



S<sup>UNIVERSIDAD</sup> SANTO TOMAS

# Análisis Estadístico con Datos Faltantes

Author: Rolando Barajas Pérez

Profesor: José López

23 de octubre de 2025

## Índice

<b>1</b>	<b>Sobre los datos</b>	<b>2</b>
<b>2</b>	<b>Parte I: Análisis de datos faltantes.</b>	<b>2</b>
2.1	Datos faltantes por variable . . . . .	2
2.2	Mapa de calor de los patrones de datos faltantes . . . . .	4
2.3	Patrón de datos faltantes con mice . . . . .	5
2.4	Test de Little para MCA . . . . .	6
<b>3</b>	<b>Parte II: Estudio y tratamiento de datos faltantes.</b>	<b>7</b>
3.1	Cálculo de las Medidas de Severidad: Reglas de Rubin . . . . .	12
3.1.1	Modelo random forest . . . . .	12
3.1.2	Modelo knn . . . . .	13
3.1.3	Modelo pmm . . . . .	14
3.1.4	Modelo de regresión . . . . .	15
3.1.5	¿Cual modelo elegir? . . . . .	16
3.1.6	modelo PMM métricas de Rubin . . . . .	17
3.1.7	Modelo imputado . . . . .	18
<b>4</b>	<b>Análisis de residuales</b>	<b>19</b>
4.1	Análisis de Residuales Cuantil ( $r_Q, i$ ) . . . . .	23
4.2	Análisis de Influencia (Distancia de Cook) . . . . .	26
<b>5</b>	<b>tabla de comparativa de los 20 municipios</b>	<b>28</b>
<b>6</b>	<b>Mapa final</b>	<b>30</b>
<b>7</b>	<b>Conclusión</b>	<b>32</b>
7.1	Conclusiones y Proyecciones . . . . .	32
<b>8</b>	<b>Referencia</b>	<b>32</b>

# 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 principalmente 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 (acueducto, alcantarillado, energía eléctrica, aseo, gas natural, internet, salud y educación).

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.

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. Estos datos ausentes serán explorados en detalle en las siguientes secciones del informe.

```
datos <- read_excel("Base de datos - Indice de Pobreza.xlsx")  
  
colnames(datos) <- c("CODIGO", "Código Departamento", "Municipio", "Proporción IPM",  
"Densidad poblacional", "% población urbana", "% población rural",  
"NBI - en el área urbana", "NBI - en el área rural", "Valor agregado municipal (%)",  
"DEFVIV", "Cobertura de acueducto (%)", "Cobertura de alcantarillado (%)",  
"Cobertura de Energía Eléctrica (%)", "Cobertura de aseo (%)",  
"Cobertura de Gas Natural (%)", "Cobertura de Internet (%)",  
"Cobertura neta en educación", "% de inversión - Justicia y seguridad",  
"MDM 2018", "Cobertura salud")
```

## 2 Parte I: Análisis de datos faltantes.

Para identificar y comprender la naturaleza de los datos faltantes presentes en la base de datos, 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).

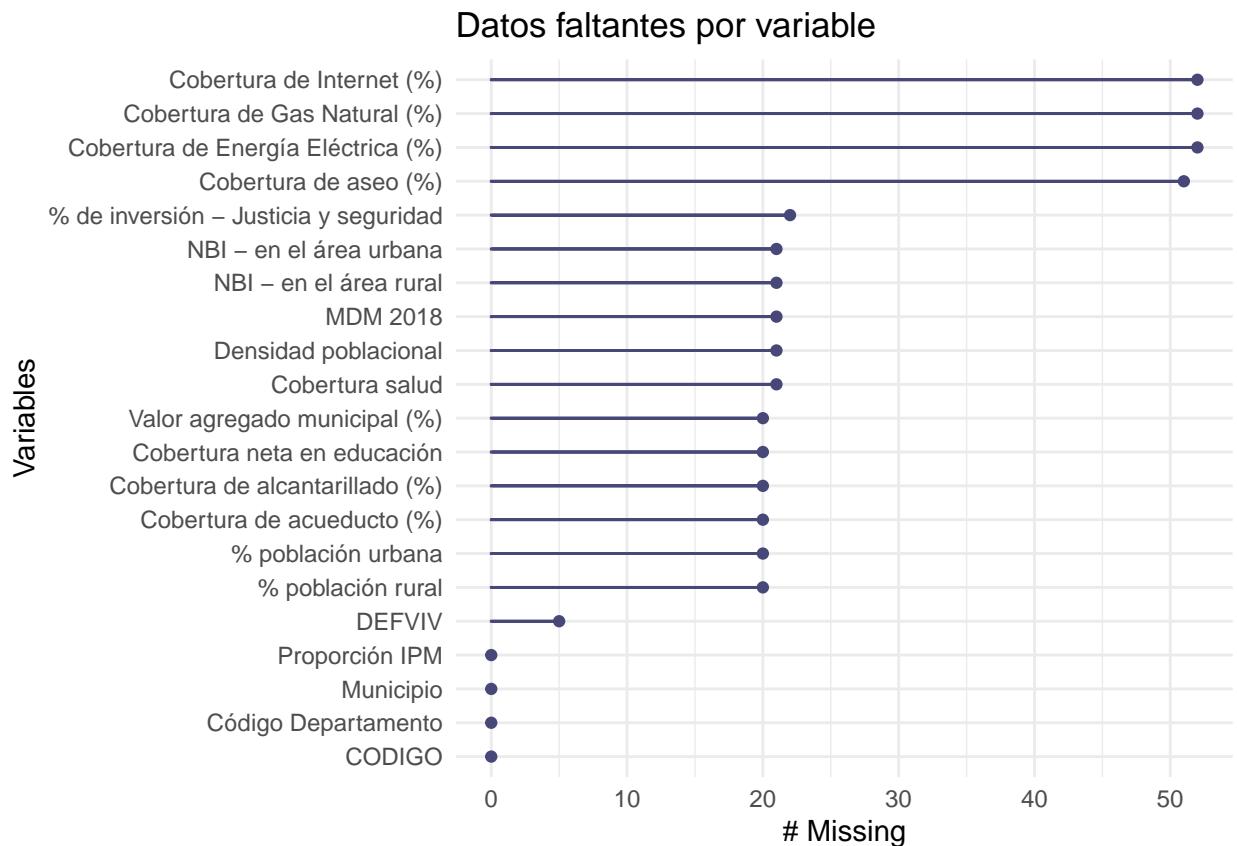
### 2.1 Datos faltantes por variable

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 (%)
- Cobertura de aseo (%)

```
# Porcentaje de datos faltantes por variable
gg_miss_var(datos) +
  labs(title = "Datos faltantes por variable")
```



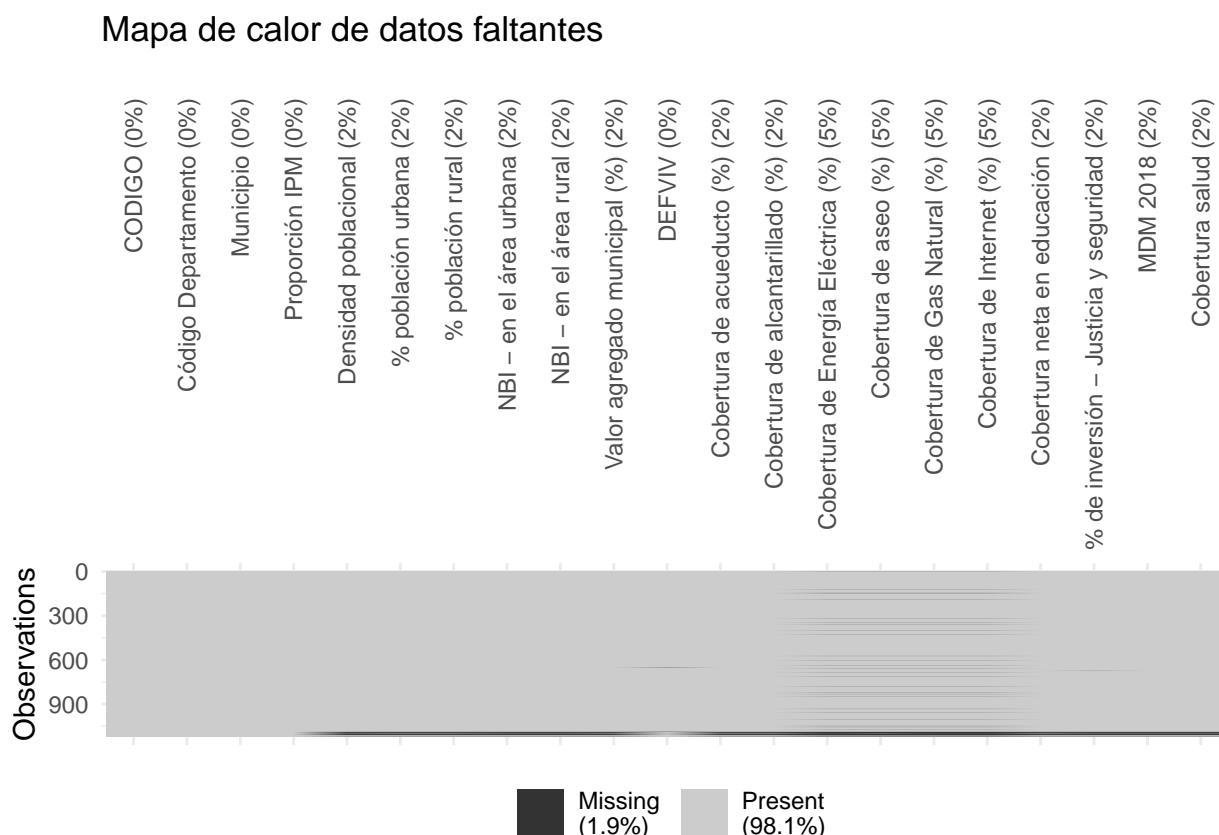
## 2.2 Mapa de calor de los patrones de datos faltantes

