

*Análisis de datos longitudinales con muestreo destructivo:  
una perspectiva desde los modelos lineales mixtos*

CAMILO ANDRÉS AVELLANEDA GARCÍA

ESTADÍSTICO

CÓDIGO: 1018465894



UNIVERSIDAD NACIONAL DE COLOMBIA

FACULTAD DE CIENCIAS

DEPARTAMENTO DE ESTADÍSTICA

BOGOTÁ, D.C.

FEBRERO DE 2020

*Análisis de datos longitudinales con muestreo destructivo:  
una perspectiva desde los modelos lineales mixtos*

CAMILO ANDRÉS AVELLANEDA GARCÍA

ESTADÍSTICO

CÓDIGO: 1018465894

DISERTACIÓN PRESENTADA PARA OPTAR AL TÍTULO DE  
MAGISTER EN ESTADÍSTICA

DIRECTOR

OSCAR ORLANDO MELO MARTÍNEZ, PH.D.

DOCTOR EN ESTADÍSTICA

LÍNEA DE INVESTIGACIÓN

DISEÑO EXPERIMENTAL



UNIVERSIDAD NACIONAL DE COLOMBIA

FACULTAD DE CIENCIAS

DEPARTAMENTO DE ESTADÍSTICA

BOGOTÁ, D.C.

FEBRERO DE 2020

**Título en español**

Análisis de datos longitudinales con muestreo destructivo: una perspectiva desde los modelos lineales mixtos.

**Title in English**

Analysis of longitudinal data with destructive sampling: a perspective using linear mixed models.

**Resumen:** En este documento se realiza una comparación de modelos de regresión para el caso en donde se tienen datos longitudinales con muestreo destructivo de unidades observacionales, las cuales provienen de unidades experimentales que son medidas en todos los tiempos del análisis. La comparación se hace a partir de modelos de regresión con efectos fijos y mixtos, entre los cuales se encuentra un símil que se utiliza para datos denominados como pseudo-panel, y uno de análisis de varianza multivariado. Para comparar los modelos se utilizó el cuadrado medio del error. Esto se realizó mediante simulación y una aplicación a datos de la vida real que hacen referencia a los puntajes en las pruebas Saber 11 aplicadas a estudiantes en Colombia.

**Abstract:** In this work I make a comparison of regression models for the case where we have longitudinal data with destructive sampling of observational units which come from experimental units that are measured in every time of the analysis. The comparison is made from linear models of fixed and mixed effects, using a model used for pseudo-panel data too, and one corresponding to the multivariate analysis of variance case. To carry out the comparison between the different models I use the mean square error. It was made through simulation and using the scores of the test Saber 11 applied to students in Colombia.

**Palabras clave:** Datos longitudinales, muestreo destructivo, unidades observacionales y experimentales, efectos fijos, efectos mixtos, análisis de varianza multivariado, cuadrado medio del error.

**Keywords:** Longitudinal data, destructive sampling, observational and experimental units, mixed effects, multivariate analysis of variance, mean square error.

# Nota de aceptación

Trabajo de tesis

“Mención ”

---

Jurado

---

Jurado

---

Director  
Oscar Orlando Melo Martínez

---

Codirector

Bogotá, D.C., Febrero 06 de 2020

---

---

## Dedicado a

---

---

A mis padres, Andrea y Hernán.

---

---

## Agradecimientos

---

---

Agradezco principalmente a la Universidad Nacional de Colombia por su contribución a mi formación profesional, a mi director de trabajo de grado Oscar Orlando Melo por su atención y guía durante la elaboración de este documento, y por último, a Mónica Cañón por su apoyo incondicional.

---

---

# Índice general

---

---

<b>Índice general</b>	<b>I</b>
<b>Índice de tablas</b>	<b>III</b>
<b>Índice de figuras</b>	<b>IV</b>
<b>Introducción</b>	<b>VI</b>
<b>1. Marco teórico</b>	<b>1</b>
1.1. Modelo lineal multivariado . . . . .	3
1.1.1. Pruebas de hipótesis . . . . .	4
1.1.1.1. Prueba de paralelismo . . . . .	6
1.1.1.2. Prueba entre grupos . . . . .	6
1.1.1.3. Prueba entre tiempos . . . . .	6
1.2. Modelo lineal mixto . . . . .	7
1.2.1. Estimación de los MLM mediante máxima verosimilitud . . . . .	9
1.2.2. Prueba del cociente de verosimilitudes . . . . .	14
1.3. Modelos para datos pseudo-panel . . . . .	14
1.4. Paquetes estadísticos . . . . .	15
<b>2. Simulación de datos</b>	<b>17</b>
2.1. Descripción de la simulación . . . . .	17
2.1.1. Modelo exclusivo de efectos fijos . . . . .	19
2.1.2. Aplicación del modelo para pseudo-panel . . . . .	21
2.1.3. Modelo de efectos mixtos propuesto . . . . .	22
2.1.4. Modelo de análisis de varianza multivariado . . . . .	24
2.1.5. Funciones en R . . . . .	25
2.2. Resultados de la simulación . . . . .	26

---

<b>3. Aplicación a datos reales</b>	<b>37</b>
3.1. Análisis descriptivo . . . . .	38
3.2. Aplicación de los modelos estadísticos . . . . .	40
<b>Conclusiones</b>	<b>46</b>
<b>Trabajo futuro</b>	<b>47</b>
<b>Código de las funciones programadas en R</b>	<b>48</b>

---

---

## Índice de tablas

---

---

1.1. Conformaciones de los vectores respuesta por grupos para el modelo lineal multivariado. . . . .	4
1.2. Formulas para las estadísticas de prueba para el análisis de varianza multivariado, las cuales se calculan a partir de los valores propios de la matriz $Q_e^{-1}Q_h$ se denotan por $\lambda_i$ , mientras que el valor propio más alto se denota como $\lambda_1$ . . . . .	5
2.1. Tabla para ilustrar el esquema de muestreo considerado en la simulación. . .	18
2.2. Tabla de análisis de varianza para el modelo dado en la ecuación 2.2. . . . .	21
2.3. Tabla de análisis de varianza para el modelo dado en la ecuación (2.3). . . .	22
2.4. Tabla de análisis de varianza para el modelo dado en la ecuación (2.4). . . .	25
3.1. Tabla de análisis de varianza para el modelo de la ecuación (3.1). . . . .	40
3.2. Tabla de análisis de varianza para el modelo de la ecuación (3.2). . . . .	41
3.3. Tabla de análisis de varianza para el modelo de la ecuación (3.3). . . . .	42
3.4. Tabla de análisis de varianza para el Manova cuya respuesta son los promedios de puntajes en matemáticas en los grupos conformados por estudiantes dentro de los colegios. . . . .	42
3.5. Resumen de los modelos aplicados a los datos del puntaje en matemáticas, en términos del cuadrado medio del error y la correlación entre las estimaciones y los valores observados. . . . .	43
3.6. Tabla con los valores $p$ asociados a la prueba de Anderson-Darling para juzgar la hipótesis de normalidad de los efectos aleatorios incluidos en el modelo de la ecuación (3.3). . . . .	43
3.7. Tabla con los valores $p$ asociados a la prueba de Bartlett para juzgar la hipótesis de homocedasticidad de los efectos aleatorios incluidos en el modelo de la ecuación (3.3). . . . .	44

---

---

## Índice de figuras

---

---

1.1.	Diagrama de dispersión de las variables $y_{ij}$ y $x_{ij}$ simuladas con una recta estimada para todo el conjunto de datos. . . . .	8
1.2.	Diagrama de dispersión de las variables $y_{ij}$ y $x_{ij}$ simuladas con rectas estimadas para cada una de las UE. . . . .	9
2.1.	Diagrama de dispersión que ilustra el esquema de muestreo utilizado, cuyas líneas punteadas representan los perfiles completos para cada UO y los puntos son las observaciones seleccionadas en cada uno de los tiempos. . . . .	20
2.2.	Diagrama de dispersión que ilustra el esquema de muestreo utilizado, cuyas líneas punteadas representan los perfiles completos para cada UO y los colores de los puntos son agrupaciones de valores superiores e inferiores de UO en cada tiempo. . . . .	24
2.3.	<i>CME</i> para cien simulaciones del esquema de muestreo destructivo para los cuatro modelos. . . . .	27
2.4.	Diagrama de dispersión de los puntos observados en el tiempo junto con las estimaciones realizadas por los modelos de promedios (línea roja) y el propuesto en este documento (línea azul), el cual considera dos subgrupos dentro de las UE, uno de valores altos y otro de valores bajos. . . . .	28
2.5.	<i>CME</i> para cien simulaciones del esquema de muestreo destructivo para el modelo descrito en Deaton (1985) y el propuesto en este documento. . . . .	29
2.6.	Diagramas de cajas y bigotes para las distribuciones de los valores $p$ de las pruebas de significancia de la interacción para cada uno de los modelos y bajo distintas diferencias máximas entre los coeficientes $AT_{ij}$ . . . . .	30
2.7.	Diagramas de cajas y bigotes para las distribuciones de los valores $p$ de las pruebas realizadas para juzgar $H_0^{(2)}$ para cada uno de los modelos y bajo distintas diferencias máximas entre los coeficientes $A_i$ . . . . .	31
2.8.	Diagramas de cajas y bigotes para las distribuciones de los valores $p$ de las pruebas realizadas para juzgar $H_0^{(3)}$ para cada uno de los modelos y bajo distintas diferencias máximas entre los coeficientes $T_j$ . . . . .	32
2.9.	Correlogramas para los subgrupos de UO dentro de las diferentes UE y para los dos tratamientos en consideración en los residuales de uno de los ajustes del modelo dado en la ecuación (2.4). . . . .	33

---

2.10. Correlogramas para las diferentes UE y para los dos tratamientos en consideración en los residuales de uno de los ajustes del modelo dado en la ecuación (2.3). . . . .	34
2.11. Correlogramas para los subgrupos de UO dentro de las diferentes UE y para los dos tratamientos en consideración en los residuales de uno de los ajustes del modelo dado en la ecuación (2.4) cuando se tiene un porcentaje de UO fijo. . . . .	35
2.12. Correlogramas para las diferentes UE y para los dos tratamientos en consideración en los residuales de uno de los ajustes del modelo dado en la ecuación (2.3) cuando se tiene un porcentaje de UO fijo. . . . .	36
3.1. Puntaje promedio de las pruebas Saber 11 en matemáticas por género y zona de ubicación (gráfico a la izquierda), y por periodo de medición y género (gráfico a la derecha). . . . .	38
3.2. Puntaje promedio de las pruebas Saber 11 en matemáticas por periodo de medición y por género de los estudiantes del colegio. . . . .	39
3.3. Puntaje promedio de las pruebas Saber 11 en matemáticas por periodo de medición y naturaleza del colegio. . . . .	39
3.4. Densidad estimada para los efectos aleatorios $b_n$ en el modelo dado en la ecuación (3.3). . . . .	43
3.5. Densidad estimada para los efectos aleatorios $\eta_{nm}$ en el modelo dado en la ecuación (3.3). . . . .	44
3.6. Densidad estimada para los residuales $\epsilon_{ijklmnr}$ en el modelo dado en la ecuación (3.3). . . . .	45

---

---

## Introducción

---

---

En experimentación se dan situaciones en las cuales es de interés para el investigador detectar diferencias en una variable determinada por el estudio, entre grupos conformados de manera particular, considerando uno o múltiples factores. A su vez, en ocasiones las unidades experimentales involucradas en el estudio son observadas en repetidas situaciones, lo cual conlleva a que los modelos de regresión usuales no sean los más apropiados, ya que dentro de sus supuestos se encuentra la independencia entre mediciones, lo que se incumple al observar a los mismos individuos en repetidas ocasiones. Dependiendo del tipo de experimento, hay situaciones en donde la medición implica la destrucción de la unidad, por ejemplo en control de calidad de ciertos aparatos, observación de características de frutos, ramas u hojas de árboles que no podrían ser medidas en momentos posteriores a su recolección o en estudios de tipo panel que consideran individuos diferentes en cada tiempo por dificultades o ausencias al momento de requerir a los mismos sujetos en todos los periodos del análisis, entre otros.

A pesar que las mediciones realizadas se hacen sobre diferentes unidades, en ciertas ocasiones éstas comparten características, es decir que provienen de una unidad más grande, que es considerada en todos los tiempos. Esto permite suponer que sus observaciones repetidas podrían ser similares. Para ejemplificar este concepto, suponga el caso en donde se desea evaluar el rendimiento en matemáticas de los estudiantes de un salón de clase, para lo cual se toman dos muestras aleatorias diferentes para dos momentos del calendario académico. A pesar de que las muestras aleatorias no tengan estudiantes en común se puede pensar en que su rendimiento será similar, considerando que comparten profesores, instalaciones y posiblemente condiciones socioeconómicas relacionadas con el plantel educativo. Este hecho hace que la regresión lineal simple no sea la más adecuada para describir variables de interés en estos contextos.

Dentro de la literatura se encuentran diversos ejemplos de este tipo de datos donde se hacen mediciones o encuestas a grupos de individuos periódicamente, pero no necesariamente a las mismas personas, lo cual imposibilita hacer seguimientos sobre unidades en particular. Deaton (1985) propone un modelo para una muestra *“pseudo-panel”*, que se construye con lo que denomina como cohortes, las cuales representan grupos de individuos con características similares, con la condición que cada individuo pertenezca únicamente a una cohorte durante todo el análisis y mediante estos grupos se ajustan efectos fijos correspondientes a cada una de las diferentes cohortes. El ejemplo que se trabaja allí es la creación de las agrupaciones mediante la fecha de nacimiento de las personas, de tal forma que cada uno pertenece únicamente a un grupo en el análisis realizado. Muchas publicaciones que abordan diferentes contextos se han elaborado a partir del modelo propuesto por Deaton (1985) (Ver Tsai et al. (2014), Tovar et al. (2012), Sprietsma (2012), Antman

& McKenzie (2007), entre otros). Este documento aborda las situaciones en donde las unidades observadas (UO) en un tiempo no son consideradas en tiempos posteriores por condiciones propias del experimento, lo cual se denomina como “*muestreo destructivo*” de la UO, con la salvedad que todas las unidades provienen de unidades experimentales (UE) de manera preestablecida, que son consideradas en todos los tiempos, a diferencia del modelo de “*pseudo-panel*”, donde las agrupaciones son construidas de manera adecuada o conveniente, de acuerdo a criterios del investigador.

A partir de la propuesta de Deaton (1985) se particiona la población en las cohortes mencionadas y considerando que dichos segmentos abarcan todo el conjunto, los efectos incluidos en los modelos lineales de regresión son efectos fijos, a diferencia del caso que se trabaja en este documento donde se utiliza el modelo propuesto allí, pero además se incluyen efectos aleatorios asociados a las diferentes cohortes, debido a que generalmente se trabaja únicamente con un subconjunto de UE pertenecientes a la población objetivo y no la totalidad de unidades.

En este trabajo se hace una propuesta de un modelo lineal de regresión de efectos mixtos para el análisis de mediciones longitudinales de UO con muestreo destructivo provenientes de diferentes UE desde un tiempo inicial. Para determinar el mejor modelo estadístico a aplicar en este contexto se tuvieron en cuenta un modelo lineal de efectos mixtos, un símil con el modelo de “*pseudo-panel*”, un modelo lineal multivariado y uno que asume independencia entre todas las observaciones ajustado mediante métodos de regresión tradicionales. La comparación de los diferentes modelos utilizados se lleva a cabo mediante el cuadrado medio del error (*CME*).

A pesar de que las pruebas relacionadas con el análisis de varianza multivariado tienen un rendimiento similar al de los modelos lineales de efectos mixtos en este contexto, se determinó que su rendimiento en cuanto a la estimación de las observaciones no es buena, ya que presenta un *CME* más grande, seguida de los modelos que solo incluyen efectos fijos, y por último, aquellos en los que se tienen efectos mixtos. Dentro de esta última categoría se tuvo en cuenta el modelo descrito por Deaton (1985) y uno propuesto en este documento que establece grupos de UO, con el objetivo de obtener estadísticas de prueba a partir del *CME* más pequeño.

Este documento está compuesto de la siguiente manera: El capítulo 1 describe las metodologías estadísticas a utilizar en el desarrollo del modelo propuesto, en el capítulo 2 se describe el proceso para llevar a cabo la comparación entre metodologías mediante simulación, mientras que en el capítulo 3 se realiza una aplicación de éstas metodologías a una base de datos que contiene los puntajes de las pruebas Saber 11 realizadas por estudiantes colombianos de los diferentes municipios durante los años 2013-2018, considerando que la mayoría de estudiantes que presentan el examen no vuelven a hacerlo en años posteriores, seguido a esto se presentan algunas conclusiones, recomendaciones para trabajos futuros y las respectivas referencias bibliográficas consultadas para este documento.

# CAPÍTULO 1

---

---

## Marco teórico

---

---

El diseño experimental es una rama de la estadística que se encarga de determinar si factores particulares influyen en el cambio de una variable de interés. De esta manera en aquellos casos donde se determine que hay una influencia importante, esta área proporciona herramientas para estimar su magnitud y determinar la existencia de diferencias significativas entre los niveles del factor en cuestión. En casos donde la experimentación incluye por lo menos una fuente de variación, se denomina “*tratamiento*” como cada combinación específica de niveles de los factores en consideración (Melo et al. 2007, p. 9).

De acuerdo con Federer & King (2007, p. 2) la unidad experimental (UE) es la cantidad más pequeña de material experimental a la cual se puede aplicar un tratamiento. La unidad observacional (UO) es la parte más pequeña en que se puede recolectar la medición de la variable de interés. Para ejemplificar la diferencia entre UE y UO se utiliza el contexto del ejemplo 1, en cuyo caso los tratamientos serán las metodologías a aplicar, que a su vez se aplicarán sobre la clase completa (UE) y sus resultados se evaluarán sobre cada uno de los estudiantes (UO).

**Ejemplo 1.** *Suponga el caso en donde se desea evaluar el rendimiento en matemáticas de los estudiantes de varios salones de clase considerando dos metodologías de enseñanza. Para este ejercicio se seleccionan muestras diferentes de estudiantes en cinco meses diferentes del año escolar, considerando que la selección de uno de ellos imposibilita que sea seleccionado posteriormente.*

Para la obtención de conclusiones efectivas Hinkelmann (2011, p. 45) enuncia los tres principios a tener en cuenta dentro de la planeación del experimento, los cuales son replicación, aleatorización y control local. La replicación se tiene para obtener una medición del error experimental. La aleatorización es necesaria ya que provee las bases para obtener un test válido de significancia al destruir cualquier sistema de correlación que pueda existir entre las unidades. La presencia de correlación positiva entre unidades supone una varianza más grande que el caso en que sean independientes, lo cual se traduce en un aumento de la probabilidad de cometer el error tipo I. Por último, la función principal del control local es la de eliminar los efectos de fuentes conocidas de variación (Melo et al. 2007, p. 17-21).

Las metodologías utilizadas dentro del contexto del diseño experimental hacen uso de los modelos regresión, los cuales tratan de explicar la dispersión de  $y_i$  denominada como

variable dependiente (respuesta) en términos de un vector de  $p$  variables independientes denotado como  $\mathbf{x}_i$ , mediante la ecuación (1.1), donde  $\mathbf{y} = [y_1 \ y_2 \ \cdots \ y_n]^t$  es el vector de respuestas, el cual se descompone mediante una parte determinista dada por  $\mathbf{X}\boldsymbol{\beta}$ , con  $\mathbf{X} = [\mathbf{1}_n \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_n]$  una matriz de información auxiliar o de variables independientes y  $\boldsymbol{\beta}^t = [\beta_0 \ \beta_1 \ \beta_2 \ \cdots \ \beta_p]$  un vector de coeficientes donde cada uno corresponde a una variable independiente y representan el crecimiento o decrecimiento esperado en la respuesta por el incremento de una unidad en cada covariable manteniendo las otras constantes. Por otro lado la parte aleatoria corresponde a un vector de errores  $\mathbf{e} = [\epsilon_1 \ \epsilon_2 \ \cdots \ \epsilon_n]^t$ , que para el caso general se asumen independientes e idénticamente distribuidos, es decir que  $\epsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$ . De los supuestos considerados sobre el modelo de regresión usual depende el proceso de estimación y evaluación de la significancia de parámetros. Para una lectura y revisión más detallada relacionada con el modelo de regresión lineal simple el lector puede referirse a Carmona (2005) o Montgomery et al. (2012).

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad (1.1)$$

Para fines de este documento las “réplicas” se definen como el número de UE asignadas a un tratamiento particular y se denotarán por  $r_j$  para el  $j$ -ésimo tratamiento. El modelo básico dentro de la teoría de diseño de experimentos se denomina diseño completamente aleatorizado y es aquel que considera un único factor  $A$  con  $a$  niveles como variable independiente, cada uno con  $r$  réplicas, en cuyo caso se desea determinar si hay diferencias estadísticamente significativas entre los grupos determinados por el factor. En este caso la ecuación (1.2) se asocia al modelo lineal, donde  $P \otimes Q$  es el producto kronecker entre las matrices  $P$  y  $Q$ ,  $\mathbf{1}_t$  es un vector  $t$  dimensional cuyas componentes son 1 en todas sus entradas e  $\mathbf{I}_t$  es la matriz idéntica de orden  $t$ . Por otro lado,  $\mu$  es la media global y los  $A_j$  son los efectos de cada uno de los  $a$  niveles del factor en consideración. Para determinar la presencia o ausencia de diferencias significativas, la varianza de la variable respuesta se divide de manera adecuada en sumas de cuadrados debidas al modelo y otra asociada al error, lo cual junto a los grados de libertad conforman la tabla de análisis de varianza (Montgomery 2017).

$$\mathbf{Y} = [\mathbf{1}_{a \times r} \ \mathbf{I}_a \otimes \mathbf{1}_r] [\mu \ A_1 \ A_2 \ \cdots \ A_a]^t + \mathbf{E} \quad (1.2)$$

De una manera análoga en la literatura se puede encontrar una gran variedad de diseños estadísticos para diferentes tipo de experimentos, los cuales dependen de los factores a incluir como variables independientes, su tipo, los controles locales, el número de réplicas e inclusive posibles interacciones entre los diferentes factores. Una manera de representar los factores incluidos en los modelos lineales utilizados en este contexto es mediante diagramas de estructura, a partir de los que se pueden deducir las tablas de análisis de varianza para cualquier modelo (Melo et al. 2007, cap. 4).

En situaciones en que se tienen múltiples mediciones sobre la misma unidad se dice que se tienen medidas repetidas sobre algún índice y cuando éste se describe a través del tiempo se tienen datos con una estructura longitudinal, que es el caso que se aborda en este documento. Para llevar a cabo un análisis de este tipo de unidades es posible asumir independencia entre éstas, lo cual puede ser ventajoso, pero se puede reflejar disminuyendo la potencia de las pruebas que se utilicen para detectar diferencias significativas de interés (Davis 2002, p. 103). En este contexto, el supuesto de aleatoriedad entre diferentes UO no necesariamente se cumple ya que las mediciones sobre las mismas UE estarán relacionadas.

Es decir, en el contexto del ejemplo 1, si se realizan mediciones a grupos de diferentes estudiantes en dos tiempos, puede que dichas observaciones estén relacionadas, a pesar de que no sean los mismos alumnos, ya que comparten una serie de características como se ha mencionado previamente. Por este motivo, se deben utilizar metodologías alternas a los modelos lineales de regresión simple. Por tal motivo un acercamiento es considerar una estructura de covarianza entre las mediciones realizadas a la misma unidad, lo cual se puede abordar mediante el modelo descrito en la sección 1.1.

## 1.1. Modelo lineal multivariado

La extensión del modelo de la ecuación (1.1) al caso donde a cada individuo se asocia un vector de observaciones y no una sola observación se realiza mediante la distribución normal multivariada, denotada por  $N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  para vectores  $q$ -dimensionales con vector de medias  $\boldsymbol{\mu}$ , y matriz de varianzas y covarianzas  $\boldsymbol{\Sigma}$ . Los detalles, propiedades y pruebas de hipótesis relacionadas con dicha distribución se pueden consultar en Rencher (2003, cap. 4-5). De manera análoga al caso univariado, el objetivo es determinar si hay diferencias significativas entre los vectores de promedios correspondientes a diferentes grupos conformados por los tratamientos en consideración o en los vectores de promedios asociados a los diferentes tiempos o inclusive determinar si hay una interacción entre los tratamientos y los tiempos, lo cual se realiza mediante un análisis de varianza multivariado (Manova por sus siglás en inglés).

$$\tilde{\mathbf{Y}} = \mathbf{X}\mathbf{B} + \mathbf{E} \quad (1.3)$$

El modelo lineal multivariado considerando  $a$  grupos, cada uno con  $r$  réplicas medidas en  $t$  tiempos, se describe en la ecuación (1.3), donde

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \mathbf{y}_{11} \\ \vdots \\ \mathbf{y}_{1r} \\ \mathbf{y}_{21} \\ \vdots \\ \mathbf{y}_{2r} \\ \vdots \\ \mathbf{y}_{a1} \\ \vdots \\ \mathbf{y}_{ar} \end{bmatrix} = \begin{bmatrix} y_{111} & y_{112} & \cdots & y_{11t} \\ \vdots & & \ddots & \\ y_{1r1} & y_{1r2} & \cdots & y_{1rt} \\ y_{211} & y_{212} & \cdots & y_{21t} \\ \vdots & & \ddots & \\ y_{2r1} & y_{2r2} & \cdots & y_{2rt} \\ \vdots & & \ddots & \\ y_{a11} & y_{a12} & \cdots & y_{a1t} \\ \vdots & & \ddots & \\ y_{ar1} & y_{ar2} & \cdots & y_{art} \end{bmatrix}, \mathbf{E} = \begin{bmatrix} \boldsymbol{\epsilon}_{11} \\ \vdots \\ \boldsymbol{\epsilon}_{1r} \\ \boldsymbol{\epsilon}_{21} \\ \vdots \\ \boldsymbol{\epsilon}_{2r} \\ \vdots \\ \boldsymbol{\epsilon}_{a1} \\ \vdots \\ \boldsymbol{\epsilon}_{ar} \end{bmatrix} = \begin{bmatrix} \epsilon_{111} & \epsilon_{112} & \cdots & \epsilon_{11t} \\ \vdots & & \ddots & \\ \epsilon_{1r1} & \epsilon_{1r2} & \cdots & \epsilon_{1rt} \\ \epsilon_{211} & \epsilon_{212} & \cdots & \epsilon_{21t} \\ \vdots & & \ddots & \\ \epsilon_{2r1} & \epsilon_{2r2} & \cdots & \epsilon_{2rt} \\ \vdots & & \ddots & \\ \epsilon_{a11} & \epsilon_{a12} & \cdots & \epsilon_{a1t} \\ \vdots & & \ddots & \\ \epsilon_{ar1} & \epsilon_{ar2} & \cdots & \epsilon_{art} \end{bmatrix},$$

$$\mathbf{B} = [\boldsymbol{\beta}_0 \quad \boldsymbol{\beta}_1 \quad \cdots \quad \boldsymbol{\beta}_t] = \begin{bmatrix} \beta_{10} & \beta_{11} & \cdots & \beta_{1t} \\ \beta_{20} & \beta_{22} & \cdots & \beta_{2t} \\ \vdots & & \ddots & \\ \beta_{a0} & \beta_{a1} & \cdots & \beta_{at} \end{bmatrix} \text{ y}$$

	Grupo 1	Grupo 2	...	Grupo a
	$N_t(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$	$N_t(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$	...	$N_t(\boldsymbol{\mu}_a, \boldsymbol{\Sigma})$
	$\mathbf{y}_{11}$	$\mathbf{y}_{21}$	...	$\mathbf{y}_{a1}$
	$\mathbf{y}_{12}$	$\mathbf{y}_{22}$	...	$\mathbf{y}_{a2}$
	$\vdots$	$\vdots$		$\vdots$
	$\mathbf{y}_{1r}$	$\mathbf{y}_{2r}$	...	$\mathbf{y}_{ar}$
Total	$\mathbf{y}_{1\cdot}$	$\mathbf{y}_{2\cdot}$	...	$\mathbf{y}_{a\cdot}$
Promedio	$\bar{\mathbf{y}}_1$	$\bar{\mathbf{y}}_2$	...	$\bar{\mathbf{y}}_a$

TABLA 1.1. Conformaciones de los vectores respuesta por grupos para el modelo lineal multivariado.

$$\mathbf{X} = [\mathbf{1}_{ar} : \mathbf{I}_a \otimes \mathbf{1}_r],$$

donde  $\mathbf{1}_{ar} = \mathbf{1}_a \otimes \mathbf{1}_r$ ,  $\mathbf{y}_{ij}$  y  $\boldsymbol{\epsilon}_{ij}$  son los vectores de medidas repetidas y de errores respectivamente cuyas entradas corresponden a las  $t$  observaciones asociadas al  $j$ -ésimo individuo en el  $i$ -ésimo tratamiento. Los  $\boldsymbol{\beta}_k$  son vectores que tienen  $a$  entradas, que corresponden a los coeficientes asociados a los diferentes grupos considerados en el estudio.

