Comparación de algunos R^2 como medidas de bondad de ajuste en modelos lineales mixtos

Diana Guzmán Aguilar

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2012

Comparación de algunos R^2 como medidas de bondad de ajuste en modelos lineales mixtos

Diana Guzmán Aguilar

Tesis o trabajo de grado presentado como requisito parcial para optar al título de:
Magister en Estadística

Director:
Juan Carlos Salazar Uribe, Ph.D University of Kentucky, Profesor Asociado, Universidad Nacional de Colombia

Universidad Nacional de Colombia
Facultad Ciencias, Escuela de Estadística
Medellín, Colombia
2012

Agradecimientos

Infinitas gracias doy a Dios por darme la sabiduría y la paciencia suficiente para lograr una meta más en mi vida.
Agradezco a al profesor Juan Carlos Salazar por haber confiado en mi persona, por la paciencia y por la dirección de este trabajo. Al profesor Juan Carlos Correa por sus comentarios en todo el proceso de elaboración de la Tesis y valiosos aportes.
A Juan Felipe Díaz que me acompaño en esta aventura que significó la maestría. A mi esposo, que desde un principio hasta el día hoy sigue dándome ánimo para terminar este proceso.

Gracias a todos.

Resumen

Los modelos lineales mixtos han recibido gran atención en los últimos años, lo cual ha impulsado su desarrollo hasta convertirlos en una herramienta indispensable para analizar datos longitudinales. La utilidad de estos modelos en aplicaciones como ensayos clínicos y estudios epidemiológicos ha generado gran interés por parte de los investigadores en describir como el modelo se ajusta a los datos observados. En este sentido para evaluar la calidad del ajuste de un modelo específico se han propuesto distintas medidas R^2. En este trabajo se proponen dos estadísticos R^2 para evaluar bondad de ajuste en modelos lineales mixtos con datos longitudinales de medidas repetidas. Se llevó a cabo un estudio de simulación para evaluar y comparar los estadísticos R^2 propuestos junto con otros estadísticos R^2 reportados en la literatura. Finalmente, se aplican los distintos R^2 a datos reales. Se llega a la conclusión que los estadísticos R^2 propuestos tiene una interpretación intuitiva, rango bien definido y son medidas que se pueden interpretar en un sentido absoluto sin hacer referencia a un modelo nulo. Sin embargo, el estudio de simulación evidencia que al igual que los otros estadísticos, los estadísticos R^2 propuestos, en algunas ocasiones son incapaces de discriminar cuando las variables eliminadas del modelo son importantes.

Palabras clave: Modelos lineales mixtos; datos longitudinales; bondad de ajuste; coeficiente de determinación.

Abstract

Linear mixed models have received considerable attention in recent years. This has promoted its development until converting them in an indispensable tool to analyze longitudinal data. The usefulness of these models in applications such as clinical trials and epidemiological studies have generated high interest among researchers for describing the fit of the model. Consequently, to evaluate the quality of the fit in a specific model they have proposed several measures R^2. This paper proposes two R^2 statistics to assess the goodness of fit in linear mixed models with longitudinal data. A simulation study was conducted to evaluate and compare the R^2 statistics proposed together with other R^2 statistics reported in the literature. Finally, the studied R^2 statistics are illustrated using real data. In conclusion, the two proposed R^2 statistics have the following properties: 1) intuitive interpretation, 2) well defined range and 3) they can be interpreted in an absolute scale without doing reference to a null model. However, the simulation study evidences that, as well as other R^2 statistics, the two proposed R^2 statistics sometimes are not able to discriminate when the removed variables of the model are important.

small Keywords: Linear mixed models, longitudinal data, goodness of fit, coefficient of determination.

Contenido

Lista de Tablas

1 Introducción

Los modelos lineales con efectos mixtos son eficientes para analizar datos longitudinales y medidas repetidas, (Liu et. al [8],Vonesh [17]). Una gran cantidad de investigadores ha enfocado sus estudios en la modelación de la varianza, la estimación de los parámetros, las pruebas de significancia y las pruebas de bondad de ajuste. En este último aspecto, se han desarrollado varias medidas tales como el criterio de información de Akaike (AIC), el criterio de información bayesiano (BIC) y la prueba de razón de verosimilitud (LRT), comúnmente utilizadas en modelos lineales mixtos (mlm) y que se encuentran disponibles en la mayoría de los paquetes estadísticos. Sin embargo, su uso se limita a la comparación de modelos con el fin de seleccionar algunos de ellos. A diferencia de los estadísticos R^2 tradicionales que además de utilizarse como criterios de selección ayudan a medir la proporción de la variación explicada de una manera intuitiva.

Como el coeficiente de determinación, R^2, del modelo lineal clásico tiene una forma de interpretación que es simple e intuitiva y además es fácil de calcular, hay un gran interés en desarrollar medidas de este tipo que sean extensiones naturales de los R^2 de los modelos tradicionales. Aunque se han propuesto estadísticos R^2 como medidas de bondad de ajuste que puedan evaluar los modelos, son pocos los que pueden evaluarlos sin hacer referencia a un modelo nulo, es decir, en un sentido absoluto, Liu et. al [8].