El mapa de calor de datos faltantes ofrece una representación visual clara del patrón de omisión en la base de datos. En él, cada fila representa un municipio (observación) y cada columna, una variable. Las celdas en color negro indican la presencia de datos faltantes, mientras que las grises representan valores presentes.

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.

Se observa que muchos de los registros con valores ausentes tienen múltiples variables con datos faltantes simultáneamente. Este patrón sugiere que no se trata de errores aleatorios en la recolección, sino de ausencias relacionadas posiblemente con características estructurales de los municipios. Esto puede ser evidencia frente a la hipótesis de que los datos no son MCAR (Missing Completely At Random), sino que podrían estar condicionados por otras variables observadas, lo cual se enmarca dentro de la categoría MAR (Missing At Random).

```
# Mapa de calor de los patrones de datos faltantes  
vis_miss(datos) +  
  labs(title = "Mapa de calor de datos faltantes") +  
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



## 2.3 Patrón de datos faltantes con mice

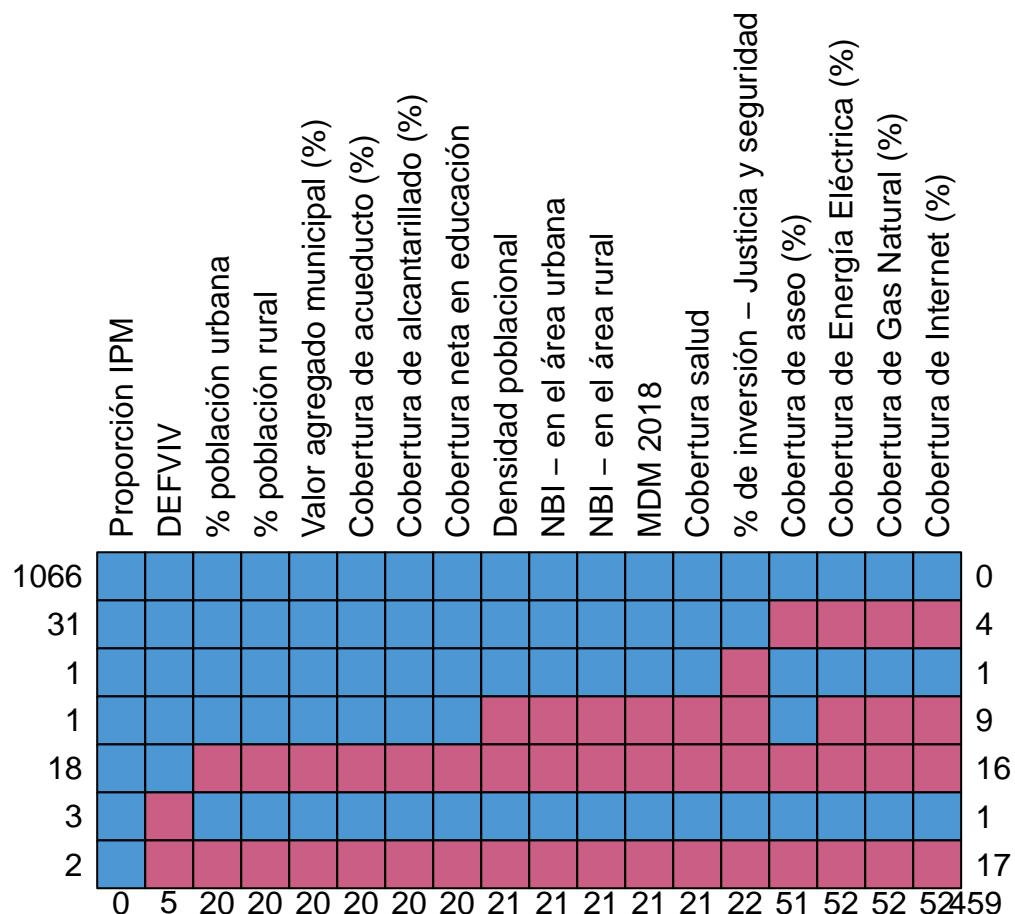
El análisis fue complementado mediante la función `md.pattern()` de la librería `mice`, que permite identificar patrones específicos de combinación entre variables faltantes. En la Figura siguiente se muestran los distintos grupos de observaciones que comparten estructuras comunes de omisión.

El patrón más frecuente corresponde a las observaciones completamente completas (1066 municipios), seguidas por combinaciones de ausencias concentradas en las variables de cobertura de servicios públicos —especialmente Internet, Gas Natural, Energía Eléctrica y Aseo— que aparecen de forma recurrente entre los registros incompletos.

La visualización permite observar que existen 17 municipios con valores faltantes simultáneamente en las cuatro variables mencionadas, así como otros grupos con entre 1 y 3 de esas variables ausentes. Este tipo de patrones sistemáticos refuerza la hipótesis de que los valores faltantes no son completamente aleatorios (no MCAR), ya que las ausencias tienden a presentarse en conjunto y afectan principalmente variables de infraestructura y servicios.

```
# Selección solo de variables numéricas (ignorar columnas categóricas para este análisis)
datos_numericos <- datos %>%
  dplyr::select(where(is.numeric))

# Solo gráfico, sin impresión en consola
invisible(md.pattern(datos_numericos, rotate.names = TRUE))
```



## 2.4 Test de Little para MCA

Para complementar el análisis visual y descriptivo de los datos faltantes, se aplicó 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.

Los resultados obtenidos fueron los siguientes:

- **Para las variables numéricas:** Estadístico chi-cuadrado = 1083, grados de libertad = 60, p-valor = 0
- **Para el conjunto completo de datos:** Estadístico chi-cuadrado = 1312, grados de libertad = 78, p-valor = 0

El p-valor, siendo menor a 0.05 en ambos casos, 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.

```
#mcar_test(datos)

#mcar_test(datos_numericos)

# Crear el data frame con los resultados
resultados <- data.frame(
  Variable = c("datos_numericos", "datos"),
  Estadistico = c(1083, 1312),
  Grados_de_libertad = c(60, 78),
  Valor_p = c(0, 0),
  Patrones_faltantes = c(7, 7)
)

# Mostrar la tabla con kable
kableExtra::kable(resultados,
  col.names = c("Variable", "Estadístico", "Grados de libertad (df)", "Valor p (p.value)"),
  caption = "Resultados del test MCAR") |>
  kable_styling(latex_options = c("HOLD_position"),
    full_width = F)
```

Cuadro 1: Resultados del test MCAR

Variable	Estadístico	Grados de libertad (df)	Valor p (p.value)	Patrones de datos faltantes
datos_numericos	1083	60	0	7
datos	1312	78	0	7

### 3 Parte II: Estudio y tratamiento de datos faltantes.

Con base en los hallazgos presentados en la Parte I, 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 evaluar diferentes técnicas de imputación de datos faltantes con el fin de restaurar el conjunto de datos.