Análogo al caso univariado se asume que  $\mathbf{y}_{ij} \sim N_t(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ , lo cual se sintetiza en la tabla 1.1, donde se supone que  $\boldsymbol{\Sigma}$  es igual para todos los grupos,  $\boldsymbol{\mu}_i$  es el vector de medias asociado al  $i$ -ésimo grupo; los vectores de totales y promedios respectivamente se calculan como

$$\mathbf{y}_i = \sum_{j=1}^r \mathbf{y}_{ij} \text{ y } \bar{\mathbf{y}}_i = \sum_{j=1}^r \mathbf{y}_{ij}/r.$$

De acuerdo con Davis (2002, p. 75) el estimador de máxima verosimilitud de  $\mathbf{B}$  viene dado en la ecuación (1.4).

$$\hat{\mathbf{B}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \tilde{\mathbf{Y}} \quad (1.4)$$

Generalmente es de interés realizar inferencia acerca de un subconjunto de coeficientes de la matriz  $\mathbf{B}$ . Las pruebas descritas en la sección 1.1.1 corresponden a pruebas de hipótesis que por lo general son de interés en ambito del modelo (1.3).

### 1.1.1. Pruebas de hipótesis

Tal como lo plantea Davis (2002, p. 76), la siguiente hipótesis general con respecto a los coeficientes de la matriz  $\mathbf{B}$

$$H_0 : \mathbf{ABC} = \mathbf{D} \text{ vs } H_a : \mathbf{ABC} \neq \mathbf{D}$$

permite llevar a cabo las pruebas de hipótesis correspondientes a los diferentes tiempos, grupos e interacciones. Las matrices  $\mathbf{A}$  y  $\mathbf{C}$  dependen de la prueba que se vaya a llevar a cabo. Al igual que en el caso univariado se debe dividir la suma de cuadrados total de manera apropiada para llevar a cabo las distintas pruebas, pero a diferencia de dicha

Estadística de prueba	Formula
Pillai	$V^{(a)} = \sum_{i=1}^a \frac{\lambda_i}{1 + \lambda_i}$
Lawley-Hotelling	$U^{(a)} = \sum_{i=1}^a \lambda_i$
Lambda de Wilk	$\Lambda = \prod_{i=1}^a \frac{1}{1 + \lambda_i}$
La raíz de Roy	$\theta = \frac{\lambda_{(1)}}{1 + \lambda_{(1)}}$

TABLA 1.2. Formulas para las estadísticas de prueba para el análisis de varianza multivariado, las cuales se calculan a partir de los valores propios de la matriz  $\mathbf{Q}_e^{-1}\mathbf{Q}_h$  se denotan por  $\lambda_i$ , mientras que el valor propio más alto se denota como  $\lambda_1$ .

situación en este contexto se tiene una matriz de suma de cuadrados total, la cual se divide en las matrices “entre” ( $\mathbf{Q}_h$ ) y “dentro” ( $\mathbf{Q}_e$ ), las cuales se calculan como se muestra en las ecuaciones (1.5) y (1.6) respectivamente.

$$\mathbf{Q}_h = \left( \mathbf{A}\widehat{\mathbf{B}}\mathbf{C} - \mathbf{D} \right)^t \left[ \mathbf{A}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{A}^t \right]^{-1} \left( \mathbf{A}\widehat{\mathbf{B}}\mathbf{C} - \mathbf{D} \right) \quad (1.5)$$

$$\mathbf{Q}_e = \mathbf{C}^t \left[ \mathbf{Y}^t\mathbf{Y} - \widehat{\mathbf{B}}^t(\mathbf{X}^t\mathbf{X})\widehat{\mathbf{B}} \right] \mathbf{C} \quad (1.6)$$

Una vez calculadas las matrices  $\mathbf{Q}_h$  y  $\mathbf{Q}_e$ , como señala Rencher (2003, p. 162) se calcula  $\mathbf{Q}_e^{-1}\mathbf{Q}_h$  junto con sus respectivos valores propios, ya que éstos serán utilizados para el cálculo de las estadísticas de prueba. La tabla 1.2 muestra un resumen de las cuatro estadísticas de prueba utilizadas, las cuales se pueden aproximar a una distribución F (Monroy & Rivera 2012). Todos los detalles y una amplia discusión con respecto a cuál estadística es mejor en qué situación se pueden consultar en Rencher (2003, p. 176-178) o en Davis (2002, p. 77).

Acorde a lo que comenta Davis (2002, p. 78-81) las tres hipótesis de interés por lo general son:

- $H_0^{(1)}$  : No existe interacción entre los grupos y los tiempos de medición (los perfiles son paralelos),
- $H_0^{(2)}$  : No hay diferencias entre los grupos,
- $H_{03}^{(3)}$  : No hay diferencias entre los tiempos de medición.

Es importante tener en cuenta que  $H_0^{(1)}$  debe juzgarse primero, ya que dependiendo de su resultado las pruebas asociadas a  $H_0^{(2)}$  y  $H_0^{(3)}$  pueden tener o carecer de sentido. Dependiendo de cuál prueba se quiera llevar a cabo se definen las matrices  $\mathbf{A}$  y  $\mathbf{C}$  para calcular las matrices dadas en las ecuaciones (1.5) y (1.6) y así proceder a calcular las diferentes estadísticas de prueba con todo lo que esto conlleva.

### 1.1.1.1. Prueba de paralelismo

La hipótesis de no interacción entre grupos,  $H_{01}$ , es equivalente a juzgar

$$H_{01} : \begin{bmatrix} \mu_{11} - \mu_{12} \\ \mu_{12} - \mu_{13} \\ \vdots \\ \mu_{1(t-1)} - \mu_{13} \end{bmatrix} = \begin{bmatrix} \mu_{21} - \mu_{22} \\ \mu_{22} - \mu_{23} \\ \vdots \\ \mu_{2(t-1)} - \mu_{2t} \end{bmatrix} = \dots = \begin{bmatrix} \mu_{s1} - \mu_{s2} \\ \mu_{s2} - \mu_{s3} \\ \vdots \\ \mu_{s(t-1)} - \mu_{2t} \end{bmatrix},$$

lo cual, para términos de la hipótesis general  $H_0 : \mathbf{ABC} = \mathbf{D}$  se utilizan las matrices de la forma

$$\mathbf{A}_{(a-1) \times a} = (\mathbf{I}_{a-1} : -\mathbf{1}_{a-1}), \quad \mathbf{C}_{t \times (t-1)} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ -1 & 1 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & -1 \end{bmatrix} \text{ y } \mathbf{D} = \mathbf{0}_{(a-1) \times (t-1)}$$

### 1.1.1.2. Prueba entre grupos

La hipótesis de no diferencia entre grupos,  $H_{02}$ , es equivalente a probar

$$H_{02} : \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1t} \end{bmatrix} = \begin{bmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2t} \end{bmatrix} = \dots = \begin{bmatrix} \mu_{a1} \\ \mu_{a2} \\ \vdots \\ \mu_{at} \end{bmatrix},$$

lo cual para escribirse en términos de la hipótesis general  $H_0 : \mathbf{ABC} = \mathbf{D}$ , se toma

$$\mathbf{A}_{(a-1) \times a} = [\mathbf{I}_{a-1} : -\mathbf{1}_{a-1}], \quad \mathbf{C}_{t \times 1} = \mathbf{1}_t \text{ y } \mathbf{D}_{(a-1) \times 1} = \mathbf{0}_{a-1}$$

### 1.1.1.3. Prueba entre tiempos

De manera análoga a las pruebas anteriores, para la hipótesis general  $H_0 : \mathbf{ABC} = \mathbf{D}$  correspondiente a la verificación de diferencias entre tiempos se toma

$$\mathbf{A}_{1 \times s} = [1, \dots, 1], \quad \mathbf{C}_{t \times (t-1)} = \begin{bmatrix} \mathbf{I}_{t-1} \\ -\mathbf{1}_{t-1}^t \end{bmatrix}, \text{ y } \mathbf{D}_{1 \times (t-1)} = \mathbf{0}_{t-1}^t.$$

Una metodología alterna al Manova que se trabaja en el contexto de datos longitudinales son los modelos lineales de efectos mixtos en donde se adicionan efectos aleatorios que permiten discriminar la variabilidad de la respuesta de una manera diferente.

## 1.2. Modelo lineal mixto

Los modelos lineales son una forma de explicar la dispersión de una o más respuestas aleatorias en términos de una serie de variables independientes (también conocidas como exógenas). En West et al. (2014, p. 5) se encuentra un recuento histórico referente a los modelos lineales desde sus inicios en 1861 hasta la actualidad, desde una perspectiva teórica, mencionando al igual avances importantes en cuanto a la implementación de las diferentes metodologías a nivel de paquetes estadísticos. Éstos tienen en cuenta una serie de supuestos lo que hace posible su planteamiento, estimación, interpretación y su respectiva evaluación. Las variables independientes pueden ser clasificadas como efectos fijos o efectos aleatorios. Acorde con Melo et al. (2007, p. 6), cuando al finalizar el experimento las conclusiones se formulan sobre un número preestablecido de tratamientos el modelo se denomina de efectos fijos y en este caso la inferencia estadística se hace sobre los efectos medios de los tratamientos, por lo cual aquellas situaciones en que se desean realizar comparaciones o contrastes entre niveles de un factor en búsqueda de diferencias, éste se considera como fijo. Si los niveles de un atributo son una muestra aleatoria de una población de posibles selecciones, es decir, las conclusiones se formulan sobre un número mayor de tratamientos a los usados en el experimento, el modelo se dice de efectos aleatorios, y en este caso, la inferencia estadística se hace sobre las varianzas de los mismos. Los modelos que incluyen ambos se denominan de efectos mixtos (MLM).

En el contexto longitudinal se tienen medidas repetidas sobre diferentes UE las cuales corresponden a distintos tratamientos. En estos casos puede que cada UE tenga un efecto particular sobre la variable respuesta que no se deba directamente al tratamiento en cuestión; por lo cual es necesario evaluar la inclusión de efectos aleatorios, que corresponden a las UE incluidas, ya que usualmente los análisis de muestras longitudinales abarcan únicamente una muestra de unidades provenientes de una población más grande y se desea obtener conclusiones con respecto al conjunto completo. A manera de ejemplo, un modelo de regresión lineal simple es de la forma

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}, \quad (1.7)$$

donde  $y_{ij}$ ,  $x_{ij}$  y  $\epsilon_{ij}$  representan la respuesta, la variable independiente y el residual asociados a la  $j$ -ésima medición de la  $i$ -ésima UE, respectivamente. Allí se observa que a todos los individuos se les asigna la misma pendiente  $\beta_1$  y el mismo intercepto  $\beta_0$  de la recta, pero al considerar que todas las UE tienen características propias que influyen sobre la relación entre las variables  $x_{ij}$  y  $y_{ij}$ , puede que una mejor manera de explicar la variabilidad de la respuesta en términos de  $x_{ij}$  sea asignando un intercepto o una pendiente propia a cada UE, es decir mediante una ecuación de la forma

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + b_i + \epsilon_{ij} \\ &= (\beta_0 + b_i) + \beta_1 x_{ij} + \epsilon_{ij}, \end{aligned} \quad (1.8)$$

donde  $b_i$  es el efecto propio de la  $i$ -ésima UE sobre la respuesta. Las figuras 1.1 y 1.2 ilustran el concepto anterior mediante datos simulados. En la primera se observa una recta en donde todas las UE comparten el mismo intercepto, mientras que la segunda muestra un intercepto diferente para cada UE, logrando un mejor ajuste de la respuesta en términos de la variable exógena. Es importante notar que en el caso de la figura 1.1 el error experimental sería más grande que el correspondiente a la figura 1.2, dado que en

esta última se observan errores más pequeños, lo cual se traduce en estadísticas de prueba más pequeñas y por ende afectando los valores  $p$  asociados a las pruebas en consideración, de manera que se pueden afectar las conclusiones directamente. El ejemplo anterior puede replicarse de la misma manera utilizando diferentes pendientes de la recta para cada UE, de la forma

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + b_i x_{ij} + \epsilon_{ij} \\ &= \beta_0 + (\beta_1 + b_i) x_{ij} + \epsilon_{ij}. \end{aligned}$$

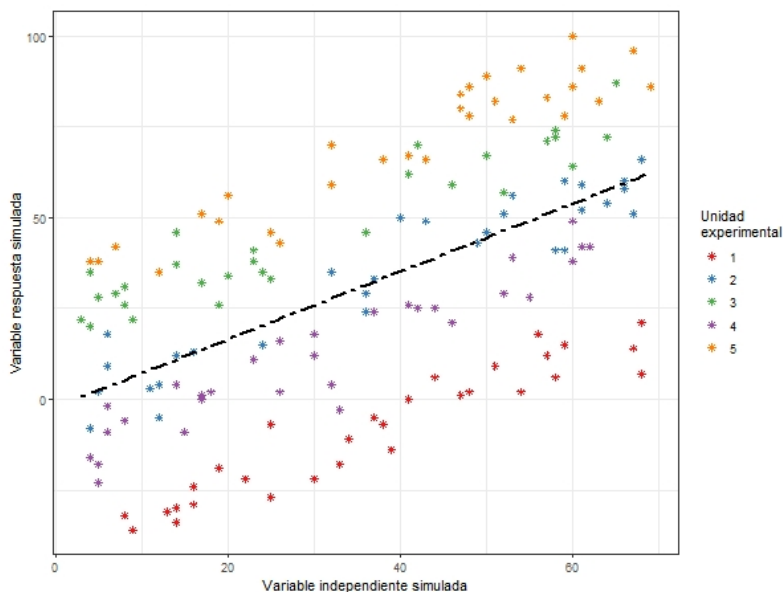


FIGURA 1.1. Diagrama de dispersión de las variables  $y_{ij}$  y  $x_{ij}$  simuladas con una recta estimada para todo el conjunto de datos.

En términos matriciales, según Pinheiro & Bates (2006, p. 58), el modelo lineal general con efectos mixtos viene dado de la forma

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad \text{con } i = 1, \dots, n \text{ y } n = r \times a \quad (1.9)$$

donde  $\mathbf{y}_i$  es el vector de respuestas,  $\mathbf{X}_i$  y  $\mathbf{Z}_i$  son las matrices de diseño de los efectos fijos y aleatorios asociadas al  $i$ -ésimo individuo, respectivamente. Los vectores  $\boldsymbol{\beta}$  y  $\mathbf{b}_i$  de dimensiones  $a$  y  $q$ , que contienen los coeficientes asociados a los efectos fijos y aleatorios considerados en la  $i$ -ésima UE, respectivamente, mientras que  $\boldsymbol{\epsilon}_i$  es el vector de errores de mediciones dentro de cada UE. Como supuestos distribucionales sobre este modelo se asume que  $\mathbf{b}_i \sim \mathbf{N}_q(\mathbf{0}_q, \boldsymbol{\Psi}_{q \times q})$  y  $\boldsymbol{\epsilon}_i \sim \mathbf{N}_t(\mathbf{0}_t, \boldsymbol{\Sigma}_{t \times t})$ . A su vez los vectores  $\boldsymbol{\epsilon}_i$  y  $\mathbf{b}_i$  se asumen independientes, es decir  $Cov(\boldsymbol{\epsilon}_i, \mathbf{b}_i) = 0$ . La matriz  $\boldsymbol{\Sigma}$  asociada a los errores dentro de las mediciones de la misma unidad puede considerar diferentes situaciones, es decir que se puede asociar a procesos autocorrelacionados, por ejemplo los que se encuentran en la literatura relacionada con series de tiempo, como los modelos *ARIMA* (Wei 2006), o como los que se estudian en la estadística espacial, que se caracterizan mediante su función de variograma (Schabenberger & Gotway 2017). En este documento únicamente aborda el

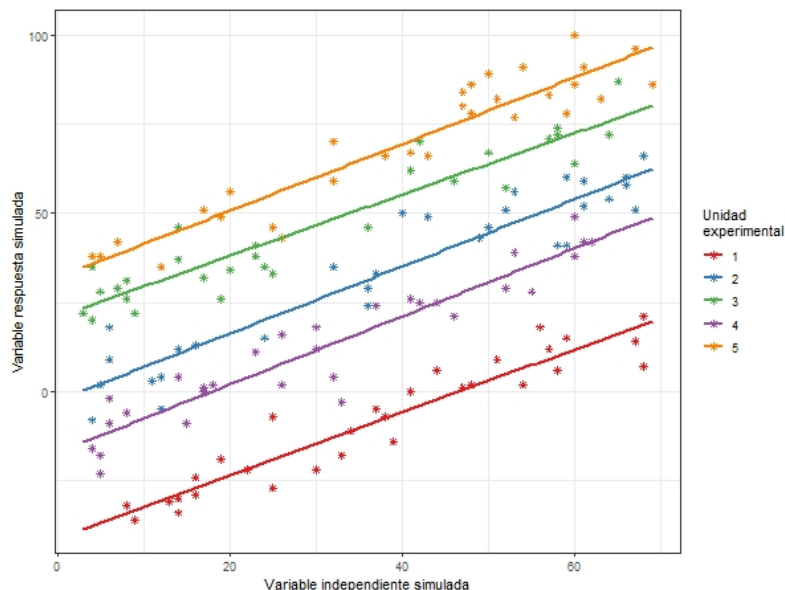


FIGURA 1.2. Diagrama de dispersión de las variables  $y_{ij}$  y  $x_{ij}$  simuladas con rectas estimadas para cada una de las UE.

contexto de datos longitudinales o datos panel, por lo cual únicamente se tendrán en cuenta aquellos procesos relacionados con series temporales en cuanto a la matriz de errores  $\Sigma$ .

### 1.2.1. Estimación de los MLM mediante máxima verosimilitud

La literatura estadística abarca numerosas técnicas de estimación, entre las cuales se encuentra la estimación por máxima verosimilitud (MV), la cual se encarga de buscar puntos máximos a la función de distribución conjunta de las observaciones o vectores asociados a la variable respuesta con respecto a un conjunto de parámetros en consideración. De acuerdo con Rizzo (2007, p. 335), siendo  $X_1, \dots, X_n$  una muestra aleatoria de variables con un parámetro, o vector de parámetros,  $\theta \in \Theta$ , donde  $\Theta$  es la representación de todo el espacio paramétrico. La *función de verosimilitud*  $L(\theta)$  de las variables aleatorias  $X_1, \dots, X_n$  evaluadas en  $x_1, \dots, x_n$  es definida como la función de probabilidad conjunta. Bajo el supuesto de independencia y considerando que las  $X_i$  son variables idénticamente distribuidas se tiene

$$\begin{aligned} L(\theta) &= f(x_1, \dots, x_n; \theta) \\ &= \prod_{i=1}^n f(x_i; \theta). \end{aligned} \quad (1.10)$$

Un estimador máximo verosímil es aquel que maximiza la función  $L(\theta)$  o lo que es igual, maximiza  $l(\theta)$ , donde

$$\begin{aligned}
l(\theta) &= \ln(L(\theta)) \\
&= \ln \left[ \prod_{i=1}^n f(x_i; \theta) \right] \\
&= \sum_{i=1}^n \ln [f(x_i; \theta)], \tag{1.11}
\end{aligned}$$

ya que el logaritmo natural es una función monótona y creciente en todo su dominio.

Retomando el MLM dado en la ecuación (1.9), asumiendo  $\Sigma = \sigma^2 \mathbf{I}$  y con el objetivo de simplificar algunos cálculos, Pinheiro & Bates (2006, p. 59) define el *factor de precisión relativa*,  $\Delta$ , como aquel que satisface

$$\sigma^2 \Psi^{-1} = \Delta^t \Delta.$$

Esto quiere decir que el factor de Cholesky, el cual puede ser consultado en Gentle (2012, p. 73), es una de las maneras de encontrar  $\Delta$ ; teniendo en cuenta que la matriz  $\Psi$  es definida positiva por el hecho de ser una matriz de varianzas y covarianzas, lo cual asegura la existencia de  $\Delta$ . Para el siguiente desarrollo es importante tener en cuenta que la matriz  $\Delta$  es triangular, por lo cual su determinante es igual al determinante de su transpuesta. A partir de los supuestos realizados sobre el modelo dado en (1.9) se sabe que

$$\begin{aligned}
\mathbf{y}_i | \mathbf{b}_i &\sim \mathcal{N}_t(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i, \sigma^2 \mathbf{I}) \\
\mathbf{b}_i &\sim \mathcal{N}_q(\mathbf{0}_q, \Psi).
\end{aligned}$$

Las funciones de densidad de  $\mathbf{y}_i | \mathbf{b}_i$  y  $\mathbf{b}_i$  se denotan como  $g(\mathbf{y}_i | \mathbf{b}_i)$  y  $h(\mathbf{b}_i)$ , respectivamente, con

$$\begin{aligned}
h(\mathbf{b}_i) &= \frac{\exp \{ -\mathbf{b}_i^t \Psi^{-1} \mathbf{b}_i / 2 \}}{(2\pi)^{q/2} \sqrt{|\Psi|}} \\
&= \frac{\exp \{ -\mathbf{b}_i^t \Delta^t \Delta \mathbf{b}_i / 2\sigma^2 \}}{(2\pi)^{q/2} \sqrt{|\sigma^{-2} \Delta^t \Delta|^{-1}}} \\
&= \frac{\exp \{ -\|\Delta \mathbf{b}_i\|^2 / 2\sigma^2 \}}{(2\pi)^{q/2} \sigma^{2q/2} |\Delta^t \Delta|^{-1/2}} \\
&= \frac{\exp \{ -\|\Delta \mathbf{b}_i\|^2 / 2\sigma^2 \}}{(2\pi\sigma^2)^{q/2} |\Delta|^{-1}} \tag{1.12}
\end{aligned}$$

$$g(\mathbf{y}_i | \mathbf{b}_i) = \frac{\exp \{ -\|\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{Z}_i \mathbf{b}_i\|^2 / (2\sigma^2) \}}{(2\pi\sigma^2)^{t/2}} \tag{1.13}$$

donde  $\|\cdot\|$  es la norma de un vector. La función de densidad conjunta de los vectores  $\mathbf{y}_i$  y  $\mathbf{b}_i$ ,  $f(\mathbf{y}_i, \mathbf{b}_i)$  se calcula de la siguiente manera

$$\begin{aligned} f(\mathbf{y}_i, \mathbf{b}_i) &= g(\mathbf{y}_i|\mathbf{b}_i)h(\mathbf{b}_i) \\ &= \frac{\exp\left\{-\left(\|\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i\|^2 + \|\boldsymbol{\Delta}\mathbf{b}_i\|^2\right) / (2\sigma^2)\right\}}{(2\pi\sigma^2)^{t/2} (2\pi\sigma^2)^{q/2} |\boldsymbol{\Delta}|^{-1}} \\ &= \frac{\exp\left\{-\left(\|\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\boldsymbol{\beta} - \tilde{\mathbf{Z}}_i\mathbf{b}_i\|^2\right) / (2\sigma^2)\right\}}{(2\pi\sigma^2)^{t/2} (2\pi\sigma^2)^{q/2} |\boldsymbol{\Delta}|^{-1}} \end{aligned} \quad (1.14)$$

donde

$$\tilde{\mathbf{y}}_i = \begin{bmatrix} \mathbf{y}_i \\ \mathbf{0}_{q \times q} \end{bmatrix}, \quad \tilde{\mathbf{X}}_i = \begin{bmatrix} \mathbf{X}_i \\ \mathbf{0}_{q \times q} \end{bmatrix} \quad \text{y} \quad \tilde{\mathbf{Z}}_i = \begin{bmatrix} \mathbf{Z}_i \\ \boldsymbol{\Delta} \end{bmatrix}.$$

Se observa que el exponente de la ecuación (1.14) es similar al vector de errores en el caso de mínimos cuadrados ordinarios, en donde se desea minimizar la *SCE* con respecto a un vector de coeficientes. De esta manera para maximizar  $f(\mathbf{y}_i, \mathbf{b}_i)$  en términos del vector  $\mathbf{b}_i$  se requiere minimizar el exponente de la ecuación (1.14). De forma análoga a la obtención de estimadores obtenidos vía máxima verosimilitud en el caso de la regresión lineal múltiple (Ver Montgomery et al. (2012, p. 78)) se obtiene lo siguiente:

$$\hat{\mathbf{b}}_i = \left(\tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i\right)^{-1} \tilde{\mathbf{Z}}_i^t \left(\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\boldsymbol{\beta}\right). \quad (1.15)$$

