# Estimación de los coeficientes de regresión en áreas pequeñas utilizando valores plausibles en muestras probabilísticas

**Stalyn Yasid Guerrero Gómez**

Universidad Nacional de Colombia
Facultad de Ciencias,
Departamento de Estadística
2019

# Estimación de los coeficientes de regresión en áreas pequeñas utilizando valores plausibles en muestras probabilísticas

**Stalyn Yasid Guerrero Gómez**

Tesis como requisito parcial para optar al título de:
**Magister en Ciencias - Estadísticas**

Director:
Ph.D. Leonardo Trujillo

Universidad Nacional de Colombia
Bogotá, Colombia
2019

# Resumen

Los diseños de muestreo nacionales tienen por objetivo asegurar que las inferencias de totales y tasas para los niveles de agregación nacional, regional y departamental puedan ser obtenidas con una precisión preestablecida. Sin embargo, cuando se hacen desagregaciones de la población en dominios no planeados estás precisiones pueden verse afectadas. Por lo tanto, haciendo uso de inferencias basadas en modelos es posible obtener una predicción de los parámetros de interés en estos dominios. Ahora, cuando se desea explicar la relación lineal de la variable $Y$ con las covariables $X$'s, en dominios no planeados y las cuales presentan información perdida en los dominios planeados hizo necesario proponer una metodología donde es posible emplear los valores plausibles para estimar los coeficientes que relacionan a $Y$ y $X$; para luego emplear estas estimaciones como insumo de las inferencias basadas en modelos y obtener una predicción de los coeficientes en los dominios no planeados.
**Palabras clave: Regresión lineal, valores plausibles, estimación basada en modelos** .

# Abstract

National sampling designs are intended to ensure that total and rate inferences for national, regional and departmental levels of aggregation can be obtained with pre-established accuracy. However, when disaggregations of the population are made in unplanned domains hese accuracies can be affected. Therefore, using model-based inferences it is possible to obtain a prediction of the parameters of interest in these domains. Now, when you want to explain the linear relationship of the variable $Y$ with the covariates $X$'s, in unplanned domains and which present information lost in the planned domains made it necessary to propose a methodology where it is possible to use plausible values to estimate the coefficients that relate to $Y$ and $X$; then use these estimates as input of the inferences based on models and obtain a prediction of the coefficients in the unplanned domains.
**Keywords: Linear regression, plausible values, model-based estimation**

# Contenido

**Resumen.** III

**Notación.** IX

**Introducción.** XIV

**1. Marco teórico** 2

1.1. Definiciones de introducción al muestreo . . . . . 2
1.2. Métodos de estimación para muestras probabilísticas . . . . . 4
1.3. Modelo de regresión múltiple en muestras probabilísticas . . . . . 6
1.4. Estimación en áreas pequeñas . . . . . 8
1.5. El problema de datos faltantes en las encuestas y sus posibles soluciones. . . 12

**2. Metodología** 21

2.1. Prueba Saber 3°, 5° y 9° . . . . . 21
2.2. Estimación de los parámetros de un modelo lineal en dominios no observados. 25

**3. Estudio por simulación** 28

3.1. Caso de interés en la simulación . . . . . 28
3.2. Simulación de la población . . . . . 30
3.3. Simulación de datos faltantes. . . . . 31
3.4. Selección de la muestra . . . . . 32
3.5. Métodos de imputación . . . . . 32
3.6. Diseño del estudio de simulación . . . . . 34
3.7. Medida de calidad de las estimaciones. . . . . 34
3.8. Resultados de simulación . . . . . 35

**4. Aplicación** 49

4.1. Cuestionario del rector o coordinador de sede . . . . . 49
4.2. Procesamiento de los datos . . . . . 50
4.3. Modelo de regresión en los departamentos . . . . . 52

4.4. Predicción de los coeficientes del modelo lineal empleando el modelo Fay-Herriot. 53

**5. Conclusiones y recomendaciones** **58**

**A. Anexos** **60**

**Anexos A.** **60**

**Bibliografía.** **68**

# Lista de Tablas


**1-1**. Listado de paquetes en $R$ que realizan imputación . . . . . 18

**3-1**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$ y muestra seleccionada por dominio del 20 %. . . . . 37
**3-2**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$, muestra seleccionada por dominio del 20 %. . . . . 39
**3-3**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$ cuando el $n$ esperado de la muestra es fijo. . . . . 40
**3-4**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$ y el $n$ esperado de la muestra es fijo. . . . . 41
**3-5**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, porcentaje de valores faltantes del 10 % y $\rho_x \in (0{,}25, 0{,}50]$ . . . . . 43
**3-6**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, porcentajes de valores faltantes del 10 % y $\rho_x \in (0{,}50, 0{,}75]$ . . . . . 44
**3-7**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, porcentaje de valores faltantes del 10 % y $\rho_x \in (0{,}75, 1]$ . . . . . 45
**3-8**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, porcentaje de valores faltantes del 30 % y $\rho_x \in (0{,}25, 0{,}50]$ . . . . . 46
**3-9**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, porcentaje de valores faltantes del 30 % y $\rho_x \in (0{,}50, 0{,}75]$ . . . . . 47
**3-10**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, porcentaje de valores faltantes del 30 % y $\rho_x \in (0{,}75, 1]$ . . . . . 48

**4-1**. Distribución promedio de establecimientos educativos por departamento en la muestra . . . . . 50
**4-2**. Correlación del índice con los resultados de las pruebas Saber 5° . . . . . 51
**4-3**. Ítems para el factor subyacente . . . . . 52
**4-4**. Estimación de $\hat{\boldsymbol{\beta}}^{(IM)}$ y $\widehat{Var}\left(\hat{\boldsymbol{\beta}}^{(IM)}\right)$ por departamento . . . . . 54
**4-5**. Parámetros estimados para el modelo de Fay-Herriot . . . . . 56

**4-6**. Predicciones del parámetro $\boldsymbol{\beta}$ con el modelo de Fay-Herriot . . . . . 57

**A-1**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, $\rho_x \in (0{,}50, 0{,}75]$ y $\rho_z = 0{,}6$, en muestras de tamaño fijo y muestras del 20 % por dominio. . . . . 60

**A-2**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, $\rho_x \in (0{,}50, 0{,}75]$,$\rho_z = 0{,}6$, el 10 % de datos faltantes y muestras del 20 % por dominio. . . . . 60

**A-3**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$, porcentaje de valores faltantes del 10 % y $\rho_x \in (0{,}25, 0{,}50]$ . . . . . 61

**A-4**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$, porcentaje de valores faltantes del 10 % y $\rho_x \in (0{,}50, 0{,}75]$ . . . . . 62

**A-5**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$, porcentaje de valores faltantes del 10 % y $\rho_x \in (0{,}75, 1]$ . . . . . 63

**A-6**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$, porcentaje de valores faltantes del 30 % y $\rho_x \in (0{,}25, 0{,}50]$ . . . . . 64

**A-7**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$, un porcentaje de valores faltantes del 30 % y $\rho_x \in (0{,}50, 0{,}75]$ . . . . . 65

**A-8**. Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$, porcentaje de valores faltantes del 30 % y $\rho_x \in (0{,}75, 1]$ . . . . . 66

# Lista de Figuras

**3-1**. Sesgo relativo por dominio para $\beta_1$ cuando $\rho_x \in (0{,}50, 0{,}75]$ y $\rho_z = 0{,}6$ . . . 38
**3-2**. Representación gráfica de la raíz del error cuadrático medio para una muestra de 150 dominios con el 30 % de datos faltantes. . . . . 48

**4-1**. Lineas de regresión estimadas con el estimador directo y el modelo multivariado de Fay-Herriot . . . . . 56

**A-1**. Sesgos relativos promedios por dominio para $\beta_1$ cuando $P = 3$, $\rho_x \in (0{,}50, 0{,}75]$, $\rho_z = 0{,}6$ y un 10 % de datos faltantes. . . . . 67

# Notación

- $U$: Población finita.
- $N$ : Número total de observaciones en $U$
- $i, l$ : Rótulo que identifica las observaciones en $U$.
- $U_d$ : Sub-población o dominios poblacionales en que se divide $U$.
- $N_d$ : Número total de observaciones en $U_d$.
- $D$ : Número total de dominios en los que se divide $U$.
- $d$ : Rótulo que identifica los dominio de $U$.
- $P$ : Número total de variables regresoras en un modelo lineal.
- $p$: Rótulo que identifica a la $p-$ésima variable regresora en un modelo lineal.
- $R$ : Número total de variables de interés.
- $r$ : Rótulo que identifica a la $r-$ésima variable de interés.
- $K$ : Número total de variables auxiliares asociadas al $d-$ésimo dominio de $U$.
- $k$ : Rótulo que identifica a la $k-$ésima variable auxiliar asociada al $d-$ésimo dominio de $U$.
- $M$ : Número total de valores plausibles obtenidos para un valor faltante.
- $m$ : Rótulo que identifica el $m-$ésimo valor plausible para un valor faltante.
- $S$ : Muestra aleatoria.
- $s$: Una realización aleatoria de $S$.
- $Q$: Representa el conjunto de todas las muestras posibles que se pueden obtener empleando un método de selección, es decir, $Q = \{s_1.s_2, \ldots, s_q, \ldots, s_{Q_n}\}$.

- $Q_n$: Número de muestras posibles contenidas en $Q$.
- $q$ : Rótulo que identifica a la $q-$ésima muestra de $Q$.
- $n_s$ : Número de observaciones que conforman $S$.
- $S_d$ : Sub conjunto de observaciones de $S$ que pertenecen a $U_d$, es decir $S_d = S \cap U_d$.
- $n_d$ : Número de observaciones que conforman $S_d$.
- $p(\cdot)$ : Diseño de muestreo probabilístico arbitrario empleado para seleccionar $S$.
- $\pi_i$ : Probabilidad de que el elemento $i-$ésimo pertenezca $S$
- $\pi_{il}$ : Probabilidad de que los elementos $i$ y $l$ pertenezcan a $S$ simultáneamente.
- $\boldsymbol{W} = \text{diag}(w_1, \ldots, w_i, \ldots, w_{n_s})$ es una matriz diagonal de $n_s \times n_s$ donde $w = \pi_i^{-1}$ representa los factores de expansión.
- $\boldsymbol{Y}$ : Matriz de tamaño $N \times R$ que representa las variables de interés en $U$.
- $\boldsymbol{Y}_r$ : Vector columna de tamaño $N \times 1$ que representa la $r-$ésima variable de interés en $U$.
- $\boldsymbol{Y}_i$ : Vector fila de tamaño $1 \times R$ que representa el $i-$ésimo elemento en $U$.
- $\boldsymbol{Y}_s$ : Matriz de tamaño $n_s \times R$ que representa las variables de interés observadas en $S$, sea $\boldsymbol{Y}_s = (\boldsymbol{Y}_{s1}, \ldots, \boldsymbol{Y}_{sd}, \ldots, \boldsymbol{Y}_{sD})$.
- $\boldsymbol{Y}_{sd}$ : Matriz de tamaño $n_d \times R$ que representa las variables de interés observadas en $S_d$, sea $\boldsymbol{Y}_{sd} = (\boldsymbol{Y}_{sd1}, \ldots, \boldsymbol{Y}_{sdi}, \ldots, \boldsymbol{Y}_{sdn_d})$.
- $\boldsymbol{Y}_{sdi}$ : Vector fila de tamaño $1 \times R$ que representa los valores observados para la $i-$ésima observación en el $d-$ésimo domino de $U$.
- $\boldsymbol{Y}_{sr}$ : Vector columna de tamaño $n_s \times 1$ que representa la $r-$ésima variable de interés observada en $S$, sea $\boldsymbol{Y}_{sr} = (\boldsymbol{Y}_{s1r}, \ldots, \boldsymbol{Y}_{sdr}, \ldots, \boldsymbol{Y}_{sDr})$.
- $\boldsymbol{Y}_{sdr}$ : Vector columna de tamaño $n_d \times 1$ que representa $r-$ésima variable de interés observada en $S_d$, sea $\boldsymbol{Y}_{sdr} = (y_{d1r}, \ldots, y_{dir}, \ldots, y_{dn_dr})$.
- $y_{dir}$ : Valor obtenido de la $i-$ésima observación en la $r-$ésima variable de interés del $d-$ésimo domino de $U$.

- $\bar{\boldsymbol{Y}}_r$: Representa el vector de medias para la $r-$ésima variable de interés del $d-$ésimo domino de $U$. Sea $\bar{y}_{dr}$ la media de la $r-$ésima variable de interés en el $d-$ésimo domino de $U$, luego, $\bar{\boldsymbol{Y}}_r = (\bar{y}_{1r}, \ldots, \bar{y}_{dr}, \ldots, \bar{y}_{Dr})$.

- $\bar{\boldsymbol{Y}}_d$: Representa el vector de medias para el $d-$ésimo domino de $U$, luego $\bar{\boldsymbol{Y}}_d = (\bar{y}_{d1}, \ldots, \bar{y}_{dr}, \ldots, \bar{y}_{dr})$.

- $\hat{\bar{\boldsymbol{Y}}}_{sr}$: Representa el vector de medias estimado para la $r-$ésima variable de interés del $d-$ésimo domino de $U$. Sea $\hat{\bar{y}}_{sdr}$ la media estimada para $r-$ésima variable de interés en el $d-$ésimo domino de $U$, luego, $\hat{\bar{\boldsymbol{Y}}}_r = \left(\hat{\bar{y}}_{s1r}, \ldots, \hat{\bar{y}}_{sdr}, \ldots, \hat{\bar{y}}_{sDr}\right)$.

- $Y_i$ Representa una variable aleatoria, $y_i$ una realización aleatoria de $Y_i$ y $\boldsymbol{y}^t = (y_1, \ldots, y_i, \ldots y_n)$ un vector de tamaño $1 \times n$ que denota las realizaciones aleatoria $Y_1, \ldots, Y_n$.

- $\boldsymbol{X}$ : Matriz de tamaño $N \times P$ que representa las variables regresoras de un modelo lineal.

- $\boldsymbol{X}_s$ : Matriz de tamaño $n_s \times P$ que representa las variables regresoras de un modelo lineal observadas en $S$, sea $\boldsymbol{X}_s = (\boldsymbol{X}_1, \ldots \boldsymbol{X}_i, \ldots, \boldsymbol{X}_D)$.

- $\boldsymbol{X}_i$ : Vector fila de tamaño $1 \times P$ que representa la $i-$ésima observación de la variable regresora asociada a $\boldsymbol{Y}_i$.

- $\boldsymbol{X}_{sp}$ : Vector columna de tamaño $n_s \times 1$ que representa la $p-$ésima variable regresora asociada a $\boldsymbol{Y}_i$. $S$, sea $\boldsymbol{Y}_{sr} = (\boldsymbol{Y}_{s1r}, \ldots, \boldsymbol{Y}_{sdr}, \ldots, \boldsymbol{Y}_{sDr})$.

- $\boldsymbol{X}_{sd}$ : Matriz de tamaño $n_d \times P$ que representa las variables regresoras de un modelo lineal observadas en $s_d$, sea $\boldsymbol{X}_{sd} = (\boldsymbol{X}_{sd1}, \ldots, \boldsymbol{X}_{sdi}, \ldots, \boldsymbol{X}_{sdn_d})$.

- $\boldsymbol{X}_{sdi}$ : Vector fila de tamaño $1 \times P$ representa los valores observados para la $i-$ésima observación en el $d-$ésimo domino de $U$. Sea $\boldsymbol{X}_{sdi} = (x_{di1}, \ldots, x_{dip}, \ldots, x_{diP})$.

- $\boldsymbol{X}_{sdp}$ : Vector columna de tamaño $n_d \times 1$ que representa los valores observados para la $p-$ésima variable regresora en el $d-$ésimo domino de $U$. Sea $\boldsymbol{X}_{sdp} = (x_{d1p}, \ldots, x_{dip} \ldots, x_{dn_dp})$.

- $\boldsymbol{X}^*$ Un vector aleatorio

- $\boldsymbol{X}_{obs}$ , $\boldsymbol{X}_{miss}$ son los valores observados y no observados de $\boldsymbol{X}_s$.

- $\boldsymbol{X}_s^{(m)}$: $m-$ésima matriz plausible de $\boldsymbol{X}_s$.

- $x_{dip}$ : Valor obtenido de la $i-$ésima observación en la $p-$ésima variable regresora del $d-$ésimo domino de $U$.
- $\boldsymbol{\beta}$ : Matriz de tamaño $P \times D$ de parámetros a estimar en un modelo de regresión lineal. Sea $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d, \ldots, \boldsymbol{\beta}_D)$.
- $\boldsymbol{\beta}_d^t$ : Un vector de tamaño $1 \times P$ que representa los parámetros del $d-$ésimo domino de $U$. Sea $\boldsymbol{\beta}_d^t = (\beta_{d1}, \ldots, \beta_{dp}, \ldots, \beta_{dP})$.
- $\hat{\boldsymbol{\beta}}$ : Matriz de tamaño $P \times D$ estimada para la matriz de parámetro $\boldsymbol{\beta}$.
- $\hat{\boldsymbol{\beta}}_d^t$ : Un vector de tamaño $1 \times P$ que representa la estimación de $\boldsymbol{\beta}_d$ en el $d-$ésimo domino de $U$. Sea $\hat{\boldsymbol{\beta}}_d = \left(\hat{\beta}_{d1}, \ldots, \hat{\beta}_{dp}, \ldots, \hat{\beta}_{dP}\right)$.
- $\hat{\beta}_{dp}$ : El valor estimado para el parámetro $\beta_{dp}$ en la $p-$ésima variable regresora del $d-$ésimo domino de $U$. Tenga presente que, cuando no se indique el sub-índice $d$, es porque no esta divididas en dominios, luego, $\boldsymbol{\beta}_d^t = \boldsymbol{\beta}^t$, $\beta_{dp} = \beta_p$, $\hat{\boldsymbol{\beta}}_d^t = \hat{\boldsymbol{\beta}}^t$, $\hat{\beta}_{dp} = \hat{\beta}_p$.
- $\hat{\beta}_p^{(m)}$ : La $m-$ésima estimación de $\beta_p$ obtenida a partir de $\boldsymbol{X}_s^{(m)}$
- $\hat{\boldsymbol{\beta}}_{(m)}^t$ : Un vector de tamaño $1 \times P$ que representa $m-$ésima estimación de $\boldsymbol{\beta}^t$ que es calculada a partir de $\boldsymbol{X}_s^{(m)}$.
- $\widehat{Var_m}\left(\hat{\boldsymbol{\beta}}\right)$ : $m-$ésima varianza estimada para $\hat{\boldsymbol{\beta}}$ a partir de $\boldsymbol{X}_s^{(m)}$.
- $\hat{\boldsymbol{\beta}}_{IM}$ : La estimación de $\boldsymbol{\beta}$ obtenida empleando la metodología de valores plausible, donde $\hat{\boldsymbol{\beta}}_{IM}^t = \left(\hat{\beta}_1^{(IM)}, \ldots, \hat{\beta}_p^{(IM)}, \ldots, \hat{\beta}_P^{(IM)}\right)$
- $\widehat{Var}\left(\hat{\boldsymbol{\beta}}_{IM}\right)$ varianza estimada para $\hat{\boldsymbol{\beta}}_{IM}$.
- $\sigma_\epsilon$ : La desviación de los errores en un modelo de regresión lineal.
- $\boldsymbol{I}_R$ : Es la matriz identidad de orden $R$.
- $\mathbf{1}$ : Es un vector de unos de tamaño adecuado.
- $\boldsymbol{\mu}_d$ : Vector de tamaño $1 \times R$ que representa las medias para $d-$ésimo dominio de $U$. Sea $\boldsymbol{\mu}_d^t = (\mu_{d1} \ldots, \mu_{dr}, \ldots, \mu_{dR})$.
- $\mu_{dr}$ : La media obtenida para $r-$ésima variable de interés en el $d-$ésimo domino de $U$.
- $\boldsymbol{e}_d$ Vector de tamaño $N_d \times 1$ que representa los errores aleatorios para el $d-$ésimo dominio de $U$.

- $\boldsymbol{\Sigma}_d$ es la matriz de varianzas para los errores $\boldsymbol{e}_d$.
- $\boldsymbol{\Sigma}_e$ es la matriz de varianzas para los errores $\boldsymbol{e}_d$.
- $\boldsymbol{u}$ : Vector de tamaño $D\times 1$ que representa los errores aleatorios, luego, $\boldsymbol{u} = (u_1, \ldots, u_d, \ldots, u_D)$.
- $u_d$ : Error asociado al $d-$ésimo dominio, el cual se asumen independientes del error aleatorio $\boldsymbol{e}_d$ y $\boldsymbol{\delta}_d = \mathbf{1}_R u_d$
- $\boldsymbol{\Sigma}_u = \sigma_u^2 \boldsymbol{I}_D$ es la matriz de varianzas para los errores $u_d$.
- $\boldsymbol{Z}^t$ : Matriz de tamaño $DR \times RK$ que representa la matriz de información auxiliar en el modelo lineal mixto.
- $Z_d^t$ : Matriz de tamaño $R \times RK$ que representa la matriz de información auxiliar asociada a las $R$ variables $\boldsymbol{Y}_d$ en el $d-$ésimo dominio de $U$. Sea $Z_d^t = I_R \otimes \boldsymbol{z}_d$.
- $\boldsymbol{z}_d$ : Vector de tamaño $1\times K$ que identifica la información auxiliar asociada a la $r-$ésima variable $\boldsymbol{Y}_d$ en el $d-$ésimo dominio. Sea $\boldsymbol{z}_d = (z_{d1}, \ldots, z_{dk}, \ldots, z_{dK})$.
- $\boldsymbol{\Gamma}$ : Vector de tamaño $RK \times 1$ que representa el conjunto de todos los coeficientes del modelo lineal mixto multivariado los cuales relacionan las $R$ variables de $\boldsymbol{Y}_d$ con las $K$ variables auxiliares de $Z_d^t$. Sea $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_r, \ldots, \boldsymbol{\gamma}_R)$.
- $\boldsymbol{\gamma}_r$ : Es un vector de coeficientes de tamaño $K \times 1$. Sea $\boldsymbol{\gamma}_r = (\gamma_{1r}, \ldots, \gamma_{kr}, \ldots, \gamma_{Kr})$.
- $\gamma_{kr}$ : Es un coeficiente de un modelo lineal mixto multivariado modela la asociación entre la $r-$ésima variable $\boldsymbol{Y}_{di}$ con la $k-$ésima variable de $\boldsymbol{z}_{di}$.
- Sea $\boldsymbol{\zeta}_d = (\zeta_{d1}, \ldots, \zeta_{di}, \ldots, \zeta_{dn_d})$ luego, $Cor\left(\boldsymbol{\zeta}_d, \boldsymbol{x}_{dp}\right) = \rho_{x\zeta}$ y sea $\rho_z$ es la correlación entre $\boldsymbol{\beta}_p$ y $\boldsymbol{z}_k$.
- $\boldsymbol{V}$: Es la matriz de varianza y covarianzas obtenida en el modelo multivariado de Fay-Herriot.
- $\boldsymbol{V}_d$ : Es la matriz de varianza y covarianzas obtenida en el modelo multivariado de Fay-Herriot, para el $d-$ésimo domino de $U$.

# Introducción

El diseño de muestreo en una encuesta nacional tiene como objetivo asegurar que las inferencias de totales y tasas para los niveles de agregación nacional, regional y departamental puedan ser obtenidas con una precisión preestablecida (Longford 2004). Sin embargo, en niveles de desagregados territoriales como municipios, u otros dominios no planificados, pueda que estas no lleguen a cumplirse con la precisión debida; es decir, que el número de observaciones dentro de cada dominio no sea el adecuado para obtener estimaciones directas con una precisión determinada. Por lo tanto, haciendo uso de información auxiliar de estos dominios, es posible que los estimadores indirectos pueden proporcionar mayor precisión sin aumentar los tamaños muestrales; sin omitir que, en ocasiones la información auxiliar para los dominios podría no estar disponible para todos estos.

Cuando la información auxiliar no esta disponible para todos los dominios, los métodos de estimación de área pequeña que se alinean con la inferencia basada en el modelo, surgen como una herramienta para realizar inferencia sobre dominios o subdominios no planeados en la encuesta. Tal es el caso del modelo clásico propuesto por Fay & Herriot (1979) donde se emplea un modelo lineal mixto con efectos aleatorios por área para estimar el ingreso per-cápita en el censo de 1960 en poblaciones con menos de 1000 habitantes. Posterior a esta publicación han surgido extensiones para esté modelo. Por ejemplo, Prasad & Rao (1990); Datta & Lahiri (2000); Das, Jiang & Rao (2004); González-Manteiga, Lombarda, Molina, Morales & Santamaría (2010); Jiang, Nguyen & Rao (2011); Datta, Kubokawa, Molina & Rao (2011) y Kubokawa (2012) proponen herramientas para medir la incertidumbre de los estimadores de área pequeña basados en modelos. Chandra & Chambers (2009) muestran mediante resultados empíricos que los estimadores basados en modelos directos para áreas pequeñas tienen un buen desempeño a través de un rango de variables que a menudo son de interés en las encuestas de negocios. Herrador, Esteban, Hobza & Morales (2011) trató situaciones donde las áreas pequeñas se dividen en dos grupos y los efectos aleatorios del dominio tienen diferentes varianzas entre los estos. Slud & Maiti (2011) estaban interesados en la estimación de áreas pequeñas basada en datos censurados a izquierda. Para revisiones más exhaustivas se pueden ver los trabajos de Ghosh & Rao (1994); Rao (2005); Jiang &

Lahiri (2006*a*,*b*); Rao (2008) Datta (2009); Lehtonen & Veijanen (2009); Pfeffermann (2013).

Cabe mencionar, que las investigaciones donde es necesario emplear un diseño muestral para acceder a la información mediante el uso de instrumentos de medición (encuestas, exámenes, toma de muestras biológicas, etc). Los instrumentos se construyen pensando en medir más de una medida interés y así obtener medidas descriptivas sobre estás, las cuales pueden estar o no correlacionadas entre sí. Es aquí cuando los modelos multivariantes son adecuados para modelar estas situaciones. Por ejemplo, Fay (1987) y Datta, Fay & Ghosh (1991) compararon los estimadores de área pequeña obtenidos de modelos univariados para cada variable respuesta con los obtenidos por un modelo multivariado, mostrando que la precisión mejora utilizando el modelo multivariado. Datta, Ghosh, Nangia & Natarajan (1996) utilizaron el mismo tipo de modelo para calcular las estimaciones del ingreso medio de las familias con cuatro miembros para los 50 estados de EE.UU. González-Manteiga, Lombardía, Molina, Morales & Santamaría (2008) emplean la versión multivariada del modelo de Fay-Herriot para ayudar a la predicción de parámetros de variables dependientes en áreas pequeñas con datos de muestra agregados a nivel de área; Benavent & Morales (2016) introduce el modelo Fay-Herriot multivariante para estimar indicadores en áreas pequeñas.

Por otro lado, un problema adicional al que se deben enfrentan los investigadores surge cuando la matriz de datos presenta información faltante, en los casos donde, se presentan problemas en la recolección de la información; edición de datos; porque las personas no responden a las preguntas, por falta de información, desconocimiento, vergüenza o temor a contestar, entre otras. En este sentido, Longford (2004) aplica los modelos multivariantes para estimar las tasas de desempleo y la inactividad económica en un área local, utilizando los datos de la encuesta sobre la fuerza de trabajo del Reino Unido. Esté mismo autor propone para el tratamiento de la información faltantes en la matriz de datos implementar el método de imputación múltiple propuesto por Rubin (1987) entre otras alternativas. Ngaruye, von Rosen & Singull (2017) abordan la estimación de áreas pequeñas con datos multivariantes que siguen un patrón escalonado o monótono de datos faltantes, siendo característico de estudios longitudinales, donde se pierde información de la variable de interés en forma progresiva, por lo que formulan un enfoque basado en la probabilidad para estimar los parámetros desconocidos en modelos de curva de crecimiento de efectos aleatorios.

En particular, el estudio de Contexto Escolar y Social del Aprendizaje en Colombia (CESAC) implementado en Colombia desde el 2011, es un programa de investigación desarrollado en

el marco de las pruebas Saber 3°, 5° y 9° que tiene entre sus principales propósitos enriquecer el análisis de los resultados de los estudiantes en las pruebas Saber a partir de un conjunto de indicadores, los cuales miden aspectos contextuales que complementan y enriquecen los resultados de las evaluaciones cognitivas, además de aportar información a los profesionales y responsables de la toma de decisiones de política pública (Icfes 2014).

En este sentido, el Icfes extiende sus procesos de evaluación e inicia un estudio de factores asociados al rendimiento escolar en conjunto con la prueba Saber 3°, 5° y 9°. A partir de esté estudio se busca evaluar el desarrollo de las competencias cognitivas y no cognitivas de los estudiantes (Icfes 2016*a*). El estudio principal de factores asociados fue aplicado en el 2012, periodo en el cual se recogió abundante información sobre las variables de la escuela, los docentes y los estudiantes que pudiesen afectar el aprendizaje. Además, para el periodo comprendido entre el 2012 y el 2015, en conjunto con la prueba Saber 3°, 5° y 9 se aplicó una serie de cuestionarios que recogían información sobre los estudiantes, sus familias y las instituciones educativas a las que asistían (Icfes 2014, 2016*a*).

En la Figura **0-1** se observan las dimensiones por las cuales se indaga en las escuelas. Ahora, asuma que $y_{dir}$ es el rendimiento promedio de los estudiantes en el $r-$ésimo grado del $i-$ésimo establecimiento educativo (EE) para el $d-$ésimo departamento de Colombia. Sea $x_{di2}$ el Índice de Ausencia Injustificada y Deserción (IAID) obtenido para el $i-$ésimo establecimiento educativo del $d-$ésimo departamento de Colombia. Sea $N_d$ el número total de EE en el $d-$ésimo departamento, sin perdida de generalidad, asuma $\boldsymbol{X}_{d2}^t = (x_{d12}, \ldots, x_{di2}, \ldots, x_{dN_d2})$ es un vector de tamaño $1 \times N_d$ que esta relacionado con $\boldsymbol{Y}_{dr}^t = (y_{d1r}, \ldots, y_{dir}, \ldots, y_{dN_dr})$ a través de la función $f$, es decir, $\boldsymbol{Y}_{dr} = f(\boldsymbol{X}_{d2})$, luego considere que la relación entre $\boldsymbol{X}_{d2}$ e $\boldsymbol{Y}_{dr}$ es lineal, para todos los departamentos, esto es:

$$\begin{array}{rcl} \boldsymbol{Y}_{1r} & = & \beta_{10} + \beta_{11}\boldsymbol{X}_{11} = \boldsymbol{X}_1^t\boldsymbol{\beta}_1 \\ \boldsymbol{Y}_{2r} & = & \beta_{20} + \beta_{21}\boldsymbol{X}_{21} = \boldsymbol{X}_2^t\boldsymbol{\beta}_2 \\ \vdots \quad \vdots & & \vdots \\ \boldsymbol{Y}_{dr} & = & \beta_{d0} + \beta_{d1}\boldsymbol{X}_{d1} = \boldsymbol{X}_d^t\boldsymbol{\beta}_d \\ \vdots \quad \vdots & & \vdots \\ \boldsymbol{Y}_{Dr} & = & \beta_{D0} + \beta_{D1}\boldsymbol{X}_{D1} = \boldsymbol{X}_D^t\boldsymbol{\beta}_D \end{array} \tag{0-1}$$

donde $\boldsymbol{X}_d^t = (\mathbf{1}_d, \boldsymbol{X}_{d1})$ es la matriz diseño de las variables regresoras para el $d-$ésimo departamento, $\mathbf{1}_d$ es el vector unitario y $\boldsymbol{\beta}_d = (\beta_{d0}, \beta_{d1})$ es el vector de coeficientes de

regresión para el modelo lineal para el $d-$ésimo departamento.

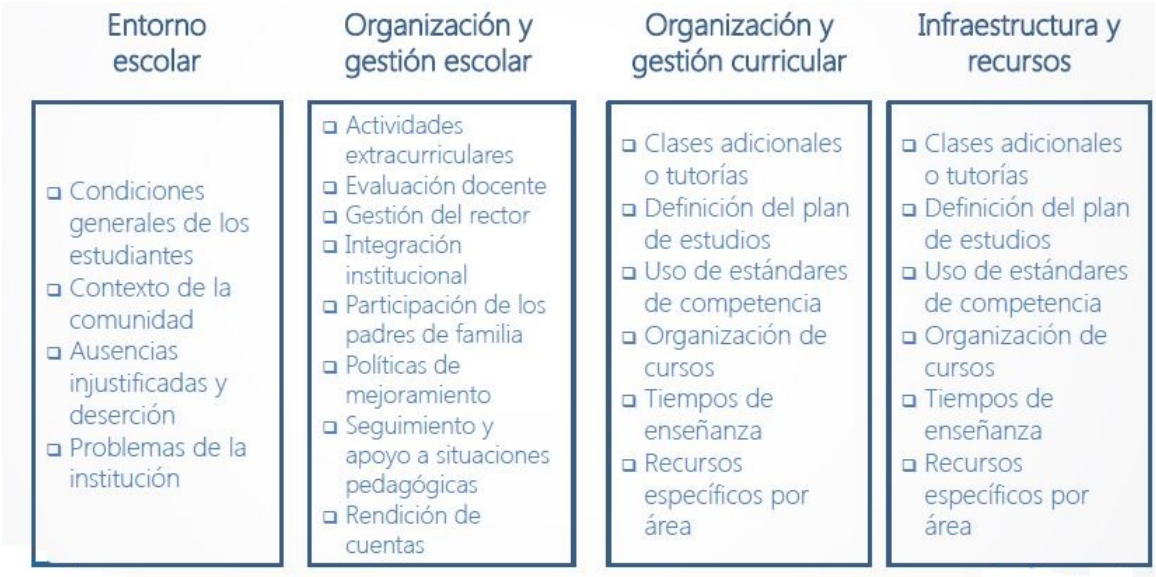


**Figura 0-1**.: Dimensiones en el establecimiento educativo, tomado de Icfes (2014)

Pero dadas las limitaciones económicas, los problemas logísticos y los errores no controlados que pueden llegar a presentarse cuando se realiza este tipo de investigaciones el Icfes, opta por ejecurtar el estudio de factores asociados enmarcado dentro del diseño muestral utilizado para las pruebas Saber 3°, 5° y 9°, en el cual empleando un diseño de muestreo complejo para seleccionar una "muestra representativa" a nivel nacional de EE, la cual se denomina como *muestra controlada*. (Icfes 2016*b*). Sin embargo, esta muestra no garantiza los resultados de la prueba Saber 3°, 5° y 9° para todos los departamentos y/o secretarias de educación. Por tal motivo, el Icfes, delega en las secretarías de Educación y las instituciones educativas la responsabilidad de aplicar la prueba en los colegios que No fueron seleccionados en la muestra controlada, a ésta aplicación se le denomina como *aplicación censal*.

Es preciso señalar que los cuestionarios de factores asociados son asignados de forma que es posible incluir preguntas diferentes entre grupos de estudiantes. Por ejemplo, algunos de los cuestionarios aplicados presentan preguntas sobre la disponibilidad y el uso de tecnologías de información, mientras que otros contienen preguntas asociadas a las características de la biblioteca escolar y el ambiente de disciplina en el salón de clases (Icfes 2017). Esta asignación de cuadernillos hace que la matriz $\mathscr{P}$ de datos que resulta presenta información faltante.

