UNIVERSIDAD NACIONAL DE COLOMBIA

Comparación de metodologías utilizadas para abordar el problema de datos faltantes en estudios longitudinales

Andrés Felipe Viloria Rodriguez

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2024

Comparación de metodologías utilizadas para abordar el problema de datos faltantes en estudios longitudinales

Andrés Felipe Viloria Rodriguez

Trabajo final de maestría presentado como requisito parcial para optar al título de:
Magister en Ciencias - Estadística

Director:
Mauricio Alejandro Mazo Lopera
Doctor en Ciencias - Estadística

Línea:
Profundización

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2024

El aleteo de las alas de una mariposa puede provocar un huracán al otro lado del mundo.

Teoria del Caos

Honra a tu padre y a tu madre, que es el primer mandamiento con promesa.

(Efesios. 6:1-2)

Agradecimientos

Quiero expresar un reconocimiento especial a toda mi familia, quienes me brindaron apoyo en todos los aspectos durante esta aventura académica. También deseo agradecer a Jehová Dios y a su Espíritu Santo por su constante guía y bendiciones.

A la escuela de estadística de Universidad Nacional de Colombia sede Medellín y a sus profesores, que con su formación logró formar un profesional con las capacidades para lograr la meta planteada en el presente trabajo. En especial quiero dar un reconocimiento al profesor Mauricio Alejandro Mazo Lopera quien me acompaño y asesoró en el camino de obtener este producto.

Resumen

En este estudio se compararon distintas metodologías utilizadas en la literatura para abordar el problema de datos faltantes en estudios longitudinales. Se consideraron tres condiciones propuestas por D. Rubin (1976) para la generación de datos faltantes: Perdidos Completamente al Azar (MCAR), Perdidos al Azar (MAR) y No Perdidos al Azar (NMAR). Se utilizaron dos bases de datos longitudinales provenientes del Portal de Datos Abiertos del Estado Colombiano, las cuales fueron ajustadas mediante modelos lineales mixtos. Posteriormente, se generaron datos faltantes bajo las tres condiciones antes mencionadas y se aplicaron diferentes métodos de imputación. Se compararon los modelos imputados utilizando el Root Mean Squared Error (RMSE) y se observó que el método Last Observation Carried Forward (LOCF) tuvo un mejor rendimiento en la mayoría de los casos. Además, se analizó la variabilidad en la precisión del modelo por departamento y se encontraron diferencias significativas entre los métodos de imputación. Se concluyó que la elección del método de imputación puede tener un impacto en la interpretación de los resultados del modelo y se hicieron recomendaciones para futuras investigaciones, como explorar otros métodos de imputación y considerar el impacto de la imputación en la precisión de las predicciones del modelo en estudios longitudinales. En resumen, este estudio destaca la importancia de abordar cuidadosamente el problema de datos faltantes y seleccionar el método de imputación más adecuado para obtener resultados precisos y fiables en estudios longitudinales.

Palabras clave: Datos faltantes, Estudios longitudinales, Modelos lineales mixtos, Metodologías de imputación, Condiciones de datos faltantes.

Comparison of Methodologies Used to Address the Problem of Missing Data in Longitudinal Studies

Abstract

This study compared different methodologies used in the literature to address the problem of missing data in longitudinal studies. Three conditions proposed by Šciterubin1976 for missing data generation were considered: Missing Completely At Random (MCAR), Missing At Random (MAR), and Not Missing At Random (NMAR). Two longitudinal databases from the Colombian State's Open Data Portal were used, which were adjusted using linear mixed models. Subsequently, missing data were generated under the three aforementioned conditions and different imputation methods were applied. The imputed models were compared using the Root Mean Squared Error (RMSE) and it was observed that the Last Observation Carried Forward method (LOCF) performed better in most cases. In addition, the variability in model accuracy by department was analyzed and significant differences were found between the imputation methods. It was concluded that the choice of imputation method may have an impact on the interpretation of model results and recommendations for future research were made, such as exploring other imputation methods and considering the impact of imputation on the accuracy of model predictions in longitudinal studies. In summary, this study highlights the importance of carefully addressing the missing data problem and selecting the most appropriate imputation method to obtain accurate and reliable results in longitudinal studies.

Keywords: Missing data, Longitudinal studies, Mixed linear models, Imputation methodologies, Missing data conditions

Contenido

Agradecimientos vii

Resumen ix

Lista de figuras xii

Lista de tablas xiii

1 Introducción 1

2 Datos longitudiales 5

2.1 Notacion . . . 5
2.2 Modelos lineales . . . 8
2.3 Modelos lineales mixtos (LMM) . . . 10
2.3.1 Estimación en Modelos Lineales Mixtos . . . 12
2.3.2 Estimación de máxima verosimilitud (MLE) . . . 13
2.3.3 Caso especial θ es conocido . . . 13
2.3.4 Caso general θ es desconocido . . . 14
2.3.5 Estimación REML . . . 15
2.3.6 Predicciones de los modelos lineales mixtos . . . 16
2.3.7 Error Cuadrático Medio (RMSE) . . . 16
2.4 Datos faltantes en estudios longitudinales . . . 17
2.4.1 Mecanismos de datos faltantes . . . 17
2.4.2 Implicaciones de los datos faltantes . . . 19
2.5 Métodos para el tratamiento de datos faltantes en estudios longitudinales . . . 20
2.5.1 Análisis de casos completos . . . 20
2.5.2 Análisis de datos disponibles . . . 21
2.5.3 Imputación . . . 21

3 Estudios de caso 27

3.1 Rendimiento de cultivos en Colombia por año . . . 27
3.1.1 Propuesta de un Modelo Lineal Mixto para la Base de Datos originales 28
3.1.2 RMSE para los datos originales . . . 31
3.1.3 Generación de Datos Faltantes bajo las Condiciones de D. Rubin (1976) 32
3.1.4 Tratamiento para la Condición MCAR . . . 33

3.1.5 Tratamiento para la condición MAR 40
3.1.6 Tratamiento para la condición MNAR 45
3.1.7 Discusión general para las tres condiciones 51
3.2 Producción de Aguacate 52
3.2.1 Propuesta de un Modelo Lineal Mixto para la Base de Datos originales 53
3.2.2 RMSE para los datos originales 56
3.2.3 Generación de Datos Faltantes bajo las Condiciones de D. Rubin (1976) 57
3.2.4 Tratamiento para la Condición MCAR 58
3.2.5 Tratamiento para la condición MAR 65
3.2.6 Tratamiento para la condición MNAR 70
3.2.7 Discusión general para las tres condiciones 77

4 Conclusiones y recomendaciones 79
4.1 Conclusiones 79
4.2 Recomendaciones 80

Lista de Figuras

3-1 Comportamiento de los departamentos por año 28
3-2 Datos Originales vs. Predicciones 31
3-3 Porcentaje de datos faltantes por departamento para la condición MCAR . . 33
3-4 Porcenta de falta por departamente para la condicinon MAR 40
3-5 Porcenta de falta por departamente para la condicinon MNAR 45
3-6 Comparación de RMSE de Métodos de Imputación 52
3-7 Comportamiento de los departamentos por año 53
3-8 Datos Originales vs. Predicciones 56
3-9 Porcentaje de datos faltantes por departamento para la condición MCAR . . 58
3-10 Porcenta de falta por departamente para la condicinon MAR 65
3-11 Porcenta de falta por departamente para la condicinon MNAR 71
3-12 Comparación de RMSE de Métodos de Imputación 78

Lista de Tablas

3-1 Descripción de efectos y parámetros del modelo 29
3-2 Valores de intercepto aleatorio por Departamento 30
3-3 RMSE por Departamento 32
3-4 RMSE_MCAR_locf por Departamento 35
3-5 RMSE_MCAR_media por Departamento 36
3-6 RMSE_MCAR_regresión por Departamento 37
3-7 RMSE_MCAR_match por Departamento 38
3-8 Comparacion de los RMSE para la condicion MCAR de métodos de imputacion 38
3-9 Comparacion de los RMSE_MCAR de los métodos de imputacion por departamentos que presentan datos faltantes 39
3-10 Comparacion de los RMSE_MCAR de los métodos de imputacion por departamentos que no presentan datos faltante 39
3-11 RMSE_MAR_locf por Departamento 41
3-12 RMSE_MAR_media por Departamento 42
3-13 RMSE_MAR_regresión por Departamento 43
3-14 RMSE_MAR_match por Departamento 44
3-15 Comparacion de los RMSE para la condicion MAR de métodos de imputacion 44
3-16 RMSE_MNAR_locf por Departamento 46
3-17 RMSE_MNAR_media por Departamento 47
3-18 RMSE_MNAR_regresión por Departamento 48
3-19 RMSE_MNAR_match por Departamento 49
3-20 Comparacion de los RMSE para la condicion MNAR de métodos de imputacion 50
3-21 Comparacion de los RMSE_MNAR de los métodos de imputacion por departamentos que presentan datos faltantes 50
3-22 Comparacion de los RMSE_MNAR de los métodos de imputacion por departamentos que no presentan datos faltante 51
3-23 Descripción de efectos y parámetros del modelo 54
3-24 Valores de intercepto aleatorio por Departamento 55
3-25 RMSE por departamento 57
3-26 RMSE_MCAR_locf por Departamento 60
3-27 RMSE_MCAR_media por Departamento 61
3-28 RMSE_MCAR_regresión por Departament 62
3-29 RMSE_MCAR_match por Departamento 63

3-30 Comparacion de los RMSE para la condicion MCAR de métodos de imputacion 63
3-31 Comparacion de los RMSE_MCAR de los métodos de imputacion por departamentos que presentan datos faltantes 64
3-32 Comparacion de los RMSE_MCAR de los métodos de imputacion por departamentos que no presentan datos faltante 64
3-33 RMSE_MAR_locf por Departamento 66
3-34 RMSE_MAR_media por Departamento 67
3-35 RMSE_MAR_regresión por Departamento 68
3-36 RMSE_MAR_match por Departamento 69
3-37 Comparacion de los RMSE para la condicion MAR de métodos de imputacion 70
3-38 RMSE_MNAR_locf por Departamento 72
3-39 RMSE_MNAR_media por Departamento 73
3-40 RMSE_MNAR_regresión por Departamento 74
3-41 RMSE_MNAR_match por Departamento 75
3-42 Comparacion de los RMSE para la condicion MNAR de métodos de imputacion 76
3-43 Comparacion de los RMSE_MNAR de los métodos de imputacion por departamentos que presentan datos faltantes 76
3-44 Comparacion de los RMSE_MNAR de los métodos de imputacion por departamentos que no presentan datos faltante 77

1 Introducción

La investigación longitudinal se configura para recopilar datos en distintos momentos durante un periodo determinado, siendo común que algunos individuos no asistan en ciertos momentos de medición o abandonen el estudio antes de su conclusión, generando así datos faltantes. Es crucial comprender las razones por las cuales los participantes no participan en ciertos momentos de medición o abandonan el estudio. Según D. Rubin (1976), se pueden identificar tres mecanismos clave para explicar la generación de datos faltantes.

El primero es el de "Desaparecidos Completamente al Azar (MCAR, por sus siglas en inglés)", donde la probabilidad de que falte una medición no está vinculada con ninguna otra variable, ya sea que esté incluida o no en los datos. El segundo es el de "Desaparecidos al Azar (MAR, por sus siglas en inglés)", que implica que la probabilidad de falta de medición se relaciona con otras variables en los datos, pero no con las variables no incluidas en el conjunto de datos o con la propia variable que carece de valores. El tercer mecanismo es "No Desaparecidos al Azar (NMAR)", donde la probabilidad de que falten valores en una variable está relacionada con las variables no incluidas en los datos o con los valores hipotéticos de la variable de interés si no faltaran valores (Laird, Ware, y Fitzmaurice, 2004).

Según Laird, Ware, y Fitzmaurice (2004), el mecanismo MCAR se considera una causa de pérdida de datos ignorables, ya que cualquier método de análisis que genere inferencias válidas para casos completos puede utilizarse. En algunos casos, el mecanismo MAR también se considera una causa ignorable, pero esto no implica que cualquier método de análisis pueda aplicarse. La ignorabilidad se refiere a la obtención de un análisis válido basado en el método de máxima verosimilitud siempre que se disponga de un modelo correcto. En cambio, el mecanismo NMAR se considera una causa no ignorable, lo que significa que no se pueden pasar por alto las faltas de medición si se busca realizar inferencias válidas sobre la distribución de las mediciones completas.

Varios estudios han abordado la cuestión de los datos faltantes, explorando los métodos más efectivos. Nooraee, Molenberghs, Ormel, y Edwin R (2018) señalan las deficiencias de los métodos de sustitución simple y se centran en enfoques avanzados como la imputación múltiple (MI, por sus siglas en inglés), la máxima verosimilitud de información completa (FIML, por sus siglas en inglés) y los métodos de ponderación de probabilidad inversa. Otros investigadores (Schafer y Graham, 2002; Van Buuren, 2012; van Ginkel, Linting, Rippe, y van der Voort, 2020) sostienen que la imputación múltiple supera a los métodos de eliminación por lista y a la imputación única para abordar los datos faltantes en estudios longitudinales. En este estudio, nos proponemos comparar estas diversas metodologías para

enfrentar el problema de los datos faltantes en estudios longitudinales.

El desafío de los datos faltantes es recurrente en estudios longitudinales, y autores como Diggle, Heagerty, Liang, y Zeger (2002), Laird y cols. (2004), y Fitzmaurice, Davidian, Verbeke, y Molenberghs (2008) han proporcionado una conceptualización detallada del problema, junto con metodologías que permiten obtener inferencias válidas para cada mecanismo subyacente en la pérdida de datos. Además, se encuentran artículos relevantes, como los de Troxel, Harrington, y Lipsitz (1998), Diggle y Kenward (1994), y Roderick y Little (1995), que proponen métodos específicos para abordar diferentes patrones de datos faltantes, ya sea monótonos, de abandono o no monótonos. Por otro lado, Ibrahim, Chen, Lipsitz, y Herring (2005) ofrecen una descripción detallada y comparativa de cuatro paradigmas para el manejo de datos faltantes: máxima verosimilitud (ML, por sus siglas en inglés), imputación múltiple (MI, por sus siglas en inglés), métodos bayesianos y ecuaciones de estimación ponderadas (WEE, por sus siglas en inglés).

Entre las conclusiones de van Ginkel y cols. (2020) y McNeish (2017), destaca que el método más sofisticado y efectivo para abordar el problema de datos faltantes es la imputación múltiple. Contrariamente, Rahim, Chen, y Lipsitz (2001) proponen un enfoque alternativo centrado en la estimación de parámetros en modelos mixtos lineales generalizados (GLMM, por sus siglas en inglés) mediante un modelo de selección que tiene en cuenta datos faltantes no ignorables. Estas perspectivas divergentes ofrecen diversas opciones para abordar la problemática de datos faltantes en estudios longitudinales. Los modelos lineales mixtos (LMM, por sus siglas en inglés) surgen como una herramienta flexible para abordar la correlación entre observaciones en datos longitudinales. Estos modelos, que integran tanto efectos fijos como aleatorios, permiten la estimación de componentes de varianza mediante el método de Máxima Verosimilitud Restringida (REML, por sus siglas en inglés) (Laird y Ware, 1982). Es esencial considerar tanto la variabilidad intra-sujeto como la variabilidad entre sujetos para comprender la respuesta individual en un contexto longitudinal (Faraway, 2016). En resumen, la elección de la metodología para abordar los datos faltantes en estudios longitudinales debe basarse en una comprensión detallada de la naturaleza y el contexto específico de los datos en cuestión.

En este trabajo, se llevará a cabo una revisión literaria inicial con el objetivo de explorar las diversas metodologías propuestas por distintos autores para abordar los diferentes mecanismos subyacentes a la pérdida de datos en estudios longitudinales. Se realizará una exhaustiva búsqueda de artículos, libros y otras publicaciones que ofrezcan enfoques y estrategias para el manejo de datos faltantes en este contexto específico.

Esta revisión no solo se centrará en identificar las metodologías propuestas, sino también en comprender las consecuencias que los datos faltantes pueden tener en los estudios longitudinales. Se explorarán las implicaciones de estos vacíos en la calidad y validez de las inferencias derivadas de dichos estudios.

Posteriormente, se llevará a cabo la aplicación de algunas de las metodologías identificadas en la literatura, con el propósito de realizar una comparación detallada entre ellas. Este

análisis comparativo permitirá evaluar la eficacia y las limitaciones de cada enfoque en la gestión de datos faltantes en estudios longitudinales.

Finalmente, con base en los resultados obtenidos de la revisión y la comparación, se propondrá un método sistemático para evaluar y comparar las distintas alternativas identificadas en la literatura para abordar el problema de datos faltantes en estudios longitudinales. Este enfoque de comparación proporcionará una guía práctica y fundamentada para investigadores y profesionales que se enfrentan a esta problemática en sus estudios.

2 Datos longitudiales

Los estudios longitudinales, centrados en la toma repetida de mediciones en los mismos individuos a lo largo del tiempo, permiten observar y analizar los cambios en las respuestas individuales a lo largo de la duración del estudio. La característica distintiva de estos estudios es la medición repetida de la variable de interés en los participantes, lo que facilita la evaluación directa de los cambios temporales.

El análisis longitudinal se enfoca en evaluar los cambios intraindividuales en la respuesta a lo largo del tiempo, dividiéndose en dos etapas: descripción de los cambios intraindividuales y su relación con las diferencias interindividuales en covariables seleccionadas. Este diseño elimina la influencia de factores constantes en el tiempo, proporcionando estimaciones precisas del cambio intraindividual y permitiendo predicciones más confiables(D. Rubin, 1976).

En un estudio longitudinal, los participantes, ya sean humanos o animales, se someten a mediciones repetidas en diferentes momentos. La variabilidad en el número de observaciones y la presencia de datos faltantes son comunes, lo que puede afectar la interpretación de los cambios intraindividuales. La incompletitud de los datos tiene implicaciones para el análisis, afectando la eficiencia y precisión de las estimaciones de cambios a lo largo del tiempo.

Los datos longitudinales, donde el resultado se observa repetidamente en diferentes momentos para cada sujeto, pueden modelarse con enfoques específicos de sujetos o modelos de población media. La elección depende de los objetivos del estudio. Para datos correlacionados, como los longitudinales, los modelos lineales de efectos mixtos se presentan como una herramienta flexible y versátil para modelar estas relaciones, basándose en la función de verosimilitud (Fitzmaurice y cols., 2008).

Los estudios longitudinales son por tanto fundamentales para comprender las variaciones individuales a lo largo del tiempo, pero la gestión cuidadosa de la incompletitud de datos es esencial para obtener interpretaciones precisas y confiables.

2.1. Notacion

Sea Y_{ij} la variable de respuesta para el i-ésimo individuo ($i = 1, \ldots, n$) en la j-ésima ocasión ($j = 1, \ldots, n_i$). Si se supone que las medidas repetidas están igualmente separadas en el tiempo, esta notación será suficiente. Dado que tenemos n_i medidas repetidas de la variable de respuesta en el mismo individuo, podemos representarlo en un vector de respuesta $n_i \times 1$, denotado por:

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}, \quad \text{para} \quad i = 1, 2, \ldots, n.$$

En el contexto de datos longitudinales, el vector $\boldsymbol{Y}_i$ puede ser visto como una colección ordenada en el tiempo de las mediciones de variables aleatoria Y para el i-ésimo sujeto.
Se puede denotar la media de cada respuesta de Y_{ij}, fijando un momento j, por:

$$\mu_j = E(Y_{ij}),$$

Donde $E(\cdot)$ es la media a largo plazo de Y_{ij}, es decir, es la media ponderada de todos los valores posibles de Y_{ij}, siendo asi las ponderaciones las probabilidades de ocurrenca de cada valor posible. Ahora, nuestra discusión sobre la media de Y_{ij} ha asumido que la respuesta media puede cambiar con el tiempo, lo cual se refleja en nuestro uso de un subíndice de una sola letra para la media, μ_i. Como hemos venido diciendo en los estudios longitudinales, el objetivo principal es establecer relaciones entre los cambios en la respuesta media a lo largo del tiempo y las covariables. Para permitir que la respuesta media, y en particular los cambios en la respuesta media, varíen entre individuos, como función de las covariables a nivel individual, es necesario emplear subíndices de dos letras:

$$\mu_{ij} = E(Y_{ij}).$$

En esta notación, la respuesta media puede cambiar con el tiempo (denotada por la dependencia de μ_{ij} del subíndice j) y los cambios en la respuesta media pueden relacionarse con covariables a nivel individual (denotadas por la dependencia de μ_{ij} del subíndice i).
Ahora introduciremos las notaciones de varianza y covarianza para llegar a la definicion de carrelacion, entonces la varianza condicional de Y_{ij} se define como:

$$\sigma_j^2 = E\left[(Y_{ij} - \mu_{ij})^2\right]$$

Mientras que μ_{ij} proporciona una medida de la ubicación del centro de la distribución de Y_{ij}, la varianza condicional proporciona una medida de la dispersión de los valores de Y_{ij} en torno a su media condicional. La raíz cuadrada positiva de la varianza condicional, σ_j, se conoce como desviación típica condicional.
A continuación consideramos la dependencia entre las respuestas en estudios longitudinales. La covarianza condicional entre las respuestas en dos ocasiones diferentes, digamos Y_{ij} e Y_{ik}, se denota por:

$$\sigma_{ik} = E\left[(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\right],$$

y proporciona una medida de la dependencia lineal entre Y_{ij} e Y_{ik}, dadas las covariables.

La covarianza entre Y_{ij} e Y_{ik} puede tomar valores tanto positivos como negativos. Cuando la covarianza es cero, no existe dependencia lineal entre las respuestas en las dos ocasiones (dadas las covariables). La magnitud de la covarianza depende no sólo del grado de dependencia entre las dos variables, sino también de sus unidades de medida. Cabe señalar que la covarianza de una variable consigo misma es simplemente la varianza de la variable. Mientras que el signo de la covarianza indica si existe dependencia positiva o negativa entre las dos variables. Por ejemplo, si $\sigma_{jk} = 10$, esta información por sí sola indica que existe dependencia entre Y_{ij} e Y_{ik} y que la dependencia es positiva, es decir, Y_{ij} aumenta a medida que Y_{ik} aumenta, y viceversa. Para proporcionar una medida de dependencia lineal entre Y_{ij} e Y_{ik} que esté en cierto sentido libre de las unidades de medida (o variabilidad) de las dos variables, se utiliza ampliamente la correlación.