De esta manera, el exponente de la ecuación (1.14) mediante (1.15) se puede escribir como

$$\begin{aligned} \left\|\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\boldsymbol{\beta} - \tilde{\mathbf{Z}}_i\mathbf{b}_i\right\|^2 &= \left\|\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\boldsymbol{\beta} - \tilde{\mathbf{Z}}_i\mathbf{b}_i + \tilde{\mathbf{Z}}_i\hat{\mathbf{b}}_i - \tilde{\mathbf{Z}}_i\hat{\mathbf{b}}_i\right\|^2 \\ &= \left\|\left(\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\boldsymbol{\beta} - \tilde{\mathbf{Z}}_i\hat{\mathbf{b}}_i\right) + \tilde{\mathbf{Z}}_i\left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right)\right\|^2 \\ &= \left(\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\boldsymbol{\beta} - \tilde{\mathbf{Z}}_i\hat{\mathbf{b}}_i\right)^t \left(\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\boldsymbol{\beta} - \tilde{\mathbf{Z}}_i\hat{\mathbf{b}}_i\right) + \left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right)^t \tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i \left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right) \\ &+ \left(\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\boldsymbol{\beta} - \tilde{\mathbf{Z}}_i\hat{\mathbf{b}}_i\right)^t \tilde{\mathbf{Z}}_i \left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right) + \left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right)^t \tilde{\mathbf{Z}}_i^t \left(\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\boldsymbol{\beta} - \tilde{\mathbf{Z}}_i\hat{\mathbf{b}}_i\right), \end{aligned} \quad (1.16)$$

donde se observa que los últimos términos

$$\left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right)^t \tilde{\mathbf{Z}}_i^t \left(\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\boldsymbol{\beta} - \tilde{\mathbf{Z}}_i\hat{\mathbf{b}}_i\right) \quad \text{y} \quad \left(\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\boldsymbol{\beta} - \tilde{\mathbf{Z}}_i\hat{\mathbf{b}}_i\right)^t \tilde{\mathbf{Z}}_i \left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right)$$

son escalares y a su vez son iguales entre ellos. Reemplazando la ecuación (1.15) en uno de ellos se tiene lo siguiente:

$$\begin{aligned}
(\mathbf{b}_i - \hat{\mathbf{b}}_i)^t \tilde{\mathbf{Z}}_i^t (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta} - \tilde{\mathbf{Z}}_i \hat{\mathbf{b}}_i) &= (\mathbf{b}_i - \hat{\mathbf{b}}_i)^t \tilde{\mathbf{Z}}_i^t (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta}) - (\mathbf{b}_i - \hat{\mathbf{b}}_i)^t \tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i \hat{\mathbf{b}}_i \\
&= \left( \mathbf{b}_i - \left( \tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i \right)^{-1} \tilde{\mathbf{Z}}_i^t (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta}) \right)^t \tilde{\mathbf{Z}}_i^t (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta}) \\
&\quad - \left( \mathbf{b}_i - \left( \tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i \right)^{-1} \tilde{\mathbf{Z}}_i^t (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta}) \right)^t \tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i \left( \tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i \right)^{-1} \tilde{\mathbf{Z}}_i^t (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta}) \\
&= \mathbf{b}_i^t \tilde{\mathbf{Z}}_i^t (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta}) - (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta})^t \tilde{\mathbf{Z}}_i \left( \tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i \right)^{-1} \tilde{\mathbf{Z}}_i^t (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta}) \\
&\quad - \mathbf{b}_i^t \tilde{\mathbf{Z}}_i^t (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta}) + (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta})^t \tilde{\mathbf{Z}}_i \left( \tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i \right)^{-1} \tilde{\mathbf{Z}}_i^t (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta}) \\
&= 0.
\end{aligned}$$

Por lo tanto el exponente de la ecuación (1.14) se puede escribir como

$$\begin{aligned}
\left\| \tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta} - \tilde{\mathbf{Z}}_i \mathbf{b}_i \right\|^2 &= (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta} - \tilde{\mathbf{Z}}_i \hat{\mathbf{b}}_i)^t (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta} - \tilde{\mathbf{Z}}_i \hat{\mathbf{b}}_i) + (\mathbf{b}_i - \hat{\mathbf{b}}_i)^t \tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i (\mathbf{b}_i - \hat{\mathbf{b}}_i) \\
&= \left\| \tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta} - \tilde{\mathbf{Z}}_i \hat{\mathbf{b}}_i \right\|^2 + (\mathbf{b}_i - \hat{\mathbf{b}}_i)^t \tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i (\mathbf{b}_i - \hat{\mathbf{b}}_i).
\end{aligned}$$

Como la ecuación (1.14) es la función conjunta de probabilidad entre los vectores  $\mathbf{y}_i$  y  $\mathbf{b}_i$ , y la función de verosimilitud se basa en la función de probabilidad de  $\mathbf{y}_i$ , es decir  $f(\mathbf{y}_i)$ , se debe integrar la ecuación (1.14) con respecto a  $\mathbf{b}_i$ . De esta manera se tiene que

$$\begin{aligned}
f(\mathbf{y}_i) &= \int f(\mathbf{y}_i, \mathbf{b}_i) d\mathbf{b}_i \\
&= \int \frac{\exp \left\{ \left[ \left\| \tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta} - \tilde{\mathbf{Z}}_i \hat{\mathbf{b}}_i \right\|^2 + (\mathbf{b}_i - \hat{\mathbf{b}}_i)^t \tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i (\mathbf{b}_i - \hat{\mathbf{b}}_i) \right] / 2\sigma^2 \right\}}{(2\pi\sigma^2)^{t/2} (2\pi\sigma^2)^{q/2} |\boldsymbol{\Delta}|^{-1}} d\mathbf{b}_i \\
&= \frac{\exp \left\{ \left\| \tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta} - \tilde{\mathbf{Z}}_i \hat{\mathbf{b}}_i \right\|^2 / 2\sigma^2 \right\}}{(2\pi\sigma^2)^{t/2} |\boldsymbol{\Delta}|^{-1}} \int \frac{\exp \left\{ (\mathbf{b}_i - \hat{\mathbf{b}}_i)^t \tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i (\mathbf{b}_i - \hat{\mathbf{b}}_i) / 2\sigma^2 \right\}}{(2\pi\sigma^2)^{q/2}} d\mathbf{b}_i \\
&= \frac{\exp \left\{ \left\| \tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta} - \tilde{\mathbf{Z}}_i \hat{\mathbf{b}}_i \right\|^2 / 2\sigma^2 \right\}}{(2\pi\sigma^2)^{t/2} |\boldsymbol{\Delta}|^{-1}} \int \frac{\exp \left\{ (\mathbf{b}_i - \hat{\mathbf{b}}_i)^t \tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i (\mathbf{b}_i - \hat{\mathbf{b}}_i) / 2\sigma^2 \right\}}{(2\pi\sigma^2)^{q/2} \left( |\tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i| \right)^{-1/2}} d\mathbf{b}_i \\
&= \frac{\exp \left\{ \left\| \tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta} - \tilde{\mathbf{Z}}_i \hat{\mathbf{b}}_i \right\|^2 / 2\sigma^2 \right\}}{(2\pi\sigma^2)^{t/2} |\boldsymbol{\Delta}|^{-1} \sqrt{|\tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i|}}. \tag{1.17}
\end{aligned}$$

Como plantea Pinheiro & Bates (2006, p. 64) la función de verosimilitud para los MLM descritos en la ecuación (1.9) es de la siguiente forma

$$\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) &= \prod_{i=1}^n \frac{\exp \left\{ \|\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta} - \tilde{\mathbf{Z}}_i \hat{\boldsymbol{\beta}}_i\| / 2\sigma^2 \right\}}{(2\pi\sigma^2)^{t/2} |\boldsymbol{\Delta}|^{-1} \sqrt{|\tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i|}} \\
&= \frac{\exp \left\{ \sum_{i=1}^n \|\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta} - \tilde{\mathbf{Z}}_i \hat{\boldsymbol{\beta}}_i\| / 2\sigma^2 \right\}}{(2\pi\sigma^2)^{nt/2}} \prod_{i=1}^n \frac{|\boldsymbol{\Delta}|}{\sqrt{|\tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i|}}
\end{aligned} \tag{1.18}$$

donde  $\boldsymbol{\theta}$  es el vector que contiene los parámetros utilizados en  $\boldsymbol{\Psi}$  (en el caso en que  $\boldsymbol{\Sigma}$  no sea una matriz diagonal, el vector  $\boldsymbol{\theta}$  también incluye los parámetros ubicados fuera de la diagonal). La función (1.11) para los MLM, a partir de la ecuación (1.2.1) se calcula de la forma

$$\begin{aligned}
l(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) &= \frac{1}{2\sigma^2} \sum_{i=1}^n \|\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \boldsymbol{\beta} - \tilde{\mathbf{Z}}_i \hat{\boldsymbol{\beta}}_i\| \\
&\quad - \frac{n}{2} \ln(2\pi\sigma^2) + \sum_{i=1}^n \left( \ln|\boldsymbol{\Delta}| - \frac{1}{2} \ln|\tilde{\mathbf{Z}}_i^t \tilde{\mathbf{Z}}_i| \right).
\end{aligned} \tag{1.19}$$

Una manera en que se pueden optimizar computacionalmente los procedimientos es mediante la “*factorización QR*”, ya que descompone una matriz como un producto de dos matrices, de las cuales una es ortogonal, mientras la otra es triangular superior, lo cual simplifica algunos cálculos. El proceso de describir la función  $l(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$  a través de esta descomposición se muestra en Pinheiro & Bates (2006, p. 68). Una descripción detallada de la factorización matricial utilizada se puede consultar en Gentle (2012, p. 95).

La función de “*máxima verosimilitud restringida*” es un método alternativo de estimación mediante el cual se obtienen estimadores insesgados para los componentes de la varianza. El proceso de la deducción de la función a optimizar para esta metodología se presenta en Pinheiro & Bates (2006, p. 75) y al igual que en el caso de la máxima verosimilitud se obtiene una función a optimizar mediante métodos iterativos con respecto a un grupo de parámetros. Por otro lado Gałecki & Burzykowski (2013, p. 257) muestran el mismo proceso de deducción para la metodología de mínimos cuadrados penalizados (PLS por sus siglas en inglés), la cual se desarrolla de manera muy similar a lo presentado en la sección 1.9.

El proceso de estimación de parámetros de los MLM requiere una serie de supuestos que inclusive hacen posible su posterior evaluación. Una vez se determine el modelo adecuado para un conjunto de datos requiere una evaluación de las premisas realizadas para dar soporte a todo el proceso realizado previamente. Toda la validación pertinente a los supuestos realizados sobre estos modelos se puede consultar en Gałecki & Burzykowski (2013, p. 268).

La ecuación (1.19) muestra la función a optimizar con respecto a los parámetros utilizados en el modelo estadístico. Para este fin se utilizan métodos iterativos como lo son el algoritmo EM, Newton Raphson o Fisher Scoring, entre otros.

### 1.2.2. Prueba del cociente de verosimilitudes

Para llevar a cabo la prueba de significancia de un factor en particular, ya sea uno de efectos fijos o aleatorios se puede hacer uso de la prueba del cociente de verosimilitudes. Acorde a Faraway (2016, p. 174) mediante la teoría estándar, se puede derivar un test para comparar dos hipótesis anidadas para comparar dos hipótesis anidadas, denotadas por  $H_0$  y  $H_a$ , calculando la estadística de prueba:

$$\xi^2 = 2\ln\left(\frac{L_{H_a}}{L_{H_0}}\right) = 2(\ln(L_{H_a}) - \ln(L_{H_0})) \quad (1.20)$$

donde  $L_{H_a}$  es el máximo valor de función de verosimilitud bajo el espacio paramétrico correspondiente a  $H_a$  (generalmente  $L_{H_a}$  es calculada a partir de los estimadores de máxima verosimilitud) y de manera análoga  $L_{H_0}$  es el valor de la función de verosimilitud calculada a partir de los parámetros incluidos en  $H_0$ . La distribución bajo  $H_0$  de la estadística de prueba dada en (1.20) es aproximadamente Chi-cuadrado con grados de libertad igual a la diferencia en las dimensiones de los dos espacios paramétricos (diferencia en el número de parámetros estimados bajo los dos escenarios).

### 1.3. Modelos para datos pseudo-panel

Muchos tipos de modelos pueden ser estimados a partir del supuesto de independencia de los modelos de regresión estándar, mientras que otros consideran datos de tipo panel, generando la necesidad de inclusión de efectos que permitan incluir los efectos individuales como variables independientes adicionales. De acuerdo con Verbeek (2008) la mayor limitación de datos con secciones transversales repetidas es que la totalidad de los individuos no son seguidos en el tiempo, por lo cual las historias individuales no están disponibles para su inclusión en un modelo estadístico, la construcción de instrumentos o para transformar un modelo a uno que considere desviaciones sobre efectos grupales. Según Deaton (1985) si hay una relación entre dos variables (una explicativa y una respuesta) entre los individuos de una población, existirá una versión grupal de esta relación del mismo tipo al caso individual. Esto quiere decir que por ejemplo considerando el modelo dado en la ecuación (1.7), calculando los promedios por UE, se mantendrá la misma relación lineal, es decir mediante el modelo descrito en la ecuación (1.21).

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + \bar{\epsilon}_i. \quad (1.21)$$

El modelo usado por Deaton (1985) es usado en la literatura, construyendo grupos de individuos de manera adecuada los cuales denominan “*cohortes*”, donde cada individuo solo puede pertenecer a una de éstas. Un ejemplo de estas agrupaciones es a partir del año de nacimiento de los individuos, de esta manera las cohortes pueden ser los nacidos en los 70’s, en los 80’s, en los 90’s y así sucesivamente. Propper et al. (2001) realiza un análisis que examina los factores principales que determinan la demanda de los seguros de salud privados en el Reino Unido desde 1978 hasta 1996 mediante un modelo lineal que incluye una muestra que denominan de pseudo-panel. Gardes et al. (2005) investiga los sesgos en las elasticidades de ingresos y gastos estimados a partir de metodologías utilizadas sobre diferentes tipos de muestra dentro de las cuales consideran datos de pseudo-panel, datos transversales y paneles verdaderos. Verbeek (2008) hace una revisión exhaustiva de

modelos lineales, su identificación y estimación desde una aproximación a datos de tipo panel a partir de mediciones repetidas en secciones transversales, adicionando rezagos de la variable respuesta a las variables explicativas como es usual en metodologías de series temporales. Sprietsma (2012) realiza una revisión de la influencia del acceso a centros de computo y al internet en los efectos como herramientas pedagógicas para la adquisición de habilidades en matemáticas a partir de muestras que consisten en secciones transversales de estudiantes de colegio en los años 1999, 2001 y 2003 en Brasil. La estimación del modelo se lleva a cabo mediante el estimador propuesto por Deaton (1985) junto con una corrección propuesta por Verbeek & Nijman (1993) para el caso en que se tiene un número pequeño de tiempos de medición. De manera análoga a los contextos anteriores Tovar et al. (2012) considera el caso en que las muestras seleccionadas en diferentes tiempos comparten algunos individuos, es decir que las diferentes secciones transversales tienen personas en común, por lo cual el modelo ajustado es el considerado por Deaton (1985) incluyendo una autocorrelación temporal en los residuales al tener mediciones provenientes sobre algunos individuos, lo que se realiza para modelar la fuerza laboral en el país Vasco desde 1993 hasta 1999. En la literatura se pueden encontrar numerosos casos adicionales a los mencionados previamente donde a partir de efectos fijos asociados a cohortes modelan el comportamiento de una variable respuesta medida en diferentes tiempos (ver Himaz & Aturupane (2016), Canavire-Bacarreza & Robles (2017), Urdinola & Ospino (2015), entre otros).

## 1.4. Paquetes estadísticos

Todos los análisis pertinentes a este documento se realizaron en el software R (R Core Team 2020), por ser de libre acceso. Éste consiste en una serie de paquetes que se elaboran con objetivos específicos de acuerdo a las necesidades de sus usuarios. Para las figuras construidas a partir de los análisis realizados se utilizó el paquete “*ggplot2*” (Wickham 2016). Para llevar a cabo las pruebas descritas en la sección 1.1 se utilizó la función *manova*(·) del paquete “*stats*”, además de las librerías “*dplyr*” y “*tidyr*” (Wickham & Henry 2018, Wickham et al. 2019), las cuales se utilizan para la manipulación de bases de datos.

Para la metodología descrita en la sección 1.2, R cuenta con dos librerías principales “*lme4*” y “*nlme*” (ver Bates et al. (2015) y Pinheiro et al. (2018) respectivamente) que permiten llevar a cabo los ajustes, estimación y evaluación de los modelos descritos. Finch et al. (2016, p. 90) hace una explicación del uso de las funciones principales de éstos dos paquetes, las cuales son *lme*(·) para “*nlme*” y *lmer*(·) para “*lme4*”, es decir una descripción de las expresiones requeridas para las formulas utilizadas en la inclusión de factores fijos y aleatorios mediante bases de datos preestablecidas, además de los diferentes argumentos y funciones que complementan su uso. La prueba de hipótesis asociada a la sección 1.2.2 se lleva a cabo mediante la función *anova*(·), que requiere los modelos bajo  $H_0$  y  $H_a$  como sus argumentos. La evaluación de los supuestos requeridos por los MLM en el software se describe en Pinheiro & Bates (2006, p. 174).

Los procedimientos realizados en el capítulo 2 requieren la generación de muestras aleatorias de la distribución normal multivariada, por lo cual se utilizó la librería “*mvtnorm*” (Genz et al. 2016).

---

Para todos lo referente a este trabajo se usó la librería “*nlme*”, ya que como lo nota Finch et al. (2016, p. 96) el paquete “*lme4*” no cuenta por ahora con la opción de incluir correlaciones en las mediciones en repetidas ocasiones del mismo individuo, para su posterior evaluación. Pinheiro & Bates (2006) realiza una revisión completa y detallada de cada uno de los escenarios que pueden tenerse en cuenta a partir del paquete “*nlme*”. West et al. (2014) por otro lado hace una revisión de incluyendo diferentes programas estadísticos, dentro de la cual se abarcan las librerías de R mencionadas previamente, mientras que Faraway (2016) ejemplifica el uso de la librería “*lme4*”, a partir de ejemplos, complementando la teoría.

---

---

## Simulación de datos

---

---

Para llevar a cabo la comparación entre diferentes tipos de modelos de regresión bajo condiciones controladas se hace uso de la simulación, de manera que se conocen los valores verdaderos de los parámetros. Más específicamente en el contexto de este documento se sabe de antemano si hay o no diferencias significativas en la respuesta entre los niveles del factor en consideración previo a realizar el ajuste de los diferentes modelos, con el fin de evaluar la eficacia de las metodologías a la hora de detectarlas. Este ejercicio utiliza el promedio de cuadrados de los errores de las estimaciones de cada modelo, es decir su cuadrado medio del error (*CME*), como una manera para realizar la comparación entre las diferentes metodologías descritas.

### 2.1. Descripción de la simulación

La simulación a realizar se basa en el esquema de muestreo destructivo de UO, es decir, aquellas que son medidas en un momento particular no son observadas posteriormente, considerando que ciertos grupos de éstas provienen de una misma UE desde un tiempo inicial. A manera de ejemplificación, la tabla 2.1 muestra el esquema de muestreo con dos UE y de cada una de ellas dos UO medidas en cada uno de los  $t$  tiempos, es decir que el símbolo \* corresponden al momento donde se realiza la medición sobre las unidades. Allí se puede observar que las UO medidas en un momento no son medidas posteriormente.

Sin pérdida de generalidad, por simplicidad en este ejercicio se tendrá en cuenta un único tratamiento con dos grupos, cuyos efectos sobre la respuesta simulada se denotarán como  $A_i$  con  $i = 1, 2$ . Por otro lado los efectos aleatorios considerados son  $b_{k(i)}$ , el cual es el efecto de la  $k$ -ésima UE en el  $i$ -ésimo tratamiento,  $\eta_{l(ik)}$  es el efecto de la  $l$ -ésima UO dentro de la  $k$ -ésima UE del  $i$ -ésimo tratamiento, y por último,  $\epsilon_{ijkl}$  un error aleatorio asociado a las mediciones dentro de las UO.

De esta manera el modelo estadístico simulado es

$$y_{ijkl} = \mu + A_i + b_{k(i)} + \eta_{l(ik)} + T_j + AT_{ij} + \epsilon_{ijkl} \quad (2.1)$$

con  $i = 1, 2$ ,  $j = 1, \dots, t$ ,  $k = 1, \dots, K$  y  $l = 1, \dots, R$ , donde  $K$  es el número de UE en cada tratamiento,  $R$  es el número de UO simuladas originalmente dentro de cada UE,  $t$  es

TABLA 2.1. Tabla para ilustrar el esquema de muestreo considerado en la simulación.

Unidad experimental	Unidad observacional	Tiempo			
		$T_1$	$T_2$	$\dots$	$T_t$
UE <sub>1</sub>	UO <sub>1</sub>	*			
UE <sub>1</sub>	UO <sub>2</sub>	*			
UE <sub>1</sub>	UO <sub>3</sub>		*		
UE <sub>1</sub>	UO <sub>4</sub>		*		
⋮	⋮			⋱	
UE <sub>1</sub>	UO <sub>L-1</sub>				*
UE <sub>1</sub>	UO <sub>L</sub>				*
UE <sub>2</sub>	UO <sub>1</sub>	*			
UE <sub>2</sub>	UO <sub>2</sub>	*			
UE <sub>2</sub>	UO <sub>3</sub>		*		
UE <sub>2</sub>	UO <sub>4</sub>		*		
⋮	⋮			⋱	
UE <sub>2</sub>	UO <sub>L-1</sub>				*
UE <sub>2</sub>	UO <sub>L</sub>				*

el número de tiempos considerados,  $\mu$  es un promedio global. Por otra parte  $A_i$  y  $T_j$  son los efectos del  $i$ -ésimo tratamiento y del  $j$ -ésimo tiempo, respectivamente, mientras que  $AT_{ij}$  es el efecto de su interacción. Por último  $y_{ijkl}$  es la variable respuesta medida en la  $l$ -ésima UO de la  $k$ -ésima UE en el  $j$ -ésimo tiempo y correspondiente al  $i$ -ésimo tratamiento. Los paréntesis en los subíndices de los efectos  $b_{k(i)}$  y  $\eta_{(ik)}$  hacen referencia a factores vivos e inertes. Por ejemplo,  $b_{k(i)}$  es el coeficiente de la  $k$ -ésima UE dentro del  $i$ -ésimo tratamiento, por lo cual, en este caso el tratamiento es un factor inerte, mientras que el de la UE es un factor vivo. En adelante se hará omisión a esta notación por simplicidad, es decir que éstos factores se notarán como  $b_{ik}$  y  $\eta_{ikl}$ . Este tipo de clasificación de factores puede consultarse al detalle en Melo et al. (2007).

Por condiciones de estimabilidad, en este modelo se suponen las siguientes restricciones:

$$\sum_{i=1}^2 A_i = \sum_{j=1}^t T_j = \sum_{i=1}^2 AT_{ij} = \sum_{j=1}^t AT_{ij} = 0,$$

y que  $b_{ik} \sim N(0, \sigma_b^2)$ ,  $\eta_{ikl} \sim N(0, \sigma_\eta^2)$  y  $\epsilon_{ijk} = [\epsilon_{ijkl}] \sim N(0, \Sigma(\xi))$  los cuales se asumen independientes entre ellos, es decir que  $Cov(b_{ik}, \epsilon_{ijkl}) = Cov(\eta_{ikl}, \epsilon_{ijkl}) = Cov(b_{ik}, \eta_{ikl}) = 0$  para todo  $i, j, k$  y  $l$ . En esta simulación  $\Sigma(\xi)$  corresponde a una matriz de varianzas y covarianzas de un proceso autoregresivo de orden uno ( $AR(1)$ ) y  $\xi = [\sigma^2, \rho]$ .

Como este trabajo se encuentra dentro del contexto de datos longitudinales se puede asumir que los errores dentro de las UO ( $\epsilon_{ijkl}$ ) pueden ser autocorrelacionados, por lo cual dentro del ejercicio se simulan los  $\epsilon_{ijkl}$  provenientes de un proceso  $AR(1)$ , aunque el ejercicio puede extenderse a diferentes tipos de autocorrelación temporal o inclusive espacial.

La primera parte del ejercicio se concentra en simular las variables independientes, tanto factores fijos como aleatorios, además de la variable respuesta  $y_{ijkl}$  mediante la ecuación (2.1). La segunda parte corresponde a simular la pérdida de información, de tal

manera que se llegue a una estructura similar a la descrita por la tabla 2.1, para lo cual se construyó la función en R *muestreo\_destructivo*(·) dada en el apéndice 3.2. Una vez se tenga una única medición de cada una de las UO, el problema principal es que no se podría hacer un seguimiento individual en el tiempo, por lo cual se estaría desconociendo una fuente de error aleatoria, es decir que no sería posible obtener una estimación de los efectos  $\eta_{ikl}$  y una de sus posibles consecuencias sería sobre las estadísticas de prueba, afectando los valores  $p$  y errores estándar correspondientes a los diferentes coeficientes.

A manera de ilustración se simuló inicialmente una tabla de datos asociada al modelo dado en (2.1). La figura 2.1 muestra este procedimiento, en donde las líneas punteadas hacen referencia a los perfiles en el tiempo de cada una de las UO. Los colores de las líneas identifican el tratamiento asignado y los títulos en cada recuadro corresponden a la numeración de las UE y sus tratamientos asignados, mientras que los puntos son las observaciones seleccionadas. Esto quiere decir que en el eje de las ordenadas se ilustran los valores simulados de  $y_{ijkl}$ , mientras que en el eje de las abscisas se representa los tiempos de medición. En cada UE y para cada uno de los tiempos se seleccionan cuatro UO, las cuales no vuelven a ser consideradas en tiempos posteriores.

Para realizar el ajuste de un modelo estadístico bajo el esquema y simulación considerados se plantean los escenarios descritos en las secciones 2.1.1, 2.1.2, 2.1.3 y 2.1.4 para realizar su ajuste, evaluación y respectiva comparación.

