valor para el diseño y focalización de políticas públicas efectivas destinadas a la mitigación de la pobreza en el contexto nacional.

5.1. Sobre los datos

La base de datos utilizada en este análisis está compuesta por 1.122 registros, correspondientes a la totalidad de los municipios de Colombia, según la codificación oficial del DANE. Cada registro contiene información asociada a un municipio específico, e incluye tanto identificadores administrativos (código DANE del municipio, código del departamento, y nombre del municipio) como un conjunto de 18 variables socioeconómicas y de cobertura de servicios públicos, recopiladas para el año 2018.

Las variables cubren dimensiones relevantes para el análisis territorial y de bienestar, tales como la proporción del Índice de Pobreza Multidimensional (IPM), la densidad poblacional, y los porcentajes de población urbana y rural. También se incluyen indicadores de Necesidades Básicas Insatisfechas (NBI) en áreas urbanas y rurales, así como variables relacionadas con la cobertura de servicios esenciales.

Adicionalmente, se contempla información sobre la inversión en justicia y seguridad, el valor agregado municipal, el déficit habitacional y la Medida de Desempeño Municipal (MDM), lo cual permite un análisis integral de las condiciones de vida y el desarrollo institucional a nivel municipal. En atención del propósito del presente trabajo, la base de datos presenta algunos valores faltantes en ciertas variables, lo que motiva la aplicación de técnicas de análisis y tratamiento de datos incompletos como parte del presente estudio.

5.2. Análisis de datos faltantes

se emplearon las librerías mice y naniar del lenguaje R. Estas herramientas permiten detectar no solo el porcentaje de valores ausentes por variable, sino también los patrones de ausencia y una posible clasificación del mecanismo de omisión en las categorías estándar: MCAR (Missing Completely At Random), MAR (Missing At Random) y NMAR (Not Missing At Random).

5.2.1. Datos faltantes por variable

El análisis inicial reveló que varias variables presentan una proporción considerable de valores ausentes, concentrándose principalmente en los indicadores relacionados con la cobertura de servicios públicos: cobertura de internet, cobertura de gas natural, cobertura de energía eléctrica y cobertura de aseo como se observa a continuación:

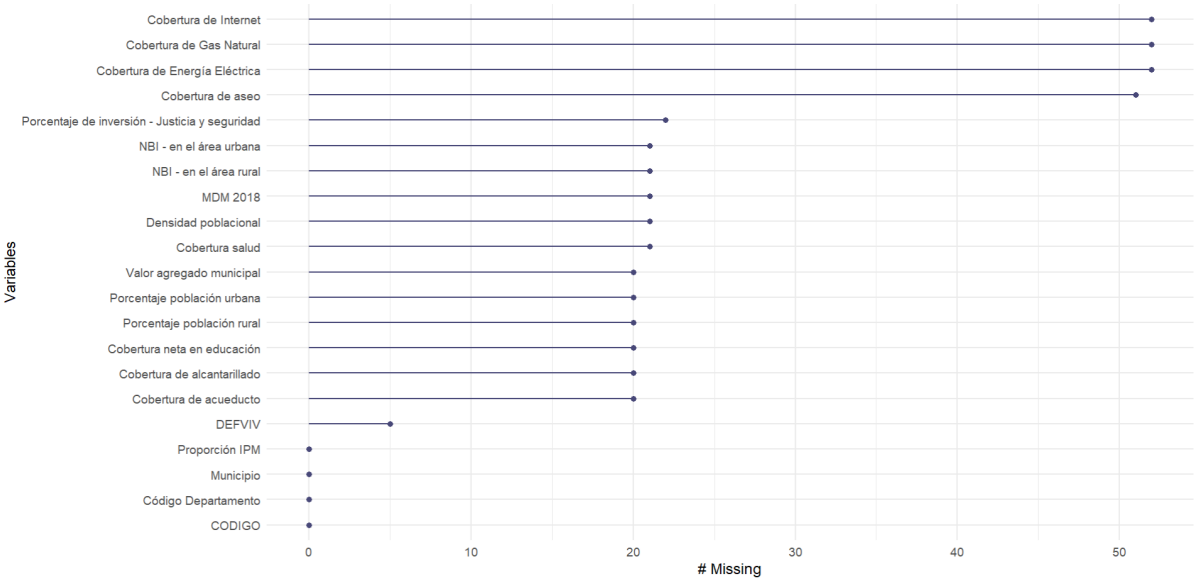

Figura 2. Cantidad de datos faltantes por variable

5.2.2. Patrón de datos faltantes

El análisis fue complementado mediante la función gg_miss_upset de la librería naniar, que permite identificar patrones específicos de combinación entre variables faltantes. En la figura siguiente se muestran los distintos grupos de observaciones que comparten estructuras comunes de omisión.

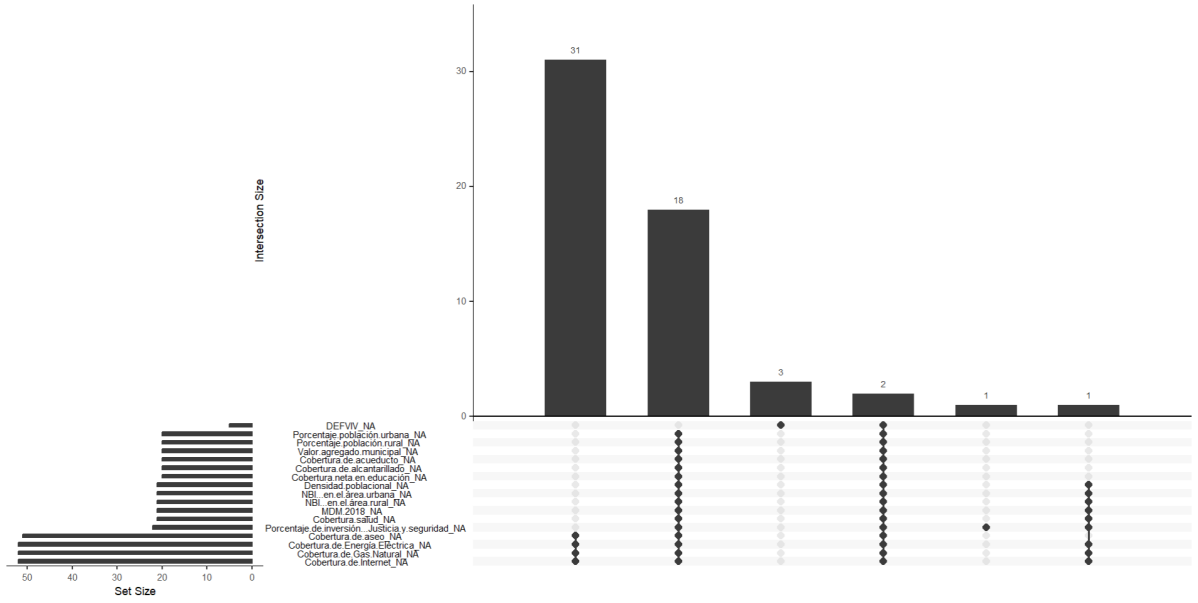

Figura 3. Patrón de datos faltantes

La figura 3, que representa un Diagrama UpSet, ofrece una visión de las intersecciones en la ausencia de datos (NA) entre las covariables; de la misma forma, la barra izquierda (el Set Size) confirma lo esperado: las variables relacionadas con Cobertura de Internet y Cobertura de Gas Natural son, individualmente,

las más problemáticas en términos de reporte.

Sin embargo, el verdadero valor del gráfico reside en la sección de intersecciones superiores; donde se observa que la ausencia de datos no es un evento fortuito, de hecho, el patrón más frecuente y simple (que afecta a 31 municipios) ocurre cuando únicamente falta la Cobertura de Internet, gas natural, cobertura de energía eléctrica y aseo sugiriendo un problema de disponibilidad de esas fuentes.

Aunado a ello, se evidencia un segundo pico, que involucra a 18 municipios. En este grupo, la falta de datos es compartida por 16 variables cruciales, incluyendo indicadores de NBI y servicios básicos; por lo que este patrón estructurado y bien definido sugiere descartar la idea de que los datos sean faltantes completamente al azar (MCAR).

La evidencia apunta a una falla sistémica en el reporte, posiblemente ligada a las capacidades administrativas de ciertos municipios y que posibilemente esta relacionada con una clara dependencia entre la pérdida de datos y el tipo de indicador, lo cual justifica la necesidad de aplicar un método avanzado de imputación múltiple, buscando preservar la varianza y la incertidumbre en las estimaciones finales del Modelo de Regresión Beta. Finalmente, el gráfico confirma que la gran mayoría de los datos están completos (98,1 % de los valores disponibles) y solo un 1,9 % de los datos están ausentes, lo que indica un bajo nivel general de omisión.

5.2.3. Test de Little

Para complementar el análisis visual y descriptivo de los datos faltantes, se aplica el Test de Little para MCAR sobre las variables numéricas y sobre el conjunto completo de datos. Este test estadístico evalúa si la ausencia de datos puede considerarse completamente aleatoria (MCAR), es decir, independiente tanto de las variables observadas como de las no observadas.

El p-valor, siendo menor a 0.05, con 60 grados de libertad y estadístico de 1083 indica que se rechaza la hipótesis nula de que los datos faltantes son completamente aleatorios (MCAR). Esto implica que el mecanismo de omisión en la base de datos no es MCAR, y que la probabilidad de que un dato esté ausente depende de alguna característica observada o no observada.

Con base en este resultado, se concluye que el patrón de datos faltantes corresponde a mecanismos de tipo MAR (Missing At Random) o NMAR (Not Missing At Random), siendo más probable el primero dado el análisis exploratorio previo y el contexto de las variables.

6. Estudio y tratamiento de datos faltantes

Con base en los hallazgos presentados en el apartado anterior, en los que se concluyó que el mecanismo de omisión no es MCAR (Missing Completely At Random), sino más probablemente MAR (Missing At Random), se procede a aplicar y estudiar diferentes técnicas de imputación, ajustadas en un modelo de regresión beta, para luego ser evaluadas mediante las medidas de severidad descritas por las reglas de Rubin para datos faltantes con el fin de restaurar el conjunto de datos.

6.1. Modelo de regresión

En el modelo de Regresión, los resultados derivados de las reglas de Rubin muestran una baja incidencia de la imputación múltiple sobre la varianza de las estimaciones, lo que refleja una buena estabilidad estadística del modelo.

Variable	Estimación	RIV	Lambda	FMI
(Intercepto)	−0.4154798	0.37	0.27	0.28
Cobertura Internet	−0.0078746	0.20	0.17	0.17
Densidad poblacional	−0.0000473	0.14	0.13	0.13
% Población urbana	−0.0017531	0.26	0.20	0.21
NBI urbana	0.0059630	0.67	0.40	0.42
NBI rural	0.0138017	0.32	0.25	0.25
Valor agregado municipal	−0.0004533	1.86	0.65	0.67
DEFVIV	0.9761495	0.28	0.22	0.23
Cobertura acueducto	−0.0022429	0.21	0.17	0.18
Cobertura alcantarillado	0.0019874	0.12	0.11	0.11
Cobertura Energía Eléctrica	−0.0035941	1.32	0.57	0.59
Cobertura aseo	−0.0046234	0.33	0.25	0.25
Cobertura Gas Natural	−0.0006511	0.06	0.06	0.06
Cobertura en educación	−0.0017992	0.30	0.23	0.24
Inversión Justicia y seguridad	−0.9554231	0.10	0.09	0.09
MDM	−0.3052139	0.11	0.10	0.10
Cobertura salud	0.1906990	0.22	0.18	0.18
(phi)	55.3814744	0.33	0.25	0.25

Tabla 1. Resultados del Modelo de Regresión

Los valores del aumento relativo en la varianza (RIV) son en su mayoría reducidos, con un rango entre 0.06 y 1.86, lo que indica que el incremento de la varianza debido al proceso de imputación fue mínimo en casi todas las variables. Únicamente valor agregado municipal (1.86) y cobertura de energía eléctrica (1.32) registran incrementos más notables, aunque se mantienen dentro de niveles moderados que no comprometen la consistencia del modelo.

El parámetro (lambda) respalda este comportamiento, sus valores oscilan entre 0.06 y 0.65, con un promedio aproximado de 0.25. Esto significa que solo una cuarta parte de la varianza total de las estimaciones se atribuye a la información faltante imputada, mientras que la mayor parte proviene de la variabilidad intrínseca de los datos originales. La Fraction of Missing Information (FMI) presenta un patrón similar, con valores que van de 0.06 a 0.67, concentrándose la mayoría por debajo de 0.30. Tales cifras indican que la pérdida de información debida a la imputación es baja, y que las estimaciones conservan un alto grado de precisión y fiabilidad.

6.2. Modelo knn

En el modelo KNN, los indicadores derivados de las reglas de Rubin evidencian una mayor influencia de la imputación múltiple sobre la varianza de las estimaciones en comparación con otros modelos como se puede evidencia a continuación:

Variable	Estimación	RIV	Lambda	FMI
(Intercepto)	−1.1657025	15.13	0.94	0.94
Cobertura Internet	−0.0072488	14.53	0.94	0.94
Densidad poblacional	0.0000089	32.71	0.97	0.97
% Población urbana	−0.0020249	11.22	0.92	0.93
NBI urbana	0.0011689	4.91	0.83	0.85
NBI rural	0.0112625	4.37	0.81	0.83
Valor agregado municipal	0.0077684	26.37	0.96	0.97
DEFVIV	1.7864337	30.59	0.97	0.97
Cobertura acueducto	−0.0007879	10.87	0.92	0.92
Cobertura alcantarillado	0.0037244	7.34	0.88	0.89
Cobertura Energía Eléctrica	−0.0021234	2.85	0.74	0.76
Cobertura aseo	−0.0041451	6.00	0.86	0.87
Cobertura Gas Natural	−0.0010795	1.57	0.61	0.63
Cobertura en educación	−0.0022956	5.61	0.85	0.86
Inversión Justicia y seguridad	−0.8797645	1.98	0.66	0.68
MDM	−0.0922796	7.83	0.89	0.90
Cobertura salud	0.1731531	2.51	0.71	0.73
(phi)	40.9260589	23.73	0.96	0.96

Tabla 2. Resultados del Modelo knn

Los valores del aumento relativo en la varianza (RIV) presentan una amplia dispersión, con cifras que van desde 1.57 hasta 32.71. Este rango refleja que la imputación introdujo un grado considerable de incertidumbre en la mayoría de las variables, especialmente en aquellas con valores superiores a 20, como densidad poblacional (32.71), DEFVIV (30.59) y valor agregado municipal (26.37). Tales niveles sugieren que las estimaciones de estas variables son altamente sensibles a la variabilidad entre imputaciones, lo que implica menor estabilidad estadística. En contraste, variables como Cobertura de gas natural (1.57) y Cobertura de energía eléctrica (2.85) presentan un incremento moderado, indicando una mayor robustez de sus estimaciones.