```
### 1. Reglas de rubin para los modelos colapsados Utilizando la matrix de cov

### paquetes necesarios

library(mice)
library(missForest) # para imputación random forest
library(VIM)         # para imputación k-NN
library(readxl)
library(tidyverse)
library(betareg)
library(car)
library(sf)
library(openxlsx)
library(ggspatial)

## numero de imputaciones

m <- 15
maxit <- 30

## cargar datos

datos <- read_excel("Base de datos - Indice de Pobreza.xlsx")

## cambiar nombres

colnames(datos) <- c("CODIGO", "Código Departamento", "Municipio", "Proporción IPM",
                      "Densidad poblacional", "% población urbana", "% población rural",
                      "NBI - en el área urbana", "NBI - en el área rural", "Valor agregado",
                      "DEFVIV", "Cobertura de acueducto (%)", "Cobertura de alcantarillado",
                      "Cobertura de Energía Eléctrica (%)", "Cobertura de aseo (%)",
                      "Cobertura de Gas Natural (%)", "Cobertura de Internet (%)",
                      "Cobertura neta en educación", "% de inversión - Justicia y seguridad",
                      "MDM 2018", "Cobertura salud")



datos_numericos <- datos %>%
  dplyr::select(where(is.numeric))

names(datos_numericos) <- make.names(names(datos_numericos))

# Quitar la variable porcentaje de población rural de la base de datos.
```

```

datos_numericos <- datos_numericos[, -4]

#####
##### Plantear modelos
##### Plantear modelos

# random forest
metodos <- make.method(datos_numericos)
metodos[] <- "rf"

# Ejecutar imputación múltiple con random forest
set.seed(123)
imp_rf <- mice(datos_numericos, m = m, maxit = maxit,
                 method = metodos, seed = 123, printFlag = F)

# Extraer las m imputaciones completas como lista
imputations_rf <- complete(imp_rf, action = "all")

## Knn 5 vecinos
imputations_knn <- vector("list", m)
n <- nrow(datos_numericos)

for (i in 1:m) {
  set.seed(i)
  idx <- sample(seq_len(n), size = n, replace = TRUE)
  boot_data <- datos_numericos[idx, ]
  imputations_knn[[i]] <- kNN(boot_data, k = 5, imp_var = FALSE)
}

## pmm

metodos[] <- "pmm"

# Ejecutar imputación múltiple con PMM
set.seed(123)
imp_pmm <- mice(datos_numericos, m = m, method = metodos, seed = 123,
                  printFlag = F)

# Extraer las m imputaciones completas como lista
imputations_pmm <- complete(imp_pmm, action = "all")

### Modelo de regresión

# Definir la variable objetivo

```

```

variable_objetivo <- "Proporción.IPM"

# Método de imputación estocástico con regresión
metodos[variable_objetivo] <- "norm"

set.seed(123)
imp_Regresion <- mice(datos_numericos, m = m, maxit = maxit, method = metodos, seed = 123)

imputations_Regresion <- complete(imp_Regresion, action = "all")

form <- Proporción.IPM ~ Cobertura.de.Internet.... + Densidad.poblacional +
X..población.urbana +
NBI...en.el.área.urbana + NBI...en.el.área.rural +
Valor.agregado.municipal.... + DEFVIV +
Cobertura.de.acueducto.... + Cobertura.de.alcantarillado.... +
Cobertura.de.Energía.Electrónica.... + Cobertura.de.aseo.... +
Cobertura.de.Gas.Natural.... +
Cobertura.neta.en.educación + X..de.inversión...Justicia.y.seguridad +
MDM.2018 + Cobertura.salud

## ajustar el modelo con las imputaciones de cada modelo

fits_rf <- lapply(imputations_rf, function(data) betareg(formula = form,
data = data))

fits_pmm <- lapply(imputations_pmm, function(data) betareg(formula = form,
data = data))

fits_knn <- lapply(imputations_knn, function(data) betareg(formula = form,
data = data))

fits_reg <- lapply(imputations_Regresion, function(data) betareg(formula = form,
data = data))

rubin_pool_multivariate_simple <- function(coefs_list, vcovs_list, conf.level = 0.95) {

# --- 0. VALIDACIÓN Y PREPARACIÓN ---
if (length(coefs_list) != length(vcovs_list)) {
  stop("coefs_list y vcovs_list deben tener la misma longitud (m).")
}
m <- length(coefs_list)
if (m < 2) {
  stop("Se requieren al menos 2 imputaciones (m >= 2).")
}

coef_names <- names(coefs_list[[1]])

```

```

q <- length(coef_names)

if (!requireNamespace("MASS", quietly = TRUE)) {
  stop("El paquete 'MASS' es necesario para la robustez de las inversas matriciales. Por favor, instalelo")
}

# --- 1. CÁLCULO DE COMPONENTES DE RUBIN ---
coefs_mat <- do.call(cbind, coefs_list)
qbar <- rowMeans(coefs_mat)

Sigma_bar <- Reduce("+", vcovs_list) / m
diffs <- coefs_mat - qbar
B <- (diffs %*% t(diffs)) / (m - 1)

V_qbar <- Sigma_bar + (1 + 1/m) * B
se <- sqrt(diag(V_qbar))

W_inv <- tryCatch(solve(Sigma_bar), error = function(e) MASS::ginv(Sigma_bar))
V_inv <- tryCatch(solve(V_qbar), error = function(e) MASS::ginv(V_qbar))

# --- 2. MÉTRICAS MULTIVARIADAS GLOBALES ---
trace_B_W_inv <- sum(diag(B %*% W_inv))
r <- (1 + 1/m) * (1/q) * trace_B_W_inv

trace_B_V_inv <- sum(diag(B %*% V_inv))
lambda <- (1 + 1/m) * (1/q) * trace_B_V_inv

nu <- (q * (m - 1)) / r^2
gamma <- (1/(1+r)) * (r + 2/(nu + 3))

# --- 3. MÉTRICAS UNIVARIADAS ---
riv_vec <- (1 + 1/m) * diag(B) / diag(Sigma_bar)

df_vec <- (m - 1) * (1 + 1 / riv_vec)^2
df_vec[riv_vec == 0 | is.infinite(df_vec)] <- NA

lambda_vec <- riv_vec / (1 + riv_vec)
fmi_vec <- (riv_vec + 2 / (df_vec + 3)) / (1 + riv_vec)

t_val <- qbar / se
p_val <- 2 * pt(-abs(t_val), df = df_vec)

# --- 4. SALIDA ---
out <- data.frame(
  term = coef_names,
  estimate = as.numeric(qbar),
  std.error = as.numeric(se),
  t.value = as.numeric(t_val),
  df = as.numeric(df_vec),
  p.value = as.numeric(p_val),
)

```

```

riv_uni = as.numeric(riv_vec),
lambda_uni = as.numeric(lambda_vec),
fmi_uni = as.numeric(fmi_vec),
row.names = NULL,
stringsAsFactors = FALSE
)

# Atributo con métricas multivariadas globales
global_metrics <- list(
  m = m,
  q = q,
  r_multivariate = r,
  lambda_multivariate = lambda,
  df_multivariate = nu,
  gamma_multivariate = gamma,
  V_qbar = V_qbar
)

attr(out, "global_metrics") <- global_metrics
class(out) <- c("mi_pool_multivariate", class(out))

return(out)
}

# Lista con los nombres de los conjuntos de modelos
model_names <- c("rf", "knn", "pmm", "reg")

# Bucle para aplicar rubin_pool_cov automáticamente con las matrices de cov
for (name in model_names) {
  fits_obj <- get(paste0("fits_", name))
  coefs_list <- lapply(fits_obj, coef)
  vcovs_list <- lapply(fits_obj, vcov)
  pooled_result <- rubin_pool_multivariate_simple(coefs_list, vcovs_list)
  assign(paste0("pooled_", name), pooled_result)
}

get_stats <- function(datalist, varname) {

  # Validar que datalist no esté vacío
  if (length(datalist) < 2) stop("Se requieren al menos 2 imputaciones (m >= 2).")
  m <- length(datalist)

  # 1. Extraer medias y varianzas de la variable en cada imputación
  means <- sapply(datalist, function(data) mean(data[[varname]], na.rm=TRUE))
  vars <- sapply(datalist, function(data) var(data[[varname]], na.rm=TRUE))
}

```

```

# 2. Componentes de Rubin (q=1)
# MW (Within-imputation variance, W) - Varianza dentro
MW <- mean(vars)
# MB (Between-imputation variance, B) - Varianza entre
MB <- var(means)

# 3. Varianza total combinada (MT)
MT <- MW + (1 + 1/m) * MB

# 4. Relative Increase in Variance (RIV = r)
RIV <- (1 + 1/m) * MB / MW

# 5. Proporción de varianza debida a la imputación (lambda)
lambda <- RIV / (1 + RIV) # Equivalente a (1 + 1/m) * MB / MT

# 6. Grados de libertad (df) de Barnard & Rubin
df <- (m - 1) * (1 + 1/RIV)^2

# 7. Fraction of Missing Information (FMI = gamma)
FMI <- (RIV + 2/(df + 3)) / (1 + RIV)

# 8. Devuelve un listado de métricas clave
list(
  mean = mean(means),           # Media combinada
  std.error = sqrt(MT),         # Error estándar combinado
  lambda = lambda,              # Proporción de varianza imputada
  RIV = RIV,                   # Incremento relativo de la varianza
  FMI = FMI,                   # Fracción de información faltante
  df = df                      # Grados de libertad
)
}