### 2.1.1. Modelo exclusivo de efectos fijos

Considerando que las UO solamente son medidas en una ocasión, un primer acercamiento es el modelo de regresión lineal que asume independencia entre todas sus mediciones en la variable respuesta, únicamente considerando un error aleatorio además de los efectos fijos. Este modelo descrito se muestra en la siguiente ecuación:

$$y_{ijkl} = \mu + A_i + T_j + AT_{ij} + \epsilon_{ijkl}^*, \quad (2.2)$$

con  $i = 1, 2$ ,  $j = 1, \dots, t$ ,  $k = 1, \dots, K$ ,  $l = 1, \dots, R$ , y además,

$$\sum_{i=1}^2 A_i = \sum_{j=1}^t T_j = \sum_{i=1}^2 AT_{ij} = \sum_{j=1}^t AT_{ij} = 0 \text{ y } \epsilon_{ijkl}^* \stackrel{i.i.d.}{\sim} N(0, \sigma_{\epsilon^*}^2).$$

Las hipótesis que se pueden evaluar a partir de la tabla de análisis de varianza del modelo dado en la ecuación (2.2) y que se considerarán de igual forma en los siguientes casos son las que se relacionan con los efectos  $AT_{ij}$ ,  $A_i$ ,  $T_j$ , es decir que es de interés evaluar si hay diferencias significativas entre sus respectivos niveles. De esta manera, las hipótesis nulas para cada caso son de la forma:

$$\begin{aligned} H_0^{(1)}: & AT_{ij} = AT_{i'j'} \text{ con } i \neq i' \text{ y } j \neq j'. \\ H_0^{(2)}: & A_i = A_{i'} \text{ con } i \neq i'. \\ H_0^{(3)}: & T_i = T_j \text{ con } j \neq j'. \end{aligned}$$

La tabla 2.2 corresponde al análisis de varianza del modelo de la ecuación (2.2), en donde únicamente se incluye una fuente de variación aleatoria, denotada por  $\epsilon_{ijkl}^*$ . La

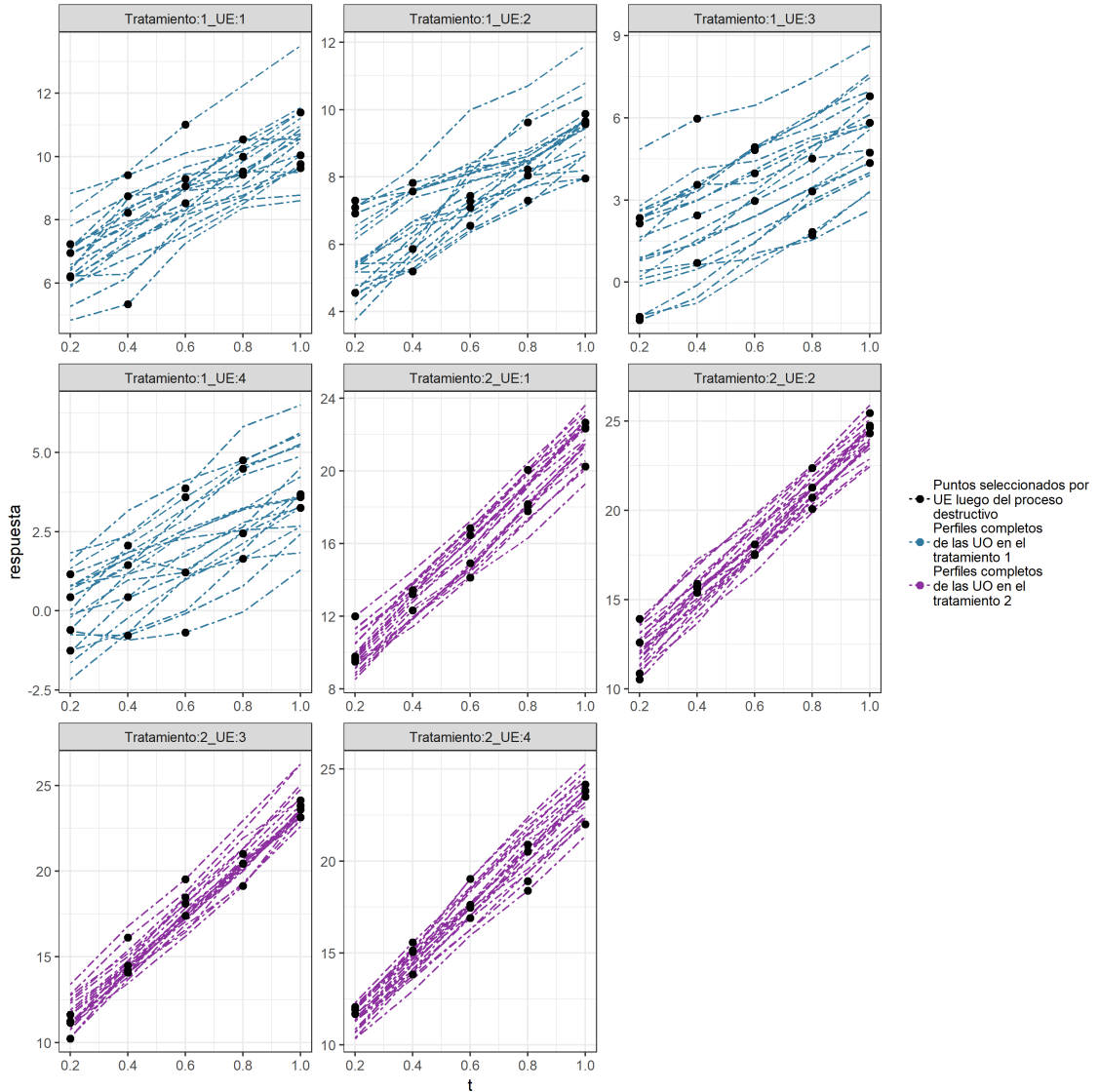


FIGURA 2.1. Diagrama de dispersión que ilustra el esquema de muestreo utilizado, cuyas líneas punteadas representan los perfiles completos para cada UO y los puntos son las observaciones seleccionadas en cada uno de los tiempos.

variable respuesta se denota por  $y_{ijkl}$  que corresponde a la medición de la  $l$ -ésima UO posterior a la selección de muestreo destructivo, en la  $k$ -ésima UE, para el  $j$ -ésimo tiempo y en el  $i$ -ésimo tratamiento. La interpretación de los coeficientes  $A_i$ ,  $T_j$  y  $AT_{ij}$  es exactamente la misma a la del modelo descrito en la ecuación (2.1). Para efectos de este documento los cuadrados medios de los factores se calculan dividiendo las respectivas sumas de cuadrado sobre sus grados de libertad. Las estadísticas de prueba para los efectos en la ecuación (2.2) se calculan dividiendo los respectivos cuadrados medios asociados a los factores por el cuadrado medio del error, es decir que las estadísticas de prueba relacionadas con  $H_0^{(1)}$ ,  $H_0^{(2)}$  y  $H_0^{(3)}$  se calculan de la siguiente forma:

$$F_{AT} = \frac{CM(AT)}{CME}, \quad F_A = \frac{CM(A)}{CME} \text{ y } F_T = \frac{CM(T)}{CME}, \text{ respectivamente.}$$

TABLA 2.2. Tabla de análisis de varianza para el modelo dado en la ecuación 2.2.

Factor	Grados de l.	Suma de cuadrados	Esperanza de los CM
$A_i$	$I - 1$	$\sum_i \frac{y_{i...}^2}{JKR} - \frac{y_{...}^2}{IJKR}$	$\sigma_{\epsilon^*}^2 + \frac{JKR}{(I-1)} \sum_i A_i^2$
$T_j$	$J - 1$	$\sum_j \frac{y_{.j..}^2}{IKR} - \frac{y_{...}^2}{IJKR}$	$\sigma_{\epsilon^*}^2 + \frac{IKR}{(J-1)} \sum_j T_j^2$
$AT_{ij}$	$(I-1)(J-1)$	$\sum_{ij} \frac{y_{ij.}^2}{KR} - \sum_j \frac{y_{.j..}^2}{IKR} - \sum_i \frac{y_{i...}^2}{JKR} + \frac{y_{...}^2}{IJKR}$	$\sigma_{\epsilon^*}^2 + \frac{KR}{(I-1)(J-1)} \sum_{ij} AT_{ij}^2$
$\epsilon_{ijkl}^*$	$IJ(KR-1)$	$\sum_{ijkl} y_{ijkl}^2 - \sum_{ij} \frac{y_{ij.}^2}{KR}$	$\sigma_{\epsilon^*}^2$
Total	$IJKR - 1$		

### 2.1.2. Aplicación del modelo para pseudo-panel

En Deaton (1985) se plantea un modelo de regresión propuesto para el caso en que las mediciones se hacen periódicamente a grupos de individuos, entre los cuales se tiene una rotación, es decir que las personas que se observan en un tiempo particular pueden o no presentarse para mediciones posteriores por motivos externos a la investigación. Esta rotación puede ser hasta de un 100% de las unidades. Por ejemplo cuando es de interés realizar mediciones a partir de extracciones de sangre en individuos, en los cuales solamente se puede hacer una única extracción para no afectar su bienestar. El planteamiento descrito se basa en hacer una regresión lineal con base en los promedios de variables dependientes e independientes con respecto a cohortes establecidas de una manera particular.

En el contexto de la simulación, considerando que las UO son medidas en un único tiempo por condiciones propias de la experimentación, se puede aplicar un modelo para datos de tipo pseudo-panel, como el descrito en la ecuación (1.21), donde las agrupaciones serán conformadas por las UE, ya que éstas son consideradas en todos los tiempos en el estudio. Para este trabajo, el factor asociado a las UE, que es incluido dentro del modelo de regresión es de tipo aleatorio, ya que únicamente se trabaja con una muestra aleatoria de estas unidades y las conclusiones obtenidas se desean generalizar a toda la población.

Para este modelo se calculan los promedios por UE en cada uno de los tiempos a partir de la ecuación (2.1), de donde se obtiene la ecuación

$$\begin{aligned} \bar{y}_{ijk.} &= \sum_{l=1}^R \{ \mu + A_i + b_{ik} + \eta_{ikl} + T_j + AT_{ij} + \epsilon_{ijkl} \} / R \\ &= \mu + A_i + b_{ik}^* + T_j + AT_{ij} + \bar{\epsilon}_{ijk.} \end{aligned} \quad (2.3)$$

con  $i = 1, 2, j = 1, \dots, t$ , y  $k = 1, 2, \dots, K$ , donde  $b_{ik}^* = \sum_{l=1}^R \frac{(b_{ik} + \eta_{ikl})}{R}$ ,  $\bar{y}_{ijk.} = \sum_{l=1}^R \frac{y_{ijkl}}{R}$  y  $\bar{\epsilon}_{ijk.} = \sum_{l=1}^R \frac{\epsilon_{ijkl}}{R}$ . Por condiciones de estimabilidad se tienen las siguientes restricciones:

$$\sum_{i=1}^2 A_i = \sum_{j=1}^t T_j = \sum_{i=1}^2 AT_{ij} = \sum_{j=1}^t AT_{ij} = 0.$$

Por otro lado, sobre los factores aleatorios se supone que  $b_{ik}^* \sim N(0, \sigma_{b^*}^2)$  y  $\bar{\epsilon}_{ijk} \stackrel{i.i.d}{\sim} N(0, \sigma_{\epsilon}^2)$ . Para el modelo dado en la ecuación (2.3)  $\bar{y}_{ijk}$  y  $\bar{\epsilon}_{ijk}$  son la respuesta y el error aleatorio promedio para la  $k$ -ésima UE en el  $j$ -ésimo tiempo asignada en el  $i$ -ésimo tratamiento, respectivamente, mientras que  $b_{ik}^*$  es el efecto aleatorio asociado a la  $k$ -ésima UE en el  $i$ -ésimo tratamiento. La interpretación de los coeficientes  $A_i$ ,  $T_j$  y  $AT_{ij}$  es exactamente la misma a la del modelo descrito en la ecuación (2.1).

A partir del modelo dado en la ecuación (2.3) se deduce su tabla de análisis de varianza dada en la tabla 2.3, en donde se adiciona un efecto aleatorio asociado a las UE medidas en el tiempo, con respecto a la tabla 2.2 del modelo en la ecuación (2.2). La inclusión de este efecto aleatorio hace que los  $CME$  de las tablas 2.2 y 2.3 sean diferentes, por ende las estadísticas de prueba y conclusiones pueden ser diferentes. Para evaluar  $H_0^{(1)}$ ,  $H_0^{(2)}$  y  $H_0^{(3)}$ , las estadísticas de prueba se calculan de la forma:

$$F_{AT} = \frac{CM(AT)}{CME}, \quad F_T = \frac{CM(T)}{CME} \text{ y } F_A = \frac{CM(A)}{CM(b^*)}, \text{ respectivamente.}$$

El modelo dado en la ecuación (2.3) se conoce en la literatura del diseño experimental como un caso específico de parcelas divididas. Todo el detalle y una gran variedad de este tipo de diseño se pueden consultar en Federer & King (2007). Comparando este modelo con la regresión original descrito en la ecuación (2.1), se observa la ausencia del efecto aleatorio asociado a las UO medidas a través del tiempo denotado como  $\eta_{ikl}$ , por lo cual se debe evaluar el impacto que tiene la ausencia de este factor. Por esto, en la sección 2.1.3 se propone un modelo para su evaluación y comparación.

TABLA 2.3. Tabla de análisis de varianza para el modelo dado en la ecuación (2.3).

Factor	Grados de l.	Suma de cuadrados	Esperanza de los CM
$A_i$	$I - 1$	$\sum_i \frac{y_{i...}^2}{JKR} - \frac{y_{...}^2}{IJKR}$	$\sigma_{\epsilon}^2 + J\sigma_{b^*}^2 + \frac{JKR}{(I-1)} \sum_i A_i^2$
$b_{ik}^*$	$I(KR - 1)$	$\sum_{ijk} \frac{y_{i.kl}^2}{J} - \frac{y_{i...}^2}{JKR}$	$\sigma_{\epsilon}^2 + J\sigma_{b^*}^2$
$T_j$	$J - 1$	$\sum_j \frac{y_{.j..}^2}{IKR} - \frac{y_{...}^2}{IJKR}$	$\sigma_{\epsilon}^2 + \frac{IKR}{(J-1)} \sum_j T_j^2$
$AT_{ij}$	$(I-1)(J-1)$	$\sum_{ij} \frac{y_{ij..}^2}{KR} - \sum_j \frac{y_{.j..}^2}{IKR} - \sum_i \frac{y_{i...}^2}{JKL} + \frac{y_{...}^2}{IJKR}$	$\sigma_{\epsilon}^2 + \frac{KR}{(I-1)(J-1)} \sum_{ij} AT_{ij}^2$
$\epsilon_{ijkl}$	$I(J-1)(KR-1)$	$\sum_{ijkl} y_{ijkl}^2 - \sum_{ij} \frac{y_{ij..}^2}{KR} - \sum_{ikl} \frac{y_{i.kl}^2}{J} + \sum_i \frac{y_{i...}^2}{JKR}$	$\sigma_{\epsilon}^2$
Total	$IJKR - 1$		

### 2.1.3. Modelo de efectos mixtos propuesto

El problema principal considerando que las mismas UO no son medidas en todos los tiempos es que no se puede hacer seguimientos individuales. Con el modelo dado en la

sección 2.1.2 se puede ajustar un efecto promedio de las UE, pero en comparación con el modelo original simulado de la ecuación (2.1) se puede estar omitiendo un efecto aleatorio asociado a las UO, denotado por  $\eta_{ikl}$ , el cual se recupera en el modelo descrito en esta sección. Es claro que cada unidad se mide una única vez y con una sola medición no es viable estimar un parámetro. De esta manera lo que se propone es conformar subgrupos de UO dentro de cada UE, de tal forma que se pueda estimar el efecto de cada uno de estos subgrupos con respecto a la variable respuesta, lo cual se hace con el fin de no omitir fuentes de variación. Siguiendo el ejemplo 1 se puede pensar que dentro de cada salón de clases se van a conformar diferentes grupos de acuerdo a su género, fecha de nacimiento o alguna característica propia de los estudiantes.

En este caso se busca ajustar un modelo que incluya efectos aleatorios correspondientes a las UO dentro de las UE, lo cual se dificulta por no tener las mismas unidades en todos los tiempos. Para hacerlo se deben conformar agrupaciones de UO que sean similares, de tal forma que se tengan mediciones de éstas en todos los tiempos. En el contexto del ejemplo 1 del capítulo 1 lo que se haría es equiparar aquellos estudiantes de buen rendimiento en cada uno de los tiempos, por otro lado, hacer lo mismo con aquellos de mal rendimiento. Es decir que en cada uno de los salones se pueden conformar dos subgrupos, los de buen y mal rendimiento, a los cuales se les podría hacer un seguimiento, a pesar de que los estudiantes no sean los mismos en los diferentes momentos de medición. Aplicando esta idea a los resultados ilustrados en la gráfica 2.1, se obtiene la gráfica 2.2, en donde dentro de cada UE se obtiene dos grupos de observaciones, que se identifican mediante los puntos de colores rojo y negro. La idea es que mediante aquellos puntos de color rojo se ajuste un coeficiente para aquellas UO con valores superiores, por otro lado, con aquellos de color negro se ajuste un valor para las que tienen valores inferiores, lo cual se realiza con el fin de poder incluir el factor de variación entre UO. Es de resaltar que un acercamiento inicial es con dos grupos dentro de cada UE, pero la idea se puede trabajar con más grupos, en cuyo caso se podrían conformar mediante divisiones de los valores más pequeñas como cuartiles, determinados percentiles o a partir de información auxiliar de las UO que no haya sido incluida dentro del modelo estadístico como por ejemplo el género de cada estudiante para el ejemplo mencionado.

De esta manera el modelo considerado viene dado en la ecuación (2.4), en donde  $b_{ik}$  es el efecto aleatorio asociado a las UE, mientras que  $\eta_{ijk}^*$  es el efecto asociado a las agrupaciones conformadas dentro de cada UE. En este modelo  $y_{ijklm}$  es la respuesta asociada a la  $m$ -ésima réplica en el  $l$ -ésimo grupo de UO de la  $k$ -ésima UE para el  $j$ -ésimo tiempo y en el  $i$ -ésimo tratamiento y  $\epsilon_{ijklm}$  su respectivo residual. La tabla 2.4 muestra el análisis de varianza correspondiente al modelo en cuestión.

$$y_{ijklm} = \mu + A_i + b_{ik} + \eta_{ikl}^* + T_j + AT_{ij} + \epsilon_{ijklm} \quad (2.4)$$

con  $i = 1, 2, j = 1, \dots, t, k = 1, 2, \dots, K, l = 1, 2, m = 1, \dots, M$ . Las condiciones de estimabilidad son las siguientes:

$$\sum_{i=1}^2 A_i = \sum_{j=1}^t T_j = \sum_{i=1}^2 AT_{ij} = \sum_{j=1}^t AT_{ij} = 0.$$

Por último, se asume que los efectos aleatorios son independientes entre ellos y que cada uno de ellos sigue distribución normal, de la siguiente forma:

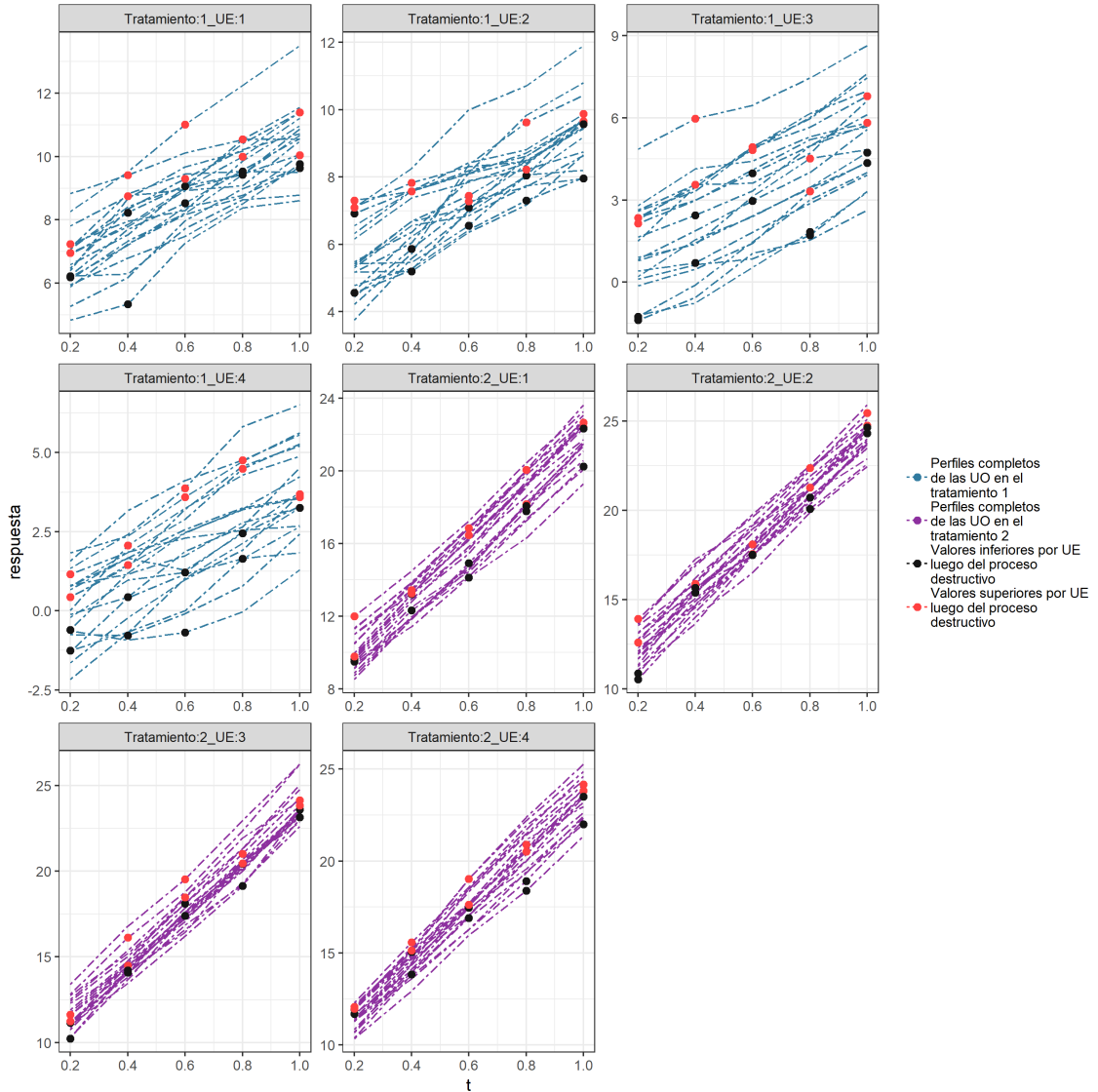


FIGURA 2.2. Diagrama de dispersión que ilustra el esquema de muestreo utilizado, cuyas líneas punteadas representan los perfiles completos para cada UO y los colores de los puntos son agrupaciones de valores superiores e inferiores de UO en cada tiempo.

$$b_{ik} \sim N(0, \sigma_b^2), \quad \eta_{ikl}^* \stackrel{i.i.d}{\sim} N(0, \sigma_{\eta^*}^2) \quad \text{y} \quad \epsilon_{ijklm} \stackrel{i.i.d}{\sim} N(0, \sigma_\epsilon^2).$$

#### 2.1.4. Modelo de análisis de varianza multivariado

Un último modelo considerado en este ejercicio es el considerado en la sección 1.1, donde se describió el modelo de análisis de varianza multivariado. A partir del esquema de muestreo no se puede hacer un seguimiento individual a las UO para la conformación de los vectores de variable respuesta, se hace uso de las agrupaciones realizada en la sección 2.1.3, donde se conformaban dos grupos dentro de cada UE. Con estos lo que se hace es calcular el promedio en cada uno de ellos, es decir

TABLA 2.4. Tabla de análisis de varianza para el modelo dado en la ecuación (2.4).

Factor	Grados de l.	Suma de cuadrados	Esperanza de los CM
$A_i$	$I - 1$	$\sum_i \frac{y_{i..}^2}{JKL} - \frac{y_{...}^2}{IJKL}$	$\sigma_\epsilon^2 + J\sigma_{\eta^*}^2 + JL\sigma_b^2 + \frac{JKL}{(I-1)} \sum_i A_i^2$
$b_{ik}$	$I(K-1)$	$\sum_{ijk} \frac{y_{i.k.}^2}{J} - \frac{y_{i..}^2}{JKL}$	$\sigma_\epsilon^2 + J\sigma_{\eta^*}^2 + JL\sigma_b^2$
$\eta_{ikl}^*$	$IK(L-1)$	$\sum_{ijk} \frac{y_{i.kl}^2}{J} - \frac{y_{i.k.}^2}{JL}$	$\sigma_\epsilon^2 + J\sigma_{\eta^*}^2$
$T_j$	$J - 1$	$\sum_j \frac{y_{.j..}^2}{IKL} - \frac{y_{...}^2}{IJKL}$	$\sigma_\epsilon^2 + \frac{IKL}{(J-1)} \sum_j T_j^2$
$AT_{ij}$	$(I-1)(J-1)$	$\sum_{ij} \frac{y_{ij..}^2}{KL} - \sum_j \frac{y_{.j..}^2}{IKL} - \sum_i \frac{y_{i..}^2}{JKL} + \frac{y_{...}^2}{IJKL}$	$\sigma_\epsilon^2 + \frac{KL}{(I-1)(J-1)} \sum_{ij} AT_{ij}^2$
$\epsilon_{ijkl}$	$I(J-1)(KL-1)$	$\sum_{ijkl} y_{ijkl}^2 - \sum_{ij} \frac{y_{ij..}^2}{KL} - \sum_{ikl} \frac{y_{i.kl}^2}{J} + \sum_i \frac{y_{i..}^2}{JKL}$	$\sigma_\epsilon^2$
Total	$IJKL - 1$		

$$\bar{y}_{ijkl} = \sum_m \frac{y_{ijklm}}{M},$$

el cual es el promedio de la  $l$ -ésima agrupación de la  $k$ -ésima UE en el  $j$ -ésimo tiempo. De esta manera, los vectores de respuestas en el tiempo vienen dados de la forma

$$\mathbf{y}_{ikl} = [\bar{y}_{i1kl} \ \bar{y}_{i2kl} \ \cdots \ \bar{y}_{itkl}].$$

La matriz de respuestas del modelo dado en la ecuación (1.3) en este contexto es de la forma

$$\mathbf{Y} = [\mathbf{y}_{111}^t \ \mathbf{y}_{112}^t \ \mathbf{y}_{121}^t \ \mathbf{y}_{122}^t \ \cdots \ \mathbf{y}_{1K1}^t \ \mathbf{y}_{1K2}^t \ \mathbf{y}_{211}^t \ \mathbf{y}_{212}^t \ \mathbf{y}_{221}^t \ \mathbf{y}_{222}^t \ \cdots \ \mathbf{y}_{2K1}^t \ \mathbf{y}_{2K2}^t]^t,$$

mientras que la matriz de coeficientes  $\mathbf{B}$  y la de diseño  $\mathbf{X}$  son de la forma

$$\mathbf{B} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1t} \\ A_{21} & A_{22} & \cdots & A_{2t} \end{bmatrix} \text{ y } \mathbf{X} = [\mathbf{I}_2 \otimes \mathbf{1}_{2 \times 2 \times K}],$$

donde  $A_{ij}$  representa el efecto del  $i$ -ésimo tratamiento en el  $j$ -ésimo tiempo.