Considerando el diseño muestral implementado por Icfes y las observaciones sobre esté surgen las interogantes: ¿Es posible obtener estimadores aproximadamente insesgados para la es-

timación de los parámetros ($\boldsymbol{\beta}_d$) de un modelo de regresión lineal múltiple y sus varianzas en dominios no observados en una muestra probabilística con presencia de datos faltantes?, ¿Cuál sería el comportamiento de los estimadores y sus varianzas bajo diferentes métodos de imputación?, ¿Cómo se afecta el modelo multivariado Fay-Herriot cuando se emplean valores plausibles en la variable respuesta ($\boldsymbol{\beta}_d$)?, ¿Los estimadores propuestos se pueden aplicar al problema de factores asociados dada la información disponible en el Icfes?; entonces, dar solución a éstos interrogantes fue la principal motivación para la realización del presente trabajo.

El contenido del documento se organizó por capítulos, el Capítulo 1 muestra un marco de referencia para las técnicas de estimación en muestras probabilísticas empleando la metodología de estimación en áreas pequeñas e incluye un apartado para los métodos de imputación y valores plausibles, así como la implementación en el paquete R. El Capítulo 2 introduce el método estimación de áreas pequeñas combinando con los valores plausibles, además presenta la metodología propuesta para estimar los coeficientes del modelo de regresión lineal. El Capitulo 3 describe un estudio de simulación desarrollado para evaluar la metodología propuesta. En el Capitulo 4 se muestra la aplicación de la metodología desarrollada sobre los datos recolectados por el Icfes para el estudio de factores asociados y en Capitulo 5 se realizan las conclusiones sobre los resultados obtenidos.

# 1. Marco teórico

En este capítulo se recopilan las definiciones de muestreo, modelos de regresión múltiple en muestras probabilísticas y técnicas de imputación esenciales para el desarrollo de este trabajo.

## 1.1. Definiciones de introducción al muestreo

Las definiciones recopiladas en el presente apartado y empleadas para introducir los métodos de inferencia en el campo de las muestras probabilísticas fueron extraídas de Särdnal, Swensson & Wretman (1992) y Gutiérrez (2009).

**Definición 1** (*Población Finita*)**.** Una población finita es un conjunto de $N$ elementos donde cada elemento puede ser identificado por un conjunto de rótulos. Sea $U = \{1, 2, \ldots, i, \ldots, N\}$ el conjunto de rótulos de la población finita.

**Definición 2** (*Muestra aleatoria*)**.** Es un subconjunto de la población que ha sido extraído mediante un mecanismo aleatorio de selección. Notaremos con una letra mayúscula $S$ a la muestra aleatoria y $s = \{1, 2, \ldots, n(s)\}$ será la realización aleatoria de $S$.

**Definición 3** (*Muestras probabilísticas*)**.** Una muestra es de tipo probabilístico sí:

1. Es posible construir un soporte $Q$, tal que $Q = \{s_1.s_2, \ldots, s_q, \ldots, s_{Q_n}\}$ de todas las muestras posibles obtenidas por un método de selección. En donde $s_q$, $q = 1, \ldots Q_n$, es una muestra perteneciente al soporte $Q$.

2. Las probabilidades de selección que el proceso aleatorio le otorga a cada posible muestra perteneciente al soporte son conocidas de antemano a la selección de la muestra final.

**Definición 4** (*Diseño de muestreo*)**.** Un diseño de muestreo $p(\cdot)$ sobre un soporte $Q$ es una distribución de probabilidad multivariante sobre $Q$; es decir, $p(\cdot)$ es una función que va desde

el soporte simétrico[1] $Q$ hasta $[0,1]$ tal que $p(\cdot) > 0$ para todo $s \in Q$

$$\sum_{s \in Q} p(s) \quad = \quad 1$$

Dado el soporte $Q$, un diseño de muestreo es una función $p(\cdot)$, tal que $p(s)$ arroja la probabilidad de selección de la muestra realizada $s$ bajo un esquema de selección particular.

**Definición 5** (*Probabilidad de inclusión*)**.** Dado un soporte $Q$, la probabilidad de inclusión $\pi_i$ es la probabilidad de que el elemento $i-$ésimo pertenezca a la muestra aleatoria $S$ y se puede escribir de la siguiente manera:

$$\pi_i \quad = \quad E(I_i(S)) = \sum_{s \in Q} I_i(s)\, p(s) = \sum_{s \ni i} p(s)$$

con $I_i(s)$ una función indicadora tal que

$$I_i(s) \quad = \quad \begin{cases} 1 & \text{si } i \in s \\ 0 & \text{si } i \notin s \end{cases}$$

de forma análoga se puede escribir la probabilidad de inclusión de segundo orden la cual se denota como $\pi_{il}$ e indica la probabilidad de que los elementos $i$ y $l$ pertenezcan a la muestra,

$$\pi_{il} \quad = \quad Pr(i \in s \,\&\, l \in s) = Pr(I_i(s)\, I_l(s) = 1) = \sum_{s \ni i,l} p(s)$$

**Definición 6** (*Pesos no calibrados*)**.** Corresponden a los pesos teóricos del diseño muestral corregidos por la falta de respuesta (factores sin calibrar). En el caso en que la no respuesta se modele mediante variables aleatorias Bernoulli independientes, los pesos no calibrados serían las inversas de las probabilidades finales de inclusión de individuos; es decir $w_i = \frac{1}{\pi_i}$.

**Definición 7** (*Dominio*)**.** Considere una partición de la población $U$ en $D$ sub-poblaciones o sub-grupos poblacionales. Sea $N_d$ el tamaño de $U_d$. Se tienen las ecuaciones de partición

$$U = \bigcup_{d=1}^{D} U_d \quad \text{y} \quad N = \sum_{d=1}^{D} N_d$$

donde $N_d$ es llamado tamaño absoluto del dominio.

[1] Un soporte se llama simétrico si para cualquier $s \in Q$, todas las permutaciones de s están también en $Q$

No está de más mencionar que para la estimación por dominios generalmente se desconoce la pertenencia de las unidades poblacionales al dominio.

## 1.2. Métodos de estimación para muestras probabilísticas

Hartley & Sielken Jr (1975) afirman que la inferencia en muestreo se puede dividir en cuatro tipos de inferencia. Para describir cada uno de los caso es necesario suponer un modelo de superpoblación, en el cual se considera a la población finita $U$ como una muestra de $N$ observaciones de una población infinita y la muestra $s$ es obtenida a partir de $U$ mediante diseño de muestreo $p(\cdot)$, de ésta forma se define un muestro en dos pasos, así: *Paso 1.* Extraer un "muestra grande" de tamaño $N$ de una superpoblación infinita. *Paso 2.* Extraer una muestra de tamaño $n < N$ de la muestra grande de tamaño $N$ obtenida en el *Paso 1.* Cabe mencionar que, el Paso 1 es un paso imaginario, por lo cual, generalmente se asume que los elementos de muestra resultantes están distribuidos de manera independiente e idéntica, luego los cuatro tipos de inferencias que resultan serian los que se describen a continuación:

- En el *Caso 1*, se asume una población finita $U$ con valores fijos y las inferencias se hacen sobre ciertas funciones de parámetros que son computables a partir de una muestra $s$, extraídas mediante un diseño de muestreo $p(\cdot)$ de la población finita $U$. En este tipo de inferencia encontramos el estimador de Horvitz & Thompson (1952) y sus variantes, que se agrupan dentro de los *métodos de estimación basados en el diseño.* Estos son considerados como el conjunto de estimadores más básicos dentro de los métodos de estimación para poblaciones finitas, dado que, su éxito depende de lo representativa que sea la muestra de la población objeto de estudio.

  Cuando además de contar con un diseño de muestreo óptimo, se dispone de información auxiliar para la población, entonces, se tienen las condiciones para utilizar los estimadores de Calibración (Deville & Särndal 1992) o el estimador de regresión generalizado (GREG) (Särndal 1982, Robinson & Särndal 1983) para obtener estimadores consistentes desde el punto de vista del diseño. El conjunto de estimadores que se apoyan en los modelos de regresión pueden ser agrupados dentro de los *métodos de estimación asistidos por modelo.* Sin embargo, estos métodos comparten las mismas restricciones que los estimadores basados en el diseño dado que se asume que la población objeto de estudio es fija.

- En el *Caso 2,* las inferencias se realizan sobre las "estadísticas" de la población finita,

haciendo de lado los parámetros del modelo de superpoblación dado que son irrelevantes. En esta inferencia se agrupan los *estimadores basados en modelos,* los cuales buscan relacionar las variables a través de un modelo de distribución de probabilidad conjunta, por tanto, las estadísticas de la población finita son predichas basados en la existencia del modelo de superpoblación. Estos estimadores generalmente poseen una varianza pequeña, pero suelen ser sesgados si los supuestos del modelo no se cumplen. Cabe mencionar que los estimadores basados en modelos toman mucha relevancia cuando nos enfrentamos al problema de hacer estimaciones precisas en dominios donde el tamaño de la muestra es pequeño o inexistente, por lo cual se toman como la base para los estimadores de áreas pequeñas.

En este sentido, debemos mencionar que los estimadores basados en modelos para estimación de áreas pequeñas se clasifican generalmente en dos clases dependiendo de la disponibilidad de datos sobre la variable respuesta de interés: Los *Modelos a nivel de unidad,* estos modelos fueron propuestos por Battese, Harter & Fuller (1988) quienes emplearon un modelo lineal mixto para estimar el área con cultivo de maíz y soja en 12 condados de Iowa Norte-Central, el segundo grupo es formado por los *Modelos de área,* los cuales se emplean cuando la información de las covariables se encuentra sólo a nivel de área y fueron propuestos por Fay & Herriot (1979). Este último es el de mayor interés para el desarrollo del presente trabajo, por lo tanto, se estudiarán con más detalle en las secciones siguientes.

- En el *Caso 3,* los parámetros del modelo superpoblación es el objetivo de interés, por ejemplo, considere que está interesado en modelar una población conceptual de personas que viven en Colombia, para ellos podemos suponer que la población $U$ es una realización aleatoria del modelo de superpoblación (población conceptual). Ahora, suponga que la muestra $s$ es obtenida de $U$ mediante un diseño de muestreo $p(\cdot)$, luego, la inferencia sobre el modelo superpoblación se hacen teniendo presente las fuentes de aleatoriedad provenientes del modelo de superpoblación y la aleatoriedad inducida $p(\cdot)$. Este caso es conocido como inferencia doble.

- El *Caso 4,* es cuando se desea hacer inferencia sobre los parámetros de un modelo superpoblación, pero se asume que la población finita es fija y no proviene del modelo de superpoblación, siendo un caso inviable para hacer inferencia.

## 1.3. Modelo de regresión múltiple en muestras probabilísticas

En la presente sección se describe de forma muy general los modelos de regresión múltiple así, como los procesos para la estimación de los parámetros de los mismos en muestras probabilísticas. La principal referencia empleada en esta sección es Pessoa & Silva (1998, Cap. 5).

### 1.3.1. Método de máxima pseudo-verosimilitud (MPV)

Sea $y_i$ una realización aleatoria de la $i-$ésima variable aleatoria $Y_i$, para todo $i \in U$, además asuma que $Y_1, \ldots, Y_i, \ldots, Y_N$ son variables aleatorios independientes e idénticamente distribuidos con función de densidad $h(y_i, \boldsymbol{\varphi})$, donde $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_R)$ es el vector de parámetros. Sea $\boldsymbol{A}_i(\boldsymbol{\varphi}) = \frac{\partial}{\partial \boldsymbol{\varphi}} \log[h(y_i, \boldsymbol{\varphi})]$ el vector de scores calculado para el $i-$ésimo elemento. Cuando se conocen todos los elementos de la población finita $U$, la función de verosimilitud y log-verosmilitud está dada por:

$$\ell_U(\boldsymbol{\varphi}) = \prod_{i \in U} h(y_i, \boldsymbol{\varphi}) \tag{1-1}$$

$$L_U(\boldsymbol{\varphi}) = \sum_{i \in U} \log[h(y_i, \boldsymbol{\varphi})] \tag{1-2}$$

y la ecuación de verosimilitud de la población es dada por

$$\sum_{i \in U} \boldsymbol{A}_i(\boldsymbol{\varphi}) = 0$$

En condiciones de regularidad (Véase, Cox & Hinkley 1974, pp. 281), la solución $\boldsymbol{\varphi}_U$ de este sistema es el Estimador Máximo Verosímil para $\boldsymbol{\varphi}$.
Sea $S$ una muestra probabilística seleccionada de $U$ mediante el diseño de muestreo $p(\cdot)$ con probabilidades de inclusión $\pi_i$. Sea $\boldsymbol{T} = \sum_{i \in U} \boldsymbol{A}_i(\boldsymbol{\varphi})$ el vector de totales para los scores en la población. Para estimar $\boldsymbol{T}$, se puede utilizar un estimador lineal de la forma $\hat{\boldsymbol{T}} = \sum_{i \in S} w_i \boldsymbol{A}_i(\boldsymbol{\varphi})$ con $w_i = \pi_i^{-1}$ para todo $i \in S$. El estimador de máxima pseudo-verosimilitud $\hat{\boldsymbol{\varphi}}_{MPV}$ para $\boldsymbol{\varphi}_U$ será la solución a las ecuaciones de pseudo-verosimilitud dadas por

$$\hat{\boldsymbol{T}} = \sum_{i \in S} w_i \boldsymbol{A}_i(\boldsymbol{\varphi}) = \boldsymbol{0}.$$

Las propiedades del estimador $\hat{\boldsymbol{\varphi}}_{MPV}$ se puede consultar en Pessoa & Silva (1998).

### 1.3.2. Estimación de los parámetros para un modelo de regresión lineal

El problema considerado en esta sección es la estimación de los coeficientes del modelo de regresión lineal normal para una muestra probabílistica.

Sea $S$ una muestra probabilística de tamaño $n_s$ seleccionada de $U$ mediante el diseño de muestreo $p(\cdot)$ con probabilidades de inclusión $\pi_i$. Sea $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{iP})$ un vector fila de tamaño $1 \times P$ y sea $\boldsymbol{X}_s = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i, \ldots, \boldsymbol{X}_{n_s})$ la matriz de tamaño $n_s \times P$ que representa las variables regresoras de un modelo lineal observadas en $S$. Sea $f$ la función lineal que relaciona la variable aleatoria $Y_i$ con el vector $\boldsymbol{X}_i$. La función $f$ tiene como objetivo explicar o predecir los valores de una variable aleatoria $Y_i$, que se considera como la variable de respuesta. Sea $y_i$ la realización aleatoria de la variable aleatoria $Y_i$ para todo $i \in U$. Ahora, asuma que para todo $i \in U$, la pareja ordena $(y_i, \boldsymbol{X}_i)$ son realizaciones aleatorias independientes e idénticamente distribuidos tales que

$$f(y_i \mid \boldsymbol{X}_i; \boldsymbol{\beta}, \sigma_\epsilon) = (2\pi\sigma_\epsilon)^{-\frac{1}{2}} \exp\left[-\frac{(y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2}{2\sigma_\epsilon}\right] \tag{1-3}$$

donde $\boldsymbol{\beta}^t = (\beta_1, \ldots, \beta_p, \ldots, \beta_P)$ de tamaño $1 \times P$ y $\sigma_\epsilon > 0$ denotan los parámetros desconocidos del modelo. Este modelo ha sido considerado por varios autores, por ejemplo Holt, Smith & Winter (1980), Nathan & Holt (1980), Skinner (1989), Chambers (1986, 1995).

La función de score para $\boldsymbol{\beta}$ y $\sigma_\epsilon$ definida a partir del modelo (1-3) está dada por

$$\begin{aligned} \frac{\partial f(y_i \mid \boldsymbol{X}_i; \boldsymbol{\beta}, \sigma_\epsilon)}{\partial \boldsymbol{\beta}} &\approx \boldsymbol{X}_i^t (y_i - \boldsymbol{X}_i\boldsymbol{\beta}) = \boldsymbol{A}_i(\boldsymbol{\beta}) \\ \frac{\partial f(y_i \mid \boldsymbol{X}_i; \boldsymbol{\beta}, \sigma_\epsilon)}{\partial \sigma_\epsilon} &\approx (y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2 - \sigma_\epsilon = A_i(\sigma_\epsilon) \end{aligned}$$

Si los pesos $w_i = \pi_i^{-1}$ satisfacen las condiciones de regularidad dadas en Cox & Hinkley (1974, pp. 281), se obtendrán inmediatamente las ecuaciones de pseudo-verosimilitud

$$\sum_{i \in S} w_i \boldsymbol{A}_i\left(\hat{\boldsymbol{\beta}}_{MPV}\right) = \boldsymbol{X}_s^t \boldsymbol{W} \boldsymbol{Y}_s - \left(\boldsymbol{X}_s^t \boldsymbol{W} \boldsymbol{X}_s\right)\hat{\boldsymbol{\beta}}_{MPV} = \boldsymbol{0} \tag{1-4}$$

$$\sum_{i \in S} w_i A_i\left(\hat{\sigma}_\epsilon^{MPV}\right) = \left(\boldsymbol{Y}_s - \boldsymbol{X}_s\hat{\boldsymbol{\beta}}_{MPV}\right)^t \boldsymbol{W}\left(\boldsymbol{Y}_s - \boldsymbol{X}_s\hat{\boldsymbol{\beta}}_{MPV}\right) - \boldsymbol{1}_{n_s}^t \boldsymbol{W} \boldsymbol{1}_{n_s} \hat{\sigma}_\epsilon^{MPV} = 0 \tag{1-5}$$

donde $\boldsymbol{Y}_s^t = (y_1, \ldots, y_{n_s})$ es de tamaño $1 \times n_s$ y representa las observaciones en la muestra $S$, $\boldsymbol{W} = \text{diag}\,(w_1, \ldots, w_i, \ldots, w_{n_s})$ es una matriz diagonal de $n_s \times n_s$ con los factores de expansión en su diagonal, $\mathbf{1}_{n_s}$ el vector de unos de tamaño $n_s \times 1$ y $\hat{\boldsymbol{\beta}}_{MPV}$ y $\hat{\sigma}_e^{MPV}$ son los estimadores de máxima pseudo-verosimilitud. Suponiendo $\boldsymbol{X}_s^t \boldsymbol{W} \boldsymbol{X}_s$ es una matriz no singular, pueden resolverse las ecuaciones (1-4) y (1-5) para $\hat{\boldsymbol{\beta}}_{MPV}$ y $\hat{\sigma}_\epsilon^{MPV}$ así:

$$\hat{\boldsymbol{\beta}}_{MPV} = \left(\boldsymbol{X}_s^t \boldsymbol{W} \boldsymbol{X}_s\right)^{-1} \boldsymbol{X}_s^t \boldsymbol{W} \boldsymbol{Y}_s \tag{1-6}$$

$$\hat{\sigma}_\epsilon^{MPV} = \left(\mathbf{1}_{n_s}^t \boldsymbol{W} \mathbf{1}_{n_s}\right)^{-1} \boldsymbol{Y}_s^t \left[\boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}_s \left(\boldsymbol{X}_s^t \boldsymbol{W} \boldsymbol{X}_s\right)^{-1} \boldsymbol{X}_s^t \boldsymbol{W}\right] \boldsymbol{Y}_s \tag{1-7}$$

y su varianza estimada esta dada por:

$$\widehat{Var}\left(\hat{\boldsymbol{\beta}}_{MPV}\right) = \left(\boldsymbol{X}_s^t \boldsymbol{\Pi}^{-1} \boldsymbol{X}_s\right)^{-1} \widehat{Var}\left(\sum_{i \in s} w_i \boldsymbol{X}_i \epsilon_i\right) \left(\boldsymbol{X}_s^t \boldsymbol{\Pi}^{-1} \boldsymbol{X}_s\right)^{-1} \tag{1-8}$$

donde

$$\widehat{Var}\left(\sum_{i \in s} w_i \boldsymbol{X}_i \epsilon_i\right) = \sum_{i \in s} \sum_{l \in s} \left(\pi_i^{-1} \pi_l^{-1} - \pi_{il}^{-1}\right) \hat{\epsilon}_i \boldsymbol{X}_i \boldsymbol{X}_l^t \hat{\epsilon}_l$$

con $\boldsymbol{\Pi}^{-1} = \text{diag}\,(\pi_i : i \in s)$ y $\hat{\epsilon}_i = y_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_{MPV}$ para $i \in S$.

## 1.4. Estimación en áreas pequeñas

Los estimadores de áreas pequeñas (SAE, Small Area Estimation) son técnicas estadísticas motivadas por la necesidad de información precisa para pequeños dominios donde el tamaño de la muestra es pequeño o inexistente para alcanzar una precisión preestablecida.
En el caso de la inferencia basada en el modelo para estimación de áreas pequeñas, Jiang & Lahiri (2006*b*) realizan una amplia discusión sobre los modelos lineales mixtos y modelos lineales generalizados relevantes en los problemas ligados a la estimación de áreas pequeñas y los problemas derivados de estos. Esta discusión ha llevado al uso del Mejor Predictor Empírico Lineal Insesgado (EBLUP, Empirical Best Linear Unbiased Predictors) derivados de los modelos mixtos y que son de crucial importancia en la estimación de área pequeña (Rao & Molina 2015).

### 1.4.1. Modelo multivariado de Fay-Herriot

El desarrollo de este apartado está enfocado en la estimación de parámetros de interés en las áreas pequeñas. En esta sección se emplea la notación y definiciones utilizadas por González-Manteiga, Lombardía, Molina, Morales & Santamaría (2008) y Rao & Molina (2015).
Dada la población $U$ de tamaño $N$ la cual se divide en $D$ dominios de tamaño $N_d$, $d = 1, \ldots, D$. Sea $\boldsymbol{Y}^t = \left(\boldsymbol{Y}_1^t, \ldots, \boldsymbol{Y}_d^t, \ldots, \boldsymbol{Y}_D^t\right)$ una matriz aleatoria de tamaño $N \times R$ que representa las variables de interés en $U$, donde $\boldsymbol{Y}_d^t = (\boldsymbol{Y}_{d1}, \ldots, \boldsymbol{Y}_{di}, \ldots, \boldsymbol{Y}_{N_d})$ es una matriz aleatoria de tamaño $N_d \times R$ que representa las variables de interés en el $d-$ésimo dominio de $U$ y $\boldsymbol{Y}_{di} = (y_{di1}, \ldots, y_{dir}, \ldots, y_{diR})$ es un vector de tamaño $1 \times R$ que representa los valores del $i-$ésimo individuo en el $d-$ésimo dominio de la población, donde $y_{dir}$ es el valor observado para la $r-$ésima variable del $i-$ésimo individuo en el $d-$ésimo dominio de la población con $i = 1, \ldots, N_d$, $d = 1, \ldots, D$ y $r = 1, \ldots, R$. Sea $\boldsymbol{Y}_{dr}^t = (y_{d1r}, \ldots, y_{dir}, \ldots, y_{dN_dr})$ el vector columna de tamaño $1 \times N_d$ que representa las observaciones en la $r-$ésima variable de interés en el $d-$ésimo dominio de $U$. Sea $\mu_{dr} = N_d^{-1} \sum_{i=1}^{N_d} y_{dir}$ la media $\boldsymbol{Y}_{dr}^t$ para $r = 1, \ldots, R$ y denote por $\boldsymbol{\mu}_d = (\mu_{d1} \ldots, \mu_{dr}, \ldots, \mu_{dR})$ el vector fila de tamaño $1 \times R$ que representa las medias para $d-$ésima área cuyo estimador directo es $\bar{\boldsymbol{Y}}_d^t = \left(\bar{Y}_{d1}, \ldots, \bar{Y}_{dr}, \ldots, \bar{Y}_{dR}\right)$, $d = 1, \ldots, D$. El modelo multivariado de Fay-Herriot supone que:

1. El estimador directo $\bar{\boldsymbol{Y}}_d$ es independiente del verdadero valor del parámetro $\boldsymbol{\mu}_d$, esto es:

$$\bar{\boldsymbol{Y}}_d \mid \boldsymbol{\mu}_d \;\; \sim \;\; N\left(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d\right), \; d = 1, \ldots, D$$

   donde $\boldsymbol{\Sigma}_d$ es una matriz de $R \times R$ conocida, de donde se sigue que $\bar{\boldsymbol{Y}}_d$ puede ser expresado como

$$\bar{\boldsymbol{Y}}_d = \boldsymbol{\mu}_d + \boldsymbol{e}_d, \quad \boldsymbol{e}_d \sim N\left(\boldsymbol{0}, \boldsymbol{\Sigma}_d\right) \quad d = 1, \ldots, D$$

   donde los errores aleatorios $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_D$ son independientes e idénticamente distribuidos, además se asume que los promedios $\mu_{dr}$ están linealmente relacionados con $P$ variables explicativas.

Sea la matriz de información auxiliar $\boldsymbol{Z}^t = \left(\boldsymbol{Z}_1^t, \ldots, \boldsymbol{Z}_d^t, \ldots, \boldsymbol{Z}_D^t\right)$ de tamaño $RK \times DR$ la

cual toma la estructura

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_{11} & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{z}_{1r} & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{z}_{1R} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ \boldsymbol{z}_{D1} & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{z}_{Dr} & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{z}_{DR} \end{bmatrix}_{DR \times RK} = \begin{bmatrix} \boldsymbol{I}_R \otimes \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{I}_R \otimes \boldsymbol{z}_d \\ \vdots \\ \boldsymbol{I}_R \otimes \boldsymbol{z}_D \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_1^t \\ \vdots \\ \boldsymbol{Z}_d^t \\ \vdots \\ \boldsymbol{Z}_D^t \end{bmatrix}$$

donde $\boldsymbol{z}_d = (z_{d1}, \ldots, z_{dk}, \ldots, z_{dK})$ es un vector fila de tamaño $1 \times K$ el cual representa la información auxiliar asociada a $\bar{\boldsymbol{Y}}_d^t$ en el $d-$ésimo dominio, $\boldsymbol{I}_R$ es la matriz identidad de orden $R$ y $\boldsymbol{Z}_d^t = \boldsymbol{I}_R \otimes \boldsymbol{z}_d$ donde $\otimes$ representa el kronecker, note que $\boldsymbol{Z}_d^t$ es una matriz que en su diagonal contiene los valores de $\boldsymbol{z}_d$. Note que, $\boldsymbol{Z}_d^t$ es una matriz de tamaño $R \times RK$, además, cuando $R = 1$ se tiene que $\boldsymbol{Z}_d^t = \boldsymbol{z}_d$.
Sea $\boldsymbol{\gamma}_r^t = (\gamma_{1r}, \ldots, \gamma_{kr}, \ldots, \gamma_{Kr})$ un vector de tamaño $1 \times K$ para $r = 1, \ldots, R$, donde $\gamma_{kr}$ es el coeficiente de un modelo lineal mixto multivariado que asocia $r-$ésimo termino de $\bar{\boldsymbol{Y}}_d^t$ con la $k-$ésima variable de $\boldsymbol{z}_d$. Sea $\boldsymbol{\Gamma}^t = (\boldsymbol{\gamma}_1^t, \ldots, \boldsymbol{\gamma}_r^t, \ldots, \boldsymbol{\gamma}_R^t)$ es un vector de tamaño $1 \times RK$ que representa el conjunto de todos los coeficientes del modelo lineal mixto multivariado que relaciona las $R$ valores de $\bar{\boldsymbol{Y}}_d^t$ con las $K$ variables auxiliares en $\boldsymbol{Z}_d^t$. Sea $\boldsymbol{1}_R$ el vector de unos de tamaño $R$.

2. El vector de $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_d, \ldots, \boldsymbol{\mu}_D)$ puede expresarse como

$$\begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_d \\ \vdots \\ \boldsymbol{\mu}_D \end{bmatrix}_{DR \times 1} = \begin{bmatrix} \boldsymbol{Z}_1^t \\ \vdots \\ \boldsymbol{Z}_d^t \\ \vdots \\ \boldsymbol{Z}_D^t \end{bmatrix}_{DR \times RK} \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_d \\ \vdots \\ \boldsymbol{\gamma}_D \end{bmatrix}_{RK \times 1} + \begin{bmatrix} \boldsymbol{1}_R u_1 \\ \vdots \\ \boldsymbol{1}_R u_d \\ \vdots \\ \boldsymbol{1}_R u_D \end{bmatrix}_{DR \times 1}$$

en particular, suponga que $K = 2$, $R = 3$ y $D = 3$, así:

$$\underbrace{\begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \\ \mu_{31} \\ \mu_{32} \\ \mu_{33} \end{bmatrix}}_{\boldsymbol{\mu}} = \underbrace{\begin{bmatrix} z_{11} & z_{12} & 0 & 0 & 0 & 0 \\ 0 & 0 & z_{11} & z_{12} & 0 & 0 \\ 0 & 0 & 0 & 0 & z_{11} & z_{12} \\ z_{21} & z_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & z_{21} & z_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 & z_{21} & z_{22} \\ z_{31} & z_{32} & 0 & 0 & 0 & 0 \\ 0 & 0 & z_{31} & z_{32} & 0 & 0 \\ 0 & 0 & 0 & 0 & z_{31} & z_{32} \end{bmatrix}}_{\boldsymbol{Z}} \underbrace{\begin{bmatrix} \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{31} \\ \gamma_{32} \end{bmatrix}}_{\boldsymbol{\Gamma}} + \underbrace{\begin{bmatrix} u_1 \\ u_1 \\ u_1 \\ u_2 \\ u_2 \\ u_2 \\ u_3 \\ u_3 \\ u_3 \end{bmatrix}}_{\boldsymbol{u}}$$

y para el $d-$ésimo dominio se sigue que

$$\boldsymbol{\mu}_d = \boldsymbol{Z}_d^t \boldsymbol{\Gamma} + \mathbf{1}_R u_d, \quad u_d \sim N\left(0, \sigma_u^2\right) \quad d = 1, \ldots, D$$

donde $\sigma_u^2 \in (0, \infty)$ es conocida y $u_d$ es el error asociado al $d-$ésimo dominio y representa las variaciones entre dominios no explicadas por $\boldsymbol{Z}_{d\cdot}^t$ , el cual se asumen independientes del error aleatorio $\boldsymbol{e}_d$ .

Sea $\boldsymbol{u}^t = (u_1, \ldots, u_d, \ldots, u_D)$ el vector de tamaño $D \times 1$ que denota el error asociado al $d-$ésimo domino el cual se asume es independientes entre dominios e independiente de los errores aleatorios $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_D$. Sea $\boldsymbol{\delta}_d^t = \mathbf{1}_R u_d$ y $\boldsymbol{\delta}^t = \left(\boldsymbol{\delta}_1^t, \ldots, \boldsymbol{\delta}_d^t, \ldots, \boldsymbol{\delta}_D^t\right)$ los efecto aleatorio no explicados por $\boldsymbol{Z}^t$ y el cual se asocia con el $d-$ésimo dominio de $U$. Las dimensiones de $\boldsymbol{\delta}_d^t$ y $\boldsymbol{\delta}^t$ son de $1 \times R$ y $1 \times DR$ respectivamente.
Luego, basado en las condiciones 1 y 2 se puede escribir el modelo mixto

$$\bar{\boldsymbol{Y}} \;=\; \boldsymbol{Z}^t \boldsymbol{\Gamma} + \boldsymbol{\delta} + \boldsymbol{e}, \tag{1-9}$$

con

$$\begin{aligned} \boldsymbol{u} &\sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_u\right), \qquad \boldsymbol{\Sigma}_u = \sigma_u^2 \boldsymbol{I}_D \\ \boldsymbol{e} &\sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_e\right), \qquad \boldsymbol{\Sigma}_e = \operatorname{diag}\left\{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_d, \ldots, \boldsymbol{\Sigma}_D\right\} \end{aligned}$$

El vector de medias y la matriz de covarianzas para $\bar{\boldsymbol{Y}}$ son respectivamente

$$E\left(\bar{\boldsymbol{Y}}\right) = \boldsymbol{Z}^t\boldsymbol{\Gamma} \quad \text{y} \quad Var\left(\bar{\boldsymbol{Y}}\right) = \boldsymbol{a}\boldsymbol{\Sigma}_u\boldsymbol{a}^t + \boldsymbol{\Sigma}_e \triangleq \boldsymbol{V}$$

donde $\boldsymbol{V} = \text{diag}\left\{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_d, \ldots, \boldsymbol{V}_D\right\}$ es una matriz de tamaño $DR \times RD$, con $\boldsymbol{V}_d$ una matriz de tamaño $R \times R$ dada por $\boldsymbol{V}_d = \sigma_u^2\mathbf{1}_R\mathbf{1}_R^t + \boldsymbol{\Sigma}_d$, con $\boldsymbol{\Sigma}_d$ la matriz de varianzas y covarianzas asociada a los errores aleatorios del $d-$ésimo dominio y $\boldsymbol{a} = \text{diag}\left(\mathbf{1}_{1R}, \ldots, \mathbf{1}_{2R}, \ldots \mathbf{1}_{DR}\right)$. Cuando la matriz de covarianzas $\boldsymbol{V}$ se asume conocida Henderson (1975) propuso el mejor predictor líneal insesgado (BLUP, Best Linear Unbiased Predictor) para cualquier combinación lineal de efectos fijos $\boldsymbol{\Gamma}$ y efectos aleatorios $\boldsymbol{u}$ en el modelo general dado en (1-9). Sin embargo, en la práctica, la matriz de covarianzas $\boldsymbol{V}$ es desconocida, entonces se requiere calcular una estimación $\hat{\sigma}_u^2$ para $\sigma_u^2$. Cuando se utiliza esta estimación, se obtiene un $BLUP$ Empírico ($EBLUP$) para $\boldsymbol{\mu}$, el cual toma la forma

$$\hat{\boldsymbol{\mu}}_{EBLUP} = \hat{\boldsymbol{\mu}}\left(\hat{\sigma}_u^2, \bar{\boldsymbol{Y}}\right) \quad = \quad \boldsymbol{Z}^t\hat{\boldsymbol{\Gamma}}_{EBLUP} + \delta\hat{\boldsymbol{u}}_{EBLUP} \tag{1-10}$$

donde

$$\hat{\boldsymbol{\Gamma}}_{EBLUP} = \hat{\boldsymbol{\Gamma}}\left(\hat{\sigma}_u^2, \bar{\boldsymbol{Y}}\right) \quad = \quad \left(\boldsymbol{Z}^t\boldsymbol{V}^{-1}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^t\boldsymbol{V}^{-1}\bar{\boldsymbol{Y}} \tag{1-11}$$

$$\hat{\boldsymbol{u}}_{EBLUP} = \hat{\boldsymbol{u}}\left(\hat{\sigma}_u^2, \bar{\boldsymbol{Y}}\right) \quad = \quad \Sigma_u\boldsymbol{Z}^t\boldsymbol{V}^{-1}\left(\bar{\boldsymbol{Y}} - \boldsymbol{Z}^t\hat{\boldsymbol{\Gamma}}_{EBLUP}\right) \tag{1-12}$$

En el trabajo de González-Manteiga, Lombardía, Molina, Morales & Santamaría (2008) muestran dos posibles estimadores de $\sigma_u^2$ así como algunas propiedades asintóticas.