# Calcular métricas para variable imputada con RF
metrics_rf <- get_stats(imputations_rf, "Proporción.IPM")
metrics_knn <- get_stats(imputations_knn, "Proporción.IPM")
metrics_pmm <- get_stats(imputations_pmm, "Proporción.IPM")
metrics_regresion <- get_stats(imputations_regresion, "Proporción.IPM")

```

## 3.1 Cálculo de las Medidas de Severidad: Reglas de Rubin

### 3.1.1 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. 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

Cuadro 2: Resultados modelo random forest

term	estimate	riv_uni	lambda_uni	fmi_uni
(Intercept)	-0.4434599	1.03	0.51	0.52
Cobertura.de.Internet....	-0.0072546	0.44	0.31	0.32
Densidad.poblacional	-0.0000369	0.10	0.09	0.09
X..población.urbana	-0.0018872	1.08	0.52	0.54
NBI...en.el.área.urbana	0.0052417	3.30	0.77	0.78
NBI...en.el.área.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.de.acueducto....	-0.0017752	0.67	0.40	0.41
Cobertura.de.alcantarillado....	0.0009914	1.59	0.61	0.63
Cobertura.de.Energía.Eléctrica....	-0.0038483	1.89	0.65	0.67
Cobertura.de.aseo....	-0.0038517	1.24	0.55	0.57
Cobertura.de.Gas.Natural....	-0.0005411	0.56	0.36	0.37
Cobertura.neta.en.educación	-0.0021448	1.15	0.54	0.55
X..de.inversión...Justicia.y.seguridad	-1.0024323	0.52	0.34	0.35
MDM.2018	-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

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 Information) 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.

```
pooled_rf[, c(7:9)] <- round(pooled_rf[, c(7:9)], 2)
pooled_rf[, c(1:2, 7:9)] |> kable(caption = "Resultados modelo random forest")
```

### 3.1.2 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. Los valores del aumento relativo en la varianza (RIV) presentan una amplia dispersión,

Cuadro 3: Resultados modelo knn

term	estimate	riv_uni	lambda_uni	fmi_uni
(Intercept)	-1.1657025	15.13	0.94	0.94
Cobertura.de.Internet....	-0.0072488	14.53	0.94	0.94
Densidad.poblacional	0.0000089	32.71	0.97	0.97
X..población.urbana	-0.0020249	11.22	0.92	0.93
NBI...en.el.área.urbana	0.0011689	4.91	0.83	0.85
NBI...en.el.área.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.de.acueducto....	-0.0007879	10.87	0.92	0.92
Cobertura.de.alcantarillado....	0.0037244	7.34	0.88	0.89
Cobertura.de.Energía.Eléctrica....	-0.0021234	2.85	0.74	0.76
Cobertura.de.aseo....	-0.0041451	6.00	0.86	0.87
Cobertura.de.Gas.Natural....	-0.0010795	1.57	0.61	0.63
Cobertura.neta.en.educación	-0.0022956	5.61	0.85	0.86
X..de.inversión...Justicia.y.seguridad	-0.8797645	1.98	0.66	0.68
MDM.2018	-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

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.

```
pooled_knn[, c(7:9)] <- round(pooled_knn[, c(7:9)], 2)
pooled_knn[, c(1:2, 7:9)] |> kable(caption = "Resultados modelo knn")
```

### 3.1.3 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 estadísticamente estables y poco afectadas por los datos faltantes. 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

Cuadro 4: Resultados modelo Pmm

term	estimate	riv_uni	lambda_uni	fmi_uni
(Intercept)	-0.3949127	0.47	0.32	0.33
Cobertura.de.Internet....	-0.0078325	0.12	0.11	0.11
Densidad.poblacional	-0.0000472	0.07	0.06	0.07
X..población.urbana	-0.0016074	0.24	0.20	0.20
NBI...en.el.área.urbana	0.0057036	0.81	0.45	0.46
NBI...en.el.área.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.de.acueducto....	-0.0023195	0.29	0.23	0.23
Cobertura.de.alcantarillado....	0.0019090	0.10	0.09	0.09
Cobertura.de.Energía.Eléctrica....	-0.0038566	1.11	0.53	0.54
Cobertura.de.aseo....	-0.0046603	0.14	0.12	0.12
Cobertura.de.Gas.Natural....	-0.0006479	0.18	0.15	0.16
Cobertura.neta.en.educación	-0.0018700	0.26	0.21	0.21
X..de.inversión...Justicia.y.seguridad	-0.8747598	0.43	0.30	0.31
MDM.2018	-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

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.

```
pooled_pmm[, c(7:9)] <- round(pooled_pmm[, c(7:9)], 2)
pooled_pmm[, c(1:2, 7:9)] |> kable(caption = "Resultados modelo Pmm")
```

### 3.1.4 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. 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.

Cuadro 5: Resultados modelo Regresión

term	estimate	riv_uni	lambda_uni	fmi_uni
(Intercept)	-0.4154798	0.37	0.27	0.28
Cobertura.de.Internet....	-0.0078746	0.20	0.17	0.17
Densidad.poblacional	-0.0000473	0.14	0.13	0.13
X..población.urbana	-0.0017531	0.26	0.20	0.21
NBI...en.el.área.urbana	0.0059630	0.67	0.40	0.42
NBI...en.el.área.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.de.acueducto....	-0.0022429	0.21	0.17	0.18
Cobertura.de.alcantarillado....	0.0019874	0.12	0.11	0.11
Cobertura.de.Energía.Eléctrica....	-0.0035941	1.32	0.57	0.59
Cobertura.de.aseo....	-0.0046234	0.33	0.25	0.25
Cobertura.de.Gas.Natural....	-0.0006511	0.06	0.06	0.06
Cobertura.neta.en.educación	-0.0017992	0.30	0.23	0.24
X..de.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

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.

```
pooled_reg[, c(7:9)] <- round(pooled_reg[, c(7:9)], 2)
pooled_reg[, c(1:2, 7:9)] |> kable(caption = "Resultados modelo Regresión")
```

### 3.1.5 ¿Cual modelo elegir?

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.

Desde el punto de vista empírico, los resultados obtenidos mediante las reglas de Rubin confirman la robustez del PMM: 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 & Groothuis-Oudshoorn, 2011; Morris et al., 2014). 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.

### 3.1.6 modelo PMM métricas de Rubin

```
pooled_pmm[, -1] <- round(pooled_pmm[, -1], 2)
```

```
pooled_pmm
```

		term	estimate	std.error	t.value	df
1		(Intercept)	-0.39	0.17	-2.35	136.55
2		Cobertura.de.Internet....	-0.01	0.00	-5.95	1194.39
3		Densidad.poblacional	0.00	0.00	-2.35	3356.85
4		X..población.urbana	0.00	0.00	-1.81	364.57
5		NBI...en.el.área.urbana	0.01	0.00	5.02	70.07
6		NBI...en.el.área.rural	0.01	0.00	15.39	230.85
7		Valor.agregado.municipal....	0.00	0.00	-0.35	64.18
8		DEFVIV	0.98	0.09	10.87	1534.60
9		Cobertura.de.acueducto....	0.00	0.00	-3.35	271.82
10		Cobertura.de.alcantarillado....	0.00	0.00	2.45	1706.77
11		Cobertura.de.Energía.Eléctrica....	0.00	0.00	-2.75	50.49
12		Cobertura.de.aseo....	0.00	0.00	-5.41	982.45
13		Cobertura.de.Gas.Natural....	0.00	0.00	-1.19	596.73
14		Cobertura.neta.en.educación	0.00	0.00	-2.47	325.52
15	X..de.inversión...Justicia.y.seguridad	-0.87	0.47	-1.85	152.61	
16		MDM.2018	-0.29	0.14	-2.05	376.32
17		Cobertura.salud	0.20	0.06	3.24	292.35
18		(phi)	55.04	2.69	20.46	200.20
		p.value	riv_uni	lambda_uni	fmi_uni	
1	0.02	0.47	0.32	0.33		
2	0.00	0.12	0.11	0.11		
3	0.02	0.07	0.06	0.07		
4	0.07	0.24	0.20	0.20		
5	0.00	0.81	0.45	0.46		
6	0.00	0.33	0.25	0.25		

7	0.73	0.88	0.47	0.48
8	0.00	0.11	0.10	0.10
9	0.00	0.29	0.23	0.23
10	0.01	0.10	0.09	0.09
11	0.01	1.11	0.53	0.54
12	0.00	0.14	0.12	0.12
13	0.23	0.18	0.15	0.16
14	0.01	0.26	0.21	0.21
15	0.07	0.43	0.30	0.31
16	0.04	0.24	0.19	0.20
17	0.00	0.28	0.22	0.22
18	0.00	0.36	0.26	0.27

### 3.1.7 Modelo imputado

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. 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  $\mu_i$ . Este alto valor de  $\phi$  ( $\hat{\phi} > 1$ ) 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 ( $\mu_i$ ), 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 ( $RIV \approx 0.03$  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.