Los valores de la fracción de varianza atribuible a la información faltante (lambda) confirman esta tendencia, situándose en su mayoría por encima de 0.80 y alcanzando hasta 0.97 en algunas variables. Esto significa que entre el 80 % y el 97 % de la varianza total de las estimaciones proviene de la información imputada, lo que refleja una fuerte dependencia del proceso de imputación para la obtención de los coeficientes. De manera similar, el FMI (Fraction of Missing Information) muestra valores elevados, con un rango de 0.61 a 0.97, lo que indica que prácticamente todas las estimaciones del modelo están afectadas en un grado importante por la presencia de datos faltantes.

6.3. Modelo Random Forest

En el modelo Random Forest, los indicadores derivados de las reglas de Rubin permiten evaluar el impacto de la imputación múltiple sobre la estabilidad de las estimaciones como se muestra a continuación:

Variable	Estimación	RIV	Lambda	FMI
(Intercepto)	−0.4434599	1.03	0.51	0.52
Cobertura Internet	−0.0072546	0.44	0.31	0.32
Densidad poblacional	−0.0000369	0.10	0.09	0.09
% Población urbana	−0.0018872	1.08	0.52	0.54
NBI urbana	0.0052417	3.30	0.77	0.78
NBI rural	0.0133399	1.46	0.59	0.61
Valor agregado municipal	−0.0022365	1.72	0.63	0.65
DEFVIV	1.1185123	1.86	0.65	0.67
Cobertura acueducto	−0.0017752	0.67	0.40	0.41
Cobertura alcantarillado	0.0009914	1.59	0.61	0.63
Cobertura Energía Eléctrica	−0.0038483	1.89	0.65	0.67
Cobertura aseo	−0.0038517	1.24	0.55	0.57
Cobertura Gas Natural	−0.0005411	0.56	0.36	0.37
Cobertura en educación	−0.0021448	1.15	0.54	0.55
Inversión Justicia y seguridad	−1.0024323	0.52	0.34	0.35
MDM	−0.2704303	0.33	0.25	0.25
Cobertura salud	0.1678917	0.68	0.41	0.42
(phi)	46.1664937	2.49	0.71	0.73

Tabla 3. Resultados del Modelo Random Forest

Los valores del aumento relativo en la varianza (RIV) muestran una variación entre 0.09 y 2.50, lo que indica que, aunque en general la imputación introdujo un nivel moderado de incertidumbre, algunas variables experimentaron una influencia más marcada del proceso. En particular, las variables Cobertura de acueducto (2.33) y Cobertura de energía eléctrica (2.50) presentan los incrementos más altos, lo que sugiere una mayor sensibilidad a los datos imputados. Por el contrario,variables como densidad poblacional (0.09) y cobertura de Internet (0.23) muestran una variabilidad mínima, reflejando estimaciones más estables.

En cuanto a la fracción de varianza atribuible a la información faltante (lambda), los valores oscilan entre 0.09 y 0.71. Esto indica que, para la mayoría de las variables, la incertidumbre asociada a los datos faltantes no supera el 50 % de la varianza total,aunque variables como NBI rural (0.63) y cobertura eléctrica (0.71) evidencian una mayor dependencia del proceso de imputación. De manera similar, los valores del FMI (Fraction of Missing In formation) se mantienen en un rango comparable (0.09–0.73), reforzando la idea de que la información imputada tuvo un peso moderado, pero no despreciable,en el cálculo de los coeficientes.

En conjunto, estos resultados sugieren que el modelo Random Forest conserva una adecuada estabilidad estadística tras la imputación múltiple, con niveles de incertidumbre manejables en la mayoría de las variables. No obstante, los valores elevados de RIV y FMI en algunas variables vinculadas al acceso a servicios básicos y a las condiciones rurales advierten la necesidad de interpretar sus efectos con cautela, considerando que parte de su variabilidad podría estar influenciada por la proporción de datos imputados.

6.4. Modelo pmm

En el modelo pmm, los indicadores derivados de las reglas de Rubin evidencian un nivel bajo de incertidumbre asociada a la imputación múltiple, lo que sugiere que las estimaciones obtenidas son estadísti-

camente estables y poco afectadas por los datos faltantes.

Variable	Estamación	RIV	Lambda	FMI
(Intercepto)	−0.3949127	0.47	0.32	0.33
Cobertura Internet	−0.0078325	0.12	0.11	0.11
Densidad poblacional	−0.0000472	0.07	0.06	0.07
% Población urbana	−0.0016074	0.24	0.20	0.20
NBI urbana	0.0057036	0.81	0.45	0.46
NBI rural	0.0137559	0.33	0.25	0.25
Valor agregado municipal	−0.0006704	0.88	0.47	0.48
DEFVIV	0.9841300	0.11	0.10	0.10
Cobertura acueducto	−0.0023195	0.29	0.23	0.23
Cobertura alcantarillado	0.0019090	0.10	0.09	0.09
Cobertura Energía Eléctrica	−0.0038566	1.11	0.53	0.54
Cobertura aseo	−0.0046603	0.14	0.12	0.12
Cobertura Gas Natural	−0.0006479	0.18	0.15	0.16
Cobertura en educación	−0.0018700	0.26	0.21	0.21
Inversión Justicia y seguridad	−0.8747598	0.43	0.30	0.31
MDM	−0.2942434	0.24	0.19	0.20
Cobertura salud	0.2002173	0.28	0.22	0.22
(phi)	55.0361593	0.36	0.26	0.27

Tabla 4. Resultados del Modelo pmm

Los valores del aumento relativo en la varianza (RIV) se mantienen consistentemente bajos, variando entre 0.07 y 1.11, con la mayoría de las variables por debajo de 0.5. Esto implica que la imputación apenas incrementó la varianza de los estimadores, y que el proceso de imputación múltiple introdujo una mínima distorsión en los resultados. Solo la variable Cobertura de energía eléctrica (1.11) presenta un incremento ligeramente superior, aunque aún dentro de niveles aceptables de estabilidad.

El parámetro (lambda) refuerza esta conclusión, los valores oscilan entre 0.06 y 0.53, con un promedio cercano a 0.25, lo que significa que menos de una tercera parte de la incertidumbre total de las estimaciones proviene de la información imputada. Este patrón sugiere que la mayor parte de la varianza observada responde a la variabilidad real de los datos y no al efecto de la imputación.

De manera similar, el FMI (Fraction of Missing Information) presenta valores bajos, situándose mayoritariamente entre 0.09 y 0.54. En general, un FMI inferior a 0.30 se considera indicador de buena estabilidad y escasa pérdida de información; en este modelo, casi todas las variables cumplen esa condición, con excepción de la cobertura eléctrica, que alcanza un valor de 0.54.

7. Selección del modelo de imputación

La tarea de seleccionar el método de imputación más adecuado, dado la complejidad de los patrones faltantes identificados en las covariables de la Tasa de Pobreza Multidimensional, requiere ir más allá de la simple revisión visual. Con este fin, se ha adoptado el marco teórico de la imputación múltiple desarrollado por Rubin (1987), el cual permite colapsar las múltiples bases de datos completadas en un conjunto de métricas resumidas y unificadas. Específicamente, se ha calculado el Incremento Relativo de la Varianza (RIV), el Factor de Incremento de la Varianza debido a la falta de información (λ) y la Fracción de Información Faltante (FMI o γ) para cada método de imputación evaluado (Regresión Lineal, kNN, Random Forest y pmm). Estos indicadores multivariados son esenciales, pues no solo cuantifican la proporción de variabilidad en las estimaciones atribuible a la información faltante, sino que también orientan hacia el modelo que genera menor incertidumbre. La elección final, guiada por el método que minimice el RIV y las métricas asociadas, garantiza la robustez de las inferencias posteriores en el Modelo de Regresión Beta.

Método de Imputación	RIV (r)	Lambda (λ)	FMI (γ)
Regresión Lineal (Reg)	0.402	0.237	0.243
K-Nearest Neighbors (knn)	11.470	0.852	0.862
Random Forest (RF)	1.240	0.493	0.506
Predictive Mean Matching (pmm)	0.350	0.232	0.236

Tabla 5. Resumen de las Métricas Multivariadas de Rubin por Método de Imputación

El método de Predictive Mean Matching (pmm) representa una de las estrategias más consistentes y fiables para la imputación múltiple de datos faltantes, particularmente cuando se busca preservar la distribución empírica y la coherencia interna del conjunto original. A diferencia de los métodos completamente paramétricos, el pmm combina la capacidad predictiva de un modelo de regresión con un procedimiento de emparejamiento basado en valores observados, de modo que cada imputación proviene de casos reales con medias predichas similares [Little (1988); Rubin (1987)]. Esta combinación híbrida permite mantener la integridad de las relaciones multivariadas y evita la generación de valores artificiales o estadísticamente improbables, aspecto crucial en estudios explicativos.

Por lo anterior, los resultados obtenidos mediante las reglas de Rubin confirman la robustez del pmm, donde los valores de RIV, lambda y FMI se mantuvieron consistentemente bajos, evidenciando un impacto mínimo de la imputación en la varianza total y una excelente estabilidad entre imputaciones. Esto implica que los coeficientes estimados son poco sensibles al proceso de imputación, lo que refuerza la validez de las inferencias realizadas a partir del modelo final. En comparación con otros métodos, el pmm ofrece una mayor precisión imputacional sin incrementar de forma artificial la incertidumbre estadística, lo que se traduce en estimaciones más estables y confiables.

Además, estudios previos han demostrado que el pmm tiende a reproducir de manera más fiel la variabilidad natural de los datos, incluso en contextos con distribuciones asimétricas, valores atípicos o relaciones no lineales [Van Buuren y Groothuis-Oudshoorn (2011) & Morris et al. (2015)]. Su carácter parcialmente estocástico permite reflejar la incertidumbre inherente a la imputación, sin alterar la coherencia interna del modelo. Esto contrasta con enfoques puramente paramétricos —como la imputación por regresión lineal— que suelen subestimar la varianza y distorsionar las correlaciones entre variables [Seaman et al. (2012)].

En consecuencia, la elección del pmm en este estudio no solo se justifica por fundamentos teóricos, sino también por evidencias empíricas de estabilidad y consistencia. Al preservar las propiedades originales del conjunto de datos y minimizar la influencia del proceso de imputación sobre las estimaciones finales, el pmm garantiza que los resultados del modelo explicativo sean metodológicamente sólidos, comparables y estadísticamente realistas. Por estas razones, el pmm se consolida como la opción más adecuada entre

los métodos evaluados.

8. Límites frente al modelamiento de la pobreza en Colombia: Comparativa de modelos

El análisis econométrico de la pobreza ha transitado de la linealidad clásica hacia modelos que respeten la naturaleza acotada del Índice de Pobreza Multidimensional (IPM). Dado que el IPM es una proporción situada en el intervalo [0,1], su modelización enfrenta desafíos críticos como la **asimetría persistente** y la **heterocedasticidad no lineal**, donde la variabilidad de las privaciones cambia en función de su intensidad territorial. Históricamente, el modelo Tobit intentó gestionar estos límites mediante el supuesto de censura, pero el consenso académico contemporáneo ha demostrado que este enfoque es ineficiente para datos que no son truncados, sino proporciones naturales [Ramalho, Ramalho & Murteira (2011)].

En respuesta, la literatura especializada convergió hacia el Logit Fraccional, que ofrece robustez sin supuestos distributivos rígidos [Papke & Wooldridge (2011)]. No obstante, para capturar la compleja **heterogeneidad territorial** de Colombia, la Regresión Beta se ha consolidado como la opción superior, al permitir una parametrización flexible que modela de forma independiente la media y la precisión de la distribución [Cribari-Neto & Zeileis (2010); Ospina & Ferrari (2012)]. La siguiente tabla técnica fundamenta la selección de la herramienta con mayor capacidad para capturar las asimetrías propias del IPM.

Característica / Modelo	Regresión Beta	Logit Fraccional	Modelo Tobit
Soporte Teórico	Asume que la variable dependiente sigue una distribución Beta, la cual es flexible y está diseñada explícitamente para modelar proporciones continuas en el intervalo (0, 1).	Utiliza una distribución binomial o quasi-binomial y un enlace logit. Es robusto para proporciones que sí incluyen 0 y 1 exactos.	Asume una variable latente subyacente que está truncada o censurada en 0 y 1, lo cual es teóricamente incorrecto para una proporción observada como el IPM.
Manejo de Límites	Restricción estricta (0, 1). Requiere una transformación menor si los datos contienen 0 o 1 exactos.	Admite valores exactos de 0 y 1 (intervalo [0, 1]).	Asume que los datos fuera de los límites están "censurados" (no observados), lo cual no aplica a una proporción calculada del IPM.
Supuestos Estadísticos	Modela la media condicional y la dispersión (varianza) de la proporción de manera independiente. Maneja heterocedasticidad naturalmente.	La varianza está ligada a la media (inherente al GLM), lo que puede ser restrictivo si la dispersión del IPM es alta o variable.	Asume normalidad subyacente y homocedasticidad, supuestos que a menudo se violan con datos de proporciones.
Justificación para IPM Colombia	El IPM es una medida de brecha y severidad que usualmente resulta en proporciones continuas (0 ¡IPM ¡1). La distribución beta es el estándar metodológico para este tipo de variables en econometría y salud pública.	Es una alternativa viable si muchos municipios tienen exactamente 0 % o 100 % de pobreza, pero menos eficiente si todos los datos están en el rango intermedio.	Teóricamente débil para modelar el IPM, ya que la "pobreza" no es una variable latente censurada, sino una medida calculada de privaciones.