## 1.5. El problema de datos faltantes en las encuestas y sus posibles soluciones.

Cochran (1977) asegura que la presencia de datos faltantes en una encuesta se puede presentar por

- *Cobertura*: No se pudo ubicar algunas de las unidades seleccionadas debido a problemas de acceso.
- *Localización*: Cuando no fue posible localizar a ningún miembro del hogar durante la visita a la vivienda seleccionada.
- *Informante inadecuado*: Cuando la persona entrevistada no está en posibilidad de

proporcionar la información que se le demanda.

- *Rechazo*: Cuando los hogares seleccionados se niegan a participar en la encuesta.

El no realizar un tratamiento adecuado de esta falta de respuesta parcial (cuando el entrevistado no responde la totalidad de la encuesta por falta de información, desconocimiento, vergüenza o temor a contestar, entre otras) o totales (cuando el entrevistado no es localizado o rechaza realizar la encuesta), puede generar situaciones no deseadas en la fase de inferencia estadística, tales como la aparición de sesgos, subestimación de las varianzas, sobreestimación de las correlaciones y subcobertura de los intervalos de confianza (García, Albaladejo & Fernández 2006). Por tanto, es necesario la implementación de procedimientos para la imputación de la información faltante, pero la inadecuada utilización de estos procedimientos también nos lleva a problemas en las inferencias, tales como el sobredimensionar el poder explicativo de los modelos (Acock 2005).

Para hacer un manejo adecuado de los datos faltantes se hace necesario saber cuáles son las causas que llevan a la presencia de estos, así como la naturaleza de los valores faltantes y su relación con los valores de otras variables dentro de la matriz de observaciones. Hasta los años setenta los métodos más utilizados para tratar el problema de los datos faltantes eran la eliminación de los casos con valores perdidos y la sustitución/imputación de estos por valores plausibles como la media de la variable o la predicción obtenida mediante regresión sobre las demás variables del vector. Sin embargo, estos métodos clásicos no suelen tener una base teórica sólida. Luego, Rubin (1976) propuso un marco conceptual para el análisis de datos faltantes sustentado en métodos de inferencia estadística, el cual se describe a continuación: Sea $\boldsymbol{X}^*$ un vector aleatoria de tamaño $1 \times P$, sea $\boldsymbol{X}_i$ una realización aleatoria de $\boldsymbol{X}^*$ y sea $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{ip}, \ldots, \lambda_{iP})$ un vector de tamaño $1 \times P$ formado por variables aleatorias binarias tal que

$$\lambda_{ip} = \begin{cases} 1 & \text{si } x_{ip} \text{ es observado} \\ 0 & \text{si } x_{ip} \text{ no es observado} \end{cases}$$

Llamaremos mecanismo de no respuesta a la distribución de probabilidad de $\boldsymbol{\lambda}_i$, con lo cual tenemos que, para cada realización aleatoria de $\boldsymbol{X}^*$ tendremos una realización de $\boldsymbol{\lambda}_i$, cuya forma dependerá de la complejidad del patrón de no respuesta. Así, cuando se tengan $n_s$ realizaciones de $\boldsymbol{X}^*$, resultarán las matrices $\boldsymbol{X}_s = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i \ldots, \boldsymbol{X}_{n_s})$ y $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_i, \ldots, \boldsymbol{\lambda}_{n_s})$ los cuales tendrían dimensiones de $n_s \times P$ donde $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{ip}, \ldots, x_{iP})$ es la realización aleatoria de la variable $\boldsymbol{X}^*$ y $\Lambda$ es una matriz de ceros y unos, de donde se sigue que $\boldsymbol{X}_s = (\boldsymbol{X}_{obs}, \boldsymbol{X}_{miss})$ donde $\boldsymbol{X}_{obs}$ son los casos en que $\lambda_{ip} = 1$ y $\boldsymbol{X}_{miss}$ los casos en

que $\lambda_{ip} = 0$.
El supuesto principal que debe hacerse es sobre la naturaleza del mecanismo que genera $\mathbf{\Lambda}$, los cuales son clasificados por Rubin (1976) y Little & Rubin (1987) en los siguientes grupos:

- Cuando la probabilidad de que un dato sea considerado perdido es la misma para todos los casos, es decir que: $Pr(\mathbf{\Lambda} \mid \boldsymbol{X}_s, \boldsymbol{\xi}) = Pr(\mathbf{\Lambda} \mid \boldsymbol{\xi})$ con $\boldsymbol{\xi}$ un vector de parámetros desconocidos asociado al mecanismo que genera la no respuesta. Bajo está condición la presencia de datos faltantes es por completa aleatoriedad y es denotada como MCAR (Missing Completely At Random) por sus siglas en inglés.

- Un mecanismo menos restrictivo que el anterior es aquel donde la falta de datos depende sólo de $\boldsymbol{X}_{obs}$, es decir que: $Pr(\mathbf{\Lambda} \mid \boldsymbol{X}_s, \boldsymbol{\xi}) = Pr(\mathbf{\Lambda} \mid \boldsymbol{X}_{obs}, \boldsymbol{\xi})$ para todo $\boldsymbol{X}_{miss}$, los datos se conocen como faltantes por aleatoriedad y se denotan como MAR (Missing At Random) por sus siglas en inglés.

- Cuando la variable de interés determina el patrón de ausencia, entonces se define el mecanismo como no aleatorio, es decir que: $Pr(\mathbf{\Lambda} \mid \boldsymbol{X}_s, \boldsymbol{\xi}) = Pr(\mathbf{\Lambda} \mid \boldsymbol{X}_{obs}, \boldsymbol{X}_{miss}, \boldsymbol{\xi})$ el mecanismo se denota por NMAR (Not Missing At Random) por sus siglas en ingles.

Durante las últimas décadas han propuesto distintas metodologías para estimar los datos faltantes (Ver Little 1992, Little & Rubin 1987). Entre estas clasificaciones se encuentra el método de Imputación Múltiple (IM) propuesto por Rubin (1987) que se sustenta en la premisa de que cada dato faltante debe ser reemplazado a partir de $M > 1$ valores plausibles. La implementación de esta técnica se facilitó con los actuales avances computacionales y el desarrollo de métodos bayesianos de simulación que aparecieron hacia finales de los ochenta. A continuación se hace una introducción al método de Imputación Múltiple propuestos por Rubin (1987)

### 1.5.1. Modelo de imputación múltiple

El método de imputación múltiple puede abordarse desde dos enfoques: i) realizar varias imputaciones de las observaciones faltantes para luego analizar los conjuntos de datos completados y combinar los resultados obtenidos para obtener una estimación final. ii) obtener la distribución posterior para el parámetro de interés y de esta empleando el muestreo de Gibbs (Geman & Geman 1987) y así obtener valores plausibles del parámetro los cuales se combinan para obtener una estimación final, estas dos técnicas son descritas en la presente sección.

Sea $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{ip}, \ldots, x_{iP})$ el vector fila de tamaño $1 \times P$ que representa la $i-$ésima observaciones de la matriz $\boldsymbol{X}_s$ la cual es de tamaño $n_s \times P$. Considere que la distribución conjunta de $\boldsymbol{X}_s$ es la distribución de probabilidad $Pr\left(\boldsymbol{X}_s \mid \boldsymbol{\Theta}\right)$ condicionada al vector de parámetros $\boldsymbol{\Theta}$.
Sea $Pr\left(\lambda_{ip} = 0 \mid x_{ip}\right) = Pr\left(x_{ip}\,\text{No observado} \mid x_{ip}\right) = \xi_{ip}$, entonces la distribución de probabilidad de $\boldsymbol{\Lambda}$ está dada por $P\left(\boldsymbol{\Lambda} \mid \boldsymbol{X}_s, \boldsymbol{\xi}\right)$. Bajo este supuesto, la distribución de probabilidad conjunta de $\boldsymbol{X}_s$ y $\boldsymbol{\Lambda}$ puede ser expresada como:

$$Pr\left(\boldsymbol{X}_s, \boldsymbol{\Lambda} \mid \boldsymbol{\Theta}, \boldsymbol{\xi}\right) \;=\; Pr\left(\boldsymbol{X}_s \mid \boldsymbol{\Theta}\right) Pr\left(\boldsymbol{\Lambda} \mid \boldsymbol{X}_s, \boldsymbol{\xi}\right), \tag{1-13}$$

donde $Pr\left(\boldsymbol{X}_s \mid \boldsymbol{\Theta}\right)$ es la distribución marginal de $\boldsymbol{X}_s$ y $Pr\left(\boldsymbol{\Lambda} \mid \boldsymbol{X}_s, \boldsymbol{\xi}\right)$ es la distribución condicional de los datos faltantes dado $\boldsymbol{X}_s$.
Note que, en el modelo de probabilidad (1-13), hay dos conjuntos de parámetros, los parámetros de interés $\boldsymbol{\Theta}$ y los parámetros $\boldsymbol{\xi}$ asociados a la no respuesta. En general, las inferencias sobre $\boldsymbol{\Theta}$ deben realizarse sobre la base del modelo de probabilidad conjunta (1-13). Por lo tanto, estas inferencias dependen de como se defina el modelo de probabilidad para los datos faltantes.
Para un conjunto de datos incompletos la función de probabilidad conjunta esta dada por $Pr\left(\boldsymbol{X}_{obs}, \boldsymbol{X}_{miss}, \boldsymbol{\Lambda} \mid \boldsymbol{\Theta}, \boldsymbol{\xi}\right)$, dado que $\boldsymbol{X}_{miss}$ es desconocido, la función de verosimilitud de esta distribución puede ser escrita como:

$$L\left(\boldsymbol{\Theta}, \boldsymbol{\xi} \mid \boldsymbol{X}_{obs}, \boldsymbol{\Lambda}\right) \;\propto\; Pr\left(\boldsymbol{X}_{obs}, \boldsymbol{\Lambda} \mid \boldsymbol{\Theta}, \boldsymbol{\xi}\right) \tag{1-14}$$

donde

$$\begin{aligned} Pr\left(\boldsymbol{X}_{obs}, \boldsymbol{\Lambda} \mid \boldsymbol{\Theta}, \boldsymbol{\xi}\right) &= \int Pr\left(\boldsymbol{X}_{obs}, \boldsymbol{X}_{miss}, \boldsymbol{\Lambda} \mid \boldsymbol{\Theta}, \boldsymbol{\xi}\right) d\boldsymbol{X}_{miss} \\ &= \int Pr\left(\boldsymbol{\Lambda} \mid \boldsymbol{X}_{obs}, \boldsymbol{X}_{miss}, \boldsymbol{\xi}\right) Pr\left(\boldsymbol{X}_{obs}, \boldsymbol{X}_{miss} \mid \boldsymbol{\Theta}\right) d\boldsymbol{X}_{miss} \end{aligned} \tag{1-15}$$

Bajo las definiciones de MCAR y MAR dadas en la sección (1.5), la ecuación (1-15) se expresa como

$$Pr\left(\boldsymbol{X}_{obs}, \boldsymbol{\Lambda} \mid \boldsymbol{\Theta}, \boldsymbol{\xi}\right) \;=\; \begin{cases} Pr\left(\boldsymbol{\Lambda} \mid \boldsymbol{\xi}\right) Pr\left(\boldsymbol{X}_{obs} \mid \boldsymbol{\Theta}\right) & \text{Si } MCAR \\ Pr\left(\boldsymbol{\Lambda} \mid \boldsymbol{\xi}, \boldsymbol{X}_{obs}\right) Pr\left(\boldsymbol{X}_{obs} \mid \boldsymbol{\Theta}\right) & \text{Si } MAR \end{cases}$$

si $\boldsymbol{\Theta}$ y $\boldsymbol{\xi}$ son distintos, las inferencias basadas en la verosimilitud de $\boldsymbol{\Theta}$, pueden realizarse

basadas en $Pr\left(\boldsymbol{X}_{obs} \mid \boldsymbol{\Theta}\right)$, esto quiere decir que la distribución de los datos observados $Pr\left(\boldsymbol{X}_{obs}, \boldsymbol{\Lambda} \mid \boldsymbol{\Theta}, \boldsymbol{\xi}\right)$ puede ser reemplazada por la distribución marginal de los datos observados $Pr\left(\boldsymbol{X}_{obs} \mid \boldsymbol{\Theta}\right)$ con el propósito de hacer inferencias sobre $\Theta$, y la función de verosimilitud de los datos observados, ignorando el mecanismo aleatorio que determina los datos faltantes,

$$L\left(\boldsymbol{\Theta} \mid \boldsymbol{X}_{obs}\right) \quad \propto \quad Pr\left(\boldsymbol{X}_{obs} \mid \boldsymbol{\Theta}\right). \tag{1-16}$$

Ahora, si suponemos que $g\left(\boldsymbol{\Theta}\right)$ es una distribución a priori para $\boldsymbol{\Theta}$ entonces, se tiene que la distribución posterior de $\boldsymbol{\Theta}$ condicionada a los datos observados esta dada por

$$Pr\left(\boldsymbol{\Theta} \mid \boldsymbol{X}_{obs}\right) \quad \propto \quad L\left(\boldsymbol{\Theta} \mid \boldsymbol{X}_{obs}\right) \times g\left(\boldsymbol{\Theta}\right) \tag{1-17}$$

Las inferencias de verosimilitud para $\boldsymbol{\Theta}$ se basan en la función de verosimilitud observada (1-16) y las inferencias bayesianas para $\boldsymbol{\Theta}$ se basan en la distribución posterior dada en (1-17). Estas inferencias utilizan los datos observados y son válidas siempre y cuando el mecanismo que genera los patrones de no respuesta sea MCAR. Sin embargo, la precisión de la inferencia se reduce si falta una gran cantidad de información.

La metodología antes mencionada la utiliza PISA (Programme for International Student Assessment), TERCE (Tercer Estudio Regional Comparativo y Explicativo), SESS (Special Education Support Service), entre otras pruebas internacionales, para entregar resultados a los países que participan en estos estudios. En Colombia, está metodología fue adoptada por el Icfes para calificar la prueba Saber 3° 5° y 9°; y así poder entregar resultados a los establecimientos educativos evaluados. En la Sección 2.1 se hace una breve explicación de la implementación de esta metodología por el Icfes.

La segunda forma de abordar el problema de datos faltantes es; imputar los datos perdidos $\boldsymbol{X}_{miss}$ y luego aplicar métodos tradicionales a la matriz que ha sido imputada, para realizar las inferencias sobre los parámetros $\boldsymbol{\Theta}$. La distribución de probabilidad condicional de $\boldsymbol{X}_{miss}$ dado $\boldsymbol{X}_{obs}$ puede derivarse integrando sobre el espacio de parámetros $\boldsymbol{\Theta}$, esto es

$$Pr\left(\boldsymbol{X}_{miss} \mid \boldsymbol{X}_{obs}\right) \quad = \quad \int Pr\left(\boldsymbol{X}_{miss} \mid \boldsymbol{X}_{obs}, \boldsymbol{\Theta}\right) Pr\left(\boldsymbol{\Theta} \mid \boldsymbol{X}_{obs}\right) d\boldsymbol{\Theta} \tag{1-18}$$

donde $Pr\left(\boldsymbol{\Theta} \mid \boldsymbol{X}_{obs}\right)$ es la distribución posterior de $\boldsymbol{\Theta}$ y $Pr\left(\boldsymbol{X}_{miss} \mid \boldsymbol{X}_{obs}, \boldsymbol{\Theta}\right)$ es denominada por Rubin (1987) como la distribución condicional predictiva de $\boldsymbol{X}_{miss}$ dado $\boldsymbol{X}_{obs}$ y $\boldsymbol{\Theta}$; y la distribución condicional $Pr\left(\boldsymbol{X}_{miss} \mid \boldsymbol{X}_{obs}\right)$ se denomina distribución predictiva posterior de $\boldsymbol{X}_{miss}$ dado $\boldsymbol{X}_{obs}$. Está metodología de estimación es la utilizada en el presente trabajo para obtener las estimaciones del vector de parámetro $\boldsymbol{\beta}$. En la Sección 1.5.2 se muestra las

ecuaciones para obtener las estimaciones de $\boldsymbol{\beta}$ y $Var\left(\boldsymbol{\beta}\right)$.
La estimación de las ecuaciones 1-17 y 1-18 se hace mediante el uso del algoritmo EM (Dempster, Laird & Rubin 1977) el cual es un algoritmo particularmente importante para el análisis de datos faltantes bajo los dos siguientes pasos:

- **Paso E:** obtener el valor esperado para los datos faltantes usando una estimación inicial del parámetro.
- **Paso M:** obtener la estimación máximo verosímil de los parámetros usando los datos obtenidos en el paso E.
- Repetir los pasos E y M hasta lograr la convergencia.

Para una mayor información e implementación del algoritmo EM puede consultar Ibrahim, Chen, Lipsitz & Herring (2005), McLachlan & Krishnan (2007). Por otro lado, en la Tabla **1-1** se muestra una lista corta de paquetes en R (R Core Team 2018) que realizan la imputación de valores perdidos empleando una variedad de herramientas. En la Sección 3.5 se describen los métodos de imputación múltiple implementados en el desarrollo del presente trabajo.

### 1.5.2. Estimación con valores plausibles

Rubin (1987) a partir de las ecuaciones 1-17 y 1-18 perfiló fórmulas relativamente sencillas que están basadas en la teoría asintótica, para obtener las estimaciones de los parámetros y los errores estándar. Por ejemplo, suponga que $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ son variables aleatorias independientes e idénticamente distribuidas las cuales se relacionan a través de la función lineal $f$, tal y como se indicó en la Sección 1.3.2. Ahora, asuma que $\boldsymbol{Y}_s^t = (y_1, \ldots, y_7)$ es un vector fila de tamaño $1 \times 7$ y

$$\boldsymbol{X}_s = \begin{bmatrix} x_{11} & ? & x_{13} \\ x_{21} & x_{22} & ? \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & ? \\ x_{51} & x_2 & x_{53} \\ x_{61} & x_{62} & x_{63} \\ x_{71} & ? & x_{73} \end{bmatrix} \tag{1-19}$$

luego, $\boldsymbol{X}_s$ puede descomponer de la forma $\boldsymbol{X}_s = (\boldsymbol{X}_{obs}, \boldsymbol{X}_{miss})$, el interés es obtener una estimación de los parámetros desconocidos del modelo $\boldsymbol{\beta}^t = (\beta_1, \beta_2, \beta_3)$ y $\sigma_\epsilon > 0$ dada la información disponible.

| Paquete | Autor(s) | Métodos |
|---|---|---|
| `MissingDataGUI` | Cheng, Cook & Hofmann (2016) | Se aplican una variedad de métodos de imputación, incluyendo las imputaciones univariadas como los valores fijos o aleatorios, las imputaciones multivariadas como el vecino más cercanos, imputaciones múltiples e imputaciones condicionadas a una variable categórica |
| `Amelia` | Honaker, King & Blackwell (2011) | Las imputaciones se realizan a través del algoritmo EMB (expectation-maximization with bootstrapping). |
| `BaBooN` | Meinfelder & Schnapp (2015) | Proporciona dos variantes de la aproximación Bootstrap bayesiana para imputar múltiples valores faltantes. Originalmente el método se desarrolló para datos de encuestas. |
| `Hmisc` | Harrell Jr, Dupont & many others. (2018) | El paquete realiza la imputación múltiple mediante la regresión aditiva, el bootstrap y la comparación de medias predictiva. |
| `missMDA` | Josse & Husson (2016) | Realiza métodos de componentes principales en conjuntos de datos incompletos obteniendo puntuaciones, cargas y representaciones gráficas a pesar de los valores perdidos. |
| `MICE` | Van Buuren, Stef and Groothuis-Oudshoorn, Karin (2011) | Realiza imputación múltiple utilizando Fully Conditionally Specification (FCS) implementado por el algoritmo MICE (Multiple Imputation by Chained Equations). |

**Tabla 1-1**.: Listado de paquetes en *R* que realizan imputación

El primer paso es obtener una estimación de la función de distribución dada en 1-18, la cual es posible estimar utilizando el algoritmo EM. Una vez se tiene está estimación podemos obtener $M$ realizaciones aleatorias para $\boldsymbol{X}_{miss}$ condicionado a la distribución probabilidad. Ahora, sea $x_{ip}^{(m)}$ la $m-$ésima realización aleatoria (valor plausible) obtenida para la observación $x_{ip}$la cual no ha sido observada. Sea $\boldsymbol{X}_s^{(m)}$ con $m = 1, \ldots, M$ la $m-$ésima matriz plausible de $\boldsymbol{X}_s$, que resulta después de reemplazar todos los $x_{ip}$ perdidos por su respectivo $m-$ésimo valores plausible. Los $M$ valores plausibles se obtienen como una realización aleatoria de la función de distribución dada en 1-18. En particular, para la matriz $\boldsymbol{X}_s$ dada 1-19 y $M = 4$ se tiene que:

$$\underbrace{\begin{bmatrix} x_{11} & x_{12}^{(1)} & x_{13} \\ x_{21} & x_{22} & x_{23}^{(1)} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43}^{(1)} \\ x_{51} & x_2 & x_{53} \\ x_{61} & x_{62} & x_{63} \\ x_{71} & x_{72}^{(1)} & x_{73} \end{bmatrix}}_{\boldsymbol{X}_s^{(1)}} \quad \underbrace{\begin{bmatrix} x_{11} & x_{12}^{(2)} & x_{13} \\ x_{21} & x_{22} & x_{23}^{(2)} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43}^{(2)} \\ x_{51} & x_2 & x_{53} \\ x_{61} & x_{62} & x_{63} \\ x_{71} & x_{72}^{(2)} & x_{73} \end{bmatrix}}_{\boldsymbol{X}_s^{(2)}} \quad \underbrace{\begin{bmatrix} x_{11} & x_{12}^{(3)} & x_{13} \\ x_{21} & x_{22} & x_{23}^{(3)} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43}^{(3)} \\ x_{51} & x_2 & x_{53} \\ x_{61} & x_{62} & x_{63} \\ x_{71} & x_{72}^{(3)} & x_{73} \end{bmatrix}}_{\boldsymbol{X}_s^{(3)}} \quad \underbrace{\begin{bmatrix} x_{11} & x_{12}^{(4)} & x_{13} \\ x_{21} & x_{22} & x_{23}^{(4)} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43}^{(4)} \\ x_{51} & x_2 & x_{53} \\ x_{61} & x_{62} & x_{63} \\ x_{71} & x_{72}^{(4)} & x_{73} \end{bmatrix}}_{\boldsymbol{X}_s^{(4)}} \tag{1-20}$$

El segundo paso es, obtener la estimación de $\boldsymbol{\beta}$ y su varianza, utilizando las ecuaciones 1-6 y 1-8 respectivamente a partir de $\boldsymbol{X}_s^{(m)}$, de está forma es posible obtener $M$ estimaciones para $\boldsymbol{\beta}$. La $m-$ésima estimación de $\boldsymbol{\beta}$ la denotaremos como $\hat{\boldsymbol{\beta}}_{(m)}^t = \left(\hat{\beta}_1^{(m)}, \ldots, \hat{\beta}_p^{(m)}, \ldots, \hat{\beta}_P^{(m)}\right)$. Por ultimo, calcular $\hat{\boldsymbol{\beta}}_{IM}^t = \left(\hat{\beta}_1^{(IM)}, \ldots, \hat{\beta}_p^{(IM)}, \ldots, \hat{\beta}_P^{(IM)}\right)$ que es el estimador insesgado para $\boldsymbol{\beta}$ y viene por:

$$\hat{\boldsymbol{\beta}}_{IM} \quad = \quad \frac{1}{M} \sum_{m=1}^{M} \hat{\boldsymbol{\beta}}_{(m)} \tag{1-21}$$

La varianza total estimada del estimador $\hat{\boldsymbol{\beta}}_{IM}^t$ está formada por el promedio de las varianzas obtenida para los $M$ valores plausibles y la varianza de las varianzas calculadas en cada valor plausible, esto es:

$$\widehat{Var}\left(\hat{\boldsymbol{\beta}}_{IM}\right) \quad = \quad W + \left(1 + \frac{1}{M}\right) B \tag{1-22}$$

donde

$$\begin{aligned} W &= \frac{1}{M}\sum_{m=1}^{M}\widehat{Var_m}\left(\hat{\boldsymbol{\beta}}\right) \\ B &= \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\boldsymbol{\beta}}_{(m)}-\hat{\boldsymbol{\beta}}_{IM}\right)\left(\hat{\boldsymbol{\beta}}_{(m)}-\hat{\boldsymbol{\beta}}_{IM}\right)^{t} \end{aligned}$$

Donde, $\widehat{Var_m}\left(\hat{\boldsymbol{\beta}}\right)$ es la varianza estimada de $\hat{\boldsymbol{\beta}}_{(m)}$ calculada a partir de la ecuación 1-8. En particular, para $p$ se tiene que:

$$\hat{\beta}_p^{(IM)} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_p^{(m)}$$

$$\begin{aligned} W &= \frac{1}{M}\sum_{m=1}^{M}\widehat{Var_m}\left(\hat{\beta}_p\right) \\ B &= \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_p-\hat{\beta}_p^{(IM)}\right)^2 \end{aligned}$$

Para una discusión más detallada sobre la metodología de los valores plausibles puede consultar Rubin (1987), Zhang (2003), Longford (2005).

# 2. Metodología

En este Capítulo se presenta la metodología mediante la cual se entrega respuesta a los interrogantes que dieron origen a este trabajo. En la Sección 2.1 se introduce la metodológica desarrollada por el Icfes, para combinar el método de imputación múltiple de Rubin (1987) con el modelo de estimación en áreas pequeñas de Fay & Herriot (1979) para pronosticar el promedio académico en el área de matemáticas en establecimientos no observados. En la Sección 2.2 es descrita la metodología propuesta en este trabajo para predecir los coeficientes de un modelo de regresión ($\boldsymbol{\beta}$) en dominios no observados, y se toma la metodología propuesta por el Icfes (Ver Sección 2.1) para hacer el desarrollo de está.

## 2.1. Prueba Saber 3°, 5° y 9°

La prueba Saber 3°, 5° y 9° es una evaluación estandarizada, realizada por el Icfes, con el objeto obtener información sobre las competencias de los estudiantes de educación básica por medio de una evaluación estandarizada. Esta prueba representa un componente fundamental de la estrategia de mejoramiento de la calidad de la educación, puesto que permite valorar si los estudiantes están alcanzando las metas y si están mejorando con el paso del tiempo. La aplicación de la prueba contempla la evaluación de todos los estudiantes que se encuentran cursando los grados tercero, quinto y noveno; sobre las áreas de lenguaje, matemáticas, ciencias naturales y competencias ciudadanas.

El Icfes, durante el 2015 realizó un estudio donde combinó la metodología de valores plausibles de Rubin y SAE para pronosticar los promedios en matemáticas y lenguaje el quinto grado de los establecimientos educativos que presentaron la prueba de Saber 3°, 5° y 9° en el 2013. El interés en este estudio era estimar la habilidad promedio ($\mu$) de los estudiantes que cursaban el 5 grado y pertenecían a un establecimiento educativo con más de 5 de estudiantes en este grado. [1]

[1] Los establecimientos con 5 o menos estudiantes son excluidos, dado que la estructura de la prueba Saber 3°, 5° y 9° implementada por el Icfes, exige más de 5 estudiantes por establecimiento educativo para realizar el reporte de resultados (Icfes 2011).

El marco muestral utilizado por el Icfes, fue restringido a los establecimientos educativos con resultados publicados en matemáticas y lenguaje, esto con el objetivo de evaluar la eficiencia del modelo SAE para predecir la habilidad promedio ($\mu$) de los establecimientos educativos no incluidos en la muestra.
El Icfes, en sus bases de datos registraba 11.226 establecimientos educativos a los cuales se les reportó la calificación promedio en matemáticas y/o lenguaje para la prueba de Saber 3°, 5° y 9° en el 5 grado. Para los 11.226 establecimientos del marco muestral del Icfes, había información de los estudiantes que presentaron las pruebas Saber en 2013 y variables asociadas al establecimiento como la zona, sector, entidad territorial, resultados históricos en matemáticas y lenguaje para las pruebas Saber 3°, 5° y 9°. Para los establecimientos que cuentan con el un décimo grado se tienen los resultados promedios en las pruebas Saber 11° de 2012.
Así mismo, el Icfes seleccionó a 477 para presentar las pruebas Saber 3°, 5° y 9 en 2013 bajo la vigilancia de un veedor, contratado por la entidad. Con está muestra se realizó la predicción de la habilidad promedio ($\mu$) de los 10.749 establecimientos que no fueron seleccionados. En el Icfes, para poder implementar la metodología SAE, se consideró a los establecimientos como dominios de interés, de tal manera que aquellos que no resultaron seleccionados en la muestra de 2013 sean considerados como áreas pequeñas (Rao & Molina 2015).

### 2.1.1. Calificación de la prueba.

Las pruebas Saber 3°, 5° y 9° evalúan a cada establecimiento educativo de Colombia en las pruebas y sub-pruebas de Matemáticas y Lenguaje, además se alternan las pruebas de Ciencias Naturales y Pensamiento Ciudadano entre aplicaciones. El Icfes para poder evaluar de manera adecuada a los establecimientos debe diseñar una cantidad suficiente de preguntas (ítems).
Teniendo en cuenta la gran cantidad de ítems y el tiempo que tienen los estudiantes para contestar la prueba, se haría imposible que un estudiante pudiese responder toda la prueba. De así que, los ítems sean distribuidos de manera aleatoria entre los estudiantes empleando un diseño de bloques incompletos balanceados (BIBD). Para que la implementación del diseño sea adecuado, el Icfes debe garantizar que cada bloque contenga ítems con las mismas características de contenido y dificultad.
Debido a que los estudiantes responden solo un subconjunto de la cantidad de ítems desarrollados para la prueba, la medición de su habilidad individual esta sujeta a un error de medición. Los enfoques tradicionales de estimación de la habilidad, como la máxima verosimilitud marginal (MML) y estadística bayesiana, arrojan estimaciones óptimas para

la habilidad individual pero sesgadas a nivel de grupo.
En Icfes (2011) se explica a detalle el procedimiento utilizado por la institución para realizar el proceso de calificación de las pruebas Saber 3°, 5° y 9°. Para ilustrar el proceso de calificación presentado en Icfes (2011) suponga que se quiere estimar el promedio en Matemáticas para el $d-$ésimo establecimiento el cual se denota por $\mu_d$. Ahora, asuma que se dispone de $M = 5$ valores plausibles para los $n_d$ estudiantes que presentaron la prueba de Matemáticas. Sea $y_{di}^{(m)}$ el $m-$ésimo valor plausible asignado al $i-$ésimo estudiante evaluado en el $d-$ésimo establecimiento educativo de Colombia, donde $y_{di}^{(m)}$ fue obtenido a partir de la ecuación 1-17 para $i = 1, \ldots, n_d$.
Luego, la estimación $\mu_d$ está dada por la ecuación 1-21, así:

$$\begin{aligned} \hat{\mu}_d = \hat{\bar{y}}_{d,(IM)} &= \frac{1}{5}\sum_{m=1}^{5}\left(\frac{1}{n_d}\sum_{i=1}^{n_d} y_{di}^{(m)}\right) \\ &= \frac{1}{5}\sum_{m=1}^{5}\hat{\bar{y}}_d^{(m)} \end{aligned} \tag{2-1}$$

donde, $\hat{\bar{y}}_d^{(m)} = \frac{1}{n_d}\sum_{i=1}^{n_d} y_{di}^{(m)}$. La varianza estimada para $\hat{\bar{y}}_{d\,(IM)}$ se obtiene con la ecuación 1-22, así:

$$\widehat{Var}_d\left(\hat{\bar{y}}_{d\,(IM)}\right) = W_d + \left(1 + \frac{1}{5}\right)B_d \tag{2-2}$$

con

$$\begin{aligned} W_d &= \frac{1}{5}\sum_{m=1}^{5}\widehat{Var}_d\left(\hat{\bar{y}}_d^{(m)}\right) \\ B_d &= \frac{1}{5-1}\sum_{m=1}^{5}\left(\hat{\bar{y}}_d^{(m)} - \hat{\bar{y}}_{d\,(IM)}\right)^2 \end{aligned}$$

donde $\widehat{Var}_d\left(\hat{\bar{y}}_d^{(m)}\right)$ es la varianza estimada para $\hat{\bar{y}}_d^{(m)}$ en el $d-$ésimo establecimiento a partir del diseño muestral. Los parámetros $\hat{\bar{y}}_d^{(m)}$ y $\widehat{Var}_d\left(\hat{\bar{y}}_d^{(m)}\right)$ son obtenidos a partir del el estimador de Horvitz-Thompson, el estimador general de regresión y el estimador compuesto. Además, de forma alternativa se empleó el método Jackknife para estimar $Var_d\left(\hat{\bar{y}}_d^{(m)}\right)$ en cada uno de los establecimiento educativos.

### 2.1.2. Estimación de los resultados empleando el modelo de Fay-Herriot

Ajustaremos un modelo a nivel de área, suponiendo los establecimientos como áreas pequeñas, que enlace la información auxiliar del $d-$ésimo establecimiento con la estimación de $\mu_d$. Recordemos que, para el ajuste del modelo de Fay & Herriot (1979) se requiere de dos insumos principales.

1. Una estimación para el resultado en "Matemáticas" de la prueba Saber 5° a nivel de establecimiento educativo (el nivel superior).

2. Una varianza del resultado en matemáticas de la prueba Saber 5° para cada establecimiento educativo obtenida con los estudiantes en cada establecimientos.