```

modelo_beta_imputado <- with(
  imp_pmm,
  betareg(
    Proporción.IPM ~ Cobertura.de.Internet.... + Densidad.poblacional +
    X..población.urbana +
    NBI...en.el.área.urbana + NBI...en.el.área.rural +
    Valor.agregado.municipal.... + DEFVIV +
    Cobertura.de.acueducto.... + Cobertura.de.alcantarillado.... +
    Cobertura.de.Energía.Eléctrica.... + Cobertura.de.aseo.... +
    Cobertura.de.Gas.Natural.... +
    Cobertura.neta.en.educación + X..de.inversión...Justicia.y.seguridad +
    MDM.2018 + Cobertura.salud,
    link = "logit" # También puedes usar "probit", "cloglog", etc.
  )
)

modelo_beta_imputado[[1]]

```

```

with.mids(data = imp_pmm, expr = betareg(Proporción.IPM ~ Cobertura.de.Internet.... +
  Densidad.poblacional + X..población.urbana + NBI...en.el.área.urbana +
  NBI...en.el.área.rural + Valor.agregado.municipal.... +
  DEFVIV + Cobertura.de.acueducto.... + Cobertura.de.alcantarillado.... +
  Cobertura.de.Energía.Eléctrica.... + Cobertura.de.aseo.... +
  Cobertura.de.Gas.Natural.... + Cobertura.neta.en.educación +
  X..de.inversión...Justicia.y.seguridad + MDM.2018 + Cobertura.salud,
  link = "logit"))

```

```

# Extraer residuales de cada imputación (tipo Pearson)
resid_imputado_list <- lapply(modelo_beta_imputado$analyses,
                                function(m) residuals(m, type = "pearson"))

fit_imputado_list <- lapply(modelo_beta_imputado$analyses,
                            function(m) fitted(m))

```

## 4 Análisis de residuales

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:

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

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** ( $\mu_i$ ) y de **homocedasticidad** (varianza constante o adecuadamente modelada) se cumplen de manera consistente. Este hallazgo respalda

la estabilidad del parámetro de precisión ( $\phi$ ) 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.

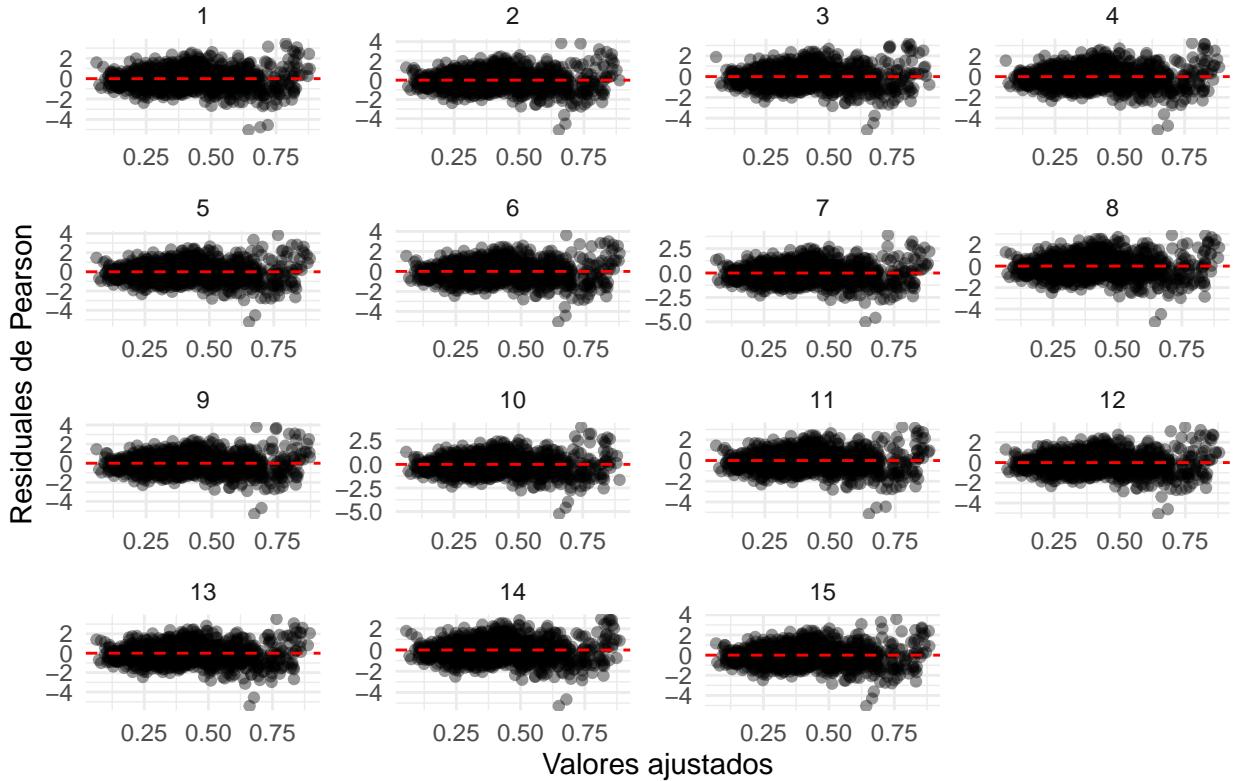
En conjunto, los resultados del diagnóstico indican que los modelos de Regresión Beta ajustados sobre los conjuntos imputados son **estadísticamente robustos y consistentes**. La estabilidad de los residuales, la ausencia de patrones sistemáticos y la adecuada aproximación a la normalidad confirman que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

```
# Combinar en un solo data.frame
resid_df <- do.call(rbind, lapply(seq_along(resid_imputado_list), function(i) {
  data.frame(
    imputacion = i,
    residuales = resid_imputado_list[[i]],
    fitted = fit_imputado_list[[i]]
  )
}))
```

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, de manera satisfactoria, 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 los modelos empleados para el proceso de imputación son estadísticamente válidos y los conjuntos de datos resultantes son apropiados para el subsiguiente análisis combinado.