### 2.1.5. Funciones en R

El código construido consideró un esquema de muestreo destructivo, es decir que una vez que se simuló el modelo dado en la ecuación (2.1) se tomó una muestra aleatoria de UO, de manera como se ha descrito previamente. Para obtener conclusiones más acertadas, cada vez que se simularon los datos se repitió cien veces el proceso de generar la pérdida de información, obteniendo una muestra final tal y como ilustran los puntos de la gráfica 2.2 y el ajuste de los cuatro modelos considerados. De esta manera la función

```
Simulacion_modelo(n_times,n_uobs_final,n_uexp,n_trat,coefficients,
variances,correlations,dif_A,correlation_type,porcentaje_panel,intercept)
```

se basa en una serie de parámetros, mediante los cuales genera la base de datos completa, ajusta un modelo con todos los datos para una comparación posterior y a partir de la función

```
Muestreo_destrutivo(UE,porcentaje_panel)
```

se genera la pérdida de información para cada UE. La función

```
Ejecucion_modelos(df1,coefficients,porcentaje_panel,intercept)
```

se encarga de realizar el ajuste de los cuatro modelos para cada muestra seleccionada. Los códigos correspondientes a las tres funciones mencionadas se encuentran en el apéndice 3.2.

## 2.2. Resultados de la simulación

A partir de los cuatro modelos descritos se realizó la respectiva evaluación para determinar cuál era el mejor. Ya que los tres primeros planteamientos en las secciones 2.1.1, 2.1.2 y 2.1.3 consideran los mismos efectos fijos, las estimaciones de los coeficientes  $A_i$ ,  $T_j$ ,  $AT_{ij}$  son las mismas, pero lo que cambia son las estimaciones de sus errores estándar y en general el cuadrado medio del error con el cual se calculan las estadísticas de prueba. Es por este motivo que la comparación se realizó con base al *CME* y no con respecto a las estimaciones de estos parámetros

Para realizar la simulación del modelo dado en la ecuación (2.1) se considero  $t = 10$ ,  $K = 4$ ,  $L = 10$ , y además, se tomaron los siguientes valores para las varianzas:

$$\sigma_b^2 = 5, \sigma_\eta^2 = 2 \text{ y } \sigma_\epsilon^2 = 1.$$

En primer instancia, los valores de  $\epsilon_{ijkl}$  fueron simulados a partir de un proceso autoregresivo de orden uno (*AR*(1)) con un parámetro  $\rho = 0.8$ . Se tomó una diferencia entre tratamientos  $|A_2 - A_1| = 10$ , un incremento en el tiempo de aproximadamente cinco unidades en la respuesta por cada unidad de incremento en el tiempo. Por último, se incluyó una interacción entre los factores de tiempo y tratamiento, en la que se consideraron varios casos, es decir que se simularon diferentes escenarios dependiendo de una diferencia máxima entre los niveles de dicha interacción denotada como  $\Delta AT_{max}$ , la cual toma valores desde cero hasta diez.

En la primera parte de la comparación se simuló un modelo general (sin pérdida de información) a partir de los coeficientes dados, y para éste se generó el muestreo destructivo de UO, de tal forma que únicamente se tenga una medición de cada una en un tiempo específico (datos incompletos). A partir de estos se ajustaron los cuatro modelos descritos y se procede a calcular el *CME* para cada uno de ellos. El anterior procedimiento se repitió 200 veces, de forma que para cada caso se recolectan cuatro *CME*, uno para cada modelo. La figura 2.3 muestra las distribuciones de los 200 valores por grupo, mediante

diagramas de cajas y bigotes. Los encabezados de cada uno de los recuadros corresponden a los valores tomados por  $\Delta AT_{m\acute{a}x}$ . Allí se observa que el modelo con mayores valores en los  $CME$  es el de análisis de varianza multivariado, seguido por el de efectos fijos, luego el propuesto por Deaton (1985) y finalmente el propuesto en este documento, el cual ajusta un efecto aleatorio adicional a partir de subgrupos de  $UO$ . Para una mejor visualización de éstos dos últimos modelos, análogo a la figura 2.3, la figura 2.5 muestra las distribuciones de los  $CME$  a partir del modelo de Deaton y el propuesto, en donde se observa que el último tiene menores valores.

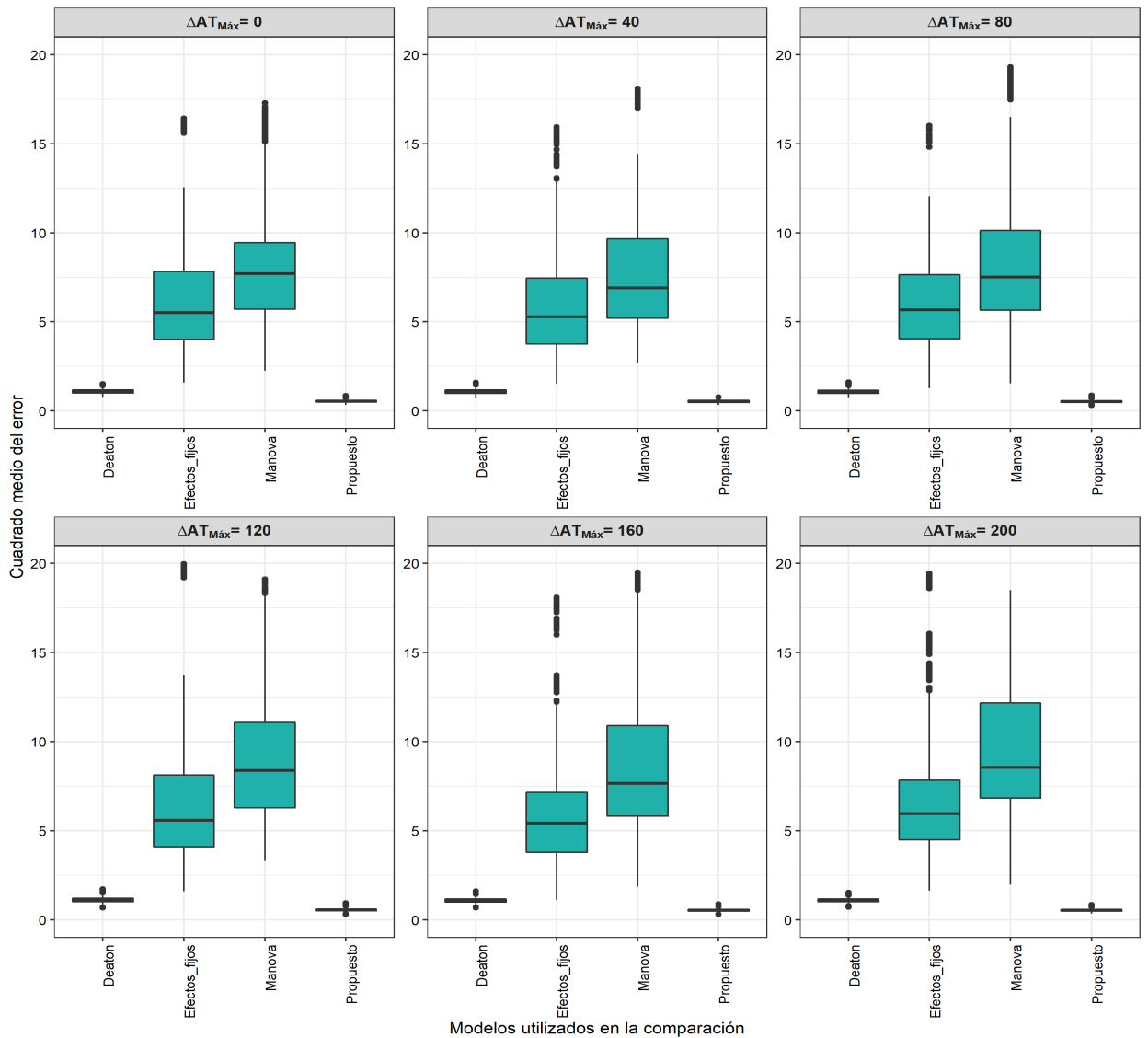


FIGURA 2.3.  $CME$  para cien simulaciones del esquema de muestreo destructivo para los cuatro modelos.

A manera ilustrativa y de forma análoga a la figura 2.2, la figura 2.4 muestra los valores observados (puntos negros) y los ajustados mediante los modelos descritos en las secciones 2.1.2 (línea roja) y 2.1.3 (líneas azules). Es importante notar que los valores observados del modelo propuesto se dividen en dos grupos (uno superior y otro inferior), con el objetivo de hacer la estimación de un efecto aleatorio  $\eta_{ikl}^*$ , tratando de hacer una analogía a los coeficientes  $\eta_{ikl}$  inicialmente simulados y ese es el motivo por el cual se visualizan dos líneas

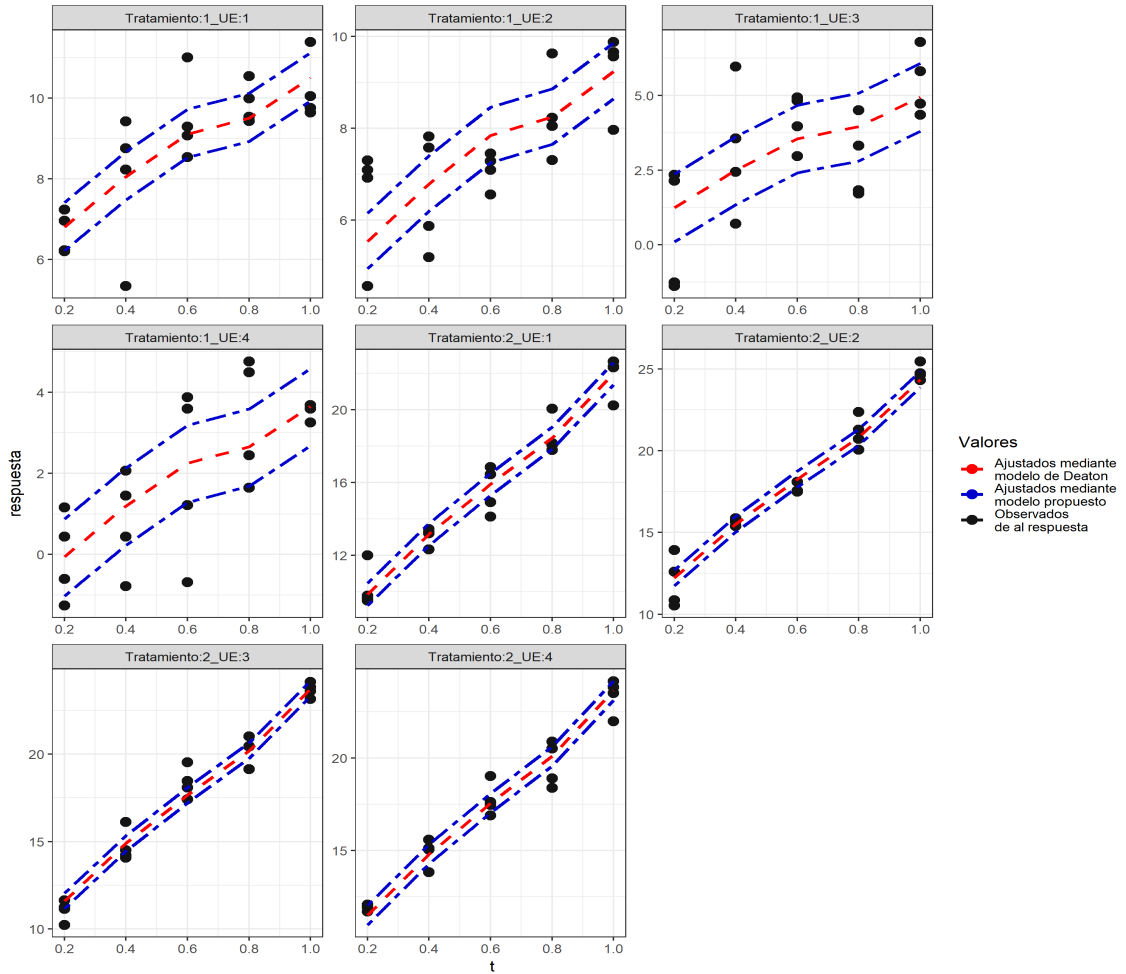


FIGURA 2.4. Diagrama de dispersión de los puntos observados en el tiempo junto con las estimaciones realizadas por los modelos de promedios (línea roja) y el propuesto en este documento (línea azul), el cual considera dos subgrupos dentro de las UE, uno de valores altos y otro de valores bajos.

azules en dicha figura. Adicionalmente, se observa que mediante el modelo propuesto se obtiene un mejor ajuste sobre los puntos observados.

A partir del modelo simulado se puede plantear inicialmente la hipótesis de diferencias significativas sobre los coeficientes de la interacción  $AT_{ij}$ , es decir  $H_0^{(1)}$ , la cual fue descrita previamente. Si  $H_0^{(1)}$  no se rechaza, es de interés juzgar las hipótesis de diferencias significativas entre tratamientos y entre tiempos, es decir, evaluar  $H_0^{(2)}$  y  $H_0^{(3)}$ , respectivamente (Hinkelmann 2011). Se debe tener en cuenta que dependiendo de la conclusión sobre  $H_0^{(1)}$  se procede a juzgar  $H_0^{(2)}$  o  $H_0^{(3)}$ .

Con el objetivo de juzgar  $H_0^{(1)}$  se tiene un valor  $p$  proveniente del modelo original que utiliza los datos completos, el cual sería el ideal, pero que en la práctica no se puede obtener por condiciones propias del experimento. A pesar de esto, dicho coeficiente es el que más se acerca a la realidad y por ende se utilizará como referencia en la siguiente comparación. Considerando las cuatro metodologías implementadas, se comparan sus resultados con el valor  $p$  de referencia obtenido a partir de la regresión simulada. Para ello, con cada

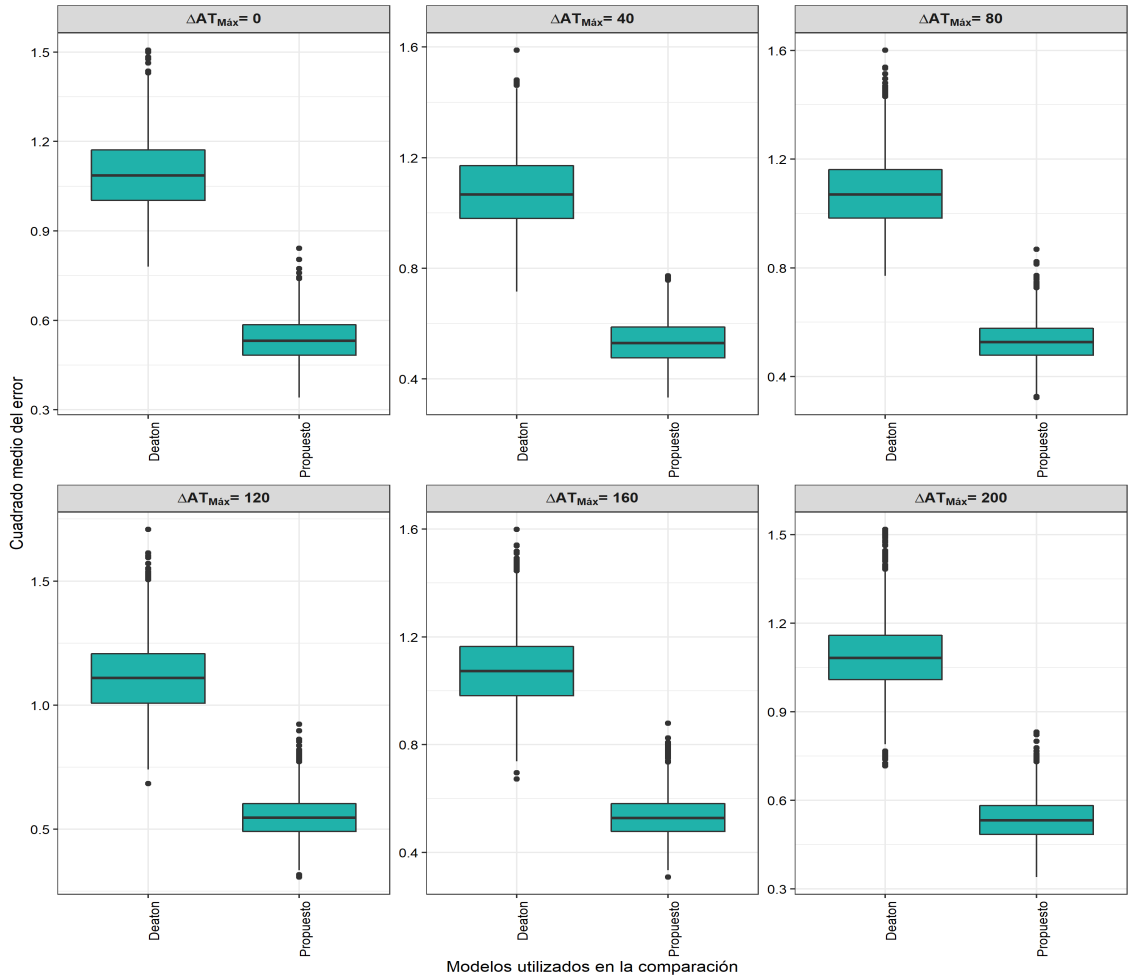


FIGURA 2.5. *CME* para cien simulaciones del esquema de muestreo destructivo para el modelo descrito en Deaton (1985) y el propuesto en este documento.

conjunto de datos completos se simuló la pérdida de información en cien ocasiones y de esta manera en cada una de estas repeticiones se recolectó el valor  $p$  asociado a  $H_0^{(1)}$  en cada uno de los modelos sometidos a la comparación. Este procedimiento se realizó para diferentes escenarios que dependen del valor asignado a  $\Delta AT_{max}$ . La figura 2.6 ilustra esta comparación, donde cada recuadro hace referencia a un  $\Delta AT_{max}$ . La línea roja horizontal representa el valor  $p$  de referencia, el cual se obtuvo mediante la regresión con todos los valores, es decir sin pérdida de información. Por otra parte, los diagramas de cajas y bigotes muestran las distribuciones de los valores  $p$  resultantes de las diferentes metodologías en cada uno de los escenarios. Es importante notar que en la prueba del análisis de varianza multivariado se tuvieron en cuenta las cuatro estadísticas de prueba descritas en la tabla 1.2. En la figura 2.6 se observa que los valores más similares al valor  $p$  de referencia, en cada uno de los casos son los asociados al modelo propuesto lo cual se debe principalmente a que en éste se incluye un factor aleatorio adicional. Luego del modelo propuesto, las pruebas realizadas con el análisis multivariado son las que más se acercan a la línea determinada por el modelo con todas las observaciones, seguidas del modelo de Deaton y por último, el que únicamente incluye efectos fijos. Además, como es de esperarse a medida que se

incrementan  $\Delta AT_{max}$  (valores en los encabezados) todas las pruebas tienden a rechazar  $H_0^{(1)}$ .

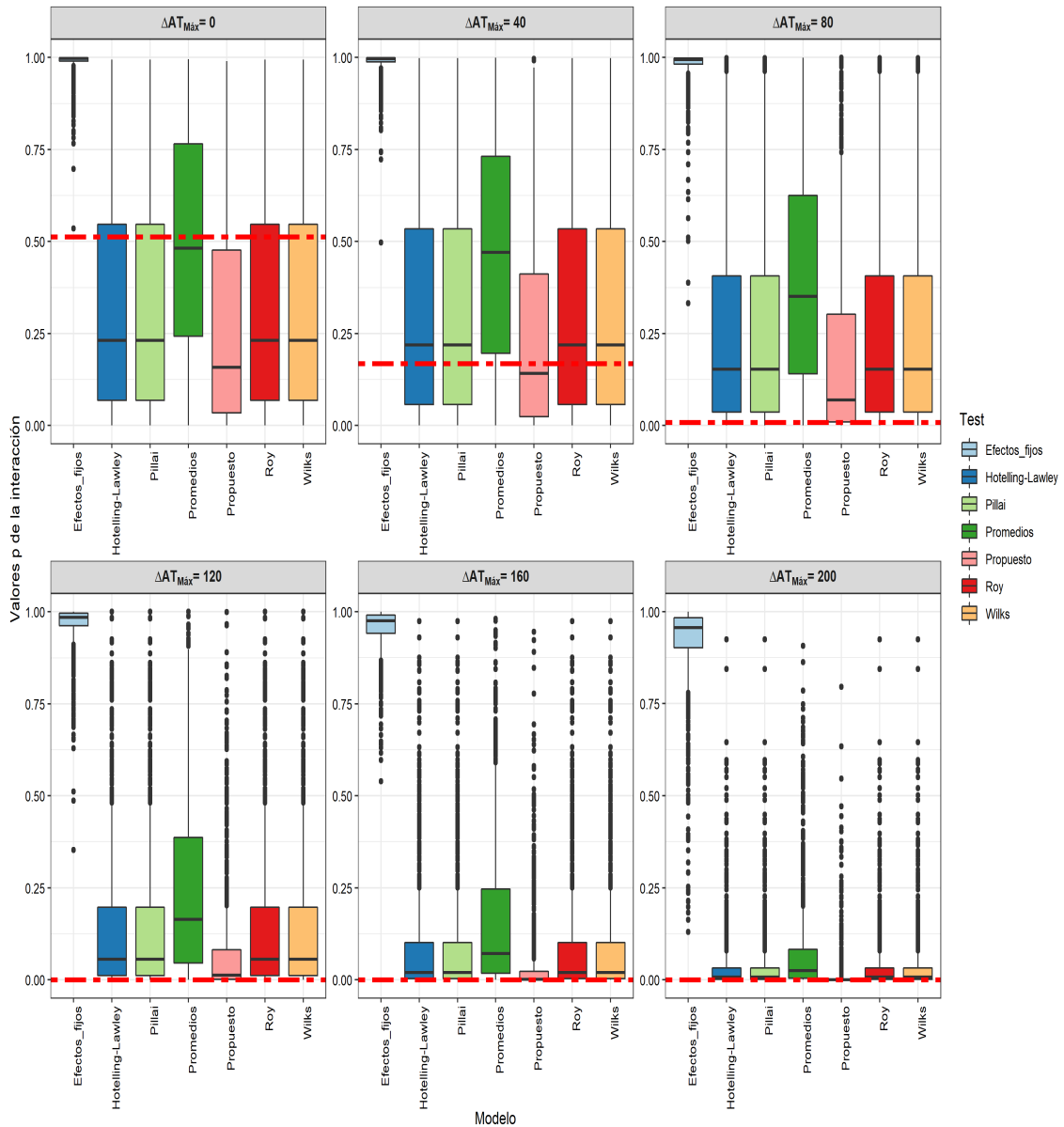


FIGURA 2.6. Diagramas de cajas y bigotes para las distribuciones de los valores  $p$  de las pruebas de significancia de la interacción para cada uno de los modelos y bajo distintas diferencias máximas entre los coeficientes  $AT_{ij}$ .

De manera análoga al ejercicio realizado con las interacciones  $AT_{ij}$  se realizó el mismo ejercicio de la figura 2.7, pero replicándolo para los efectos principales  $A_i$  de los tratamientos, es decir para juzgar  $H_0^{(2)}$  en cada una de las repeticiones, en donde la línea roja está al nivel del valor  $p$  utilizando el modelo con los datos completos, denominado como valor de referencia. De esta forma se hizo el ejercicio a partir de diferentes escenarios considerando una diferencia entre los dos niveles del factor  $A$ , es decir  $\Delta A_{max}$ . Esto con el fin de juzgar  $H_0^{(2)}$ . Por otro lado, los diagramas de cajas y bigotes representan las distribuciones de

valores  $p$  en las cien simulaciones para cada uno de los coeficientes (valores en los encabezados de los recuadros). En este caso los modelos con mayor similitud con cada línea roja son aquellos de efectos mixtos. Esto quiere decir que si el interés se plantea únicamente en detectar diferencias entre los efectos del factor principal  $A_i$ , el modelo de Deaton (1985) y el propuesto tendrán los mismos resultados en cuanto a las estadísticas de prueba y sus respectivos valores  $p$ . Seguidos de éstos modelos se encuentran las pruebas realizadas mediante el análisis de varianza multivariado y por último, nuevamente la regresión que únicamente incluye efectos fijos.