Recientemente, Orelien y Edwards [11], reportaron resultados de simulación para evaluar la habilidad de varios estadísticos R^2 propuestos en la literatura en la elección del modelo más parsimonioso[1]. Ellos comparan el valor de los estadísticos R^2 de un modelo completo con los R^2 de otros modelos a los cuales se les han removido covariables de efectos fijos. También comparan los R^2 del modelo completo con los de un modelo sobreajustado. En este estudio, todos los modelos tienen los mismos efectos aleatorios. Ellos encuentran que los R^2 marginales son capaces de seleccionar el modelo más parsimonioso, en tanto que los R^2 condicionales son incapaces de discriminar adecuadamente entre el modelo correcto y uno en el cual se han omitido covariables importantes de efectos fijos.

Como una extensión del trabajo propuesto por Orelien y Edwards [11] en este estudio se realiza un análisis de simulación para evaluar el desempeño de algunos estadísticos R^2 condicionales propuestos en la literatura, aplicados a modelos lineales mixtos, ajustando modelos

[1]Un modelo es parsimonioso si captura un buen porcentaje de información con pocas covariables

en los cuales no se asume que las covariables de efectos aleatorios están bien especificadas. Además, se proponen dos estadísticos R^2 y se evalúa el desempeño de los mismos como medidas de bondad de ajuste en un sentido absoluto, para el caso de los modelos lineales mixtos aplicados a datos longitudinales con respuesta continua, mediante diferentes escenarios de simulación donde se generan conjuntos de datos balanceados y desbalanceados. Estas propuestas junto con el estudio de simulación constituyen los aportes más importante de este trabajo.

Este trabajo se compone de una parte teórica donde se muestran los conceptos a usar a lo largo del estudio y de una parte práctica donde se indica la metodología empleada para la simulación de los conjuntos de datos a ser estudiados. Todas las simulaciones se llevan a cabo usando el software estadístico R [13]. En el capítulo 2, se realiza una revisión de la literatura referente a los estadísticos R^2 propuestos para algunos modelos, en especial en el modelo lineal con efectos mixtos. Además, se presenta la especificación matemática del modelo y algunos conceptos relacionados con él. También se presentan los cinco estadísticos R^2 seleccionados de la literatura para llevar a cabo el estudio de simulación. En el capítulo 3, se presentan los dos estadísticos R^2 propuestos junto con sus propiedades. En el capítulo 4 se presenta el estudio de simulación bajo diferentes escenarios para evaluar el desempeño de los estadísticos como medidas de bondad de ajuste y como criterios de selección de modelos. En el capítulo 5 se estiman los siete estadísticos R^2 para evaluar el ajuste de diferentes modelos ajustados a dos conjuntos de datos: El primer conjunto es un estudio longitudinal que analiza el crecimiento de 27 niños de 8 a 14 años de edad. El segundo conjunto de datos analiza el secado al aire de unos tablones ciprés en función del tiempo y su grosor. Se pretende encontrar los modelos que mejor se ajusten a estos dos conjuntos de datos utilizado los diferentes estadísticos R^2 como medidas de bondad de ajuste y criterios de selección para los predictores. Finalmente, en el capítulo 6 se exponen las conclusiones derivadas de los resultados obtenidos a través de la investigación. También se ponen algunas recomendaciones para estudios futuros.

2 Marco teórico

En este capítulo se realiza una revisión de la literatura donde se presentan avances recientes relacionados con los estadísticos R^2 dentro del contexto de modelos lineales mixtos y se muestran los conceptos generales para contextualizar el desarrollo de este estudio.

2.1. Estado del arte

En los últimos años, para diversos modelos estadísticos, se han propuesto diferentes tipos de estadísticos R^2 como medidas de bondad de ajuste. A continuación se hace una revisión de algunos de estos trabajos.

Magee [9], propone dos métodos para generar medidas de R^2 para una amplia clase de modelos. Esas medidas están conectadas a los modelos de regresión lineal estándar a través de los estadísticos Wald y razón de verosimilitud que se utilizan para chequear la significancia conjunta de las variables explicativas. El autor muestra algunos usos corrientes de la utilización del R^2 para casos especiales de esos modelos.

Nagelkerke [10], discute una generalización del coeficiente de correlación R^2 para modelos de regresión en general. Él propone una definición concerniente a modelos con respuesta discreta.

Cameron y Windmeijer [2], propusieron un estadístico R^2 como medida de bondad de ajuste para los modelos de regresión de la familia exponencial. Estos incluyen transformaciones logit, probit, Poisson, geométrica, gamma y exponencial. Este R^2 está definido como la reducción proporcional en certeza, medida por la divergencia de Kullback-Leibler, debida a la inclusión de regresores.

Cameron y Windmeijer [3], propusieron varias medidas R^2 basadas en varias definiciones de residuales para el modelo de regresión Poisson básico y para modelos más generales tales como el binomial negativo que acomoda datos sobredispersos. La medida R^2 preferida está basada en la desviación residual.

Vonesh et al. [17], presentaron un estadístico de bondad de ajuste en regresión no lineal que puede ser utilizado de manera similar al criterio R^2 en regresión lineal, para la evaluación de la estructura de covarianza y de la media supuesta. Además, introducen una aproximación

al pseudo-test de razón de verosimilitud, para chequear la hipótesis de la estructura de covarianza.