La correlación condicional entre Y_{ij} e Y_{ij} se denota:

$$\rho_{jk} = \frac{E\left[(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\right]}{\sigma_j \sigma_k},$$

donde σ_j y σ_k son las desviaciones estándar condicionales de Y_{ij} e Y_{ik}, respectivamente. La correlación, a diferencia de la covarianza, es una medida de dependencia sin unidades o libre de las escalas de medición de Y_{ij} e Y_{ik}, lo que se consigue dividiendo cada variable por su respectiva desviación estándar. Esto hace que sea una medida más fácilmente interpretable de la dependencia lineal entre dos variables. Obsérvese que cuando la covarianza es cero, también lo es la correlación. La correlación debe tomar valores entre -1 y 1. Recordemos que se obtiene una correlación de 1 o -1 cuando existe una relación lineal perfecta entre las dos variables. Una correlación positiva implica que una variable aumenta a medida que aumenta la otra. Aunque dos variables que son estadísticamente independientes entre sí no estarán necesariamente correlacionadas, las variables pueden no estar correlacionadas sin ser independientes, ya que la correlación sólo mide la dependencia lineal. En los datos longitudinales, se antevé que las medidas repetidas del mismo individuo estén correlacionadas positivamente. Cuando las n_i medidas repetidas se recogen en un vector Y_{ij}, podemos definir la matriz de varianza-covarianza como la siguiente matriz bidimensional de varianzas y covarianzas condicionales:

$$\text{cov}(\boldsymbol{Y}_i) = \text{Cov}\left[\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}\right] = \begin{pmatrix} \text{Var}(Y_{i1}) & \text{Cov}(Y_{i1}, Y_{i2}) & \cdots & \text{Cov}(Y_{i1}, Y_{in_i}) \\ \text{Cov}(Y_{i2}, Y_{i1}) & \text{Var}(Y_{i2}) & \cdots & \text{Cov}(Y_{i2}, Y_{in_i}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_{in_i}, Y_{i1}) & \text{Cov}(Y_{in_i}, Y_{i2}) & \cdots & \text{Var}(Y_{in_i}) \end{pmatrix}$$
$$= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n_i} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n_i1} & \sigma_{n_i2} & \cdots & \sigma_{n_i}^2 \end{pmatrix}$$

Donde $\text{Cov}(Y_{ij}, Y_{ik}) = \sigma_{jk}$. Obsérvese que esta matriz es simétrica en el sentido de que $\text{Cov}(Y_{ij}, Y_{ik}) = \sigma_{jk} = \sigma_{kj} = \text{Cov}(Y_k, Y_1)$. Recordemos también que la covarianza de una variable consigo misma es la varianza. Así, podemos denotar $\sigma_{kk} = \text{Cov}(Y_{ik}, Y_{ik}) = \text{Var}(Y_{ik}) = \sigma_k^2$. También podemos definir la matrix de correlación, $\text{corr}(\boldsymbol{Y}_i)$, en términos de una matriz bidimensional similar:

$$\text{corr}(\boldsymbol{Y}_i) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n_i} \\ \rho_{21} & 1 & \cdots & \rho_{2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n_i1} & \rho_{n_i2} & \cdots & 1 \end{pmatrix}$$

Se dice que esta matriz es simétrica en el sentido de que $\text{Corr}(Y_{ij}, Y_{ik}) = \rho_{jk} = \rho_{kj} = \text{Corr}(Y_{ik}, Y_{ij})$. Los elementos diagonales de la matriz son todos iguales a 1, ya que denotan la correlación de una variable consigo misma.

2.2. Modelos lineales

Los modelos de lineal para datos longitudinales generalmente pueden formularse para encapsular las principales preguntas de investigación de interés en términos de un conjunto de parámetros de regresión. Antes de discutir los enfoques para modelar la respuesta media a lo largo del tiempo, es importante aclarar la distinción entre parámetros sustantivos y molestos en el contexto de un estudio longitudinal.
A Continuación, consideremos un conjunto de k covariables que permitan explicar o predecir a $\boldsymbol{Y}_i$, este conjunto sera un vector de covariables $\boldsymbol{X}_{ij}$ de dimensiones $k \times 1$ denotado por:

$$\boldsymbol{X}_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijk} \end{pmatrix}, \quad i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, n_i.$$

Esta notación permite observar que $\boldsymbol{X}_{ij}$ representa un vector de covariables asociadas con $\boldsymbol{Y}_{ij}$ en la j-ésima medición del i-ésimo sujeto, y las k filas de $\boldsymbol{X}_{ij}$ corresponden a las diferentes covariables.
Ahora, estos vectores los podemos agrupar en una matriz $n_i \times k$:

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{X}_{i1}^\top \\ \boldsymbol{X}_{i2}^\top \\ \vdots \\ \boldsymbol{X}_{in_i}^\top \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1k} \\ X_{i21} & X_{i22} & \cdots & X_{i2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ik} \end{pmatrix},$$

Donde $\boldsymbol{X}_{ij}^\top$ es la transpuesta del vector de covariable $\boldsymbol{X}_{ij}$. Por tanto, $\boldsymbol{X}_{ij}^\top$ es un vector de $1 \times k$ de covariables para el i-ésimo sujeto en la j-ésima medición.
A continuación, se puede plantear un modelo de lineal que relacione a Y_{ij} con $\boldsymbol{X}_i$ y además para ver los cambios en la respuesta media de las mediciones de $\boldsymbol{Y}_i$ a lo largo del tiempo y relacionemos esos cambios con $\boldsymbol{X}_i$:

$$Y_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_k X_{ijk} + e_{ij}; j = 1, ..., n_i; \tag{2-1}$$

Donde $\beta_1, \cdots, \beta_k$ son los paramentros que relacionan la media de Y_{ij} con sus covariables correspondientes. Este modelo describe cómo se relacionan la variable de interés en cada ocasión de medición con las covariables. Es decir, hay n_i ecuaciones separadas para la variable de respuesta en cada una de las n_i ocasiones. Por tanto:

$$\begin{aligned} Y_{i1} &= \beta_1 X_{i11} + \beta_2 X_{i12} + \cdots + \beta_k X_{i1k} + e_{i1} = \boldsymbol{X}_{i1}^\top \boldsymbol{\beta} + e_{i1}, \\ Y_{i2} &= \beta_1 X_{i21} + \beta_2 X_{i22} + \ldots + \beta_k X_{i2k} + e_{i2} = \boldsymbol{X}_{i2}^\top \boldsymbol{\beta} + e_{i2}, \\ &\vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots \\ Y_{in_i} &= \beta_1 X_{in_i1} + \beta_2 X_{in_i2} + \cdots + \beta_k X_{in_ik} + e_{in_i} = \boldsymbol{X}_{in_i}^\top \boldsymbol{\beta} + e_{in_i}, \end{aligned} \tag{2-2}$$

Donde los parametros desconocidos se agrupan en un vector transpuesto $k \times 1$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_k)^\top$ y e_{ij} son errores aleatorios con media cero que representan desviaciones de las mediciones de la variable de interés respecto a sus correspondientes medias previstas, dadas por:

$$E(Y_{ij} \mid \boldsymbol{X}_i) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijk}$$

Normalmente, West, Welch, y Galecki (2022) plantean que $\boldsymbol{X}_{ij} = 1$ para todos i y j, mientras que β_i representa el término de intercepción del modelo.
Utilizando la notación vectorial y matricial dada por (2-1) o (2-2), podemos reescribir el modelo de regresión de manera más compacta:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{e}_i, \tag{2-3}$$

Que es lo mismo que:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1k} \\ X_{i21} & X_{i22} & \cdots & X_{i2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ik} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ik} \end{pmatrix}$$

Donde los errores aleatorios se agrupan en un vector transpuesto $n_i \times 1$, $\boldsymbol{e}_i = (e_{i1}, e_{i2}, ..., e_{in_i})^\top$. Estos errores aleatorios están asociados a los elementos correspondientes del vector de la variable de interes del i-ésimo sujeto.

2.3. Modelos lineales mixtos (LMM)

Los modelos lineales mixtos fueron propuestos por Laird y Ware (1982). El nombre de modelos lineales mixtos proviene del hecho de que estos modelos son lineales en los parámetros y que las covariables o variables independientes pueden involucrar efectos fijos y efectos aleatorios. Consideremos el vector de respuestas del i-ésimo individuo, $\boldsymbol{Y}_i$ con n_i observaciones. Un modelo lineal mixto individual está dado por:

$$\begin{aligned} \boldsymbol{Y}_i &= \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i \\ \boldsymbol{b}_i &\sim N(\boldsymbol{0}, \boldsymbol{D}) \\ \boldsymbol{e}_i &\sim N(\boldsymbol{0}, \boldsymbol{R}_i) \end{aligned}$$

donde los parámetros $\boldsymbol{\beta}$ son un vector $k \times 1$ de efectos fijos, $\boldsymbol{b}_i$ es un vector $q \times 1$ de efectos aleatorios, $\boldsymbol{X}_i$ es una matriz de $n_i \times k$ de covariables, y $\boldsymbol{Z}_i$ es una matriz $n_i \times q$, con $q \leqslant k$. Aquí, $\boldsymbol{Z}_i$ es una matriz de diseño conocida que vincula el vector de efectos aleatorios $\boldsymbol{b}_i$ con $\boldsymbol{Y}_i$; En particular, para muchos modelos de análisis longitudinal, las columnas de $\boldsymbol{Z}_i$ son un subconjunto de las columnas de $\boldsymbol{X}_i$.
El vector $\boldsymbol{b}_i = (b_{i1}, b_{i2}, \cdots, b_{iq})^\top$ para el i-ésimo sujeto representa un vector de orden q de efectos aleatorios asociados con las q covariables de la matriz $\boldsymbol{Z}_i$. Suponemos que los efectos de orden aleatorio en el vector $\boldsymbol{b}_i$ siguen una distribución normal multivariada, con vector de medias $\boldsymbol{0}$ y una matriz de varianzas-covarianzas denotada por $\boldsymbol{D}$. Los elementos a lo largo de la diagonal principal de la matriz $\boldsymbol{D}$ representan las variaciones de cada efecto aleatorio en $\boldsymbol{b}_i$, y los elementos fuera de la diagonal representan las covarianzas entre dos efectos aleatorios correspondientes. Debido a que hay q efectos aleatorios en el modelo asociado con el i-ésimo sujeto, $\boldsymbol{D}$ es una matriz $q \times q$ simétrica y positiva. Los elementos de esta matriz se muestran a continuación:

$$\boldsymbol{D} = \text{Var}(\boldsymbol{b}_i) = \begin{pmatrix} \text{Var}(b_{1i}) & \text{cov}(b_{1i}, b_{2i}) & \cdots & \text{cov}(b_{1i}, b_{qi}) \\ \text{cov}(b_{1i}, b_{2i}) & \text{Var}(b_{2i}) & \cdots & \text{cov}(b_{2i}, b_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(b_{1i}, b_{qi}) & \text{cov}(b_{2i}, b_{qi}) & \cdots & \text{Var}(b_{qi}) \end{pmatrix}$$

Los elementos (varianzas y covarianzas) de la matriz D se definen como funciones de un conjunto pequeño de parámetros de covarianza almacenados en un vector denotado por $\boldsymbol{D}$. Finalmente, el vector $\boldsymbol{e}_i$ es un vector de n_i residuales, con cada elemento en $\boldsymbol{e}_i$ que denota el residuo asociado con una respuesta observada en t o en n_i ocasiones para el i-ésimo sujeto. Debido a que algunos sujetos pueden tener más observaciones que otros, el vector $\boldsymbol{e}_i = (e_{i1}, e_{i2}, \cdots, e_{ini})^\top$ puede tener un número diferente de elementos. Supóngase que los errores n_i en el vector $\boldsymbol{e}_i$ para un sujeto dado, i, son variables aleatorias que siguen una distribución normal multivariada, con un vector de medias $\mathbf{0}$ y una matriz de varianza-covarianza definida positiva $\boldsymbol{R}_i$. Se asume que los errores asociados a diferentes sujetos son independientes entre sí, y que los vectores de errores $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_{n_i}$, y los efectos aleatorios $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_{n_i}$, son independientes. La forma general de la matriz $\boldsymbol{R}_i$ es:

$$\boldsymbol{R}_i = \text{Var}(\boldsymbol{e}_i) = \begin{pmatrix} \text{Var}(e_{1i}) & \text{cov}(e_{1i}, e_{2i}) & \cdots & \text{cov}(e_{1i}, e_{ni}) \\ \text{cov}(e_{1i}, e_{2i}) & \text{Var}(e_{2i}) & \cdots & \text{cov}(b_{2i}, e_{ni}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(e_{1i}, e_{ni}) & \text{cov}(e_{2i}, e_{ni}) & \cdots & \text{Var}(e_{ni}) \end{pmatrix}$$

Los elementos (varianzas y covarianzas) de la matriz $\boldsymbol{R}_i$ se definen como funciones de otro conjunto de parámetros de covarianza en un vector denotado por $\boldsymbol{\theta}_{R_i}$. Al completar la notación para el modelo lineal mixto, el vector $\boldsymbol{\theta}$ combina todos los parámetros de covarianza contenidos en los vectores $\boldsymbol{\theta}_D$ y $\boldsymbol{\theta}_{R_i}$. Correa Morales y Salazar Uribe (2016) Estos parámetros se conocen con el nombre técnico de componentes de varianza y covarianza.

Modelo de interceptos aleatorios

Para Correa Morales y Salazar Uribe (2016) este es uno de los modelo más usados en modelos lineales mixtos. Cada sujeto es aproximadamente paralelo a la tendencia de su grupo. El modelo se expresa de manera genérica así:

$$Y_{ij} = Tiempo + Grupo + (Grupo \times Sujetos) + Sujeto + error$$

Que en símbolos es:

$$Y_{ij} = \beta_0 + \beta_1 T_{ij} + \beta_e G_i \times T_{ij} + b_{0i} + e_{ij}$$

Donde:

- El sujeto es representado por $i : i = 1, 2, \ldots, n$.
- Las observaciones para el sujeto $i : j = 1, 2, \ldots, n_i$.
- $b_{0i} \sim N(0, \sigma_b^2)$
- $e_{ij} \sim N(0, \sigma_e^2)$

Modelo de interceptos y pendiente aleatorias

Para Correa Morales y Salazar Uribe (2016) esta es otra de las formulaciones más usada en modelos lineales mixtos. Cada sujeto es diferente de los demás en términos del intercepto y la pendiente de su grupo. El modelo se expresa así:

$$Y_{ij} = Tiempo + Grupo + (Grupo \times Sujetos) + Sujeto + (Sujeto \times Tiempo) + error$$

Que en símbolos es:

$$Y_{ij} = \beta_0 + \beta_1 T_{ij} + \beta_e G_i \times T_{ij} + b_{0i} + b_{1i} \times T_{ij} + e_{ij}$$

Donde:

$$e_{ij} \sim N(0, \sigma_e^2)$$

Y

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{b_0}^2 & \sigma_{b_0,b_1} \\ \sigma_{b_0,b_1} & \sigma_{b_1}^2 \end{bmatrix} \right)$$

2.3.1. Estimación en Modelos Lineales Mixtos

En el modelo lineal mixto, estimamos los parámetros de efectos fijos, $\boldsymbol{\beta}$, y los parámetros de covarianza, $\boldsymbol{\theta} = (\boldsymbol{\theta}_D, \boldsymbol{\theta}_{R_i})^\top$ para las matrices $\boldsymbol{D}$ y $\boldsymbol{R}_i$. Usualmente, esto se hace por medio de la estimación de máxima verosimilitud (MLE) y máxima verosimilitud restringida (REML).

2.3.2. Estimación de máxima verosimilitud (MLE)

En general, la estimación de máxima verosimilitud (MLE) es un método para obtener estimaciones de parámetros desconocidos mediante la optimización de una función de verosimilitud (Casella y Berger, 2002). Para aplicar la estimación de MLE, primero construimos la probabilidad como una función de los parámetros en el modelo especificado, basado en supuestos de distribución. Las estimaciones de máxima verosimilitud (MLE) de los parámetros son los valores de los argumentos que maximizan la función de probabilidad (es decir, los valores de los parámetros que hacen que los valores observados de la variable dependiente sean más probables, dados los supuestos de distribución). En el contexto de los LMM, construimos la función de probabilidad de β y θ haciendo referencia a la distribución marginal de la variable dependiente Y_i. La función de densidad de probabilidad normal multivariada correspondiente, $f(\boldsymbol{Y}_i|\boldsymbol{\beta},\boldsymbol{\theta})$, es:

$$f(\boldsymbol{Y}_i|\boldsymbol{\beta},\boldsymbol{\theta}) = (2\pi)^{-\frac{n_i}{2}} \det(\boldsymbol{V}_i)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right)$$

Donde *det* se define como el determinante de la matriz y $\boldsymbol{V}_i = \boldsymbol{Z}_i^\top \boldsymbol{D}\boldsymbol{Z}_i + \boldsymbol{R}_i$. Basado en la función de densidad de probabilidad y dado los datos observados $\boldsymbol{Y}_i = \boldsymbol{y}_i$, la contribución de la función de verosimilitud para el i-ésimo sujeto se define de la siguiente manera:

$$L_i(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y}_i) = (2\pi)^{-\frac{n_i}{2}} \det(\boldsymbol{V}_i)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right)$$

Escribimos la función de verosimilitud, $L(\boldsymbol{\beta},\boldsymbol{\theta})$ como el producto de las n contribuciones independientes para los individuos ($i = 1,\ldots,n$).

$$L(\boldsymbol{\beta},\theta) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta},\boldsymbol{\theta}) = \prod_{i=1}^{n} (2\pi)^{-\frac{n_i}{2}} \det(\boldsymbol{V}_i)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right)$$

La función correspondiente a la log-verosimilitud $l(\boldsymbol{\beta},\boldsymbol{\theta})$, se define como:

$$l(\boldsymbol{\beta},\boldsymbol{\theta}) = \ln L(\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{1}{2}n\times\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\ln(\det(\boldsymbol{V}_i)) - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$$

2.3.3. Caso especial θ es conocido

Consideramos un caso especial de estimación de ML para LMM, en el cual asumimos que $\boldsymbol{\theta}$, y por ende la matriz $\boldsymbol{V_i}$, son conocidos. Dado que asumimos que $\boldsymbol{\beta}$ es conocido, los únicos parámetros que estimamos son los efectos fijos, $\boldsymbol{\beta}$. La función log-verosimilitud, $l(\boldsymbol{\beta},\boldsymbol{\theta})$, se convierte así en una función de $\boldsymbol{\beta}$ y su optimización es equivalente a encontrar un mínimo de una función objetivo $q(\boldsymbol{\theta})$.

$$q(\boldsymbol{\theta}) = \frac{1}{2}\sum_i (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$$

Es importante destacar que la optimización de $q(\boldsymbol{\theta})$ con respecto a $\boldsymbol{\beta}$ se puede realizar aplicando el método de mínimos cuadrados generalizados (GLS). El valor de $\boldsymbol{\beta}$ se puede obtener analíticamente:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_i \boldsymbol{X}_i^T \boldsymbol{V}_{bi}^{-1} \boldsymbol{X}_i\right)^{-1} \sum_i \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{y}_i$$

Según Correa Morales y Salazar Uribe (2016), $\widehat{\boldsymbol{\beta}}$ tiene la propiedad estadística de ser el mejor estimador lineal insesgado BLUE.

2.3.4. Caso general θ es desconocido

Se considera la estimación de ML de los parámetros de covarianza, $\boldsymbol{\theta}$, y el valor de efectos fijo, $\boldsymbol{\beta}$, suponiendo que $\boldsymbol{\theta}$ es desconocido. Primero, para obtener estimaciones para los parámetros de covarianza en $\boldsymbol{\theta}$, construimos una función de log-verosimilitud $l_{ML}(\boldsymbol{\theta})$. La función $l_{ML}(\boldsymbol{\theta})$ se deriva de $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ reemplazando el parámetro $\boldsymbol{\beta}$ por la expresión que define $\widehat{\boldsymbol{\beta}}$. La función resultante es

$$l_{ML}(\boldsymbol{\theta}) = -\frac{1}{2} n \ln(2\pi) - \frac{1}{2}\sum_i \ln(\det(\boldsymbol{V}_i)) - \frac{1}{2}\sum_i \operatorname{tr}(\tau_i \boldsymbol{V}_i^{-1} \mathbf{r}_i)$$

donde,

$$\boldsymbol{r}_i = \boldsymbol{y}_i - \boldsymbol{X}_i \left(\left(\sum_i \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right)^{-1} \sum_i \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{y}_i\right)$$

Después de las estimaciones de ML de los parámetros de covarianza en $\boldsymbol{\theta}$ (y, en consecuencia, las estimaciones de las varianzas y covarianzas en r y $\boldsymbol{R}_i$) calculamos $\widehat{\boldsymbol{\beta}}$. Primero, reemplazamos las matrices $\boldsymbol{D}$ y $\boldsymbol{R}_i$ por sus estimaciones de ML $\widehat{D}$ y $\widehat{R_i}$, para calcular $\widehat{V_i}$, una estimación

$$\widehat{V_i} = \boldsymbol{Z}_i \widehat{D} \boldsymbol{Z}_i^T + \widehat{R_i}$$

Luego, usamos la fórmula generalizada de mínimos cuadrados para $\widehat{\boldsymbol{\beta}}$, con $\boldsymbol{V}_i$ reemplazado por su estimación para obtener $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_i \boldsymbol{X}_i^T \widehat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i \right)^{-1} \sum_i \boldsymbol{X}_i^T \widehat{\boldsymbol{V}}_i^{-1} \boldsymbol{y}_i$$

Reemplazamos $\boldsymbol{V}_i$ por su estimación $\widehat{V}_i$, decimos que $\widehat{\boldsymbol{\beta}}$ es el mejor estimador lineal insesgado empírico EBLUE de $\boldsymbol{\beta}$.
La varianza de $\widehat{\boldsymbol{\beta}}$, $\text{Var}(\widehat{\boldsymbol{\beta}})$, es una matriz de varianza-covarianza $p \times p$ calculada de la siguiente manera:

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = \left(\sum_i \boldsymbol{X}_i^T \widehat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i \right)^{-1}.$$

Según Verbeke, Molenberghs, y Verbeke (1997), Las estimaciones de ML de $\boldsymbol{\theta}$ están sesgadas porque no tienen en cuenta la pérdida de grados de libertad que resulta de estimar los parámetros de efectos fijos en $\boldsymbol{\beta}$. Una forma alternativa del método de máxima verosimilitud conocido como estimación REML se usa con frecuencia para eliminar el sesgo en las estimaciones de ML de los parámetros de covarianza. Discutimos la estimación REML en la siguiente subsección.