El modelo empleado por el Icfes corresponde al caso univariado de Fay-Herriot, es decir, cuando el número de variables $R = 1$ en la ecuación 1-9 de la sección 1.4.1, en particular para el $d-$ésimo establecimiento se tiene que

$$\hat{\mu}_d = \boldsymbol{z}_d^t \boldsymbol{\gamma} + u_d + e_d, \quad u_d \sim N\left(0, \sigma_u^2\right), e_d \sim N\left(0, \sigma_e^2\right) \quad d = 1, \ldots, D \tag{2-3}$$

donde $\boldsymbol{\gamma}^t = (\gamma_1, \ldots, \gamma_k, \ldots, \gamma_K)$ es el vector de coeficientes de tamaño $K \times 1$ y $\boldsymbol{z}_d = (z_{d1}, \ldots, z_{dk}, \ldots, z_{dK})$ es de tamaño $K \times 1$ y representa la información auxiliar asociada a $\mu_d$ en el $d-$ésimo establecimiento, $e_d$ y $u_d$ son efectos aleatorios asociados al $d-$ésimo establecimiento, los cuales se asumen independientes e idénticamente distribuidos.[2]
Para el ajuste del modelo (2-3) asume que $\hat{\sigma}_d^2 = \widehat{Var_d}\left(\hat{\bar{y}}_{d(IM)}\right)$, luego el $EBLUP$ para $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_d, \ldots, \hat{\mu}_D)$, toma la forma

$$\hat{\mu}_{d,EBLUP} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_d^2} \hat{\bar{y}}_{d(IM)} + \frac{\hat{\sigma}_d^2}{\hat{\sigma}_u^2 + \hat{\sigma}_d^2} \boldsymbol{z}_d^t \hat{\boldsymbol{\gamma}} \tag{2-4}$$

[2] Observación: $R$ representa el número de variables respuestas en el modelo, $K$ representa la número de variables auxiliares a utilizar en el modelo y $P$ representa la información auxiliar utilizada para estimar $\hat{\mu}_d$.

donde

$$
\begin{aligned}
\hat{\gamma}_{EBLUP} = \hat{\gamma} &= \left(\boldsymbol{Z}^t \boldsymbol{V}^{-1} \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^t \boldsymbol{V}^{-1} \hat{\bar{\boldsymbol{y}}}_{IM} \\
\hat{\boldsymbol{u}}_{EBLUP} = \hat{\boldsymbol{u}} &= \hat{\Sigma}_u \boldsymbol{Z}^t \boldsymbol{V}^{-1} \left(\hat{\bar{\boldsymbol{y}}}_{IM} - \boldsymbol{Z}^t \hat{\gamma}\right) \\
\hat{\bar{\boldsymbol{y}}}_{IM} &= \left(\hat{\bar{y}}_{1\,(IM)}, \ldots, \hat{\bar{y}}_{d\,(IM)}, \ldots, \hat{\bar{y}}_{D\,(IM)}\right) \\
\boldsymbol{Z}^t &= \left(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_d, \ldots, \boldsymbol{z}_D\right) \\
\hat{\boldsymbol{u}} &= \left(u_d, \ldots, u_d, \ldots, u_D\right) \\
\boldsymbol{V}^{-1} &= \operatorname{diag}\left(\left(\hat{\sigma}_u^2 + \hat{\sigma}_1^2\right)^{-1}, \ldots, \left(\hat{\sigma}_u^2 + \hat{\sigma}_d^2\right)^{-1}, \ldots, \left(\hat{\sigma}_u^2 + \hat{\sigma}_D^2\right)^{-1}\right) \\
\hat{\Sigma}_u &= \hat{\sigma}_u^2 \boldsymbol{I}_D
\end{aligned}
$$

Como estrategia para la comparación de los resultados obtenidos con las estimaciones obtenidas para las ecuaciones 2-1 y 2-2 con estimadores de Horvitz-Thompson, de regresión y compuestos se utilizó el error cuadrático medio (MSE, por sus siglas en inglés) definido mediante la expresión:

$$
MSE = \frac{1}{D} \sum_{d=1}^{D} \left(\hat{\mu}_{d,EBLUP} - \bar{\boldsymbol{Y}}_d\right)^2
$$

en el cual, $\bar{\boldsymbol{Y}}_d$ corresponde a los resultados reales obtenido por el $d-$ésimo establecimientos educativos en el área de matemáticas en la prueba Saber 5° de 2013 y $D = 11{,}226$ representa el número de los establecimientos educativos para los cuales se reportaron los resultados en el 2013. El Icfes, concluyó que el modelo de SAE puede emplearse para predecir los resultados en las pruebas Saber 5° en establecimientos no observados.
En la siguiente sección se desarrolla la mitología propuesta en el presente trabajo para realizar la estimación del parámetro $\boldsymbol{\beta}$ en dominios no observados, tomando como base los desarrollos implementados por el Icfes.

## 2.2. Estimación de los parámetros de un modelo lineal en dominios no observados.

En está sección se propone una metodología para lograr estimar el vector de parámetros $\boldsymbol{\beta}_d$ de un modelo de regresión lineal en dominios con presencia de datos faltantes y empleando la versión multivariante del modelo Fay & Herriot (1979) (ver sección 1.4.1) realizar la predicción de estos parámetros en dominios no observados. Para lograr esto se parte de las

siguientes consideraciones:

- Sea $S = \cup_{d=1}^{D} S_d$ una muestra probabilística seleccionada de $U = \cup_{d=1}^{D} U_d$, tal que $S \cap U_d = S_d$. Sea $n_s = \sum_{d=1}^{D} n_d$ el número de observaciones en $S$ y $n_d$ el número de observaciones en el $d-$ésimo dominio de interés.

- Sea $p(\cdot)$ el diseño de muestreo mediante el cual fue seleccionada la muestra $s$ con probabilidades de inclusión $\pi_i$.

- Sea $\boldsymbol{X}_{sdi} = (x_{di1}, \ldots, x_{dip}, \ldots, x_{diP})$ un vector fila de tamaño $1 \times P$ que representa las realizaciones aleatorias del vector aleatorio la $\boldsymbol{X}^*$. Sea $\boldsymbol{X}_{sd}^t = (\boldsymbol{X}_{d1}, \ldots, \boldsymbol{X}_{di}, \ldots, \boldsymbol{X}_{dn_d})$ una matriz de tamaño $P \times n_d$ que representa las realizaciones aleatorias de los vectores aleatorios $\boldsymbol{X}_{d1}^*, \ldots, \boldsymbol{X}_{di}^*, \ldots, \boldsymbol{X}_{dn_d}^*$ en el $d-$ésimo dominio y sea $\boldsymbol{X}_s = (\boldsymbol{X}_{s1}, \ldots \boldsymbol{X}_{sd}, \ldots, \boldsymbol{X}_{sD})$ una matriz de tamaño $n_s \times P$

- Sea $Y_{d1}, \ldots, Y_{di}, \ldots, Y_{dn_d}$ variables aleatorias independientes e idénticamente distribuidas. Sea $\boldsymbol{y}_d = (y_{d1}, \ldots, y_{di}, \ldots y_{dn_d})$ un vector columna de tamaño $1 \times n_d$ donde $y_{di}$ es la realización aleatoria de $Y_{di}$ para el $i-$ésimo elemento de $S_d$.

- Sea $f$ la función lineal que relaciona la variable aleatoria $Y_{di}$ con $\boldsymbol{X}_{sdi}$. Está función tiene como objetivo explicar o predecir los valores de una variable $Y_{di}$, que se considera como la variable de respuesta. Ahora, asuma que $(Y_{di}, \boldsymbol{X}_{sdi})$ son variables aleatorias independientes e idénticamente distribuidas tales que

$$f\left(y_{di} \mid \boldsymbol{X}_{sdi}; \boldsymbol{\beta}_d, \sigma_d\right) = (2\pi\sigma_d)^{-\frac{1}{2}} \exp\left[-\frac{\left(y_{di} - \boldsymbol{X}_{di}\boldsymbol{\beta}_d\right)^2}{2\sigma_d}\right] \tag{2-5}$$

  observe que la relación entre $Y_{di}$ y $\boldsymbol{X}_{sdi}$ es definida de forma independiente para cada dominio, luego, van a estimar tantos $\boldsymbol{\beta}$'s de tamaño $P \times 1$ como dominios tenga la población $U$.

Ahora, se plantean algunos casos que pueden surgir cuando se selecciona la muestra $S$ y en la cual se tiene como interés estimar $\boldsymbol{\beta}_d$:

### Caso 1: La muestra $S$ es observada completamente

Suponga que $\boldsymbol{X}_{di}$ es observado para todo elemento de $S_d \subset S$ , entonces, es posible estimar $\boldsymbol{\beta}_d$ y su varianza mediante las ecuaciones 1-6 y 1-8; la estimación de $\sigma_d$ se obtiene con la ecuación 1-7.

### Caso 2: La muestra $S_d$ no fue observada en todos los dominios

Asuma que para algún domino $d$, $S_d$ no fue observada, entonces, bajo este escenario se utilizara la versión multivariante del modelo Fay & Herriot (1979) (Ver sección 1.4.1), haciendo a $\bar{\boldsymbol{Y}} = \left(\hat{\boldsymbol{\beta}}_{1,MPV}, \ldots, \hat{\boldsymbol{\beta}}_{d,MPV}, \ldots, \hat{\boldsymbol{\beta}}_{D,MPV}\right)$ y $\Sigma_d = \widehat{Var}_d\left(\hat{\boldsymbol{\beta}}_{d,MPV}\right)$ donde $\hat{\boldsymbol{\beta}}_{d,MPV}$ y $\widehat{Var}_d\left(\hat{\boldsymbol{\beta}}_{d,MPV}\right)$ están dados por las ecuaciones 1-6 y 1-8 respectivamente. Luego, para el $d-$ésimo domino $\hat{\boldsymbol{\beta}}_{d,EBLUP} = \boldsymbol{Z}_d^t \hat{\boldsymbol{\Gamma}}_{EBLUP}$ donde $\hat{\boldsymbol{\Gamma}}_{EBLUP}$ esta dado por la ecuación 1-11.

### Caso 3: La muestra $S_d$ es observada para todos los dominios pero tiene información faltante en $\boldsymbol{X}_{sd}$

Considere que $\boldsymbol{X}_{di}$ no puede ser observado para todo $i \in S_d$, entonces, $\boldsymbol{X}_{sd} = (\boldsymbol{X}_{d,obs}, \boldsymbol{X}_{d,miss})$ donde $\boldsymbol{X}_{d,obs}$ son los elementos observados ($\lambda_{ip} = 1$) y $\boldsymbol{X}_{d,miss}$ los elementos no observados ($\lambda_{ip} = 0$) (Ver Sección 1.5), bajo este escenario la estimación de $\boldsymbol{\beta}_d$ y su varianza puede obtenerse utilizando la metodología de valores plausibles descrita en la Sección 1.5.2.

### Caso 4: La muestra $S_d$ no fue observada en todos los dominios y $\boldsymbol{X}_{sd}$ tiene información faltante cuando $S_d$ es observada

En este caso, combinamos los casos 2 y 3, dado que consideramos que $S_d$ y $\boldsymbol{X}_{sd}$ no pueden ser observados en su totalidad, sea por que surgieron dominios no planeados o la información auxiliar $\boldsymbol{X}_{di}$ no esta disponible para todo $i \in S_d$ que fue seleccionada.
Suponga que $\boldsymbol{X}_{di}$ no es observado para todo $i \in S_d$, luego, $\boldsymbol{X}_d = (\boldsymbol{X}_{d,obs}, \boldsymbol{X}_{d,miss})$. Además, sin perdida de generalidad, suponga que sean observado los primeros $D_0$ dominios, es decir, $D - D_0$ dominios no son observados, por tanto $S_d = \emptyset$ para $d = D_0 + 1, \ldots, D$. Bajo estás condiciones se emplea la metodología de valores plausibles para obtener el estimador $\hat{\boldsymbol{\beta}}_{d,IM}$ de $\boldsymbol{\beta}_d$ para $d = 1, \ldots, D_0$ y para obtener una estimación de $\boldsymbol{\beta}$ en los $D - D_0$ dominios utiliza el modelo multivariante Fay-Herriot, tomando a $\bar{\boldsymbol{Y}} = \left(\hat{\boldsymbol{\beta}}_{1,IM}, \ldots, \hat{\boldsymbol{\beta}}_{d,IM}, \ldots, \hat{\boldsymbol{\beta}}_{D_0,IM}\right)$ y $\Sigma_d = \widehat{Var_d}\left(\hat{\boldsymbol{\beta}}_{d,IM}\right)$ donde $\hat{\boldsymbol{\beta}}_{d,IM}$ y $\widehat{Var_d}\left(\hat{\boldsymbol{\beta}}_{d,IM}\right)$ están dados por las ecuaciones 1-21 y 1-22 respectivamente. Ahora, la predicción de $\boldsymbol{\beta}_d$ para los dominios no observados se obtiene como: $\hat{\boldsymbol{\beta}}_{d,EBLUP} = \boldsymbol{Z}_d^t \hat{\boldsymbol{\Gamma}}_{EBLUP}$.
Una vez sea realizado el cambio de la variable respuesta $\bar{\boldsymbol{Y}}$ se debe estudiar el comportamiento del modelo Fay-Herriot para predecir a $\boldsymbol{\beta}_d$ en los dominios no observados, para lograr esto se propone el estudio de simulación que se muestra en el Capítulo siguiente.

# 3. Estudio por simulación

El interés en este apartado es examinar mediante un estudio de simulación el comportamiento de la metodología propuesta para pronosticar los coeficientes $\boldsymbol{\beta}$ de un modelo de regresión linea múltiple en áreas no observadas haciendo uso de la metodología de estimación en áreas pequeñas multivariada.
Para evaluar la propuesta metodológica hecha en los casos 2 y 4 de la Sección 2.2, es necesario hacer la estimación del parámetro $\boldsymbol{\beta}$ empleando la metodología valores plausibles descrita en la Sección 1.5.2. Para hacer esto posible, la matriz $\boldsymbol{X}$ debe tener valores perdidos, de tal forma que el mecanismo de no respuesta siga un comportamiento aleatorio MCAR, posteriormente se realiza la imputación de estos valores empleando los métodos de PMM (Predictive Mean Matching), Regresión lineal bayesiana (RLB) y Combinación bootstrap-imputación múltiple (RLBoot) que están disponibles en la librería `mice` del paquete R. Por último, se evalúan los resultados obtenidos mediante comparaciones respecto a la raíz del error cuadrático medio.

## 3.1. Caso de interés en la simulación

En la Sección 2.2, se contemplan algunos de los casos que pueden ser encontrados en la práctica, a continuación, mencionaremos las soluciones que se han desarrollados y que pueden encontrarse en la literatura.

#### Solución Caso 1

De los casos que presentaremos a continuación, este caso es el más sencillo que se pueden presentar, puesto que $S$ no presenta información faltante. Adicional a esto, todos los dominios tienen suficiente información para obtener estimaciones directas. Pessoa & Silva (1998) muestran las expresiones matemáticas para la estimación de $\boldsymbol{\beta}$, y adicional a esto, presentan las bondades de dicho estimador. Estas ecuaciones están implementadas en la librería `survey` (Lumley 2016) del software R, y por lo anterior, este caso no será abordado en este trabajo.

### Solución Caso 2

En este caso, se aborda el problema que surge cuando se desea obtener estimaciones directas en dominios no planeados (o cuando el tamaño de la muestra en el dominio es insuficiente para tener estimaciones de calidad). En la literatura especializada es posible encontrar metodologías que se pueden combinar para obtener estimaciones o predicciones del parámetro de interés. Por ejemplo, es posible mezclar el modelo multivariante de Fay-Herriot y las expresiones matemáticas de Pessoa & Silva (1998), para realizar la predicción del parámetro $\boldsymbol{\beta}$ en los dominios no observados, para ello se debe asumir que $\bar{\boldsymbol{Y}} = \hat{\boldsymbol{\beta}}_{MPV}$, luego, el modelo de Fay-Herriot multivariante se expresa así,

$$\hat{\boldsymbol{\beta}}_{MPV} \quad = \quad \boldsymbol{Z}^t\boldsymbol{\Gamma} + \boldsymbol{\delta} + \boldsymbol{e} \tag{3-1}$$

donde $\hat{\boldsymbol{\beta}}_{MPV}$ esta dado por la ecuación 1-6.
En literatura especializada no fue posible encontrar documentación donde se aborde esté problema. Por lo tanto, es abordo como un caso particular dentro del estudio de simulación propuesto en la Sección 3.6.

### Solución Caso 3

El problema de datos faltantes descrito en éste caso es posible abordarlo como un problema de datos faltantes tradicional y que cuentan con abundante literatura especializada en la solución del mismo, en este sentido puede remitirse a Rubin (1976), Little & Rubin (1987), Little (1992), Raghunathan, Lepkowski, Van Hoewyk & Solenberger (2001), Longford (2005), Ibrahim, Chen, Lipsitz & Herring (2005), Van Buuren, Stef and Groothuis-Oudshoorn, Karin (2011) y hacer uso de la técnica de imputación de valores perdidos que mejor se ajuste a la muestra $S$ seleccionada. La implementación de algunas de estas técnicas en el software R están consignadas en la Tabla **1-1**.

### Solución Caso 4

De los casos presentados, esté tiene la mayor dificultad pues asume que:

1. La muestra $S$ tiene información faltante en la matriz $\boldsymbol{X}_s$.
2. Para algún dominio $d$ se tiene que $S_d = \emptyset$, es decir, no fue posible observar todos los dominios de la población $U$, lo que dificulta obtener una estimación directa $\boldsymbol{\beta}_d$.

Las condiciones antes mencionadas hacen que el caso 4 sea el mayor interés en este trabajo, puesto que, la estimación de $\boldsymbol{\beta}$ en los dominios observados se debe hacer implementado la

metodología de valores plausibles (Ver Sección 1.5.2) y la predicción de $\boldsymbol{\beta}$ en los dominios no observados se hace empleando el modelo multivariante de Fay-Herriot, que es el objetivo de este trabajo.

Para implementar el modelo de Fay-Herriot multivariante, debemos asumir que la variable respuesta $\bar{\boldsymbol{Y}}$ en la ecuación 1-9 toma los valores de $\boldsymbol{\beta}$, cuyo estimador es $\hat{\boldsymbol{\beta}}_{IM}$ el cual esta dado por la ecuación 1-21 y la varianza estimada viene dada por la ecuación 1-22, luego el modelo Fay-Herriot se escribe como:

$$\hat{\boldsymbol{\beta}}_{IM} \quad = \quad \boldsymbol{Z}^t\boldsymbol{\Gamma} + \boldsymbol{\delta} + \boldsymbol{e} \tag{3-2}$$

Note que, al realizar el cambio de variable respuesta en el modelo 1-9, el contador del número de variables en $\boldsymbol{Y}$ y el contador de las covariables $\boldsymbol{X}_s$ se hacen iguales, es decir, $R = P$.

## 3.2. Simulación de la población

El proceso para simular la población es realizado en dos etapas, para esto debe recordar que el modelo a nivel de área esta dado por la ecuación 3-2 y el modelo de regresión lineal para el $i-$ésimo elemento del $d-$ésimo dominio esta dado por

$$f\left(y_{di} \mid \boldsymbol{X}_{di}; \boldsymbol{\beta}_d, \sigma_d\right) \quad = \quad \left(2\pi\sigma_d\right)^{-\frac{1}{2}} \exp\left[-\frac{\left(y_{di} - \boldsymbol{X}_{di}\boldsymbol{\beta}_d\right)^2}{2\sigma_d}\right], \tag{3-3}$$

por lo que es necesario: *i)* Se debe simula la información agregada a nivel de área, es decir, $\boldsymbol{\beta}_d = \boldsymbol{Z}_d^t\boldsymbol{\Gamma} + \boldsymbol{\delta}_d + e$, con $\boldsymbol{\Gamma}^t = \left(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p, \ldots, \boldsymbol{\gamma}_P\right)$ un vector de tamaño $1 \times RK$ y $\boldsymbol{\gamma}_d^t = (\gamma_1, \ldots, \gamma_k, \ldots, \gamma_K)$ de tamaño $1 \times K$ un vector de constantes arbitrarias. *ii)* Dado el $\boldsymbol{\beta}_d$ simulado en la etapa 1, simular $\boldsymbol{Y}_d$ tal que $\boldsymbol{Y}_d = \boldsymbol{X}_d^t\boldsymbol{\beta}_d + \epsilon_d$, que replica el modelo dado en 3-3. Cada etapa del proceso se detalla a continuación:

- **Etapa 1:** Para el $d-$ésimo dominio, asuma que $\boldsymbol{z}_d = (z_{d1}, \ldots, z_{dk}, \ldots, z_{dK})$ es un vector de tamaño $K \times 1$ que sigue una distribución $N_K\left(\boldsymbol{0}, \boldsymbol{R}_z\right)$. Luego, el $p-$ésimo elemento del vector $\boldsymbol{\beta}_d^t = (\beta_1, \ldots, \beta_p, \ldots, \beta_P)$ se define como $\beta_{dp} = c_p + \boldsymbol{z}_d^t\boldsymbol{\gamma}_p + u_d + e$ donde $c_p$ sigue una distribución $U\left(a, b\right)$, $u_d \sim n\left(0, \sigma_u\right)$ y $e \sim n\left(0, \sigma_e\right)$, observe que $E\left(\beta_{dp}\right) = \frac{a+b}{2}$. Ahora, la matriz $\boldsymbol{R}_z$ se toma de forma similar a la presentada por Benavent & Morales

(2016), así

$$\boldsymbol{\mathcal{R}}_z = \begin{bmatrix} 1 & \rho_z^1 & \dots & \rho_z^{K-1} \\ \rho_z^1 & 1 & \dots & \rho_z^{K-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_z^{K-1} & \rho_z^{K-2} & \dots & 1 \end{bmatrix}$$

donde, $\rho_z$ toma los valores de $0,4$; $0,6$ y $0,9$; note que $\boldsymbol{\mathcal{R}}_z$ es $AR\,(1)$.

- **Etapa 2:** Dentro del $d-$ésimo domino se define $\boldsymbol{Y}_d = \boldsymbol{X}_d^t L\boldsymbol{\beta}_d + \epsilon_d$, donde$\boldsymbol{\beta}_d$ fue obtenido en la etapa 1, $\boldsymbol{X}_d \sim N_P\,(\boldsymbol{\mu}, \boldsymbol{\mathcal{R}}_d)$, $\boldsymbol{\mu}^t = (\mu_1, \dots, \mu_P)$ es un vector con valores arbitrario, $\boldsymbol{L}$ se obtiene de la descomposición de Cholesky tal que $\boldsymbol{\mathcal{R}}_d = \boldsymbol{L}^t\boldsymbol{L}$ es obtenida mediante métodos iterativos de tal forma que la $Cor\,(Y_d, \boldsymbol{X}_{dp}) = \rho_x$ y $\epsilon_d \sim N\,(0, \sigma_\epsilon)$.

  Note que, cuando $P > K$ el número de parámetros $\boldsymbol{\beta}$ será mayor que el número de variables auxiliares $\boldsymbol{Z}$ asociadas a $d-$ésimo dominio de $U$. Ahora, cuando $P \leq K$ sucederá que, el número de variables auxiliares $\boldsymbol{Z}$ asociadas a $d-$ésimo dominio de $U$ entonces, el número de parámetros $\boldsymbol{\beta}$ es menor. Estos dos casos fueron parte de los escenarios de simulación evaluados en este trabajo.

Sea $U$ la población la objetivo, la cual tiene $N = 40{,}000$ observaciones agrupadas en $D = 200$ dominios tal que $E\,(N_d) = 200$ observaciones. Sea $P = 3$ y $5$ el número de variables regresoras $\boldsymbol{X}$ en la población $U$, de esta forma, el número de coeficientes de regresión $\beta$'s a predecir en el $d-$ésimo dominio es $P$. Sea $K = 4$ el número de variables auxiliares $\boldsymbol{Z}$ asociadas con el $d-$ésimo dominio de $U$, tales que $Cor\,(\boldsymbol{\beta}, \boldsymbol{Z}) \approx \boldsymbol{\mathcal{R}}_z$. Los elementos de la matriz $\boldsymbol{\mathcal{R}}_d$ se define como; Bajo (B) cuando $\rho_x \in (0{,}25, 0{,}50]$; Medios (M) si, $\rho_x \in (0{,}50, 0{,}75]$ y Altos (A) para $\rho_x \in (0{,}75, 1]$. Los valores para $\rho_x \in (0, 0{,}25]$ no son considerados, dado que cuando $\rho_x$ es "pequeño" es mejor emplear el modelo univarido de Fay-Herriot.

## 3.3. Simulación de datos faltantes.

En la población $U$ simulada en sección anterior, todas los valores son completamente observados luego, se remueven algunos de ellos de tal forma que el patrón de no respuesta reproduzca el mecanismo MCAR, para así tener un conjunto con datos faltantes.

Ahora, para reproducir el mecanismo MCAR, asuma que $\lambda_{ip}$ es una variable aleatoria Bernoulli tal que $x_{ip}$ es observado si, $\lambda_{ip} = 1$ y $x_{ip}$ es un valor no observado si $\lambda_{ip} = 0$, de esta forma se garantiza que dentro del conjunto de datos, todas las observaciones tiene

exactamente la misma probabilidad de estar ausente. Ahora, sea $\alpha = 0{,}1$ y 0,3 la probabilidad $\lambda_{ip} = 1$, así se introduce un 10 % y 30 % información faltante a las variables regresoras $\boldsymbol{X}$ de la población $U$.

## 3.4. Selección de la muestra

El proceso de selección de la muestra será mediante un diseño en dos etapas, donde las *Unidades Primarias de Muestreo* (UPM) son los dominios y se seleccionan $n_I = (50, 100, 150)$, mediante un muestreo aleatorio simple sin reemplazo (M.A.S.). Las *Unidades Secundarias de Muestreo* (USM) corresponden a las observaciones pertenecientes al $d-$ésimo dominio seleccionado en la etapa anterior son seleccionadas utilizando el M.A.S.

## 3.5. Métodos de imputación

En la Sección 3.3 se remueven observaciones de la población $U$, de tal forma que estos repliquen el mecanismo MCAR, de esta forma tener una población conjunto de datos faltantes. Está Sección se introducen las técnicas de imputación Predictive Mean Matching, Regresión lineal bayesiana y Combinación bootstrap-imputación múltiple disponibles en la librería `mice` del software R y fueron empleando durante el estudio de simulación para generar los valores plausibles.

### 3.5.1. Coincidencia de media predictiva

El método conocido como Predictive Mean Matching (PMM) (Rubin 1986, Little 1988), ha tenido auge recientemente debido a los avancen en la tecnología e implementación de esté en paquetes estadísticos. La idea básica del modelo de imputación PMM es utilizar métodos de regresión para llegar a una estimación del dato faltante de la variable $\boldsymbol{X}_s = (\boldsymbol{X}_1, \ldots \boldsymbol{X}_d, \ldots, \boldsymbol{X}_D)$. Sin embargo, en lugar de utilizar dicha estimación, la emplea para construir una métrica para emparejar registros con datos faltantes a registros similares ($\kappa$) con datos presentes. Está método se encuentra disponible en el software estadístico R en las librerías `mice` (Van Buuren, Stef and Groothuis-Oudshoorn, Karin 2011), `Hmisc` (Harrell Jr, Dupont & many others. 2018), `BaBooN` (Meinfelder & Schnapp 2015) y `mi` (Gelman & Hill 2011).

### 3.5.2. Regresión lineal bayesiana

Las funciones `mice.impute.norm()` y `mice.impute.norm.nob()` del la librería `mice` imputan las observaciones faltantes de acuerdo a modelos de regresión lineal bayesiano y no bayesiano para la imputación, y son rápidas y eficientes si los residuos del modelo son aproximadamente normales. El segundo modelo ignora cualquier incertidumbre de muestreo del modelo de imputación, por lo que solo es apropiado para muestras muy grandes.
La regresión lineal bayesiana (RLB) para la imputación de valores perdidos se deriva del metodología desarrollada por Rubin (1987). A continuación ilustramos cómo usar el algoritmo MICE para extraer iterativamente muestras para realizar la imputación múltiple.

- Calcular la matriz $\boldsymbol{\mathcal{S}} = \boldsymbol{X}_{obs}^{t}\boldsymbol{X}_{obs}$.
- Calcular $\boldsymbol{\mathcal{V}} = (\boldsymbol{\mathcal{S}} + \operatorname{diag}(\mathcal{S})\,\kappa)$ , donde $\kappa$ es el numero de observaciones coincidentes.
- Calcular los coeficientes de regresión $\hat{\boldsymbol{\beta}} = \boldsymbol{\mathcal{V}}\boldsymbol{X}_{obs}^{t}\boldsymbol{Y}_{obs}$.
- Simular una variable aleatoria $\tau \sim \chi_{v}^{2}$ con $v = n_s - P$
- Calcular $\dot{\sigma} = (\boldsymbol{Y}_{obs} - \boldsymbol{X}_{obs}\boldsymbol{\beta})^{t}(\boldsymbol{Y}_{obs} - \boldsymbol{X}_{obs}\boldsymbol{\beta})\,/\tau$
- Simular $P$ variables independientes $N(0,1)$ en el vector $\dot{\boldsymbol{x}}_1$.
- Calcular $\boldsymbol{\mathcal{V}}^{1/2}$ por descomposición de Cholesky.
- Calcular $\dot{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \dot{\sigma}\dot{\boldsymbol{x}}_1\boldsymbol{\mathcal{V}}^{1/2}$.
- Extraer $n_{miss}$ variables aleatorias independientes $N(0,1)$ en el vector $\dot{\boldsymbol{x}}_2$.
- Calcular los $n_{miss}$ valores de $\boldsymbol{Y}_{miss} = \boldsymbol{X}_{miss}\dot{\boldsymbol{\beta}} + \dot{\boldsymbol{x}}_2\dot{\sigma}$.

### 3.5.3. Combinación bootstrap-imputación múltiple

Schomaker & Heumann (2016) consideran varias formas posibles de combinar bootstrap e imputación múltiple. Los dos enfoques principales son: *i)* imputar primero los datos faltantes y luego usar el bootstrap para obtener una estimación del error estándar dentro de la imputación para cada conjunto de datos imputados *ii)* Seleccionar muestras bootstrap de los datos, y aplicar técnicas de imputación múltiple por separado a cada muestra seleccionada. La estimación puntual encontrada con cada muestra se usa luego para construir el error estándar bootstrap e intervalos de confianza de la manera habitual.

## 3.6. Diseño del estudio de simulación

En el estudio de simulación se desea evaluar los estimadores propuestos en la Sección 2.2. Por tal motivo, se divide el estudio en dos partes:

- **Parte 1:** Generar una población $U$ como se indica en la Sección 3.2, luego, realizar la selección de la muestra empleando el diseño de muestreo descrito en 3.4. El estudio de simulación se realiza utilizando dos configuraciones diferentes para los tamaños de las muestras: *i)* Se fija un tamaño de muestra del 20 % de las observaciones para $d-$ésimo domino seleccionado, luego, el número de observaciones esperadas $n_s$ será de 2000, 4000 y 6000 para $n_I = 50, 100$ y 150 respectivamente. *ii)* Sea $n_s = 4000$ el número de observaciones esperadas en la muestra y asigne $n_{sd} = n_s \times \frac{N_d}{N}$, por tanto el porcentaje de muestra esperado para el $d-$ésimo domino seleccionado será de 40 % 20 % y 13 % para cada $n_I$ respectivamente. Esta parte de la simulación se emplea para dar respuesta al caso 2.2 de la Sección 2.2.

- **Parte 2:** En esta parte del estudio se quiere dar respuesta al caso 2.2 de la Sección 2.2, por lo tanto, para una población artificial construida a partir del pseudocódigo descrito en la Sección 3.2 se le remueven valores tales que repliquen el comportamiento de un mecanismo MCAR (ver Sección 3.3). El paso a seguir es seleccionar muestras de tamaño $n_s$ empleando el diseño de muestro descrito en la Sección 3.4. Para el estudio se fijo que $n_s$ será de 2000, 4000 y 6000 que corresponde a muestras del 20 % de las observaciones del $d-$ésimo domino seleccionado.

  Los $M = 5$ valores plausibles se obtienen utilizando los tres métodos de imputación descritos en la Sección 3.5, por último, se emplea la ecuación 3-2 para predecir el parámetro $\boldsymbol{\beta}$ en los dominios no observados.

Los procesos descritos anteriormente se repiten $B = 1000$ veces, para poblaciones con el 0 %, 10 % y 30 % de información faltante.

## 3.7. Medida de calidad de las estimaciones.

El último paso es evaluar los resultados obtenidos de las estimaciones, para ello nos valdremos de las medidas de calidad que se muestran a continuación:
Sea $\vec{\boldsymbol{\beta}} = (\beta_{11}, \ldots, \beta_{1P}, \ldots, \beta_{D1}, \ldots, \beta_{DP})$ el vector de parámetros de interés y sea $\vec{\boldsymbol{\beta}}_{b,EBLUP}$ la estimación de $\vec{\boldsymbol{\beta}}$ obtenido mediante el estimador propuestos, para la $b-$ésima simulación

donde $b = 1, \ldots, B$, se define la raíz del error cuadrático medio (RMSE) como

$$RMSE = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}_{b,EBLUP}\right)^{t}\left(\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}_{b,EBLUP}\right)}$$

Ahora, se considera de interés el $p-$ésimo elemento de $\boldsymbol{\beta}_d$. Sea $E(\beta_{dp}) = \bar{\beta}_{dp,EBLUP} = B^{-1}\sum_{b=1}^{B}\beta_{bdp,EBLUP}$ y sea $\vec{\boldsymbol{\beta}}_p = (\beta_{1p}, \ldots, \beta_{dp}, \ldots, \beta_{Dp})$, se define el sesgo relativo promedio como:

$$\overline{BiasR_p} = \left[\frac{1}{D}\sum_{d=1}^{D}\frac{\bar{\beta}_{bdp,EBLUP} - \beta_{dp}}{\beta_{dp}}\right] \times 100$$

y

$$RMSE_p = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\vec{\boldsymbol{\beta}}_p - \vec{\boldsymbol{\beta}}_{bp,EBLUP}\right)^{t}\left(\vec{\boldsymbol{\beta}}_p - \vec{\boldsymbol{\beta}}_{bp,EBLUP}\right)}$$

## 3.8. Resultados de simulación

Los resultados de la simulación para la Parte 1 se muestran en las Tablas **3-1**, **3-2**, **3-3** y **3-4**. Aquí debemos recordar que en las Tablas **3-1** y **3-2** se muestran los resultados cuando el porcentaje de observaciones muestreadas por domino es constante, lo cual hace que la muestra efectiva aumente a medida que el número de dominios aumenta; mientras que en las Tablas **3-3** y **3-4** el número de observaciones por dominio disminuye a medida que el número de dominios seleccionados aumenta (ver Sección 3.6).
De las Tablas **3-5** a la **3-10** observamos los resultados obtenidos en la Parte 2 da la simulación, donde la muestra efectiva aumente a medida que el número de dominios aumenta. A continuación se realiza un breve análisis de los resultados obtenidos en los distintos escenarios de simulación.

### 3.8.1. Resultados cuando la población está completa.

En la Tabla **3-1** se muestran los resultados obtenidos cuando $K \geqslant P$, bajo los resultados mostrados en esta tabla es posible notar que el RMSE del modelo es menor que 1 para todos los escenarios evaluados. En cuanto al sesgo relativo promedio, podemos notar que para la mayoría de los escenarios el valor obtenido es pequeño y de valor negativo, indicando que hay sobreestimación de los parámetros.

En los resultados obtenidos cuando la $\rho_x$ está en $M$ y $\rho_z = 0{,}6$ vemos que el sesgo relativo promedio de $\beta_0$ presenta los valores más altos dentro de los resultados reportados en esta tabla.

El comportamiento descrito anteriormente se repite cuando $P = 3$ y el $n$ esperado de la muestra efectiva es fijo (ver Tabla **3-3**), por lo que se decide hacer un análisis más detallado de estos resultados reportados.

En la Figura **3-1** se muestran los sesgos relativos promedios para los 200 dominios de la población $U$ los cuales están en el intervalo $(0, 1)$ (linea negra en la base del gráfico), sin embargo, el dominio $D135$ tiene un valor atípico en todos los escenarios es evaluados, resultado que puede estar asociado al mecanismo aleatorio con el cual se genero la población $U$. Dado el comportamiento atípico del dominio $D135$ es omitido; y se realiza un nuevo cálculo de la RMSE (ver Anexo, Tabla **A-1**), donde es posible notar la concordancia con lo mostrado en las Tablas **3-1** y **3-3**.

Por los resultados mostrados en las Tablas **3-1** y **3-3**, es posible asumir que la metodología presenta un buen comportamiento bajo las condiciones evaluadas por estos escenarios de simulación.