Xu [18], generaliza la bien conocida medida R^2 para modelos de regresión lineal con efectos mixtos. Él cuantifica la variabilidad en la variable respuesta que es explicada por las covariables bajo el modelo lineal mixto y estima tres tipos de medida para cuantificar tales cantidades. El primer tipo de medida hace uso directo de la varianza estimada; el segundo tipo de medida usa la suma de cuadrados de los residuales en analogía a la regresión lineal; y el tercer tipo de medida está basado en la ganancia de información de Kullback-Leibler. Xu estudia el rendimiento de las medidas a través de simulaciones de Monte Carlo e ilustra su utilidad con un conjunto de datos de un experimento clínico para el tratamiento de la esquizofrenia.

Orelien y Edwards [11], reportan resultados de simulación para evaluar la habilidad de varios estadísticos R^2 en la elección del modelo más parsimonioso. Ellos comparan el valor de los estadísticos R^2 de un modelo completo con los R^2 de otros modelos a los cuales se les han removido covariables de efectos fijos. También comparan los R^2 del modelo completo con los de un modelo sobreajustado. En este estudio, todos los modelos tienen los mismos efectos aleatorios. Ellos encuentran que los R^2 marginales son capaces de seleccionar el modelo más parsimonioso, en tanto que los R^2 condicionales son incapaces de discriminar adecuadamente entre el modelo correcto y uno en el cual se han omitido covariables importantes de efectos fijos.

Edwards et al [4], definen y describen el modo de calcular un estadístico R^2 para el modelo lineal mixto mediante el uso de un solo modelo. El estadístico propuesto mide la asociación multivariada entre la variable respuesta y los efectos fijos en el modelo mixto. El estadístico surge como una función uno a uno de un estadístico F apropiado para chequear todos los efectos fijos, excepto el intercepto, pero manteniendo la misma estructura de covarianza. Además, el estadístico conduce inmediatamente a una definición natural de un R^2 parcial.

Liu, Zheng y Shen [8], proponen tres coeficientes de determinación R^2 como medidas de bondad de ajuste para el modelo lineal con efectos mixtos aplicados a datos longitudinales con medidas repetidas. Ellos presentan teoremas que describen las propiedades de los tres R^2 y las relaciones entre ellos. Para comparar los R^2 junto con otros criterios de la literatura llevaron a cabo un estudio completo de simulación. Por último, aplican cada uno de los R^2 a datos reales de respuesta virológica de pacientes con VIH.

2.2. Datos longitudinales

Uno de los objetivos más interesantes de la estadística aplicada es poder construir modelos que expliquen las relaciones entre los datos obtenidos derivados de la práctica estadística. En el contexto clásico, una de las hipótesis planteadas, a la hora de modelar, es la independencia entre observaciones. Sin embargo, muchos estudios tienen diseños que implican datos correlacionados. Un caso particular de datos correlacionados son los diseños longitudinales, que se obtienen cuando la misma característica es medida repetidamente en el tiempo. El análisis longitudinal presenta importantes ventajas respecto a otros diseños, una de las más importantes es que permite distinguir variaciones en el tiempo de la característica que se mide en un individuo (o variación intra-individual) junto con las variaciones debidas a las diferencias entre los individuos (o variación inter-individual).

Los métodos para variable respuesta continua bajo suposiciones de normalidad son los que han tenido un mayor desarrollo. En concreto, el modelo lineal mixto (Laird y Ware[6]) ha jugado un papel importante en extender el modelo lineal general para tratar con datos continuos correlacionados. Debido a las propiedades de la distribución normal multivariada, su teoría e implementación se simplifica bastante. Programas de software tales como el R han desarrollado librerías como el *nlme* que son ampliamente utilizados para ajustar esta clase de modelos y han facilitado la difusión de esta metodología.

2.3. EL modelo lineal mixto (MLM)

Los modelos lineales mixtos han generado un creciente interés en la literatura estadística en los últimos años, debido a que representan una herramienta rica y poderosa para el análisis de datos de medidas repetidas con datos balanceados o desbalanceados entre otras aplicaciones. Dichos modelos son usados para describir la relación entre una variable respuesta y una o más covariables en datos con medidas repetidas de acuerdo a uno o más factores de clasificación. Al tratar de modelar datos con medidas repetidas hay que considerar que aparecen dos fuentes de variabilidad, una entre las medidas del mismo sujeto (variación intraindidividual), otra es la variación aleatoria entre los sujetos (variación interindividual). El modelo propuesto por Laird y Ware [6] reconoce ambas fuentes de variabilidad y se puede definir en dos etapas.

ETAPA 1 (Variación intra-individual)

Se supone que se tienen m sujetos y que cada sujeto i ha sido observado n_i veces, por tanto, se tienen disponibles $N{=}\sum_{i=1}^{m} n_i$ observaciones en total. Sea $\boldsymbol{Y}_i$ el vector de respuestas $n_i \times 1$ del $i-$*ésimo* sujeto. Por tanto se satisface que,

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i. \tag{2-1}$$

Donde $\mathbf{b}_i$ es el efecto aleatorio del $i-ésimo$ individuo, $\boldsymbol{\beta}$ es el vector de parámetros $p \times 1$ correspondiente a los efectos fijos, $\mathbf{X}_i$ es la matriz de diseño $n_i \times p$ específica para el $i-ésimo$ sujeto, $\mathbf{Z}_i$ es una matriz de diseño $n_i \times k$ y $\boldsymbol{e}_i$ es el vector de errores intra-individual. Se asume que $\mathbf{e}_i \sim N(0, \mathbf{R}_i)$ donde $\mathbf{R}_i$ es la matriz de covarianzas intra-individual de dimensión $n_i \times n_i$. Con los $\mathbf{b}_i$ y los $\mathbf{e}_i$ estadísticamente independientes.