```
ggplot(resid_df, aes(x = fitted, y = residuales)) +
  geom_point(alpha = 0.4) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  facet_wrap(~imputacion, scales = "free") +
  labs(x = "Valores ajustados", y = "Residuales de Pearson",
       title = "Residuales vs Ajustados por imputación") +
  theme_minimal()
```

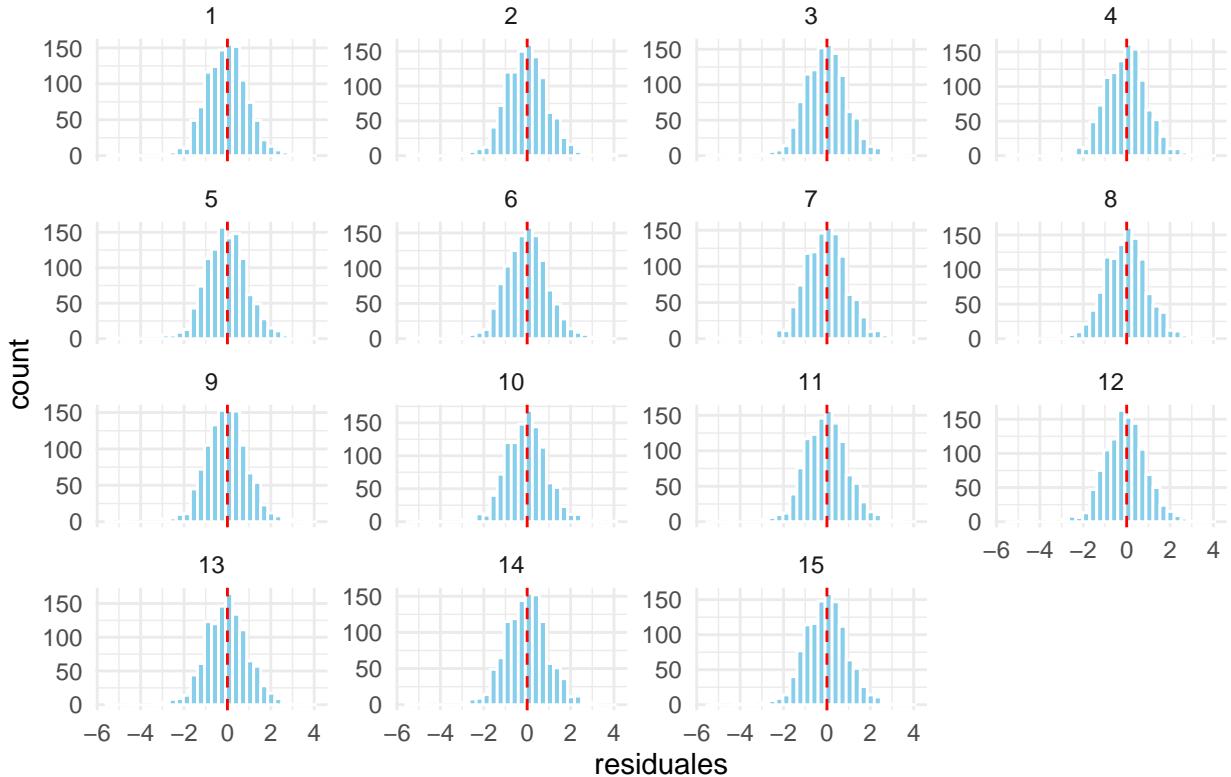
## Residuales vs Ajustados por imputación



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.

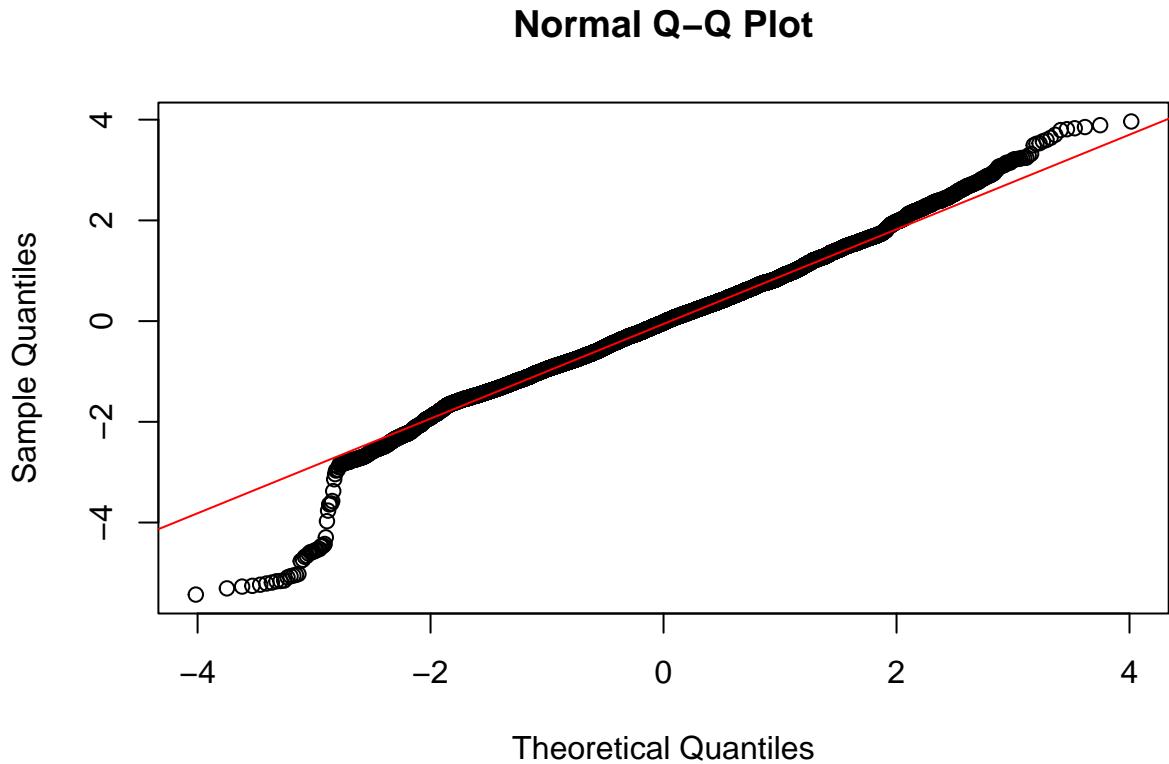
```
ggplot(resid_df, aes(x = residuales)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "white") +
  facet_wrap(~imputacion, scales = "free_y") +
  geom_vline(xintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Distribución de residuales de Pearson por imputación") +
  theme_minimal()
```

## 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). 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.

```
qqnorm(resid_df$residuales)
qqline(resid_df$residuales, col = "red")
```



#### 4.1 Análisis de Residuales Cuantil ( $r_{Q,i}$ )

El análisis de los **Residuales Cuantil** es una extensión fundamental en el diagnóstico de modelos de **Regresión Beta**, ya 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:

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

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.

```

# El objeto 'modelo_beta_imputado$analyses' contiene la lista de los modelos ajustados

# Extraer Residuales Cuantil de cada imputación
resid_quantile_list <- lapply(modelo_beta_imputado$analyses,
                                function(m) residuals(m, type = "quantile"))

# Extraer Valores Ajustados (Fitted Values) de cada imputación (útil para el R vs. Aju
fit_imputado_list <- lapply(modelo_beta_imputado$analyses,
                             function(m) fitted(m))

# Combinar los resultados en un solo data.frame para graficar
resid_df_quantile <- do.call(rbind, lapply(seq_along(resid_quantile_list), function(i) +
  data.frame(
    imputacion = i,
    residuales = resid_quantile_list[[i]],
    fitted = fit_imputado_list[[i]]
  )
))
}

# Para simplificar el nombre de la columna:
names(resid_df_quantile)[2] <- "r_Qi"

```

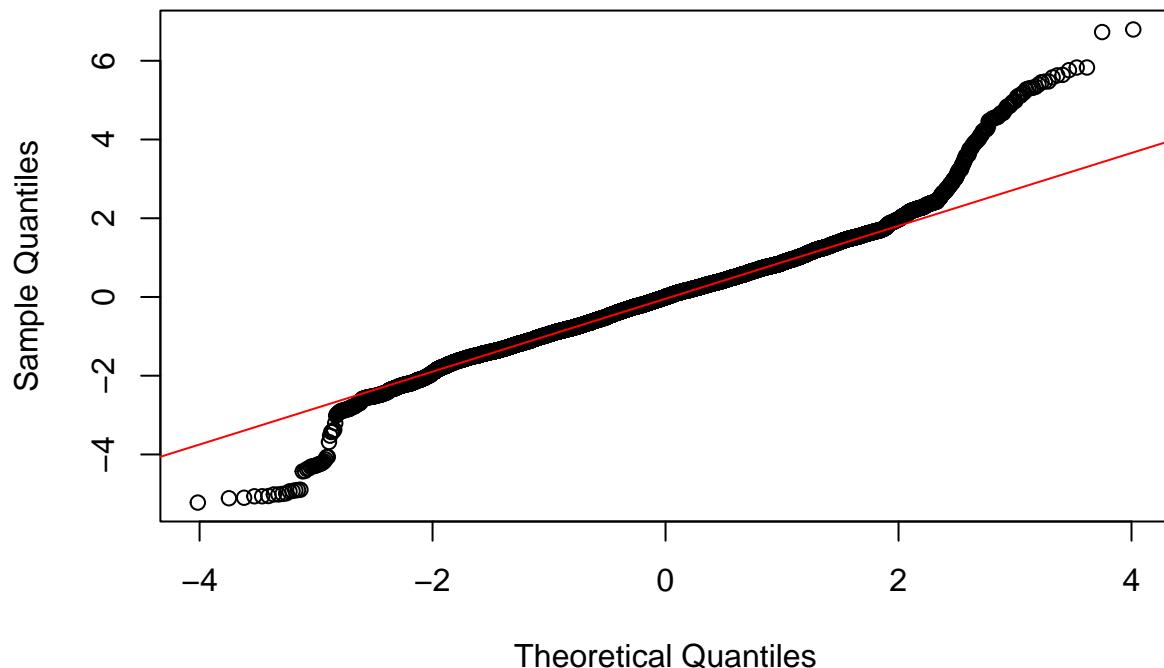
Si bien existe una excelente alineación con la línea recta de referencia en el rango central (cerca de cero), lo cual sugiere que la mayoría de los errores son adecuadamente modelados, se observan fuertes y preocupantes desviaciones en ambas colas. Específicamente, en la cola inferior (izquierda), los 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.