En la Tabla **3-2** se muestran los resultados obtenidos cuando $K < P$. A la luz de los resultados que se muestran en esta tabla, vemos que cuando $\rho_x$ tiene valores bajos, la RMSE disminuye a medida que $\rho_z$ aumenta. Esta tendencia se puede apreciar en la RMSE calculada de forma individual para los $K$ parámetros estimados. Otro aspecto a rescatar de la tabla es que cuando $\rho_x$ y $\rho_z$ tienen valores bajos es necesario tener un mayor número de dominios en la muestra para tener poca varianza en las estimaciones.

| $\rho_x$ | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ ($BiasR_1$) | $RMSE_2$ ($BiasR_2$) | $RMSE_3$ ($BiasR_3$) |
|---|---|---|---|---|---|---|
| | | 0.4 | 0.739 | 0.768 ( -0.073 ) | 0.699 ( 0.236 ) | 0.747 ( 0.107 ) |
| | 50 | 0.6 | 0.788 | 0.842 ( 0.071 ) | 0.797 ( -0.014 ) | 0.721 ( 0.331 ) |
| | | 0.9 | 0.728 | 0.762 ( -0.004 ) | 0.662 ( -0.157 ) | 0.755 ( 0.008 ) |
| | | 0.4 | 0.641 | 0.673 ( -0.027 ) | 0.634 ( 0.111 ) | 0.614 ( 0.078 ) |
| B | 100 | 0.6 | 0.672 | 0.722 ( 0.038 ) | 0.692 ( -0.023 ) | 0.594 ( 0.287 ) |
| | | 0.9 | 0.615 | 0.651 ( 0.005 ) | 0.581 ( -0.117 ) | 0.612 ( 0.010 ) |
| | | 0.4 | 0.552 | 0.605 ( 0.068 ) | 0.583 ( -0.039 ) | 0.458 ( 0.055 ) |
| | 150 | 0.6 | 0.564 | 0.632 ( -0.028 ) | 0.595 ( -0.035 ) | 0.447 ( 0.246 ) |
| | | 0.9 | 0.489 | 0.529 ( 0.007 ) | 0.494 ( -0.100 ) | 0.441 ( 0.003 ) |
| | | 0.4 | 0.758 | 0.780 ( 0.153 ) | 0.768 ( -0.477 ) | 0.724 ( 0.036 ) |
| | 50 | 0.6 | 0.783 | 0.771 ( **5.764** ) | 0.822 ( 0.561 ) | 0.753 ( 0.128 ) |
| | | 0.9 | 0.746 | 0.730 ( 0.763 ) | 0.716 ( 0.119 ) | 0.789 ( 0.127 ) |
| | | 0.4 | 0.651 | 0.670 ( 0.145 ) | 0.685 ( -0.540 ) | 0.593 ( 0.026 ) |
| M | 100 | 0.6 | 0.675 | 0.681 ( **4.905** ) | 0.725 ( 0.512 ) | 0.613 ( 0.174 ) |
| | | 0.9 | 0.626 | 0.617 ( 0.481 ) | 0.628 ( 0.193 ) | 0.634 ( 0.082 ) |
| | | 0.4 | 0.548 | 0.579 ( 0.139 ) | 0.605 ( -0.659 ) | 0.448 ( 0.026 ) |
| | 150 | 0.6 | 0.568 | 0.585 ( **4.390** ) | 0.637 ( 0.466 ) | 0.468 ( 0.236 ) |
| | | 0.9 | 0.506 | 0.503 ( 0.246 ) | 0.549 ( 0.216 ) | 0.461 ( 0.036 ) |
| | | 0.4 | 0.680 | 0.649 ( -0.301 ) | 0.652 ( 0.427 ) | 0.736 ( -0.049 ) |
| | 50 | 0.6 | 0.712 | 0.651 ( -0.312 ) | 0.749 ( -0.120 ) | 0.733 ( 0.098 ) |
| | | 0.9 | 0.717 | 0.709 ( 0.420 ) | 0.751 ( 0.215 ) | 0.689 ( 0.146 ) |
| | | 0.4 | 0.593 | 0.551 ( -0.213 ) | 0.588 ( 0.328 ) | 0.636 ( -0.046 ) |
| A | 100 | 0.6 | 0.619 | 0.554 ( -0.141 ) | 0.672 ( -0.079 ) | 0.626 ( 0.088 ) |
| | | 0.9 | 0.589 | 0.579 ( 0.268 ) | 0.617 ( 0.168 ) | 0.570 ( 0.124 ) |
| | | 0.4 | 0.515 | 0.465 ( -0.047 ) | 0.537 ( 0.251 ) | 0.539 ( -0.061 ) |
| | 150 | 0.6 | 0.524 | 0.455 ( 0.052 ) | 0.594 ( -0.051 ) | 0.515 ( 0.088 ) |
| | | 0.9 | 0.453 | 0.438 ( 0.094 ) | 0.477 ( 0.121 ) | 0.443 ( 0.116 ) |

**Tabla 3-1**.: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$ y muestra seleccionada por dominio del 20 %.

La Tabla **3-2** muestra que tener $\rho_x$ altos influye de forma favorable los resultados de la RMSE, dado que este disminuye a medida que $\rho_x$ aumenta. En cuanto al sesgo relativo promedio, podemos notar que para la mayoría de los escenarios el valor obtenido es pequeño, salvo el caso cuando $\rho_x \in B$, la muestra por dominios fue de 50 y $\rho_z = 0{,}4$ siendo este un valor acorde con la RMSE mostrado en el tabla.

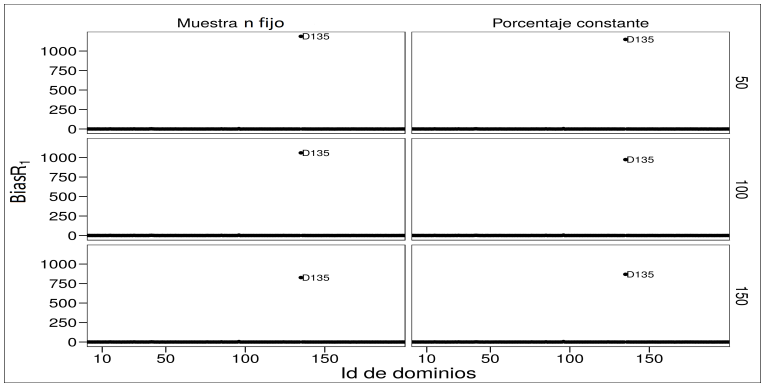


**Figura 3-1**.: Sesgo relativo por dominio para $\beta_1$ cuando $\rho_x \in (0{,}50, 0{,}75]$ y $\rho_z = 0{,}6$

En la Tabla **3-4** se observa los resultados obtenidos cuando $K < P$ y la muestra esperada es constante. Aquí, se nota que para valores bajos de $\rho_x$ y muestras pequeñas por domino (13 % de las observaciones) a media que $\rho_z$ toma valores altos, la RMSE se afecta de forma negativa. Y en mayor media la RMSE del $\beta_1$, sin embargo, estos mejoran a medida que $\rho_x$ aumenta, haciendo que la RMSE y los sesgos relativos promedios no cambian en comparación con los valores obtenidos para el resto de los escenarios.

En conclusión, es posible afirmar que la metodología propuesta tiene sesgos relativos bajos en todos los escenarios evaluados, la RMSE se comporta bien cuando el $K \geqslant P$ independiente de los valores que tome $\rho_x$, $\rho_z$ y el número de observaciones muestreadas por dominio. En el caso que $K < P$ y se tengan valores bajos de $\rho_x$ y $\rho_z$ es necesario tener más observaciones por dominio en la muestra para tener RMSE pequeños.

| $\rho_x$ | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ ($BiasR_1$) | $RMSE_2$ ($BiasR_2$) | $RMSE_3$ ($BiasR_3$) | $RMSE_4$($BiasR_4$) | $RMSE_5$($BiasR_5$) |
|---|---|---|---|---|---|---|---|---|
| | | 0.4 | **44.557** | **96.353** ( **-11.026** ) | **11.320** ( 2.258 ) | **9.097** ( 0.524 ) | **14.832** ( 0.767 ) | **14.551** ( -0.045 ) |
| | 50 | 0.6 | 9.116 | 19.815 ( -1.035 ) | 1.895 ( 0.117 ) | 2.947 ( -0.106 ) | 2.008 ( 0.053 ) | 2.564 ( 0.074 ) |
| | | 0.9 | 2.051 | 3.476 ( -0.403 ) | 1.170 ( -0.234 ) | 1.460 ( 0.393 ) | 1.618 ( 0.171 ) | 1.685 ( 0.305 ) |
| | | 0.4 | 11.609 | 25.467 ( 2.879 ) | 2.547 ( -1.026 ) | 2.538 ( -0.086 ) | 2.088 ( 0.426 ) | 2.814 ( 0.035 ) |
| B | 100 | 0.6 | 2.713 | 5.254 ( -0.620 ) | 1.274 ( 0.118 ) | 1.352 ( -0.064 ) | 1.794 ( 0.097 ) | 1.587 ( 0.047 ) |
| | | 0.9 | 1.648 | 2.555 ( -0.562 ) | 1.098 ( -0.253 ) | 1.346 ( 0.399 ) | 1.427 ( 0.154 ) | 1.415 ( 0.212 ) |
| | | 0.4 | 2.330 | 4.641 ( 0.737 ) | 1.083 ( 0.958 ) | 1.169 ( 0.041 ) | 1.288 ( 0.384 ) | 1.186 ( 0.022 ) |
| | 150 | 0.6 | 1.541 | 2.423 ( -0.381 ) | 1.093 ( 0.033 ) | 1.232 ( -0.066 ) | 1.347 ( 0.103 ) | 1.213 ( 0.050 ) |
| | | 0.9 | 1.446 | 2.214 ( -0.664 ) | 1.055 ( -0.299 ) | 1.259 ( 0.402 ) | 1.249 ( 0.141 ) | 1.139 ( 0.132 ) |
| | | 0.4 | 2.043 | 3.514 ( -2.992 ) | 1.217 ( 0.130 ) | 1.378 ( -0.360 ) | 1.560 ( 0.118 ) | 1.647 ( 0.319 ) |
| | 50 | 0.6 | 2.224 | 3.781 ( 0.202 ) | 1.273 ( 0.706 ) | 1.424 ( 0.077 ) | 1.720 ( 0.171 ) | 1.954 ( 0.240 ) |
| | | 0.9 | 2.119 | 3.548 ( 0.072 ) | 1.183 ( -0.110 ) | 1.321 ( 0.129 ) | 1.735 ( 0.407 ) | 1.926 ( -0.151 ) |
| | | 0.4 | 1.639 | 2.605 ( -2.398 ) | 1.129 ( 0.156 ) | 1.230 ( -0.150 ) | 1.356 ( 0.105 ) | 1.418 ( 0.302 ) |
| M | 100 | 0.6 | 1.769 | 2.766 ( 0.379 ) | 1.150 ( 0.531 ) | 1.302 ( 0.097 ) | 1.485 ( 0.151 ) | 1.664 ( 0.182 ) |
| | | 0.9 | 1.682 | 2.579 ( 0.130 ) | 1.091 ( -0.017 ) | 1.199 ( 0.160 ) | 1.491 ( 0.381 ) | 1.624 ( 0.068 ) |
| | | 0.4 | 1.406 | 2.177 ( 0.259 ) | 1.073 ( 0.157 ) | 1.110 ( -0.048 ) | 1.167 ( 0.092 ) | 1.185 ( 0.283 ) |
| | 150 | 0.6 | 1.481 | 2.212 ( 0.382 ) | 1.052 ( 0.257 ) | 1.197 ( 0.122 ) | 1.266 ( 0.117 ) | 1.389 ( 0.141 ) |
| | | 0.9 | 1.445 | 2.206 ( 0.209 ) | 1.035 ( 0.073 ) | 1.108 ( 0.193 ) | 1.235 ( 0.405 ) | 1.321 ( 0.142 ) |
| | | 0.4 | 2.170 | 3.517 ( -0.477 ) | 1.333 ( 2.208 ) | 1.349 ( 0.315 ) | 1.686 ( 0.095 ) | 2.174 ( -0.011 ) |
| | 50 | 0.6 | 2.179 | 3.737 ( 0.687 ) | 1.204 ( 0.085 ) | 1.409 ( 0.039 ) | 1.570 ( 0.048 ) | 1.968 ( 0.503 ) |
| | | 0.9 | 2.273 | 3.903 ( 0.310 ) | 1.313 ( -0.048 ) | 1.413 ( 1.006 ) | 1.595 ( 0.034 ) | 2.082 ( 0.083 ) |
| | | 0.4 | 1.801 | 2.690 ( -0.488 ) | 1.196 ( 1.697 ) | 1.201 ( 0.224 ) | 1.382 ( 0.069 ) | 2.050 ( -0.031 ) |
| A | 100 | 0.6 | 1.745 | 2.697 ( 0.790 ) | 1.115 ( 0.069 ) | 1.243 ( 0.020 ) | 1.305 ( 0.028 ) | 1.857 ( 0.469 ) |
| | | 0.9 | 1.858 | 2.936 ( 0.331 ) | 1.219 ( -0.034 ) | 1.239 ( 0.868 ) | 1.311 ( 0.019 ) | 1.977 ( 0.077 ) |
| | | 0.4 | 1.550 | 2.192 ( -0.446 ) | 1.082 ( 1.561 ) | 1.076 ( 0.165 ) | 1.069 ( 0.047 ) | 1.933 ( -0.031 ) |
| | 150 | 0.6 | 1.567 | 2.377 ( 0.695 ) | 1.070 ( 0.064 ) | 1.116 ( 0.002 ) | 1.010 ( 0.026 ) | 1.791 ( 0.421 ) |
| | | 0.9 | 1.689 | 2.605 ( 0.305 ) | 1.178 ( -0.011 ) | 1.113 ( 0.713 ) | 1.024 ( -0.003 ) | 1.951 ( 0.088 ) |

**Tabla 3-2.**: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$, muestra seleccionada por dominio del 20 %.

| $\rho_x$ | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ ($BiasR_1$) | $RMSE_2$ ($BiasR_2$) | $RMSE_3$ ($BiasR_3$) |
|---|---|---|---|---|---|---|
| | | 0.4 | 0.731 | 0.744 ( -0.136 ) | 0.703 ( 0.219 ) | 0.746 ( 0.102 ) |
| | 50 | 0.6 | 0.775 | 0.813 ( 0.075 ) | 0.791 ( -0.017 ) | 0.720 ( 0.358 ) |
| | | 0.9 | 0.723 | 0.748 ( -0.004 ) | 0.666 ( -0.144 ) | 0.751 ( 0.005 ) |
| | | 0.4 | 0.642 | 0.670 ( -0.029 ) | 0.641 ( 0.068 ) | 0.613 ( 0.084 ) |
| B | 100 | 0.6 | 0.675 | 0.730 ( 0.041 ) | 0.696 ( -0.027 ) | 0.592 ( 0.342 ) |
| | | 0.9 | 0.611 | 0.643 ( -0.002 ) | 0.579 ( -0.122 ) | 0.610 ( 0.004 ) |
| | | 0.4 | 0.564 | 0.625 ( 0.049 ) | 0.588 ( -0.019 ) | 0.465 ( 0.056 ) |
| | 150 | 0.6 | 0.575 | 0.651 ( -0.040 ) | 0.603 ( -0.039 ) | 0.451 ( 0.235 ) |
| | | 0.9 | 0.499 | 0.546 ( 0.001 ) | 0.497 ( -0.130 ) | 0.449 ( 0.006 ) |
| | | 0.4 | 0.749 | 0.760 ( 0.152 ) | 0.762 ( -0.496 ) | 0.725 ( 0.022 ) |
| | 50 | 0.6 | 0.776 | 0.763 ( **5.965** ) | 0.815 ( 0.563 ) | 0.747 ( 0.142 ) |
| | | 0.9 | 0.737 | 0.720 ( 0.780 ) | 0.707 ( 0.126 ) | 0.781 ( 0.130 ) |
| | | 0.4 | 0.652 | 0.676 ( 0.164 ) | 0.684 ( -0.570 ) | 0.593 ( 0.025 ) |
| M | 100 | 0.6 | 0.673 | 0.675 ( **5.339** ) | 0.725 ( 0.508 ) | 0.615 ( 0.173 ) |
| | | 0.9 | 0.625 | 0.616 ( 0.396 ) | 0.625 ( 0.142 ) | 0.633 ( 0.080 ) |
| | | 0.4 | 0.560 | 0.595 ( 0.146 ) | 0.613 ( -0.661 ) | 0.460 ( 0.024 ) |
| | 150 | 0.6 | 0.573 | 0.599 ( **4.175** ) | 0.638 ( 0.450 ) | 0.468 ( 0.239 ) |
| | | 0.9 | 0.510 | 0.513 ( 0.264 ) | 0.551 ( 0.186 ) | 0.462 ( 0.031 ) |
| | | 0.4 | 0.671 | 0.637 ( -0.362 ) | 0.646 ( 0.451 ) | 0.725 ( -0.045 ) |
| | 50 | 0.6 | 0.716 | 0.638 ( -0.248 ) | 0.759 ( -0.104 ) | 0.744 ( 0.101 ) |
| | | 0.9 | 0.716 | 0.706 ( 0.375 ) | 0.749 ( 0.217 ) | 0.692 ( 0.129 ) |
| | | 0.4 | 0.593 | 0.551 ( -0.167 ) | 0.589 ( 0.292 ) | 0.636 ( -0.063 ) |
| A | 100 | 0.6 | 0.619 | 0.552 ( -0.103 ) | 0.670 ( -0.076 ) | 0.630 ( 0.093 ) |
| | | 0.9 | 0.591 | 0.579 ( 0.277 ) | 0.619 ( 0.165 ) | 0.576 ( 0.128 ) |
| | | 0.4 | 0.525 | 0.477 ( -0.087 ) | 0.546 ( 0.204 ) | 0.549 ( -0.056 ) |
| | 150 | 0.6 | 0.530 | 0.463 ( 0.055 ) | 0.596 ( -0.055 ) | 0.522 ( 0.084 ) |
| | | 0.9 | 0.465 | 0.450 ( 0.113 ) | 0.489 ( 0.131 ) | 0.454 ( 0.102 ) |

**Tabla 3-3**.: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$ cuando el $n$ esperado de la muestra es fijo.

| $\rho_x$ | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ $(BiasR_1)$ | $RMSE_2$ $(BiasR_2)$ | $RMSE_3$ $(BiasR_3)$ | $RMSE_4(BiasR_4)$ | $RMSE_5(BiasR_5)$ |
|---|---|---|---|---|---|---|---|---|
| | | 0.4 | 1.675 | 2.466 ( 0.880 ) | 1.201 ( -1.882 ) | 1.224 ( 0.053 ) | 1.508 ( 0.607 ) | 1.654 ( 0.036 ) |
| | 50 | 0.6 | 1.874 | 2.819 ( -1.098 ) | 1.150 ( 0.175 ) | 1.439 ( -0.074 ) | 1.734 ( 0.054 ) | 1.790 ( 0.054 ) |
| | | 0.9 | 1.751 | 2.558 ( -0.506 ) | 1.164 ( -0.230 ) | 1.442 ( 0.439 ) | 1.603 ( 0.178 ) | 1.671 ( 0.332 ) |
| | | 0.4 | 3.988 | 8.448 ( 0.916 ) | 1.422 ( -0.555 ) | 1.277 ( 0.023 ) | 1.484 ( 0.506 ) | 1.512 ( 0.025 ) |
| B | 100 | 0.6 | 1.729 | 2.688 ( -0.765 ) | 1.124 ( 0.099 ) | 1.327 ( -0.079 ) | 1.539 ( 0.088 ) | 1.523 ( 0.051 ) |
| | | 0.9 | 1.668 | 2.618 ( -0.532 ) | 1.098 ( -0.293 ) | 1.352 ( 0.431 ) | 1.421 ( 0.162 ) | 1.419 ( 0.209 ) |
| | | 0.4 | 32.012 | 70.930 ( -3.680 ) | 4.912 ( 0.289 ) | 4.285 ( 0.093 ) | 3.722 ( 0.473 ) | 6.022 ( -0.032 ) |
| | 150 | 0.6 | 29.143 | 62.943 ( -0.122 ) | 11.709 ( 0.123 ) | 8.855 ( -0.213 ) | 5.264 ( -0.131 ) | 6.442 ( 0.137 ) |
| | | 0.9 | 337.388 | 731.371 ( 23.886 ) | 72.067 ( -2.291 ) | 39.533 ( 0.312 ) | 106.586 ( -1.858 ) | 127.017 ( -1.540 ) |
| | | 0.4 | 1.764 | 2.660 ( -3.102 ) | 1.210 ( 0.150 ) | 1.375 ( -0.412 ) | 1.565 ( 0.127 ) | 1.635 ( 0.348 ) |
| | 50 | 0.6 | 1.912 | 2.829 ( 0.312 ) | 1.258 ( 0.709 ) | 1.417 ( 0.060 ) | 1.709 ( 0.163 ) | 1.938 ( 0.247 ) |
| | | 0.9 | 1.788 | 2.505 ( 0.102 ) | 1.166 ( -0.111 ) | 1.301 ( 0.122 ) | 1.725 ( 0.398 ) | 1.920 ( -0.213 ) |
| | | 0.4 | 1.671 | 2.713 ( -1.173 ) | 1.132 ( 0.164 ) | 1.231 ( -0.251 ) | 1.355 ( 0.118 ) | 1.401 ( 0.303 ) |
| M | 100 | 0.6 | 1.728 | 2.626 ( 0.346 ) | 1.150 ( 0.491 ) | 1.294 ( 0.090 ) | 1.486 ( 0.135 ) | 1.684 ( 0.183 ) |
| | | 0.9 | 1.624 | 2.376 ( -0.012 ) | 1.099 ( -0.035 ) | 1.210 ( 0.160 ) | 1.487 ( 0.395 ) | 1.628 ( 0.001 ) |
| | | 0.4 | 2.271 | 4.438 ( -1.394 ) | 1.200 ( 0.168 ) | 1.216 ( -0.055 ) | 1.256 ( 0.087 ) | 1.264 ( 0.275 ) |
| | 150 | 0.6 | 1.797 | 3.115 ( 0.092 ) | 1.082 ( 0.233 ) | 1.219 ( 0.115 ) | 1.316 ( 0.126 ) | 1.436 ( 0.160 ) |
| | | 0.9 | 2.588 | 5.217 ( -0.237 ) | 1.096 ( 0.064 ) | 1.171 ( 0.203 ) | 1.323 ( 0.404 ) | 1.396 ( 0.141 ) |
| | | 0.4 | 1.851 | 2.548 ( -0.593 ) | 1.276 ( 2.143 ) | 1.309 ( 0.314 ) | 1.644 ( 0.105 ) | 2.143 ( -0.038 ) |
| | 50 | 0.6 | 1.929 | 3.020 ( 0.708 ) | 1.170 ( 0.079 ) | 1.368 ( 0.023 ) | 1.551 ( 0.042 ) | 1.958 ( 0.474 ) |
| | | 0.9 | 1.899 | 2.820 ( 0.321 ) | 1.267 ( -0.044 ) | 1.368 ( 1.044 ) | 1.550 ( 0.028 ) | 2.049 ( 0.096 ) |
| | | 0.4 | 1.730 | 2.479 ( -0.522 ) | 1.183 ( 1.768 ) | 1.188 ( 0.230 ) | 1.372 ( 0.063 ) | 2.031 ( -0.041 ) |
| A | 100 | 0.6 | 1.775 | 2.797 ( 0.696 ) | 1.123 ( 0.066 ) | 1.238 ( 0.015 ) | 1.302 ( 0.036 ) | 1.857 ( 0.456 ) |
| | | 0.9 | 1.872 | 2.955 ( 0.330 ) | 1.224 ( -0.023 ) | 1.255 ( 0.834 ) | 1.311 ( 0.007 ) | 2.001 ( 0.083 ) |
| | | 0.4 | 1.692 | 2.617 ( -0.425 ) | 1.112 ( 1.637 ) | 1.103 ( 0.148 ) | 1.091 ( 0.042 ) | 1.955 ( -0.044 ) |
| | 150 | 0.6 | 1.753 | 2.920 ( 0.781 ) | 1.091 ( 0.059 ) | 1.140 ( 0.002 ) | 1.048 ( 0.025 ) | 1.804 ( 0.415 ) |
| | | 0.9 | 1.851 | 3.074 ( 0.368 ) | 1.187 ( -0.011 ) | 1.139 ( 0.722 ) | 1.042 ( -0.008 ) | 1.971 ( 0.079 ) |

**Tabla 3-4.**: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$ y el $n$ esperado de la muestra es fijo.

### 3.8.2. Resultados cuando la población está incompleta.

Las Tablas **3-5**, **3-6** y **3-7** muestran los resultados obtenidos en el estudio de simulación, empleando los métodos de imputación descritos en la Sección 3.5, para $K \geqslant P$, un porcentaje de missing del 10 % y diferentes valores para $\rho_x$. Dado los resultados de la simulación es posible notar que los métodos de imputación evaluados tienen una RMSE mayor que 1 cuando $\rho_x$ tiene valores bajos, donde $\beta_1$ es quien mayor aporte hace a la RMSE del modelo. Sin embargo, estos se reducen a medida que la correlación $\rho_x$ aumentan.

Cuando se comparan los resultados mostrados en las Tablas **3-1** y **3-3** con los mostrados en las Tablas **3-5** a la **3-7** vemos que las RMSE's aumentan. Esto se puede atribuir al efecto aleatorio introducido por los métodos de imputación utilizados para subsanar los valores perdidos.

Por otro lado, en las tablas se observa que el sesgo relativo promedio no está siendo influenciado por el porcentaje de faltantes o método de imputación. No obstante, cuando los valores de $\rho_x$ van creciendo los sesgos relativos promedios tienden a disminuir.

Un caso particular de los resultados se presenta en la Tabla **3-6** cuando $\rho_x \in (0{,}50, 0{,}75]$, $\rho_z = 0{,}6$, por lo cual se realiza un análisis gráfico (ver Figura **A-1**) de la sesgo relativo promedio para los 200 dominios con el 10 % de valores faltantes. El análisis muestra como el dominio $D135$ tiene valores atípico ($> 1000$), mientras los restantes dominios tienen sesgos relativos promedios pequeños (lineas negras en la base del gráfico) siendo esto una tendencias para todos los tamaños de muestras y las diferentes técnicas de imputación. Con estos resultados se recalculan las RMSE y los sesgos relativos promedios excluyendo el dominio $D135$, logrando así reducir el error cuadrático medio del modelo. Los valores para este escenario se muestran en el Tabla **A-2** del anexo.

Las Tablas **3-8**, **3-9** y **3-10** muestran los resultados obtenidos para los métodos de imputación en los casos cuando $K \geqslant P$, un porcentaje de missing del 30 % y diferentes valores para $\rho_x$, aquí podemos observar que la RMSE toma valores entre 1 y 2; el cual muestra leve reducción a medida que los valores de $\rho_x$ crecen.

En cuanto a los sesgos relativos promedios vemos que estos presentan valores bajos, sin embargo, cuando se tiene $\rho_z = 0{,}4$, $\rho_x \in (0{,}50, 0{,}75]$, el sesgo relativo promedio de $\beta_2$ presenta los valores más grandes dentro de los resultados registrados en la Tabla **3-9**, pero estos disminuyen a medida que $\rho_z$ aumenta. Seguidamente podemos ver que para $\rho_x \in (0{,}75, 1]$, $\rho_z = 0{,}4$ y una muestra de 50 dominios, el sesgo relativo promedio para $\beta_1$ fue alto.

Bajo un escenario donde la muestra fue de 150 dominios, un 30 % de datos faltantes, $\rho_z = 0{,}4$, $\rho_x \in (0{,}75, 1]$ y $P = 3$, la RMSE muestra un valor de 6,215 para el método de imputación PMM, siendo esté un valor atípico dentro de los valores obtenidos en la Tabla **3-9**. Dado

| Método | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ ($BiasR_1$) | $RMSE_2$ ($BiasR_2$) | $RMSE_3$ ($BiasR_3$) |
|---|---|---|---|---|---|---|
| | | 0.4 | 1.372 | 1.998 ( -0.310 ) | 0.824 ( 0.020 ) | 0.990 ( -0.078 ) |
| | 50 | 0.6 | 1.342 | 1.891 ( -0.048 ) | 0.970 ( -0.170 ) | 0.940 ( 0.088 ) |
| | | 0.9 | 1.514 | 2.294 ( -0.221 ) | 0.819 ( -0.290 ) | 0.974 ( -0.086 ) |
| | | 0.4 | 1.136 | 1.507 ( -0.275 ) | 0.809 ( 0.005 ) | 0.973 ( -0.095 ) |
| PMM | 100 | 0.6 | 1.162 | 1.515 ( -0.123 ) | 0.968 ( -0.186 ) | 0.906 ( 0.134 ) |
| | | 0.9 | 1.367 | 2.017 ( -0.319 ) | 0.810 ( -0.298 ) | 0.938 ( -0.080 ) |
| | | 0.4 | 1.033 | 1.291 ( -0.267 ) | 0.801 ( 0.005 ) | 0.946 ( -0.084 ) |
| | 150 | 0.6 | 1.136 | 1.470 ( -0.187 ) | 0.953 ( -0.183 ) | 0.897 ( 0.171 ) |
| | | 0.9 | 1.346 | 1.984 ( -0.351 ) | 0.805 ( -0.295 ) | 0.921 ( -0.078 ) |
| | | 0.4 | 1.396 | 2.045 ( -0.237 ) | 0.822 ( 0.026 ) | 0.994 ( -0.075 ) |
| | 50 | 0.6 | 1.390 | 1.996 ( -0.014 ) | 0.966 ( -0.169 ) | 0.940 ( 0.112 ) |
| | | 0.9 | 1.534 | 2.325 ( -0.230 ) | 0.826 ( -0.281 ) | 0.985 ( -0.085 ) |
| | | 0.4 | 1.186 | 1.618 ( -0.168 ) | 0.811 ( -0.005 ) | 0.973 ( -0.090 ) |
| RLB | 100 | 0.6 | 1.154 | 1.506 ( -0.083 ) | 0.961 ( -0.188 ) | 0.899 ( 0.125 ) |
| | | 0.9 | 1.393 | 2.067 ( -0.288 ) | 0.807 ( -0.295 ) | 0.947 ( -0.082 ) |
| | | 0.4 | 1.085 | 1.409 ( -0.182 ) | 0.799 ( 0.008 ) | 0.953 ( -0.086 ) |
| | 150 | 0.6 | 1.141 | 1.482 ( -0.174 ) | 0.953 ( -0.184 ) | 0.893 ( 0.169 ) |
| | | 0.9 | 1.321 | 1.931 ( -0.364 ) | 0.809 ( -0.308 ) | 0.924 ( -0.078 ) |
| | | 0.4 | 1.433 | 2.121 ( -0.364 ) | 0.819 ( 0.042 ) | 0.997 ( -0.078 ) |
| | 50 | 0.6 | 1.372 | 1.967 ( -0.051 ) | 0.962 ( -0.170 ) | 0.924 ( 0.134 ) |
| | | 0.9 | 1.569 | 2.393 ( -0.243 ) | 0.827 ( -0.276 ) | 0.987 ( -0.084 ) |
| | | 0.4 | 1.170 | 1.589 ( -0.292 ) | 0.805 ( 0.019 ) | 0.967 ( -0.087 ) |
| RLBoot | 100 | 0.6 | 1.155 | 1.511 ( -0.129 ) | 0.965 ( -0.187 ) | 0.888 ( 0.138 ) |
| | | 0.9 | 1.397 | 2.079 ( -0.323 ) | 0.806 ( -0.291 ) | 0.942 ( -0.077 ) |
| | | 0.4 | 1.090 | 1.425 ( -0.337 ) | 0.800 ( 0.012 ) | 0.947 ( -0.080 ) |
| | 150 | 0.6 | 1.108 | 1.417 ( -0.215 ) | 0.950 ( -0.182 ) | 0.877 ( 0.137 ) |
| | | 0.9 | 1.335 | 1.962 ( -0.370 ) | 0.805 ( -0.296 ) | 0.920 ( -0.084 ) |

**Tabla 3-5**.: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, porcentaje de valores faltantes del 10 % y $\rho_x \in (0{,}25, 0{,}50]$

este valor, se opta por realizar un análisis descriptivo por dominio (ver Figura **3-2**). El análisis mostró, como el dominio $D151$ tiene una RMSE atípica ($> 80$), mientras los restantes dominios tienen una RMSE pequeño ($< 5$). Con estos resultados se recalculan las RMSE y los sesgos relativos promedios del modelo y de los parámetros $\beta$'s excluyendo el dominio $D151$. El resultado muestra que el modelo tiene una RMSE de 1,82 y de 1,64, 1,87 y 1,95 para los coeficientes $\beta_1$, $\beta_2$ y $\beta_3$ respectivamente.

En el anexo encontramos las Tablas **A-3**, **A-3** y **A-5** donde se muestran los resultados obtenidos cuando $K < P$, porcentaje de missing del 10 % y valores crecientes para $\rho_x$. En estas se observa que los métodos de imputación evaluados presentan varianzas altas y la

| Método | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ ($BiasR_1$) | $RMSE_2$ ($BiasR_2$) | $RMSE_3$ ($BiasR_3$) |
|---|---|---|---|---|---|---|
| | | 0.4 | 1.289 | 1.819 ( 0.310 ) | 0.922 ( -0.415 ) | 0.907 ( -0.115 ) |
| | 50 | 0.6 | 0.977 | 1.136 ( **7.236** ) | 0.915 ( 0.241 ) | 0.859 ( -0.059 ) |
| | | 0.9 | 1.200 | 1.550 ( 2.063 ) | 0.911 ( -0.106 ) | 1.041 ( -0.034 ) |
| | | 0.4 | 1.031 | 1.259 ( 0.120 ) | 0.895 ( -0.436 ) | 0.898 ( -0.132 ) |
| PMM | 100 | 0.6 | 0.881 | 1.002 ( **5.898** ) | 0.844 ( 0.239 ) | 0.781 ( -0.022 ) |
| | | 0.9 | 1.124 | 1.406 ( 2.092 ) | 0.884 ( -0.067 ) | 1.015 ( -0.043 ) |
| | | 0.4 | 0.959 | 1.102 ( 0.105 ) | 0.877 ( -0.457 ) | 0.881 ( -0.125 ) |
| | 150 | 0.6 | 0.773 | 0.848 ( **5.874** ) | 0.774 ( 0.193 ) | 0.689 ( 0.018 ) |
| | | 0.9 | 1.059 | 1.267 ( 2.041 ) | 0.874 ( -0.070 ) | 0.996 ( -0.042 ) |
| | | 0.4 | 1.161 | 1.526 ( 0.266 ) | 0.927 ( -0.472 ) | 0.923 ( -0.128 ) |
| | 50 | 0.6 | 1.023 | 1.238 ( **6.449** ) | 0.921 ( 0.267 ) | 0.871 ( -0.055 ) |
| | | 0.9 | 1.202 | 1.567 ( 2.116 ) | 0.904 ( -0.101 ) | 1.031 ( -0.031 ) |
| | | 0.4 | 1.049 | 1.274 ( 0.135 ) | 0.915 ( -0.475 ) | 0.916 ( -0.148 ) |
| RLB | 100 | 0.6 | 0.898 | 1.040 ( **5.849** ) | 0.850 ( 0.227 ) | 0.785 ( -0.015 ) |
| | | 0.9 | 1.126 | 1.418 ( 2.234 ) | 0.881 ( -0.075 ) | 1.007 ( -0.037 ) |
| | | 0.4 | 0.956 | 1.078 ( 0.107 ) | 0.888 ( -0.469 ) | 0.888 ( -0.136 ) |
| | 150 | 0.6 | 0.791 | 0.878 ( **5.869** ) | 0.782 ( 0.187 ) | 0.701 ( 0.044 ) |
| | | 0.9 | 1.030 | 1.207 ( 2.028 ) | 0.867 ( -0.064 ) | 0.988 ( -0.042 ) |
| | | 0.4 | 1.196 | 1.620 ( 0.228 ) | 0.913 ( -0.448 ) | 0.913 ( -0.115 ) |
| | 50 | 0.6 | 0.964 | 1.111 ( **7.273** ) | 0.909 ( 0.248 ) | 0.853 ( -0.048 ) |
| | | 0.9 | 1.190 | 1.526 ( 2.047 ) | 0.913 ( -0.132 ) | 1.041 ( -0.035 ) |
| | | 0.4 | 1.044 | 1.303 ( 0.066 ) | 0.887 ( -0.475 ) | 0.884 ( -0.121 ) |
| RLBoot | 100 | 0.6 | 0.885 | 1.006 ( **6.363** ) | 0.849 ( 0.215 ) | 0.785 ( -0.024 ) |
| | | 0.9 | 1.123 | 1.395 ( 2.171 ) | 0.894 ( -0.111 ) | 1.020 ( -0.048 ) |
| | | 0.4 | 0.955 | 1.102 ( 0.043 ) | 0.872 ( -0.489 ) | 0.873 ( -0.120 ) |
| | 150 | 0.6 | 0.773 | 0.871 ( **6.066** ) | 0.760 ( 0.179 ) | 0.674 ( 0.049 ) |
| | | 0.9 | 1.065 | 1.276 ( 2.104 ) | 0.879 ( -0.069 ) | 1.002 ( -0.047 ) |

**Tabla 3-6**.: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, porcentajes de valores faltantes del 10 % y $\rho_x \in (0{,}50, 0{,}75]$

mayor influencia a la variabilidad del modelo proviene de la estimación obtenida para el parámetro $\beta_1$. Sin embargo, el sesgo relativo promedio muestra una tendencia similar a las mostradas en las tablas anteriores. Como un futuro tema de investigación se propone estudiar estos escenarios aumentando el número de valores plausibles, implementar otros métodos de imputación, para así lograr reducir la varianza de las estimaciones por dominio.