A partir del modelo (2-1) se tiene que,

$$\begin{aligned} E(\mathbf{Y}_i \mid \mathbf{b}_i) &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i \\ Cov(\mathbf{Y}_i \mid \mathbf{b}_i) &= \mathbf{R}_i. \end{aligned}$$

ETAPA 2 (Variación inter-individual)

Se supone que el vector de efectos aleatorios $\mathbf{b}_i$ proviene de una distribución normal con media cero y matriz de dispersión $\mathbf{D}_{k\times k}$; además se asume que los $\mathbf{b}_i$, $i = 1, 2, ..., m$ son independientes entre sí y del vector de errores $\mathbf{e}_i$. Con estos supuestos se tiene que,

$$\begin{aligned} E[\mathbf{Y}_i] &= E[E(\mathbf{Y}_i \mid \mathbf{b}_i)] = \mathbf{X}_i\boldsymbol{\beta} \\ Cov[\mathbf{Y}_i] &= E[Cov(\mathbf{Y}_i \mid \mathbf{b_i})] + Cov[E(\mathbf{Y}_i \mid \mathbf{b}_i)] \\ &= \mathbf{R}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z'}_i \\ &= \mathbf{V}_i. \end{aligned}$$

La forma de $\mathbf{V}_i$ implica que el modelo tiene dos fuentes de variabilidad: La primera se refiere a la variación dentro de los individuos representada por la matriz $\mathbf{R}_i$ y la segunda a la variación entre los individuos representada por la matriz $\mathbf{D}$.

El modelo (2-1) junto con los supuestos del vector de errores $\mathbf{e}_i$ y el vector de efectos aleatorios $\mathbf{b}_i$ implica que los vectores tienen distribuciones normales multivariadas de dimensión n_i; para precisar, para cada $i = 1, 2, ..., m$ se tiene que $\mathbf{Y}_i|\mathbf{b}_i \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \mathbf{R}_i)$ y por tanto $\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i)$.

A veces es más conveniente escribir el modelo para todos los individuos en un vector,

$$\begin{cases} \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}b + \mathbf{e} \\ var[Y] = \mathbf{R} + \mathbf{Z}\dot{\mathbf{D}}\mathbf{Z}' = \mathbf{V}_{N\times N} = diag(\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_m). \end{cases} \tag{2-2}$$

Donde,

$\mathbf{Y} = [\mathbf{Y}'_1, \mathbf{Y}'_2, \ldots, \mathbf{Y}'_m]'$,$\mathbf{X} = [\mathbf{X}'_1, \mathbf{X}'_2, \ldots, \mathbf{X}'_m]'$,$\mathbf{b} = [\mathbf{b}'_1, \mathbf{b}'_2, \ldots, \mathbf{b}'_m]'$,
$\mathbf{e} = [\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_m,]'$, $\mathbf{Z} = diag(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_m)_{(N\times km)}$, $\mathbf{R} = diag(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_m)_{(N\times N)}$,
$\dot{\mathbf{D}} = diag(\mathbf{D}, \mathbf{D}, \ldots, \mathbf{D})_{(km\times km)}$

Estimación de los efectos y los parámetros de covarianza

Para ajustar un modelo lineal mixto se deben estimar los parámetros que caracterizan la estructura de la media, $\boldsymbol{\beta}$, y la estructura de la varianza, descritas por las matrices $\mathbf{D}$ y $\mathbf{R}$. Bajo el supuesto de normalidad multivariada es frecuente usar los métodos de máxima verosimilitud, (**ML**) y máxima verosimilitud restringida, (**REML**), para estimar los efectos fijos especificados en el modelo.

Máxima verosimilitud (MLE)

El método de máxima verosimilitud es el siguiente. Si se quieren estimar los parámetros que caracterizan el modelo con base en los datos disponibles, un enfoque sería utilizar como estimadores de los parámetros los valores que maximizan la probabilidad de que la muestra obtenida ocurra. La función de verosimilitud es la función de densidad conjunta de los datos pero mirada como función de los parámetros.