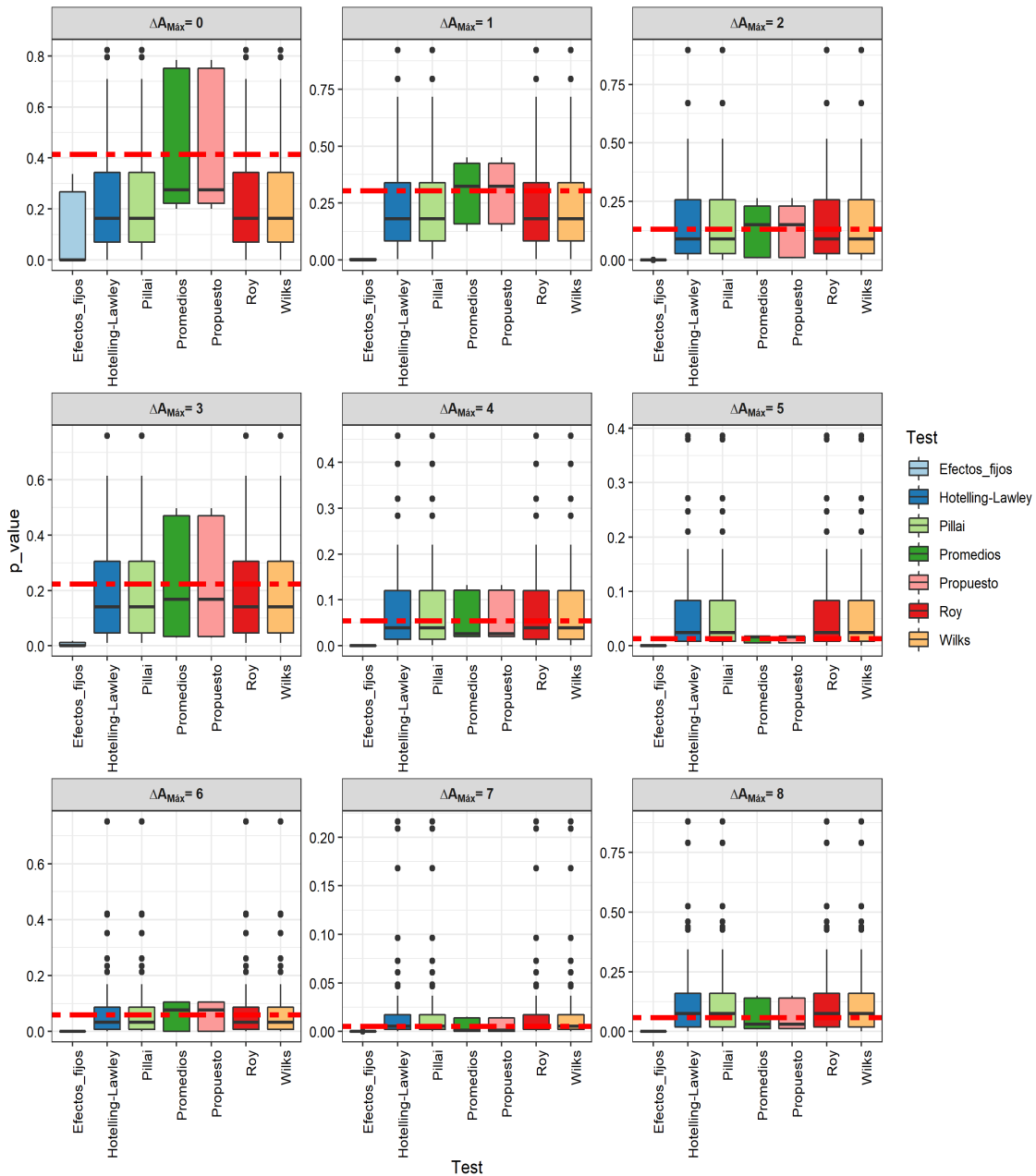


FIGURA 2.7. Diagramas de cajas y bigotes para las distribuciones de los valores  $p$  de las pruebas realizadas para juzgar  $H_0^{(2)}$  para cada uno de los modelos y bajo distintas diferencias máximas entre los coeficientes  $A_i$ .

Por último, se realizó el mismo ejercicio para evaluar  $H_0^{(3)}$ . Es decir que se asignaron diferentes valores para la diferencia máxima entre efectos asociados a los diferentes  $T_j$ , denotada en este caso como  $\Delta T_{max}$ . La comparación de valores  $p$  asociados a las cuatro metodologías, junto con el valor de referencia calculado a partir del modelo que incluyó la totalidad de los datos se muestra en la figura 2.8, donde nuevamente se observa que los valores asociados al modelo de promedios y el propuesto son los que más se aproximan al valor de referencia.

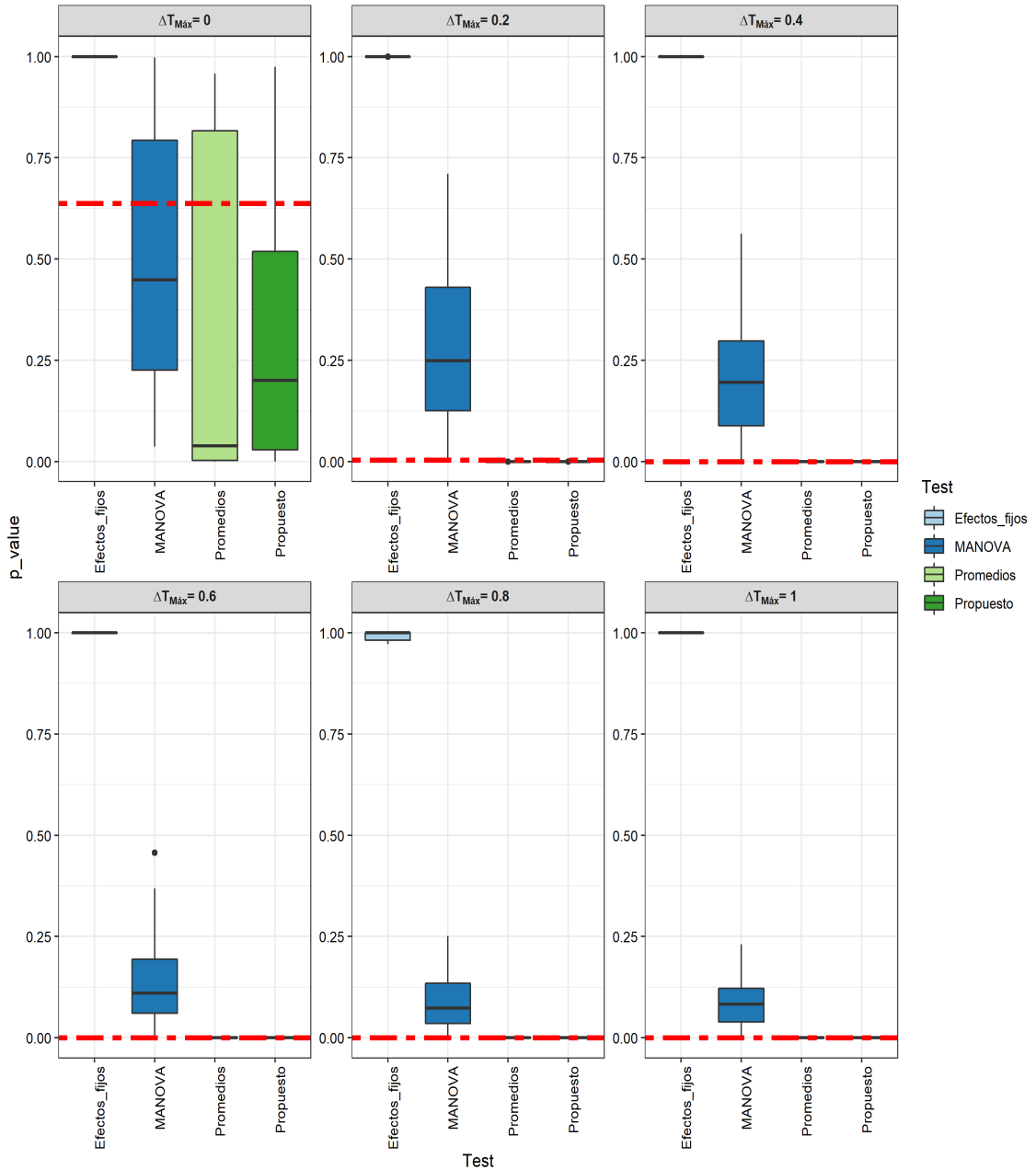


FIGURA 2.8. Diagramas de cajas y bigotes para las distribuciones de los valores  $p$  de las pruebas realizadas para juzgar  $H_0^{(3)}$  para cada uno de los modelos y bajo distintas diferencias máximas entre los coeficientes  $T_j$ .

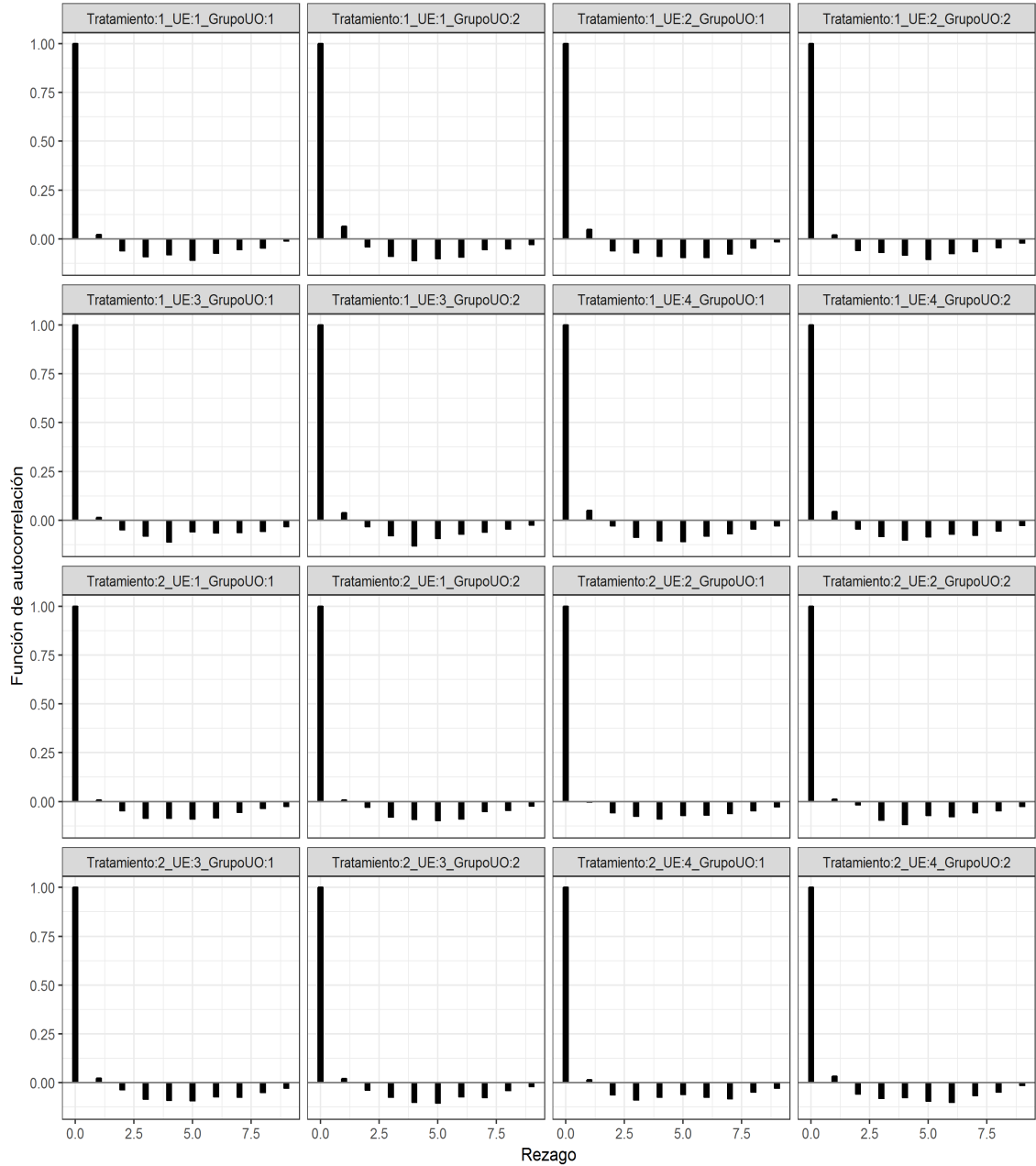


FIGURA 2.9. Correlogramas para los subgrupos de UO dentro de las diferentes UE y para los dos tratamientos en consideración en los residuales de uno de los ajustes del modelo dado en la ecuación (2.4).

Una observación importante a partir del modelo propuesto se da a partir de la evaluación de la autocorrelación temporal incluida dentro del modelo simulado. Es decir que a partir del modelo dado en la ecuación (2.1) en donde se incluyó una correlación temporal, más específicamente un proceso autocorrelacionado de primer orden ( $AR(1)$ ), se puede pensar en evaluar la presencia o ausencia de esta misma luego del proceso destructivo de UO en el modelo final. De esta manera la figura 2.9 muestra los correlogramas asociados a los residuales de modelo propuesto, descrito en la ecuación (2.4) para los diferentes grupos de UO conformadas. De manera análoga, la figura 2.10 muestra los autocorrelogramas de

los residuales por UE en el modelo dado para pseudo-panel, descrito en la ecuación (2.3). En ambas figuras no se observan autocorrelaciones importantes más allá del rezago cero, que por definición es igual a uno en todos los casos. Por ende no hay evidencia de que la autocorrelación temporal se recupera mediante los modelos de efectos mixtos (propuesto y de pseudo-panel) con las agrupaciones conformadas.

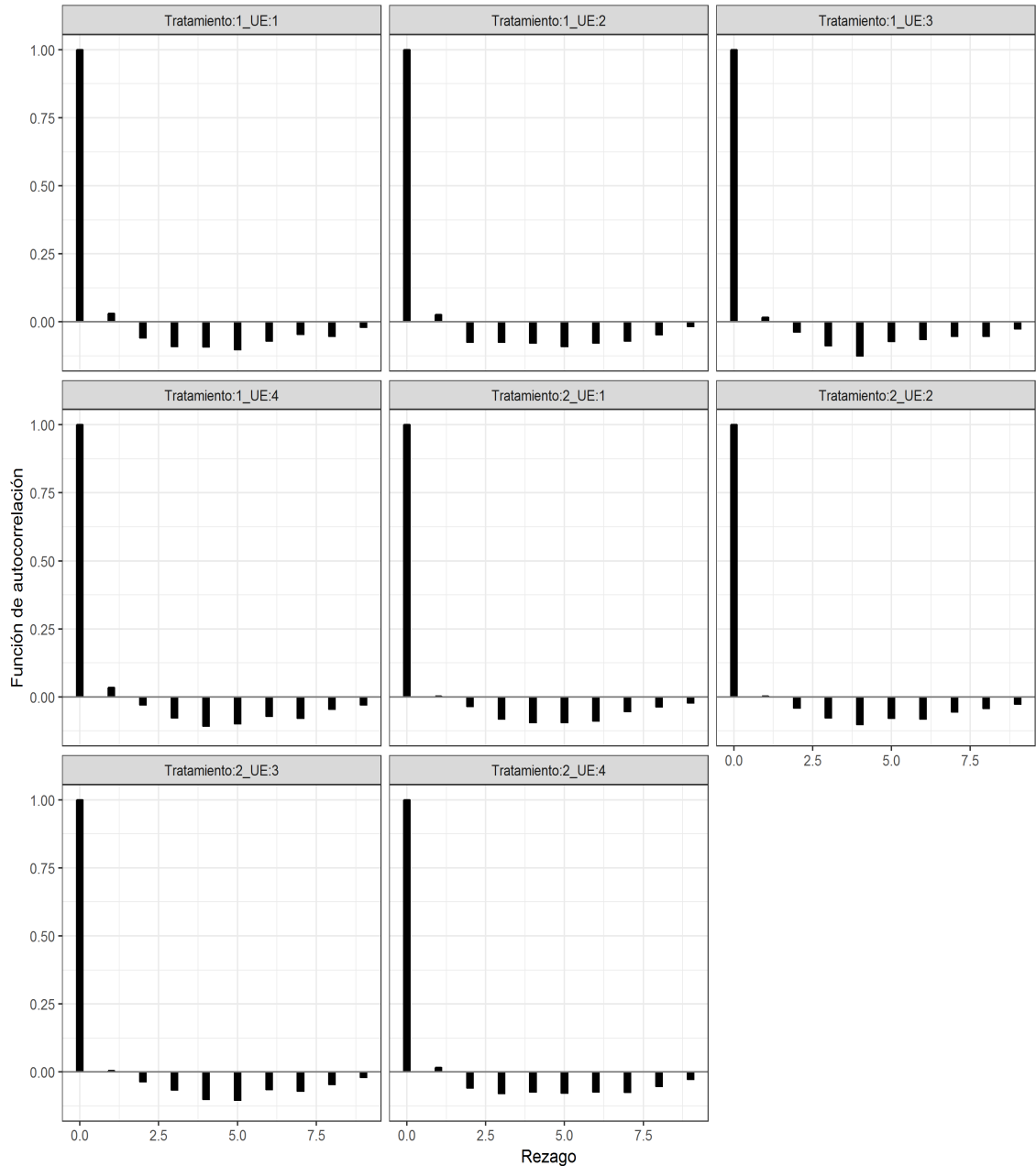


FIGURA 2.10. Correlogramas para las diferentes UE y para los dos tratamientos en consideración en los residuales de uno de los ajustes del modelo dado en la ecuación (2.3).

Con el objetivo de complementar el ejercicio, se realizó el ajuste de los mismos modelos descritos previamente, pero esta vez dejando un porcentaje de la muestra de UO fija, es decir que este porcentaje no será medido en un único tiempo, si no que será considerado en todos los tiempos en consideración. Esto se hace con el fin de evaluar los autocorrelogramas,

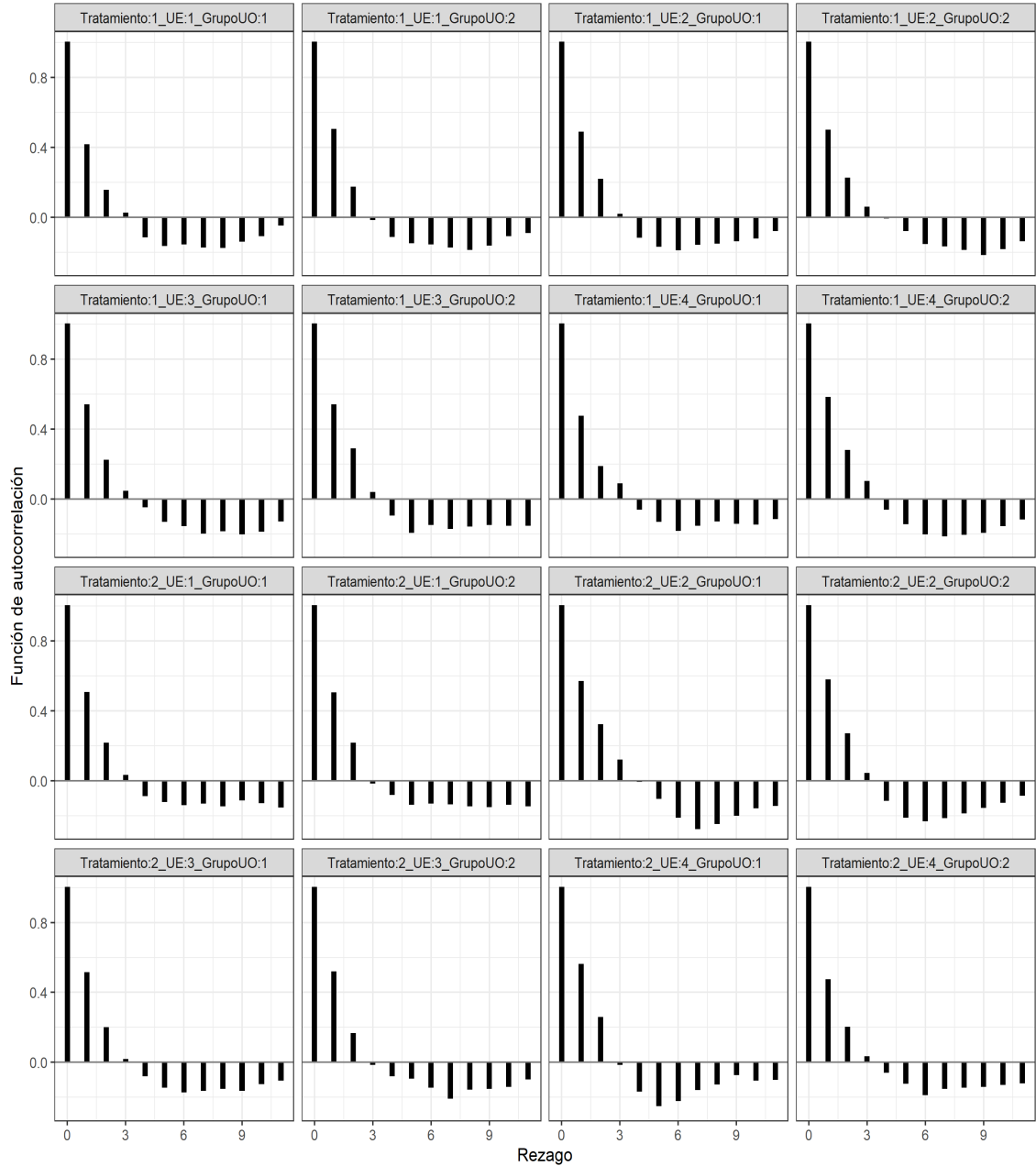


FIGURA 2.11. Correlogramas para los subgrupos de UO dentro de las diferentes UE y para los dos tratamientos en consideración en los residuales de uno de los ajustes del modelo dado en la ecuación (2.4) cuando se tiene un porcentaje de UO fijo.

cuando no se hace una rotación total de toda la muestra. Las figuras 2.11 y 2.12 muestran las funciones de autocorrelación de los residuales de los modelos propuesto y de promedios para los diferentes rezagos. El porcentaje de muestra panel (UO que no se destruyen) en este caso es del 50%. En estas dos figuras se observa como el rezago número uno tiene una autocorrelación importante, por lo cual en este caso se evidencia la presencia de correlación temporal cuando un porcentaje de la muestra es fijo y no rota, a diferencia del caso en que todas las unidades son diferentes en cada tiempo.

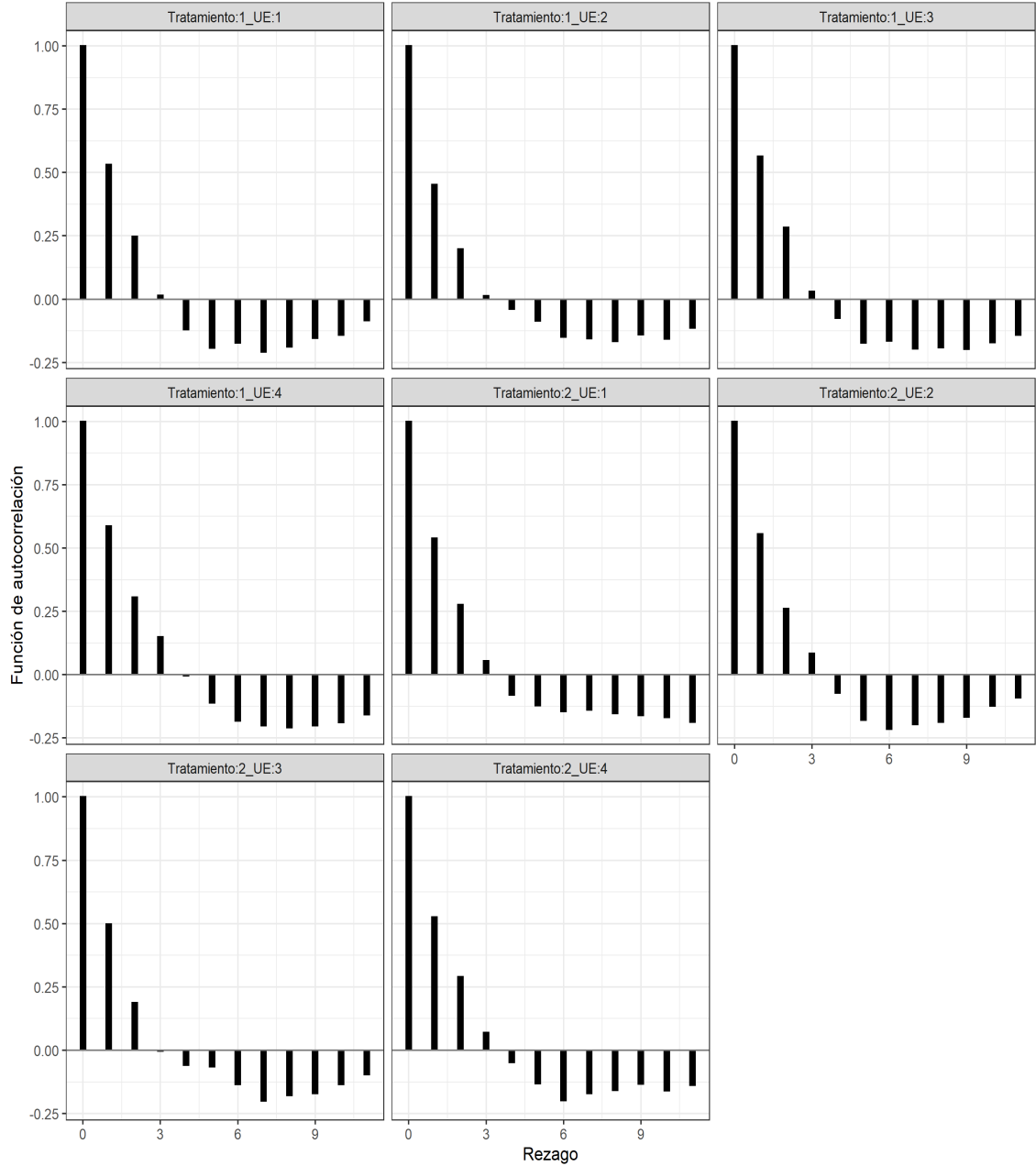


FIGURA 2.12. Correlogramas para las diferentes UE y para los dos tratamientos en consideración en los residuales de uno de los ajustes del modelo dado en la ecuación (2.3) cuando se tiene un porcentaje de UO fijo.

---

---

### Aplicación a datos reales

---

---

Para realizar una aplicación de los modelos considerados en este documento a datos de la vida real, se tomaron los resultados de matemáticas correspondientes a las pruebas ICFES o Saber 11 de colegios ubicados en los diferentes municipios de Colombia, las cuales deben ser presentadas por todos los estudiantes al finalizar el grado undécimo como prerrequisito para su grado. La particularidad de esta información es que la mayoría de alumnos que una vez presentan dicha prueba no la presentan en años posteriores y se puede suponer que a pesar de que los estudiantes de un colegio específico no son los mismos en los diferentes periodos, sus resultados pueden ser similares, ya que tienen una formación similar. Esto se debe a que compartieron la institución educativa, docentes, metodologías de enseñanza e instalaciones, entre otros posibles factores que puedan influir sobre el aprendizaje. Considerando que todas las instituciones son representadas mediante sus estudiantes en todos los periodos de medición, cada colegio representa una UE, mientras que los estudiantes, de los cuales provienen las mediciones son las UO.

Los datos se encontraron en la página de datos abiertos proporcionada por el Ministerio de Tecnologías de la Información y las Comunicaciones (ver MinTic (2020)). Como variable respuesta se tomó el puntaje en matemáticas y como variables independientes en los modelos se tomaron las siguientes:

- área de ubicación de la institución (rural o urbano),
- género de los estudiantes en la institución (femenino, masculino o mixto),
- naturaleza del colegio (oficial o privado) y
- periodo de medición.

Es necesario tener en cuenta que la mayoría de colegios en Colombia inician su año escolar en enero y finaliza en cercanías al mes de diciembre, lo cual indica que éstas escuelas presentan el examen considerado en el segundo semestre del calendario académico. Es por este motivo que únicamente se tomaron observaciones cuyo periodo de medición corresponde al segundo semestre. Los años considerados van desde 2013 hasta 2018.