En las Tablas **A-6**, **A-7** y **A-8** del anexo, están consolidados los resultados de la simulación cuando $K \geq P$, el porcentaje de missing es del 30 % y valores crecientes para $\rho_x$. Aquí, es posible notar que las estimaciones obtenidas para el $\beta_1$ son las que introducen la mayor variabilidad a la RMSE del modelo. Además, es posible notar las primeras diferencias entre

| Método | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ ($BiasR_1$) | $RMSE_2$ ($BiasR_2$) | $RMSE_3$ ($BiasR_3$) |
|---|---|---|---|---|---|---|
| | | 0.4 | 0.799 | 0.768 ( 0.061 ) | 0.760 ( 0.242 ) | 0.866 ( -0.272 ) |
| | 50 | 0.6 | 0.890 | 0.794 ( -0.040 ) | 0.933 ( -0.290 ) | 0.935 ( -0.169 ) |
| | | 0.9 | 0.781 | 0.771 ( 0.644 ) | 0.809 ( -0.003 ) | 0.761 ( -0.117 ) |
| | | 0.4 | 0.730 | 0.670 ( 0.149 ) | 0.717 ( 0.142 ) | 0.797 ( -0.277 ) |
| PMM | 100 | 0.6 | 0.822 | 0.735 ( 0.112 ) | 0.863 ( -0.234 ) | 0.862 ( -0.157 ) |
| | | 0.9 | 0.672 | 0.645 ( 0.504 ) | 0.699 ( -0.053 ) | 0.670 ( -0.154 ) |
| | | 0.4 | 0.675 | 0.601 ( 0.329 ) | 0.682 ( 0.071 ) | 0.735 ( -0.296 ) |
| | 150 | 0.6 | 0.754 | 0.657 ( 0.277 ) | 0.803 ( -0.214 ) | 0.792 ( -0.161 ) |
| | | 0.9 | 0.565 | 0.523 ( 0.330 ) | 0.592 ( -0.094 ) | 0.578 ( -0.156 ) |
| | | 0.4 | 0.793 | 0.759 ( 0.086 ) | 0.763 ( 0.184 ) | 0.855 ( -0.267 ) |
| | 50 | 0.6 | 0.907 | 0.799 ( -0.044 ) | 0.959 ( -0.298 ) | 0.954 ( -0.180 ) |
| | | 0.9 | 0.781 | 0.771 ( 0.637 ) | 0.808 ( -0.001 ) | 0.762 ( -0.115 ) |
| | | 0.4 | 0.740 | 0.677 ( 0.208 ) | 0.733 ( 0.092 ) | 0.803 ( -0.279 ) |
| RLBoot | 100 | 0.6 | 0.814 | 0.724 ( 0.122 ) | 0.855 ( -0.236 ) | 0.856 ( -0.173 ) |
| | | 0.9 | 0.673 | 0.646 ( 0.510 ) | 0.701 ( -0.054 ) | 0.671 ( -0.151 ) |
| | | 0.4 | 0.678 | 0.610 ( 0.394 ) | 0.694 ( 0.018 ) | 0.724 ( -0.286 ) |
| | 150 | 0.6 | 0.772 | 0.664 ( 0.258 ) | 0.826 ( -0.223 ) | 0.817 ( -0.184 ) |
| | | 0.9 | 0.566 | 0.526 ( 0.339 ) | 0.594 ( -0.094 ) | 0.577 ( -0.157 ) |
| | | 0.4 | 0.812 | 0.778 ( 0.125 ) | 0.776 ( 0.196 ) | 0.878 ( -0.274 ) |
| | 50 | 0.6 | 0.906 | 0.789 ( -0.091 ) | 0.959 ( -0.298 ) | 0.959 ( -0.179 ) |
| | | 0.9 | 0.782 | 0.770 ( 0.638 ) | 0.811 ( -0.002 ) | 0.765 ( -0.124 ) |
| | | 0.4 | 0.773 | 0.714 ( 0.193 ) | 0.766 ( 0.092 ) | 0.833 ( -0.277 ) |
| RLB | 100 | 0.6 | 0.826 | 0.730 ( 0.081 ) | 0.869 ( -0.241 ) | 0.870 ( -0.174 ) |
| | | 0.9 | 0.677 | 0.649 ( 0.499 ) | 0.706 ( -0.055 ) | 0.675 ( -0.156 ) |
| | | 0.4 | 0.693 | 0.618 ( 0.339 ) | 0.704 ( 0.043 ) | 0.751 ( -0.279 ) |
| | 150 | 0.6 | 0.840 | 0.749 ( 0.224 ) | 0.887 ( -0.220 ) | 0.876 ( -0.175 ) |
| | | 0.9 | 0.572 | 0.528 ( 0.324 ) | 0.602 ( -0.098 ) | 0.583 ( -0.163 ) |

**Tabla 3-7**.: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, porcentaje de valores faltantes del 10 % y $\rho_x \in (0{,}75, 1]$

los métodos de imputación, por ejemplo, cuando $\rho_x \in (0{,}25, 0{,}50]$ y $\rho_z$ aumenta, los métodos de imputación RLBoot y PMM muestran un crecimiento de la RMSE, mientras que RLB presenta una tendencia a disminuir a medida que $\rho_z$ crece. También cabe mencionar que el método PMM es quien muestra menor RMSE cuando se tienen muestras de 50 dominios.

En la Tabla **A-7** del anexo vemos que el aumentar el número de dominios en la muestra disminuye la RSME. En particular, cuando el número de dominios en la muestras es de 50, $\rho_z = 0{,}9$ la RMSE disminuye sustancialmente y los sesgos relativos promedios son pequeños a medida que $\rho_z$ crece.

En la Tabla **A-8** del anexo vemos que el método PMM tiene buenos resultados con muestras

| Método | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ ($BiasR_1$) | $RMSE_2$ ($BiasR_2$) | $RMSE_3$ ($BiasR_3$) |
|---|---|---|---|---|---|---|
| | | 0.4 | 1.723 | 2.453 ( 0.726 ) | 1.137 ( -0.417 ) | 1.266 ( 0.117 ) |
| | 50 | 0.6 | 1.822 | 2.704 ( 1.811 ) | 1.024 ( -0.385 ) | 1.267 ( -0.380 ) |
| | | 0.9 | 1.882 | 2.739 ( 0.849 ) | 1.185 ( 0.153 ) | 1.312 ( -0.344 ) |
| | | 0.4 | 1.581 | 2.128 ( 0.671 ) | 1.137 ( -0.434 ) | 1.297 ( 0.074 ) |
| PMM | 100 | 0.6 | 1.573 | 2.175 ( 1.765 ) | 1.026 ( -0.396 ) | 1.280 ( -0.397 ) |
| | | 0.9 | 1.784 | 2.520 ( 0.808 ) | 1.196 ( 0.132 ) | 1.329 ( -0.363 ) |
| | | 0.4 | 1.504 | 1.959 ( 0.607 ) | 1.131 ( -0.431 ) | 1.292 ( 0.079 ) |
| | 150 | 0.6 | 1.454 | 1.917 ( 1.713 ) | 1.026 ( -0.397 ) | 1.269 ( -0.393 ) |
| | | 0.9 | 1.748 | 2.451 ( 0.828 ) | 1.194 ( 0.107 ) | 1.317 ( -0.362 ) |
| | | 0.4 | 1.721 | 2.457 ( 0.758 ) | 1.136 ( -0.411 ) | 1.247 ( 0.134 ) |
| | 50 | 0.6 | 1.799 | 2.661 ( 1.765 ) | 1.031 ( -0.389 ) | 1.252 ( -0.371 ) |
| | | 0.9 | 1.883 | 2.753 ( 0.860 ) | 1.174 ( 0.149 ) | 1.294 ( -0.337 ) |
| | | 0.4 | 1.583 | 2.144 ( 0.710 ) | 1.139 ( -0.438 ) | 1.274 ( 0.089 ) |
| RLB | 100 | 0.6 | 1.530 | 2.091 ( 1.660 ) | 1.026 ( -0.398 ) | 1.264 ( -0.391 ) |
| | | 0.9 | 1.741 | 2.438 ( 0.781 ) | 1.184 ( 0.154 ) | 1.321 ( -0.359 ) |
| | | 0.4 | 1.497 | 1.965 ( 0.610 ) | 1.131 ( -0.431 ) | 1.260 ( 0.107 ) |
| | 150 | 0.6 | 1.430 | 1.880 ( 1.584 ) | 1.026 ( -0.399 ) | 1.244 ( -0.384 ) |
| | | 0.9 | 1.692 | 2.348 ( 0.773 ) | 1.178 ( 0.130 ) | 1.301 ( -0.354 ) |
| | | 0.4 | 1.709 | 2.421 ( 0.735 ) | 1.140 ( -0.429 ) | 1.264 ( 0.119 ) |
| | 50 | 0.6 | 1.867 | 2.795 ( 1.818 ) | 1.039 ( -0.397 ) | 1.253 ( -0.372 ) |
| | | 0.9 | 1.924 | 2.843 ( 0.858 ) | 1.176 ( 0.149 ) | 1.279 ( -0.329 ) |
| | | 0.4 | 1.561 | 2.084 ( 0.687 ) | 1.143 ( -0.451 ) | 1.287 ( 0.084 ) |
| RLBoot | 100 | 0.6 | 1.560 | 2.157 ( 1.754 ) | 1.031 ( -0.401 ) | 1.258 ( -0.387 ) |
| | | 0.9 | 1.777 | 2.537 ( 0.824 ) | 1.178 ( 0.156 ) | 1.285 ( -0.344 ) |
| | | 0.4 | 1.488 | 1.923 ( 0.584 ) | 1.139 ( -0.448 ) | 1.284 ( 0.090 ) |
| | 150 | 0.6 | 1.526 | 2.079 ( 1.829 ) | 1.037 ( -0.406 ) | 1.259 ( -0.390 ) |
| | | 0.9 | 1.706 | 2.397 ( 0.800 ) | 1.178 ( 0.129 ) | 1.265 ( -0.339 ) |

**Tabla 3-8**.: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, porcentaje de valores faltantes del 30 % y $\rho_x \in (0{,}25, 0{,}50]$

pequeñas, dado que la varianza del modelo y el sesgo relativo tiende a disminuir a medida que $\rho_z$ aumenta. Los métodos RLB y RLBoot muestran sesgos relativos pequeños cuando el $\rho_z = 0{,}9$ y al aumentar el número de dominios en la muestra RMSE disminuye en forma progresiva.

En general, del estudio de simulación podemos concluir que la metodología propuesta para la predicción de los parámetros $\beta'$s de un modelo de regresión lineal tiene un buen comportamiento cuando la matriz de datos observada no tiene información faltante, puesto que la RMSE y los sesgo relativo promedio fueron pequeños bajo los escenarios estudiados (Ver Tablas **3-1**, **3-2**, **3-3** y **3-4**).

| Método | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ ($BiasR_1$) | $RMSE_2$ ($BiasR_2$) | $RMSE_3$ ($BiasR_3$) |
|---|---|---|---|---|---|---|
| | | 0.4 | 1.811 | 2.497 ( 0.307 ) | 1.248 ( **4.707** ) | 1.430 ( -0.365 ) |
| | 50 | 0.6 | 1.529 | 1.837 ( -1.057 ) | 1.406 ( -0.416 ) | 1.289 ( -0.685 ) |
| | | 0.9 | 1.428 | 1.753 ( 1.353 ) | 1.290 ( -0.327 ) | 1.177 ( -0.599 ) |
| | | 0.4 | 1.607 | 2.043 ( 0.602 ) | 1.242 ( **4.407** ) | 1.425 ( -0.375 ) |
| PMM | 100 | 0.6 | 1.400 | 1.515 ( -0.981 ) | 1.361 ( -0.414 ) | 1.317 ( -0.749 ) |
| | | 0.9 | 1.330 | 1.491 ( 0.334 ) | 1.304 ( -0.353 ) | 1.176 ( -0.605 ) |
| | | 0.4 | 1.527 | 1.852 ( 0.576 ) | 1.242 ( **4.400** ) | 1.421 ( -0.379 ) |
| | 150 | 0.6 | 1.334 | 1.314 ( -1.026 ) | 1.383 ( -0.428 ) | 1.303 ( -0.756 ) |
| | | 0.9 | 1.251 | 1.263 ( 0.275 ) | 1.319 ( -0.375 ) | 1.166 ( -0.630 ) |
| | | 0.4 | 1.796 | 2.448 ( 0.266 ) | 1.260 ( **4.573** ) | 1.446 ( -0.367 ) |
| | 50 | 0.6 | 1.547 | 1.873 ( -1.025 ) | 1.406 ( -0.422 ) | 1.300 ( -0.720 ) |
| | | 0.9 | 1.405 | 1.712 ( 1.456 ) | 1.284 ( -0.321 ) | 1.159 ( -0.584 ) |
| | | 0.4 | 1.632 | 2.091 ( 0.567 ) | 1.250 ( **4.228** ) | 1.435 ( -0.381 ) |
| RLB | 100 | 0.6 | 1.358 | 1.413 ( -0.818 ) | 1.379 ( -0.423 ) | 1.278 ( -0.738 ) |
| | | 0.9 | 1.326 | 1.478 ( 0.605 ) | 1.311 ( -0.355 ) | 1.171 ( -0.609 ) |
| | | 0.4 | 1.532 | 1.851 ( 0.550 ) | 1.254 ( **4.142** ) | 1.429 ( -0.382 ) |
| | 150 | 0.6 | 1.330 | 1.292 ( -0.914 ) | 1.399 ( -0.436 ) | 1.297 ( -0.783 ) |
| | | 0.9 | 1.251 | 1.228 ( 0.134 ) | 1.341 ( -0.363 ) | 1.178 ( -0.627 ) |
| | | 0.4 | 1.802 | 2.468 ( 0.318 ) | 1.257 ( **4.420** ) | 1.439 ( -0.370 ) |
| | 50 | 0.6 | 1.534 | 1.819 ( -0.984 ) | 1.424 ( -0.428 ) | 1.313 ( -0.737 ) |
| | | 0.9 | 1.460 | 1.793 ( 1.187 ) | 1.323 ( -0.333 ) | 1.197 ( -0.606 ) |
| | | 0.4 | 1.572 | 1.948 ( 0.465 ) | 1.246 ( **4.218** ) | 1.437 ( -0.381 ) |
| RLBoot | 100 | 0.6 | 1.321 | 1.347 ( -0.890 ) | 1.356 ( -0.413 ) | 1.258 ( -0.712 ) |
| | | 0.9 | 1.350 | 1.472 ( 0.297 ) | 1.355 ( -0.359 ) | 1.212 ( -0.627 ) |
| | | 0.4 | 1.507 | 1.783 ( 0.498 ) | 1.255 ( **4.091** ) | 1.436 ( -0.385 ) |
| | 150 | 0.6 | 1.305 | 1.232 ( -0.951 ) | 1.388 ( -0.430 ) | 1.289 ( -0.795 ) |
| | | 0.9 | 1.263 | 1.238 ( -0.014 ) | 1.353 ( -0.369 ) | 1.192 ( -0.625 ) |

**Tabla 3-9**.: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, porcentaje de valores faltantes del 30 % y $\rho_x \in (0{,}50, 0{,}75]$

En el caso cuando la matriz de observaciones tenga datos faltantes, se debe tener cuidado en la elección del método de imputación, puesto que una mala elección de esté puede inflar el RMSE. Además, es recomendable tener más variables explicativas que estén altamente correlacionadas con el parámetro de interés $\boldsymbol{\beta}$ para que los sesgos relativos promedios sean pequeños.

| Método | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ ($BiasR_1$) | $RMSE_2$ ($BiasR_2$) | $RMSE_3$ ($BiasR_3$) |
|---|---|---|---|---|---|---|
| | | 0.4 | 1.556 | 1.380 ( **-4.017** ) | 1.705 ( -0.585 ) | 1.566 ( -0.591 ) |
| | 50 | 0.6 | 1.011 | 1.015 ( -0.127 ) | 0.997 ( -0.530 ) | 1.021 ( -2.290 ) |
| | | 0.9 | 1.432 | 1.292 ( 1.938 ) | 1.560 ( -0.464 ) | 1.433 ( -0.538 ) |
| | | 0.4 | 1.497 | 1.296 ( -2.772 ) | 1.660 ( -0.531 ) | 1.511 ( -0.561 ) |
| PMM | 100 | 0.6 | 0.938 | 0.900 ( -0.127 ) | 0.951 ( -0.457 ) | 0.963 ( -1.442 ) |
| | | 0.9 | 1.380 | 1.239 ( 1.676 ) | 1.508 ( -0.468 ) | 1.381 ( -0.529 ) |
| | | 0.4 | **6.215** | **6.152** ( -2.229 ) | **6.162** ( -0.480 ) | **6.329** ( -0.590 ) |
| | 150 | 0.6 | 0.874 | 0.841 ( -0.111 ) | 0.889 ( -0.351 ) | 0.891 ( 0.493 ) |
| | | 0.9 | 1.328 | 1.170 ( 1.457 ) | 1.466 ( -0.478 ) | 1.330 ( -0.533 ) |
| | | 0.4 | 1.591 | 1.406 ( **-3.994** ) | 1.747 ( -0.595 ) | 1.602 ( -0.603 ) |
| | 50 | 0.6 | 1.013 | 1.013 ( -0.136 ) | 0.997 ( -0.531 ) | 1.028 ( -2.313 ) |
| | | 0.9 | 1.424 | 1.290 ( 1.932 ) | 1.542 ( -0.459 ) | 1.430 ( -0.541 ) |
| | | 0.4 | 1.498 | 1.293 ( -2.782 ) | 1.663 ( -0.539 ) | 1.516 ( -0.561 ) |
| RLB | 100 | 0.6 | 0.941 | 0.904 ( -0.123 ) | 0.950 ( -0.458 ) | 0.968 ( -1.468 ) |
| | | 0.9 | 1.369 | 1.235 ( 1.661 ) | 1.489 ( -0.463 ) | 1.371 ( -0.529 ) |
| | | 0.4 | 1.495 | 1.300 ( -2.181 ) | 1.666 ( -0.502 ) | 1.496 ( -0.546 ) |
| | 150 | 0.6 | 0.875 | 0.836 ( -0.106 ) | 0.889 ( -0.354 ) | 0.899 ( 0.440 ) |
| | | 0.9 | 1.324 | 1.173 ( 1.449 ) | 1.455 ( -0.475 ) | 1.329 ( -0.535 ) |
| | | 0.4 | 1.560 | 1.374 ( **-3.888** ) | 1.716 ( -0.593 ) | 1.570 ( -0.592 ) |
| | 50 | 0.6 | 1.013 | 1.016 ( -0.132 ) | 0.997 ( -0.533 ) | 1.027 ( -2.300 ) |
| | | 0.9 | 1.423 | 1.284 ( 1.928 ) | 1.544 ( -0.455 ) | 1.428 ( -0.532 ) |
| | | 0.4 | 1.490 | 1.282 ( -2.691 ) | 1.657 ( -0.533 ) | 1.507 ( -0.559 ) |
| RLBoot | 100 | 0.6 | 0.942 | 0.904 ( -0.126 ) | 0.953 ( -0.456 ) | 0.968 ( -1.441 ) |
| | | 0.9 | 1.370 | 1.232 ( 1.669 ) | 1.492 ( -0.462 ) | 1.374 ( -0.525 ) |
| | | 0.4 | 1.467 | 1.264 ( -2.016 ) | 1.642 ( -0.497 ) | 1.472 ( -0.541 ) |
| | 150 | 0.6 | 0.878 | 0.843 ( -0.109 ) | 0.891 ( -0.350 ) | 0.900 ( 0.500 ) |
| | | 0.9 | 1.321 | 1.166 ( 1.458 ) | 1.455 ( -0.472 ) | 1.327 ( -0.529 ) |

**Tabla 3-10**.: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, porcentaje de valores faltantes del 30 % y $\rho_x \in (0{,}75, 1]$

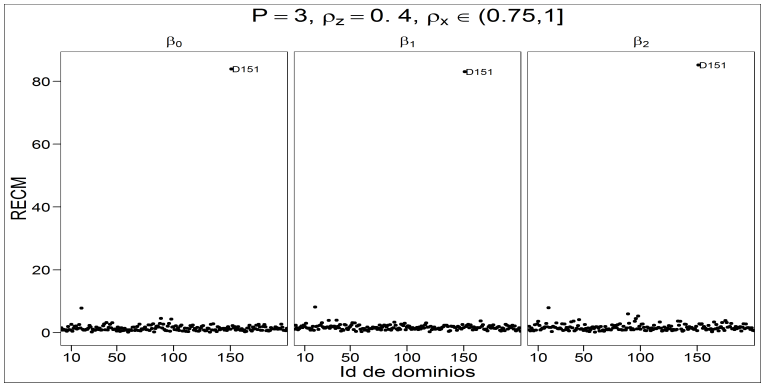


**Figura 3-2**.: Representación gráfica de la raíz del error cuadrático medio para una muestra de 150 dominios con el 30 % de datos faltantes.

# 4. Aplicación

La aplicación de la metodología propuesta en esta tesis se realiza para los datos suministrados por el Icfes, los cuales fueron recolectados para el estudio principal de factores asociados que se realizó en el 2012, en conjunto con la prueba Saber 3°, 5° y 9°, donde se hizo el levantamiento de abundante información sobre las variables de la escuela, los docentes y los estudiantes que se presumen pueden afectar el aprendizaje (Icfes 2014, 2016*a*). La información fue recolectada mediante la aplicación de una serie de cuestionarios; en el caso particular de este trabajo se realizó el estudio sobre la información diligenciada por los rectores y/o coordinadores.
En la base de datos, están registradas 2407 encuestas, las cuales estaban distribuidas en 562 establecimientos educativos, distribuidos por departamento como se muestra en la Tabla **4-1**. Hay que tener en cuenta que los establecimientos educativos pueden contar con más de una encuesta, dado que se realizaron encuestas por sede.

## 4.1. Cuestionario del rector o coordinador de sede

El cuestionario aplicado en las instituciones costaba de 156 preguntas de opción múltiple en escala nominal y ordinal, las cuales recolectaban información sobre el entrono social de la institución, el compromiso de los estudiantes, padres de familia y profesores con la educación. Además, las preguntas formuladas permiten construir un perfil de las personas encargadas de liderar la gestión administrativa de la institución. De allí que la encuesta estuviese dividida en tres apartados: *i)* La formación y experiencia del rector, en la cual se indaga por la edad, género, nivel de formación, estudios recientes, años de experiencia en el cargo, entre otras. *ii)* Características sociales de la sede. Las preguntas de este apartado indagan por la población estudiantil de la sede, las problemáticas sociales dentro y fuera la sede, infraestructura, recursos tecnológicos, entre otros aspectos. *iii)* Organización y gestión escolar. Acá se abordaron los temas de la deserción escolar, vinculación de los padres en la formación de los niños, innovación de los docentes dentro y fuera del aula de clases.

| Departamento | Colegios en la muestra | Pensamiento | Lenguaje | Matemáticas |
|---|---|---|---|---|
| Bogotá | 71 | 332 | 336 | 328 |
| Valle | 57 | 309 | 322 | 310 |
| Antioquia | 49 | 295 | 294 | 283 |
| Cundinamarca | 34 | 309 | 310 | 308 |
| Santander | 33 | 314 | 324 | 319 |
| Atlántico | 30 | 290 | 289 | 295 |
| Bolívar | 27 | 269 | 274 | 264 |
| Córdoba | 27 | 263 | 266 | 259 |
| Magdalena | 24 | 258 | 256 | 257 |
| Nariño | 20 | 306 | 301 | 302 |
| Cauca | 19 | 288 | 290 | 295 |
| Tolima | 19 | 286 | 287 | 292 |
| Norte de Santander | 18 | 293 | 300 | 300 |
| Boyacá | 15 | 318 | 313 | 316 |
| Huila | 15 | 282 | 290 | 284 |
| Sucre | 13 | 249 | 252 | 241 |
| Cesar | 12 | 275 | 280 | 278 |
| Caquetá | 11 | 268 | 262 | 268 |
| Meta | 11 | 287 | 293 | 291 |
| Caldas | 10 | 314 | 311 | 296 |
| Risaralda | 9 | 320 | 324 | 317 |
| La Guajira | 8 | 277 | 279 | 286 |
| Arauca | 7 | 279 | 287 | 282 |
| Casanare | 7 | 290 | 291 | 289 |
| Chocó | 7 | 259 | 255 | 245 |
| Putumayo | 3 | 253 | 255 | 256 |
| Quindío | 2 | 369 | 389 | 359 |
| Guainía | 1 | 208 | 192 | 189 |
| Guaviare | 1 | 237 | 237 | 252 |
| San Andrés y Providencia | 1 | 241 | 261 | 278 |
| Vaupés | 1 | 235 | 236 | 256 |

**Tabla 4-1**.: Distribución promedio de establecimientos educativos por departamento en la muestra

## 4.2. Procesamiento de los datos

Para realizar la construcción del índice, se utilizó un Análisis de Correspondencia Múltiple (MCA, por sus siglas en inglés), empleando el paquete `homals` creado por de Leeuw & Mair

(2009) y disponible en el software R. Este paquete realiza un análisis de homogeneidad que combina la idea de maximizar las correlaciones entre las variables de un conjunto de datos multivariados con la escala óptima que permite tener una representación numérica de las opciones de respuestas en cuestionario. Homals puede agrupar las variables de modo que el análisis de homogeneidad se extienda al análisis de correlación canónica no lineal o a modelos predictivos que emulen el análisis discriminante y los modelos de regresión. Para una mayor documentación del paquete puede consultar *Gifi Methods for Optimal Scaling in R: The Package homals,* donde presentan los aspectos metodológicos y prácticos del paquete y además muestran varias extensiones del mismo.

Una vez realizado el análisis de homogeneidad se hizo la imputación de los valores perdidos en la matriz numérica resultante. El método de imputación utilizado es PMM de la librería `mice` de *R*, con el cual se obtienen 5 valores plausibles para las variables. La tasa promedio de información perdida por variables era del 5 %.

Una vez realizado el proceso de imputación a la matriz de datos transformada se pasó a hacer un Análisis Factorial Exploratorio (EFA, por sus siglas en inglés) para cada matriz plausible resultante, con el propósito de identificar una estructura de factores subyacentes dentro del conjunto de ítems y de estos se seleccionó el factor que mostró la mayor correlación con los resultados en las pruebas Saber de 5° (ver Tabla **4-2**). Los ítems que conforman este factor subyacente se muestran en la Tabla **4-3**.

| Departamento | Lenguaje | Matemáticas | Pensamiento |
|---|---|---|---|
| Colombia | -0.470 | -0.399 | -0.459 |
| Antioquia | -0.552 | -0.462 | -0.533 |
| Atlántico | -0.573 | -0.639 | -0.495 |
| Bogotá | -0.439 | -0.298 | -0.475 |
| Bolívar | -0.624 | -0.608 | -0.629 |
| Boyacá | -0.529 | -0.437 | -0.675 |
| Cauca | -0.344 | -0.423 | -0.392 |
| Córdoba | -0.364 | -0.158 | -0.431 |
| Cundinamarca | -0.428 | -0.376 | -0.453 |
| Huila | 0.024 | 0.155 | -0.008 |
| Magdalena | -0.568 | -0.472 | -0.740 |
| Nariño | -0.666 | -0.642 | -0.609 |
| Norte de Santander | -0.305 | -0.264 | -0.445 |
| Santander | -0.371 | -0.177 | -0.058 |
| Tolima | -0.534 | -0.337 | -0.665 |
| Valle | -0.456 | -0.447 | -0.425 |

**Tabla 4-2**.: Correlación del índice con los resultados de las pruebas Saber 5°

| Contexto | Enunciado | Opción de respuesta |
|---|---|---|
| ¿Qué tanto las siguientes situaciones relacionadas con los estudiantes afectan esta sede y jornada escolar? | Eludir clases (horas/periodos) | No se presenta, Poco, Moderadamente, Mucho |
| | Deserción o retiro durante el transcurso del año escolar | |
| ¿Con qué frecuencia los estudiantes de esta sede y jornada abandonan sus estudios por alguna de las siguientes razones? | Falta de recursos económicos de las familias | Nunca, Casi Nunca, Algunas Veces, Siempre o casi siempre |
| | Enfermedad | |
| | Extraedad[1] | |
| | Desinteres y falta de motivación con el estudio | |
| | Dificultades académicas de los estudiantes (pérdida de cursos o logros, dificultades para aprender, etc.) | |
| | Los estudiantes necesitan o quieren trabajar | |

**Tabla 4-3**.: Ítems para el factor subyacente

## 4.3. Modelo de regresión en los departamentos

Para iniciar con el procesamiento de los datos, debemos recordar que el estudio de factores asociados se encuentra enmarcado dentro de la muestra controlada para la aplicación de la prueba Saber 3°, 5° y 9° del 2012, por tal motivo se excluirán del presente análisis los colegios que no tienen reportado resultados para la prueba de Lenguaje en 5° (los resultados de la prueba Saber no se reportan cuando las colegios y/o sedes tienen menos de 6 estudiantes que presentan el examen).

Siguiendo con las restricciones, se ha considerado que para estimar el modelo de regresión (ver las ecuaciones 1-6, 1-7 y 1-8) los departamentos deben tener una muestra mayor o igual a 15 colegios con resultados en la prueba Saber. Bajo esta restricción vemos que el numero de departamentos en la muestra es 15 (ver Tabla **4-1** )

Los resultados consolidados en la Tabla **4-2** muestran que el índice (factor) seleccionado está relacionado inversamente con el rendimiento académico de los estudiantes, con un comportamiento diferencial entre los departamentos, siendo el departamento de Nariño el que tiene la mayor correlación promedio. Note que a nivel nacional las correlaciones más altas se presentan con el área de Lenguaje, por lo cual se toma está como variable respuesta

[1] Estudiantes con edades muy por encima del promedio del grado.

($\boldsymbol{Y}_d$) en el modelo de regresión lineal por departamento y el índice como la variable respuesta ($\boldsymbol{X}_d$). Así la ecuación del modelo para el $d-$ésimo departamento estada dado por:

$$y_{sid} \quad = \quad \beta_{d0} + \beta_{d1} x_{di} = \boldsymbol{X}_{sid}^t \boldsymbol{\beta}_d \tag{4-1}$$

Dado que se cuentan con 5 valores plausibles, es necesario estimar el modelo dado en la ecuación 4-1 este mismo número de veces, para ellos se utiliza la metodología presentada en la Sección 1.3.2 la cual está implementada en la librería *survey* de Lumley (2004). Seguidamente, se emplea la metodología dada en la Sección 1.5.2, para estimar $\hat{\boldsymbol{\beta}}_{IM} = \left(\hat{\boldsymbol{\beta}}_{1,IM}, \ldots, \hat{\boldsymbol{\beta}}_{d,IM}, \ldots, \hat{\boldsymbol{\beta}}_{D_0,IM}\right)$ y $\Sigma_d = \widehat{Var_d}\left(\hat{\boldsymbol{\beta}}_{d,IM}\right)$ donde $\hat{\boldsymbol{\beta}}_{d,IM}$ y $\widehat{Var_d}\left(\hat{\boldsymbol{\beta}}_{d,IM}\right)$ están dados por las ecuaciones 1-21 y 1-22 respectivamente.

## 4.4. Predicción de los coeficientes del modelo lineal empleando el modelo Fay-Herriot.

En la presente sección se muestran los resultados obtenidos con el modelo Fay-Herriot dado por

$$\hat{\boldsymbol{\beta}}_{IM} \quad = \quad \boldsymbol{Z}^t \boldsymbol{\Gamma} + \boldsymbol{\delta} + \boldsymbol{e} \tag{4-2}$$

para predecir $\boldsymbol{\beta}$ en los $D - D_0$ departamentos no son observados, por tal razón se debe encontrar el modelo que minimice la RMSE.