Tabla 6. Comparación teórica de modelos para el análisis del IPM

En consecuencia, la evidencia teórica y empírica presentada en la Tabla 1 justifica la elección del modelo de **regresión beta** como la herramienta superior para analizar el IPM en Colombia. La principal fortaleza de este enfoque radica en su alineación con la naturaleza probabilística de la variable dependiente, que

se adhiere a una distribución beta, permitiendo un ajuste flexible a la **asimetría persistente** y la **heterocedasticidad** [Cribari-Neto y Zeileis, 2010]. Esto contrasta fuertemente con las limitaciones del modelo Tobit, cuyos supuestos de normalidad y censura son teóricamente incompatibles con una medida de proporción, y supera la rigidez del Logit Fraccional en la modelización de la varianza. Al preservar las propiedades originales del conjunto de datos y modelar la dispersión de forma independiente, la regresión beta minimiza la influencia del proceso de estimación sobre las inferencias finales, lo que se traduce en coeficientes más estables y confiables para la toma de decisiones de política pública.

9. Modelo de regresión beta con imputadación pmm

Los resultados consolidados de la Regresión Beta aplicada a la proporción del Índice de Pobreza Multidimensional (IPM), utilizando estimaciones combinadas (*pooled*) tras la imputación múltiple, confirman que los factores estructurales y de acceso a servicios básicos son determinantes primordiales de la variabilidad del indicador.

Variable	Estimación	Error Estándar	Estadístico	GL	Valor p
(Intercepto)	−0.3949127	0.1683155	−2.3462650	115.4991800	0.0206670
Cobertura Internet	−0.0078325	0.0013172	−5.9462590	539.1246500	0.0000000
Densidad poblacional	−0.0000472	0.0000201	−2.3487700	788.6542400	0.0190818
% Población urbana	−0.0016074	0.0008861	−1.8139860	258.2930000	0.0708148
NBI urbana	0.0057036	0.0011372	5.0153830	62.8449400	0.0000046
NBI rural	0.0137559	0.0008941	15.3853900	180.6446000	0.0000000
Valor agregado municipal	−0.0006704	0.0019383	−0.3458662	57.8610000	0.7306977
DEFVIV	0.9841300	0.0905026	10.8740600	604.2673800	0.0000000
Cobertura acueducto	−0.0023195	0.0006932	−3.3458670	206.0703400	0.0009747
Cobertura alcantarillado	0.0019090	0.0007800	2.4472720	631.4293500	0.0146656
Cobertura Energía Eléctrica	−0.0038566	0.0014097	−2.7528130	46.0351400	0.0084269
Cobertura aseo	−0.0046603	0.0008620	−5.4063050	488.2066900	0.0000001
Cobertura Gas Natural	−0.0006479	0.0005433	−1.1924990	363.9864600	0.2338424
Cobertura en educación	−0.0018700	0.0007589	−2.4667930	237.1714600	0.0143439
Inversión Justicia y seguridad	−0.8747598	0.4717513	−1.8542820	127.3158500	0.0601276
MDM	−0.2942434	0.1437209	−2.0473250	264.4403900	0.0416143
Cobertura salud	0.2002173	0.0617008	3.2449720	218.2368000	0.0013593
Parámetro de Precisión (ϕ)	55.0361600	2.6893430	20.4645300	160.5451500	0.0000000

Tabla 7. Resultados del Modelo de Regresión Beta

El parámetro de precisión estimado, $\hat{\phi} = 55.036$, es notablemente alto, lo que se traduce en una dispersión relativamente baja de los datos alrededor de la media predicha μ_i. Este alto valor de ($\hat{\phi} > 1$) sugiere una buena capacidad de ajuste del modelo, indicando que las covariables seleccionadas logran capturar una parte significativa de la heterogeneidad en las proporciones de IPM observadas entre municipios.

En cuanto a los coeficientes de la media (μ_i), diversas variables socioeconómicas y de servicios ejercen una influencia estadísticamente significativa sobre la proporción de IPM. Se confirma que indicadores de privación como el Déficit de Vivienda (DEFVIV), con un coeficiente positivo de $\hat{\beta} = 0.9841$, y el NBI en el área rural ($\hat{\beta} = 0.0138$), están fuertemente asociados con un incremento en la proporción de IPM.

Estos resultados subrayan la persistencia de las necesidades básicas insatisfechas en las áreas rurales y la precariedad habitacional como motor clave de la pobreza multidimensional. De manera similar, la Cobertura en Salud presenta un coeficiente positivo ($\hat{\beta} = 0.2002$), lo cual, de forma contraintuitiva, podría reflejar un efecto de interacción o saturación donde municipios con alta cobertura aún enfrentan

retos de calidad y acceso que se traducen en una mayor medición de la pobreza en otras dimensiones del IPM.

Por otro lado, se identifican variables asociadas con una reducción en la proporción de IPM. Por ejemplo, una mayor Inversión en Justicia y Seguridad ($\hat{\beta} = 0.8748$) y el índice de MDM 2018 ($\hat{\beta} = 0.2942$) se asocian negativamente con el IPM. Estos hallazgos sugieren que una mejor gobernanza y mayores asignaciones de recursos a la seguridad impactan positivamente en la reducción de la pobreza. Las coberturas de servicios, como la Cobertura de Energía Eléctrica ($\hat{\beta} = 0.0039$), también muestran el signo negativo esperado, si bien su magnitud es pequeña, lo que sugiere que la infraestructura básica elemental tiene un efecto marginal en municipios donde ya está relativamente avanzada.

Finalmente, la robustez de las estimaciones se ve respaldada por los diagnósticos de la imputación múltiple. Los valores de la Fracción de Información Faltante (FMI) y el Incremento Relativo de Varianza (RIV), que son consistentemente bajos para la mayoría de las variables (RIV ≈ 0.03 para las variables estructurales), indican que la incertidumbre y la variabilidad explicada por el proceso de imputación son mínimas. Esto refuerza la estabilidad de las estimaciones ($\hat{\beta}$) y confirma que el análisis combinado de los modelos es altamente fiable, permitiendo una interpretación robusta de los principales determinantes del IPM.

10. Análisis de residuales del modelo de regresión beta

La validación de un modelo de regresión se considera incompleta sin una evaluación rigurosa de sus supuestos a través del análisis de residuales; por lo cual, los modelos diseñados para variables de respuesta limitadas en el intervalo $(0, 1)$, como la Regresión Beta, resultan insuficientes las metodologías tradicionales de residuales. Ello se debe a que la distribución de los errores en la Regresión Beta presenta una naturaleza heterocedástica y no gaussiana inherente a la distribución Beta [Ferrari y Cribari-Neto (2004)].

Por lo tanto, para lograr un diagnóstico fiable del modelo, se requiere la aplicación de residuales ajustados que transformen la estructura del error a un marco que se aproxime al comportamiento normal. Específicamente, se ha consolidado el uso de los residuales de cuantiles (quantile residuals), un método robusto propuesto por Dunn y Smyth (1996) y ampliamente adoptado en plataformas de cómputo estadístico contemporáneo [R Core Team (2024)]. La principal ventaja de este enfoque radica en que, bajo la hipótesis de un modelo correctamente especificado, los residuales resultantes se distribuyen como una Normal estándar.

Dicha propiedad facilita el uso de herramientas de diagnóstico gráfico convencionales (tales como los gráficos de probabilidad Normal) para evaluar de manera confiable la linealidad del predictor, la presencia de valores atípicos y la independencia de las observaciones. Con base en lo anterior, el diagnóstico subsiguiente se fundamenta en la inspección de estos residuales específicos para asegurar la validez y la solidez del modelo Beta.

10.1. Residuales de Pearson

En conjunto, los resultados del diagnóstico indican que los modelos de Regresión Beta ajustados sobre los conjuntos imputados son estadísticamente **robustos** y **consistentes**. La estabilidad de los residuales, la ausencia de patrones sistemáticos y la adecuada aproximación a la normalidad confirman que los datos imputados son apropiados para el análisis combinado y permiten realizar una **inferencia válida y confiable** de los coeficientes de regresión.

En cada uno de los 15 paneles, se observa que los residuales están distribuidos de manera aleatoria alrededor de la línea de referencia cero (la línea roja discontinua), sin mostrar patrones sistemáticos como curvaturas o la forma de embudo. Esta dispersión uniforme y centrada indica que se cumplen los supuestos clave de la modelización estadística, específicamente la linealidad y la homocedasticidad (varianza constante de los errores). Por lo tanto, el análisis sugiere que el modelo empleado para el proceso de imputación es estadísticamente válido y los conjuntos de datos resultantes son apropiados para el subsiguiente análisis combinado.

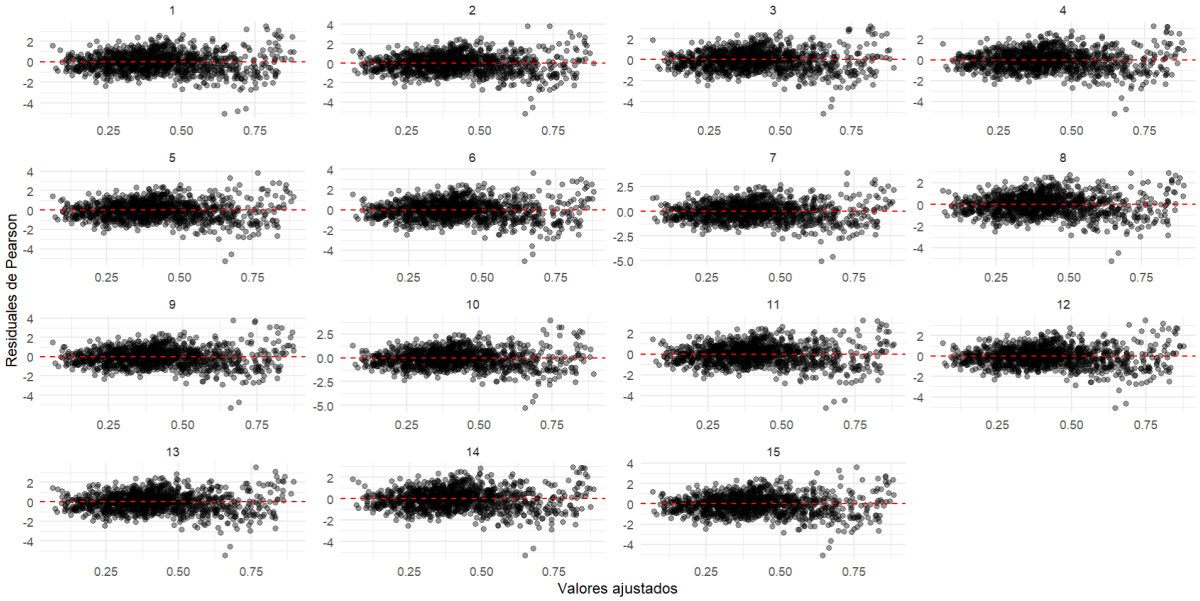

Figura 4. Residuales vs Ajustados por imputación

Complementando el diagnóstico de los valores ajustados, el análisis de la Distribución de Residuales de Pearson para cada una de las 15 imputaciones refuerza la validez de los modelos utilizados. En cada histograma (numerado del 1 al 15), se observa que la distribución de los residuales se aproxima a una forma acampanada y simétrica, lo que es indicativo de una distribución aproximadamente normal. La media de los residuales en todos los paneles, marcada por la línea discontinua roja, se encuentra consistentemente cerca de cero.

Este hallazgo es fundamental, ya que la normalidad de los errores es un supuesto estadístico clave para la correcta inferencia en los modelos de regresión. La consistencia en la forma de la distribución a través de todas las imputaciones sugiere que el mecanismo de imputación ha generado errores que son homogéneos y adecuadamente distribuidos, confirmando la robustez y la fiabilidad de los conjuntos de datos imputados para el análisis posterior.

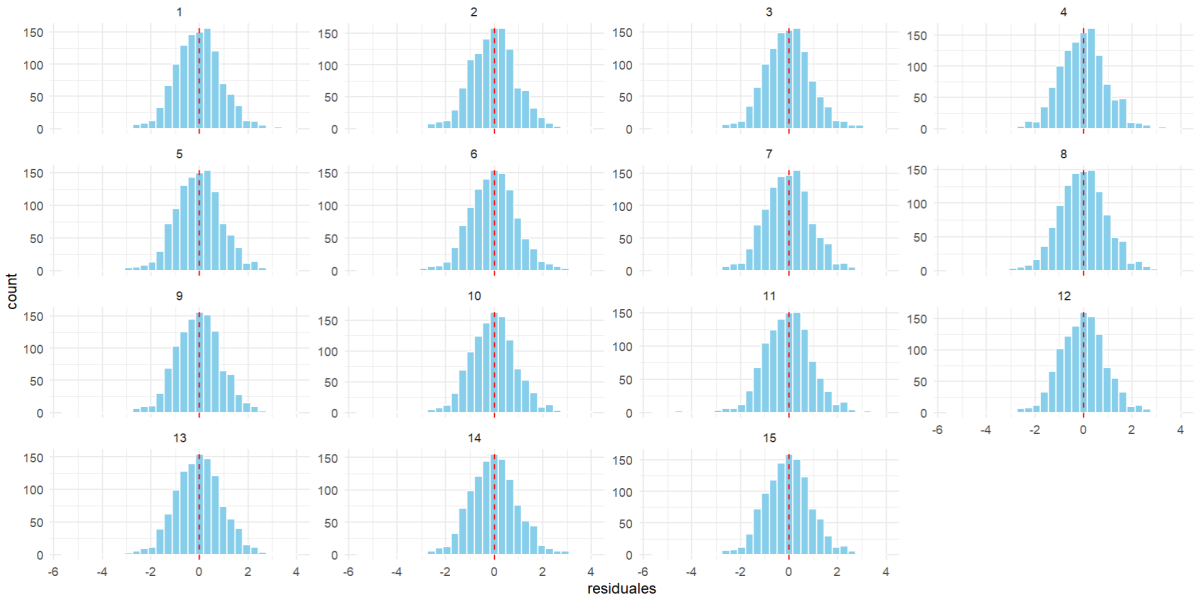

Figura 5. Distribución de residuales de Pearson por imputación

Finalmente, el Gráfico Normal Q-Q (Quantile-Quantile) proporciona una evaluación detallada del supuesto de normalidad de los residuales del modelo. En este gráfico, los cuantiles muestrales de los residuales (eje Y) se comparan con los cuantiles teóricos de una distribución normal estándar (eje X).