### 3.1. Análisis descriptivo

Con miras a saber qué interacciones entre factores se deben incluir en los modelos a ajustar se graficaron los perfiles teniendo en cuenta cada pareja de variables independientes. La figura 3.1 muestra los promedios según género de los estudiantes y área de ubicación (gráfico izquierdo), y los promedios por área de ubicación y periodo de medición (gráfico derecho). Allí se observa que los hombres tienen mejor rendimiento en promedio que los colegios mixtos y femeninos, tanto en instituciones ubicadas en zonas urbanas como rurales. Por otra parte los colegios urbanos tienden a tener mejores rendimientos, con respecto a los rurales en todos los periodos.

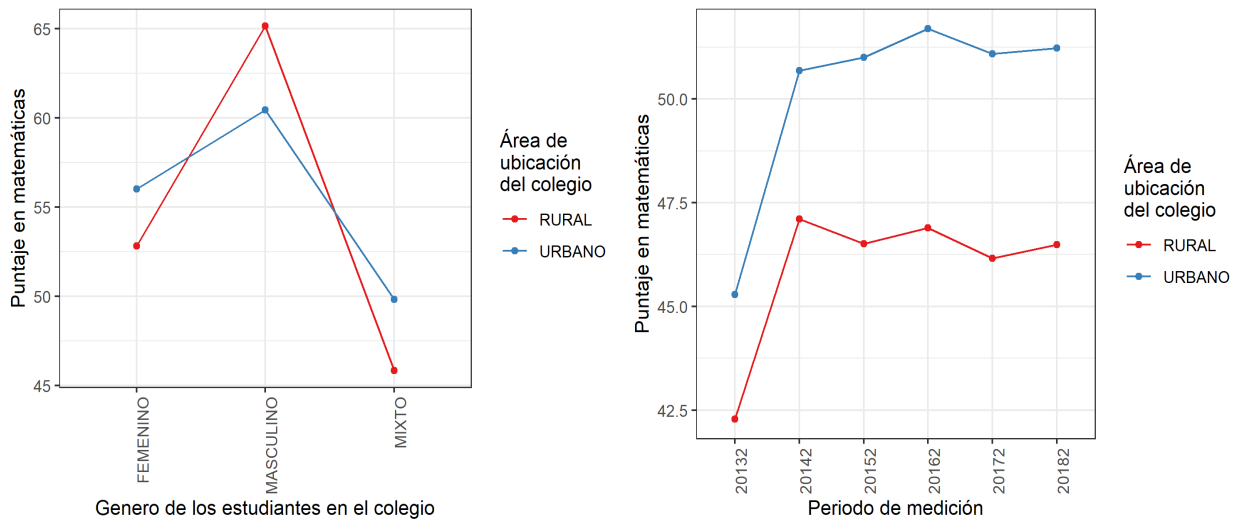


FIGURA 3.1. Puntaje promedio de las pruebas Saber 11 en matemáticas por género y zona de ubicación (gráfico a la izquierda), y por periodo de medición y género (gráfico a la derecha).

La figura 3.2 muestra en su gráfico izquierdo los promedios de puntajes por género de los estudiantes y por área de ubicación, mientras que en el derecho se observan los promedios por periodo de medición y género de los estudiantes. Allí se nota un mejor rendimiento para aquellos colegios que son exclusivamente de género masculino, seguido de aquellos que son exclusivamente de mujeres y por último los mixtos, tanto para instituciones oficiales como no oficiales. Adicionalmente se nota un mejor rendimiento en promedio a partir del año 2014, en comparación con el periodo 2013 para los tres niveles de la variable género.

En la figura 3.3 se observan los promedios por tipo de institución y área de ubicación, y por periodo de medición y naturaleza del colegio en los gráficos de la izquierda y la derecha, respectivamente. Se observa que en los colegios no oficiales se tiene un mayor promedio en colegios rurales que en los urbanos, mientras que en los oficiales ocurre lo contrario. Por último se nota un mejor rendimiento en colegios no oficiales en todos los periodos de medición, además de un aumento considerable en los años 2013-2014.

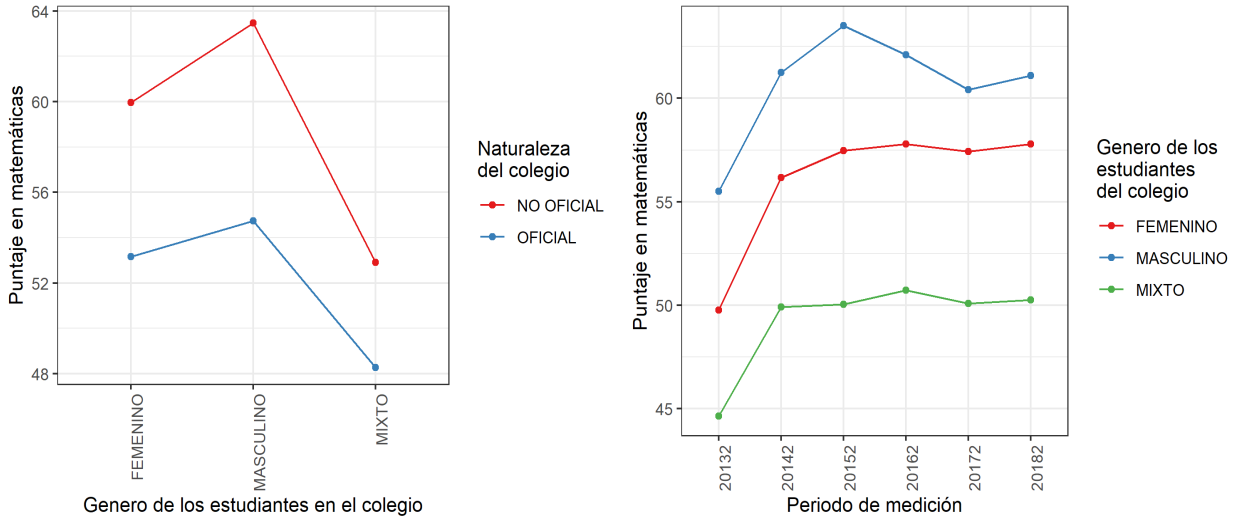


FIGURA 3.2. Puntaje promedio de las pruebas Saber 11 en matemáticas por periodo de medición y por género de los estudiantes del colegio.

En las figuras 3.1, 3.2 y 3.3 se puede visualizar si los perfiles son o no paralelos en búsqueda de interacciones potenciales entre factores para incluirlas en los modelos de regresión. El género y el área de ubicación, además de la naturaleza de la institución y el área son los casos en donde los perfiles se cruzan, motivo por el cual se puede suponer que no son paralelos y se pueden incluir dentro de los modelos de regresión.

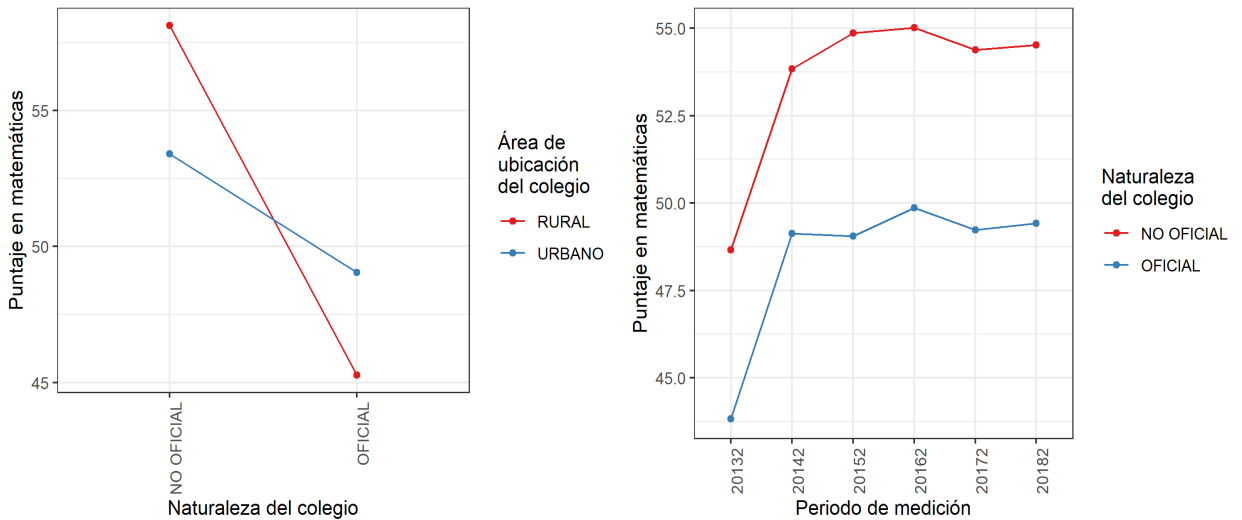


FIGURA 3.3. Puntaje promedio de las pruebas Saber 11 en matemáticas por periodo de medición y naturaleza del colegio.

### 3.2. Aplicación de los modelos estadísticos

De manera análoga al modelo exclusivo de efectos fijos descrito en la sección 2.1.1, el primer modelo considerado es

$$y_{ijklm} = \mu + N_i + G_j + A_k + P_l + NG_{ij} + AG_{jk} + \epsilon_{ijklm}^* \quad (3.1)$$

con  $i = 1, 2, j = 1, 2, 3, k = 1, 2, l = 1, \dots, 6$  y  $m = 1, \dots, M$ . Donde además se suponen las siguientes condiciones de estimabilidad:

$$\sum_{i=1}^2 N_i = \sum_{j=1}^3 G_j = \sum_{k=1}^2 A_k = \sum_{l=1}^6 P_l = \sum_{i=1}^2 NG_{ij} = \sum_{j=1}^3 NG_{ij} = \sum_{j=1}^3 AG_{jk} = \sum_{k=1}^2 AG_{ij} = 0$$

y por último  $\epsilon_{ijklm}^* \stackrel{i.i.d}{\sim} N(0, \sigma_\epsilon^{2*})$ .

En el modelo de la ecuación (3.1)  $y_{ijklm}$  es el puntaje en matemáticas de la  $m$ -ésimo estudiante en el  $l$ -ésimo periodo, en el  $i$ -ésimo tipo (naturaleza del colegio) con el  $j$ -ésimo grupo de la variable de género en la  $k$ -ésima área de ubicación. Por otro lado  $N_i, G_j, A_k$  y  $P_l$  son los efectos del tipo de colegio, del género de estudiantes en la institución, del área de ubicación de la institución y del periodo de medición sobre el puntaje en matemáticas de los estudiantes, respectivamente. Además  $NG_{ij}$  y  $AG_{jk}$  son los efectos sobre la respuesta de las interacciones de las variables naturaleza y área con el género, respectivamente. Por último,  $\epsilon_{ijklm}^*$  es el respectivo error del modelo. Al construir la tabla de análisis de varianza de este modelo, se obtuvo la tabla 3.1, en la cual se observa que todos los factores que hacen referencia a características propias de los planteles educativos, incluyendo las interacciones son significativas bajo el modelo dado en (3.1).

TABLA 3.1. Tabla de análisis de varianza para el modelo de la ecuación (3.1).

Variable	Gl.	SC	CM	Estadística F	Valor $p$
Área de ubicación	1	6572064	6572064	53300.63	< 0.000001
Género	2	6347994	3173997	25741.69	< 0.000001
Naturaleza del colegio	1	9733908	9733908	78943.76	< 0.000001
Periodo	1	6743197	6743197	54688.55	< 0.000001
Género*Naturaleza	2	290592	145296	1178.38	< 0.000001
Género*Área	2	36785	18393	149.17	< 0.000001
Error	3087878	380740935	123		

El modelo de promedios análogo al descrito en la sección 2.1.2 aplicado en este contexto se describe en la ecuación (3.2).

$$\bar{y}_{ijkln} = \mu + N_i + G_j + A_k + NG_{ij} + AG_{jk} + P_l + b_n^* + \bar{\epsilon}_{ijkln}. \quad (3.2)$$

con  $i = 1, 2, j = 1, 2, 3, k = 1, 2, l = 1, \dots, 6$  y  $n = 1, \dots, N$ ,

$$\sum_{i=1}^2 N_i = \sum_{j=1}^3 G_j = \sum_{k=1}^2 A_k = \sum_{l=1}^6 P_l = \sum_{i=1}^2 NG_{ij} = \sum_{j=1}^3 NG_{ij} = \sum_{j=1}^3 AG_{jk} = \sum_{k=1}^2 AG_{ij} = 0$$

y por último  $\bar{\epsilon}_{ijkln} \stackrel{i.i.d}{\sim} N(0, \sigma_\epsilon^2)$  y  $b_n^* \sim N(0, \sigma_b^2)$ , los cuales se asumen independientes entre ellos. En este modelo  $\bar{y}_{ijkln}$  es el puntaje promedio del  $n$ -ésimo colegio en el  $l$ -ésimo periodo, en el  $i$ -ésimo tipo (naturaleza del colegio) con el  $j$ -ésimo grupo de la variable de género en la  $k$ -ésima área de ubicación,  $b_n^*$  es el efecto aleatorio asociado al  $n$ -ésimo establecimiento.

TABLA 3.2. Tabla de análisis de varianza para el modelo de la ecuación (3.2).

Variable	Gl.	Estadística F	Valor $p$
Área de ubicación	1	895.2	< 0.0001
Género	2	243.5	< 0.0001
Naturaleza del colegio	1	1069.8	< 0.0001
Periodo	1	74179.5	< 0.0001
Género*Naturaleza	2	9.9	< 0.0001
Género*Área	2	1.4	0.2549

La tabla de análisis de varianza para el modelo de la ecuación (3.2) se muestra en la tabla 3.2. Allí se observa una disminución considerable en las estadísticas de prueba y en los valores  $p$ , lo cual se refleja por ejemplo en la interacción entre género y área de ubicación la cual no es significativa. Esto se debe a la inclusión de un efecto aleatorio adicional  $b_n^*$ .

Por otro lado, considerando que la UO en este caso son los estudiantes, más no los planteles educativos se puede pensar en la inclusión de un efecto aleatorio con el objetivo de discriminar la variabilidad entre y dentro de los colegios. De esta manera lo que se hace en el modelo propuesto es particionar a los estudiantes por aquellos de buen rendimiento y los de mal rendimiento a partir de la mediana de puntajes en cada colegio, de tal forma que a estos grupos se les puede hacer un seguimiento en el tiempo, con el objetivo de incluir un efecto aleatorio referente a dichos grupos. Por esto la ecuación (3.3) muestra el modelo análogo al descrito en la sección 2.1.3 aplicado en este contexto.

$$y_{ijklnmr} = \mu + N_i + G_j + A_k + P_l + NG_{ij} + AG_{jk} + b_n + \eta_{nm}^* + \epsilon_{ijklnmr} \quad (3.3)$$

con  $i = 1, 2$ ,  $j = 1, 2, 3$ ,  $k = 1, 2$ ,  $l = 1, \dots, 6$  y  $m = 1, 2$ ,

$$\sum_{i=1}^2 N_i = \sum_{j=1}^3 G_j = \sum_{k=1}^2 A_k = \sum_{l=1}^6 P_l = \sum_{i=1}^2 NG_{ij} = \sum_{j=1}^3 NG_{ij} = \sum_{j=1}^3 AG_{jk} = \sum_{k=1}^2 AG_{ij} = 0$$

y por último,  $\epsilon_{ijklmr} \stackrel{i.i.d}{\sim} N(0, \sigma_\epsilon^2)$ ,  $b_n \sim N(0, \sigma_b^2)$  y  $\eta_{nm}^* \sim N(0, \sigma_\eta^2)$ , los cuales se asumen independientes entre ellos.

En el modelo (3.3)  $y_{ijklnmr}$  es el puntaje en matemáticas del  $r$ -ésimo estudiante, del  $m$ -ésimo grupo de estudiantes en el  $n$ -ésimo colegio para el  $l$ -ésimo periodo, en el  $i$ -ésimo tipo (naturaleza del colegio) con el  $j$ -ésimo grupo de la variable de género en la  $k$ -ésima área de ubicación,  $b_n$  y  $\eta_{nm}^*$  son los efectos aleatorios del  $n$ -ésimo establecimiento y del  $m$ -ésimo grupo en el  $n$ -ésimo colegio, respectivamente. El sub-índice  $m$  únicamente toma los valores 1 y 2, ya que las mediciones de cada institución se dividieron en dos grupos

(buen y mal rendimiento) a partir de su mediana, pero a pesar de esto las particiones pueden hacerse de manera que se conformen más grupos, mediante percentiles o cuartiles por ejemplo. La respectiva tabla de análisis de varianza se muestra en la tabla 3.3, en donde se concluye lo mismo que en la tabla 3.2.

TABLA 3.3. Tabla de análisis de varianza para el modelo de la ecuación (3.3).

Variable	Gl.	Estadística F	Valor $p$
Área de ubicación	1	427.4	< 0.0001
Género	2	203.5	< 0.0001
Naturaleza del colegio	1	649.5	< 0.0001
Periodo	1	195259.9	< 0.0001
Género*Naturaleza	2	9.8	< 0.001
Género*Área	2	1.0	0.3865

Por último, para el modelo de análisis de varianza multivariado análogo al de la sección 2.1.4 se tomó el promedio de cada uno de los grupos de estudiantes dentro de las instituciones para cada uno de los periodos, es decir se calcularon los  $\bar{y}_{ijklnm}$ . para conformar los vectores de respuestas, mientras que la ecuación que considera los efectos fijos se mantuvo igual a la de los modelos lineales univariados descritos previamente. Los resultados del Manova utilizando la estadística de Pillai (considerando que las pruebas arrojaron resultados similares se omitieron las otras tres) se muestran en la tabla 3.4.

TABLA 3.4. Tabla de análisis de varianza para el Manova cuya respuesta son los promedios de puntajes en matemáticas en los grupos conformados por estudiantes dentro de los colegios.

Variable	Gl.	Estadística Pillai	Aprox. F	Num. Gl.	Den. Gl.	Valor $p$
Área de ubicación	1	0.06031	185	6	17261	< 0.00001
Género	2	0.02624	38	12	34524	< 0.00001
Naturaleza del colegio	1	0.05518	168	6	17261	< 0.00001
Género*Naturaleza	2	0.00280	4	12	34524	< 0.00001
Género*Área	2	0.00077	1	12	34524	0.3425

Es decir que mediante este último modelo se obtuvieron resultados similares a los obtenidos mediante aquellos modelos que incluyeron efectos mixtos.

Como ejercicio complementario de comparación se presenta la tabla 3.5, en la cual se muestran los cuadrados medios del error, la correlación entre las estimaciones y la respuesta original, y ésta última elevada al cuadrado calculando lo que denominamos un *pseudo* -  $R^2$ , con el fin de obtener medidas de bondad de ajuste. Allí se observa que el modelo de mejor ajuste es el descrito en la ecuación (3.3), que incluye dos efectos aleatorios además de los errores, es decir el modelo propuesto, ya que cuenta con un cuadrado medio del error más pequeño en comparación con los otros y valores mayores en cuanto a la correlación y el *pseudo* -  $R^2$ .

Considerando que se determinó que el mejor modelo es el correspondiente a la ecuación (3.3) se realizaron algunas validaciones sobre los efectos aleatorios y sobre los residuales. Las figuras 3.4, 3.5 y 3.6 muestran las densidades estimadas y los respectivos gráficos de cuantiles para los efectos aleatorios del modelo seleccionado. Con el fin de evaluar

TABLA 3.5. Resumen de los modelos aplicados a los datos del puntaje en matemáticas, en términos del cuadrado medio del error y la correlación entre las estimaciones y los valores observados.

Modelo	$CME$	$r_{y,\hat{y}}$	$pseudo-R^2$
Efectos fijos	123.28	0.27	0.07
Deaton	93.18	0.55	0.30
Propuesto	39.61	0.84	0.70
Manova	88.04	0.38	0.14

el supuesto de normalidad. Gráficamente no se ve ninguna desviación notoria de este supuesto, por lo cual se procedió a realizar las respectivas pruebas de hipótesis, para lo cual se consideró el test de Anderson-Darling para los tres efectos aleatorios. Los respectivos valores  $p$  de estas pruebas se observan en la tabla 3.6. Como los valores  $p$  son mayores a 0.05, para los tres errores se concluye que no hay suficiente evidencia estadística para rechazar la hipótesis de normalidad en los errores aleatorios.

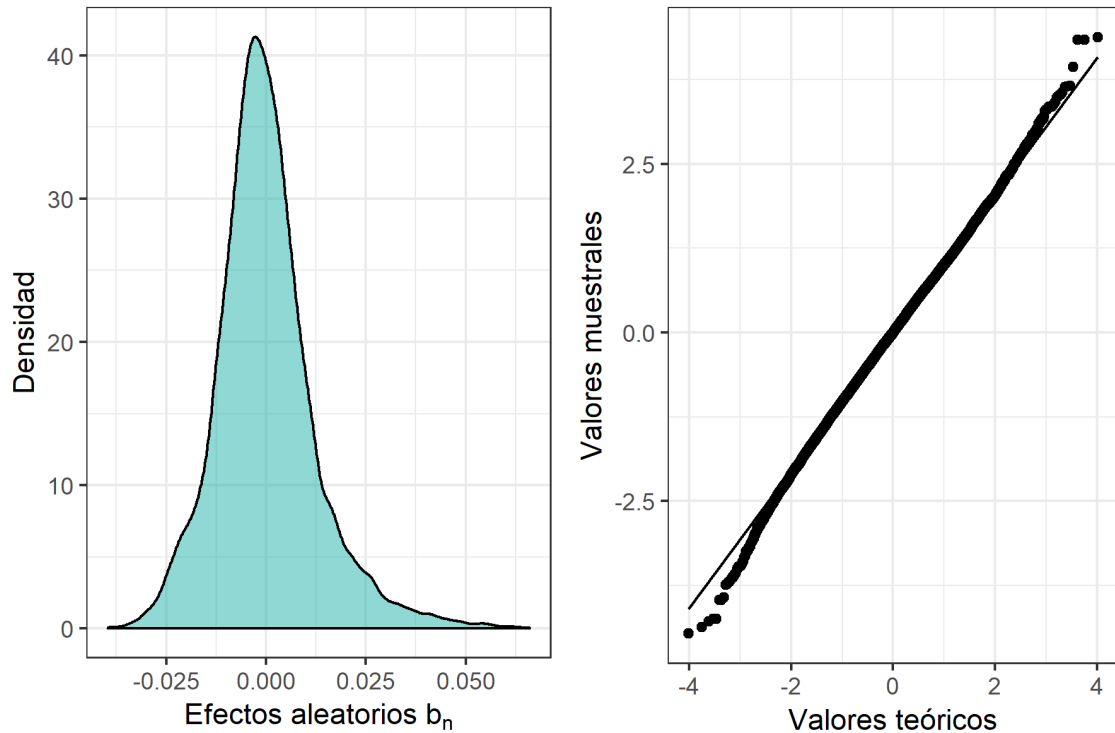


FIGURA 3.4. Densidad estimada para los efectos aleatorios  $b_n$  en el modelo dado en la ecuación (3.3).

TABLA 3.6. Tabla con los valores  $p$  asociados a la prueba de Anderson-Darling para juzgar la hipótesis de normalidad de los efectos aleatorios incluidos en el modelo de la ecuación (3.3).

Test	$\epsilon_{ijklmnr}$	$\eta_{mm}$	$b_n$
Anderson-Darling	0.06	0.13	0.11

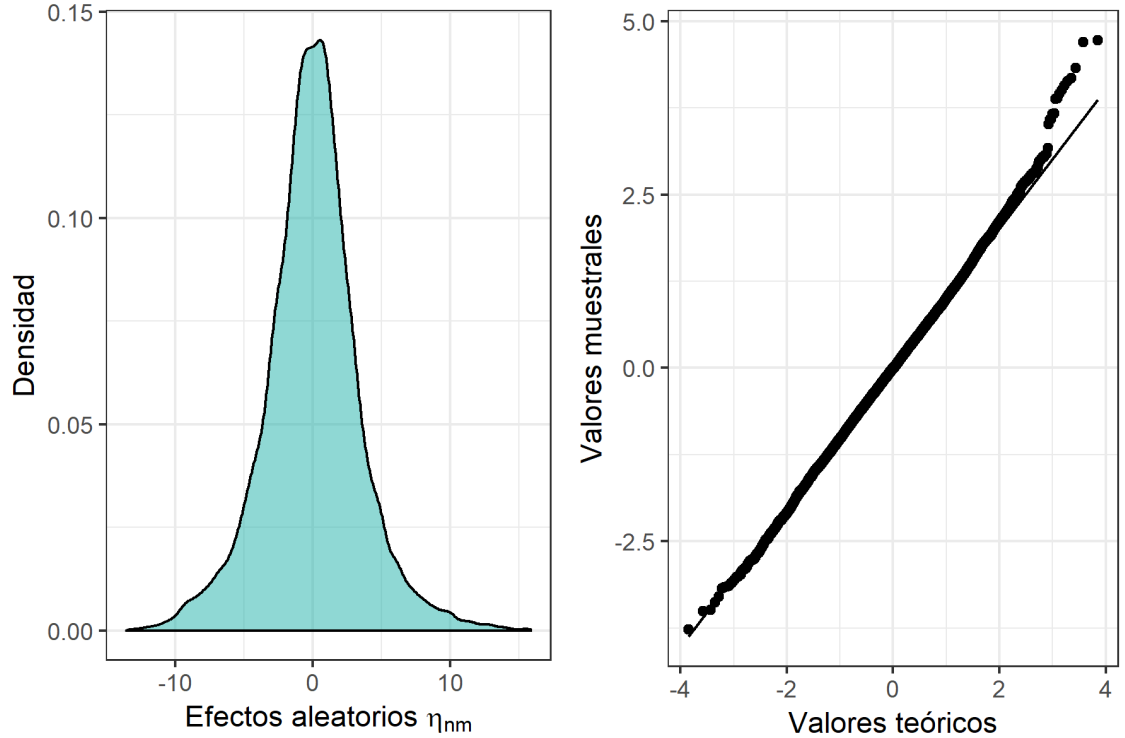


FIGURA 3.5. Densidad estimada para los efectos aleatorios  $\eta_{nm}$  en el modelo dado en la ecuación (3.3).