Debido al hecho que las observaciones para cada individuo están correlacionadas, la función de verosimilitud para un individuo se obtiene escribiendo primero la densidad de probabilidad conjunta de las observaciones del individuo en forma matricial. Luego la densidad de probabilidad conjunta de todos los individuos es el producto de las densidades de probabilidad individual, esto considerando que las observaciones de diferentes individuos son independientes. Bajo el supuesto de normalidad e independencia del error y del efecto aleatorio, $\mathbf{b}_i$ y $\mathbf{e}_i$, la función de verosimilitud tiene la forma de una distribución conjunta normal multivariada, así, la función de probabilidad resultante para todos los m individuos de acuerdo con la ecuación (2-1), es la siguiente:

$$L = \prod_{i=1}^{m} (2\pi)^{-(\frac{n_i}{2})} \mid \mathbf{V}_i \mid^{-\frac{1}{2}} \exp\{\frac{-(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})}{2}\} \tag{2-3}$$

Equivalentemente, la log-verosimilitud es:

$$\ell = -\frac{1}{2}\left[\sum_{i=1}^{m} n_i log(2\pi) + log|\mathbf{V}_i| + (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}_i^{-1}(\mathbf{Y}_i - X_i\boldsymbol{\beta})\right] \tag{2-4}$$

con $\mathbf{V}_i = \mathbf{R}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'$, donde los parámetros de covarianza intra-individual están representados por $\mathbf{R}_i$, y los distintos elementos de la covarianza entre los individuos por la matriz $\mathbf{D}$. Si $\boldsymbol{\alpha}$ denota el vector que contiene los parámetros de varianza y covarianza, usualmente llamados componentes de varianza, encontrados en $\mathbf{V}_i = \mathbf{R}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'$, entonces para obtener los estimadores primero se asume que $\boldsymbol{\alpha}$ es conocido y se maximiza ℓ con respecto $\boldsymbol{\beta}$, así se obtiene el estimador de máxima verosimilitud **MLE** de $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha})$ denotado de esta manera porque $\hat{\boldsymbol{\beta}}$ es función de las componentes de $\boldsymbol{\alpha}$, entonces

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = [\sum_{i=1}^{m} X_i' \mathbf{V}_i^{-1} X_i]^{-1} [\sum_{i=1}^{m} X_i' \mathbf{V}_i^{-1} \mathbf{Y}_i]. \tag{2-5}$$

Luego el MLE de $\boldsymbol{\alpha}$ se obtiene maximizando ℓ en función de $\boldsymbol{\alpha}$ después de reemplazar $\boldsymbol{\beta}$ por $\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha})$ en (2-4).

Máxima verosimilitud restringida

Thompson [14] introdujo la idea de máxima verosimilitud restringida, REML, con el objeto de obtener estimadores para los componentes de varianza insesgados o menos sesgados que los estimadores de máxima verosimilitud.
Este procedimiento cobra importancia cuando el número de parámetros de efectos fijos es relativamente grande con respecto al número total de observaciones, pues en este caso los estimadores obtenidos por MLE de los componentes de la varianza resultan subestimados, es decir, los parámetros de varianza son demasiado pequeños. Verbeke y Molenberghs [15] consideran que lo anterior se debe a la pérdida de información que ocurre al estimar primero (los componentes de la media), pero que aparentemente, se pueden obtener directamente estimadores insesgados para los componentes de la varianza que se basan en un procedimiento estadístico que no requiere estimación inicial de los componentes de la media.

La estimación por REML se describe en el siguiente procedimiento: Primero se combinan los m modelos de regresión de todos los individuos en un solo modelo $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}$ donde los vectores $\mathbf{Y}$, $\mathbf{b}$, $\mathbf{e}$ y la matrix $\mathbf{X}$ se obtienen superponiendo los vectores $\mathbf{Y}_i$, $\mathbf{b}_i$,$\mathbf{e}_i$ y las matrices $\mathbf{X}_i$ respectivamente una sobre la otra, y donde $\mathbf{Z}$ es una matriz con bloques en la diagonal principal cuyos bloques son las matrices $\mathbf{Z}_i$ y ceros en los otros lugares. La dimensión de $\mathbf{Y}$ es igual a $N \times 1$ donde $N = \sum_{i=1}^{m} n_i$.

La distribución marginal de $\mathbf{Y}$ es normal con media $\mathbf{X}\boldsymbol{\beta}$ y con matriz de covarianza $\mathbf{V}(\alpha)$ que es igual a una matriz con bloques en la diagonal principal cuyos bloques son $\mathbf{V}_i$ y ceros en los otros lugares.

Los estimadores para los componentes de la varianza por máxima verosimilitud restringida se obtienen maximizando la función de verosimilitud de una combinación lineal de los elementos de la variable respuesta $\mathbf{Y}$, estas combinaciones lineales se conocen como conjunto de contrastes de errores, Verbeke [15] que están dados por

$$\mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{U} = \mathbf{A}'\mathbf{Y},$$

con $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.Donde $\mathbf{A}$ es cualquier matriz de rango completo de tamaño $N \times (N-p)$ con columnas ortogonales a las columnas de $\mathbf{X}$.

Por lo anterior, el vector $\mathbf{U}$ sigue una distribución normal que tiene por media un vector de ceros y por varianza la matriz $\mathbf{A}'\mathbf{V}\mathbf{A}$, además, la distribución de $\mathbf{U}$ no depende de $\boldsymbol{\beta}$, por tanto, las inferencias sobre los componentes de la varianza basadas en la distribución del vector $\mathbf{U}$ no pierden información en la estimación de los componentes de la media contenidos en $\boldsymbol{\beta}$ como si ocurre cuando las inferencias sobre los componentes de la varianza se basan en la distribución del vector $\mathbf{Y}$.