### 4.4.1. Información auxiliar para el modelo Fay-Herriot

Cuando se inició el presente trabajo, algunos de los interrogantes que causaban mayor incertidumbre, eran: *¿Qué información auxiliar utilizar en el modelo de Fay-Harriot?, ¿Dónde encontrar la información auxiliar?, ¿El censo 2005 tendrá información idónea para la estimación?.* Con estos interrogantes en mente se inició el proceso de recopilar la información censal disponible a nivel departamental en el Icfes y en el Departamento Administrativo Nacional de Estadística (DANE), pudiendo consolidar una base de 30 variables regresoras.

| Departamento | $\hat{\boldsymbol{\beta}}_{IM}$ | Estimado | $\hat{\boldsymbol{\Sigma}}_d = \widehat{Var}_d\left(\hat{\boldsymbol{\beta}}_d^{(IM)}\right)$ |
|---|---|---|---|
| Bogotá | $\beta_0^{(IM)}$ | 404.138 | $\begin{bmatrix} 2340{,}754 & -26{,}534 \\ -26{,}534 & 0{,}480 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -1.648 | |
| Valle | $\beta_0^{(IM)}$ | 355.154 | $\begin{bmatrix} 21978{,}856 & -31{,}732 \\ -31{,}732 & 0{,}720 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -0.966 | |
| Antioquia | $\beta_0^{(IM)}$ | 436.606 | $\begin{bmatrix} 27415{,}008 & -60 \\ -60 & 1{,}224 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -2.740 | |
| Cundinamarca | $\beta_0^{(IM)}$ | 372.596 | $\begin{bmatrix} 10758{,}762 & -33{,}416 \\ -33{,}416 & 0{,}756 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -1.468 | |
| Santander | $\beta_0^{(IM)}$ | 419.276 | $\begin{bmatrix} 12771{,}968 & -41{,}450 \\ -41{,}450 & 0{,}852 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -1.992 | |
| Atlántico | $\beta_0^{(IM)}$ | 549.844 | $\begin{bmatrix} 113061{,}546 & -86{,}416 \\ -86{,}416 & 2{,}134 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -4.920 | |
| Bolívar | $\beta_0^{(IM)}$ | 446.792 | $\begin{bmatrix} 269152{,}632 & -152{,}348 \\ -152{,}348 & 4{,}542 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -3.560 | |
| Córdoba | $\beta_0^{(IM)}$ | 360.302 | $\begin{bmatrix} 57339{,}090 & -55{,}446 \\ -55{,}446 & 1{,}354 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -1.910 | |
| Magdalena | $\beta_0^{(IM)}$ | 422.560 | $\begin{bmatrix} 179369{,}410 & 27{,}582 \\ 27{,}582 & 1{,}496 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -3.090 | |
| Nariño | $\beta_0^{(IM)}$ | 573.942 | $\begin{bmatrix} 354766{,}658 & -191{,}734 \\ -191{,}734 & 5{,}822 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -5.172 | |
| Cauca | $\beta_0^{(IM)}$ | 344.818 | $\begin{bmatrix} 12970{,}372 & -75 \\ -75 & 1{,}608 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -1.136 | |
| Tolima | $\beta_0^{(IM)}$ | 386.422 | $\begin{bmatrix} 55741{,}842 & -86{,}306 \\ -86{,}306 & 1{,}966 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -1.866 | |
| Norte de Santander | $\beta_0^{(IM)}$ | 310.766 | $\begin{bmatrix} 35430{,}010 & -58{,}972 \\ -58{,}972 & 1{,}392 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -0.242 | |
| Boyacá | $\beta_0^{(IM)}$ | 407.636 | $\begin{bmatrix} 83303{,}024 & -148{,}992 \\ -148{,}992 & 4{,}208 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -2.204 | |
| Huila | $\beta_0^{(IM)}$ | 328.008 | $\begin{bmatrix} 29084{,}504 & -152{,}966 \\ -152{,}966 & 3{,}264 \end{bmatrix}$ |
| | $\beta_1^{(IM)}$ | -0.636 | |

**Tabla 4-4**.: Estimación de $\hat{\boldsymbol{\beta}}^{(IM)}$ y $\widehat{Var}\left(\hat{\boldsymbol{\beta}}^{(IM)}\right)$ por departamento

Entre las variables obtenidas del censo tenemos, el total de la población del departamento por rango de edad, actividad económica, condición de ocupación, tipo de emplazamiento del establecimiento y Necesidades Básicas Insatisfechas - NBI[2]. En el caso del Icfes se obtuvieron los resultados de las pruebas Saber 11 del 2014 al 2016, resultados del Índice de Clasificación de Planteles del 2014-2, 2015-1 y 2015-2, número de establecimientos educativos en el 2012

[2] `http://www.dane.gov.co/files/censos/resultados/NBI_total_30_Jun_2011.pdf`

y el índice socio-económico de 2014.

### 4.4.2. Selección del modelo Fay-Herriot.

Dada la naturaleza de la variable respuesta del modelo en la ecuación 4-2, fue necesario ajustar todos los modelos posibles que resultan al combinar las 30 variables disponibles en grupos de 2, 3 y 4 variables. De todos los posibles modelos resultantes es seleccionado el que mostró la menor RMSE con un valor de 49,862.

#### 4.4.2.1. Modelo para predecir $\beta$.

El modelo de Fay-Herriot multivariado ajustado está formado por las variables: Número de establecimientos educativos en el 2012 por departamento (N), resultado promedio en inglés y razonamiento cuantitativo (RC) en las pruebas Saber 11 del segundo semestre de 2014. En la Tabla **4-5** vemos los resultados obtenidos al estimar los parámetros del modelo, donde se observa que $\gamma_{11}$ y $\gamma_{21}$ no resultaron significativas, luego el modelo se puede escribir como:

$$\begin{aligned}\hat{\beta}_{0,EBLUP} &= 0{,}017N + 3{,}805Ingles + 3{,}840RC \\ \hat{\beta}_{1,EBLUP} &= -0{,}000093194N - 0{,}370Ingles + 0{,}335RC\end{aligned}$$

En la Figura **4-1** podemos comparar las lineas de regresión obtenidas a partir del estimador directo (Ver Sección 1.3.2) y la obtenida con el modelo de Fay-Herriot (Ver Sección 2.2, caso 4) en está se observa que los departamentos con la mayor diferencia entre los resultados son Atlántico y Nariño, mientras que para los departamentos Cauca, Córdoba, Cundinamarca, Santander, Tolima y Valle las estimaciones obtenidas presentan poca diferencia entre el estimador directo y el modelo SAE.
En la Tabla **4-6** se muestras las estimaciones puntuales obtenidas para los departamentos observados mediante la ecuación 1-10 y las predicciones realizadas para los no observados, de donde podemos concluir que la metodología propuesta puede ser utilizada pronosticar los coeficientes de un modelo de regresión en dominios que no pudieron ser observados por alguna razón. Además, haciendo uso de los valores plausibles fue posible obtener la estimación de los parámetros de un modelo de regresión lineal en cada dominio.

| Variable | $\hat{\mathbf{\Gamma}}_{EBLUP}$ | Estimado | Error estándar | $t-$valor | $p-$valor |
|---|---|---|---|---|---|
| N | $\hat{\gamma}_{11}$ | 0.017 | 0.045 | 0.379 | 0.704 |
| Inglés | $\hat{\gamma}_{12}$ | 3.805 | 0.686 | 5.548 | 0.000 |
| RC | $\hat{\gamma}_{13}$ | 3.840 | 0.692 | 5.552 | 0.000 |
| N | $\hat{\gamma}_{21}$ | $-9.3194\times10^{-5}$ | 0.000 | -0.284 | 0.776 |
| Inglés | $\hat{\gamma}_{22}$ | -0.370 | 0.344 | -1.076 | 0.282 |
| RC | $\hat{\gamma}_{23}$ | 0.335 | 0.341 | 0.982 | 0.326 |

**Tabla 4-5**.: Parámetros estimados para el modelo de Fay-Herriot

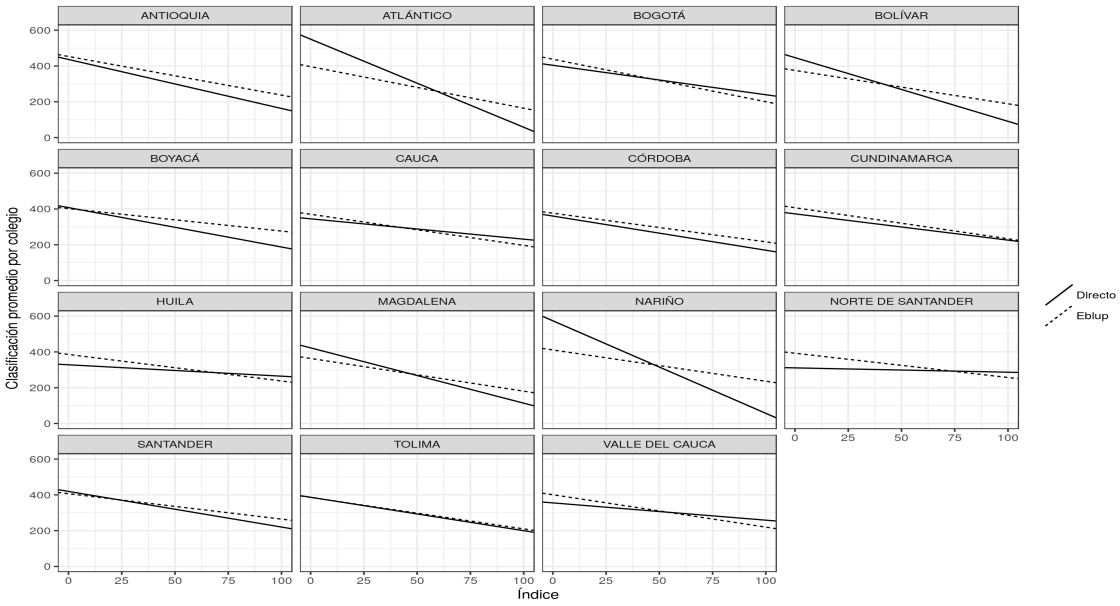


**Figura 4-1**.: Lineas de regresión estimadas con el estimador directo y el modelo multivariado de Fay-Herriot

| Dominios | Departamento | $\hat{\beta}_{0,EBLUP}$ | $\hat{\beta}_{1,EBLUP}$ | N | Inglés | Razonamiento cuantitativo |
|---|---|---|---|---|---|---|
| Observados | Antioquia | 453.29 | -2.16 | 4246 | 49.86 | 49.80 |
| | Atlántico | 396.46 | -2.32 | 913 | 50.52 | 49.12 |
| | Bogotá | 437.85 | -2.37 | 1694 | 53.99 | 53.01 |
| | Bolívar | 375.35 | -1.86 | 664 | 47.81 | 47.42 |
| | Boyacá | 401.88 | -1.25 | 511 | 50.58 | 52.26 |
| | Cauca | 370.36 | -1.74 | 765 | 46.78 | 46.69 |
| | Córdoba | 376.31 | -1.60 | 611 | 47.67 | 48.04 |
| | Cundinamarca | 406.96 | -1.73 | 1054 | 50.67 | 51.08 |
| | Huila | 385.81 | -1.48 | 391 | 49.18 | 50.00 |
| | Magdalena | 363.48 | -1.83 | 535 | 46.57 | 46.13 |
| | Nariño | 410.85 | -1.75 | 1938 | 49.19 | 49.65 |
| | Norte de Santander | 392.76 | -1.36 | 494 | 49.63 | 50.90 |
| | Santander | 406.49 | -1.42 | 762 | 50.82 | 52.11 |
| | Tolima | 385.81 | -1.76 | 637 | 49.02 | 49.07 |
| | Valle | 400.70 | -1.81 | 1670 | 48.62 | 48.76 |
| No observados | Amazonas | 350.99 | -1.95 | 46 | 46.30 | 45.31 |
| | Arauca | 376.33 | -1.22 | 102 | 48.28 | 49.70 |
| | Caldas | 385.46 | -1.73 | 323 | 49.64 | 49.76 |
| | Caquetá | 368.96 | -1.34 | 213 | 47.28 | 48.28 |
| | Casanare | 380.80 | -1.15 | 122 | 48.68 | 50.38 |
| | Cesar | 376.46 | -1.60 | 422 | 48.11 | 48.49 |
| | Chocó | 345.39 | -1.91 | 192 | 45.21 | 44.29 |
| | Guainía | 360.73 | -1.90 | 61 | 47.39 | 46.70 |
| | Guaviare | 363.13 | -1.07 | 46 | 46.54 | 48.23 |
| | La Guajira | 361.19 | -1.63 | 258 | 46.63 | 46.70 |
| | Meta | 390.66 | -1.52 | 339 | 49.95 | 50.72 |
| | Putumayo | 371.59 | -1.34 | 193 | 47.65 | 48.69 |
| | Quindío | 385.18 | -1.80 | 181 | 50.02 | 49.93 |
| | Risaralda | 391.42 | -1.88 | 318 | 50.61 | 50.37 |
| | San Andrés y Providencia | 402.70 | -4.57 | 18 | 56.51 | 48.79 |
| | Sucre | 376.65 | -1.71 | 713 | 47.65 | 47.71 |
| | Vaupés | 352.29 | -2.39 | 32 | 47.11 | 44.91 |
| | Vichada | 361.68 | -1.50 | 151 | 46.75 | 47.19 |

**Tabla 4-6**.: Predicciones del parámetro $\boldsymbol{\beta}$ con el modelo de Fay-Herriot

# 5. Conclusiones y recomendaciones

En este trabajo se desarrolla una metodología que permite obtener estimadores aproximadamente insesgados, para predecir el parámetro $\boldsymbol{\beta}_d$ de un modelo de regresión lineal en muestras probabilísticas con dominios observados y no observados. Además, se hizo la extensión de la técnica a muestras con presencia de datos faltantes. La metodología propuesta mostró un RMSE y sesgos relativos pequeños en los diferentes escenarios de simulación donde fue puesta a prueba.

Los estudios de simulación muestran que la metodología propuesta tiene RMSE pequeños cuando la matriz de datos observada no tiene información faltante. Además, para valores bajos en $\rho_x$ el RMSE se reduce a medida que $\rho_z$ aumenta. Está tendencia se repite en los $P$ parámetros $\beta$'s predichos con el modelo. En general, la metodología tiene sesgos relativos bajos en todos los escenarios evaluados, y la RMSE tiene valores pequeños cuando el número de variables respuesta $P$ es menor que el número de variables auxiliares $K$. Está tendencia se repite independientemente de los valores que tomen $\rho_x$, $\rho_z$, ni el tamaño de la muestra por dominio. Sin embargo, cuando $K < P$ y $\rho_x$, $\rho_z$ presenten valores bajos, es necesario tener un mayor número de observaciones por dominio para disminuir el RMSE.

Medina & Galván (2007) asegura que, no existe el mejor método de imputación. Cada situación es diferente y la elección del procedimiento de sustitución de datos depende de la variable de estudio, del porcentaje de datos faltantes, del tipo de encuesta que se analice y del uso que se hará de la información imputada. Por ello, no se aconseja elegir un procedimiento de imputación y aplicarlo en forma generalizada para todas las variables en todas las encuestas. Por lo anterior, cuando la matriz $\boldsymbol{X}_s$ tenga datos faltantes, se debe tener cuidado en la elección del método de imputación, puesto que una mala elección de esté puede aumentar el RMSE. Además, es recomendable tener más variables explicativas con valores altos de $\rho_x$ y así poder tener pequeños sesgos relativos y promedios. Cabe mencionar que los sesgos relativos promedios no están influenciados por el porcentaje de faltantes o método de imputación cuando $K \geqslant P$. Pero a medida que los valores de $\rho_x$ crecen los sesgos relativos promedios tienden a disminuir.

En general, se puede afirmar que es posible utilizar el modelo multivariado de Fay-Herriot para estimar y pronosticar los coeficientes de un modelo de regresión lineal con y sin la

utilización de los valores plausibles para imputar la matriz de observaciones, dado que el RMSE y sesgo relativo promedio son pequeños en los diferentes escenarios de simulación evaluados.

En relación a la aplicación con datos reales, el Icfes suministró la información necesaria para poder implementar la metodología. Los resultados obtenidos muestran la facilidad con la cual esta metodología se puede implementar en situaciones reales, dando la posibilidad de obtener resultados en dominios no observados. Sin embargo, se debe mencionar lo dispendioso que representa encontrar información auxiliar que explique el comportamiento de una variable respuesta a la cual no es posible intuir la información auxiliar que la pueda modelar de forma adecuada. En esté sentido, es importante hacer notar que en Colombia la información Censal pública disponible es escasa, aparte de las bases del Censo de 2005.

Una propuesta para futuros trabajos, es construir una prueba de hipótesis sobre el parámetro $\beta$'s, para así determinar la significancia estadística de esté en el dominio no observado. También se recomienda realizar el estudio de simulación controlando la correlación entre $\beta$'s, y extender el estudio de simulación a otros métodos de imputación y a modelos lineales generalizados.

# A. Anexos

| Tamaño de la muestra | $N_d$ | $RMSE$ | $RMSE_1$ ($BiasR_1$) | $RMSE_2$ ($BiasR_2$) | $RMSE_3$ ($BiasR_3$) |
|---|---|---|---|---|---|
| | 50 | 0.776 | 0.762 ( 0.023 ) | 0.815 ( 0.546 ) | 0.748 ( 0.140 ) |
| con $n$ fijo | 100 | 0.673 | 0.674 ( 0.038 ) | 0.725 ( 0.494 ) | 0.616 ( 0.173 ) |
| | 150. | 0.574 | 0.599 ( 0.041 ) | 0.639 ( 0.439 ) | 0.469 ( 0.241 ) |
| | 50 | 0.783 | 0.771 ( 0.019 ) | 0.822 ( 0.545 ) | 0.755 ( 0.127 ) |
| 20 % por dominio | 100 | 0.675 | 0.680 ( 0.037 ) | 0.725 ( 0.500 ) | 0.614 ( 0.174 ) |
| | 150 | 0.568 | 0.585 ( 0.048 ) | 0.638 ( 0.456 ) | 0.468 ( 0.238 ) |

**Tabla A-1**.: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, $\rho_x \in (0{,}50, 0{,}75]$ y $\rho_z = 0{,}6$, en muestras de tamaño fijo y muestras del 20 % por dominio.

| Método | $N_d$ | $RMSE$ | $RMSE_1$ ($BiasR_1$) | $RMSE_2$ ($BiasR_2$) | $RMSE_3$ ($BiasR_3$) |
|---|---|---|---|---|---|
| | 50 | 0.978 | 1.136 ( 0.221 ) | 0.916 ( 0.227 ) | 0.861 ( -0.060 ) |
| PMM | 100 | 0.882 | 1.003 ( 0.195 ) | 0.846 ( 0.227 ) | 0.783 ( -0.022 ) |
| | 150 | 0.774 | 0.847 ( 0.222 ) | 0.775 ( 0.182 ) | 0.691 ( 0.020 ) |
| | 50 | 1.025 | 1.239 ( 0.183 ) | 0.923 ( 0.252 ) | 0.873 ( -0.057 ) |
| RLB | 100 | 0.899 | 1.040 ( 0.182 ) | 0.851 ( 0.215 ) | 0.787 ( -0.015 ) |
| | 150 | 0.791 | 0.878 ( 0.195 ) | 0.783 ( 0.175 ) | 0.703 ( 0.045 ) |
| | 50 | 0.965 | 1.111 ( 0.251 ) | 0.911 ( 0.233 ) | 0.855 ( -0.049 ) |
| RLBoot | 100 | 0.886 | 1.006 ( 0.230 ) | 0.850 ( 0.203 ) | 0.786 ( -0.023 ) |
| | 150 | 0.773 | 0.871 ( 0.248 ) | 0.761 ( 0.169 ) | 0.676 ( 0.051 ) |

**Tabla A-2**.: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 3$, $\rho_x \in (0{,}50, 0{,}75]$,$\rho_z = 0{,}6$, el 10 % de datos faltantes y muestras del 20 % por dominio.

,

| Método | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ ($BiasR_1$) | $RMSE_2$ ($BiasR_2$) | $RMSE_3$ ($BiasR_3$) | $RMSE_4$($BiasR_4$) | $RMSE_5$($BiasR_5$) |
|---|---|---|---|---|---|---|---|---|
| | | 0.4 | **8.358** | 18.375 ( 0.802 ) | 1.479 ( -2.302 ) | 1.464 ( -0.182 ) | 1.793 ( 0.314 ) | 2.029 ( -0.115 ) |
| | 50 | 0.6 | **43.977** | 98.265 ( -6.027 ) | 1.388 ( -0.024 ) | 1.714 ( -0.242 ) | 2.036 ( -0.132 ) | 2.200 ( -0.099 ) |
| | | 0.9 | **13.475** | 29.881 ( -1.393 ) | 1.615 ( -0.466 ) | 1.959 ( 0.032 ) | 2.057 ( -0.053 ) | 2.092 ( 0.096 ) |
| | | 0.4 | 5.144 | 11.023 ( 0.823 ) | 1.391 ( -2.650 ) | 1.400 ( -0.156 ) | 1.719 ( 0.350 ) | 1.985 ( -0.124 ) |
| PMM | 100 | 0.6 | 6.329 | 13.713 ( -0.522 ) | 1.249 ( 0.038 ) | 1.585 ( -0.234 ) | 1.951 ( -0.112 ) | 2.092 ( -0.084 ) |
| | | 0.9 | 9.283 | 20.456 ( -0.353 ) | 1.357 ( -0.429 ) | 1.675 ( 0.069 ) | 1.934 ( -0.043 ) | 2.011 ( 0.096 ) |
| | | 0.4 | 7.137 | 15.633 ( 2.044 ) | 1.350 ( -2.653 ) | 1.365 ( -0.144 ) | 1.685 ( 0.393 ) | 1.939 ( -0.108 ) |
| | 150 | 0.6 | 4.822 | 10.211 ( -0.853 ) | 1.225 ( 0.060 ) | 1.575 ( -0.222 ) | 1.932 ( -0.106 ) | 2.069 ( -0.077 ) |
| | | 0.9 | 6.890 | 15.009 ( -1.128 ) | 1.331 ( -0.418 ) | 1.657 ( 0.081 ) | 1.914 ( -0.041 ) | 1.985 ( 0.100 ) |
| | | 0.4 | **27.584** | 61.580 ( 8.453 ) | 1.631 ( -2.446 ) | 1.487 ( -0.146 ) | 1.816 ( 0.338 ) | 2.045 ( -0.122 ) |
| | 50 | 0.6 | **10.253** | 22.619 ( -0.780 ) | 1.376 ( -0.007 ) | 1.674 ( -0.253 ) | 2.025 ( -0.134 ) | 2.274 ( -0.104 ) |
| | | 0.9 | **21.268** | 47.411 ( -0.136 ) | 1.459 ( -0.474 ) | 1.780 ( 0.052 ) | 2.051 ( -0.046 ) | 2.090 ( 0.093 ) |
| | | 0.4 | 6.398 | 13.926 ( 2.677 ) | 1.399 ( -2.553 ) | 1.390 ( -0.156 ) | 1.708 ( 0.362 ) | 1.988 ( -0.129 ) |
| RLB | 100 | 0.6 | 19.684 | 43.830 ( 0.350 ) | 1.946 ( 0.119 ) | 1.800 ( -0.199 ) | 2.151 ( -0.103 ) | 2.130 ( -0.074 ) |
| | | 0.9 | 10.866 | 23.960 ( -1.697 ) | 1.395 ( -0.452 ) | 2.264 ( 0.027 ) | 2.153 ( -0.051 ) | 2.129 ( 0.083 ) |
| | | 0.4 | 4.444 | 9.408 ( 2.337 ) | 1.352 ( -2.622 ) | 1.360 ( -0.133 ) | 1.678 ( 0.389 ) | 1.937 ( -0.110 ) |
| | 150 | 0.6 | 4.093 | 8.481 ( -1.137 ) | 1.212 ( 0.091 ) | 1.572 ( -0.220 ) | 1.915 ( -0.091 ) | 2.053 ( -0.072 ) |
| | | 0.9 | 6.212 | 13.442 ( -1.455 ) | 1.345 ( -0.407 ) | 1.660 ( 0.082 ) | 1.911 ( -0.042 ) | 2.005 ( 0.082 ) |
| | | 0.4 | **8.220** | 18.028 ( 0.908 ) | 1.556 ( -2.512 ) | 1.485 ( -0.172 ) | 1.998 ( 0.284 ) | 2.060 ( -0.119 ) |
| | 50 | 0.6 | **37.293** | 83.118 ( -4.211 ) | 1.480 ( 0.009 ) | 2.289 ( -0.215 ) | 4.231 ( -0.068 ) | 4.475 ( -0.121 ) |
| | | 0.9 | **28.384** | 63.300 ( -1.455 ) | 1.854 ( -0.335 ) | 2.606 ( 0.007 ) | 2.116 ( -0.037 ) | 2.578 ( 0.089 ) |
| | | 0.4 | **75.203** | 167.149 ( 10.800 ) | 3.278 ( -1.341 ) | 9.480 ( -0.199 ) | 13.688 ( -0.074 ) | 7.104 ( -0.149 ) |
| RLBoot | 100 | 0.6 | 6.271 | 13.535 ( -1.375 ) | 1.366 ( 0.103 ) | 1.633 ( -0.207 ) | 2.054 ( -0.084 ) | 2.155 ( -0.076 ) |
| | | 0.9 | 8.175 | 17.934 ( -1.099 ) | 1.365 ( -0.444 ) | 1.681 ( 0.061 ) | 1.936 ( -0.051 ) | 2.021 ( 0.091 ) |
| | | 0.4 | 5.793 | 12.542 ( 3.148 ) | 1.350 ( -2.544 ) | 1.406 ( -0.128 ) | 1.701 ( 0.399 ) | 1.953 ( -0.105 ) |
| | 150 | 0.6 | **76.024** | 10.264 ( -1.649 ) | 117.299 ( 6.471 ) | 109.060 ( -0.959 ) | 43.600 ( 0.265 ) | 35.199 ( -0.361 ) |
| | | 0.9 | 4.665 | 9.829 ( -1.449 ) | 1.328 ( -0.407 ) | 1.664 ( 0.061 ) | 1.912 ( -0.043 ) | 2.000 ( 0.090 ) |

**Tabla A-3.**: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$, porcentaje de valores faltantes del 10 % y $\rho_x \in (0{,}25, 0{,}50]$

| Método | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ $(BiasR_1)$ | $RMSE_2$ $(BiasR_2)$ | $RMSE_3$ $(BiasR_3)$ | $RMSE_4(BiasR_4)$ | $RMSE_5(BiasR_5)$ |
|---|---|---|---|---|---|---|---|---|
| | | 0.4 | **33.499** | 74.806 ( -16.675 ) | 1.649 ( -0.328 ) | 1.839 ( -0.455 ) | 2.086 ( -0.179 ) | 2.076 ( 0.023 ) |
| | 50 | 0.6 | **1112.666** | 2487.311 ( -44.640 ) | 29.303 ( 2.080 ) | 37.801 ( -1.176 ) | 25.594 ( -0.587 ) | 21.658 ( 0.103 ) |
| | | 0.9 | 19.059 | 42.427 ( -2.581 ) | 1.549 ( -0.413 ) | 1.781 ( -0.252 ) | 2.201 ( 0.043 ) | 2.396 ( -0.455 ) |
| | | 0.4 | 5.516 | 11.757 ( 5.466 ) | 1.562 ( -0.302 ) | 1.793 ( -0.492 ) | 2.016 ( -0.175 ) | 2.043 ( 0.004 ) |
| PMM | 100 | 0.6 | 4.743 | 9.802 ( -0.357 ) | 1.593 ( 0.236 ) | 1.842 ( -0.335 ) | 2.206 ( -0.142 ) | 2.369 ( -0.109 ) |
| | | 0.9 | 9.642 | 21.181 ( -3.188 ) | 1.600 ( -0.452 ) | 1.726 ( -0.267 ) | 2.190 ( -0.005 ) | 2.427 ( -0.442 ) |
| | | 0.4 | 5.336 | 11.339 ( 2.385 ) | 1.551 ( -0.316 ) | 1.810 ( -0.492 ) | 1.993 ( -0.172 ) | 2.034 ( -0.001 ) |
| | 150 | 0.6 | 3.952 | 7.864 ( -0.298 ) | 1.598 ( 0.206 ) | 1.829 ( -0.326 ) | 2.195 ( -0.140 ) | 2.351 ( -0.104 ) |
| | | 0.9 | 5.237 | 11.052 ( -3.044 ) | 1.495 ( -0.439 ) | 1.692 ( -0.268 ) | 2.139 ( -0.010 ) | 2.308 ( -0.458 ) |
| | | 0.4 | **14.971** | 33.180 ( -15.103 ) | 1.858 ( -0.340 ) | 2.245 ( -0.398 ) | 2.544 ( -0.201 ) | 2.175 ( 0.001 ) |
| | 50 | 0.6 | 7.466 | 16.136 ( 0.480 ) | 1.762 ( 0.206 ) | 2.004 ( -0.326 ) | 2.290 ( -0.133 ) | 2.443 ( -0.110 ) |
| | | 0.9 | 9.284 | 20.365 ( -1.563 ) | 1.555 ( -0.424 ) | 1.721 ( -0.249 ) | 2.219 ( 0.020 ) | 2.429 ( -0.444 ) |
| | | 0.4 | 9.575 | 21.083 ( -2.647 ) | 1.579 ( -0.306 ) | 1.783 ( -0.481 ) | 2.029 ( -0.178 ) | 2.041 ( 0.000 ) |
| RLB | 100 | 0.6 | 4.680 | 9.633 ( 0.256 ) | 1.643 ( 0.175 ) | 1.856 ( -0.341 ) | 2.218 ( -0.148 ) | 2.381 ( -0.114 ) |
| | | 0.9 | 6.968 | 15.069 ( -3.496 ) | 1.543 ( -0.447 ) | 1.710 ( -0.267 ) | 2.192 ( 0.011 ) | 2.368 ( -0.424 ) |
| | | 0.4 | 16.793 | 37.366 ( 9.588 ) | 1.573 ( -0.323 ) | 1.796 ( -0.500 ) | 1.994 ( -0.167 ) | 2.017 ( 0.006 ) |
| | 150 | 0.6 | 4.231 | 8.543 ( -0.278 ) | 1.615 ( 0.181 ) | 1.844 ( -0.339 ) | 2.209 ( -0.148 ) | 2.371 ( -0.118 ) |
| | | 0.9 | 5.562 | 11.809 ( -3.270 ) | 1.497 ( -0.432 ) | 1.713 ( -0.273 ) | 2.160 ( -0.009 ) | 2.316 ( -0.442 ) |
| | | 0.4 | 9.952 | 21.850 ( 0.488 ) | 1.819 ( -0.297 ) | 2.051 ( -0.529 ) | 2.361 ( -0.164 ) | 2.179 ( 0.021 ) |
| | 50 | 0.6 | 6.687 | 14.366 ( 0.686 ) | 1.630 ( 0.229 ) | 1.876 ( -0.315 ) | 2.243 ( -0.140 ) | 2.449 ( -0.113 ) |
| | | 0.9 | 6.918 | 14.944 ( -0.443 ) | 1.549 ( -0.427 ) | 1.720 ( -0.244 ) | 2.187 ( 0.048 ) | 2.413 ( -0.433 ) |
| | | 0.4 | 9.112 | 20.026 ( 16.474 ) | 1.584 ( -0.305 ) | 1.812 ( -0.493 ) | 2.040 ( -0.180 ) | 2.049 ( 0.000 ) |
| RLBoot | 100 | 0.6 | 4.415 | 8.978 ( -0.193 ) | 1.630 ( 0.189 ) | 1.856 ( -0.343 ) | 2.221 ( -0.148 ) | 2.411 ( -0.129 ) |
| | | 0.9 | 5.702 | 12.132 ( -2.996 ) | 1.528 ( -0.441 ) | 1.717 ( -0.266 ) | 2.158 ( 0.027 ) | 2.327 ( -0.414 ) |
| | | 0.4 | 39.400 | 88.020 ( 15.014 ) | 1.653 ( -0.331 ) | 1.829 ( -0.504 ) | 2.049 ( -0.180 ) | 2.039 ( -0.002 ) |
| | 150 | 0.6 | 3.688 | 7.164 ( -0.628 ) | 1.613 ( 0.188 ) | 1.862 ( -0.341 ) | 2.219 ( -0.148 ) | 2.384 ( -0.127 ) |
| | | 0.9 | 5.314 | 11.238 ( -2.681 ) | 1.513 ( -0.439 ) | 1.700 ( -0.267 ) | 2.129 ( 0.019 ) | 2.282 ( -0.426 ) |

**Tabla A-4.**: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$, porcentaje de valores faltantes del 10 % y $\rho_x \in (0{,}50, 0{,}75]$

| Método | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ ($BiasR_1$) | $RMSE_2$ ($BiasR_2$) | $RMSE_3$ ($BiasR_3$) | $RMSE_4$($BiasR_4$) | $RMSE_5$($BiasR_5$) |
|---|---|---|---|---|---|---|---|---|
| | | 0.4 | 14.190 | 31.282 ( 0.182 ) | 2.287 ( 0.006 ) | 2.183 ( -0.238 ) | 2.804 ( -0.293 ) | 3.230 ( -0.314 ) |
| | 50 | 0.6 | 6.853 | 14.577 ( 4.305 ) | 2.046 ( -0.369 ) | 2.328 ( -0.326 ) | 2.417 ( -0.393 ) | 2.617 ( -0.053 ) |
| | | 0.9 | 793.322 | 1677.398 ( 34.529 ) | 211.056 ( 4.066 ) | 70.151 ( 0.804 ) | 200.075 ( -1.548 ) | 493.601 ( -5.375 ) |
| | | 0.4 | 4.160 | 7.947 ( 0.241 ) | 2.113 ( -0.040 ) | 2.203 ( -0.201 ) | 2.498 ( -0.297 ) | 2.799 ( -0.342 ) |
| PMM | 100 | 0.6 | 4.574 | 9.069 ( 2.979 ) | 2.116 ( -0.400 ) | 2.289 ( -0.360 ) | 2.438 ( -0.416 ) | 2.594 ( -0.094 ) |
| | | 0.9 | 5.440 | 11.116 ( -0.246 ) | 2.084 ( -0.509 ) | 2.289 ( -0.144 ) | 2.532 ( -0.356 ) | 2.899 ( -0.292 ) |
| | | 0.4 | 3.629 | 6.681 ( 0.379 ) | 1.951 ( 0.049 ) | 1.981 ( -0.243 ) | 2.402 ( -0.321 ) | 2.782 ( -0.350 ) |
| | 150 | 0.6 | 4.037 | 7.720 ( 2.087 ) | 2.042 ( -0.412 ) | 2.293 ( -0.367 ) | 2.422 ( -0.436 ) | 2.568 ( -0.113 ) |
| | | 0.9 | 5.942 | 12.379 ( -0.096 ) | 2.064 ( -0.513 ) | 2.225 ( -0.108 ) | 2.468 ( -0.352 ) | 2.824 ( -0.293 ) |
| | | 0.4 | 5.811 | 12.054 ( -0.034 ) | 2.068 ( -0.042 ) | 2.074 ( -0.238 ) | 2.552 ( -0.333 ) | 2.910 ( -0.349 ) |
| | 50 | 0.6 | 7.159 | 15.251 ( 4.117 ) | 2.083 ( -0.383 ) | 2.364 ( -0.331 ) | 2.556 ( -0.400 ) | 2.688 ( -0.073 ) |
| | | 0.9 | 169.433 | 375.754 ( 3.334 ) | 18.427 ( -0.938 ) | 29.104 ( 1.576 ) | 28.738 ( -0.037 ) | 18.287 ( -0.355 ) |
| | | 0.4 | 13.526 | 29.756 ( -0.052 ) | 2.184 ( -0.102 ) | 2.346 ( -0.271 ) | 2.591 ( -0.351 ) | 3.504 ( -0.375 ) |
| RLB | 100 | 0.6 | 5.323 | 10.962 ( 2.926 ) | 2.045 ( -0.407 ) | 2.280 ( -0.350 ) | 2.377 ( -0.406 ) | 2.542 ( -0.086 ) |
| | | 0.9 | 5.044 | 10.210 ( 0.311 ) | 2.034 ( -0.486 ) | 2.222 ( -0.062 ) | 2.463 ( -0.340 ) | 2.794 ( -0.272 ) |
| | | 0.4 | 4.547 | 8.816 ( 0.258 ) | 2.289 ( -0.405 ) | 2.025 ( -0.260 ) | 2.640 ( -0.340 ) | 3.061 ( -0.362 ) |
| | 150 | 0.6 | 4.798 | 9.582 ( 2.469 ) | 2.094 ( -0.406 ) | 2.331 ( -0.359 ) | 2.430 ( -0.435 ) | 2.745 ( -0.140 ) |
| | | 0.9 | 6.054 | 12.664 ( 0.185 ) | 2.022 ( -0.494 ) | 2.225 ( -0.056 ) | 2.463 ( -0.340 ) | 2.791 ( -0.269 ) |
| | | 0.4 | 11.171 | 22.578 ( -0.621 ) | 3.464 ( 0.387 ) | 3.101 ( -0.276 ) | 5.081 ( -0.398 ) | 8.167 ( -0.388 ) |
| | 50 | 0.6 | 5.836 | 11.598 ( 3.650 ) | 3.163 ( -0.446 ) | 2.718 ( -0.353 ) | 3.261 ( -0.426 ) | 2.788 ( -0.088 ) |
| | | 0.9 | 7.290 | 15.510 ( 0.776 ) | 2.170 ( -0.508 ) | 2.345 ( -0.136 ) | 2.548 ( -0.342 ) | 2.905 ( -0.279 ) |
| | | 0.4 | 5.103 | 10.241 ( 0.186 ) | 2.047 ( 0.079 ) | 2.371 ( -0.241 ) | 2.652 ( -0.341 ) | 2.915 ( -0.358 ) |
| RLBoot | 100 | 0.6 | 4.949 | 9.956 ( 2.595 ) | 2.153 ( -0.419 ) | 2.372 ( -0.361 ) | 2.485 ( -0.433 ) | 2.623 ( -0.092 ) |
| | | 0.9 | 6.451 | 13.590 ( -0.247 ) | 2.051 ( -0.493 ) | 2.245 ( -0.121 ) | 2.484 ( -0.355 ) | 2.820 ( -0.283 ) |
| | | 0.4 | 6.761 | 13.396 ( -0.039 ) | 2.682 ( 0.406 ) | 3.031 ( -0.304 ) | 3.188 ( -0.411 ) | 4.750 ( -0.399 ) |
| | 150 | 0.6 | 4.022 | 7.675 ( 1.339 ) | 2.060 ( -0.417 ) | 2.294 ( -0.362 ) | 2.419 ( -0.433 ) | 2.572 ( -0.110 ) |
| | | 0.9 | 6.264 | 13.115 ( -0.294 ) | 2.079 ( -0.507 ) | 2.268 ( -0.104 ) | 2.561 ( -0.360 ) | 2.859 ( -0.293 ) |