Por otro lado, para evaluar el supuesto de homocedasticidad en las varianzas de los efectos aleatorios en el modelo propuesto se asignan categorías aleatoriamente a los coeficientes  $b_n$ ,  $\eta_{nm}^*$  y  $\epsilon_{ijklmnr}$ , de tal manera que al comparar los coeficientes a través de estas agrupaciones se puede determinar si se cumple el supuesto de varianzas constantes. La tabla 3.7 muestra los valores  $p$  asociados a la prueba de varianza constante para los diferentes efectos aleatorios incluidos en el modelo propuesto aplicado a los datos reales dado en la ecuación (3.3). En esta tabla se observa que para el caso de  $\epsilon_{ijklmnr}$  se rechaza la hipótesis de homocedasticidad y se concluye que la varianza no es constante, sin embargo en los otros dos factores aleatorios se concluye que no hay evidencia estadística para concluir que la varianza no es constante. Estas conclusiones se realizan utilizando un nivel de significancia de 0.05.

TABLA 3.7. Tabla con los valores  $p$  asociados a la prueba de Bartlett para juzgar la hipótesis de homocedasticidad de los efectos aleatorios incluidos en el modelo de la ecuación (3.3).

Test	$\epsilon_{ijklmnr}$	$\eta_{nm}$	$b_n$
Bartlett	0.02	0.9545	0.5337

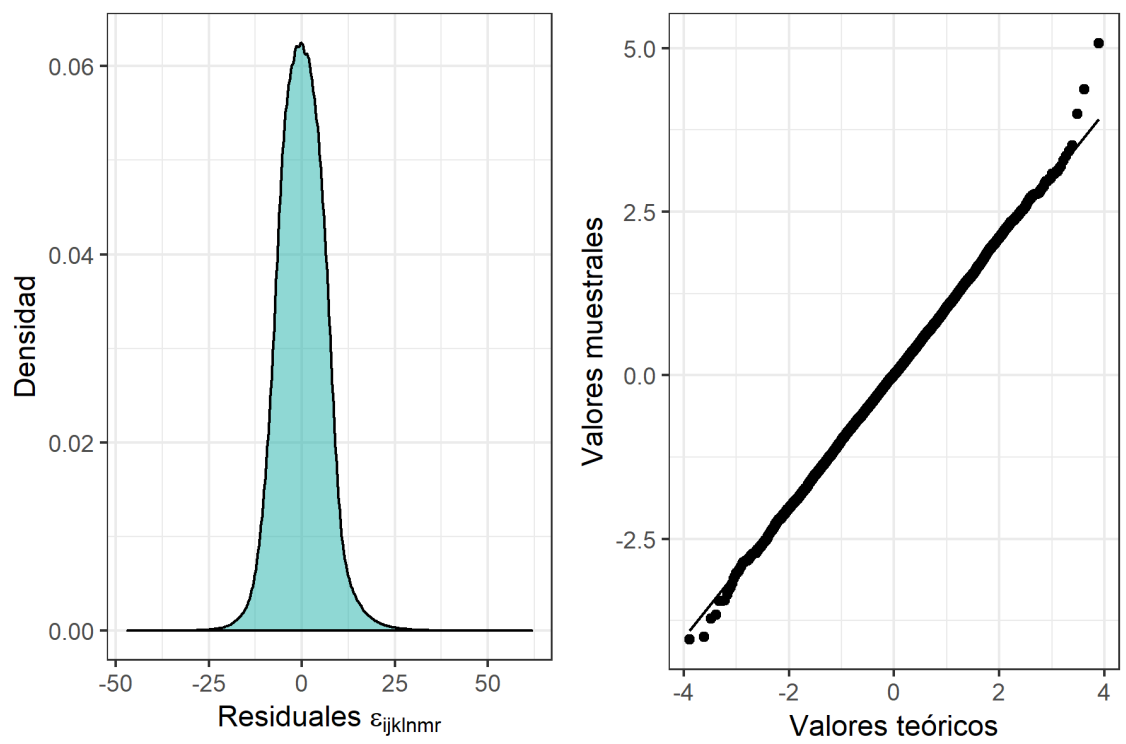


FIGURA 3.6. Densidad estimada para los residuales  $\epsilon_{ijklmnr}$  en el modelo dado en la ecuación (3.3).

---

---

## Conclusiones

---

---

Considerando el esquema de muestreo trabajado en este documento bajo el escenario de datos longitudinales se concluye que el modelo de regresión simple es el menos adecuado, ya que la omisión de factores como los asociados a las diferentes UE conlleva a conclusiones totalmente opuestas. El modelo propuesto que hace agrupaciones de UO dentro de las diferentes UE es el que menor *CME* presenta, con respecto a los otros tres modelos de regresión tenidos en cuenta. Por otra parte, este ajuste en los datos incompletos es el que más se acerca al valor  $p$  calculado para evaluar la significancia de los efectos principales y secundarios, mediante un modelo de regresión considerado sin pérdida de información, seguido de las pruebas desarrolladas con el Manova en el caso de la interacción y del modelo propuesto por Deaton (1985) para evaluar los efectos principales.

De manera adicional, la autocorrelación temporal simulada en los residuales del modelo original no se recupera en las agrupaciones conformadas, lo cual se debe a que no es la misma UO la que se mide en todos los tiempos. Sin embargo, al considerar un porcentaje de la muestra de UO fija, es decir que sea considerada en todos los tiempos, se observa una autocorrelación temporal significativa a partir de los gráficos de autocorrelación.

Con respecto a la aplicación a los datos reales se ve una contribución importante de las variables ubicación de la institución, género de los estudiantes en el colegio, su tipo, periodo de medición y de la interacción entre género y naturaleza. La interacción entre género y área de ubicación es un claro ejemplo de los riesgos de la omisión de factores, ya que en el modelo exclusivo de efectos fijos fue significativa, mientras que en los otros tres no lo fue. En este último ejercicio, al igual que en la simulación el modelo propuesto es el que tiene una mejor bondad de ajuste, lo cual se observó mediante el *CME*, la correlación entre los valores ajustados y los observados y ésta última estadística elevada al cuadrado ejerciendo un papel de un *pseudo* -  $R^2$ .

---

---

## Trabajo futuro

---

---

Como trabajo posterior se puede tener en cuenta la correlación espacial entre mediciones, ya que como en los datos del ICFES, las instituciones educativas pueden estar correlacionadas espacialmente. Es decir, que se puede plantear el mismo ejercicio, pero esta vez evaluando qué sucede con un modelo estadístico que se basa en el muestreo destructivo de UO pertenecientes a UE desde un tiempo inicial, pero esta vez considerando correlación espacio-temporal. Por otro lado, el trabajo realizado en este documento se puede extender al caso donde los tiempos de medición no sean igualmente espaciados.

## APÉNDICE

---

---

### Código de las funciones programadas en R

---

---

```
require("nlme")
require("tidyverse")
require("MASS")
require("mvtnorm")
require("lme4")
require("rlang")
require("RColorBrewer")
require("profileR")

Ejecucion_modelos<-function(df1,coefficients,porcentaje_panel,intercept){
  if(missingArg(intercept)) intercept <- TRUE
  if(missingArg(porcentaje_panel)) porcentaje_panel <- 0
  df2<-df1 %>% mutate(Key=paste0(UExp1))
  list_UE<-df2 %>% split(df2$Key)
  Base_destr_samplng<-map_df(list_UE,~Muestreo_destructivo(.x,porcentaje_panel))

  Base_destr_samplng_new<-Base_destr_samplng %>%
  mutate(key=paste0(UExp1,"_",t)) %>% split(.$key) %>%

  map_df(function(.x){
    .x %>%
    mutate(New_Uobs=cut(.x$respuesta,breaks=c(min(.x$respuesta)-1,
    median(.x$respuesta),
    max(.x$respuesta)+1),labels=paste0("U_",1:2)))
  }) %>%
  dplyr::select(-key)

  Base_averages<-Base_destr_samplng_new %>%
  group_by(Trat.,UExp,UExp1,t,New_Uobs) %>%
  summarise(respuesta=mean(respuesta)) %>% ungroup()
  Base_averages_2<-Base_destr_samplng_new %>%
  group_by(Trat.,UExp,UExp1,t) %>%
  summarise(respuesta=mean(respuesta)) %>% ungroup()

  base_spread<-Base_averages %>% spread(t,respuesta)
  respuestas<-base_spread %>%
```

```

dplyr::select(-Trat., -UExp, -UExp1, -New_Uobs) %>% as.matrix()
Manova_analysis_1<-pbg(data=respuestas, group=base_spread$Trat.,
original.names=TRUE, profile.plot=FALSE)
manova<-lm(respuestas~base_spread$Trat.)
Perfiles<-Base_averages %>% group_by(Trat.,t) %>%
summarise(mean=mean(respuesta)) %>% ungroup()

if(intercept==TRUE){
  if(length(coefficients) >1 ){
    if(porcentaje_panel>0){
      mixed_proposed<-try(lme(respuesta~Trat.+factor(t)+Trat.*factor(t),
random=~1|UExp1/New_Uobs,
data=Base_destr_samplng_new,
control=lmeControl(opt = "optim"),
correlation=corAR1(form=~1|UExp1/New_Uobs)))
mixed_deaton<-try(lme(respuesta~Trat.+factor(t)+Trat.*factor(t),
random=~1|UExp1,data=Base_averages_2,
control=lmeControl(opt = "optim"),correlation=corAR1(form=~1|UExp1)))
    }else{
      mixed_proposed<-try(lme(respuesta~Trat.+factor(t)+Trat.*factor(t)
,
random=~1|UExp1/New_Uobs,
data=Base_destr_samplng_new,
control=lmeControl(opt = "optim")))
      mixed_deaton<-try(lme(respuesta~Trat.+factor(t)+Trat.*factor(t),
random=~1|UExp1,data=Base_averages_2,control=lmeControl(opt = "optim")))
    }

    fixed_lm<-lm(respuesta~Trat.+factor(t)+Trat.*factor(t),
data=Base_averages_2)
  }else{
    if(porcentaje_panel>0){
      mixed_proposed<-try(lme(respuesta~Trat.+factor(t),
random=~1|UExp1/New_Uobs,
data=Base_destr_samplng_new,
control=lmeControl(opt = "optim"),
correlation=corAR1(form=~1|UExp1/New_Uobs)))
      mixed_deaton<-try(lme(respuesta~Trat.+factor(t),
random=~1|UExp1,data=Base_averages_2,
control=lmeControl(opt = "optim"),correlation=corAR1(form=~1|UExp1)))
    }else{
      mixed_proposed<-try(lme(respuesta~Trat.+factor(t),
random=~1|UExp1/New_Uobs,
data=Base_destr_samplng_new,
control=lmeControl(opt = "optim")))
      mixed_deaton<-try(lme(respuesta~Trat.+factor(t),
random=~1|UExp1,data=Base_averages_2,
control=lmeControl(opt = "optim")))
    }

    fixed_lm<-lm(respuesta~Trat.+factor(t),
data=Base_averages_2)
  }
}else{
  if(length(coefficients) >1 ){

```

```

if(porcentaje_panel>0){
  mixed_proposed<-try(lme(respuesta~-1+Trat.+factor(t)+Trat.*factor(t)
,
random=~1|UExp1/New_Uobs,
data=Base_destr_samplng_new,
control=lmeControl(opt = "optim"),
correlation=corAR1(form=~1|UExp1/New_Uobs)))
}else{
  mixed_proposed<-try(lme(respuesta~-1+Trat.+factor(t)+Trat.*factor(t),
random=~1|UExp1/New_Uobs,
data=Base_destr_samplng_new,
control=lmeControl(opt = "optim")))
}
mixed_deaton<-try(lme(respuesta~-1+Trat.+factor(t)+Trat.*factor(t),
random=~1|UExp1,data=Base_averages_2,control=lmeControl(opt = "optim")))
fixed_lm<-lm(respuesta~-1+Trat.+factor(t)+Trat.*factor(t),
data=Base_averages_2)
}else{
  if(porcentaje_panel>0){
    mixed_proposed<-try(lme(respuesta~-1+Trat.+factor(t),
random=~1|UExp1/New_Uobs,
data=Base_destr_samplng_new,
control=lmeControl(opt = "optim"),
correlation=corAR1(form=~1|UExp1/New_Uobs)))
  }else{
    mixed_proposed<-try(lme(respuesta~-1+Trat.+factor(t),
random=~1|UExp1/New_Uobs,
data=Base_destr_samplng_new,
control=lmeControl(opt = "optim")))
  }
  mixed_deaton<-try(lme(respuesta~-1+Trat.+factor(t),
random=~1|UExp1,
data=Base_averages_2,
control=lmeControl(opt = "optim")))
  fixed_lm<-lm(respuesta~-1+Trat.+factor(t),
data=Base_averages_2)
}
}
return(list(Manova_analysis_1=Manova_analysis_1,
Manova_analysis_2=manova,
mixed_proposed=mixed_proposed,
mixed_deaton=mixed_deaton,
fixed_lm=fixed_lm,
Base_destr_samplng=Base_destr_samplng_new
))
}

```

```

Muestreo_destructivo<-function(.x,porcentaje){
  Porc_panel<-porcentaje
  N_panel<-round(length(unique(.x$Uobs))*Porc_panel)
  N_no_panel<-length(unique(.x$Uobs))-round(length(unique(.x$Uobs))*Porc_panel)
  Sample_panel<-.x[1:N_panel]
  filter(.x$Uobs %in% sample(unique(.x$Uobs),
round(length(unique(.x$Uobs))*Porc_panel)))
}

```

```

Sample_no_panel<- .x %>%
filter(!(.x$Uobs %in% Sample_panel$Uobs))
for(tk in unique(Sample_no_panel$t)){

  if(tk==unique(Sample_no_panel$t)[1]){
    Sample_tk<-sample(unique(Sample_no_panel$Uobs),
length(unique(Sample_no_panel$Uobs))/length(unique(Sample_no_panel$t)))
Cumulating_sample<-Sample_no_panel %>%
filter(Uobs %in% Sample_tk,t==unique(Sample_no_panel$t)[1])
Residual_sample<-Sample_no_panel %>% filter(!(Uobs %in% Sample_tk))
  }else{
    Sample_tk<-sample(unique(Residual_sample$Uobs),
length(unique(Sample_no_panel$Uobs))/length(unique(Sample_no_panel$t)))
Cumulating_sample<-Cumulating_sample %>%
rbind(Residual_sample %>%
filter(Uobs %in% Sample_tk,t==tk))
Residual_sample<-Residual_sample %>%
filter(!(Uobs %in% Sample_tk))
  }
}
Final_sample<-rbind(Cumulating_sample,Sample_panel) %>%
as.data.frame()
return(Final_sample)
}

```

```

Simulacion_modelo<-function(n_times,n_uobs_final,
n_uexp,n_trat,coefficients,
variances,correlations,diff,
correlation_type,porcentaje_panel,intercept){
if(missingArg(porcentaje_panel)) porcentaje_panel <- 0
if(missingArg(correlation_type)) correlation_type <- "AR(1)"
if(missingArg(intercept)) intercept <- TRUE

# n_times=10
# n_uobs_final=10
# n_uexp=10
# n_trat=2
# coefficients=c(0.01,1)
# variances=c(3,2,0.3) #
# correlations=c(0,0.9,0.9)
# diff=1
# correlation_type="AR(1)"

##n_times<-n_times Número de tiempos observados en cada unidad observacional
#n_uobs_final<-n_uobs
n_uobs<-n_uobs_final*n_times
UExp<-n_uexp # Número de unidades experimentales por tratamiento

efecto.trat=seq(0,diff,length=n_trat)
coeficientes<-coefficients #c(0.01)

#for(k in 1:nrow(valores_2)){

```

```

df=data.frame(
  Trat.=paste0("Trat",rep(1:n_trat,each=n_times*n_uobs*UExp)),
  Trat=rep(efecto.trat,each=n_times*n_uobs*UExp),
  UExp=paste0("UE" ,rep(rep(1:UExp,each=n_times*n_uobs),n_trat)),
  UExp1=paste0(paste0("Trat",rep(1:n_trat,each=n_times*n_uobs*UExp)),
  "_",
  paste0("UE" ,rep(rep(1:UExp,each=n_times*n_uobs),n_trat)) ),
  Uobs=rep(rep(paste0("U0",1:n_uobs),each=n_times),UExp*n_trat),
  t=rep(seq(2,2*n_times,length=n_times)/(2*n_times),n_uobs*UExp*n_trat) )

variance=variances[3];varUobs=variances[2];varUexp=variances[1]
df<-df %>% arrange(Trat.,UExp1,Uobs,t)
n_uobs<-length(unique(df$Uobs))
UExp<-length(unique(df$UExp))
n_times<-length(unique(df$t))

df1<-df %>% dplyr::select(Trat.,UExp1,Uobs,t) %>% distinct()
df2<-df %>% dplyr::select(Trat.,UExp1,Uobs) %>% distinct()
df3<-df %>% dplyr::select(Trat.,UExp1) %>% distinct()

if(!(correlation_type %in% c("CompSymm","ARMA(1,1)","AR(1)"))){
  stop("This correlation model is not included in the function!!",call.=FALSE)
}else{
  if(correlation_type=="CompSymm"){
    cor_errores <-corCompSymm(correlations[3], form = ~ 1 |UExp1/Uobs)
  }else{
    if(correlation_type=="ARMA(1,1)") {
      cor_errores <-corARMA(correlations[3:4], form = ~ 1|UExp1/Uobs,p=1,q=1)
    }else{
      cor_errores <-corARMA(correlations[3], form = ~ 1|UExp1/Uobs,p=1,q=0)
    }
  }
}

cor_errores. <- Initialize(cor_errores, data = df)
sigmat<-variance*corMatrix(cor_errores.)[[1]]
epsilon=matrix(t(rmvnorm(n=UExp*n_trat*n_uobs,sigma=sigmat)),ncol=1)
df_uobs_times<-data.frame(df1,epsilon)

cor_errores_1<-corCompSymm(correlations[2], form = ~ 1 |UExp1)
cor_errores_1. <- Initialize(cor_errores_1, data = df2)
Sigma_1<-varUobs*corMatrix(cor_errores_1.)[[1]]
Error_obs<-matrix(t(rmvnorm(n=n_trat*UExp,sigma=Sigma_1)),ncol=1)
df_uobs<-data.frame(df2,Error_obs)

cor_errores_2<-corCompSymm(correlations[1], form = ~ 1 )
cor_errores_2. <- Initialize(cor_errores_2, data = df3)
Sigma_2<-varUexp*corMatrix(cor_errores_2.)
Error_exp<-matrix(t(rmvnorm(n=1,sigma=Sigma_2)),ncol=1)
df_uexp<-data.frame(df3,Error_exp)

df<- df %>%
left_join(df_uobs_times) %>%
left_join(df_uobs) %>%

```

```

left_join(df_uexp)

if(intercept==TRUE){
  if(length(coefficients) >1){
    df1<-df %>%
      mutate(respuesta=coefficients[1]*t+
        Trat+coefficients[2]*t*Trat+epsilon+
          (Error_obs+Error_exp),
        betas=coefficients[1]*t+Trat+coefficients[2]*t*Trat)
    mixed_complete<-try(lme(respuesta~Trat.+factor(t)+Trat.*factor(t),
      random=~1|UExp1/Uobs,data=df1,
      control=lmeControl(opt = "optim"),
      correlation=corAR1(form=~1|UExp1/Uobs)))
  }else{
    df1<-df %>%
      mutate(respuesta=coefficients[1]*t+Trat+
        epsilon+(Error_obs+Error_exp),
        betas=coefficients[1]*t+Trat)
    mixed_complete<-try(lme(respuesta~Trat.+factor(t),
      random=~1|UExp1/Uobs,data=df1,
      control=lmeControl(opt = "optim"),
      correlation=corAR1(form=~1|UExp1/Uobs)))
  }
}else{
  if(length(coefficients) >1){
    df1<-df %>%
      mutate(respuesta=coefficients[1]*t+
        Trat+coefficients[2]*t*Trat+epsilon+(Error_obs+Error_exp),
        betas=coefficients[1]*t+Trat+coefficients[2]*t*Trat) #+
    mixed_complete<-try(lme(respuesta~-1+Trat.+factor(t)+
      Trat.*factor(t),
      random=~1|UExp1/Uobs,
      data=df1,control=lmeControl(opt = "optim"),
      correlation=corAR1(form=~1|UExp1/Uobs)))
  }else{
    df1<-df %>%
      mutate(respuesta=coefficients[1]*t+Trat+
        epsilon+(Error_obs+Error_exp),
        betas=coefficients[1]*t+Trat)
    mixed_complete<-try(lme(respuesta~-1+Trat.+factor(t),
      random=~1|UExp1/Uobs,data=df1,
      control=lmeControl(opt = "optim"),
      correlation=corAR1(form=~1|UExp1/Uobs)))
  }
}
}
resultado<-map(1:20,
  ~Ejecucion_modelos(df1,coefficients,
    porcentaje_panel,
    intercept=intercept))
Real_coef<-df1 %>%
dplyr::select(Trat.,t,betas) %>%
  unique()
return(list(mixed_complete=mixed_complete,
  resultado20sim=resultado,Real_coef=Real_coef))
}

```

---

---

## Bibliografía

---

---

- Antman, F. & McKenzie, D. J. (2007), ‘Earnings mobility and measurement error: A pseudo-panel approach’, *Economic Development and Cultural Change* **56**(1), 125–161.
- Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015), ‘Fitting linear mixed-effects models using lme4’, *Journal of Statistical Software* **67**(1), 1–48.
- Canavire-Bacarreza, G. & Robles, M. (2017), ‘Non-parametric analysis of poverty duration using repeated cross section: an application for Peru’, *Applied Economics* **49**(22), 2141–2152.
- Carmona, F. (2005), ‘Modelos lineales’, *Publicación Universidad de Barcelona, Barcelona*.
- Davis, C. S. (2002), *Statistical methods for the analysis of repeated measurements*, Springer Science and Business Media, New York.
- Deaton, A. (1985), ‘Panel data from time series of cross-sections’, *Journal of Econometrics* **30**(1-2), 109–126.
- Faraway, J. J. (2016), *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, Chapman and Hall/CRC, Boca Raton.
- Federer, W. T. & King, F. (2007), *Variations on split plot and split block experiment designs*, Vol. 654, John Wiley and Sons, Hoboken.
- Finch, W. H., Bolin, J. E. & Kelley, K. (2016), *Multilevel modeling using R*, Chapman and Hall/CRC, Boca Raton.
- Gałecki, A. & Burzykowski, T. (2013), *Linear mixed-effects models using R: A step-by-step approach*, Springer Science and Business Media, New York.
- Gardes, F., Duncan, G. J., Gaubert, P., Gurgand, M. & Starzec, C. (2005), ‘Panel and pseudo-panel estimation of cross-sectional and time series elasticities of food consumption: The case of US and polish data’, *Journal of Business and Economic Statistics* **23**(2), 242–253.
- Gentle, J. E. (2012), *Numerical linear algebra for applications in statistics*, Springer Science and Business Media, New York.

- Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. & Hothorn, T. (2016), *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-5.  
**URL:** <http://CRAN.R-project.org/package=mvtnorm>
- Himaz, R. & Aturupane, H. (2016), ‘Returns to education in Sri Lanka: a pseudo-panel approach’, *Education Economics* **24**(3), 300–311.
- Hinkelmann, K. (2011), *Design and analysis of experiments, special designs and applications*, Vol. 3, John Wiley and Sons, Blacksburg.
- Melo, O., López, L. & Melo, S. (2007), ‘Diseño de experimentos: métodos y aplicaciones’, *Editorial Universidad Nacional de Colombia. Bogotá*.
- MinTic (2020), ‘Datos abiertos Colombia’, [urlhttps://www.datos.gov.co](https://www.datos.gov.co). Accedido 01-08-2019.
- Monroy, L. G. D. & Rivera, M. A. M. (2012), *Análisis estadístico de datos multivariados*, Universidad Nacional de Colombia, Bogotá.
- Montgomery, D. C. (2017), *Design and analysis of experiments*, John Wiley and Sons, New York.
- Montgomery, D. C., Peck, E. A. & Vining, G. G. (2012), *Introduction to Linear Regression Analysis*, Vol. 821, John Wiley and Sons, New York.
- Pinheiro, J. & Bates, D. (2006), *Mixed-effects models in S and S-PLUS*, Springer Science & Business Media, New York.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team (2018), *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131.  
**URL:** <https://CRAN.R-project.org/package=nlme>
- Propper, C., Rees, H. & Green, K. (2001), ‘The demand for private medical insurance in the UK: a cohort analysis’, *The Economic Journal* **111**(471), 180–200.
- R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.  
**URL:** <https://www.R-project.org/>
- Rencher, A. C. (2003), *Methods of multivariate analysis*, John Wiley and Sons, Provo.
- Rizzo, M. L. (2007), *Statistical computing with R*, Chapman and Hall/CRC, Bowling Green.
- Schabenberger, O. & Gotway, C. A. (2017), *Statistical methods for spatial data analysis*, Chapman and Hall/CRC, Boca Raton.
- Sprietsma, M. (2012), ‘Computers as pedagogical tools in Brazil: a pseudo-panel analysis’, *Education Economics* **20**(1), 19–32.
- Tovar, A. O., Zulaica, I. G. & Núñez-Antón, V. (2012), ‘Analysis of pseudo-panel data with dependent samples’, *Journal of Applied Statistics* **39**(9), 1921–1937.
- Tsai, C.-H., Mulley, C. & Clifton, G. (2014), ‘A review of pseudo panel data approach in estimating short-run and long-run public transport demand elasticities’, *Transport Reviews* **34**(1), 102–121.

- 
- Urdinola, B. P. & Ospino, C. (2015), ‘Long-term consequences of adolescent fertility: The colombian case’, *Demographic Research* **32**, 1487–1518.
- Verbeek, M. (2008), Pseudo-panels and repeated cross-sections, in ‘The econometrics of panel data’, Springer, pp. 369–383.
- Verbeek, M. & Nijman, T. (1993), ‘Minimum mse estimation of a regression model with fixed effects from a series of cross-sections’, *Journal of Econometrics* **59**(1-2), 125–136.
- Wei, W. W. (2006), Time series analysis, in ‘The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2’.
- West, B. T., Welch, K. B. & Galecki, A. T. (2014), *Linear mixed models: a practical guide using Statistical Software*, CRC Press, Boca Raton.
- Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag, New York.  
**URL:** <http://ggplot2.org>
- Wickham, H., François, R., Henry, L. & Müller, K. (2019), *dplyr: A Grammar of Data Manipulation*. R package version 0.8.3.  
**URL:** <https://CRAN.R-project.org/package=dplyr>
- Wickham, H. & Henry, L. (2018), *tidyr: Easily Tidy Data with ‘spread()’ and ‘gather()’ Functions*. R package version 0.8.1.  
**URL:** <https://CRAN.R-project.org/package=tidyr>