La función de log-verosimilitud de $\mathbf{U}$ o de los contrastes de errores se puede escribir como:

$$\mathrm{L}(\boldsymbol{\alpha})=(2\pi)^{-\frac{n-p}{2}} \left|\sum_{i=1}^{m} \mathbf{X}_i'\mathbf{X}_i\right|^{\frac{1}{2}} \left|\sum_{i=1}^{m} \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}\sum_{i=1}^{m}\left(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)' \mathbf{V}_i^{-1}\left(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)\right]$$

Donde $\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{m} \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{Y}_i\right)$ y $\boldsymbol{\alpha}$ denota un vector que contiene todos los componentes de la varianza.

Los estimadores de los componentes de la varianza, $\hat{\boldsymbol{\alpha}}$, obtenidos por máxima verosimilitud restringida no dependen de la escogencia de la matriz $\mathbf{A}$ (es decir, de los contrastes de error utilizados, o de la combinación lineal utilizada sobre el vector de variables respuesta Y), Verbeke [15].

Estimador empírico de Bayes

Además del interés en la estimación de los parámetros fijos del modelo, que incluye las componentes de la varianza, a menudo es también de interés la predicción de los efectos aleatorios $\mathbf{b}_i$. Una forma natural para estimar los efectos aleatorios está basado en técnicas bayesianas, dado que en el modelo (2-1) considera los efectos $\mathbf{b}_i$ como variables aleatorias. Partiendo del hecho que la distribución del vector $\mathbf{Y}_i$ del individuo i condicionado por el coeficiente $\mathbf{b}_i$ es normal multivariada con vector de medias $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$ y con matriz de covarianza $\mathbf{R}_i$. Además, la distribución marginal de $\mathbf{b}_i$ es normal multivariada con vector de medias $\mathbf{0}$ y matriz de covarianza $\mathbf{D}$. En la literatura bayesiana, esta última distribución es usualmente llamada distribución a priori de los parámetros $\mathbf{b}_i$, pues no dependen de la distribución de los datos $\mathbf{Y}_i$. Una vez recolectados los valores $\mathbf{y}_i$ del vector aleatorio $\mathbf{Y}_i$, la llamada distribución a posteriori de $\mathbf{b}_i$, definida como la distribución de $\mathbf{b}_i$ dado que $\mathbf{Y}_i = \mathbf{y}_i$, que se denota por $f(\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i)$ y se puede calcular. Si se denota la función de densidad de $\mathbf{Y}_i$ dado $\mathbf{b}_i$ y la función de densidad a priori de $\mathbf{b}_i$ por $f(\mathbf{y}_i|\mathbf{b}_i)$ y $f(\mathbf{b}_i)$ respectivamente, se tiene que la función de densidad a posteriori de $\mathbf{b}_i$ está dada por:

$$f(\mathbf{b}_i|\mathbf{y}_i) = f(\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i) = \frac{f(y_i|\mathbf{b}_i)f(\mathbf{b}_i)}{\int f(y_i|\mathbf{b}_i)f(\mathbf{b}_i)db_i}.$$

Usando la teoría bayesiana de modelos lineales generales se puede demostrar que la función de densidad anterior es normal multivariada. Es muy usual estimar $\mathbf{b}_i$ por su media en la distribución a posteriori, llamada media a posteriori de $\mathbf{b}_i$. Este estimador está dado por

$$\hat{\boldsymbol{b}}_{\boldsymbol{i}} = E[\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i] = \int_{-\infty}^{\infty} \mathbf{b}_i f(b_i|\mathbf{y}_i)db_i = \mathbf{D}\mathbf{Z}_i'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}).$$

Y la matriz de covarianza correspondiente a dicho estimador es igual a

$$V[\hat{\boldsymbol{b}}_{\boldsymbol{i}}] = \mathbf{D}\mathbf{Z}_i'[\mathbf{V}_i^{-1} - \mathbf{V}_i^{-1}\mathbf{X}_i(\textstyle\sum_{i=1}^{m} \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{V}_i^{-1}]\mathbf{Z}_i\mathbf{D},$$

Verbeke [15]

2.4. El estadístico R^2

Los R^2 como medidas de bondad de ajuste son muy populares. Son números que se pueden obtener a partir de un modelo ajustado. Tienen una interpretación natural como la proporción de la varianza explicada de la variable respuesta por el modelo, está entre 0 y 1, es adimensional y es mayor a medida que el modelo presenta mejor ajuste. Magee [9].
La metodología planteada en esta propuesta se relaciona con la teoría de los estadísticos R^2, los modelos lineales mixtos y los datos longitudinales. Por esta razón se hace necesario mostrar algunas definiciones y conceptos básicos que ayuden a entender como se desepeñan los estadísticos R^2 condicionales en el contexto de los modelos lineales mixtos.

Conceptos básicos

Las propiedades básicas generales de los R^2 fueron enunciadas por Kvalseth [5]. Él propone ocho criterios para evaluar los estadísticos R^2:

1. Los R^2 deben tener una interpretación razonable y útil como una medida de bondad de ajuste (Goodnes of fit, GOF siglas en inglés).

2. Los R^2 deben ser independientes de las unidades de medida.

3. El rango de valores potencial debe estar bien definido con criterios correspondientes para perfecto ajuste o total falta de ajuste.