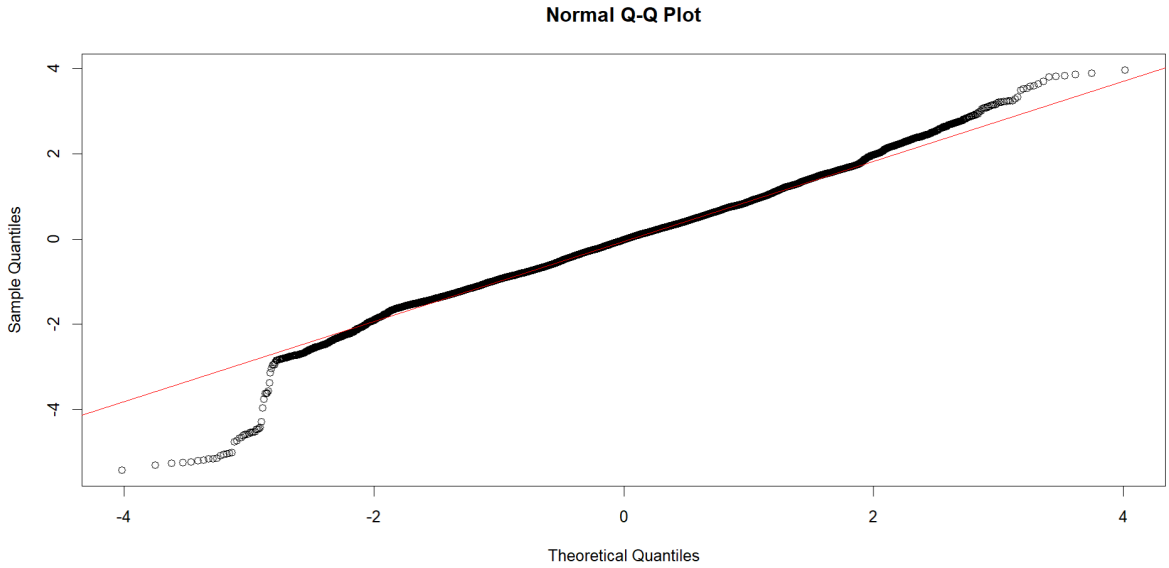

Figura 6. Gráfico Normal Q - Q

La mayoría de los puntos (los círculos negros) siguen estrechamente la línea recta de referencia (la línea roja), especialmente en el rango central, lo que es una fuerte evidencia de que los residuales se distribuyen aproximadamente de forma normal. No obstante, se observa una desviación notable en las colas inferiores (valores negativos), donde los puntos se separan de la línea recta. Esta separación indica que la distribución de los residuales presenta colas más pesadas o una asimetría leve en la parte inferior

de lo que se esperaría de una normal perfecta. Aunque la desviación en las colas merece consideración, la buena alineación en la región central sugiere que la asunción de normalidad es lo suficientemente adecuada para fines de inferencia, aunque se debe reconocer la posible presencia de valores atípicos o una ligera asimetría en los errores.

10.2. Residuales Cuantil

Si bien existe una excelente alineación con la línea recta de referencia en el rango central (cerca de cero), lo cual sugiere que la mayoría de los errores son adecuadamente modelados, se observan fuertes y preocupantes desviaciones en ambas colas. Específicamente, en la cola inferior (izquierda), los cuantiles muestrales se curvan notablemente hacia abajo, indicando una cola pesada o la presencia de valores atípicos (outliers) negativos extremos; de manera similar, en la cola superior (derecha), los puntos se curvan abruptamente hacia arriba, señalando la existencia de outliers positivos extremos.

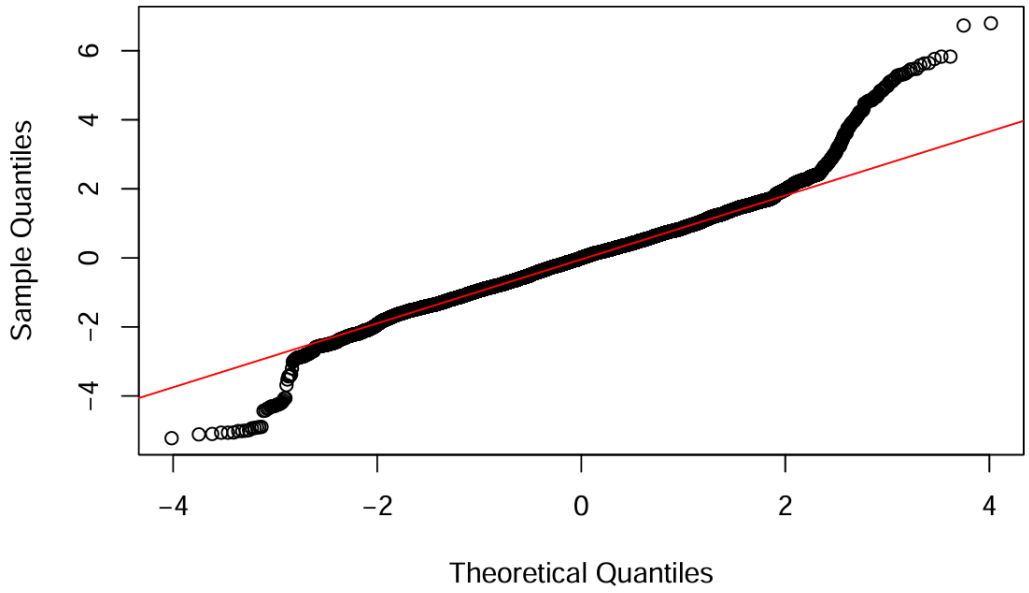

Figura 7. Gráfico QQ Normal (Residuales Cuantil Agregados)

El análisis de la distribución de Residuales Cuantil por imputación muestra un resultado diagnóstico aceptable para los modelos de Regresión Beta ajustados a los 15 conjuntos de datos imputados. En cada uno de los paneles, la distribución de los residuales se aproxima de manera muy cercana a la forma acampanada y simétrica esperada de una Distribución Normal Estándar N(0,1). La media (línea discontinua roja) está centrada consistentemente en cero, y la gran mayoría de los valores caen dentro del rango esperado de -2.5 a 2.5.

Esta uniformidad y consistencia a través de todas las imputaciones confirman que el supuesto fundamental de que los errores siguen una distribución normal se cumple. Este hallazgo es un fuerte indicativo de que el modelo está correctamente especificado para la distribución de la variable de respuesta, lo que a su vez valida la confiabilidad y la robustez de la inferencia combinada de los coeficientes de regresión.

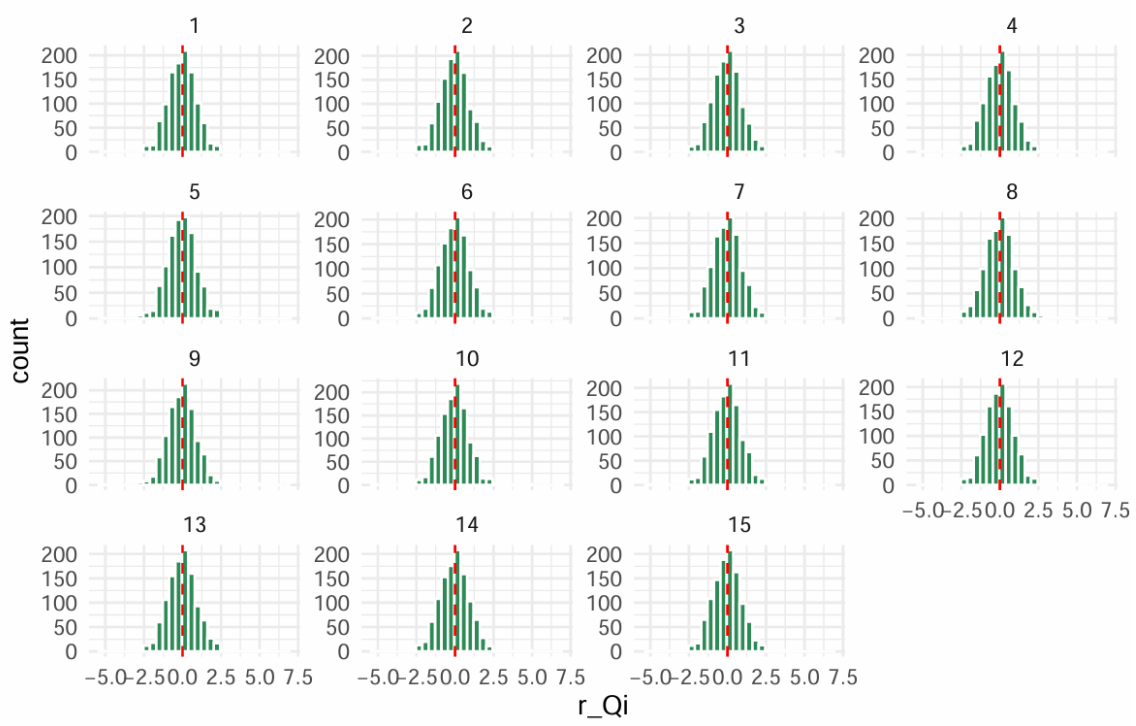

Figura 8. Distribución de Residuales Cuantil por Imputación

10.3. Análisis de Influencia (Distancia de Cook)

El gráfico de la Distancia de Cook por Observación y por Imputación indica que los modelos de Regresión Beta son robustos y estables ante la influencia de las observaciones individuales. La gran mayoría de los puntos en las 15 imputaciones se agrupan consistentemente cerca del eje cero y no superan un umbral de influencia que se consideraría problemático.

Los valores de la Distancia de Cook, que miden cuánto cambiarían las estimaciones de los coeficientes si se eliminara una observación, son mayoritariamente inferiores a 0.05.

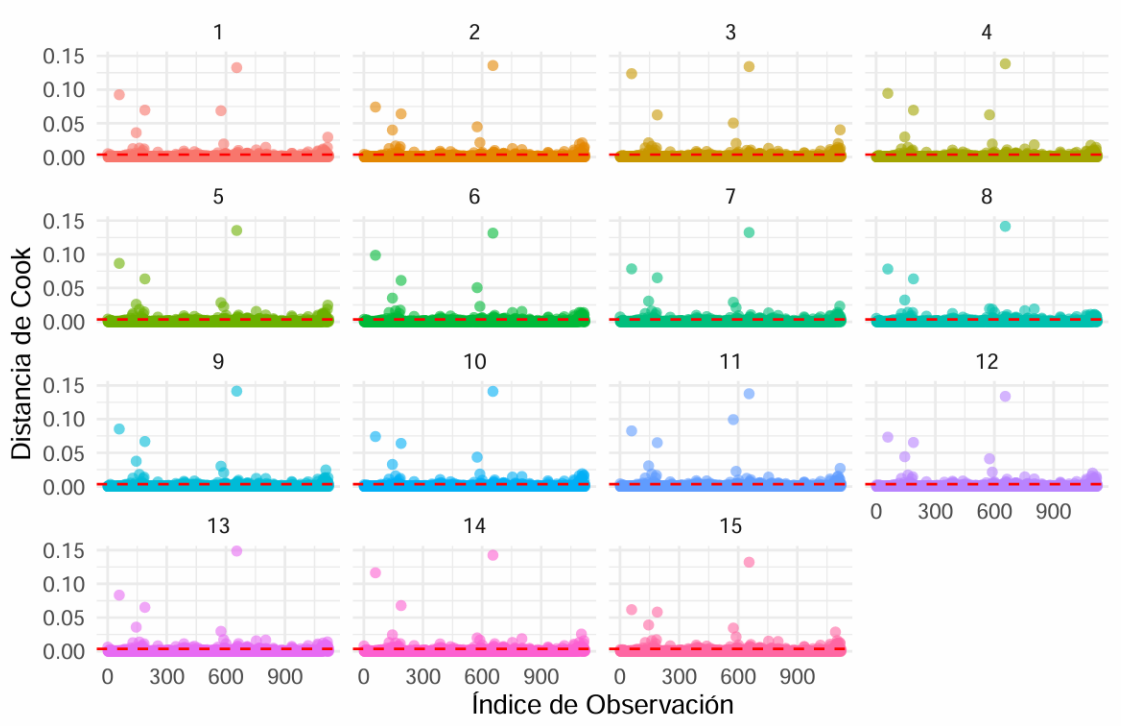

Figura 9. Distancia de Cook por Observación y por Imputación

No obstante, se identifican picos de influencia moderada (entre 0.05 y 0.15) en cada panel. Estos picos sugieren que algunas observaciones son sistemáticamente más influyentes que el promedio a través de los conjuntos de datos imputados. Aunque estos valores no son lo suficientemente altos como para requerir la eliminación o un ajuste del modelo, su consistencia valida el hallazgo de los outliers extremos observados en el Gráfico Q-Q Normal. En conjunto, este análisis confirma que, a pesar de la presencia de unos pocos valores atípicos, estos no ejercen una distorsión crítica en la estimación de los coeficientes, lo que respalda la confiabilidad del análisis combinado.

11. Municipios de interés

El análisis de la tabla que se presenta a continuación se centra en un grupo específico de 20 municipios que han exhibido las mayores porcentajes de datos faltantes (NA) en el conjunto de variables explicativas del modelo. Esta focalización no es casual; representa una validación crucial de la robustez del modelo de Regresión Beta frente a la imputación de datos faltantes a través del método Predictive Mean Matching (pmm). La tabla expone la Proporción del Índice de Pobreza Multidimensional (IPM) observada en cada municipio junto con la media predicha por el modelo, y los respectivos intervalos de confianza (límite inferior/límite superior).