2.3.5. Estimación REML

La estimación REML es una forma alterna de estimar los parámetros de covarianza en $\boldsymbol{\theta}$. según Patterson y Thompson (1971) propusieron un método alternativo llamado Estimación por Máxima Verosimilitud Restringida (REML), que también se conoce como Máxima verosimilitud residual. Este procedimiento utiliza técnicas de estimación por máxima verosimilitud a una función de verosimilitud asociada a un conjunto de errores de contraste en vez de aquella asociada a las observaciones originales. Este procedimiento compensa la pérdida de grados de libertad que resulta de la estimación de los efectos fijos en $\boldsymbol{\beta}$ y produce estimaciones menos sesgadas de las componentes de varianza. El asunto relacionado con este sesgo es importante cuando el número de parámetros no es pequeño en relación al número total de observaciones (Correa Morales y Salazar Uribe, 2016). Las derivaciones más generales de REML están dadas por Harville (1977), Cooper y Thompson (1977), y Verbyla (1990). Las estimaciones REML de $\boldsymbol{\theta}$ se basan en la optimización de la siguiente función de verosimilitud REML:

$$\ell_{REML}(\boldsymbol{\theta}) = -\frac{1}{2}(n-p)\ln(2\pi) - \frac{1}{2}\sum_i \ln(\det(\boldsymbol{V}_i)) - \frac{1}{2}\sum_i \text{tr}(\tau_i \boldsymbol{V}_i^{-1}) - \frac{1}{2}\sum_i \ln(\det(\boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i))$$

2.3.6. Predicciones de los modelos lineales mixtos

Dado que el modelo lineal de efectos mixtos distingue explícitamente entre efectos fijos y aleatorios, podemos estimar ambos tipos de efectos. Como se señaló en la sección anterior con respecto a los efectos fijos, la predicción de los efectos aleatorios se traduce en el problema de predecir la media condicional de $\boldsymbol{b_i}$, dado el vector de respuestas $\boldsymbol{Y_i}$: $E(\boldsymbol{b_i}|\boldsymbol{Y_i})$. Bajo los supuestos del modelo lineal de efectos mixtos, $\boldsymbol{Y_i}$ y $\boldsymbol{b}_i$ tienen una distribución normal multivariante conjunta. Utilizando propiedades de la distribución normal multivariante, se puede demostrar que la media condicional de $\boldsymbol{b}_i$ dado $\boldsymbol{Y}_i$ (y $\widehat{\boldsymbol{\beta}}$) es

$$E(\boldsymbol{b}_i|\boldsymbol{Y}_i) = \widehat{G}\boldsymbol{Z}_i^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}})$$

donde $\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^T + \boldsymbol{R}_i$. Esto se conoce como el mejor predictor lineal insesgado (BLUP). Cuando los parámetros de covarianza desconocidos se sustituyen por sus estimaciones REML, el predictor resultante es

$$\widehat{\boldsymbol{b_i}} = \widehat{G}\boldsymbol{Z}_i^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}})$$

Dada la predicción de b_i, la respuesta predicha del sujeto i-ésimo viene dada por

$$\widehat{\boldsymbol{Y_i}} = \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\widehat{\boldsymbol{b_i}}$$

Estos cálculos permiten obtener predicciones de las respuestas $\widehat{\boldsymbol{Y_i}}$ considerando tanto la contribución de los efectos fijos como la variabilidad de los efectos aleatorios.

2.3.7. Error Cuadrático Medio (RMSE)

El Error Cuadrático Medio de la Raíz (RMSE, por sus siglas en inglés) es una medida comúnmente utilizada para evaluar la calidad de las predicciones en un modelo estadístico. En el contexto de un modelo lineal mixto, puedes calcular el RMSE comparando las predicciones $\widehat{Y}_i$ con los valores reales de la variable dependiente $\boldsymbol{Y}_i$.
La fórmula para el RMSE se define como la raíz cuadrada de la media de los errores cuadrados entre las predicciones y los valores reales:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\boldsymbol{Y_i}} - \boldsymbol{Y}_i)^2}$$

donde n es el número total de observaciones.

2.4. Datos faltantes en estudios longitudinales

Los datos faltantes son un problema común en estudios longitudinales. Recordemos que este tipo de estudio está diseñado para recolectar datos sobre cada individuo que participa en la investigación en diferentes momentos a lo largo del tiempo. Por lo tanto, es muy común que en momentos de medición algún individuo no asista a la toma de datos, lo que genera datos faltantes. Estas faltas de datos se pueden presentar dos tipos de patrones (monótono o abandono y no monótono). Monótono o abandono se presenta cuando un individuo falta o deja de asistir a partir de cierto momento en adelante. No monótono cuando un individuo falta o no asiste a un momento y luego asiste a uno o varios momentos de medición posteriores.
Ahora, es importante conocer el motivo por el cual no asisten, ya sea por motivos relacionados con el estudio o no. D. Rubin (1976) consideró tres mecanismos de datos faltantes para entender cómo pueden haber surgido los datos faltantes. Normalmente, estos mecanismos no están bajo el control del investigador. Los mecanismos son los siguientes: Desaparición completamente al azar (MCAR), Desaparición al azar (MAR) y No desaparición al azar (NMAR).
Veamos unas notaciones previas para luego definir cada mecamismo, para Laird y cols. (2004), esos mecanismos de datos perdidos pueden considerarse como un modelo de probabilidad para la distribución de un conjunto de variables indicadoras de respuesta. Estas variables indicadoras de respuesta toma el valor 1 cuando se obtiene la medición y 0 en caso contrario.
Supongamos que en un estudio se requieren n medidas por individuo. Por tanto, un vector de medición completo es de $n \times 1$: $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in})^\top$; donde Y_{ij} son los datos observados. Cuando faltan datos, significa que algunas mediciones de $\boldsymbol{Y_i}$ no se observarán para algunos individuos. Para identificar en que momento de medicion se produce la falta ultilizaren un vector de indicadores de respuesta denota: $\boldsymbol{W_i} = (W_{i1}, W_{i2}, ..., W_{in})^\top$; donde $W_{ij} = 1$ si Y_{ij} es obtenida y $W_{ij} = 0$ si Y_{ij} no es obtenida. Por tanto a $\boldsymbol{Y_i}$ puede ser dividida en $\boldsymbol{Y_i^o}$ el vector de respuesta obtenidas y $\boldsymbol{Y_i^M}$ el vector de respuestas no obtenidas.
Además, como la población objetivo de interés puede dividirse o estratificarse en una serie de subpoblaciones distintas definidas por los patrones de datos que faltan. Por lo tanto, $\boldsymbol{W_i}$ también puede considerarse una variable de estratificación que divide la población objetivo en una serie de subpoblaciones.

2.4.1. Mecanismos de datos faltantes

Definamos los tres mecanismos de falta de datos y como se relaciona $\boldsymbol{W_i}$ con $\boldsymbol{Y_i}$:

Desaparecidos completamente al azar (MCAR)

Se dice que los datos faltantes son completamente al azar (MCAR) si la probabilidad de que falten no está relacionada con los datos observados. Es decir, Laird y cols. (2004) los datos

son MCAR cuando $\boldsymbol{W}_i$ es independiente, tanto de $\boldsymbol{Y}_i^o$ y $\boldsymbol{Y}_i^M$, las mediciones obtenidas y no obtenidas de $\boldsymbol{Y}_i$. Considero un caso bivariante en el que $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})^\top$, si Y_{i1} es un vector totalmente obtenido y $\boldsymbol{Y}_{i2}$ es un vector con algunas mediciones no obtenidas, en este caso, $W_{i2} = 1$ si $\boldsymbol{Y}_{i2}$ es obtenia y $W_{i2} = 0$ si $\boldsymbol{Y}_{i2}$ no es obtenida. Y solo datos son MCAR, entonces: $P_r\left(W_{i2} = 1|\boldsymbol{Y}_{i1}, \boldsymbol{Y}_{i2}, \boldsymbol{X}_i\right) = P_r = (W_{i2} = 1|\boldsymbol{X}_i)$;
La probabilidad de que falte $\boldsymbol{Y}_{i2}$ no depende de las mediciones obtenidas de $\boldsymbol{Y}_{i1}$ o las obtenidas de $\boldsymbol{Y}_{i2}$ que en principio debería haberse obtenido.
Una característica de los datos faltantes MCAR es que se puede utilizar cualquier método que produzca inferencias válidas en casos completos, ya que se pueden eliminar los individuos son $\boldsymbol{Y}_i$ no completos, porque $\boldsymbol{Y}_i$ se pueden considerar una nuestra aleatoria elegida al azar.

Desaparecidos al azar (MAR)

Se dice que los datos faltantes son desaparecidos al azar (MAR) cuando la probabilidad de que falten datos depende de los conjuntos de mediciones obtenidas, pero es condicionalmente independiente de los valores faltantes específicos que en principio deberían haberse obtenido. es decir los datos son MAR cuando $\boldsymbol{W}_i$ es condicionalmente independiente de $\boldsymbol{Y}_i^m$ dado $\boldsymbol{Y}_i^o$. $P_r\left(\boldsymbol{W}_i|\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, \boldsymbol{X}_i\right) = P_r = (\boldsymbol{W}_i|\boldsymbol{Y}_i^o, \boldsymbol{X}_i)$;
En caso bivariante $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})$, si Y_{i1} es un vector totalmente obtenido y $\boldsymbol{Y}_{i2}$ es un vector con algunas mediciones no obtenidas. si $\boldsymbol{Y}_{i2}$ es MAR, entonces:
$P_r\left(W_{i2} = 1|\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, \boldsymbol{X}_i\right) = P_r = (W_{i2} = 1|\boldsymbol{Y}_i^o, \boldsymbol{X}_i)$
Y la probabilidad de que falte $\boldsymbol{Y}_{i2}$ depende del valor observado de $\boldsymbol{Y}_{i1}$. sin embargo, dado que $\boldsymbol{Y}_{i1}$, la probabilidad que falte $\boldsymbol{Y}_{i2}$ no pedende del valor de $\boldsymbol{Y}_{i2}$ que deberia haberse obtenido.
D. Rubin (1976) declara que los datos MCAR y MAR se denominan mecanismos de causa ignorable, pero cuando los datos son MAR, la ignorabilidad se refiere al hecho que de que una vez que establecemos que $P_r(\boldsymbol{W}_1|\boldsymbol{Y}_i, \boldsymbol{X}_i)$ no depende de los datos no obtenidos, podemos ignorar $P_r(\boldsymbol{W}_1|\boldsymbol{Y}_i, \boldsymbol{X}_i)$ y obtener un análisis válidos basados en la verosimilitud siempre y cuando tengamos un modelo correcto para $f(\boldsymbol{Y}_i|\boldsymbol{X}_i)$.

No desaparecidos al azar (NMAR)

Se dice que los datos son no desaparecidos al azar (NMAR) cuando la probabilidad de que falten está relacionada con los las mediciones que deberían haberse obtenido. Es decir, la distribución condicional de $\boldsymbol{W}_i$ dada $\boldsymbol{Y}_i^o$ está relacionada con $\boldsymbol{Y}_i^M$, y $P_r = (\boldsymbol{W}_i|\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^M, \boldsymbol{X}_i)$
Depende de al menos algunos elementos de $\boldsymbol{Y}_i^M$. En caso bivariado $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})^\top$, si $\boldsymbol{Y}_{i1}$ es un vector totalmente obtenido y $\boldsymbol{Y}_{i2}$ es un vector con algunas mediciones no obtenidas, si la falta en $\boldsymbol{Y}_{i2}$ es NMAR, entonces:

$$P_r = (W_{i2} = 1|Y_{i1}, Y_{i2}, \boldsymbol{X}_i)$$

Además a este mecanismo de causa de datos faltantes NMAR se les denominan no ignorables, se refiere al hecho que si se quieren inferencias válidas sobre la distribución de las mediciones

no pueden ignorarse las mediciones no obtenidas.

2.4.2. Implicaciones de los datos faltantes

Los métodos estadísticos utilizados para analizar estudios longitudinales con datos completos pueden adaptarse para abordar estudios longitudinales con datos faltantes. Sin embargo, para obtener inferencias válidas a partir de datos longitudinales con datos faltantes, se requiere realizar suposiciones sobre el mecanismo de los datos ausentes.

Cuando el mecanismo de datos perdidos es MCAR, significa que los individuos con datos perdidos forman un subconjunto aleatorio de la muestra. En este caso, los valores observados de las respuestas representan una submuestra aleatoria de todos los valores de las respuestas, y no se introduce sesgo significativo en la mayoría de los métodos de análisis de datos . En particular, todos los métodos utilizados para analizar estudios longitudinales con datos completos producirán estimaciones válidas de las tendencias medias de las respuestas (y las asociaciones dentro de los sujetos) si se puede asumir que los datos faltantes son MCAR.

Cuando el mecanismo de datos faltantes es MAR (Missing at Random, por sus siglas en inglés), los individuos con datos faltantes dejan de ser un subconjunto aleatorio de la muestra. Solo cuando se estratifican en función de sus resultados observados (es decir, condicionados a $\boldsymbol{Y}_i^o$), pueden considerarse un subconjunto aleatorio de la muestra perteneciente a ese estrato. En consecuencia, los valores observados no necesariamente representan una submuestra aleatoria de las respuestas. En particular, la distribución de $\boldsymbol{Y}_i^o$, los datos observados de $\boldsymbol{Y}_i$, difiere de la distribución de los mismos datos de $\boldsymbol{Y}_i$ en la población objetivo. Esto implica que, basándose en las observaciones disponibles, las medias muestrales en cada ocasión (y las covarianzas) proporcionan estimaciones sesgadas de las medias (y covarianzas) en la población objetivo.

Cuando los datos que faltan son MAR pero no MCAR, los métodos de casos completos y los métodos de Estimación de Ecuaciones Generalizadas (EEG) estándar basados en todas las observaciones disponibles producen estimaciones sesgadas de las tendencias medias de respuesta. Por el contrario, los métodos basados en la verosimilitud que especifican correctamente toda la distribución conjunta de las respuestas producen estimaciones válidas cuando los datos que faltan son MAR. Así, cuando los datos que faltan son MAR, los métodos basados en la verosimilitud proporcionan inferencias válidas sobre los cambios en la respuesta media a lo largo del tiempo, siempre que la matriz de covarianza se haya modelado correctamente. En resumen, cuando los datos que faltan son MAR, las inferencias sobre la respuesta media son sensibles a cualquier forma de especificación errónea de la distribución conjunta del vector de respuestas. En consecuencia, si los datos longitudinales son incompletos, se debe tener más cuidado al modelar la asociación dentro del sujeto.

Por último, cuando los datos longitudinales son NMAR (Not Missing at Random, por sus siglas en inglés), la mayoría de los métodos estándar de análisis longitudinal no son válidos. Tanto los métodos GEE como los métodos estándar basados en la verosimilitud (que ignoran

el mecanismo de datos faltantes) producen estimaciones sesgadas de las tendencias medias de la respuesta. Para obtener estimaciones válidas, se requiere utilizar modelos conjuntos para la respuesta y el mecanismo de datos faltantes. De hecho, el término "no ignorable"se utiliza para enfatizar que el mecanismo de datos faltantes debe especificarse correctamente (es decir, no puede ignorarse) para realizar inferencias sobre las respuestas completas. También es importante destacar que cualquier suposición de que la falta de respuesta sea NMAR es totalmente inverificable a partir de los datos disponibles. Es decir, sin información externa sobre las razones de la falta de respuesta o el mecanismo de la falta de respuesta, los datos observados no proporcionan información que pueda respaldar o refutar un mecanismo de NMAR sobre otro. Por lo tanto, cuando se sospecha que la falta es NMAR, es importante evaluar cuidadosamente la sensibilidad de las inferencias a una variedad de supuestos plausibles sobre el proceso de falta.

2.5. Métodos para el tratamiento de datos faltantes en estudios longitudinales

En esta sección, ofrecemos una revisión exhaustiva de algunos de los métodos más empleados para abordar la problemática de los datos faltantes en el análisis longitudinal. Es crucial destacar que existen dos categorías principales de métodos: aquellos que se centran en el sujeto individual, conocidos como métodos intrasujeto (que emplean exclusivamente las mediciones del propio sujeto o individuo) y aquellos que se enfocan en la relación entre sujetos, denominados métodos intersujetos los cuales utilizan las mediciones de todos los sujetos participantes en el estudio.

2.5.1. Análisis de casos completos

Este método consiste en eliminar a todos los sujetos que no cuenten con datos en todas las ocasiones de medición programadas. Es importante destacar que este método es altamente problemático y rara vez se considera una opción aceptable para el análisis. Solo generará estimaciones imparciales de las tendencias medias de respuesta cuando se pueda asumir que la pérdida de datos es completamente al azar (MCAR). Es fundamental recordar que en situaciones de pérdida de datos MCAR, los "completados" del estudio constituyen una submuestra aleatoria de la población original.
Sin embargo, incluso en casos donde la suposición de pérdida de datos MCAR podría ser defendible, el análisis de caso completo resulta poco atractivo debido a la reducción en el número de sujetos que contribuyen al análisis. Este método puede ser altamente ineficaz, conduciendo a un análisis con una capacidad estadística disminuida.
Para implementar este enfoque en el software estadístico R, se puede utilizar la función "drop_na" del paquete "tidyr".

2.5.2. Análisis de datos disponibles

El concepto de "métodos de datos disponibles" no se refiere a un método específico, sino a un enfoque amplio que abarca diversas técnicas capaces de incorporar fácilmente vectores de medidas repetidas de longitud desigual en el análisis. Por ejemplo, las aplicaciones estándar de Mínimos Cuadrados Generalizados (GLS) o el enfoque de Ecuaciones de Estimación Generalizadas son considerados métodos de datos disponibles, ya que sustentan el análisis en todas las observaciones disponibles.

Este enfoque también puede aplicarse a variables categóricas incompletas, donde los datos faltantes se agrupan en una categoría adicional llamada "faltante". Para variables continuas incompletas, los valores faltantes se establecen en un valor fijo (generalmente cero), y se añade un indicador adicional o variable ficticia (1/0) al modelo analítico principal para indicar si falta el valor de esa variable.

En términos generales, los métodos de datos disponibles son más eficientes que los métodos de casos completos, ya que incorporan la información parcial obtenida de aquellos sujetos que presentan datos faltantes en el estudio. Sin embargo, es crucial destacar que muchos métodos de datos disponibles generarán análisis válidos solo si las medias y covarianzas condicionales, es decir, condicionadas a $\boldsymbol{X}_i$ de las mediciones observadas de $\boldsymbol{Y_i}$ entre aquellos que tienen datos faltantes coinciden con las medias y covarianzas condicionales correspondientes de $\boldsymbol{Y_i}$ en la población objetivo. En consecuencia, los métodos de datos disponibles producirán estimaciones sesgadas de las tendencias medias de respuesta a menos que la exclusión sea completamente aleatoria (MCAR).

En resumen, para que los resultados de los métodos de datos disponibles (y de casos completos) sean válidos, es necesario que la exclusión sea completamente aleatoria (MCAR).

2.5.3. Imputación

La imputación, un método ampliamente utilizado en la práctica, implica alguna forma de sustitución de respuestas faltantes mediante técnicas específicas. La premisa fundamental de la imputación es sencilla: reemplazar o completar los valores no registrados con valores imputados. Uno de los beneficios principales de los métodos de imputación radica en que, una vez construido un conjunto de datos completado, se pueden aplicar métodos estándar diseñados para datos completos. Sin embargo, los métodos que se basan en una única imputación, creando solo un conjunto de datos completado, no reconocen la incertidumbre inherente a la imputación de datos faltantes. La imputación múltiple aborda esta limitación al reemplazar los valores faltantes con un conjunto de m valores plausibles, reconociendo así la incertidumbre asociada a qué valores imputar para las respuestas faltantes.

Aunque la idea principal detrás de la imputación es sencilla, surge la pregunta de cómo generar los valores imputados para las respuestas faltantes. A continuación, examinaremos algunos de los métodos comúnmente utilizados para llevar a cabo la imputación de datos faltantes.

Antes de describir métodos específicos para imputar datos longitudinales con respuestas faltantes, vale la pena señalar que estos métodos asumen implícitamente que el conjunto de datos está estructurado en un formato " ancho" en lugar de un formato "largo", con un solo "registro" para cada individuo. En un formato "ancho", los métodos de imputación de una respuesta faltante en cualquier medición particular pueden aprovechar la correlación positiva con las respuestas en cualquiera de las mediciónes restantes.

Imputación mediante LOCF

El método conocido como "última observación arrastrada" (LOCF) es una técnica de imputación única que consiste en completar los valores faltantes con la última observación registrada del mismo sujeto. Es importante destacar que este método se clasifica como intrasujeto, ya que se basa exclusivamente en las mediciones del sujeto individual.
Este método funciona de la siguiente manera: supongamos que tenemos una serie de datos longitudinales $\boldsymbol{Y}_i = (y_{i1}, y_{12}, ..., y_{in_i})^\top$. Si encontramos un valor faltante en el tiempo k, entonces aplicando LOCF, asignamos el último valor observado antes de k al tiempo k y a todos los tiempos subsiguientes hasta que se obtenga una nueva observación. Esto se puede expresar como:

$$y_k = Y_{k-1}$$

donde y_k es el valor imputado en el tiempo k y y_{k-1} es el último valor observado antes de k.
Para aplicar este enfoque utilizando el software estadístico R, se puede emplear la función "fill" disponible en el paquete "tidyr".

Imputación de la media

Este método se caracteriza por su simplicidad, ya que implica la imputación de valores faltantes mediante la media (o promedio) de los valores observados en la variable correspondiente. Además, este enfoque puede aplicarse tanto a nivel intrasujeto como intersujeto.
En el contexto de datos longitudinales, es posible que tengas múltiples observaciones para cada individuo a lo largo del tiempo.
Este método funciona de la siguiente manera: supongamos que tenemos una serie de datos longitudinales $\boldsymbol{Y}_i = (y_{i1}, y_{12}, ..., y_{in_i})^\top$, tenemos n_i mediciones o observaciones para el i-ésimo sujeto, el promedio para este será:

$$\overline{y_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

donde y_{ij} representa al i-ésimo sujeto en la medición j. Luego, este promedio $\overline{y_i}$ se utiliza como la imputación para cualquier valor faltante correspondiente a ese individuo en cualquier momento de tiempo específico.
En el software R, la implementación de la imputación basada en la media es sencilla utilizando el método "mean" de la función "mice".