**Tabla A-5.**: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$, porcentaje de valores faltantes del 10 % y $\rho_x \in (0{,}75, 1]$

| Método | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ $(BiasR_1)$ | $RMSE_2$ $(BiasR_2)$ | $RMSE_3$ $(BiasR_3)$ | $RMSE_4(BiasR_4)$ | $RMSE_5(BiasR_5)$ |
|---|---|---|---|---|---|---|---|---|
| | | 0.4 | 13.198 | 29.180 ( 12.289 ) | 1.766 ( -0.337 ) | 2.034 ( -0.353 ) | 2.497 ( -0.351 ) | 2.447 ( -0.835 ) |
| | 50 | 0.6 | 47.726 | 106.198 ( 9.822 ) | 8.255 ( -0.309 ) | 4.729 ( -0.346 ) | 2.892 ( -0.104 ) | 3.496 ( -0.348 ) |
| | | 0.9 | 17.020 | 37.722 ( 7.435 ) | 1.934 ( -0.320 ) | 2.586 ( -0.109 ) | 2.446 ( -0.613 ) | 3.012 ( -0.286 ) |
| | | 0.4 | 25.161 | 56.096 ( 12.653 ) | 1.728 ( -0.327 ) | 1.980 ( -0.347 ) | 2.446 ( -0.345 ) | 2.421 ( -0.845 ) |
| PMM | 100 | 0.6 | 20.944 | 46.627 ( 8.411 ) | 1.769 ( -0.379 ) | 2.133 ( -0.348 ) | 2.332 ( -0.135 ) | 2.486 ( -0.337 ) |
| | | 0.9 | 18.383 | 40.809 ( 7.765 ) | 2.036 ( -0.314 ) | 2.350 ( -0.118 ) | 2.400 ( -0.586 ) | 2.979 ( -0.285 ) |
| | | 0.4 | 12.240 | 27.030 ( 11.216 ) | 1.692 ( -0.316 ) | 1.980 ( -0.349 ) | 2.425 ( -0.344 ) | 2.399 ( -0.841 ) |
| | 150 | 0.6 | 49.070 | 109.636 ( 6.803 ) | 1.760 ( -0.380 ) | 2.147 ( -0.353 ) | 2.316 ( -0.130 ) | 2.494 ( -0.342 ) |
| | | 0.9 | 39.937 | 89.180 ( 7.408 ) | 1.839 ( -0.272 ) | 2.161 ( -0.093 ) | 2.314 ( -0.553 ) | 2.898 ( -0.265 ) |
| | | 0.4 | 390.891 | 874.008 ( -21.223 ) | 6.833 ( -0.059 ) | 3.828 ( -0.307 ) | 2.621 ( -0.343 ) | 4.546 ( -0.841 ) |
| | 50 | 0.6 | 145.016 | 324.231 ( 11.699 ) | 2.067 ( -0.365 ) | 2.485 ( -0.358 ) | 2.445 ( -0.105 ) | 2.617 ( -0.320 ) |
| | | 0.9 | 106.821 | 238.793 ( 3.521 ) | 1.986 ( -0.337 ) | 3.192 ( -0.267 ) | 2.584 ( -0.607 ) | 3.205 ( -0.311 ) |
| | | 0.4 | 81.355 | 181.829 ( 21.226 ) | 2.864 ( -0.399 ) | 3.040 ( -0.370 ) | 2.823 ( -0.333 ) | 2.563 ( -0.806 ) |
| RLB | 100 | 0.6 | 77.475 | 173.180 ( -1.122 ) | 1.943 ( -0.380 ) | 2.170 ( -0.343 ) | 2.407 ( -0.131 ) | 2.563 ( -0.321 ) |
| | | 0.9 | 12.828 | 28.279 ( 7.111 ) | 1.855 ( -0.317 ) | 2.246 ( -0.136 ) | 2.390 ( -0.583 ) | 2.971 ( -0.289 ) |
| | | 0.4 | 66.909 | 149.547 ( 11.570 ) | 1.786 ( -0.343 ) | 2.085 ( -0.356 ) | 2.474 ( -0.351 ) | 2.462 ( -0.839 ) |
| | 150 | 0.6 | 12.304 | 27.170 ( 10.464 ) | 1.740 ( -0.374 ) | 2.078 ( -0.338 ) | 2.274 ( -0.115 ) | 2.488 ( -0.339 ) |
| | | 0.9 | 11.516 | 25.267 ( 6.728 ) | 1.979 ( -0.304 ) | 2.248 ( -0.107 ) | 2.590 ( -0.560 ) | 2.997 ( -0.278 ) |
| | | 0.4 | 51.206 | 114.398 ( 18.276 ) | 2.095 ( -0.367 ) | 2.359 ( -0.374 ) | 2.692 ( -0.324 ) | 2.555 ( -0.823 ) |
| | 50 | 0.6 | 217.496 | 486.312 ( 25.522 ) | 1.974 ( -0.325 ) | 2.348 ( -0.343 ) | 2.774 ( -0.092 ) | 2.555 ( -0.344 ) |
| | | 0.9 | 90.835 | 202.667 ( 1.807 ) | 4.460 ( -0.275 ) | 11.015 ( -0.544 ) | 5.182 ( -0.521 ) | 3.649 ( -0.274 ) |
| | | 0.4 | 12.329 | 27.219 ( 11.389 ) | 1.746 ( -0.346 ) | 2.006 ( -0.362 ) | 2.447 ( -0.348 ) | 2.463 ( -0.838 ) |
| RLBoot | 100 | 0.6 | 110.816 | 247.741 ( -5.861 ) | 2.172 ( -0.380 ) | 2.435 ( -0.350 ) | 2.621 ( -0.185 ) | 2.708 ( -0.356 ) |
| | | 0.9 | 17.619 | 39.108 ( 7.684 ) | 1.839 ( -0.308 ) | 2.245 ( -0.150 ) | 2.352 ( -0.574 ) | 2.966 ( -0.285 ) |
| | | 0.4 | 17.972 | 39.952 ( 10.655 ) | 1.713 ( -0.330 ) | 1.993 ( -0.351 ) | 2.419 ( -0.343 ) | 2.437 ( -0.838 ) |
| | 150 | 0.6 | 42.217 | 94.298 ( 8.034 ) | 1.755 ( -0.373 ) | 2.120 ( -0.337 ) | 2.320 ( -0.137 ) | 2.508 ( -0.343 ) |
| | | 0.9 | 12.162 | 26.770 ( 7.446 ) | 1.883 ( -0.279 ) | 2.197 ( -0.114 ) | 2.431 ( -0.575 ) | 2.931 ( -0.257 ) |

**Tabla A-6.**: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$, porcentaje de valores faltantes del 30 % y $\rho_x \in (0{,}25, 0{,}50]$

| Método | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ ($BiasR_1$) | $RMSE_2$ ($BiasR_2$) | $RMSE_3$ ($BiasR_3$) | $RMSE_4$($BiasR_4$) | $RMSE_5$($BiasR_5$) |
|---|---|---|---|---|---|---|---|---|
| | | 0.4 | 58.922 | 131.599 ( 15.528 ) | 2.804 ( -0.508 ) | 3.219 ( -0.501 ) | 3.186 ( -0.487 ) | 3.501 ( -0.507 ) |
| | 50 | 0.6 | 34.647 | 77.260 ( 577.733 ) | 2.531 ( -0.432 ) | 2.796 ( -0.645 ) | 3.212 ( -0.487 ) | 2.888 ( -0.444 ) |
| | | 0.9 | 18.809 | 41.704 ( -0.859 ) | 2.185 ( -0.527 ) | 2.564 ( -0.650 ) | 2.901 ( 0.337 ) | 3.134 ( -0.473 ) |
| | | 0.4 | 14.081 | 30.945 ( 21.875 ) | 2.554 ( -0.545 ) | 2.925 ( -0.549 ) | 2.997 ( -0.600 ) | 3.108 ( -0.535 ) |
| PMM | 100 | 0.6 | 15.984 | 35.298 ( 531.471 ) | 2.458 ( -0.468 ) | 2.764 ( -0.661 ) | 3.081 ( -0.493 ) | 2.883 ( -0.450 ) |
| | | 0.9 | 13.809 | 30.397 ( -0.705 ) | 2.181 ( -0.551 ) | 2.553 ( -0.652 ) | 2.899 ( 0.359 ) | 3.132 ( -0.484 ) |
| | | 0.4 | 10.265 | 22.218 ( 19.134 ) | 2.528 ( -0.546 ) | 2.910 ( -0.544 ) | 2.972 ( -0.555 ) | 3.087 ( -0.531 ) |
| | 150 | 0.6 | 13.523 | 29.712 ( 531.457 ) | 2.453 ( -0.469 ) | 2.788 ( -0.674 ) | 3.084 ( -0.496 ) | 2.873 ( -0.463 ) |
| | | 0.9 | 11.623 | 25.421 ( -0.691 ) | 2.145 ( -0.554 ) | 2.572 ( -0.663 ) | 2.901 ( 0.396 ) | 3.112 ( -0.479 ) |
| | | 0.4 | 212.715 | 475.565 ( 58.340 ) | 4.463 ( -0.600 ) | 3.631 ( -0.539 ) | 5.611 ( -1.122 ) | 3.328 ( -0.521 ) |
| | 50 | 0.6 | 55.208 | 123.235 ( 614.907 ) | 4.258 ( -0.466 ) | 3.917 ( -0.666 ) | 3.283 ( -0.488 ) | 2.929 ( -0.450 ) |
| | | 0.9 | 18.098 | 40.074 ( -1.336 ) | 2.326 ( -0.568 ) | 2.783 ( -0.676 ) | 2.931 ( 0.406 ) | 3.149 ( -0.476 ) |
| | | 0.4 | 12.719 | 27.860 ( 20.369 ) | 2.513 ( -0.541 ) | 2.858 ( -0.517 ) | 2.965 ( -0.617 ) | 3.058 ( -0.523 ) |
| RLB | 100 | 0.6 | 37.490 | 83.591 ( 496.696 ) | 2.559 ( -0.475 ) | 2.900 ( -0.634 ) | 3.750 ( -0.505 ) | 3.310 ( -0.446 ) |
| | | 0.9 | 12.550 | 27.536 ( -0.477 ) | 2.155 ( -0.559 ) | 2.558 ( -0.654 ) | 2.894 ( 0.317 ) | 3.114 ( -0.479 ) |
| | | 0.4 | 9.774 | 21.097 ( 17.671 ) | 2.510 ( -0.543 ) | 2.875 ( -0.525 ) | 2.961 ( -0.563 ) | 3.041 ( -0.521 ) |
| | 150 | 0.6 | 14.121 | 31.082 ( 590.947 ) | 2.433 ( -0.467 ) | 2.763 ( -0.669 ) | 3.051 ( -0.486 ) | 2.847 ( -0.443 ) |
| | | 0.9 | 21.168 | 47.023 ( -0.621 ) | 2.148 ( -0.567 ) | 2.542 ( -0.651 ) | 2.907 ( 0.339 ) | 3.110 ( -0.478 ) |
| | | 0.4 | 15.482 | 34.127 ( 26.300 ) | 2.555 ( -0.524 ) | 2.897 ( -0.527 ) | 2.980 ( -0.649 ) | 3.170 ( -0.537 ) |
| | 50 | 0.6 | 20.789 | 46.143 ( 674.079 ) | 2.441 ( -0.454 ) | 2.781 ( -0.662 ) | 3.122 ( -0.499 ) | 2.868 ( -0.452 ) |
| | | 0.9 | 16.749 | 37.045 ( -0.332 ) | 2.133 ( -0.530 ) | 2.622 ( -0.639 ) | 2.968 ( 0.362 ) | 3.174 ( -0.475 ) |
| | | 0.4 | 14.271 | 31.387 ( 23.681 ) | 2.513 ( -0.540 ) | 2.893 ( -0.545 ) | 2.998 ( -0.610 ) | 3.089 ( -0.530 ) |
| RLBoot | 100 | 0.6 | 16.824 | 37.198 ( 635.751 ) | 2.453 ( -0.471 ) | 2.754 ( -0.658 ) | 3.095 ( -0.497 ) | 2.881 ( -0.446 ) |
| | | 0.9 | 13.319 | 29.286 ( -0.446 ) | 2.151 ( -0.566 ) | 2.550 ( -0.648 ) | 2.887 ( 0.362 ) | 3.140 ( -0.481 ) |
| | | 0.4 | 10.171 | 22.006 ( 20.138 ) | 2.522 ( -0.545 ) | 2.910 ( -0.550 ) | 2.973 ( -0.593 ) | 3.052 ( -0.521 ) |
| | 150 | 0.6 | 16.467 | 36.389 ( 615.896 ) | 2.471 ( -0.479 ) | 2.774 ( -0.672 ) | 3.088 ( -0.497 ) | 2.869 ( -0.452 ) |
| | | 0.9 | 33.663 | 74.919 ( -3.086 ) | 2.549 ( -0.556 ) | 3.868 ( -0.600 ) | 3.766 ( 0.157 ) | 4.199 ( -0.499 ) |

**Tabla A-7.**: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$, un porcentaje de valores faltantes del 30 % y $\rho_x \in (0{,}50, 0{,}75]$

| Método | $N_d$ | $\rho_z$ | $RMSE$ | $RMSE_1$ $(BiasR_1)$ | $RMSE_2$ $(BiasR_2)$ | $RMSE_3$ $(BiasR_3)$ | $RMSE_4(BiasR_4)$ | $RMSE_5(BiasR_5)$ |
|---|---|---|---|---|---|---|---|---|
| | | 0.4 | 20.825 | 45.950 ( 5.261 ) | 3.248 ( -1.572 ) | 3.723 ( -0.699 ) | 3.822 ( -0.720 ) | 4.251 ( -0.606 ) |
| | 50 | 0.6 | 19.230 | 42.434 ( 7.576 ) | 2.998 ( -0.840 ) | 3.241 ( -1.161 ) | 3.648 ( -0.679 ) | 3.933 ( -0.628 ) |
| | | 0.9 | 18.900 | 41.323 ( 0.836 ) | 4.343 ( -1.057 ) | 4.087 ( -1.082 ) | 4.724 ( -0.592 ) | 4.534 ( -0.633 ) |
| | | 0.4 | 27.158 | 60.151 ( 5.740 ) | 3.871 ( -1.615 ) | 4.233 ( -0.696 ) | 4.119 ( -0.735 ) | 4.434 ( -0.595 ) |
| PMM | 100 | 0.6 | 14.209 | 30.888 ( 5.466 ) | 3.307 ( -0.848 ) | 3.578 ( -1.114 ) | 3.919 ( -0.706 ) | 4.044 ( -0.617 ) |
| | | 0.9 | 12.594 | 27.127 ( 1.016 ) | 2.991 ( -1.119 ) | 3.773 ( -1.193 ) | 4.065 ( -0.639 ) | 4.182 ( -0.670 ) |
| | | 0.4 | 50.011 | 111.336 ( 6.830 ) | 6.321 ( -1.702 ) | 5.092 ( -0.708 ) | 4.212 ( -0.748 ) | 5.114 ( -0.609 ) |
| | 150 | 0.6 | 8.866 | 18.563 ( 5.853 ) | 2.928 ( -0.840 ) | 3.218 ( -1.273 ) | 3.674 ( -0.705 ) | 4.007 ( -0.635 ) |
| | | 0.9 | 29.846 | 65.955 ( -3.878 ) | 6.403 ( -1.023 ) | 4.190 ( -1.269 ) | 5.194 ( -0.625 ) | 4.268 ( -0.671 ) |
| | | 0.4 | 195.487 | 430.467 ( 10.563 ) | 49.980 ( -3.114 ) | 50.059 ( -0.999 ) | 18.156 ( -0.777 ) | 20.996 ( -0.611 ) |
| | 50 | 0.6 | 32.817 | 72.920 ( 7.573 ) | 3.239 ( -0.867 ) | 3.328 ( -1.278 ) | 4.152 ( -0.706 ) | 5.352 ( -0.712 ) |
| | | 0.9 | 18.235 | 39.984 ( -0.682 ) | 3.242 ( -1.090 ) | 3.988 ( -1.316 ) | 4.324 ( -0.663 ) | 4.330 ( -0.662 ) |
| | | 0.4 | 59.942 | 131.268 ( 6.584 ) | 14.823 ( -0.753 ) | 20.256 ( -0.443 ) | 5.242 ( -0.776 ) | 8.758 ( -0.572 ) |
| RLB | 100 | 0.6 | 16.906 | 36.987 ( 5.360 ) | 3.069 ( -0.828 ) | 3.477 ( -1.324 ) | 3.758 ( -0.703 ) | 5.045 ( -0.607 ) |
| | | 0.9 | 9.799 | 20.602 ( -0.141 ) | 2.969 ( -1.119 ) | 3.702 ( -1.205 ) | 4.011 ( -0.612 ) | 4.124 ( -0.672 ) |
| | | 0.4 | 18.020 | 39.026 ( 5.534 ) | 6.306 ( -1.063 ) | 3.635 ( -0.674 ) | 5.479 ( -0.805 ) | 4.190 ( -0.594 ) |
| | 150 | 0.6 | 8.338 | 17.342 ( 5.766 ) | 2.921 ( -0.851 ) | 3.165 ( -1.267 ) | 3.616 ( -0.689 ) | 3.905 ( -0.642 ) |
| | | 0.9 | 9.356 | 19.543 ( -0.348 ) | 2.944 ( -1.109 ) | 3.705 ( -1.077 ) | 4.050 ( -0.652 ) | 4.121 ( -0.688 ) |
| | | 0.4 | 33.276 | 73.224 ( 5.083 ) | 8.322 ( -1.959 ) | 5.569 ( -0.696 ) | 4.386 ( -0.667 ) | 7.426 ( -0.547 ) |
| | 50 | 0.6 | 99.565 | 222.381 ( 4.114 ) | 4.255 ( -0.916 ) | 5.385 ( -0.307 ) | 6.608 ( -0.852 ) | 4.647 ( -0.635 ) |
| | | 0.9 | 43.848 | 97.687 ( -1.677 ) | 3.460 ( -1.133 ) | 4.370 ( -1.237 ) | 4.560 ( -0.685 ) | 4.300 ( -0.657 ) |
| | | 0.4 | 98.002 | 217.908 ( -0.587 ) | 9.872 ( -2.220 ) | 9.395 ( -0.816 ) | 17.674 ( -1.042 ) | 6.331 ( -0.708 ) |
| RLBoot | 100 | 0.6 | 32.576 | 71.922 ( 7.209 ) | 3.764 ( -0.829 ) | 5.751 ( -0.556 ) | 4.092 ( -0.714 ) | 8.333 ( -0.602 ) |
| | | 0.9 | 11.953 | 25.545 ( -0.119 ) | 2.937 ( -1.056 ) | 3.739 ( -0.908 ) | 4.208 ( -0.646 ) | 4.636 ( -0.696 ) |
| | | 0.4 | 310.939 | 694.482 ( 9.202 ) | 13.697 ( -2.320 ) | 19.696 ( -0.528 ) | 11.050 ( -0.620 ) | 20.296 ( -0.458 ) |
| | 150 | 0.6 | 18.748 | 41.220 ( 5.361 ) | 3.255 ( -0.830 ) | 3.425 ( -1.242 ) | 3.864 ( -0.680 ) | 4.598 ( -0.686 ) |
| | | 0.9 | 10.384 | 21.950 ( -0.121 ) | 3.013 ( -1.124 ) | 3.766 ( -1.129 ) | 4.137 ( -0.660 ) | 4.124 ( -0.694 ) |

**Tabla A-8.**: Errores cuadráticos medios de los $\beta$'s y sesgos relativos promedios para $P = 5$, porcentaje de valores faltantes del 30 % y $\rho_x \in (0{,}75, 1]$

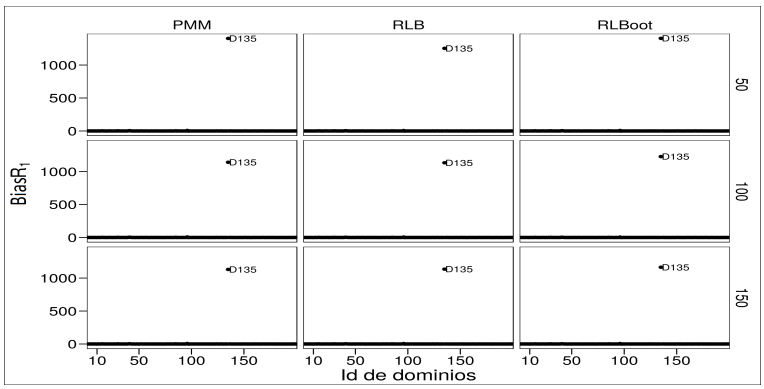


**Figura A-1**.: Sesgos relativos promedios por dominio para $\beta_1$ cuando $P = 3$, $\rho_x \in (0{,}50, 0{,}75]$, $\rho_z = 0{,}6$ y un 10 % de datos faltantes.

# Bibliografía


Acock, A. C. (2005), 'Working with missing values', *Journal of Marriage and family* **67**(4), 1012–1028.

Battese, G. E., Harter, R. M. & Fuller, W. A. (1988), 'An error-components model for prediction of county crop areas using survey and satellite data', *Journal of the American Statistical Association* **83**(401), 28–36.

Benavent, R. & Morales, D. (2016), 'Multivariate Fay–Herriot models for small area estimation', *Computational Statistics & Data Analysis* **94**, 372–390.

Chambers, R. (1986), 'Design-adjusted parameter estimation', *Journal of the Royal Statistical Society. Series A (General)* pp. 161–173.

Chambers, R. L. (1995), 'Regression analysis with sample survey data', *Manuscrito inédito cedido por cortesia do autor* **30**, 70–87.

Chandra, H. & Chambers, R. (2009), 'Multipurpose weighting for small area estimation'.

Cheng, X., Cook, D. & Hofmann, H. (2016), 'MissingDataGUI: A GUI for Missing Data Exploration'.

Cochran, W. G. (1977), *Sampling Techniques: 3d Ed*, Wiley.

Cox, D. & Hinkley, D. (1974), 'Theoretical statistics'.

Das, K., Jiang, J. & Rao, J. (2004), 'Mean squared error of empirical predictor', *Annals of Statistics* pp. 818–840.

Datta, G., Fay, R. & Ghosh, M. (1991), Hierarchical and Empirical Multivariate Bayes Analysis in Small, *in* 'Proceedings of the Annual Research Conference', US Department of Commerce, Bureau of the Census, pp. 63–78.

Datta, G., Ghosh, M., Nangia, N. & Natarajan, K. (1996), 'Estimation of median income of four-person families: a Bayesian approach', *Bayesian Analysis in Statistics and Econometrics,(DA Berry, KM Chaloner and JM Geweke Eds.), Wiley, New York* pp. 129–140.

Datta, G. S. (2009), 'Model-based approach to small area estimation', *Handbook of Statistics* **29**, 251–288.

Datta, G. S., Kubokawa, T., Molina, I. & Rao, J. N. (2011), 'Estimation of mean squared error of model-based small area estimators', *Test* **20**(2), 367–388.

Datta, G. S. & Lahiri, P. (2000), 'A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems', *Statistica Sinica* pp. 613–627.

de Leeuw, J. & Mair, P. (2009), 'Gifi methods for optimal scaling in R: The package homals', *Journal of Statistical Software* **31**(4), 1–20.
**URL:** *http://www.jstatsoft.org/v31/i04/*

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the royal statistical society. Series B (methodological)* pp. 1–38.

Deville, J. & Särndal, C. (1992), 'Calibration estimators in survey sampling', *Journal of the American Statistical Association* **87**, 376–382.

Fay, R. (1987), 'Application of multivariate regression to small domain estimation', *Small Area Statistics (R. Platek, JNK Rao, CE Sarndal and MP Singh, eds.)* pp. 91–102.

Fay, R. E. & Herriot, R. A. (1979), 'Estimates of income for small places: an application of James-Stein procedures to census data', *Journal of the American Statistical Association* **74**(366a), 269–277.

García, J. G., Albaladejo, J. P. & Fernández, J. A. M. (2006), 'Métodos de inferencia estadística con datos faltantes: estudio de simulación sobre los efectos en las estimaciones.', *Estadística española* **48**(162), 241–270.

Gelman, A. & Hill, J. (2011), 'Opening Windows to the Black Box', *Journal of Statistical Software* **40**.

Geman, S. & Geman, D. (1987), Stochastic relaxation, gibbs distributions, and the bayesian restoration of images, *in* 'Readings in computer vision', Elsevier, pp. 564–584.

Ghosh, M. & Rao, J. (1994), 'Small area estimation: an appraisal', *Statistical science* pp. 55–76.

González-Manteiga, W., Lombarda, M., Molina, I., Morales, D. & Santamaría, L. (2010), 'Small area estimation under Fay–Herriot models with non-parametric estimation of heteroscedasticity', *Statistical Modelling* **10**(2), 215–239.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. & Santamaría, L. (2008), 'Analytic and bootstrap approximations of prediction errors under a multivariate Fay–Herriot model', *Computational Statistics & Data Analysis* **52**(12), 5242–5252.

Gutiérrez, H. A. (2009), *Estrategias de muestreo. Diseño de encuestas y estimación de parámetros*, Hugo Andrés Gutiérrez Rojas.

Harrell Jr, F. E., Dupont, w. c. f. C. & many others. (2018), *Hmisc: Harrell Miscellaneous*. R package version 4.1-1.
**URL:** *https://CRAN.R-project.org/package=Hmisc*

Hartley, H. & Sielken Jr, R. L. (1975), 'A"super-population viewpoint"for finite population sampling', *Biometrics* pp. 411–422.

Henderson, C. R. (1975), 'Best linear unbiased estimation and prediction under a selection model', *Biometrics* pp. 423–447.

Herrador, M., Esteban, M., Hobza, T. & Morales, D. (2011), 'A Fay–Herriot Model with Different Random Effect Variances', *Communications in Statistics Theory and Methods* **40**(5), 785–797.

Holt, D., Smith, T. & Winter, P. (1980), 'Regression analysis of data from complex surveys', *Journal of the Royal Statistical Society. Series A (General)* pp. 474–487.

Honaker, J., King, G. & Blackwell, M. (2011), 'Amelia II: A Program for Missing Data', *Journal of Statistical Software* **45**(7), 1–47.
**URL:** *http://www.jstatsoft.org/v45/i07/*

Horvitz, D. G. & Thompson, D. J. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American statistical Association* **47**(260), 663–685.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R. & Herring, A. H. (2005), 'Missing-data methods for generalized linear models: A comparative review', *Journal of the American Statistical Association* **100**(469), 332–346.

Icfes (2011), Informe técnico SABER 5o. y 9o. 2009, Technical report, Instituto Colombiano para la Evaluación de la Educación.

Icfes (2014), ESTUDIO SOBRE EL CONTEXTO ESCOLAR Y SOCIAL DEL APRENDIZAJE EN COLOMBIA -CESAC.

Icfes (2016*a*), Marco de factores asociados Saber 3, 5 y 9 2016, Technical report.
**URL:** *http://www.icfes.gov.co/docman/instituciones-educativas-y-secretarias/pruebas-saber-3579/factores-asociados-2/2821-marco-de-factores-asociados-saber-3-5-y-9/file?force-download=1*

Icfes (2016*b*), SABER 3, 5 y 9 resultados nacionales 2009 2014, Technical report.
**URL:** *http://www2.icfes.gov.co/docman/instituciones-educativas-y-secretarias/pruebas-saber-3579/documentos/informes-saber-3-5-y-9/2323-resultados-nacionales-saber-3o-5o-y-9o-2009-2014/file?force-download=1*

Icfes (2017), EXPERIENCIA ACUMULADA EN EL LEVANTAMIENTO DE DATOS DE FACTORES ASOCIADOS, 2012-2015, SABER 3, 5 Y 9, Technical report.
**URL:** *http://www.icfes.gov.co/docman/instituciones-educativas-y-secretarias/pruebas-saber-3579/factores-asociados-2/4864-experiencia-acumulada-en-el-levantamiento-de-datos-de-factores-asociados-2012-2015-saber-3-5-y-9/file?force-download=1*

Jiang, J. & Lahiri, P. (2006*a*), 'Estimation of finite population domain means: A model-assisted empirical best prediction approach', *Journal of the American Statistical Association* **101**(473), 301–311.

Jiang, J. & Lahiri, P. (2006*b*), 'Mixed model prediction and small area estimation', *Test* **15**(1), 1.

Jiang, J., Nguyen, T. & Rao, J. S. (2011), 'Best predictive small area estimation', *Journal of the American Statistical Association* **106**(494), 732–745.

Josse, J. & Husson, F. (2016), 'missMDA: A Package for Handling Missing Values in Multivariate Data Analysis', *Journal of Statistical Software* **70**(1), 1–31.

Kubokawa, T. (2012), 'On measuring uncertainty of small area estimators with higher order accuracy', *Journal of the Japan Statistical Society* **41**(2), 93–119.

Lehtonen, R. & Veijanen, A. (2009), 'Design-based methods of estimation for domains and small areas', *Handbook of statistics* **29**, 219–249.

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

Little, R. J. (1992), 'Regression with missing X's: a review', *Journal of the American Statistical Association* **87**(420), 1227–1237.

Little, R. J. & Rubin, D. B. (1987), 'Statistical Analysis withMissing Data'.

Longford, N. (2005), 'Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician (Statistics for Social Science and Behavorial Sciences)'.

Longford, N. T. (2004), 'Missing data and small area estimation in the UK Labour Force Survey', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **167**(2), 341–373.

Lumley, T. (2004), 'Analysis of Complex Survey Samples', *Journal of Statistical Software* **9**(1), 1–19. R package verson 2.2.

Lumley, T. (2016), 'survey: analysis of complex survey samples'. R package version 3.32.

McLachlan, G. & Krishnan, T. (2007), *The EM algorithm and extensions*, Vol. 382, John Wiley & Sons.

Medina, F. & Galván, M. (2007), 'Estudios estadísticos y prospectivos', *Imputación de datos: Teoría y práctica. División Estadísitca y proyecciones Económicas Naciones Unidas. CEPAL* .

Meinfelder, F. & Schnapp, T. (2015), *BaBooN: Bayesian Bootstrap Predictive Mean Matching - Multiple and Single Imputation for Discrete Data*. R package version 0.2-0.
**URL:** *https://CRAN.R-project.org/package=BaBooN*

Nathan, G. & Holt, D. (1980), 'The effect of survey design on regression analysis', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 377–386.

Ngaruye, I., von Rosen, D. & Singull, M. (2017), 'Small area estimation with missing data using a multivariate linear random effects model'.

Pessoa, D. G. C. & Silva, P. L. N. (1998), 'Análise de dados amostrais complexos', *São Paulo: Associação Brasileira de Estatística* **112**.

Pfeffermann, D. (2013), 'New Important Developments in Small Area Estimation', *Statist. Sci.* **28**(1), 40–68.
**URL:** *https://doi.org/10.1214/12-STS395*

Prasad, N. & Rao, J. N. (1990), 'The estimation of the mean squared error of small-area estimators', *Journal of the American statistical association* **85**(409), 163–171.

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
**URL:** *https://www.R-project.org/*

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. & Solenberger, P. (2001), 'A multivariate technique for multiply imputing missing values using a sequence of regression models', *Survey methodology* **27**(1), 85–96.

Rao, J. (2005), 'Inferential issues in small area estimation: some new developments', *Statistics in Transition* **7**(3), 513–526.

Rao, J. (2008), 'Some methods for small area estimation', *Rivista Internazionale di Scienze Sociali* pp. 387–406.

Rao, J. N. & Molina, I. (2015), *Small area estimation*, John Wiley & Sons.

Robinson, P. & Särndal, C. E. (1983), 'Asymptotic properties of the generalized regression estimator in probability sampling', *Sankhyā: The Indian Journal of Statistics, Series B* pp. 240–248.

Rubin, D. (1987), 'Multiple imputation for nonresponse in surveys'.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* pp. 581–592.

Rubin, D. B. (1986), 'Statistical matching using file concatenation with adjusted weights and multiple imputations', *Journal of Business & Economic Statistics* **4**(1), 87–94.

Särdnal, C., Swensson, B. & Wretman, J. (1992), *Model assisted survey sampling*, Springer-Verlag.

Särndal, C. E. (1982), 'Implications of survey design for generalized regression estimation of linear functions', *Journal of Statistical Planning and Inference* **7**(2), 155–170.

Schomaker, M. & Heumann, C. (2016), 'Bootstrap inference when using multiple imputation', *arXiv preprint arXiv:1602.07933* .

Skinner, C. J. (1989), 'Domain means, regression and multi-variate analysis'.

Slud, E. V. & Maiti, T. (2011), 'Small-area estimation based on survey data from a left-censored Fay–Herriot model', *Journal of Statistical Planning and Inference* **141**(11), 3520–3535.

Van Buuren, Stef and Groothuis-Oudshoorn, Karin (2011), 'mice: Multivariate Imputation by Chained Equations in R', *Journal of Statistical Software* **45**(3), 1–67.
**URL:** *http://www.jstatsoft.org/v45/i03/*

Zhang, P. (2003), 'Multiple imputation: theory and method', *International Statistical Review* **71**(3), 581–592.