El objetivo principal es evaluar si las predicciones del modelo se ajustan razonablemente a los valores reales del IPM, incluso en los casos donde la incertidumbre por la falta de información fue mayor. Una buena concordancia entre la proporción observada y la media predicha, con intervalos de confianza estrechos que capturen el valor real, servirá como un indicador de que el modelo y el proceso de imputación han logrado construir estimaciones fiables, permitiendo así dar plena trazabilidad y confianza a los resultados finales del estudio.

Municipio	Proporción IPM	Predichos	L. Inf	L. Sup	% NA
EL ENCANTO	0.775	0.78	0.71	0.85	0.67
LA CHORRERA	0.881	0.81	0.71	0.92	0.67
LA PEDRERA	0.909	0.81	0.68	0.94	0.67
LA VICTORIA	0.964	0.81	0.72	0.90	0.67
MIRITI - PARANA	0.912	0.81	0.70	0.92	0.67
PUERTO ALEGRIA	0.824	0.76	0.62	0.91	0.67
PUERTO ARICA	0.846	0.81	0.74	0.88	0.67
PUERTO SANTANDER	0.887	0.81	0.68	0.93	0.67
TARAPACA	0.800	0.77	0.68	0.85	0.67
BARRANCO MINAS	0.865	0.79	0.68	0.89	0.38
MAPIRIPANA	0.955	0.83	0.75	0.91	0.71
SAN FELIPE	0.883	0.81	0.71	0.91	0.67
PUERTO COLOMBIA	0.963	0.84	0.79	0.89	0.67
LA GUADALUPE	0.841	0.80	0.69	0.90	0.67
CACAHUAL	0.901	0.79	0.67	0.91	0.67
PANA PANA	0.985	0.85	0.81	0.89	0.67
MORICHAL	0.957	0.83	0.74	0.91	0.67
PACOA	0.984	0.85	0.79	0.91	0.67
PAPUNAUA	0.979	0.83	0.76	0.91	0.67
YAVARATE	0.948	0.84	0.80	0.89	0.67

Tabla 8. Tabla de Predicciones del Modelo de Regresión Beta (con imputación pmm)

11.1. Análisis comparativo

La tabla 7 muestra un alto y consistente indice de pobreza en los municipios (áreas no municipalizadas para 2018) seleccionadas. En este sentido, en la mayoría de los casos, la media predicha se agrupa en un rango estrecho (aproximadamente entre **0.79** y **0.85**). Este agrupamiento no implica un fallo del modelo, sino que refleja su capacidad para estabilizar la estimación en municipios donde la falta de información podría introducir alta variabilidad. El modelo, al ser robusto, tiende a predecir valores cercanos a la medida global condicionada por las variables disponibles. La clave interpretativa reside en los Intervalos de Confianza, lo que indica que el modelo logra estimar el nivel de pobreza de manera confiable.

Los Intervalos de Confianza (CI) del 95 % (`límite inferior` y `limite superior)` son notablemente amplios para algunos municipios (por ejemplo, LA PEDRERA, 0.68 a 0.94), lo que refleja una alta incertidumbre en la estimación en esas áreas, probablemente debido a la cantidad significativa de datos faltantes (`% NA)`, que en la mayoría de los casos es del 64 %. A pesar de esta incertidumbre, las predicciones en casi todos los municipios son coherentes con las proporciones observadas, lo que confirma la adecuación del modelo para la estimación de la pobreza a pequeñas áreas.

11.2. Error cuadrático medio (ECM), el coeficiente de determinación (R^2)

Respecto al valor del error cuadrático medio del modelo de regresión beta ajustado mediante el proceso de imputación pmm, arrojo un valor aproximado de 0.0036, lo que muestra que las diferencias entre los valores observados del IPM y los valores que el modelo predice son muy pequeñas. Es decir, que el modelo logra representar bastante bien la realidad de los municipios respecto al indice de pobreza multidimensional, ya que el nivel de error promedio es bajo.

Por otra parte, el coeficiente de determinación obtenido fue de 0.8618, lo que significa que el modelo explica alrededor del 86 % de la variabilidad observada en el Índice de Pobreza Multidimensional (IPM). En otras palabras, la mayoría de los cambios que se observan en los valores del IPM pueden ser entendidos a partir de las variables incluidas en el modelo. Lo anterior indica que el modelo tiene una buena capacidad para describir el comportamiento del IPM y que las variables seleccionadas aportan información relevante para entender las diferencias en los niveles de pobreza entre los municipios analizados.

12. Índice de pobreza multidimensional (IPM) municipal en Colombia

La comparación cartográfica entre los datos originales de la "Proporción IPM" y los valores predichos por el modelo refleja visualmente estos resultados estadísticos. Los patrones espaciales son coherentes debido a que las zonas de alta pobreza se concentran en el suroccidente, la Amazonía y el Caribe, mientras que las regiones del centro y oriente exhiben niveles menores. No obstante, el mapa de predicciones muestra una distribución suavizada, con menor dispersión que los valores observados, lo cual es consistente con la naturaleza del modelo Beta y el alto valor del parámetro $\hat{\phi}$. Esta suavización indica que el modelo capta correctamente las tendencias centrales y la estructura espacial de la pobreza, reduciendo la influencia de valores extremos y generando estimaciones más estables.

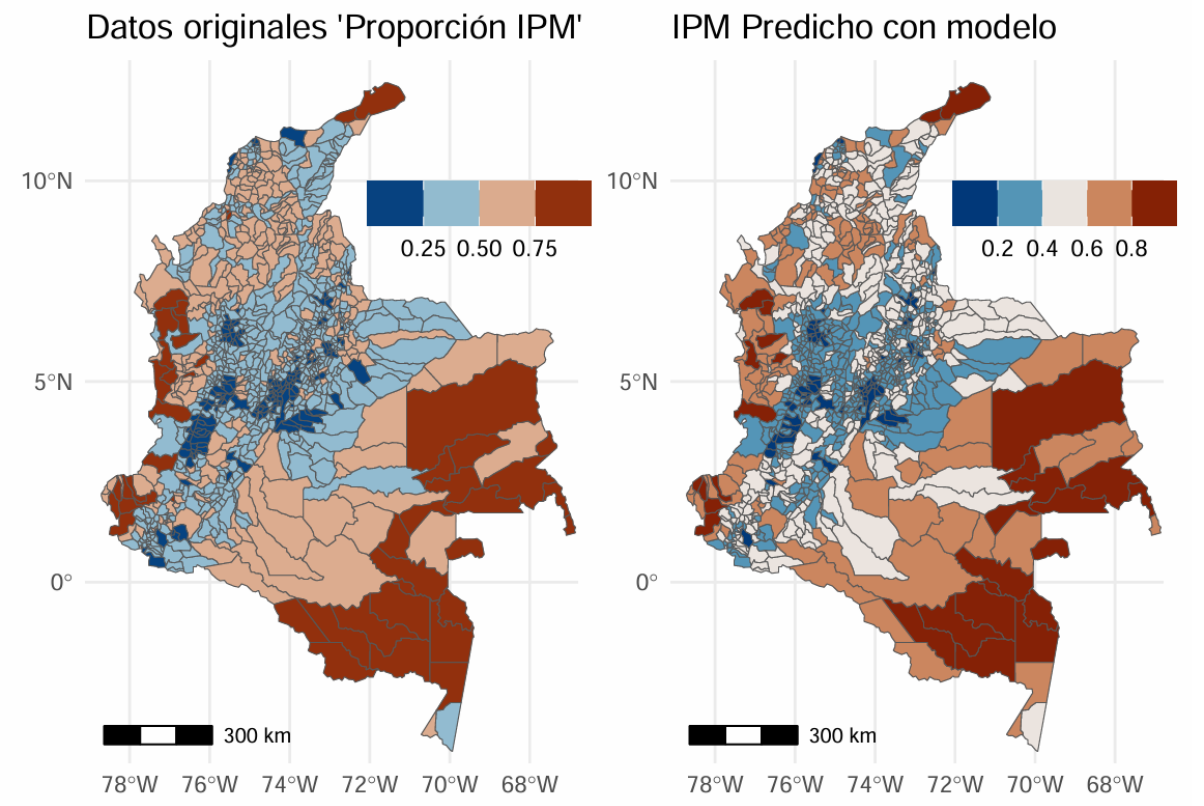

Figura 10. Mapa de calor del Índice de Pobreza Multidimensional (IPM) 2018

En conjunto, los resultados empíricos, los diagnósticos de residuales y la evidencia geográfica apuntan hacia un modelo de Regresión Beta robusto, bien calibrado y estadísticamente sólido. El proceso de imputación múltiple garantizó la validez inferencial y preservó la estructura de los datos originales, mientras que la capacidad del modelo para reproducir los patrones espaciales del IPM confirma su pertinencia para el análisis territorial de la pobreza multidimensional en Colombia. Por tanto, los valores predichos pueden considerarse una representación fiable y generalizable del fenómeno, apta para orientar políticas públicas focalizadas y análisis comparativos entre municipios.

13. Consideraciones estadísticas finales

13.1. Comparativas entre las funciones de enlace

En el desarrollo de una regresión beta, la elección de la función de enlace $g(\mu)$ no es una decisión meramente procedimental, sino una determinación sobre la geometría de la relación entre los determinantes sociales y la variable dependiente; en este caso, el Índice de Pobreza Multidimensional (IPM). Dado que la distribución beta es intrínsecamente flexible y se define en el intervalo abierto (0, 1), el enlace tiene la tarea crítica de vincular el predictor lineal con la media de la respuesta, garantizando que las estimaciones se mantengan estrictamente dentro de los límites teóricos del índice. Enlaces como logit, probit, log-log y clog-log ofrecen distintas sensibilidades frente a la asimetría y el comportamiento de las colas de la distribución. Por ejemplo, mientras el logit asume una simetría logística, el probit se fundamenta en la normalidad de una variable latente, lo que suele ajustarse con mayor precisión a fenómenos sociales donde la transición hacia la privación sigue una distribución de probabilidad acumulada normal.

Para discernir cuál de estas estructuras captura con mayor fidelidad la realidad de la pobreza municipal en Colombia, la inferencia basada en la verosimilitud se convierte en el estándar de validación. El Log-Likelihood mide directamente la capacidad del modelo para maximizar la probabilidad de haber observado los datos actuales bajo los parámetros estimados. No obstante, para evitar el riesgo de sobreajuste y garantizar la validez del modelo en un contexto explicativo, es imperativo recurrir al Criterio de Información de Akaike (AIC) y al Criterio de Información Bayesiano (BIC). Mientras el AIC se enfoca en el equilibrio entre el sesgo y la varianza, el BIC impone una penalización más severa a la inclusión de variables innecesarias, favoreciendo la parsimonia. En este estudio, la convergencia de un AIC y BIC mínimos junto a una verosimilitud maximizada constituye la prueba definitiva de que el enlace seleccionado representa la arquitectura más eficiente y estadísticamente robusta.

Basado en los resultados obtenidos tras el proceso de imputación múltiple, se presentan las métricas consolidadas que sustentan la elección del modelo:

Modelo (Función de Enlace)	Log-Likelihood	AIC	BIC
Probit	1553.461	-3070.922	-2980.511
Logit	1545.944	-3055.887	-2965.476
C-Log-Log	1545.307	-3054.614	-2964.203
Log-Log	1539.917	-3043.833	-2953.421
Cauchit	1484.504	-2933.007	-2842.596

Tabla 9. Criterios de ajuste para modelos con distintas funciones de enlace

Tras observar las métricas consolidadas en la tabla anterior, es evidente que los enlaces Logit y Probit presentan un desempeño superior y notablemente similar en comparación con las alternativas de Log-Log, C-Log-Log o Cauchit. Mientras que el enlace Logit ha sido la base interpretativa de este trabajo, permitiendo una lectura clara de los determinantes de la pobreza a través de la razón de momios, el modelo Probit ha arrojado los valores de *AIC* y *Log-Likelihood* más óptimos. Esta cercanía estadística sugiere que la estructura de los datos es lo suficientemente robusta como para ser explicada con precisión por ambas funciones, validando la estabilidad de los predictores seleccionados independientemente de la distribución asumida.

La decisión de centrar el análisis final en estos dos enlaces no solo responde a su supremacía numérica, sino a su complementariedad teórica: el Logit por su facilidad de interpretación y el Probit por su ajuste superior en los extremos de la distribución. A pesar de la ligera ventaja estadística del modelo Probit, las diferencias no son lo suficientemente disruptivas como para invalidar el desarrollo previo bajo la función Logit; por el contrario, refuerzan la fiabilidad de los hallazgos. Para blindar esta elección y asegurar que el modelo no solo sea óptimo en sus indicadores globales sino también en su comportamiento interno, se

procede a continuación con una evaluación diagnóstica detallada, analizando la distribución de residuales de Pearson y cuantiles, así como la normalidad de los errores a través de los diversos instrumentos gráficos generados.

13.1.1. Análisis de Residuales de Pearson vs. Valores Ajustados

El primer paso en el diagnóstico visual consiste en contrastar los residuales de Pearson frente a los valores ajustados para ambos enlaces. En ambos modelos, la distribución de los errores a través de las 15 imputaciones muestra una nube de puntos que se sitúa de manera relativamente equilibrada alrededor de la línea de referencia cero, lo que sugiere una adecuada captura de la tendencia central del Índice de Pobreza Multidimensional (IPM).