Imputación mediante regresión

En los métodos de regresión para imputar datos longitudinales, las observaciones que faltan se imputan secuencialmente utilizando todas las observaciones obtenidas como "predictores" en el modelo de regresión. Es decir, se ajustan a los datos observados una serie de modelos de regresión para $\boldsymbol{Y}_{ik}$, dados $Y_{i1}, ..., Y_{ik-1}$ y $\boldsymbol{X}_i$ (o un subconjunto de $\boldsymbol{X}_i$), se ajustan a los datos observados. Un modelo de regresión sería:

$$E(\boldsymbol{Y}_{ik} \mid Y_{i1}, ..., Y_{ik-1}, X_i) = \boldsymbol{Z}_{ik}^T \boldsymbol{\gamma}$$

puede ajustarse a los datos de los N_k sujetos. Donde $\boldsymbol{Z}_i k$ denota un vector compuesto por $Y_{i1}, ..., Y_{ik-1}$ y cualquier subconjunto de componentes de $\boldsymbol{X_i}$. En este modelo $\boldsymbol{\gamma}$ denota un vector $q \times 1$ de parámetros de regresión que relacionan Y_{ik} con las observaciones y las covariables. Además, abusando ligeramente de la notación, obsérvese que existe un conjunto separado de parámetros de regresión, $\boldsymbol{\gamma}$, para el modelo de regresión en cada ocasión y que las dimensiones correspondientes de $\boldsymbol{Z}_i k$ y $\boldsymbol{\gamma}$ pueden variar en los modelos para distintas ocasiones. Estos modelos de regresión considerados aquí son modelos de imputación y no de análisis. Para el modelo de regresión en la k^{th} ocasión,

$$E(\boldsymbol{Y}_{ik} \mid Y_{i1}, ..., Y_{ik-1}, X_i) = \boldsymbol{Z}_{ik}^T \boldsymbol{\gamma}$$

los parámetros de regresión (y la varianza residual) pueden estimarse mediante mínimos cuadrados ordinarios. La regresión lineal ajustada produce estimaciones de los parámetros de regresión, $\widehat{\gamma}$, y su matriz de covarianza asociada,

$$\widehat{Cov}(\widehat{\gamma}) = \widehat{\sigma}^2 \left(\sum_{i=1}^{N_k} \boldsymbol{Z}_{ik} \boldsymbol{Z}_{ik}^T \right)^{-1},$$

donde $\widehat{\sigma}^2$ es una estimación de la varianza residual y $\boldsymbol{Z}_{ik}$ es el vector de diseño para la regresión de $\boldsymbol{Y}_{ik}$ en $Y_{ik}, ..., Y_{ik-1}$ y cualquier subconjunto de los componentes de $\boldsymbol{X}_i$. En principio, la regresión ajustada podría utilizarse para producir predicciones o imputaciones de los valores que faltan de $\boldsymbol{Y}_{ik}$. Sin embargo, como la regresión ajustada produce una predicción determinista de los valores que faltan en $\boldsymbol{Y}_{ik}$ para cualquier $\boldsymbol{Z}_{ik}$ fijo, debemos incorporar la variación aleatoria para reflejar la incertidumbre de las imputaciones. En concreto, tenemos que añadir al valor predicho para $\boldsymbol{Y}_{ik}$ una extracción aleatoria de la distribución residual de $\boldsymbol{Y}_{ik}$ dada $Y_{i1}, ..., Y_{ik-1}$ y $\boldsymbol{X_i}$. Esto corresponde a añadir un "error" aleatorio del modelo de regresión. Además, debemos tener en cuenta otra fuente de incertidumbre. Recordemos que la predicción de $\boldsymbol{Y_{ik}}$ se basa en un conjunto de coeficientes de regresión estimados y una estimación de la varianza residual, σ^2 ; sin embargo, implícitamente, el proceso de imputación anterior los trataría como fijos y conocidos en lugar de como estimaciones muestrales. Una imputación "adecuada" también debería tener en cuenta esta última fuente de variabilidad.

Para incorporar esta fuente adicional de variación, cada imputación debe basarse en valores diferentes de los coeficientes de regresión y σ^2.Estos valores deben ser extracciones aleatorias de lo que se conoce como sus "distribuciones posteriore". Este último paso, la extracción aleatoria de la "distribución posterior" de los parámetros (bajo una distribución a priori no informativa para estos parámetros), es probablemente la parte más complicada del proceso de imputación. Afortunadamente, este paso se ha implementado en muchos paquetes de de software estadístico para la imputación múltiple.
En resumen, los métodos de regresión requieren el uso de los dos pasos siguientes para producir valores imputados, digamos $\boldsymbol{Y}_{ik}^*$, para los $\boldsymbol{Y}_{ik}$ que faltan:

- Los nuevos parámetros de regresión, digamos γ^*, y la varianza residual, digamos σ^{*2} , se extraen aleatoriamente de sus "distribuciones posteriores" para tener en cuenta la incertidumbre en la estimación de γ y σ^2.En concreto, la varianza residual se extrae aleatoriamente como

  $$\sigma^{*2} = (N_k - q)\widehat{\sigma}^2/\chi^2$$

  donde $N_k - q$ denota los grados de libertad para la varianza residual, y χ^2 es una extracción aleatoria de una distribución chi-cuadrado con $(Nk - q)$ grados de libertad. A continuación, los parámetros de regresión, γ^*, se extraen aleatoriamente de una distribución normal multivariante con media igual a los parámetros de regresión estimada, $\widehat{\gamma}$, y con matriz de covarianza,

  $$\widehat{Cov}(\widehat{\gamma}) = \sigma^{*2}\left(\sum_{i=1}^{N_k} \boldsymbol{Z}_{ik}\boldsymbol{Z}_{ik}^T\right)^{-1},$$

  donde $\widehat{\sigma}^2$ se ha sustituido por σ^{*2} y $\boldsymbol{Z}_{ik}$ denota el vector de diseño para la regresión de $\boldsymbol{Y}_{ik}$ en $Y_{i1}, ..., Y_{ik-1}$, y cualquier subconjunto de los componentes de $\boldsymbol{X}_i$.

- Los valores que faltan para Y_{ik}, digamos Y_{ik}^*, pueden entonces imputarse sobre la base de las siguientes predicciones:

  $$Y_{ik}^* = \boldsymbol{Z}_{\boldsymbol{ik}}^{\boldsymbol{T}}\gamma^* + e^*$$

  donde, para cada observación que falta en $\boldsymbol{Y}_{ik}$, $e*$ se extrae aleatoriamente de una distribución normal con media cero y desviación típica,σ^*.

La imputación por regresión de los restantes valores perdidos continúa de forma similar hasta que se hayan completado todos los valores perdidos. Las imputaciones múltiples se obtienen repitiendo estos pasos $m > 1$ veces.
En el software R, la implementación de la imputación basada en la media es sencilla utilizando el método "norm" de la función "mice".

Imputación mediante Coincidencia de medias predictiva

Este enfoque de la imputación se conoce como ajuste medias predictivas y está estrechamente relacionado con los métodos de regresión (D. B. Rubin, 1986); (Van Buuren, Boshuizen, y Knook, 1999). El ajuste predictivo de medias también se basa en una secuencia de modelos de regresión para $\boldsymbol{Y}_{ik}$, dados $Y_{i1}, ..., Y_{ik-1}$ y $\boldsymbol{X}_i$, ajustados a los datos observados. Sin embargo, en lugar de rellenar los valores que faltan basándose en los valores predichos de las regresiones, el ajuste predictivo de medias imputa utilizando los valores observados de los resultados de los datos disponibles que son, en cierto sentido, los más "cercanos" a los valores predichos. Al igual que con los métodos de imputación por regresión, se ajustan una serie de modelos de regresión para Y_{ik}, dados $Y_{i1}, ..., Y_{ik-1}$ y $\boldsymbol{X}_i$, se ajustan utilizando los datos observados. Cuando las respuestas son continuas, se suele utilizar la regresión lineal estándar para obtener los valores predichos. La regresión lineal ajustada produce estimaciones tanto de los parámetros de regresión como de la varianza residual. Los nuevos parámetros de regresión y la varianza residual se extraen aleatoriamente de sus "distribuciones posteriores" para tener en cuenta la incertidumbre en su estimación. Dada una extracción aleatoria de γ de su "distribución posterior", digamos γ^*, para cada valor perdido se predice un valor

$$\widehat{Y_{ik}} = \boldsymbol{Z} - \boldsymbol{ik}^{\boldsymbol{T}}\gamma^*.$$

Tenga en cuenta que, a diferencia de la imputación por regresión, en la que también se incluye un "error" aleatorio en la predicción, aquí $\boldsymbol{Y}_{ik}$ es una predicción de la media. Los valores predichos también se calculan para las observaciones con valores no ausentes para $\boldsymbol{Y}_{ik}$. A partir de la última, se genera un subconjunto de K observaciones cuyos valores predichos correspondientes son los más cercanos a $\boldsymbol{Y}_{ik}$; el método de correspondencia de medias predictivas requiere que el número de observaciones más cercanas, K, se especifique de antemano. Por tanto, para cada valor que falta hay un conjunto de K "donantes" potenciales que tienen valores predichos similares. El valor que falta se sustituye por un valor extraído aleatoriamente de estos K valores observados. Las imputaciones múltiples se obtienen repitiendo estos pasos $m > 1$ veces. Aunque el ajuste predictivo de medias está estrechamente relacionado con los métodos de regresión, debería ser algo menos sensible a cualquier error de especificación de la secuencia de modelos de regresión. Es decir, el ajuste de medias predictivo se basa en los modelos de regresión sólo para determinar la distancia entre los valores observados y los que faltan. En lugar de sustituir los valores que faltan por las predicciones de los modelos de regresión, la equiparación de medias predictiva imputa los valores observados de los resultados que están "más cerca" de los valores predichos. Cabe destacar que, dado que los valores imputados se extraen aleatoriamente de los valores observados aportados por otros sujetos, el método de ajuste predictivo de medias garantiza que todos los valores imputados sean plausibles.

En el software R, la implementación de la imputación basada en la media es sencilla utilizando el método "pmm" de la función "mice".

3 Estudios de caso

En este capítulo, la idea principal es llevar a cabo un estudio práctico para ilustrar el tratamiento de datos faltantes mediante métodos previamente descritos. Utilizaremos dos bases de datos provenientes del Portal de Datos Abiertos del Estado colombiano, específicamente de estudios longitudinales que no presentan ausencia de datos inicialmente.
El procedimiento consta de los siguientes pasos:

- Paso 1: Aplicar modelos lineales mixtos a cada base de datos sin datos faltantes.
- Paso 2: Realizar predicciones a partir de los modelos y calcular el error cuadrático medio de la raíz RMSE como medida de la calidad de las predicciones.

Después, se introducen datos faltantes bajo las condiciones de D. Rubin (1976). Estos datos faltantes se tratarán utilizando algunos de los métodos previamente mencionados. La finalidad es aplicar un modelo lineal mixto a estos nuevos conjuntos de datos, realizar predicciones y calcular el RMSE para cada metodología de tratamiento de datos faltantes.
El objetivo final es comparar los RMSE obtenidos de las predicciones realizadas sobre los datos originales sin ausencia de datos y los datos tratados con diferentes métodos para evaluar las implicaciones y eficacia de las técnicas de imputación utilizadas.

3.1. Rendimiento de cultivos en Colombia por año

Se dispone de una base de datos histórica que abarca los años 2007 a 2018, centrada en la producción agrícola a nivel nacional. Esta base contiene registros de 224 cultivos distintos distribuidos por 31 departamento de Colombia durante este periodo.
La base de datos original comprende 17 variables, de las cuales nos enfocaremos exclusivamente en las siguientes: Año, Producción (t), Área Cosechada (ha), Departamento y Rendimiento (t/ha). Para optimizar la interpretación de los datos, hemos optado por aplicar el logaritmo a las variables de Producción, Área Cosechada y Rendimiento.
Este enfoque nos permitirá explorar de manera más precisa las tendencias y patrones relacionados con la producción agrícola a lo largo de estos años, proporcionando una visión más detallada y facilitando el análisis de los datos.
A continuación, examinaremos el comportamiento de los departamentos a lo largo de los años.

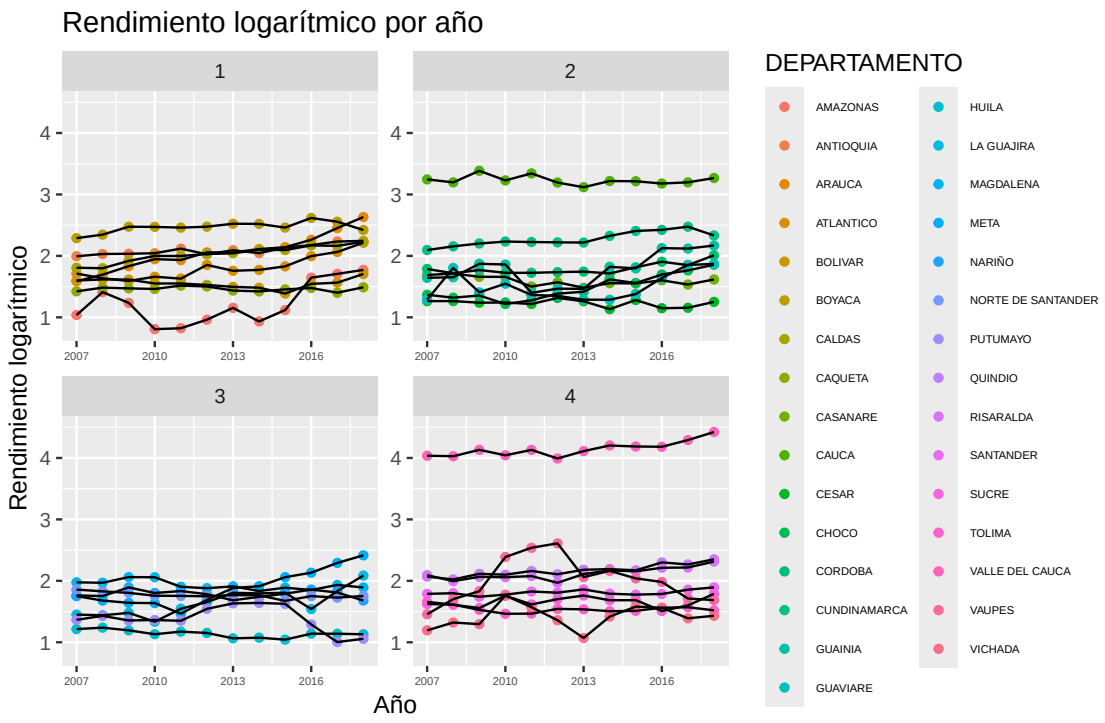

Figura 3-1: Comportamiento de los departamentos por año

3.1.1. Propuesta de un Modelo Lineal Mixto para la Base de Datos originales

Se presenta un modelo lineal mixto con intercepto aleatorio propuesto para analizar el rendimiento en la base de datos, expresado de la siguiente manera:

$$\text{Rd}_{ij} = \beta_0 + \beta_1 \times \text{Año}_{ij} + b_{0i} + e_{ij}$$

Donde:

- Rd_{ij} es el valor de la variable rendimiento para la observación j en el departamento i.
- β_0 es el intercepto fijo común para todas las observaciones. (intercepto fijo).
- β_1 es la pendiente fija asociada a la variable Año.
- b_{0i} es el efecto aleatorio del intercepto para el Departamento i.
- e_{ij} es el error residual.

Los resultados del modelo se resumen en la Tabla **3-1**, que describe los efectos y parámetros estimados.
La interpretación de los resultados se detalla a continuación:

Effect	Group	Term	Estimate	Std. Error	df	Statistic	p-value
fixed	NA	Intercept	-33.9212	4.7500	340	-7.1400	< 0.0001
fixed	NA	Año	0.0177	0.0024	340	7.5400	< 0.0001
ran_pars	Departamento	sd_Intercept	0.5900	NA	NA	NA	NA
ran_pars	Residual	sd_Observa	0.1560	NA	NA	NA	NA

Tabla 3-1: Descripción de efectos y parámetros del modelo

- Efecto Fijos (fixed)
  - Intercepto: Se observa un intercepto significativamente diferente de cero (-33.9212 con un error estándar de 4.75), lo que sugiere un impacto global en el rendimiento.
  - Año: La pendiente asociada a la variable Año es significativa (0.0177 con un error estándar de 0.0023), indicando un cambio significativo en el rendimiento a lo largo del tiempo.
- Efectos Aleatorios (ran_pars)
  - Departamento: Se introduce un efecto aleatorio para cada departamento, con un desvío estándar (sd_Intercepto) de 0.59, que representa la variabilidad entre departamentos.
- Residual: Este componente captura la variación no explicada por el modelo.

Los valores de "Estadístico" y "Valor_p" en la sección "Fixed" indican la significancia de los coeficientes fijos, confirmando que tanto el intercepto como la variable Año son estadísticamente significativos.
Los resultados del modelo, proporcionan estimaciones para los coeficientes del modelo:

$$\widehat{\text{Rd_log}}_{ij} = \widehat{\beta}_0 + \widehat{\beta}_1 \times \text{Año}_{ij} + \widehat{b}_{0j}$$

- $\widehat{\beta}_0$ (Estimación del intercepto fijo)
- $\widehat{\beta}_1$ (Estimación de la pendiente fija asociada a AÑO).
- $\widehat{b_{0i}}$ (Efecto aleatorio del intercepto para cada departamento).

La ecuación sustituyendo los coeficientes $\widehat{\beta}_0$ y $\widehat{\beta}_1$ toman los valores de -33.9212 y 0.0177 respectivamente, sería la sigueinte:

$$\widehat{\text{Rd_log}}_{ij} = -33.9212 + 0.0177 \times \text{Año}_{ij} + \widehat{b_{0i}}.$$

Esta ecuación permitira predecir el logaritmo del rendimiento para una observación específica en un departamento específico en función del año.

La estimacion de $\widehat{b_{0i}}$ que es el efecto aleatorio del intercepto los departamentos es:

DEPARTAMENTO	$\widehat{b_{0i}}$	**DEPARTAMENTO**	$\widehat{b_{0i}}$
AMAZONAS	-0.6499	ANTIOQUIA	0.2254
ARAUCA	0.1868	ATLANTICO	-0.0609
BOLIVAR	-0.3157	BOYACA	0.5935
CALDAS	0.1576	CAQUETA	-0.4069
CASANARE	-0.2748	CAUCA	1.3531
CESAR	-0.6361	CHOCO	-0.3829
CORDOBA	-0.0792	CUNDINAMARCA	0.4026
GUAINIA	-0.0879	GUAVIARE	-0.3627
HUILA	-0.7253	LA GUAJIRA	-0.2219
MAGDALENA	0.0586	META	0.0091
NARIÑO	-0.0881	NORTE DE SANTANDER	-0.1045
PUTUMAYO	-0.4799	QUINDIO	0.2968
RISARALDA	0.2472	SANTANDER	-0.0621
SUCRE	-0.2114	TOLIMA	-0.3291
VALLE DEL CAUCA	2.2623	VAUPES	0.1423
VICHADA	-0.4556		

Tabla 3-2: Valores de intercepto aleatorio por Departamento

Podemos notar como el intercepto es aleotoria para cada departamennto.

Si observamos un gráfico que ilustre los datos reales junto con las predicciones para los todos departamentos.

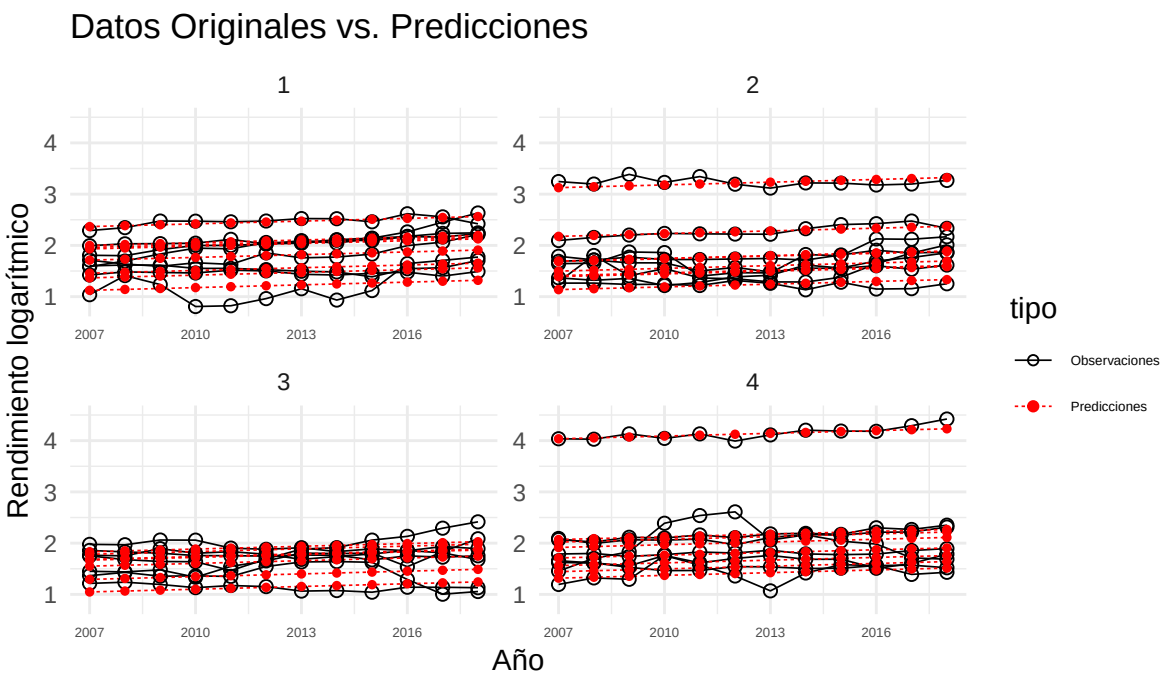

Figura 3-2: Datos Originales vs. Predicciones

Cada panel de la Figura **3-1** representa los mismo departamentos que los paneles correspondientes en la Figura **3-2**.En estos gráficos, las líneas rojas representan las predicciones. Observamos que todas las predicciones para los distintos departamentos son paralelas, lo que sugiere que se trata de un modelo lineal mixto con intercepto aleatorio (Correa Morales y Salazar Uribe, 2016). Por otro lado, las líneas negras muestran los datos reales.

3.1.2. RMSE para los datos originales

$$RMSE = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Rd_log}} - \text{Rd_log})^2} = 0.1501$$

Dado que la media del rendimiento logarítmico es 1.870873, un RMSE de 0.1501 sugiere que, en promedio, las predicciones difieren del valor real en aproximadamente 0.1501 unidades en la escala logarítmica. Este valor puede ser considerado pequeño en relación con la media, lo cual es positivo y sugiere que tu modelo mixto está haciendo predicciones bastante precisas en el contexto de tu conjunto de datos.
Finalmente obtendremos el RMSE para cada departamento

Departamento	RMSE	Departamento	RMSE
AMAZONAS	0.2959	ANTIOQUIA	0.0396
ARAUCA	0.2253	ATLANTICO	0.1347
BOLIVAR	0.1043	BOYACA	0.0648
CALDAS	0.0724	CAQUETA	0.0738
CASANARE	0.1170	CAUCA	0.1056
CESAR	0.0975	CHOCO	0.1466
CORDOBA	0.0625	CUNDINAMARCA	0.0625
GUAINIA	0.2235	GUAVIARE	0.2019
HUILA	0.1071	LA GUAJIRA	0.1636
MAGDALENA	0.1194	META	0.2282
NARIÑO	0.0956	NORTE DE SANTANDER	0.1000
PUTUMAYO	0.2270	QUINDIO	0.04280
RISARALDA	0.0570	SANTANDER	0.0511
SUCRE	0.0931	TOLIMA	0.0924
VALLE DEL CAUCA	0.0764	VAUPES	0.3508
VICHADA	0.1760		

Tabla 3-3: RMSE por Departamento

3.1.3. Generación de Datos Faltantes bajo las Condiciones de D. Rubin (1976)

Como se mencionó previamente, las condiciones de falta de datos según Rubin (1976) se dividen en tres categorías:

- MCAR (Missing Completely At Random): La probabilidad de que falten datos es la misma para todas las observaciones, es decir, la probabilidad de pérdida es completamente aleatoria y no depende de ninguna variable observada o no observada.
- MAR (Missing At Random): La probabilidad de que falten datos puede depender de las variables observadas, pero no de las variables no observadas.
- MNAR (Missing Not At Random): La probabilidad de que falten datos puede depender tanto de las variables observadas como de las no observadas.