```

# Gráfico Q-Q Normal para todos los residuales cuantil agregados
qqnorm(resid_df_quantile$r_Qi,
       main = "Gráfico Q-Q Normal (Residuales Cuantil Agregados)")
qqline(resid_df_quantile$r_Qi, col = "red")

```

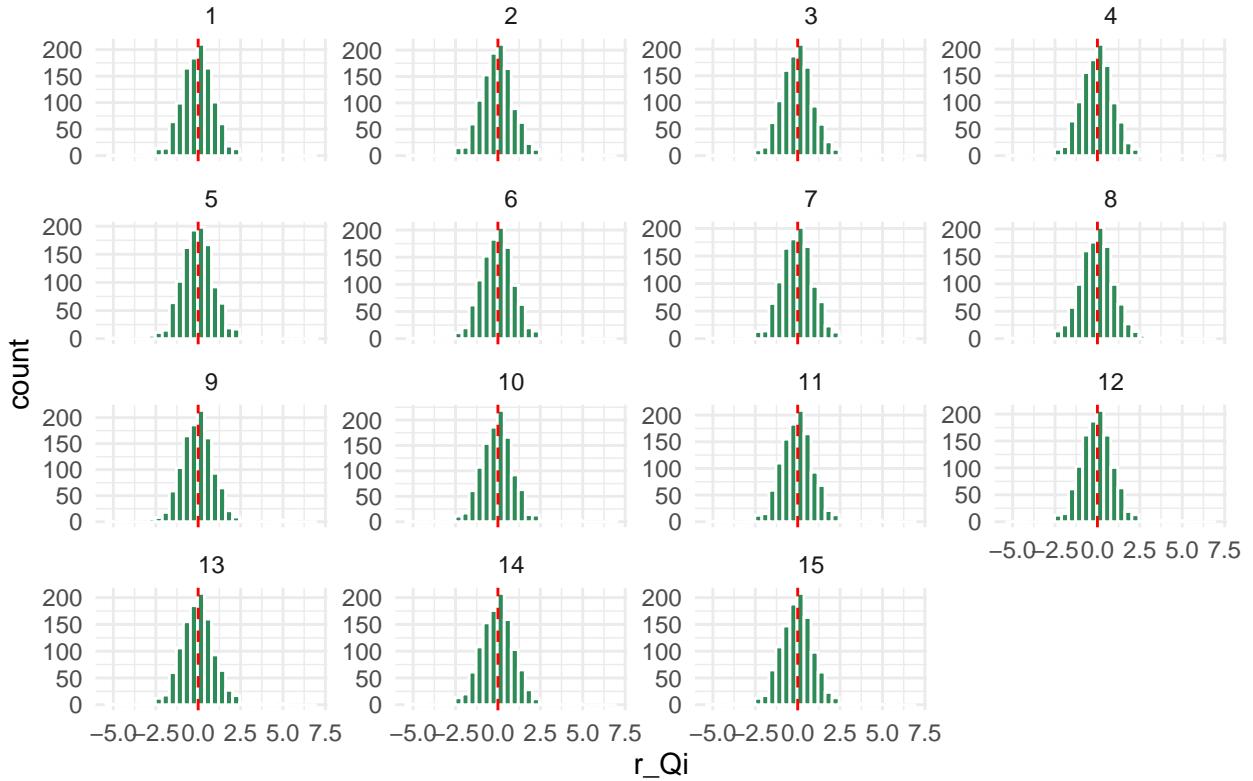
## Gráfico Q-Q Normal (Residuales Cuantil Agregados)



El análisis de la Distribución de Residuales Cuantil por imputación muestra un resultado diagnóstico excelente y muy robusto 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 satisfactoriamente. 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.

```
ggplot(resid_df_quantile, aes(x = r_Qi)) +  
  geom_histogram(bins = 30, fill = "seagreen", color = "white") +  
  facet_wrap(~imputacion, scales = "free_y") +  
  geom_vline(xintercept = 0, color = "red", linetype = "dashed") +  
  labs(title = "Distribución de Residuales Cuantil por Imputación") +  
  theme_minimal()
```

## Distribución de Residuales Cuantil por Imputación



## 4.2 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. El diagnóstico aquí presentado es crucial para asegurar que la inferencia no dependa de un puñado de puntos de datos atípicos.

```
# Extraer la Distancia de Cook de cada imputación
cooks_distance_list <- lapply(modelo_beta_imputado$analyses,
                                function(m) cooks.distance(m))

# Extraer los Residuales Cuantil para el gráfico combinado (opcional, pero útil)
resid_quantile_list <- lapply(modelo_beta_imputado$analyses,
                                function(m) residuals(m, type = "quantile"))

# Combinar los resultados en un solo data.frame
influence_df <- do.call(rbind, lapply(seq_along(cooks_distance_list), function(i) {
  data.frame(
    imputacion = i,
    observacion = 1:length(cooks_distance_list[[i]]), # Asume que el orden de las obser
    cooks_d = cooks_distance_list[[i]],
    r_Qi = resid_quantile_list[[i]]
  )
})
```

}))

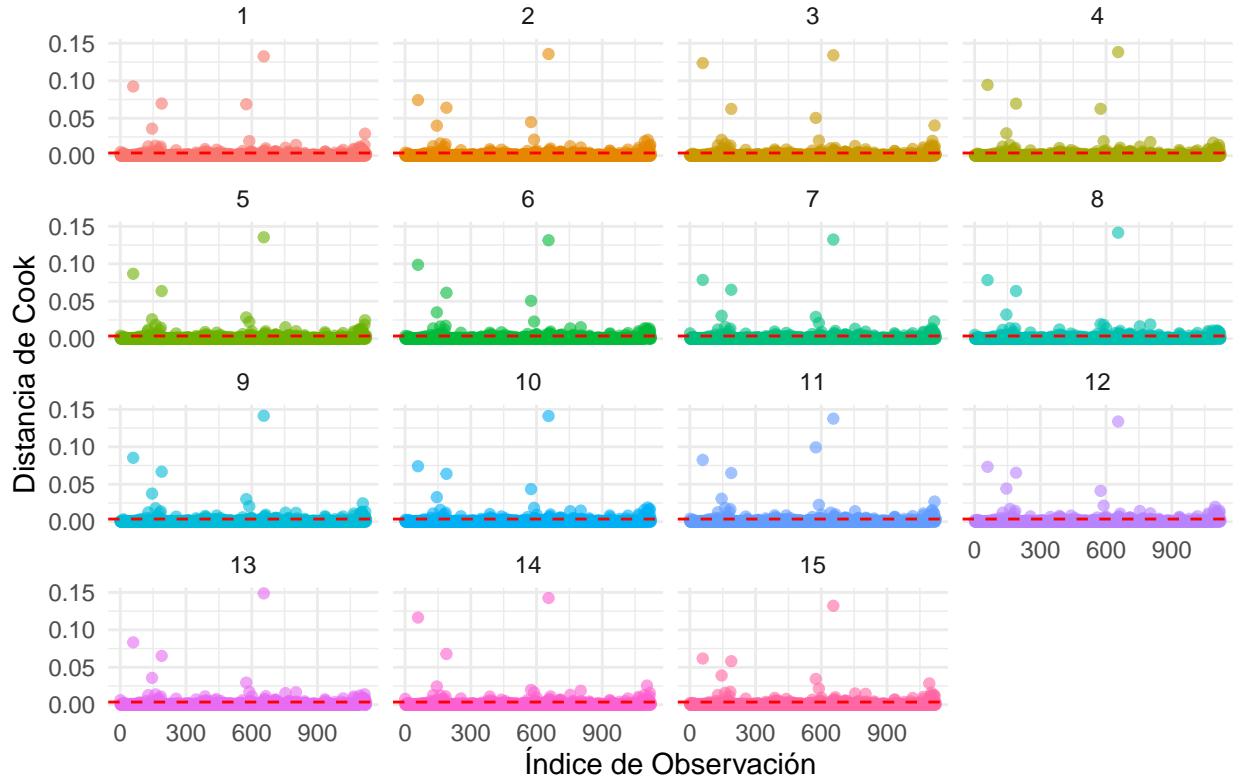
El gráfico de la Distancia de Cook por Observación y por Imputación indica que los modelos de Regresión Beta son robustos y estables ante la influencia de las observaciones individuales. La gran mayoría de los puntos en las 15 imputaciones se agrupan consistentemente cerca del eje cero y no superan un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

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.

```
# Calcular un umbral de referencia. Un umbral común es 4/n (o 4/n_efectivo)
n_obs_total <- nrow(modelo_beta_imputado$analyses[[1]]$model) # Asumiendo mismo n en cada imputacion
umbral_cook <- 4 / n_obs_total

ggplot(influence_df, aes(x = observacion, y = cooks_d)) +
  geom_point(aes(color = as.factor(imputacion)), alpha = 0.6) +
  geom_hline(yintercept = umbral_cook, color = "red", linetype = "dashed") +
  labs(title = "Distancia de Cook por Observación y por Imputación",
       x = "Índice de Observación",
       y = "Distancia de Cook") +
  theme_minimal() +
  facet_wrap(~imputacion) +
  theme(legend.position = "none")
```

## Distancia de Cook por Observación y por Imputación



## 5 tabla de comparativa de los 20 municipios

muestra una alta y consistente pobreza a nivel municipal en los departamentos seleccionados, con proporciones observadas (Proporción IPM) que oscilan entre 0.775 y 0.985. Las predicciones del modelo (predichos\_pmm) son robustas y se sitúan consistentemente en un rango estrecho (principalmente entre 0.76 y 0.85), lo que indica que el modelo logra estimar el nivel de pobreza de manera confiable. Los Intervalos de Confianza (CI) del 95% (ci\_lower y ci\_upper) 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 (prop\_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.