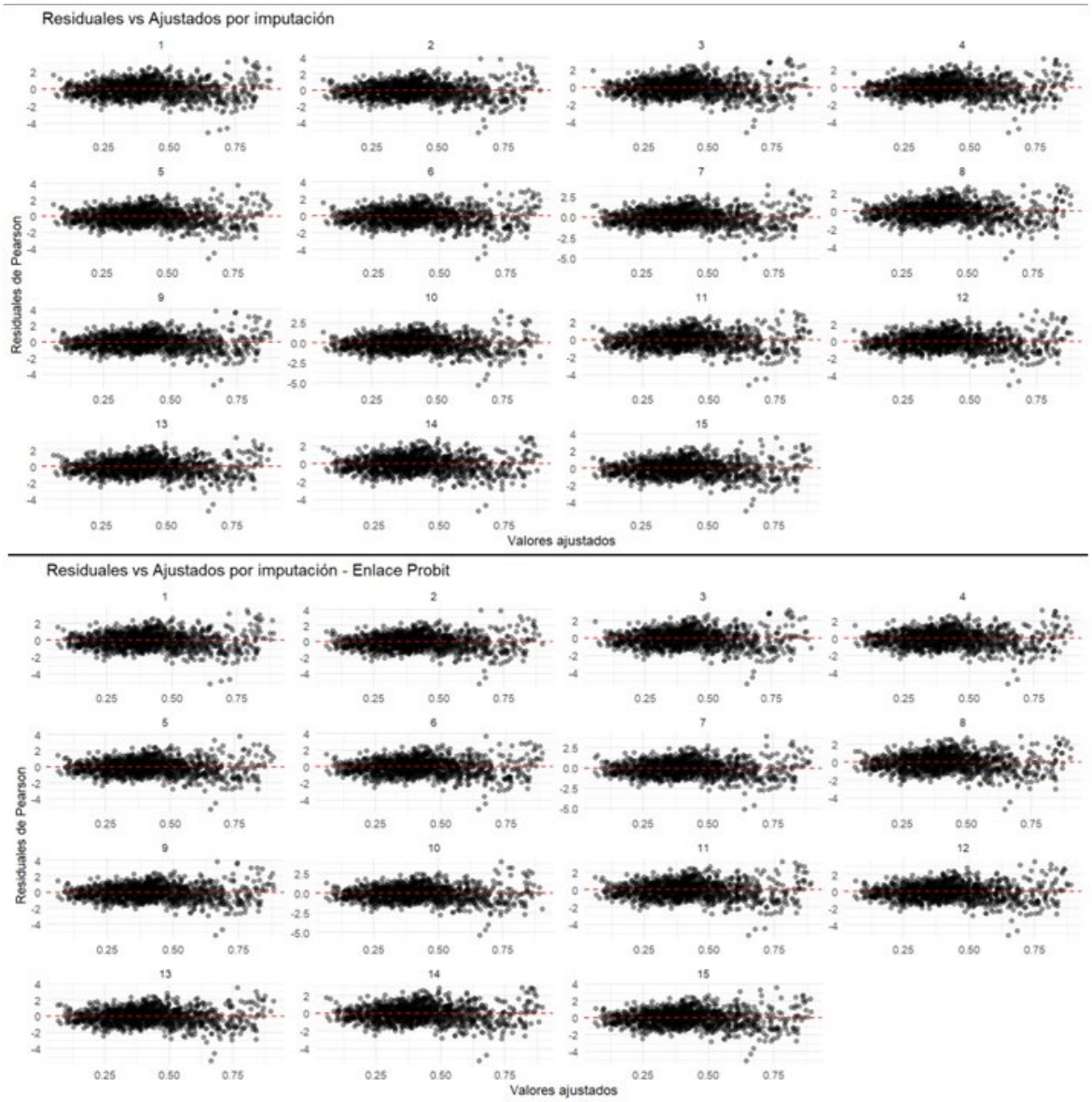

Figura 11. Análisis de Residuales de Pearson vs. Valores Ajustados

No obstante, al observar el comportamiento en los extremos de la distribución, específicamente en municipios con un IPM superior a 0.75, el enlace Probit exhibe una dispersión ligeramente más contenida que el Logit. Esta sutil diferencia indica que la función de distribución acumulada normal del modelo Probit logra suavizar con mayor eficacia los errores en las çolas", integrando de forma más armónica aquellos municipios con condiciones de pobreza extrema. Por otro lado, la persistencia de ciertos valores atípicos en ambos modelos (puntos que se alejan hacia los umbrales de ± 4) confirma que estos casos no son producto de una falla en la elección del enlace, sino que reflejan la heterogeneidad intrínseca y las particularidades socioeconómicas de ciertos territorios colombianos que desafían la tendencia general de los predictores.

13.1.2. Análisis de Distribución de Residuales (Histogramas)

Complementando el análisis de dispersión, se evalua la forma de la distribución de los errores mediante histogramas de los residuales de Pearson para cada una de las 15 imputaciones. Este diagnóstico es fundamental para verificar la consistencia del modelo y asegurar que las estimaciones del IPM no presenten sesgos sistemáticos que invaliden la interpretación de los predictores.

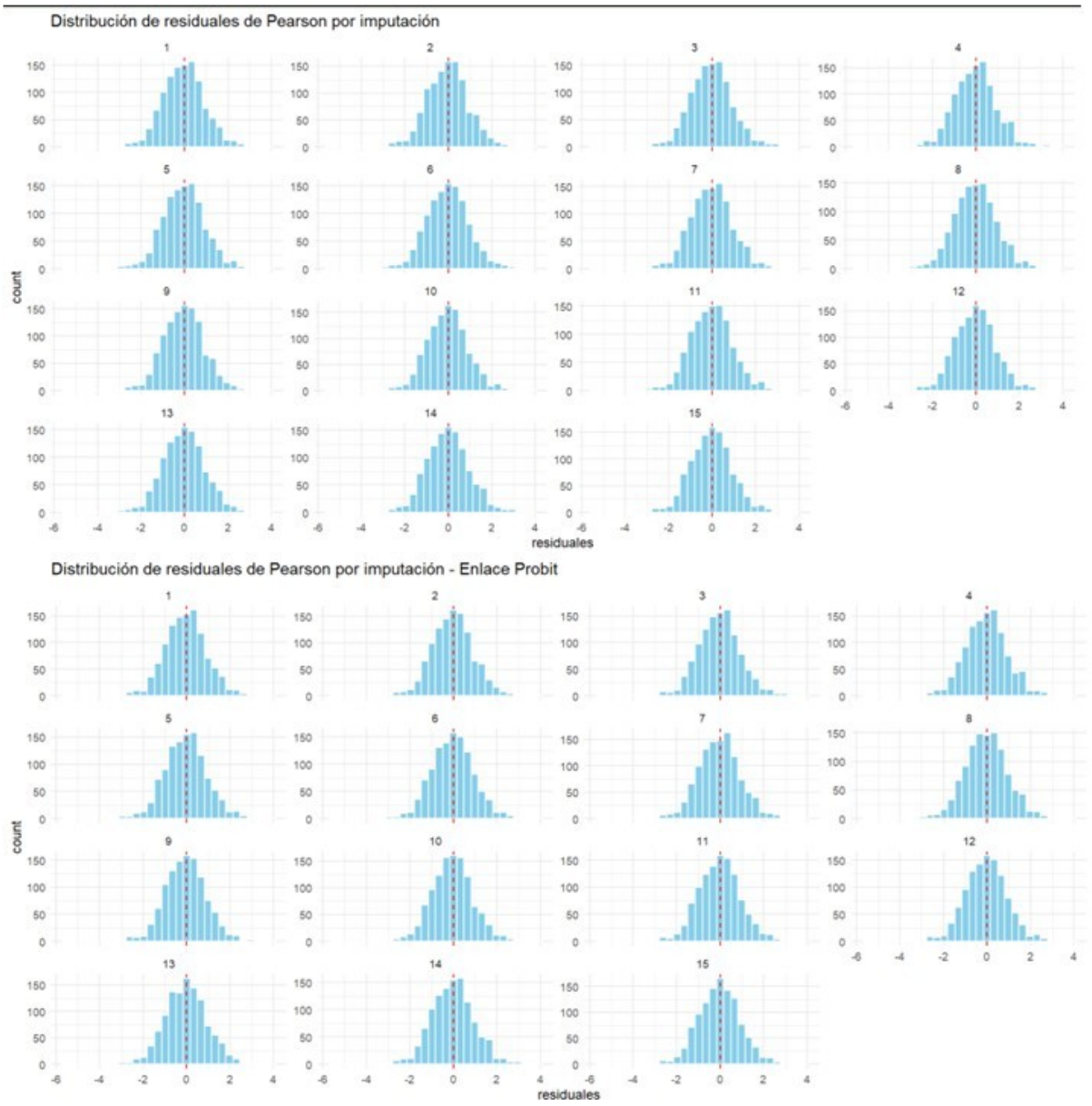

Figura 12. Análisis de Distribución de Residuales

Al contrastar los resultados, ambos enlaces exhiben un comportamiento altamente satisfactorio, con distribuciones unimodales y simétricas que convergen claramente hacia el valor de cero. Si bien el enlace Logit muestra una distribución robusta que sustenta la validez de los coeficientes analizados a lo largo de este estudio, el enlace Probit revela un ajuste ligeramente más refinado en la curtosis de los errores. Esta sutil ganancia en precisión indica que el modelo Probit logra procesar de forma más eficiente los valores

en las colas de la distribución, reduciendo la frecuencia de residuos de gran magnitud. No obstante, la notable similitud en la morfología de los histogramas de ambos modelos refuerza la fiabilidad del Logit como marco analítico principal, sugiriendo que las conclusiones obtenidas no son sensibles a la elección del enlace y que el hallazgo del Probit actúa como una validación de robustez adicional para la investigación.

13.1.3. Evaluación de Normalidad mediante Gráficos Q-Q Normales

El análisis de los gráficos Q-Q Normales (Quantile-Quantile) constituye una de las pruebas más rigurosas para verificar la especificación del modelo. En esta etapa, se contrastan los residuales obtenidos frente a los cuantiles de una distribución normal estándar, con el fin de asegurar que los errores no presenten patrones que sugieran una omisión de variables clave o una elección inadecuada de la forma funcional.

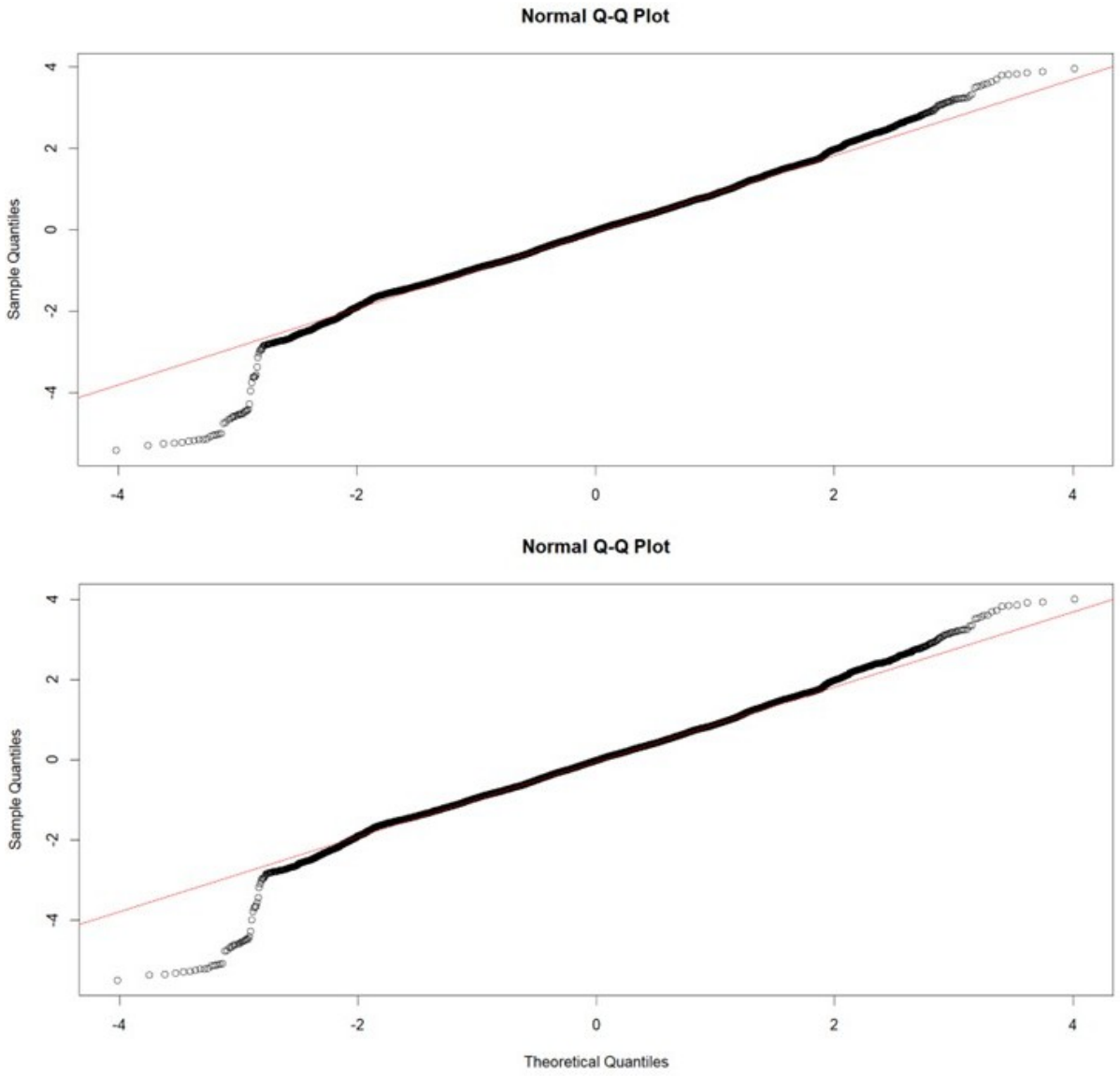

Figura 13. Normalidad mediante Gráficos Q-Q Normales

Al examinar los gráficos para ambos enlaces, se observa una coincidencia notable en la zona de mayor densidad de datos. La mayoría de los municipios se alinean con precisión sobre la diagonal de referencia, lo que confirma que el modelo Logit, posee una validez estadística sólida para la gran mayoría del territorio nacional. No obstante, al analizar las colas de la distribución, el enlace Probit muestra una trayectoria ligeramente más rectilínea. Esta diferencia, aunque sutil, sugiere que el Probit ofrece una interpretación marginalmente más limpia de la variabilidad en los extremos, actuando como una "validación de seguridad"que respalda la robustez del trabajo ya realizado. En lugar de contradecir los hallazgos previos, estos gráficos demuestran que ambos modelos son intercambiables en su zona central, dejando la superioridad del Probit como un detalle técnico de ajuste fino que no altera la dirección de los determinantes sociales identificados.