A continuación, exploraremos cómo se generan datos faltantes bajo estas condiciones utilizando nuestra base de datos:

- MCAR: se seleccionan aleatoriamente el 10 % de las observaciones y se establece su valor de 'Rd_log' como NA, lo que significa que estos datos están perdidos completamente al azar (MCAR).

- MAR: se establece el valor de 'Rd_log' como NA para todas las observaciones donde 'AÑO' sea mayor al 2007 y menor que 2011. Esto significa que la probabilidad de que falten datos depende de la variable observada 'Año', por lo que estos datos están perdidos al azar (MAR).
- MNAR: se establece el valor de 'Rd_log' como NA para todas las observaciones donde 'Rd_log' es menor que la mediana de 'Rd_log'. Esto significa que la probabilidad de que falten datos depende de la propia variable 'Rd_log', por lo que estos datos están perdidos no al azar (MNAR).

Con estos datos faltantes, aplicaremos métodos de tratamiento para cada condición y realizaremos un modelo lineal mixto, similar al utilizado para los datos originales, para luego comparar los RMSE por departamento y los RMSE de los modelos.

3.1.4. Tratamiento para la Condición MCAR

Como mencionamos anteriormente, generamos datos faltantes bajo la condición MCAR. A continuación, visualizamos gráficamente el porcentaje de datos faltantes creados por departamento.

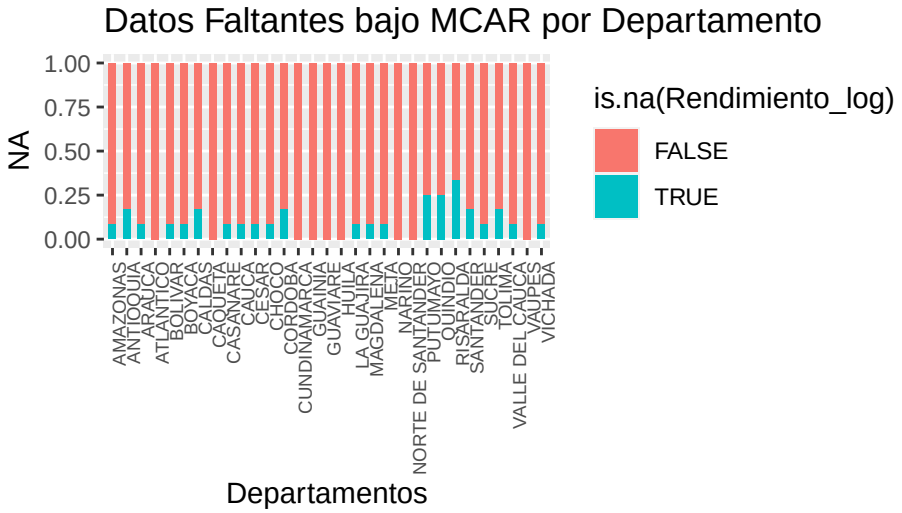

Figura 3-3: Porcentaje de datos faltantes por departamento para la condición MCAR

Se crearon datos faltantes equivalentes al 10 % de la cantidad total de datos. De los 31 departamentos, 22 presentan datos faltantes, siendo Risaraldas el departamento con mayor cantidad de datos faltantes, seguido por Quindío y Putumayo, de los 9 que no presentan datos fataltes algunos son Cudinamarca, Hulia y Vapues.

métodos De Imputación

Como se detalló en el capítulo 4, emplearemos todos los métodos de imputación previamente revisados. Posteriormente, aplicaremos el mismo modelo lineal mixto a cada conjunto de datos imputados. Finalmente, evaluaremos el RMSE para cada modelo y el RMSE para cada departamento en cada modelo.
El modelo utilizado anteriormente y que volveremos a emplear es:

$$\text{Rd}_{ij} = \beta_0 + \beta_1 \times \text{Año}_{ij} + b_{0i} + e_{ij}$$

Donde:

- Rd_{ij} es el valor de la variable respuesta Rendimiento para la observación j en el departamento i.
- β_0 es el intercepto fijo común para todas las observaciones. (intercepto fijo).
- β_1 es la pendiente fija asociada a la variable Año.
- b_{0i} es el efecto aleatorio del intercepto para el Departamento i.
- e_{ij} es el error residual.

los métodos de imputacion a usar son: Imputación mediante LOCF, Imputación de la media, Imputación mediante regresión y Imputación mediante Coincidencia de medias predictiva.

Imputación mediante LOCF

Para el conjunto de datos imputados mediante el método de la última observación arrastrada (LOCF), el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Rd_log}}_{ij} = -32.7654 + 0.0172 \times \text{Año}_{ij} + \widehat{b_{0i}}.$$

Es importante destacar que, al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$ también varían.
Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$RMSE_MCAR_locf = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Rd_log}} - \text{Rd_log})^2} = 0.1519$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante LOCF fue de 0.1519. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamentos	**RMSE_MCAR_locf**	**Departamentos**	**RMSE_MCAR_locf**
AMAZONAS	0.3087	ANTIOQUIA	0.0360
ARAUCA	0.2106	ATLANTICO	0.1361
BOLIVAR	0.1032	BOYACA	0.0696
CALDAS	0.0716	CAQUETA	0.0720
CASANARE	0.1120	CAUCA	0.0962
CESAR	0.0882	CHOCO	0.1416
CORDOBA	0.0625	CUNDINAMARCA	0.0637
GUAINIA	0.2244	GUAVIARE	0.2021
HUILA	0.1053	LA GUAJIRA	0.1687
MAGDALENA	0.1217	META	0.2519
NARIÑO	0.0941	NORTE DE SANTANDER	0.0981
PUTUMAYO	0.2258	QUINDIO	0.0514
RISARALDA	0.0575	SANTANDER	0.0472
SUCRE	0.0926	TOLIMA	0.09699
VALLE DEL CAUCA	0.0639	VAUPES	0.3504
VICHADA	0.2034		

Tabla 3-4: RMSE_MCAR_locf por Departamento

Imputación de la media

Para el conjunto de datos imputados mediante el metodo de la media, el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Rd_log}}_{ij} = -31.9991 + 0.0168 \times \text{Año}_{ij} + \widehat{b_{0i}}.$$

Como se mencionó anteriormente, al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$ también experimentan variaciones.
Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$RMSE_MCAR_media = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Rd_log}} - \text{Rd_log})^2} = 0.2157$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante la media fue de 0.2157351. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamentos	RMSE_MCAR_media	Departamentos	RMSE_MCAR_media
AMAZONAS	0.33705	ANTIOQUIA	0.0853
ARAUCA	0.2178	ATLANTICO	0.1371
BOLIVAR	0.1422	BOYACA	0.1600
CALDAS	0.0836	CAQUETA	0.0711
CASANARE	0.1404	CAUCA	0.3798
CESAR	0.1868	CHOCO	0.1627
CORDOBA	0.0691	CUNDINAMARCA	0.0648
GUAINIA	0.2249	GUAVIARE	0.2024
HUILA	0.1046	LA GUAJIRA	0.1678
MAGDALENA	0.1180	META	0.2244
NARIÑO	0.0931	N. DE SANTANDER	0.0969
PUTUMAYO	0.3005	QUINDIO	0.1503
RISARALDA	0.1419	SANTANDER	0.0570
SUCRE	0.1205	TOLIMA	0.1665
V. DEL CAUCA	0.6358	VAUPES	0.3502
VICHADA	0.2113		

Tabla 3-5: RMSE_MCAR_media por Departamento

Imputación mediante regresión

Para el conjunto de datos imputados mediante el metodo regresión , el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Rd_log}}_{ij} = -37.7644 + 0.0196 \times \text{Año}_{ij} + \widehat{b_{0i}}.$$

Al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$ también experimentan variaciones.

Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$\text{RMSE_MCAR_regresión} = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Rd_log}} - \text{Rd_log})^2} = 0.2451$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante regresión fue de 0.2451. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamentos	RMSE_MCAR_regresión	Departamentos	RMSE_MCAR_regresión
AMAZONAS	0.2900	ANTIOQUIA	0.1176
ARAUCA	0.2481	ATLANTICO	0.1299
BOLIVAR	0.1037	BOYACA	0.3227
CALDAS	0.1123	CAQUETA	0.0800
CASANARE	0.2198	CAUCA	0.5427
CESAR	0.3870	CHOCO	0.1409
CORDOBA	0.0787	CUNDINAMARCA	0.0592
GUAINIA	0.2209	GUAVIARE	0.2012
HUILA	0.1138	LA GUAJIRA	0.1803
MAGDALENA	0.2757	META	0.2689
NARIÑO	0.1009	N. DE SANTANDER	0.1063
PUTUMAYO	0.3495	QUINDIO	0.3308
RISARALDA	0.1667	SANTANDER	0.1835
SUCRE	0.1155	TOLIMA	0.4093
V. DEL CAUCA	0.3069	VAUPES	0.3519
VICHADA	0.1836		

Tabla 3-6: RMSE_MCAR_regresión por Departamento

Imputación mediante coincidencia de medias predictiva

Para el conjunto de datos imputados mediante el metodo Coincidencia de medias predictiva, el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Rd_log}}_{ij} = -35.5437 + 0.0185 \times \text{Año}_{ij} + \widehat{b_{0i}}.$$

Al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$ también experimentan variaciones.

Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$RMSE_MCAR_match = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Rd_log}} - \text{Rd_log})^2} = 0.2270$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante regresión fue de 0.2270. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamentos	RMSE_MCAR_match	Departamentos	RMSE_MCAR_match
AMAZONAS	0.2969	ANTIOQUIA	0.2142
ARAUCA	0.2469	ATLANTICO	0.1326
BOLIVAR	0.1012	BOYACA	0.3242
CALDAS	0.2125	CAQUETA	0.0765
CASANARE	0.1244	CAUCA	0.4657
CESAR	0.3577	CHOCO	0.1567
CORDOBA	0.2672	CUNDINAMARCA	0.0612
GUAINIA	0.2224	GUAVIARE	0.2016
HUILA	0.1101	LA GUAJIRA	0.1631
MAGDALENA	0.1305	META	0.2635
NARIÑO	0.0978	NORTE DE SANTANDER	0.1027
PUTUMAYO	0.2592	QUINDIO	0.2224
RISARALDA	0.1519	SANTANDER	0.0893
SUCRE	0.1139	TOLIMA	0.2107
VALLE DEL CAUCA	0.2732	VAUPES	0.3512
VICHADA	0.3461		

Tabla 3-7: RMSE_MCAR_match por Departamento

Discusión

1 RMSE de Modelos de Datos Completos y Imputados:

- El modelo original con datos completos tiene un RMSE de 0.1501, lo cual indica un buen ajuste del modelo a los datos originales.
- Al imputar datos bajo la condición MCAR, se observa que los RMSE varían dependiendo del método de imputación. El método LOCF tiene un RMSE de 0.1519, el método de la media tiene un RMSE de 0.2157, el método de regresión tiene un RMSE de 0.2451, y el método de coincidencia de medias predictiva tiene un RMSE de 0.2270.

RMSE_MCAR_locf	0.1519
RMSE_MCAR_media	0.2157
RMSE_MCAR_regresión	0.2451
RMSE_MCAR_match	0.2270

Tabla 3-8: Comparacion de los RMSE para la condicion MCAR de métodos de imputacion

2 Comparación de Métodos de Imputación:

- Los resultados sugieren que el método LOCF tiene un rendimiento superior en

términos de RMSE en comparación con el método de la media, regresión y coincidencia de medias predictiva.

- Es crucial considerar que la elección del método de imputación puede depender de la naturaleza de los datos y la validez de las suposiciones subyacentes.

3 Variabilidad por Departamento:

- Al desglosar los RMSE por departamento, se observa que la variabilidad en la precisión del modelo es diferente para cada método de imputación y departamento.
- Para los departamentos como que presentan datos faltantes los métodos sugieren que el método LOCF tiene un rendimiento superior en términos de RMSE en comparación con el método de la media, regresión y coincidencia de medias predictiva. Deparatamentos como Risaralda, Quindio Y Putumayo que son los que mas datos faltantes presenta el RMSE para el modelo original fueron 0.057042, 0.0428 y 0.2270 respectivamente. En la siguiente Tabla **3-31** veresmos las comparación pra los métodos de imputacion para los departamentes mencionados anteriormente.

	Risaralda	**Quindio**	**Putumayo**
RMSE	0.0570	0.0428	0.2270
RMSE_MCAR_locf	0.0575	0.0514	0.2258
RMSE_MCAR_media	0.1419	0.1503	0.3005
RMSE_MCAR_regresión	0.1667	0.3308	0.3495
RMSE_MCAR_match	0.1519	0.2224	0.2592

Tabla 3-9: Comparacion de los RMSE_MCAR de los métodos de imputacion por departamentos que presentan datos faltantes

notamos que el métodos con mejor rendimiento es LOCF.

- En cambio los departamentos como que no presentan datos faltantes independientemente del método de imputación tienen un rendimiento en término de RMSE parecido al de los datos originales.

	Cundinamarca	**Hulia**	**Vaupes**
RMSE	0.0625	0.1071	0.3508
RMSE_MCAR_locf	0.0637	0.1053	0.3504
RMSE_MCAR_media	0.0648	0.1046	0.3502
RMSE_MCAR_regresión	0.0592	0.1138	0.3519
RMSE_MCAR_match	0.0612	0.1101	0.3512

Tabla 3-10: Comparacion de los RMSE_MCAR de los métodos de imputacion por departamentos que no presentan datos faltante

4 Implicaciones de los datos faltantes

- La imputación puede tener un impacto en la interpretación de los coeficientes del modelo. Los cambios en los parámetros, como $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$, deben considerarse al interpretar los resultados del modelo después de la imputación.
- Los métodos de imputación tiene un impacto significativo en los departamentos donde se registran datos faltantes.

3.1.5. Tratamiento para la condición MAR

Generamos datos faltantes bajo la condición MAR. A continuación, visualizamos gráficamente el porcentaje de datos faltantes creados por departamento.

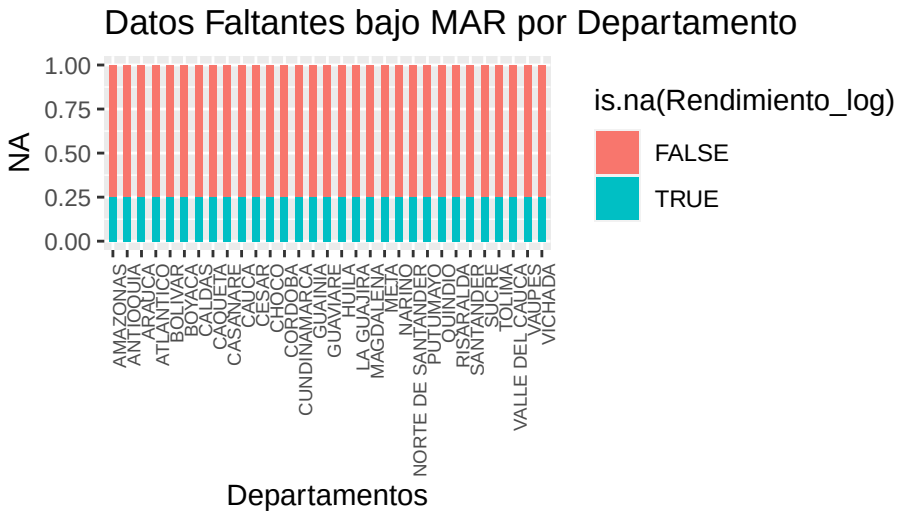

Figura 3-4: Porcenta de falta por departamente para la condicinon MAR

Se crearon datos faltantes para todas las observaciones en las que el año se encuentra en el rango de 2008 a 2010. En consecuencia, todos los departamentos muestran ausencia de datos. A continuación, procederemos con la aplicación de métodos de imputación.

Imputación mediante LOCF

Para el conjunto de datos imputados mediante el método de la ultima observación arrastrada (LOCF), el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Rd_log}}_{ij} = -41.0823 + 0.0213 \times \text{Año}_{ij} + \widehat{b_{0i}}.$$

Es importante destacar que, al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$ también varían.
Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$RMSE_MAR_locf = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Rd_log}} - \text{Rd_log})^2} = 0.1494$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante LOCF fue de 0.1494. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamentos	**RMSE_MAR_locf**	**Departamentos**	**RMSE_MAR_locf**
AMAZONAS	0.2613	ANTIOQUIA	0.0398
ARAUCA	0.2711	ATLANTICO	0.1124
BOLIVAR	0.1125	BOYACA	0.0719
CALDAS	0.0840	CAQUETA	0.0740
CASANARE	0.1299	CAUCA	0.1023
CESAR	0.1112	CHOCO	0.1206
CORDOBA	0.0703	CUNDINAMARCA	0.0730
GUAINIA	0.2093	GUAVIARE	0.1620
HUILA	0.1223	LA GUAJIRA	0.1433
MAGDALENA	0.1129	META	0.2047
NARIÑO	0.0901	NORTE DE SANTANDER	0.1262
PUTUMAYO	0.2319	QUINDIO	0.0367
RISARALDA	0.0580	SANTANDER	0.0619
SUCRE	0.0904	TOLIMA	0.1271
VALLE DEL CAUCA	0.0700	VAUPES	0.3824
VICHADA	0.1376		

Tabla 3-11: RMSE_MAR_locf por Departamento

Imputación de la media

Para el conjunto de datos imputados mediante el metodo de la media, el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Rd_log}}_{ij} = -25.2962 + 0.0135 \times \text{Año}_{ij} + \widehat{b_{0i}}.$$

Como se mencionó anteriormente, al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$ también experimentan variaciones. Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$RMSE_MAR_media = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Rd_log}} - \text{Rd_log})^2} = 0.2888$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante la media fue de 0.2888. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamentos	**RMSE_MAR_media**	**Departamentos**	**RMSE_MAR_media**
AMAZONAS	0.4105	ANTIOQUIA	0.0866
ARAUCA	0.2243	ATLANTICO	0.1281
BOLIVAR	0.1932	BOYACA	0.2387
CALDAS	0.0839	CAQUETA	0.2183
CASANARE	0.1797	CAUCA	0.5562
CESAR	0.3225	CHOCO	0.2175
CORDOBA	0.0870	CUNDINAMARCA	0.1688
GUAINIA	0.2347	GUAVIARE	0.2653
HUILA	0.3637	LA GUAJIRA	0.1653
MAGDALENA	0.0661	META	0.2159
NARIÑO	0.1006	NORTE DE SANTANDER	0.1092
PUTUMAYO	0.3249	QUINDIO	0.1178
RISARALDA	0.1043	SANTANDER	0.06725
SUCRE	0.1446	TOLIMA	0.1888
VALLE DEL CAUCA	0.9687	VAUPES	0.3195
VICHADA	0.2726		

Tabla 3-12: RMSE_MAR_media por Departamento

Imputación mediante regresión

Para el conjunto de datos imputados mediante el metodo regresión , el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Rd_log}}_{ij} = -22.7985 + 0.0122 \times \text{Año}_{ij} + \widehat{b_{0i}}.$$

Al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$ también experimentan variaciones. Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$\text{RMSE_MAR_regresión} = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Rd_log}} - \text{Rd_log})^2} = 0.4343$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante regresión fue de 0.2451. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamentos	**RMSE_MAR_regresión**	**Departamentos**	**RMSE_MAR_regresión**
AMAZONAS	0.6138	ANTIOQUIA	0.3449
ARAUCA	0.4009	ATLANTICO	0.2745
BOLIVAR	0.5153	BOYACA	0.5104
CALDAS	0.3649	CAQUETA	0.3592
CASANARE	0.2640	CAUCA	0.5056
CESAR	0.3722	CHOCO	0.1606
CORDOBA	0.4918	CUNDINAMARCA	0.2496
GUAINIA	0.4159	GUAVIARE	0.2625
HUILA	0.4694	LA GUAJIRA	0.3440
MAGDALENA	0.2419	META	0.4406
NARIÑO	0.3987	N. DE SANTANDER	0.0876
PUTUMAYO	0.3375	QUINDIO	0.6283
RISARALDA	0.2005	SANTANDER	0.3197
SUCRE	0.2303	TOLIMA	0.4399
V. DEL CAUCA	1.1121	VAUPES	0.4999
VICHADA	0.3875		

Tabla 3-13: RMSE_MAR_regresión por Departamento

Imputación mediante Coincidencia de medias predictiva

Para el conjunto de datos imputados mediante el metodo Coincidencia de medias predictiva, el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Rd_log}}_{ij} = -85.3610 + 0.0433 \times \text{Año}_{ij} + \widehat{b_{0i}}.$$

Al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$ también experimentan variaciones. Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$RMSE_MAR_match = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Rd_log}} - \text{Rd_log})^2} = 0.3263$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante regresión fue de 0.3263. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamentos	RMSE_MAR_match	Departamentos	RMSE_MAR_match
AMAZONAS	0.2610	ANTIOQUIA	0.2379
ARAUCA	0.2891	ATLANTICO	0.1693
BOLIVAR	0.2426	BOYACA	0.4311
CALDAS	0.1932	CAQUETA	0.1988
CASANARE	0.2148	CAUCA	0.6855
CESAR	0.3253	CHOCO	0.1029
CORDOBA	0.1118	CUNDINAMARCA	0.3182
GUAINIA	0.2010	GUAVIARE	0.2193
HUILA	0.2583	LA GUAJIRA	0.1614
MAGDALENA	0.2102	META	0.2394
NARIÑO	0.1689	NORTE DE SANTANDER	0.1725
PUTUMAYO	0.3432	QUINDIO	0.2841
RISARALDA	0.3192	SANTANDER	0.1540
SUCRE	0.1654	TOLIMA	0.1756
VALLE DEL CAUCA	1.0226	VAUPES	0.3799
VICHADA	0.2871		

Tabla 3-14: RMSE_MAR_match por Departamento

Discusión

1 RMSE de Modelos de Datos Completos y Imputados:

- El modelo original con datos completos tiene un RMSE de 0.1501, lo cual indica un buen ajuste del modelo a los datos originales.
- Al imputar datos bajo la condición MAR, se observa que los RMSE varían dependiendo del método de imputación. El método LOCF tiene un RMSE de 0.1494, el método de la media tiene un RMSE de 0.2888, el método de regresión tiene un RMSE de 0.4343, y el método de coincidencia de medias predictiva tiene un RMSE de 0.3263.

RMSE_MAR_locf	0.1494
RMSE_MAR_media	0.2888
RMSE_MAR_regresión	0.4343
RMSE_MAR_match	0.3263

Tabla 3-15: Comparacion de los RMSE para la condicion MAR de métodos de imputacion

2 Comparación de Métodos de Imputación:

- Los resultados sugieren que el método LOCF tiene un rendimiento superior en

términos de RMSE en comparación con el método de la media, regresión y coincidencia de medias predictiva.

- Es crucial considerar que la elección del método de imputación puede depender de la naturaleza de los datos y la validez de las suposiciones subyacentes.

3 Variabilidad por Departamento:

- Al desglosar los RMSE por departamento, se observa que la variabilidad en la precisión del modelo es diferente para cada método de imputación y departamento.