4. Los R^2 deben ser los suficientemente generales como para ser aplicados a cualquier tipo de modelo.

5. Los R^2 no deben estar confinados a cualquier técnica de ajuste del modelo específico. Es decir, sin importar que técnica de estimación del modelo se utilice debe ser factible calcular los estadísticos R^2.

6. Los R^2 se deben poder comparar cuando se aplique al mismo conjunto de datos pero con diferentes modelos ajustados (Comparar entre modelos).

7. Los valores relativos de los R^2 deben ser en general comparables con otros criterios razonables de bondad de ajuste (Comparar entre criterios de bondad de ajuste).

8. Los residuales positivos y negativos deberían tener igual ponderación.

Cameron y Windmeijer [3], propusieron otros cuatro criterios:

1. Los R^2 no deberían decrecer a medida que se agregan regresores el modelo (sin corrección de los grados de libertad).

2. Los R^2 basados en la suma de cuadrados de los residuales deben coincidir con los R^2 basados en la suma de cuadrados explicada.

3. Debe haber una correspondencia entre los R^2 y las pruebas de significancia de los parámetros ajustados y entre los cambios de los R^2 a medida que se agregan regresores y las pruebas de significancia de todos los parámetros.

4. Los R^2 deberían tener una explicación en términos de la información contenida por los datos.

2.5. Estadístico R^2 para modelos lineales mixtos

2.5.1. El estadístico R_1^2

Vonesh y Chinchilli [16] propusieron un estadístico R^2 para evaluar bondad de ajuste en un MLM generalizado, el cual se denotará R_1^2. Asumiendo $R_i = \sigma^2 I_{n_i}$, el estadístico R_1^2 se define como:

$$R_1^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)'(y_i - \hat{y}_i)}{\sum_{i=1}^{m}(y_i - \bar{y}1_{n_i})'(y_i - \bar{y}1_{n_i})}. \tag{2-6}$$

El R_1^2 coincide con el tradicional estadístico R^2 para modelos lineales cuando m=1, por lo tanto este R_1^2 es fácil de interpretar. Sin embargo, R_1^2 no toma explícitamente las componentes aleatorias del modelo.

2.5.2. El estadístico R_c

Motivado por el coeficiente de correlación de concordancia ρ_c, Vonesh y otros [17] proponen

$$R_c = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)'(y_i - \hat{y}_i)}{\sum_{i=1}^{m}(y_i - \bar{y}1_{n_i})'(y_i - \bar{y}1_{n_i}) + \sum_{i=1}^{m}(\hat{y}_i - \hat{y}1_{n_i})'(\hat{y}_i - \hat{y}1_{n_i}) + N(\bar{y} - \hat{y})^2}, \tag{2-7}$$

como una medida de bondad de ajuste para modelos no lineales con efectos mixtos generalizados, donde N es el número total de observaciones; m es el número de individuos, y_i es el vector de valores observados del i-ésimo sujeto; $\hat{y}$ es el vector de valores predichos, $\bar{y}$ es la media de los elementos de todos los valores observados y 1_{n_i} es un vector $n_i \times 1$ de unos.

2.5.3. El estadístico $\hat{\Omega}^2$

,
Xu[18] propuso el estadístico $\hat{\Omega}^2$ que sirve para explicar la variación en el modelo lineal. Está dado por:

$$\hat{\Omega}^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}, \tag{2-8}$$

$\hat{\Omega}^2$ tiene el objetivo de estimar Ω^2=$1 - \frac{V(y_{ij}|X,b)}{V(y_{ij}\text{bajo un modelo nulo})}$, donde $j \in \{1, 2, \ldots, n_i\}$ y y_{ij} es el j-ésimo elemento de y_i, $\hat{\sigma}^2$ es el estimador de σ^2 del modelo ajustado (o modelo interés), $\hat{\sigma}_0^2$ es el estimador de la varianza de los residuales de un modelo nulo de la siguiente forma:

$$\mathbf{y}_i = \mathbf{1}_{n_i}\boldsymbol{\beta}_o + \mathbf{1}_{n_i}\mathbf{b}_{oi} + \epsilon_i, \tag{2-9}$$

es decir, $\hat{\sigma}_0^2$ es el estimador de la $V(y_{ij}|b_{oi})$, donde $\boldsymbol{\beta_0}$ es un parámetro fijo desconocido, $\boldsymbol{b_{oi}}$ es un coeficiente aleatorio desconocido que tiene distribución normal con media cero y ϵ_i es error aleatorio intra-sujeto.