```
# Convertir lista de fitted values en matriz
fit_matrix <- do.call(cbind, fit_imputado_list)

# Número de imputaciones
m <- length(fit_imputado_list)

# Predicción promedio por observación
fit_mean <- rowMeans(fit_matrix)

# Varianza total según reglas de Rubin
within_var <- apply(fit_matrix, 1, function(x) mean(rep(0, m))) # para fitted values no
between_var <- apply(fit_matrix, 1, var) # varianza entre imputaciones
```

```

total_var <- within_var + (1 + 1/m) * between_var

# Intervalos de confianza 95%
ci_lower <- fit_mean - 1.96 * sqrt(total_var)
ci_upper <- fit_mean + 1.96 * sqrt(total_var)

# Asignar predicciones y CI al data frame
datos$predichos_pmm <- fit_mean
datos$ci_lower <- ci_lower
datos$ci_upper <- ci_upper

# Filtrar municipios específicos
municipios <- c(91263,91405,91407,91430,91460,91530,91536,
                 91669,91798,94343,94663,94883,94884,
                 94885,94886,94887,94888,97511,97777,97889)

# Calcular proporción de NA por fila
datos$prop_NA <- apply(datos, 1, function(x) mean(is.na(x))) |> round(2)

datos[, c(22:25)] <- round(datos[, c(22:25)], 2)

datos_filtrados <- datos[datos$CODIGO %in% municipios, c(3:4, 22:25)]

kable(datos_filtrados)

```

Municipio	Proporción IPM	predichos_pmm	ci_lower	ci_upper	prop_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
PAPUNAUUA	0.979	0.83	0.76	0.91	0.67
YAVARATE	0.948	0.84	0.80	0.89	0.67

## 6 Mapa final

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: 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.

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.

```
# Promedio de residuales y ajustados
resid_imputado <- Reduce("+", resid_imputado_list) / length(resid_imputado_list)
fit_imputado <- Reduce("+", fit_imputado_list) / length(fit_imputado_list)

# 7. El gráfico de comparativa de los resultados
## añadir los valores predichos
datos$pmm <- fit_imputado

## cargar shp con los municipios de colombia

colombia <- read_sf("co_2018_MGN_MPIO_POLITICO.geojson")

colnames(colombia) <- c("DPTO_CCDGO", "MPIO_CCDGO", "MPIO_CNMBR", "MPIO_CRSLC", "MPIO_NA"
                        "CODIGO", "MPIO_NANO", "DPTO_CNMBR", "Shape_Leng", "Shape_Area",
                        "geometry")

## merge con nuestros datos
colombia <- merge(datos[, c(1, 3:7, 22)], colombia, by = "CODIGO")

colombia <- colombia |> st_as_sf()

grafico1 <- ggplot(colombia, aes(fill = colombia$`Proporción IPM`)) + geom_sf() +
  labs(fill = "", title = "Datos originales 'Proporción IPM'") +
  coord_sf(
    xlim = c(-79.1, -66.8), # Longitudes aproximadas (Oeste a Este)
    ylim = c(-4.5, 13),     # Latitudes aproximadas (Sur a Norte)
    expand = FALSE)
```

```

) + scale_fill_stepsn(
  colours = scico::scico(11, palette = "vik"),
  guide = guide_colorbar(direction = "horizontal", title.position = "top", title.hjust
) + theme_minimal() +
annotation_scale(location = "bl") +
theme(legend.position = c(0.8, 0.8),
      legend.background = element_blank())

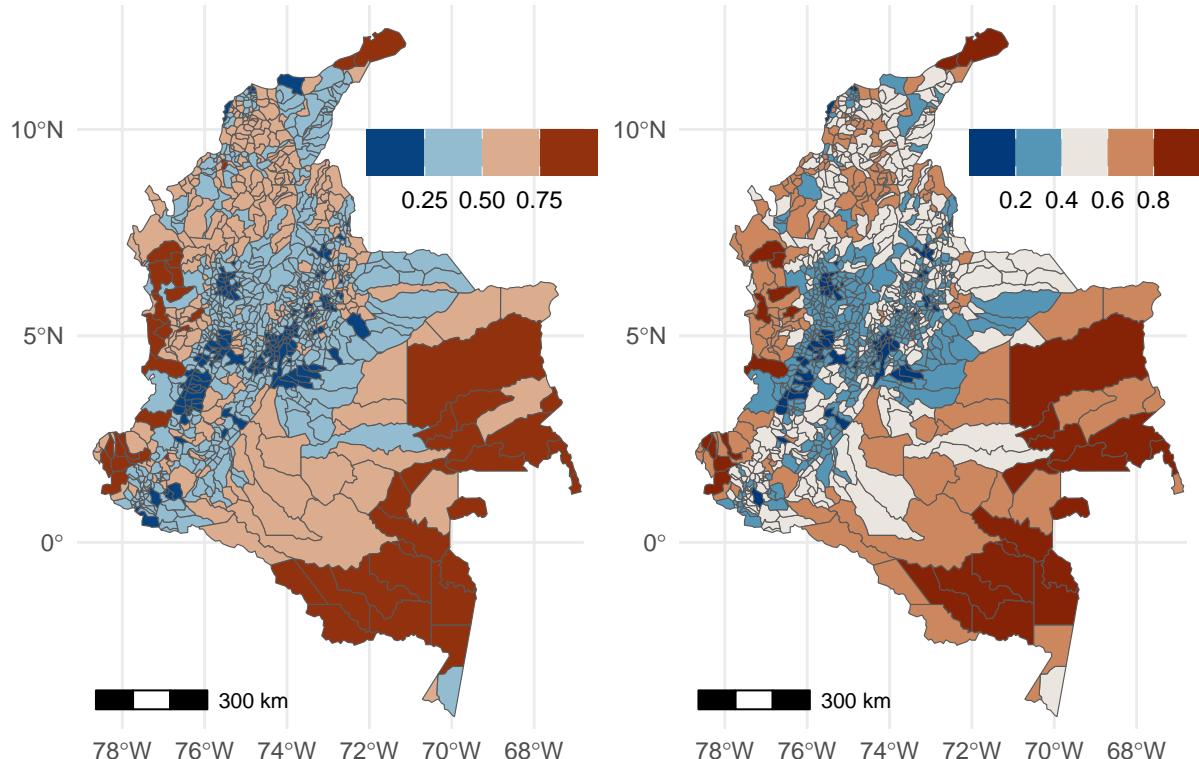
grafico2 <- ggplot(colombia, aes(fill = colombia$predichos_pmm)) +
  geom_sf() + labs(fill = "", title = "IPM Predicho con modelo") +
  coord_sf(
    xlim = c(-79.1, -66.8), # Longitudes aproximadas (Oeste a Este)
    ylim = c(-4.5, 13),     # Latitudes aproximadas (Sur a Norte)
    expand = FALSE
) + scale_fill_stepsn(
  colours = scico::scico(11, palette = "vik"),
  guide = guide_colorbar(direction = "horizontal", title.position = "top", title.hjust
) + theme_minimal() +
annotation_scale(location = "bl") +
theme(legend.position = c(0.8, 0.8),
      legend.background = element_blank())

```

library(patchwork)

grafico1 | grafico2

Datos originales 'Proporción IPM'      IPM Predicho con modelo



## 7 Conclusión

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 alta 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 alta precisión. 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 Pearson y el análisis de Distancia de Cook confirmaron la robustez estadística de las estimaciones, cumpliéndose los supuestos de linealidad y homocedasticidad, y demostrando una baja influencia de las observaciones atípicas. La evaluación rigurosa a través de los Residuales Cuantil confirmó una excelente aproximación a la normalidad en el rango central, un requisito clave para la validez de la inferencia.

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. Finalmente, 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.

### 7.1 Conclusiones y Proyecciones

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. 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: la desviación en las colas de los Residuales Cuantil y la alta incertidumbre en los CIs de los municipios con más datos faltantes requieren atención. En conclusión, 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.

## 8 Referencia

- Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1–67.
- Little, R. J. A. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, 6(3), 287–296.

- Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14(1), 75.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.
- Seaman, S. R., White, I. R., Copas, A. J., & Li, L. (2012). Combining multiple imputation and inverse-probability weighting. *Biostatistics*, 13(4), 734–745.