4 Implicaciones de los datos faltantes

- La imputación puede tener un impacto en la interpretación de los coeficientes del modelo. Los cambios en los parámetros, como $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$, deben considerarse al interpretar los resultados del modelo después de la imputación.

3.1.6. Tratamiento para la condición MNAR

Generamos datos faltantes bajo la condición MNAR. A continuación, visualizamos gráficamente el porcentaje de datos faltantes creados por departamento.

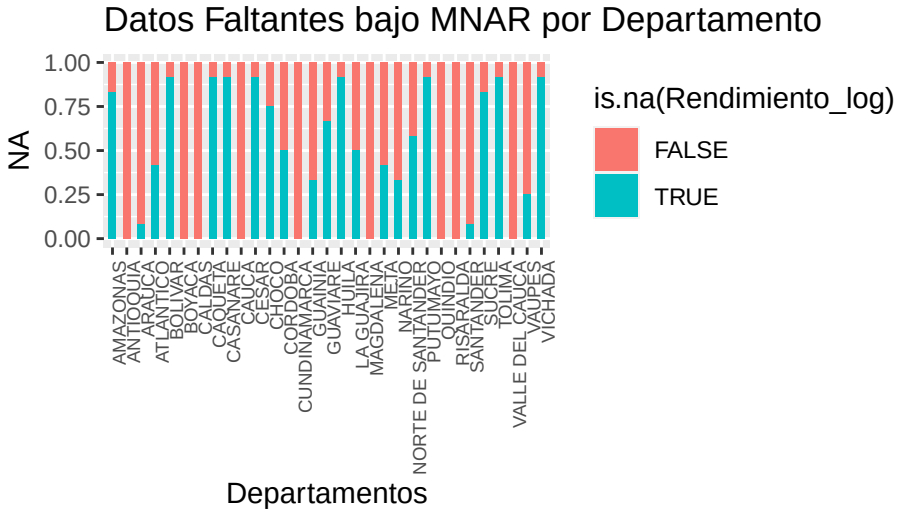

Figura 3-5: Porcenta de falta por departamente para la condicinon MNAR

Se crearon datos faltantes para las observaciones de la variable 'Rd_log' que son inferiores a la mediana de dicha variable. De los 31 departamentos, 22 presentan datos faltantes, siendo Bolivar, Caqueta, Casanares, Cesar y otros con la mayor cantidad de datos faltantes, de los 9 que no presentan datos fataltes algunos son Antioquia, Boyaca, Caldas y Cauca.

Imputación mediante LOCF

Para el conjunto de datos imputados mediante el método de la última observación arrastrada (LOCF), el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Rd_log}}_{ij} = -35.0243 + 0.0183 \times \text{Año}_{ij} + \widehat{b_{0i}}.$$

Es importante destacar que, al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$ también varían.
Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$RMSE_MNAR_locf = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Rd_log}} - \text{Rd_log})^2} = 0.1163$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante LOCF fue de 0.1163. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamentos	**RMSE_MNAR_locf**	**Departamentos**	**RMSE_MNAR_locf**
AMAZONAS	0.1811	ANTIOQUIA	0.0392
ARAUCA	0.2364	ATLANTICO	0.1075
BOLIVAR	0.0633	BOYACA	0.0649
CALDAS	0.07064	CAQUETA	0.0633
CASANARE	0.0633	CAUCA	0.1069
CESAR	0.0633	CHOCO	0.1339
CORDOBA	0.0520	CUNDINAMARCA	0.0614
GUAINIA	0.1167	GUAVIARE	0.1228
HUILA	0.0633	LA GUAJIRA	0.1604
MAGDALENA	0.1211	META	0.1718
NARIÑO	0.0683	NORTE DE SANTANDER	0.0912
PUTUMAYO	0.0633	QUINDIO	0.0417
RISARALDA	0.0565	SANTANDER	0.0546
SUCRE	0.0574	TOLIMA	0.0633
VALLE DEL CAUCA	0.0748	VAUPES	0.3346
VICHADA	0.0633		

Tabla 3-16: RMSE_MNAR_locf por Departamento

Imputación de la media

Para el conjunto de datos imputados mediante el metodo de la media, el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Rd_log}}_{ij} = -37.4092 + 0.0196 \times \text{Año}_{ij} + \widehat{b_{0i}}.$$

Como se mencionó anteriormente, al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$ también experimentan variaciones.
Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$RMSE_MNAR_media = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Rd_log}} - \text{Rd_log})^2} = 0.1829$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante la media fue de 0.1829. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamentos	**RMSE_MNAR_media**	**Departamentos**	**RMSE_MNAR_media**
AMAZONAS	0.3168	ANTIOQUIA	0.0386
ARAUCA	0.2100	ATLANTICO	0.1961
BOLIVAR	0.1445	BOYACA	0.0658
CALDAS	0.0666	CAQUETA	0.1901
CASANARE	0.1225	CAUCA	0.1118
CESAR	0.2323	CHOCO	0.2579
CORDOBA	0.1981	CUNDINAMARCA	0.0588
GUAINIA	0.1806	GUAVIARE	0.2576
HUILA	0.2454	LA GUAJIRA	0.2468
MAGDALENA	0.1253	META	0.1638
NARIÑO	0.1826	N. DE SANTANDER	0.1701
PUTUMAYO	0.2052	QUINDIO	0.0394
RISARALDA	0.0556	SANTANDER	0.1378
SUCRE	0.1896	TOLIMA	0.1303
V. DEL CAUCA	0.0793	VAUPES	0.2882
VICHADA	0.2510		

Tabla 3-17: RMSE_MNAR_media por Departamento

Imputación mediante regresión

Para el conjunto de datos imputados mediante el metodo regresión , el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Rd_log}}_{ij} = -28.4853 + 0.0152 \times \text{Año}_{ij} + \widehat{b_{0i}}.$$

Al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$ también experimentan variaciones.
Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$\textbf{RMSE_MNAR_regresión} = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Rd_log}} - \text{Rd_log})^2} = 0.4445$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante regresión fue de 0.4445. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamentos	RMSE_MNAR_regresión	Departamentos	RMSE_MNAR_regresión
AMAZONAS	0.6231	ANTIOQUIA	0.0436
ARAUCA	0.4662	ATLANTICO	0.3552
BOLIVAR	0.6882	BOYACA	0.0684
CALDAS	0.0818	CAQUETA	0.6109
CASANARE	0.4830	CAUCA	0.1338
CESAR	0.5918	CHOCO	0.4818
CORDOBA	0.4281	CUNDINAMARCA	0.0684
GUAINIA	0.5380	GUAVIARE	0.4068
HUILA	0.6332	LA GUAJIRA	0.6503
MAGDALENA	0.1142	META	0.2494
NARIÑO	0.3483	N. DE SANTANDER	0.3229
PUTUMAYO	0.6239	QUINDIO	0.0484
RISARALDA	0.0604	SANTANDER	0.0949
SUCRE	0.7599	TOLIMA	0.5796
V. DEL CAUCA	0.1884	VAUPES	0.5362
VICHADA	0.5396		

Tabla 3-18: RMSE_MNAR_regresión por Departamento

Imputación mediante Coincidencia de medias predictiva

Para el conjunto de datos imputados mediante el metodo Coincidencia de medias predictiva, el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Rd_log}}_{ij} = -80.6657 + 0.0411 \times \text{Año}_{ij} + \widehat{b_{0i}}.$$

Al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$ también experimentan variaciones.

Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$RMSE_MNAR_match = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Rd_log}} - \text{Rd_log})^2} = 0.3127$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante regresión fue de 0.3127. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamentos	**RMSE_MNAR_match**	**Departamentos**	**RMSE_MNAR_match**
AMAZONAS	0.3411	ANTIOQUIA	0.0808
ARAUCA	0.1490	ATLANTICO	0.1754
BOLIVAR	0.2054	BOYACA	0.1095
CALDAS	0.0354	CAQUETA	0.3760
CASANARE	0.3264	CAUCA	0.1820
CESAR	0.3875	CHOCO	0.2886
CORDOBA	0.7305	CUNDINAMARCA	0.0661
GUAINIA	0.1993	GUAVIARE	0.4301
HUILA	0.4888	LA GUAJIRA	0.2182
MAGDALENA	0.1950	META	0.1219
NARIÑO	0.3594	N. DE SANTANDER	0.4856
PUTUMAYO	0.4776	QUINDIO	0.0663
RISARALDA	0.0852	SANTANDER	0.1311
SUCRE	0.4446	TOLIMA	0.2980
V. DEL CAUCA	0.1193	VAUPES	0.3895
VICHADA	0.3299		

Tabla 3-19: RMSE_MNAR_match por Departamento

Discusión

1 RMSE de Modelos de Datos Completos y Imputados:

- El modelo original con datos completos tiene un RMSE de 0.1501, lo cual indica un buen ajuste del modelo a los datos originales.
- Al imputar datos bajo la condición MNAR, se observa que los RMSE varían dependiendo del método de imputación. El método LOCF tiene un RMSE de 0.1163, el método de la media tiene un RMSE de 0.1829 , el método de regresión tiene un RMSE de 0.4445, y el método de coincidencia de medias predictiva tiene un RMSE de 0.3127.

RMSE_MNAR_locf	0.1163
RMSE_MNAR_media	0.1829
RMSE_MNAR_regresión	0.4445
RMSE_MNAR_match	0.3127

Tabla 3-20: Comparacion de los RMSE para la condicion MNAR de métodos de imputacion

2 Comparación de Métodos de Imputación:

- Los resultados sugieren que el método LOCF tiene un rendimiento superior en términos de RMSE en comparación con el método de la media, regresión y coincidencia de medias predictiva.
- Es crucial considerar que la elección del método de imputación puede depender de la naturaleza de los datos y la validez de las suposiciones subyacentes.

3 Variabilidad por Departamento:

- Al desglosar los RMSE por departamento, se observa que la variabilidad en la precisión del modelo es diferente para cada método de imputación y departamento.
- Para los departamentos como que presentan datos faltantes los métodos sugieren que el método LOCF tiene un rendimiento superior en términos de RMSE en comparación con el método de la media, regresión y coincidencia de medias predictiva. Deparatamentos como Bolivar, Caqueta, Casanares, Cesar que son algunos de los que mas datos faltantes presenta el RMSE para el modelo original fueron 0.1043, 0.0738, 0.1170 y 0.0975 respectivamente. En la siguiente tabla **3-21** veresmos las comparación pra los métodos de imputacion para los departamentes mencionados anteriormente.

	Bolivar	**Caqueta**	**Casanares**	**Cesar**
RMSE	0.1043	0.0738	0.1170	0.0975
RMSE_MCAR_locf	0.0633	0.0633	0.0633	0.0633
RMSE_MCAR_media	0.1445	0.1901	0.1225	0.2323
RMSE_MCAR_regresión	0.6882	0.6109	0.4830	0.5918
RMSE_MCAR_match	0.2054	0.3760	0.3264	0.3875

Tabla 3-21: Comparacion de los RMSE_MNAR de los métodos de imputacion por departamentos que presentan datos faltantes

notamos que el métodos con mejor rendimiento es LOCF.

- En cambio los departamentos como que no presentan datos faltantes independientemente del método de imputación tienen un rendimiento en término de RMSE parecido al de los datos originales.

	Antioquia	**Boyaca**	**Caldas**	**Cauca**
RMSE	0.0396	0.0648	0.0724	0.1056
RMSE_MCAR_locf	0.0392	0.0649	0.0706	0.1069
RMSE_MCAR_media	0.0386	0.0658	0.0666	0.1118
RMSE_MCAR_regresión	0.0436	0.0684	0.0818	0.1338
RMSE_MCAR_match	0.0808	0.1095	0.0354	0.1820

Tabla 3-22: Comparacion de los RMSE_MNAR de los métodos de imputacion por departamentos que no presentan datos faltante

4 Implicaciones de los datos faltantes

- La imputación puede tener un impacto en la interpretación de los coeficientes del modelo. Los cambios en los parámetros, como $\widehat{\beta_0}$, $\widehat{\beta_1}$ y $\widehat{b_{0i}}$, deben considerarse al interpretar los resultados del modelo después de la imputación.

3.1.7. Discusión general para las tres condiciones

Para la base de datos del Rendimiento de Cultivos en Colombia por año, se generaron datos faltantes siguiendo las tres condiciones propuestas por D. Rubin (1976). Posteriormente, se aplicó un modelo lineal con intercepto aleatorio. Se llevaron a cabo métodos de imputación para los datos faltantes.

Los resultados revelaron que al comparar los errores cuadráticos medios de los modelos, estos variaron dependiendo del método de imputación utilizado. Específicamente, se encontró que el método LOCK tuvo un rendimiento superior en todas las condiciones evaluadas, tanto en el RMSE del modelo como en el RMSE entre los departamentos (intra-sujeto). En conclusión, el método LOCK demostró consistentemente ser el más efectivo en términos de rendimiento.

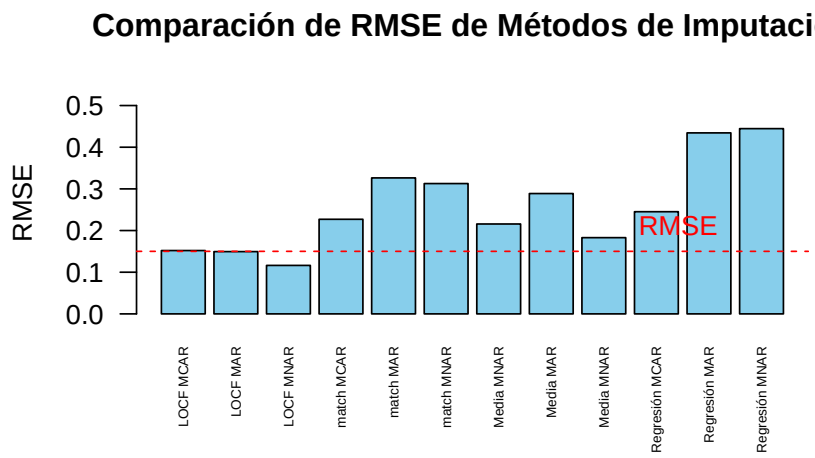

Figura 3-6: Comparación de RMSE de Métodos de Imputación

3.2. Producción de Aguacate

Esta es la segunda base de datos es una base histórica de los años 2007 a 2018, relacionada con la producción agrícola nacional. Esta base contiene registros del cultivo de aguacate distribuido por 29 departamentos de Colombia durante este periodo. De estos 29 departamentos, solo utilizaremos 19, ya que únicamente estos tienen registro en todo el periodo de tiempo.

La base de datos original comprende 17 variables, de las cuales nos enfocaremos exclusivamente en las siguientes: Año, Producción (t), Área Cosechada (ha) y Departamento. Para optimizar la interpretación de los datos, hemos optado por aplicar el logaritmo a las variables de Producción y Área Cosechada. Además, en lugar de utilizar la variable año desde el año 2007 al 2018, la utilizaremos como 1 para el año 2007, 2 para el año 2008, y así sucesivamente hasta 11 para el año 2018.

Este enfoque nos permitirá explorar de manera más precisa las tendencias y patrones relacionados con la producción agrícola a lo largo de estos años, proporcionando una visión más detallada y facilitando el análisis de los datos.

A continuación, examinaremos el comportamiento de los departamentos a lo largo de los años.

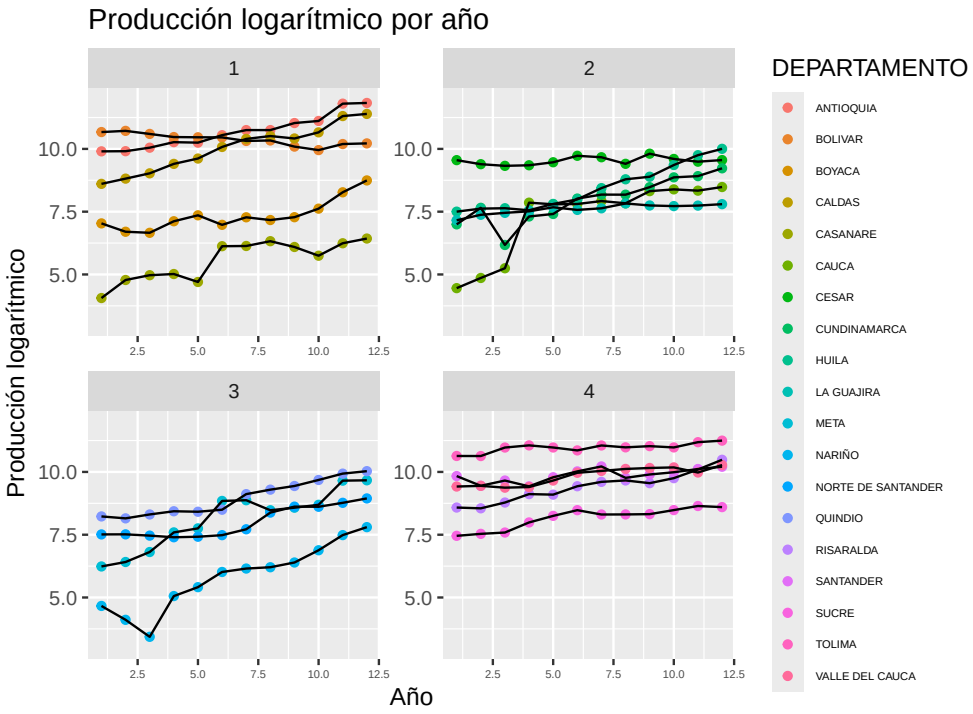

Figura 3-7: Comportamiento de los departamentos por año

3.2.1. Propuesta de un Modelo Lineal Mixto para la Base de Datos originales

Se presenta un modelo lineal mixto con intecepto y pendiente aleatoria propuesto para analizar el rendimiento en la base de datos, expresado de la siguiente manera:

$$\text{Prod}_{ij} = \beta_0 + \beta_1 \times \text{Año}_{ij} + b_{0i} + b_{1i} \times \text{Año}_{ij} + e_{ij}$$

Donde:

- Prod_{ij} es el valor de la variable respuesta Producción para la observación j en el departamento i.
- β_0 es el intercepto fijo común para todas las observaciones. (intercepto fijo).
- β_1 es la pendiente fija asociada a la variable Año.
- Año es el valorde la variable independiente Año para la observación j en el departamento i.

- b_{0i} es el efecto aleatorio del intercepto para el Departamento i.
- b_{1i} es el efecto aleatorio de la pendiente para el Departamento i.
- e_{ij} es el error residual.

Los resultados del modelo se resumen en la Tabla **3-23**, que describe los efectos y parámetros estimados.

Effect	**Group**	**Term**	**Estimate**	**Std. Error**	**df**	**Statistic**	**p-value**
fixed	NA	(Intercept)	7.5911	0.4642	208	16.3524	<0.0001
fixed	NA	Año	0.1590	0.0257	208	6.1835	<0.0001
ran_pars	Dpto	sd_(Intercept)	2.0137	NA	NA	NA	NA
ran_pars	Dpto	cor_Año.(Intercept)	-0.7730	NA	NA	NA	NA
ran_pars	Dpto	sd_Año	0.1088	NA	NA	NA	NA
ran_pars	Residual	sd_Observation	0.3220	NA	NA	NA	NA

Tabla 3-23: Descripción de efectos y parámetros del modelo

La interpretación de los resultados se detalla a continuación:

- Efectos Fijos (fixed)
  - La estimación (estimate) para la intersección (Intercept) es 7.5911, con un error estándar de 0.4642. El valor t es 16.3524 y el p-valor es prácticamente cero, lo que sugiere que la intersección es significativamente diferente de cero.
  - La estimación para el coeficiente de Año es 0.1590946, con un error estándar de 0.02572. El valor t es 6.1835 y el p-valor es prácticamente cero, indicando que el coeficiente de Año es significativamente diferente de cero.
- Efectos Aleatorios (ran_pars)
  - Para el efecto aleatorio de la intersección (sd_(Intercept)) en el nivel de departamento, la estimación es 2.0137.
  - El coeficiente de correlación entre el efecto aleatorio de la intersección y el efecto aleatorio de la pendiente es -0.7730.
  - La desviación estándar del efecto aleatorio de la pendiente (sd_Año) en el nivel de departamento es 0.1088.
- Residual: La desviación estándar del término de error (Residual) es 0.3220.

Los resultados del modelo, proporcionan estimaciones para los coeficientes del modelo:

- $\widehat{\beta}_0$ (Estimación del intercepto fijo)

- $\widehat{\beta}_1$ (Estimación de la pendiente fija asociada a AÑO).
- $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$ (Efecto aleatorio del intercepto para cada departamento).

Por lo tanto, la ecuación del modelo con las estimaciones sería:

$$\widehat{\text{Prod_log}}_{ij} = \widehat{\beta}_0 + \widehat{\beta}_1 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

La ecuación sustituyendo los coeficientes $\widehat{\beta}_0$ y $\widehat{\beta}_1$ toman los valores de 7.5911 y 0.1590 respectivamente, sería la sigueinte:

$$\widehat{\text{Prod_log}}_{ij} = 7.5911 + 0.1590 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

Esta ecuación permitira predecir el logaritmo del rendimiento para una observación específica en un departamento específico en función del año.
Las estimaciones de $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$ que son los efecto aleatorio del intercepto y pendiente los departamentos es:

Departamento	$\widehat{b_{0i}}$	$\widehat{b_{1i}}$
ANTIOQUIA	1.9699	0.0117
BOLIVAR	3.0813	-0.2051
BOYACA	-1.1839	-0.0129
CALDAS	0.8469	0.0825
CASANARE	-3.3191	0.0400
CAUCA	-2.5191	0.1798
CESAR	1.7702	-0.1332
CUNDINAMARCA	-1.0546	0.0585
HUILA	-0.6721	0.0753
LA GUAJIRA	-0.3754	-0.0982
META	-1.3765	0.1354
NARIÑO	-3.9760	0.1783
NORTE DE SANTANDER	-0.6240	-0.0017
QUINDIO	0.1669	0.0254
RISARALDA	0.7755	-0.0018
SANTANDER	1.9010	-0.1022
SUCRE	-0.1539	-0.0468
TOLIMA	3.0863	-0.1157
VALLE DEL CAUCA	1.6565	-0.0693

Tabla 3-24: Valores de intercepto aleatorio por Departamento

Podemos notar como el intercepto y la pendiente son aleotoria para cada departamento.
Si observamos un gráfico que ilustre los datos reales junto con las predicciones para los todos departamentos.