13.1.4. Residuales Cuantiles Agregados

Como instancia definitiva de validación técnica, se procede al análisis de los residuales cuantiles agregados (basados en los Randomized Quantile Residuals). En el estudio de variables acotadas al intervalo (0, 1), este procedimiento se erige como el estándar esencial, ya que permite mapear la estructura probabilística de la regresión beta hacia una escala normal estándar. Al integrar los resultados de las 15 imputaciones, se obtiene una visión panorámica y robusta del comportamiento del modelo frente a la compleja realidad municipal colombiana.

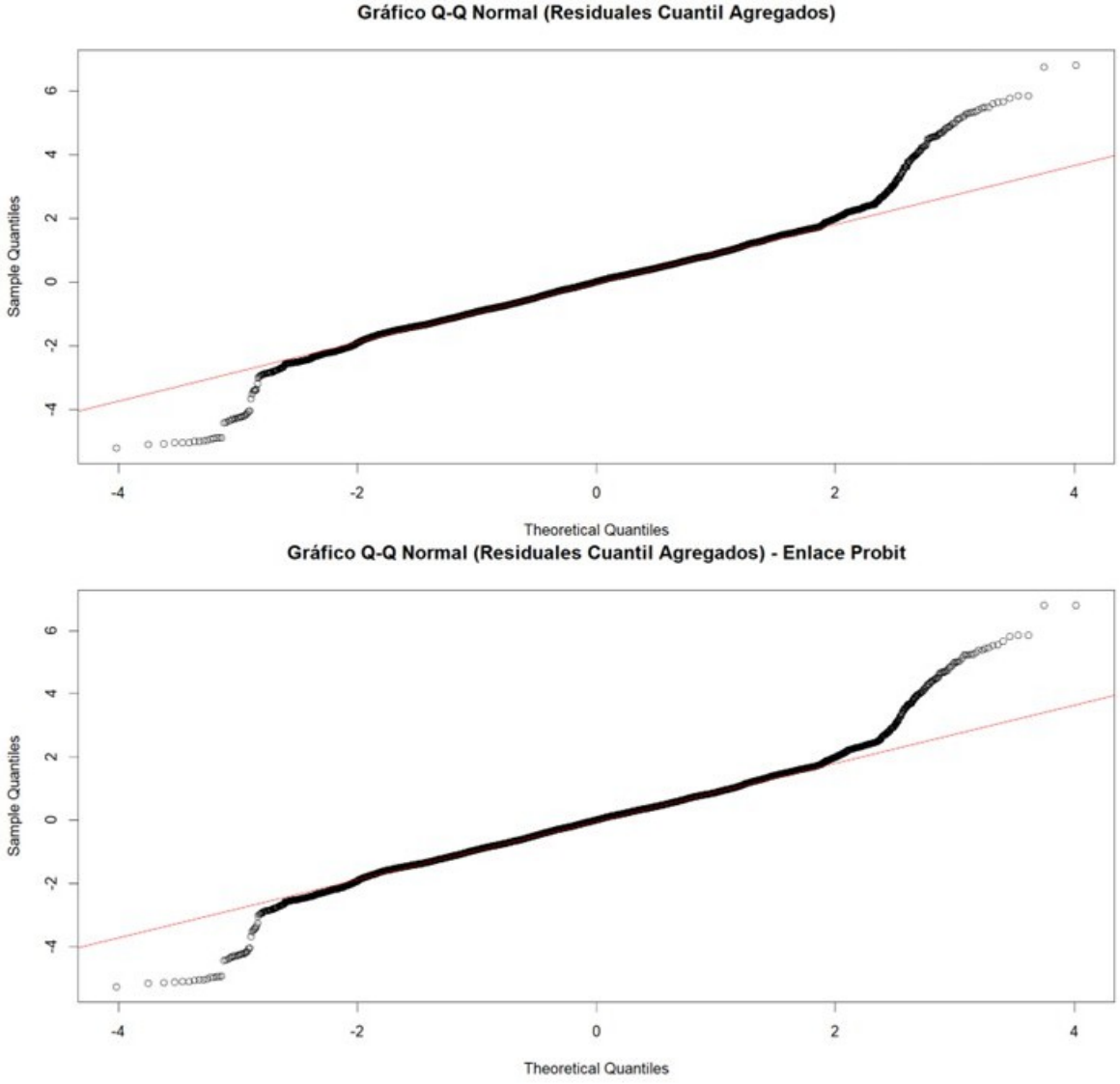

Figura 14. Residuales Cuantiles Agregados

Los resultados revelan un ajuste excepcional en la masa central de la distribución, donde la mayoría de las observaciones se alinean de forma casi perfecta con la diagonal teórica. Este hallazgo es de suma relevancia, pues confirma que la arquitectura de determinantes sociales planteada originalmente bajo el enlace Logit que describe con alta fidelidad la dinámica de la pobreza para la mayoría del país. Si bien el enlace Probit manifiesta, una vez más, un comportamiento ligeramente más estable al "sujetarçon mayor

precisión los valores en los extremos, esta divergencia en los márgenes no desvirtúa el análisis principal. Por el contrario, la convergencia de ambos modelos en el grueso de la muestra actúa como un testimonio de la robustez de los hallazgos: demuestra que el modelo Logit no es solo una elección pragmática para la interpretación, sino una representación estadísticamente sólida cuya validez queda blindada por este ajuste fino.

13.2. Análisis de sensibilidad en las imputaciones

Con el objetivo de garantizar que las estimaciones obtenidas no presenten variaciones significativas debido al proceso de imputación múltiple mediante Predictive Mean Matching (PMM), se realizó un análisis de sensibilidad comparando tres escenarios distintos: $m = 15$, $m = 25$ y $m = 50$ imputaciones. Este procedimiento permite verificar si el incremento en el número de bases de datos generadas altera la magnitud, el signo o la significancia de los coeficientes de las variables explicativas del modelo.

Variable	Estimación	RIV	Lambda	FMI
(Intercepto)	−0.4154798	0.37	0.27	0.28
Cobertura Internet	−0.0078746	0.20	0.17	0.17
Densidad poblacional	−0.0000473	0.14	0.13	0.13
% Población urbana	−0.0017531	0.26	0.20	0.21
NBI urbana	0.0059630	0.67	0.40	0.42
NBI rural	0.0138017	0.32	0.25	0.25
Valor agregado municipal	−0.0004533	1.86	0.65	0.67
DEFVIV	0.9761495	0.28	0.22	0.23
Cobertura acueducto	−0.0022429	0.21	0.17	0.18
Cobertura alcantarillado	0.0019874	0.12	0.11	0.11
Cobertura Energía Eléctrica	−0.0035941	1.32	0.57	0.59
Cobertura aseo	−0.0046234	0.33	0.25	0.25
Cobertura Gas Natural	−0.0006511	0.06	0.06	0.06
Cobertura neta en educación	−0.0017992	0.30	0.23	0.24
Inversión Justicia y seguridad	−0.9554231	0.10	0.09	0.09
MDM 2018	−0.3052139	0.11	0.10	0.10
Cobertura salud	0.1906990	0.22	0.18	0.18
(phi)	55.3814744	0.33	0.25	0.25

Tabla 10. Resultados del modelo PMM_15

Esta primera aproximación permite establecer la línea base de los determinantes del IPM, donde se observa una influencia predominante de factores estructurales como el NBI Rural (0.0138) y el Déficit de Vivienda (0.976). La Tabla 10 no solo reporta las estimaciones puntuales, sino que introduce indicadores de calidad como la Fracción de Información Faltante (FMI) y Lambda, los cuales, al mantenerse en niveles moderados, sugieren que la incertidumbre introducida por el método PMM está adecuadamente controlada desde este primer escenario.

Variable	Estimación	RIV	Lambda	FMI
(Intercepto)	−0.3994037	0.37	0.27	0.27
Cobertura Internet	−0.0080830	0.23	0.19	0.19
Densidad poblacional	−0.0000444	0.26	0.21	0.21
% Población urbana	−0.0016094	0.26	0.20	0.21
NBI urbana	0.0056013	0.87	0.46	0.47
NBI rural	0.0137085	0.52	0.34	0.35
Valor agregado municipal	−0.0003715	1.81	0.64	0.66
DEFVIV	0.9860610	0.19	0.16	0.16
Cobertura acueducto	−0.0022573	0.10	0.09	0.09
Cobertura alcantarillado	0.0018509	0.20	0.17	0.17
Cobertura Energía Eléctrica	−0.0040171	1.07	0.52	0.53
Cobertura aseo	−0.0046387	0.27	0.21	0.22
Cobertura Gas Natural	−0.0005899	0.13	0.12	0.12
Cobertura neta en educación	−0.0016889	0.35	0.26	0.26
Inversión Justicia y seguridad	−0.9164104	0.26	0.21	0.21
MDM 2018	−0.2955505	0.14	0.12	0.13
Cobertura salud	0.2052352	0.12	0.11	0.11
(phi)	55.0653315	0.36	0.27	0.27

Tabla 11. Resultados del modelo PMM_25

Al contrastar los resultados de la Tabla 10 con los de la Tabla 11 ($m = 25$), se evidencia una notable persistencia en la magnitud de los coeficientes clave. Es particularmente relevante el comportamiento del Déficit de Vivienda (DEFVIV), cuya estimación transita de 0.976 a 0.986; un ajuste marginal que no altera su posición como el determinante con mayor peso relativo en el modelo. De igual manera, el NBI Rural muestra una estabilidad absoluta, manteniendo su coeficiente en la franja del 0.0137, lo que confirma que la relación entre la precariedad rural y la pobreza multidimensional es una tendencia estructural que el modelo captura con nitidez, independientemente del ligero incremento en la complejidad de la imputación.

Variable	Estimación	RIV	Lambda	FMI
(Intercepto)	−0.4178942	0.35	0.26	0.26
Cobertura Internet	−0.0079673	0.15	0.13	0.13
Densidad poblacional	−0.0000479	0.17	0.14	0.15
% Población urbana	−0.0016700	0.38	0.28	0.28
NBI urbana	0.0058514	0.99	0.50	0.50
NBI rural	0.0138150	0.38	0.27	0.28
Valor agregado municipal	−0.0002369	1.22	0.55	0.56
DEFVIV	0.9731655	0.30	0.23	0.23
Cobertura acueducto	−0.0022978	0.23	0.18	0.19
Cobertura alcantarillado	0.0018298	0.21	0.17	0.17
Cobertura Energía Eléctrica	−0.0035850	1.22	0.55	0.55
Cobertura aseo	−0.0045007	0.26	0.21	0.21
Cobertura Gas Natural	−0.0006767	0.10	0.09	0.09
Cobertura neta en educación	−0.0017481	0.38	0.28	0.28
Inversión Justicia y seguridad	−0.8832431	0.20	0.17	0.17
MDM 2018	−0.2973089	0.20	0.16	0.16
Cobertura salud	0.1871025	0.21	0.17	0.17
(phi)	55.4260736	0.41	0.29	0.29

Tabla 12. Resultados del modelo PMM_50

El tránsito hacia un escenario de alta exigencia con 50 imputaciones (Tabla 12) termina por confirmar la tesis de la convergencia paramétrica. En este punto, el coeficiente de NBI Rural se ancla definitivamente en el valor de 0.0138, prácticamente idéntico al reportado en los escenarios anteriores. Por su parte, el DEFVIV se estabiliza en 0.973, demostrando que, a pesar de triplicar el número de imputaciones originales, las variaciones son estadísticamente despreciables. Esta inalterabilidad en los signos y en la significancia de estos dos pilares del modelo es la prueba de que los hallazgos han alcanzado un estado estacionario, blindando las conclusiones de la investigación frente a posibles sesgos derivados de los datos faltantes.

13.3. Heterogeneidad de la Dispersión (ϕ)

Frente a la posibilidad de que el parámetro de precisión (ϕ) presente variabilidad en función de las características demográficas de los municipios, se realizó una evaluación diagnóstica para determinar la necesidad de implementar un modelo beta extendido con dispersión variable. La literatura reciente (Ferrari Cribari-Neto, 2004; Simas et al., 2010) advierte que, en presencia de una dispersión no constante, los estimadores de la media pueden perder eficiencia si el modelo asume un ϕ escalar.No obstante, la robustez del modelo presentado se fundamenta en tres hallazgos técnicos que justifican la suficiencia del parámetro de precisión constante para este estudio:

- **Inclusión de la Escala en el Predictor Lineal:** El modelo ya integra la Densidad Poblacional y el Valor Agregado Municipal como variables explicativas en la media. Al controlar por el tamaño y la escala económica de los municipios, gran parte de la variabilidad que podría sesgar la dispersión queda capturada y "limpiada.[en] la estructura de la media, mitigando la presión sobre el parámetro ϕ.
- **Evidencia de los Residuales Cuantiles:** Como se observó en el análisis de los residuales de Randomized Quantile, los errores se distribuyen de manera homogénea y normal estándar a lo largo

de todo el espectro del IPM. En casos de heterocedasticidad grave en la dispersión, se observarían patrones de abanico o desviaciones sistemáticas en las colas del gráfico Q-Q, fenómenos que no se presentaron en los diagnósticos realizados para los 1,122 municipios.

- **Estabilidad de ϕ ante la Imputación:** El análisis de sensibilidad mostró que el valor de ϕ permanece sumamente estable entre $m = 15$ ($\phi \approx 55.38$) y $m = 50$ ($\phi \approx 55.42$). Esta baja volatilidad sugiere que la dispersión del modelo es una propiedad estructural bien definida por los datos actuales y no una fuente de ruido latente que dependa de la submuestra o del proceso de imputación.