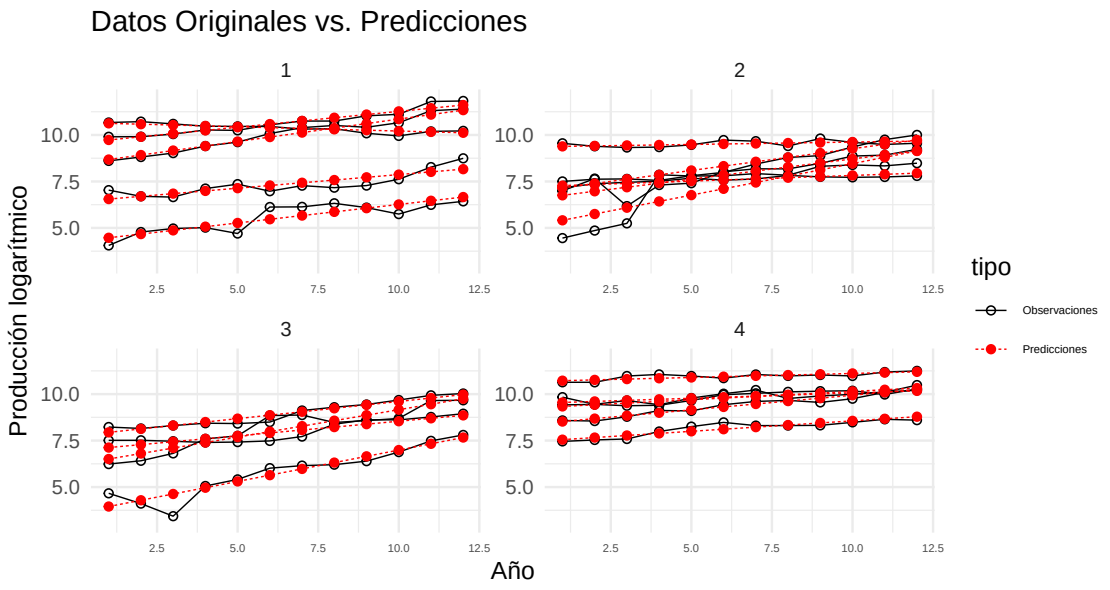

Figura 3-8: Datos Originales vs. Predicciones

Cada panel de la Figura **3-7** representa los mismo departamentos que los paneles correspondientes en la Figura **3-8**. En estos gráficos, las líneas rojas representan las predicciones. Observamos que todas las predicciones para los distintos departamentos son diferentes de los demás en términos del intercepto y la pendiente, lo que sugiere que se trata de un modelo lineal mixto con intercepto y pendiente aleatorio (Correa Morales y Salazar Uribe, 2016). Por otro lado, las líneas negras muestran los datos reales.

3.2.2. RMSE para los datos originales

$$RMSE = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Prod_log}} - \text{Prod_log})^2} = 0.2951$$

Dado que la media del producción logarítmico es 8.6252, un RMSE de 0.2951 sugiere que, en promedio, las predicciones difieren del valor real en aproximadamente 0.2951 unidades en la escala logarítmica. Este valor puede ser considerado pequeño en relación con la media, lo cual es positivo y sugiere que tu modelo mixto está haciendo predicciones bastante precisas en el contexto de tu conjunto de datos.
Finalmente obtendremos el RMSE para cada departamento

Departamento	RMSE
ANTIOQUIA	0.1578
BOLIVAR	0.1070
BOYACA	0.3264
CALDAS	0.1549
CASANARE	0.3816
CAUCA	0.7598
CESAR	0.1368
CUNDINAMARCA	0.3761
HUILA	0.2308
LA GUAJIRA	0.0975
META	0.3887
NARIÑO	0.4367
NORTE DE SANTANDER	0.2518
QUINDIO	0.1663
RISARALDA	0.1342
SANTANDER	0.1796
SUCRE	0.1667
TOLIMA	0.1070
VALLE DEL CAUCA	0.1403

Tabla 3-25: RMSE por departamento

3.2.3. Generación de Datos Faltantes bajo las Condiciones de D. Rubin (1976)

Como se mencionó previamente, las condiciones de falta de datos según Rubin (1976) se dividen en tres categorías:

- MCAR (Missing Completely At Random): La probabilidad de que falten datos es la misma para todas las observaciones, es decir, la probabilidad de pérdida es completamente aleatoria y no depende de ninguna variable observada o no observada.
- MAR (Missing At Random): La probabilidad de que falten datos puede depender de las variables observadas, pero no de las variables no observadas.
- MNAR (Missing Not At Random): La probabilidad de que falten datos puede depender tanto de las variables observadas como de las no observadas.

A continuación, exploraremos cómo se generan datos faltantes bajo estas condiciones utilizando nuestra base de datos:

- MCAR: se seleccionan aleatoriamente el 10 % de las observaciones y se establece su valor de 'Prod_log' como NA, lo que significa que estos datos están perdidos completamente al azar (MCAR).

- MAR: se establece el valor de 'Prod_log' como NA para todas las observaciones donde 'Año' sea mayor al 1 y menor que 5. Esto significa que la probabilidad de que falten datos depende de la variable observada 'Año', por lo que estos datos están perdidos al azar (MAR).

- MNAR: se establece el valor de 'Prod_log' como NA para todas las observaciones donde 'Prod_log' es menor que la mediana de 'Prod_log'. Esto significa que la probabilidad de que falten datos depende de la propia variable 'Prod_log', por lo que estos datos están perdidos no al azar (MNAR).

Con estos datos faltantes, aplicaremos métodos de tratamiento para cada condición y realizaremos un modelo lineal mixto, similar al utilizado para los datos originales, para luego comparar los RMSE por departamento y los RMSE de los modelos.

3.2.4. Tratamiento para la Condición MCAR

Como mencionamos anteriormente, generamos datos faltantes bajo la condición MCAR. A continuación, visualizamos gráficamente el porcentaje de datos faltantes creados por departamento.

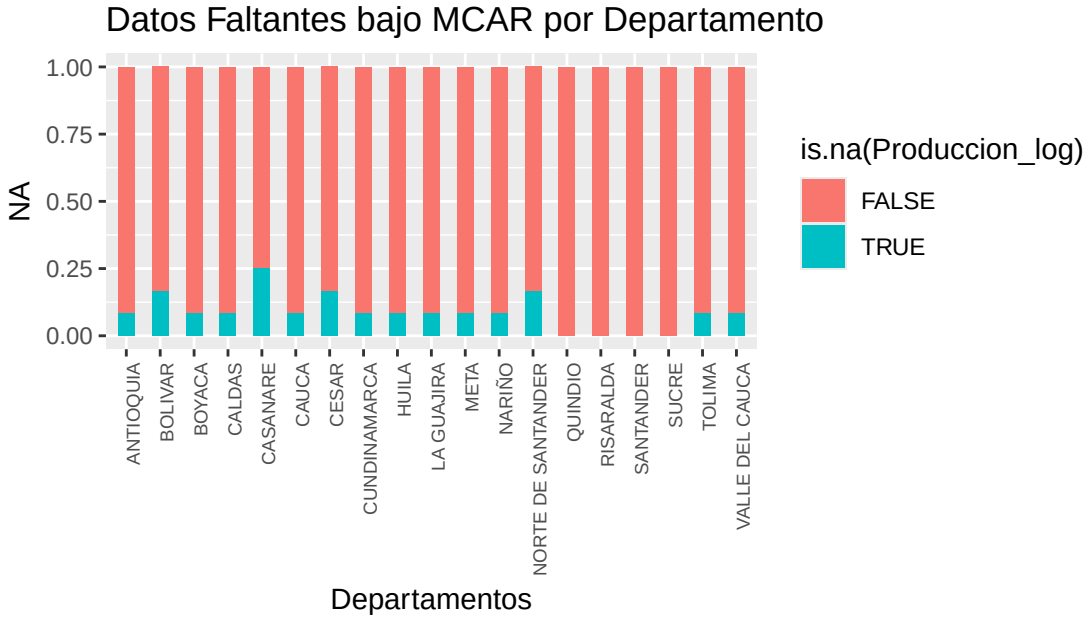

Figura 3-9: Porcentaje de datos faltantes por departamento para la condición MCAR

Se crearon datos faltantes equivalentes al 10 % de la cantidad total de datos. De los 19 departamentos, 15 presentan datos faltantes, siendo Casanare el departamento con mayor cantidad de datos faltantes, seguido por Norte de santander y Cesar,los 4 que no presentan datos fataltes son Quindio, Risaralda, Santander y Sucre.

métodos De Imputación

Como se detalló en el capítulo 4, emplearemos todos los métodos de imputación previamente revisados. Posteriormente, aplicaremos el mismo modelo lineal mixto a cada conjunto de datos imputados. Finalmente, evaluaremos el RMSE para cada modelo y el RMSE para cada departamento en cada modelo.
El modelo utilizado anteriormente y que volveremos a emplear es:

$$\widehat{\text{Prod_log}}_{ij} = \widehat{\beta}_0 + \widehat{\beta}_1 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

los métodos de imputacion a usar son: Imputación mediante LOCF, Imputación de la media, Imputación mediante regresión y Imputación mediante Coincidencia de medias predictiva.

Imputación mediante LOCF

Para el conjunto de datos imputados mediante el método de la última observación arrastrada (LOCF), el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Prod_log}}_{ij} = 7.5855 + 0.1576 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

Es importante destacar que, al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$ también varían.
Posteriormente, para evaluar el produccion del modelo, se calcula el RMSE utilizando la fórmula:

$$RMSE_MCAR_locf = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Prod_log}} - \text{Prod_log})^2} = 0.4863$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante LOCF fue de 0.4863. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamento	RMSE_MCAR_locf
ANTIOQUIA	0.1913
BOLIVAR	0.8090
BOYACA	0.3653
CALDAS	0.4797
CASANARE	0.5560
CAUCA	1.2542
CESAR	0.5876
CUNDINAMARCA	0.2644
HUILA	0.3519
LA GUAJIRA	0.3729
META	0.8005
NARIÑO	0.5211
NORTE DE SANTANDER	0.2544
QUINDIO	0.1977
RISARALDA	0.1727
SANTANDER	0.5047
SUCRE	0.2704
TOLIMA	0.4014
VALLE DEL CAUCA	0.3173

Tabla 3-26: RMSE_MCAR_locf por Departamento

Imputación de la media

Para el conjunto de datos imputados mediante el metodo de la media, el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Prod_log}}_{ij} = 8.2075 + 0.0958 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

Como se mencionó anteriormente, al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$ también experimentan variaciones.
Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$RMSE_MCAR_media = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Prod_log}} - \text{Prod_log})^2} = 0.6746$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante la media fue de 0.6746. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamento	RMSE_MCAR_media
ANTIOQUIA	0.7555
BOLIVAR	0.6167
BOYACA	0.8320
CALDAS	0.6351
CASANARE	1.5987
CAUCA	1.0559
CESAR	0.3168
CUNDINAMARCA	0.5575
HUILA	0.5226
LA GUAJIRA	0.7220
META	0.6990
NARIÑO	1.3867
NORTE DE SANTANDER	0.5695
QUINDIO	0.3097
RISARALDA	0.2469
SANTANDER	0.3819
SUCRE	0.4412
TOLIMA	0.7613
VALLE DEL CAUCA	0.3309

Tabla 3-27: RMSE_MCAR_media por Departamento

Imputación mediante regresión

Para el conjunto de datos imputados mediante el metodo regresión , el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Prod_log}}_{ij} = 7.6293 + 0.1543 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

Al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$ también experimentan variaciones.
Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$\text{RMSE_MCAR_regresión} = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Prod_log}} - \text{Prod_log})^2} = 0.7788$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante regresión fue de 0.7788. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamento	RMSE_MCAR_regresión
ANTIOQUIA	0.3385
BOLIVAR	1.0953
BOYACA	0.8299
CALDAS	1.2500
CASANARE	1.6552
CAUCA	1.0796
CESAR	0.6874
CUNDINAMARCA	0.4816
HUILA	0.3811
LA GUAJIRA	0.4312
META	0.6507
NARIÑO	0.8535
NORTE DE SANTANDER	0.6312
QUINDIO	0.2028
RISARALDA	0.1372
SANTANDER	0.3973
SUCRE	0.2341
TOLIMA	0.9968
VALLE DEL CAUCA	0.4483

Tabla 3-28: RMSE_MCAR_regresión por Departament

Imputación mediante coincidencia de medias predictiva

Para el conjunto de datos imputados mediante el metodo Coincidencia de medias predictiva, el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Prod_log}}_{ij} = 7.3184 + 0.1868 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

Al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$ también experimentan variaciones.
Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$\text{RMSE_MCAR_regresión} = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Prod_log}} - \text{Prod_log})^2} = 0{,}8820$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante regresión fue de 0.8820. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamento	RMSE_MCAR_match
ANTIOQUIA	0.6722
BOLIVAR	1.4735
BOYACA	0.7116
CALDAS	1.2479
CASANARE	2.1414
CAUCA	0.9262
CESAR	0.7341
CUNDINAMARCA	0.9402
HUILA	0.6045
LA GUAJIRA	0.3330
META	0.8874
NARIÑO	0.7624
NORTE DE SANTANDER	0.8912
QUINDIO	0.2256
RISARALDA	0.1475
SANTANDER	0.3666
SUCRE	0.2106
TOLIMA	0.5223
VALLE DEL CAUCA	0.2521

Tabla 3-29: RMSE_MCAR_match por Departamento

Discusión

1 RMSE de Modelos de Datos Completos y Imputados:

- El modelo original con datos completos tiene un RMSE de 0.2951, lo cual indica un buen ajuste del modelo a los datos originales.
- Al imputar datos bajo la condición MCAR, se observa que los RMSE varían dependiendo del método de imputación. El método LOCF tiene un RMSE de 0.4863, el método de la media tiene un RMSE de 0.6746, el método de regresión tiene un RMSE de 0.7788, y el método de coincidencia de medias predictiva tiene un RMSE de 0.8820.

RMSE_MCAR_locf	0.4863
RMSE_MCAR_media	0.6746
RMSE_MCAR_regresión	0.7788
RMSE_MCAR_match	0.8820

Tabla 3-30: Comparacion de los RMSE para la condicion MCAR de métodos de imputacion

2 Comparación de Métodos de Imputación:

- Los resultados sugieren que el método LOCF tiene un rendimiento superior en términos de RMSE en comparación con el método de la media, regresión y coincidencia de medias predictiva.
- Es crucial considerar que la elección del método de imputación puede depender de la naturaleza de los datos y la validez de las suposiciones subyacentes.

3 Variabilidad por Departamento:

- Al desglosar los RMSE por departamento, se observa que la variabilidad en la precisión del modelo es diferente para cada método de imputación y departamento.
- Para los departamentos como que presentan datos faltantes los métodos sugieren que el método LOCF tiene un rendimiento superior en términos de RMSE en comparación con el método de la media, regresión y coincidencia de medias predictiva. Deparatamentos como Casanare, Norte de santander Y Cesar que son los que mas datos faltantes presenta el RMSE para el modelo original fueron 0.3816, 0.2518 y 0.1368 respectivamente. En la siguiente Tabla **3-31** veresmos las comparación pra los métodos de imputacion para los departamentes mencionados anteriormente.

	Casanare	**Norte de santander**	**Cesar**
RMSE	0.3816	0.2518	0.1368
RMSE_MCAR_locf	0.5560	0.2544	0.5876
RMSE_MCAR_media	1.5998	0.5695	0.0559
RMSE_MCAR_regresión	1.2500	0.6312	1.0796
RMSE_MCAR_match	1.2479	0.8912	0.7341

Tabla 3-31: Comparacion de los RMSE_MCAR de los métodos de imputacion por departamentos que presentan datos faltantes

notamos que el métodos con mejor rendimiento es LOCF.

- En cambio los departamentos como que no presentan datos faltantes independientemente del método de imputación tienen un rendimiento en término de RMSE parecido al de los datos originales.

	Quindio	**Risalda**	**Sucre**
RMSE	0.1663	0.1342	0.1667
RMSE_MCAR_locf	0.1977	0.1727	0.2704
RMSE_MCAR_media	0.3097	0.2069	0.4412
RMSE_MCAR_regresión	0.2028	0.1372	0.2341
RMSE_MCAR_match	0.2256	0.3666	0.2106

Tabla 3-32: Comparacion de los RMSE_MCAR de los métodos de imputacion por departamentos que no presentan datos faltante

4 Implicaciones de los datos faltantes

- La imputación puede tener un impacto en la interpretación de los coeficientes del modelo. Los cambios en los parámetros, como $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$, deben considerarse al interpretar los resultados del modelo después de la imputación.
- Los métodos de imputación tiene un impacto significativo en los departamentos donde se registran datos faltantes.

3.2.5. Tratamiento para la condición MAR

Generamos datos faltantes bajo la condición MAR. A continuación, visualizamos gráficamente el porcentaje de datos faltantes creados por departamento.

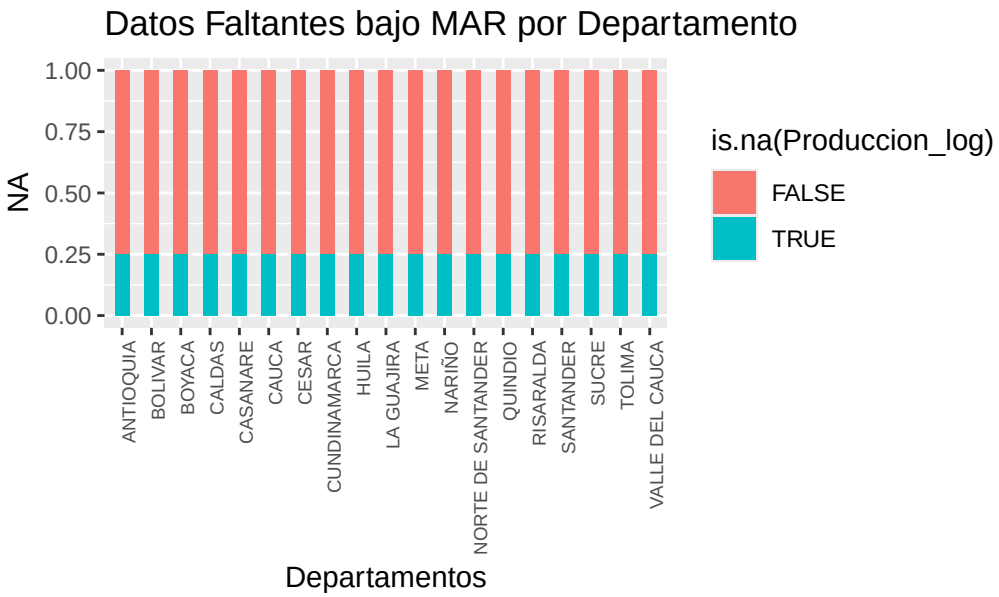

Figura 3-10: Porcenta de falta por departamente para la condicinon MAR

Se crearon datos faltantes para todas las observaciones en las que el año se encuentra en el rango de 2 a 4. En consecuencia, todos los departamentos muestran ausencia de datos. A continuación, procederemos con la aplicación de métodos de imputación.

Imputación mediante LOCF

Para el conjunto de datos imputados mediante el método de la última observación arrastrada (LOCF), el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Prod_log}}_{ij} = 7.4694 + 0.1704 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

Es importante destacar que, al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$ también varían.
Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$\text{RMSE_MCAR_regresión} = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Prod_log}} - \text{Prod_log})^2} = 0.5257$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante LOCF fue de 0.5257. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamento	**RMSE_MAR_locf**
ANTIOQUIA	0.1913
BOLIVAR	0.8090
BOYACA	0.3653
CALDAS	0.4797
CASANARE	0.5560
CAUCA	1.2542
CESAR	0.5876
CUNDINAMARCA	0.2644
HUILA	0.3519
LA GUAJIRA	0.3729
META	0.8005
NARIÑO	0.5211
NORTE DE SANTANDER	0.2544
QUINDIO	0.1977
RISARALDA	0.1727
SANTANDER	0.5047
SUCRE	0.2704
TOLIMA	0.4014
VALLE DEL CAUCA	0.3173

Tabla 3-33: RMSE_MAR_locf por Departamento

Imputación de la media

Para el conjunto de datos imputados mediante el metodo de la media, el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Prod_log}}_{ij} = 8.2075 + 0.0958 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

Como se mencionó anteriormente, al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$ también experimentan variaciones. Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$\text{RMSE_MCAR_regresión} = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Prod_log}} - \text{Prod_log})^2} = 0.7555$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante la media fue de 0.7555. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamento	**RMSE_MAR_media**
ANTIOQUIA	0.7555
BOLIVAR	0.6167
BOYACA	0.8320
CALDAS	0.6351
CASANARE	1.5987
CAUCA	1.0559
CESAR	0.3168
CUNDINAMARCA	0.5575
HUILA	0.5226
LA GUAJIRA	0.7220
META	0.6990
NARIÑO	1.3867
NORTE DE SANTANDER	0.5695
QUINDIO	0.3097
RISARALDA	0.2469
SANTANDER	0.3819
SUCRE	0.4412
TOLIMA	0.7613
VALLE DEL CAUCA	0.3309

Tabla 3-34: RMSE_MAR_media por Departamento

Imputación mediante regresión

Para el conjunto de datos imputados mediante el metodo regresión , el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Prod_log}}_{ij} = 7.8194 + 0.1349 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

Al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$ también experimentan variaciones. Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$\text{RMSE_MCAR_regresión} = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Prod_log}} - \text{Prod_log})^2} = 0.9933$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante regresión fue de 0.2451. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamento	**RMSE_MAR_regresión**
ANTIOQUIA	1.0074
BOLIVAR	1.0851
BOYACA	0.9119
CALDAS	1.0244
CASANARE	1.5648
CAUCA	1.0703
CESAR	0.5044
CUNDINAMARCA	0.9286
HUILA	0.9060
LA GUAJIRA	0.9214
META	0.6689
NARIÑO	1.5293
NORTE DE SANTANDER	1.1501
QUINDIO	0.3509
RISARALDA	0.3749
SANTANDER	0.6987
SUCRE	0.8523
TOLIMA	1.4602
VALLE DEL CAUCA	0.7582

Tabla 3-35: RMSE_MAR_regresión por Departamento

Imputación mediante Coincidencia de medias predictiva

Para el conjunto de datos imputados mediante el metodo Coincidencia de medias predictiva, el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Prod_log}}_{ij} = 7.3184 + 0.1868 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

Al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$ también experimentan variaciones. Posteriormente, para evaluar el rendimiento del modelo, se calcula

el RMSE utilizando la fórmula:

$$\text{RMSE_MCAR_regresión} = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Prod_log}} - \text{Prod_log})^2} = 1.2813$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante regresión fue de 1.2813. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamento	**RMSE_MAR_match**
ANTIOQUIA	1.5564
BOLIVAR	1.4945
BOYACA	0.9266
CALDAS	1.8407
CASANARE	1.5430
CAUCA	1.5917
CESAR	1.4894
CUNDINAMARCA	1.1933
HUILA	0.8069
LA GUAJIRA	1.0543
META	1.2633
NARIÑO	1.2681
NORTE DE SANTANDER	1.0027
QUINDIO	0.5755
RISARALDA	1.1755
SANTANDER	1.8759
SUCRE	0.6415
TOLIMA	1.1304
VALLE DEL CAUCA	0.9158

Tabla 3-36: RMSE_MAR_match por Departamento

Discusión

1 RMSE de Modelos de Datos Completos y Imputados:

- El modelo original con datos completos tiene un RMSE de 0.2951, lo cual indica un buen ajuste del modelo a los datos originales.
- Al imputar datos bajo la condición MAR, se observa que los RMSE varían dependiendo del método de imputación. El método LOCF tiene un RMSE de 0.5257, el método de la media tiene un RMSE de 0.7555, el método de regresión tiene un RMSE de 0.9933, y el método de coincidencia de medias predictiva tiene un RMSE de 1.2813.