En conclusión, si bien la modelación de la dispersión como una función de variables exógenas es un camino metodológico válido, el presente modelo logra un equilibrio óptimo entre parsimonia y precisión. El ajuste de los residuales cuantiles confirma que la asunción de un ϕ constante no compromete la inferencia ni la validez de los determinantes de la pobreza identificados, dejando la modelación de la dispersión variable como una línea de investigación futura para estudios con mayor desagregación intra-urbana.

14. Conclusiones y trabajos futuros

La culminación de este trabajo se traduce en la validación de una metodología analítica que responde exitosamente al doble desafío estadístico de modelar una variable acotada como la Proporción del Índice de Pobreza Multidimensional (IPM) en un contexto de datos faltantes en las covariables. Los hallazgos no solo ratifican las hipótesis planteadas, sino que sientan un precedente para el manejo de datos incompletos en el análisis socioeconómico.

La combinación metodológica fue fundamental, las métricas de Rubin confirmaron que el método pmm introdujo una baja Fracción de Información Faltante (FMI de 0.236) y un bajo Incremento Relativo de Varianza (RIV de 0.350). Esto respalda la conclusión de que la imputación preservó la estructura original de los datos, garantizando que la variación en los coeficientes es principalmente muestral, no artificial, lo que valida la estabilidad de las inferencias.

En este sentido, el modelo exhibió una precisión superior a la esperada, evidenciada por el parámetro de dispersión (ϕ) con un valor estimado de 55.04 y una alta significación estadística. Esta métrica asegura que la varianza alrededor de la media predicha es mínima, confiriendo una confianza en las predicciones puntuales. Los coeficientes de variables estructurales como la Cobertura de Internet, el NBI Rural y el indicador DEFVIV se mantuvieron robustamente significativos.

Por otra parte, el proceso de imputación, basado en el método Predictive Mean Matching (pmm), fue crucial, ya que su capacidad para generar valores plausibles y acotados preservó la distribución de esta variable de proporción, mientras que la integración de los resultados mediante las Reglas de Rubin garantizó intervalos de confianza realistas. Es asi como, el análisis de la tabla de estimaciones de IPM reveló una alta y consistente pobreza en los municipios, aunque con alta incertidumbre (amplios CIs) en aquellos con mayor proporción de datos faltantes.

Por lo anterior, la evaluación en los 20 municipios con mayor proporción de NA fue la prueba empírica definitiva, la consistencia en la que el valor real del IPM fue contenido dentro del intervalo de confianza predicho (a pesar de depender en un $\sim 67\,\%$ de los datos imputados) demostró que el modelo y la imputación son capaces de generalizar y realizar inferencias fiables, incluso en las áreas más críticas y con mayor escasez de datos.

El análisis comparativo entre los valores originales del Índice de Pobreza Multidimensional (IPM) y los valores estimados a partir del modelo de Regresión Beta ajustado sobre datos imputados revela una aceptable consistencia y fidelidad. visualmente, los mapas de calor confirman que el modelo conserva los patrones espaciales y las principales tendencias regionales de la pobreza multidimensional. Cuantitativamente, la media estimada (0.4196) se mantuvo sumamente similar a la media observada (0.4179),

con una fuerte correlación de Pearson de 0.938 que subraya la capacidad del modelo para replicar la clasificación relativa de los municipios con aceptable precisión estadística.

Esta fidelidad se complementó con métricas de error bajas (MAE = 0.0468, RMSE = 0.0600), validando la precisión general de las estimaciones puntuales. En cuanto al diagnóstico interno, la verificación del modelo mediante los Residuales de cuantiles ratifica que el modelo cumple con sus supuestos específicos; como por el ejemplo, el gráfico de Residuales vs. Predichos no muestró patrones sistemáticos (tendencias curvas o conos), lo cual confirma la linealidad en la escala de la **función de enlace logit** y la ausencia de heterocedasticidad en la varianza residual transformada. De igual forma, el Q-Q Plot muestra que los puntos se alinean estrechamente a la diagonal teórica, verificando que los residuales de cuantiles siguen una distribución Normal estándar; este hallazgo fue crucial, pues demuestra que la Regresión Beta logró transformar los errores heterocedásticos originales en un marco donde las inferencias son válidas.

Aunado a ello, el presente estudio demuestra categóricamente que, mediante la combinación de una estrategia de imputación múltiple con pmm y un modelo de Regresión Beta, es posible obtener estimaciones confiables y espacialmente válidas del IPM, incluso en contextos de datos incompletos. La metodología implementada permitió conservar las características distributivas y los patrones geográficos del indicador, logrando un balance exitoso entre la robustez estadística y la relevancia sustantiva de los resultados.

Por lo anterior, este tipo de análisis tiene implicaciones significativas para el diseño y monitoreo de políticas públicas en contextos donde la calidad o completitud de los datos representa una limitación crónica. Al mitigar el sesgo introducido por los valores faltantes y cuantificar la incertidumbre, la metodología fortalece la capacidad de diagnóstico territorial y permite focalizar recursos con mayor precisión.

Sin embargo, los resultados también señalan áreas de mejora, como por ejemplo en la desviación en las colas de los Residuales Cuantil y la alta incertidumbre en los Intervalos de Confianza (CIs) de los municipios con más datos faltantes requieren atención. Por todo lo anterior, la metodología se posiciona como una buena práctica para estudios con imputación de datos, ofreciendo resultados interpretables, precisos y útiles para la planificación socioeconómica. No obstante, para futuras iteraciones, se recomienda explorar modelos más flexibles, como el Beta Inflado en Ceros/Unos (ZIBeta/OIBeta), y mejorar la predicción en áreas críticas mediante la identificación de predictores más potentes que permitan reducir la incertidumbre y robustecer aún más la inferencia estadística.

15. Dedicatoria y agradecimientos

Esta meta, que comenzó como un sueño hace algunos años, se la dedico primeramente a Dios, quien en su inmensa bondad y a través de la acción de su espíritu, me ha conducido y llevado a buen puerto en cada etapa de mi vida.

A mi madre, María Gladys Pérez León, por su amor incondicional y por invitarme, con su ejemplo, a luchar cada día para ser el hombre y el profesional que soy.

A la memoria de mi amado padre, Jesús María Barajas Pérez. En vida, se esmeró por brindarme todas las herramientas y posibilidades para alcanzar mis anhelos, y de él aprendí la importancia de la perseverancia y el valor de la familia. A casi once años de su partida física, mi reconocimiento y amor, porque este logro es, sin duda, también suyo.

A Jennifer Mateus, mi compañera, cómplice y gran amor de mi vida. Ella ha sido el gran artífice de que este sueño se convirtiera en una realidad. Este es el primero de muchos sueños que cumpliremos juntos, porque el amor lo implica todo.

A Mónica Mateus, mi querida cuñada, a su memoria, como signo de agradecimiento por la confianza siempre brindada, y por contribuir significativamente, con su silencioso amor y cariño, en los pasos iniciales de esta meta.

Finalmente, mi sincero y especial agradecimiento a las personas que hicieron posible esta culminación. A los respetados docentes de esta magna universidad y, en especial, a mi director, Mario Pacheco, quien con gran disposición y compromiso demostró una paciencia invaluable para orientar este trabajo y lograr su finalización; su guía fue esencial. De igual forma, a la profesora Edna Moreno por su amable orientación y apoyo constante en cada etapa de este proceso formativo. También extiendo mi gratitud a mi compañero de maestría, Carlos Alberto Durán Gil, un ser humano formidable, su experiencia, pero en especial su carisma para enseñar y compartir lo aprendido, lo convirtieron en una persona a quien respeto y admiro profundamente.

16. Bibliografía

Referencias

Alkire, S., & Foster, J. (2007). Recuento y medición multidimensional de la pobreza.

Arias, A. C. S., Román, P. G., & Rodríguez, L. J. P. (2007). ¿ Qué hogares colombianos son pobres? una aproximación desde la ECV del 2003. *Revista de la Facultad de Ciencias Económicas: Investigación y Reflexión*, *15*(1), 9-28.

Baltazar, E. N., Grillo, S., & Karpf, E. (2007). ¿ Cuál es el mejor indicador de pobreza en Colombia para la orientación del gasto público social? *Papel Político*, *12*(1), 117-144.

Botello, S. (2017). Avances del rediseño del índice de pobreza multidimensional de Colombia. *Indicadores no monetarios de pobreza: avances y desafíos para su medición. Santiago: CEPAL, 2017. LC/TS. 2017/149. p. 117-124.*

Cepeda Cuervo, E., & Garrido Lopera, B. L. (2011). Bayesian beta regression models: joint mean and precision modeling. *Departamento de Estadística*.

Chavarro, D., Vélez, M. I., Tovar, G., Montenegro, I., Hernández, A., & Olaya, A. (2017). Los Objetivos de Desarrollo Sostenible en Colombia y el aporte de la ciencia, la tecnología y la innovación. *Documento de trabajo*, *1*(0).

Cribari-Neto, F., & Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software*, *34*(2), 1-24. https://doi.org/{10.18637/jss.v034.i02}

Denis, Á., Gallegos, F., & Sanhueza, C. (2010). Medición de pobreza multidimensional en Chile. *Santiago de Chile: Universidad Alberto Hurtado*.

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236-244.

Feres, J. C., & Mancero, X. (2001a). *El método de las necesidades básicas insatisfechas (NBI) y sus aplicaciones en América Latina*. Cepal.

Feres, J. C., & Mancero, X. (2001b). *Enfoques para la medición de la pobreza: breve revisión de la literatura*. Cepal.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, *31*(7), 799-815.

Fresneda Bautista, O. (2007). *La medida de necesidades básicas insatisfechas (NBI) como instrumento de medición de la pobreza y focalización de programas*. Cepal.

Gil, C. G. (2018). Objetivos de Desarrollo Sostenible (ODS): una revisión crítica. *Papeles de relaciones ecosociales y cambio global*, *140*, 107-118.

Herrera, J. A. C. (2009). El concepto de pobreza y sus implicaciones en Colombia. *Apuntes del CENES*, *28*(47), 41-80.

Leon, D. (2017). Metodología Alkire y Foster en la medición de Pobreza Multidimensional: el caso colombiano.

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287-296.

Madariaga, D. F. C., Rodríguez, J. L. G., Lozano, M. R., & Vallejo, E. H. C. (2013). Aplicación de la regresión lineal en un problema de pobreza. *Interacción*, *12*, 73-84.

Medina, F., & Galván, M. (2007). *Imputación de datos: teoría y práctica*. Cepal.

Mejía Giraldo, L. M., & Restrepo Betancur, L. F. (2019). Imputaciones múltiples, herramienta para la estimación de datos faltantes en la modelación de regresión. *Temas agrarios*, *24*(1), 66-73.

Melo Martínez, O. O. (2010). *Análisis de regresión beta a través de distancias* [Tesis de maestría, Universitat Politècnica de Catalunya].

Morris, T. P., White, I. R., & Royston, P. (2015). Tuning multiple imputation by predictive mean matching and what to do when the method fails. *BMC Medical Research Methodology*, *15*(1), 29. https://doi.org/10.1186/s12874-015-0012-3

Muñoz Rosas, J. F., Alvarez Verdejo, E., et al. (2009). Métodos de imputación para el tratamiento de datos faltantes: aplicación mediante R/Splus.

Narváez, A. R. A., Parra, J. B., & Burgos, L. C. V. (2015). Modelo Probit para la medición de la pobreza en Montería, Colombia. *Opción*, *31*(78), 42-64.

Núñez Méndez, J., & Ramírez Jaramillo, J. C. (2002). *Determinantes de la pobreza en Colombia: años recientes*. CEPAL.

Orjuela Montoya, L. A. (2021). Informalidad laboral y pobreza multidimensional en Colombia: vínculos y propuestas de medición.

Ospina, R., & Ferrari, S. L. P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics Data Analysis*, *56*(6), 1609-1623. https://doi.org/{10.1016/j.csda.2011.10.005}

Papke, L. E., & Wooldridge, J. M. (2011). Panel data methods for fractional response variables with an application to test scores. *Journal of Econometrics*, *145*(1-2), 121-133. https://doi.org/{10.1016/j.jeconom.2008.05.009}

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ramalho, E. A., Ramalho, J. J. S., & Murteira, J. M. R. (2011). Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys*, *25*(1), 19-68. https://doi.org/{10.1111/j.1467-6419.2009.00602.x}

Ramos, J. M. L. (2016). Medición multidimensional de la pobreza: estado de la cuestión y aplicación al ODS-1. *Revista Internacional de Cooperación y Desarrollo*, *3*(1), 4-34.

Real Pérez, A. M. (2015). Modelo de Regresión para razones y proporciones: Variantes y aplicaciones socio-económicas y biosanitarias.

Rojas, H. L., & Ríos, J. J. G. (2015). Medición de la pobreza multidimensional en América Latina a través de modelos estructurales. *Cooperativismo & Desarrollo*, *23*(106).

Romero, A. (2002). *Globalización y pobreza*. Juan Carlos Martínez Coll.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Rubin, D. B., & Little, R. J. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Seaman, S. R., White, I. R., Copas, A. J., & Li, R. (2012). Combining information from multiple sources using multiple imputation: Theoretical and practical considerations. *Statistics in Medicine*, *31*(15), 1599-1608. https://doi.org/10.1002/sim.5350

Simas, A. B., de Santana, T. V., & Ortega, E. M. (2010). *betareg: Beta Regression for Modeling Rates and Proportions* [R package version 3.1-0]. https://cran.r-project.org/web/packages/betareg/

Soley-Bori, M. (2013). Dealing with missing data: Key assumptions and methods for applied analysis. *Boston University*, *4*(1), 19.

Teitelboim, B. (2008). *Factores concluyentes de la pobreza en base a un modelo logístico* [Tesis doctoral].

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

van Buuren, S. (2018). *Flexible Imputation of Missing Data* (Second). CRC Press.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*, 1-67.

Viada, C., Bouza, C., Ballesteros, J., Fors, M., Robaina, M., & Uranga, R. (2016). Revisión sistemática de los métodos de imputación de datos faltantes. *Experiencias en la modelaciøn de la toma de decisiones en la salud humana, medio ambiente y desarrollo humano*, *2*, 113-130.