RMSE_MAR_locf	0.5257
RMSE_MAR_media	0.7555
RMSE_MAR_regresión	0.9933
RMSE_MAR_match	1.2813

Tabla 3-37: Comparacion de los RMSE para la condicion MAR de métodos de imputacion

2 Comparación de Métodos de Imputación:

- Los resultados sugieren que el método LOCF tiene un rendimiento superior en términos de RMSE en comparación con el método de la media, regresión y coincidencia de medias predictiva.

- Es crucial considerar que la elección del método de imputación puede depender de la naturaleza de los datos y la validez de las suposiciones subyacentes.

3 Variabilidad por Departamento:

- Al desglosar los RMSE por departamento, se observa que la variabilidad en la precisión del modelo es diferente para cada método de imputación y departamento.

4 Implicaciones de los datos faltantes

- La imputación puede tener un impacto en la interpretación de los coeficientes del modelo. Los cambios en los parámetros, como $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$, deben considerarse al interpretar los resultados del modelo después de la imputación.

3.2.6. Tratamiento para la condición MNAR

Generamos datos faltantes bajo la condición MNAR. A continuación, visualizamos gráficamente el porcentaje de datos faltantes creados por departamento.

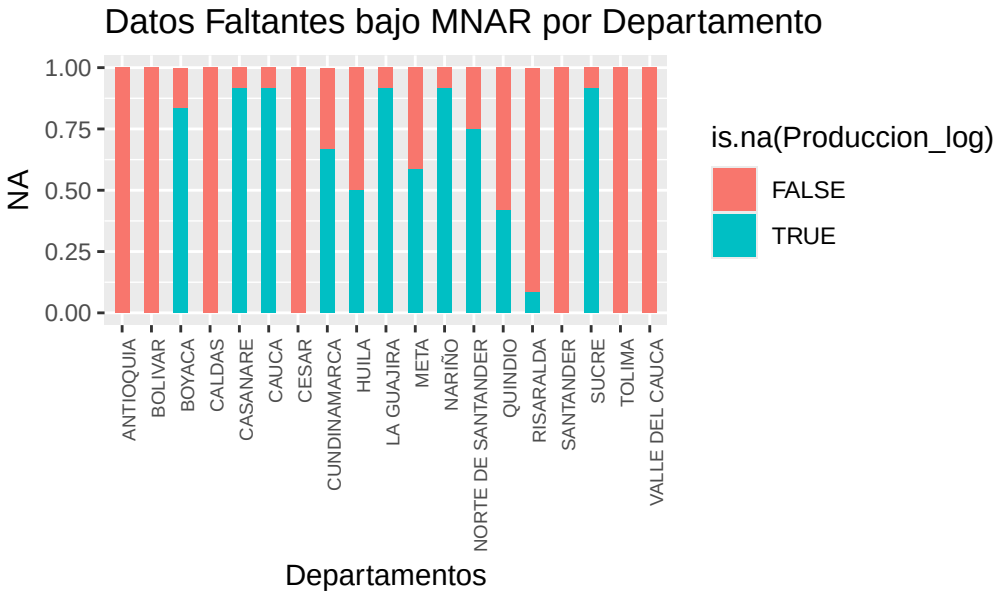

Figura 3-11: Porcenta de falta por departamente para la condicinon MNAR

Se crearon datos faltantes para las observaciones de la variable 'Prod_log' que son inferiores a la mediana de dicha variable. De los 19 departamentos, 12 presentan datos faltantes, siendo Casanares, Cauca, Huila y otros con la mayor cantidad de datos faltantes, de los 7 que no presentan datos fataltes algunos son Antioquia, Bolivar, Caldas y Cesar.

Imputación mediante LOCF

Para el conjunto de datos imputados mediante el método de la última observación arrastrada (LOCF), el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Prod_log}}_{ij} = 7.5557 + 0.1003 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

Es importante destacar que, al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$ también varían.
Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$\text{RMSE_MCAR_locf} = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Prod_log}} - \text{Prod_log})^2} = 0.4733$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante LOCF fue de 0.1163. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamento	**RMSE_MNAR_locf**
ANTIOQUIA	0.3128
BOLIVAR	0.5558
BOYACA	0.4315
CALDAS	0.5500
CASANARE	0.3469
CAUCA	0.3468
CESAR	0.3166
CUNDINAMARCA	0.6504
HUILA	0.6725
LA GUAJIRA	0.3464
META	1.1297
NARIÑO	0.3467
NORTE DE SANTANDER	0.3824
QUINDIO	0.4111
RISARALDA	0.2409
SANTANDER	0.2444
SUCRE	0.3463
TOLIMA	0.2305
VALLE DEL CAUCA	0.1468

Tabla 3-38: RMSE_MNAR_locf por Departamento

Imputación de la media

Para el conjunto de datos imputados mediante el metodo de la media, el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Prod_log}}_{ij} = 9.0017 + 0.0996 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

Como se mencionó anteriormente, al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$ también experimentan variaciones.
Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$\text{RMSE_MCAR_media} = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Prod_log}} - \text{Prod_log})^2} = 0.7196$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante la media fue de 0.7196. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamento	**RMSE_MNAR_media**
ANTIOQUIA	0.3618
BOLIVAR	0.5674
BOYACA	0.7191
CALDAS	0.5558
CASANARE	1.4143
CAUCA	1.3080
CESAR	0.3151
CUNDINAMARCA	0.7554
HUILA	0.6141
LA GUAJIRA	0.6052
META	0.8442
NARIÑO	1.2518
NORTE DE SANTANDER	0.6681
QUINDIO	0.3940
RISARALDA	0.2862
SANTANDER	0.2455
SUCRE	0.5353
TOLIMA	0.3223
VALLE DEL CAUCA	0.1494

Tabla 3-39: RMSE_MNAR_media por Departamento

Imputación mediante regresión

Para el conjunto de datos imputados mediante el metodo regresión , el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Prod_log}}_{ij} = 7.6293 + 0.1543 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

Al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta}_0$, $\widehat{\beta}_1$ y $\widehat{b_{0i}}$ también experimentan variaciones.
Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$\text{RMSE_MCAR_regresión} = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Prod_log}} - \text{Prod_log})^2} = 0.9767$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante regresión fue de 0.9767. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamento	**RMSE_MNAR_regresión**
ANTIOQUIA	0.3305
BOLIVAR	0.7968
BOYACA	1.2747
CALDAS	0.3597
CASANARE	1.7125
CAUCA	1.5538
CESAR	0.5236
CUNDINAMARCA	0.8066
HUILA	0.8545
LA GUAJIRA	1.3033
META	0.8092
NARIÑO	1.7330
NORTE DE SANTANDER	1.1249
QUINDIO	0.8982
RISARALDA	0.1592
SANTANDER	0.4279
SUCRE	0.7568
TOLIMA	0.5689
VALLE DEL CAUCA	0.3006

Tabla 3-40: RMSE_MNAR_regresión por Departamento

Imputación mediante Coincidencia de medias predictiva

Para el conjunto de datos imputados mediante el metodo Coincidencia de medias predictiva, el modelo lineal mixto se expresa de la siguiente manera:

$$\widehat{\text{Prod_log}}_{ij} = 9.0925 + 0.1022 \times \text{Año}_{ij} + \widehat{b}_{0j} + \widehat{b}_{1i} \times \text{Año}_{ij} + e_{ij}$$

Al tratarse de un conjunto de datos diferente, los parámetros $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{b_{0i}}$ y $\widehat{b_{1i}}$ también experimentan variaciones.
Posteriormente, para evaluar el rendimiento del modelo, se calcula el RMSE utilizando la fórmula:

$$\text{RMSE_MCAR_match} = \sqrt{\frac{1}{379}\sum_{i=1}^{379}(\widehat{\text{Prod_log}} - \text{Prod_log})^2} = 0.9044$$

El RMSE obtenido para el modelo construido con el conjunto de datos imputados mediante regresión fue de 0.9044. A continuación, se presentan los RMSE desglosados por departamento en la siguiente tabla:

Departamento	**RMSE_MNAR_match**
ANTIOQUIA	0.4258
BOLIVAR	0.5956
BOYACA	0.8869
CALDAS	0.5499
CASANARE	1.5677
CAUCA	1.5959
CESAR	0.3307
CUNDINAMARCA	0.7503
HUILA	0.8677
LA GUAJIRA	0.8437
META	1.1659
NARIÑO	1.4769
NORTE DE SANTANDER	1.0766
QUINDIO	0.7594
RISARALDA	0.2402
SANTANDER	0.2511
SUCRE	1.0725
TOLIMA	0.4521
VALLE DEL CAUCA	0.1507

Tabla 3-41: RMSE_MNAR_match por Departamento

Discusión

1 RMSE de Modelos de Datos Completos y Imputados:

- El modelo original con datos completos tiene un RMSE de 0.2951, lo cual indica un buen ajuste del modelo a los datos originales.

- Al imputar datos bajo la condición MNAR, se observa que los RMSE varían dependiendo del método de imputación. El método LOCF tiene un RMSE de 0.4733, el método de la media tiene un RMSE de 0.7196 , el método de regresión tiene un RMSE de 0.9767, y el método de coincidencia de medias predictiva tiene un RMSE de 0.9044.

RMSE_MNAR_locf	0.4733
RMSE_MNAR_media	0.7196
RMSE_MNAR_regresión	0.9767
RMSE_MNAR_match	0.9044

Tabla 3-42: Comparacion de los RMSE para la condicion MNAR de métodos de imputacion

2 Comparación de Métodos de Imputación:

- Los resultados sugieren que el método LOCF tiene un rendimiento superior en términos de RMSE en comparación con el método de la media, regresión y coincidencia de medias predictiva.
- Es crucial considerar que la elección del método de imputación puede depender de la naturaleza de los datos y la validez de las suposiciones subyacentes.

3 Variabilidad por Departamento:

- Al desglosar los RMSE por departamento, se observa que la variabilidad en la precisión del modelo es diferente para cada método de imputación y departamento.
- Para los departamentos como que presentan datos faltantes los métodos sugieren que el método LOCF tiene un rendimiento superior en términos de RMSE en comparación con el método de la media, regresión y coincidencia de medias predictiva. Deparatamentos como Bolivar, Caqueta, Casanares, Cesar que son algunos de los que mas datos faltantes presenta el RMSE para el modelo original fueron 0.4733, 0.71968, 0.9767 y 0.9044 respectivamente. En la siguiente tabla **3-43** veremos las comparación pra los métodos de imputacion para los departamentes mencionados anteriormente.

	Casanare	**Cauca**	**Huila**
RMSE	0.3816	0.7598	0.2308
RMSE_MCAR_locf	0.3469	0.3468	0.6725
RMSE_MCAR_media	1.414	1.3080	0.6141
RMSE_MCAR_regresión	1.7125	1.5538	0.8545
RMSE_MCAR_match	1.5677	1.5959	0.8677

Tabla 3-43: Comparacion de los RMSE_MNAR de los métodos de imputacion por departamentos que presentan datos faltantes

notamos que el métodos con mejor rendimiento es LOCF.

- En cambio los departamentos como que no presentan datos faltantes independientemente del método de imputación tienen un rendimiento en término de RMSE parecido al de los datos originales.

	Antioquia	**Bolivar**	**Caldas**
RMSE	0.1578	0.1070	0.1549
RMSE_MCAR_locf	0.3128	0.0.5558	0.5500
RMSE_MCAR_media	0.3618	0.5674	0.5558
RMSE_MCAR_regresión	0.3305	0.7968	0.3597
RMSE_MCAR_match	0.4258	0.5956	0.5499

Tabla 3-44: Comparacion de los RMSE_MNAR de los métodos de imputacion por departamentos que no presentan datos faltante

4 Implicaciones de los datos faltantes

- La imputación puede tener un impacto en la interpretación de los coeficientes del modelo. Los cambios en los parámetros, como $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{b_{0i}}$ y $\widehat{b_{0i}}$, deben considerarse al interpretar los resultados del modelo después de la imputación.

3.2.7. Discusión general para las tres condiciones

Para la base de datos del Rendimiento de Cultivos en Colombia por año, se generaron datos faltantes siguiendo las tres condiciones propuestas por D. Rubin (1976). Posteriormente, se aplicó un modelo lineal con intercepto y pendiente aleatorio. Se llevaron a cabo métodos de imputación para los datos faltantes.

Los resultados revelaron que al comparar los errores cuadráticos medios de los modelos, estos variaron dependiendo del método de imputación utilizado. Específicamente, se encontró que el método LOCK tuvo un rendimiento superior en todas las condiciones evaluadas, tanto en el RMSE del modelo como en el RMSE entre los departamentos (intra-sujeto). En conclusión, el método LOCK demostró consistentemente ser el más efectivo en términos de rendimiento.

veamos un grafico donde notemos la comparacion de los RMSE de todos lo métodos en las tres condiciones.

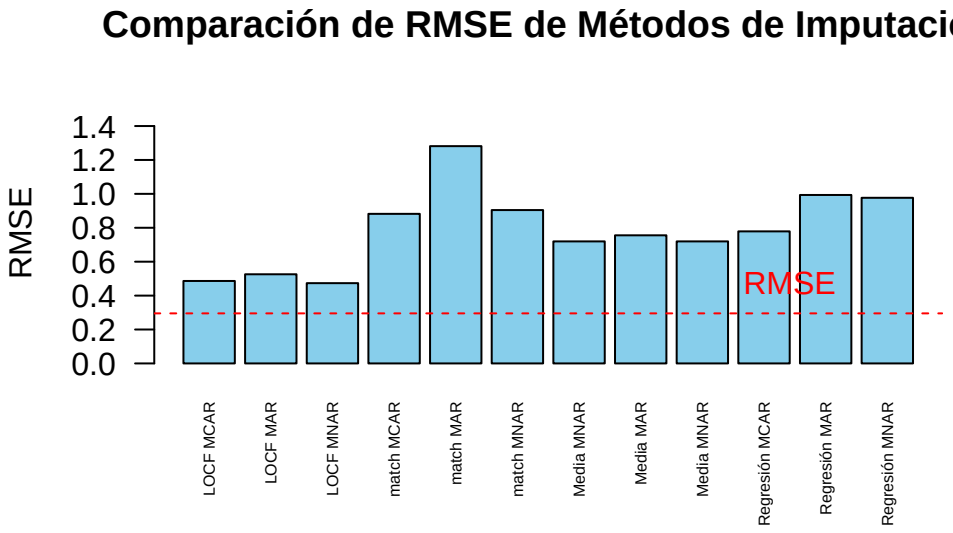

Figura 3-12: Comparación de RMSE de Métodos de Imputación

4 Conclusiones y recomendaciones

4.1. Conclusiones

Considerando el objetivo general de este trabajo, que es comparar distintas metodologías utilizadas en la literatura para abordar el problema de datos faltantes en estudios longitudinales, se observa que los datos faltantes pueden deberse a tres condiciones distintas: Perdidos Completamente al Azar (MCAR), Perdidos al Azar (MAR) y No Perdidos al Azar (NMAR), propuestas por Rubin (1976). Estas condiciones tienen implicaciones tanto en la selección de métodos de imputación como en la obtención de resultados.

En este estudio, se utilizaron dos bases de datos longitudinales provenientes del Portal de Datos Abiertos del Estado colombiano. Estas bases de datos son equilibradas, lo que significa que todos los sujetos (en este caso, los departamentos) tienen la misma cantidad de observaciones. Se ajustaron modelos lineales mixtos a ambas bases de datos: un modelo lineal mixto con intercepto aleatorio para la primera base de datos y un modelo lineal mixto con intercepto y pendiente aleatorio para la segunda base de datos. Posteriormente, se crearon datos faltantes bajo las tres condiciones mencionadas y se aplicaron diferentes métodos de imputación para ajustar los mismos modelos lineales mixtos a cada base de datos.

Para comparar el rendimiento de los modelos, se propuso calcular el RMSE de cada modelo y el RMSE intra-sujeto. Los resultados mostraron que, al imputar datos bajo cualquier condición y para ambas bases de datos, el método LOCF tuvo un mejor rendimiento que otros métodos de imputación.

Además, al desglosar los RMSE por departamento, se observó que la variabilidad en la precisión del modelo difiere para cada método de imputación y departamento. Esto indica que la elección del método de imputación puede tener un impacto significativo en la interpretación de los coeficientes del modelo, ya que estos pueden variar dependiendo de la condición de los datos faltantes.

En resumen, este estudio resalta la importancia de considerar cuidadosamente las condiciones de los datos faltantes y la selección del método de imputación en estudios longitudinales, ya que estas decisiones pueden afectar la precisión y la interpretación de los resultados del modelo.

Los resultados obtenidos en este estudio permiten establecer una guía para la selección de métodos de imputación en estudios longitudinales. Entre los métodos evaluados, el método LOCF destacó por su simplicidad y efectividad en condiciones donde los datos se pierden completamente al azar (MCAR). Sin embargo, su rendimiento fue menos consistente en

situaciones donde los datos se perdieron al azar (MAR) o no al azar (NMAR), lo que sugiere que su uso debe ser considerado con cautela en estos contextos.
El método de imputación múltiple (MI) mostró un buen rendimiento en condiciones MAR, proporcionando estimaciones más precisas y menos sesgadas que LOCF en estas situaciones. MI se beneficia de su capacidad para incorporar la incertidumbre en los datos faltantes, lo que lo hace una opción robusta cuando se espera que los datos faltantes dependan de información observada.
El método LOCK demostró ser el más efectivo en términos generales. Su capacidad para manejar datos faltantes bajo todas las condiciones (MCAR, MAR y NMAR) lo convierte en una opción versátil y potente. LOCK mostró el menor error cuadrático medio (RMSE) tanto a nivel de modelo global como intra-sujeto, lo que indica su superioridad en la recuperación de la estructura de los datos originales y en la reducción de sesgos.
En conclusión, la elección del método de imputación debe basarse en la naturaleza de los datos faltantes y en el contexto específico del estudio longitudinal. Para situaciones MCAR, LOCF puede ser una opción simple y eficaz. En condiciones MAR, el método de imputación múltiple es preferible debido a su capacidad para manejar la incertidumbre de los datos faltantes. Para una mayor robustez y precisión en una variedad de condiciones de datos faltantes, el método LOCK se recomienda como la mejor opción. Estos hallazgos subrayan la importancia de adaptar la estrategia de imputación a las características específicas de los datos faltantes para maximizar la precisión y fiabilidad de los análisis longitudinales.

4.2. Recomendaciones

Basado en los hallazgos de este estudio, se sugieren las siguientes recomendaciones para futuras investigaciones:

- Realizar análisis similares utilizando diferentes conjuntos de datos y condiciones de datos faltantes para validar los resultados obtenidos.
- Explorar otros métodos de imputación además de los utilizados en este estudio para evaluar su efectividad en diferentes escenarios.
- Considerar el impacto de la imputación de datos en la interpretación de los resultados del modelo y explorar estrategias para mitigar cualquier sesgo introducido por la imputación.
- Investigar cómo la elección del método de imputación puede afectar la precisión de las predicciones del modelo en estudios longitudinales.

Implementar estas recomendaciones podría mejorar nuestra comprensión de cómo abordar de manera efectiva el problema de los datos faltantes en estudios longitudinales y mejorar la calidad de los análisis y las inferencias realizadas en este contexto.

Referencias

Casella, G., y Berger, R. (2002). *Statistical reference.* Duxbury Advanced Series.

Cooper, D., y Thompson, R. (1977). Note concerning the akaike and hannan estimation procedures for an autoregressive-moving average processa note on the estimation of the parameters of the autoregressive-moving average process. *Biometrika*, *64*(3), 625–628.

Correa Morales, J. C., y Salazar Uribe, J. C. (2016). Introducción a los modelos mixtos. *Escuela de Estadística*.

Diggle, P., Heagerty, P., Liang, K., y Zeger, S. (2002). *Analysis of longitudinal data. london.* Oxford University Press.

Diggle, P., y Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *43*(1), 49-73. Descargado de `https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2986113` doi: https://doi.org/10.2307/2986113

Faraway, J. J. (2016). *Extending the linear model with r: generalized linear, mixed effects and nonparametric regression models.* Chapman and Hall/CRC.

Fitzmaurice, G., Davidian, M., Verbeke, G., y Molenberghs, M. (2008). *Longitudinal data analysis..* London: Chapman and Hall.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American statistical association*, *72*(358), 320–338.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., y Herring, A. H. (2005). Missing-data methods for generalized linear models. *Journal of the American Statistical Association*, *100*(469), 332-346. Descargado de `https://doi.org/10.1198/016214504000001844` doi: 10.1198/016214504000001844

Laird, N. M., y Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 963–974.

Laird, N. M., Ware, J. H., y Fitzmaurice, G. (2004). *Missing data and dropout: Overview of concepts and methods. applied longitudinal analysis, 2nd edition.* Hoboken, NJ: John Wiley & Sons, Inc,.

McNeish, D. (2017). Missing data methods for arbitrary missingness with small samples. *Journal of Applied Statistics*, *44*(1), 24-39. Descargado de `https://doi.org/10.1080/02664763.2016.1158246` doi: 10.1080/02664763.2016.1158246

Nooraee, N., Molenberghs, G., Ormel, J., y Edwin R, V. d. H. (2018).

Strategies for handling missing data in longitudinal studies with questionnaires. *Journal of Statistical Computation and Simulation*, *88*(17), 3415-3436. Descargado de https://doi.org/10.1080/00949655.2018.1520854 doi: 10.1080/00949655.2018.1520854

Patterson, H. D., y Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554.

Rahim, J., Chen, M., y Lipsitz, S. (2001). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *88*(469), 551–564.

Roderick, J., y Little, A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*(431), 1112-1121. Descargado de https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476615 doi: 10.1080/01621459.1995.10476615

Rubin, D. (1976). Inference and missing data. biometrika , 63 (3), . *Biometrika*, *63*(3), 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.

Schafer, J., y Graham, J. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

Troxel, A. B., Harrington, D. P., y Lipsitz, S. R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *47*(3), 425-438. Descargado de https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9876.00119 doi: https://doi.org/10.1111/1467-9876.00119

Van Buuren, S. (2012). *Flexible imputation of missing data. boca raton*. CRC Press.

Van Buuren, S., Boshuizen, H. C., y Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, *18*(6), 681–694.

van Ginkel, J. R., Linting, M., Rippe, R. C. A., y van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, *102*(3), 297-308. Descargado de https://doi.org/10.1080/00223891.2018.1530680 (PMID: 30657714) doi: 10.1080/00223891.2018.1530680

Verbeke, G., Molenberghs, G., y Verbeke, G. (1997). *Linear mixed models for longitudinal data*. Springer.

Verbyla, A. P. (1990). A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics*, *32*(2), 227–230.

West, B. T., Welch, K. B., y Galecki, A. T. (2022). *Linear mixed models: a practical guide using statistical software*. Crc Press.