2.5.4. El estadístico R_2^2

El estadístico R_2^2 propopuesto por Xu[13], que al igual que $\hat{\Omega}^2$ mide la proporción de la variación explicada por las covariables en el modelo. Se obtiene mediante la expresión:

$$R_2^2 = 1 - \frac{RSS}{RSS_0}. \tag{2-10}$$

Donde RSS es la suma de los residuales al cuadrado del modelo (2-1) y RSS_0 es la suma de los residuales al cuadrado de un modelo nulo dado en (2-9)

2.5.5. El estadístico $\hat{\rho}^2$

Xu [18] propuso el estadístico $\hat{\rho}^2$ como estimador de ρ^2 que es un parámetro que mide la proporción de la aleatoriedad explicada de una variable aleatoria Y.
ρ^2 es una transformación monótona de la entropía, $e^{[-2I(\theta)]}$, donde $I(\theta) = E[\log p(y;\theta)]$ es la esperanza de la log-verosimilitud bajo un modelo lineal mixto que tiene la forma dada en la

ecuación (2-1), la aleatoriedad residual está definida como $D(y_{ij}|X,b)$=exp $(-2E[\log p(y_{ij}|X,b)])$. La proporción de la aleatoriedad explicada esta dada por:

$$\rho^2 = 1 - \frac{D(y_{ij}|X,b)}{D\left(y_{ij}|b_o^*\right)} \tag{2-11}$$

donde $D(y_{ij}|b_o^*)$ es la esperanza de la log-verosimilitud del modelo nulo dado en la ecuación(2-9).

$$\hat{\rho}^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\exp\left(\frac{RSS}{\hat{\sigma}^2} - \frac{RRS_0}{\hat{\sigma}_0^2}\right), \tag{2-12}$$

donde RSS es la suma de los residuales al cuadrado del modelo (2-1) y RSS_0 es la suma de los residuales al cuadrado del modelo nulo dado en (2-9)

2.5.6. Estadísticos R^2 marginales y condicionales

Vonesh y otros [17] propusieron los conceptos de estadístico R^2 marginal y condicional. Para calcular la versión condicional del estadístico R^2, se utiliza el valor ajustado $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{Z}\hat{\mathbf{b}}$, no obstante, para la versión marginal del estadístico R^2, se utiliza el valor ajustado $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. La versión condicional cuenta con ambos efectos fijos y aleatorios para medir bondad de ajuste y el poder de predicción, mientras que la versión marginal sólo cuenta con la parte de efectos fijos, la media del modelo.

Los estadísticos R^2 descritos en la sección anterior junto con dos estadísticos que constituyen la propuesta original de este trabajo los cuales son especificados en el capítulo siguiente, poseen versión condicional, es decir, los valores ajustados son de la forma $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}}$. Todos los estadísticos son calculados en el capítulo 4 mediante un estudio de simulación y a partir del ajuste de diferentes modelos para evaluar su desempeño como medidas de bondad de ajuste y criterios de selección de variables.

3 Estadísticos R^2 propuestos

En este capítulo se proponen dos estadísticos R^2 para evaluar la bondad de ajuste de los modelos lineales mixtos, el R^2_{DG} y el R^2_{DGP}[1]. El R^2_{DG} da igual peso a cada residual mientras que el R^2_{DGP} pondera los residuales de forma tal que da mayor peso a los residuales de los individuos con mayor cantidad de observaciones. Ambos estadísticos se pueden utilizar para datos balanceados[2] o desbalanceados, sin embargo, los estadísticos coinciden cuando los datos son balanceados. En este sentido el R^2_{DGP} es una generalización de R^2_{DG}.
Para el cálculo de estos estadísticos los valores predichos se obtienen a partir de la siguiente relación $\hat{\mathbf{Y}}_i = \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\hat{\mathbf{b}}_i$. Estas propuestas, constituyen los aportes originales de esta tesis de maestría.

3.1. Los estadísticos R^2_{DG} y R^2_{DGP}

Para modelos lineales con efectos mixtos ajustados a conjuntos de datos con medidas repetidas, se proponen los estadísticos R^2_{DGP} y R^2_{DG} dados por:

$$R^2_{DG} = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)'(y_i - \hat{y}_i)}{\sum_{i=1}^{m}(y_i - \bar{y}_i 1_{n_i})'(y_i - \bar{y}_i 1_{n_i})} \tag{3-1}$$

$$R^2_{DGP} = 1 - \frac{m}{N}\frac{\sum_{i=1}^{m} n_i(y_i - \hat{y}_i)'(y_i - \hat{y}_i)}{\sum_{i=1}^{m}(y_i - \bar{y}_i 1_{n_i})'(y_i - \bar{y}_i 1_{n_i})} \tag{3-2}$$

Donde:
m es el número de individuos.
y_i es el vector de valores observados del individuo i, $i = 1, 2, \ldots, m$.
N es el total de observaciones.
$\hat{y}_i$ es el vector de los valores ajustados o predichos del individuo i, $i = 1, 2, \ldots, m$.
$\bar{y}_i$ es la media de los valores observados del individuo i, $i = 1, 2, \ldots, m$.
n_i es el número de observaciones del individuo i, $i = 1, 2, \ldots, m$.
1_{n_i} es un vector de unos de tamaño $n_i \times 1$, $i = 1, 2, \ldots, m$.

[1]Los subíndices de los estadístico R^2_{DG} yR^2_{DGP} corresponde a las iniciales del nombre y el apellido del autor DG: Diana Guzmán y la P hace alusión a la versión ponderada del estadístico R^2_{DG}
[2]Los datos son balanceados cuando el número de observaciones es el mismo para cada individuo

3.2. Justificación

Para los modelos lineales mixtos generalizados, Vonesh y Chichilli [16] propusieron el R_1^2 definido en la ecuación (2-6) el cual es una extensión natural del R^2 tradicional para modelos lineales. A partir de la modificación del estadístico R_1^2, se han propuesto dos estadísticos R^2 descritos en las ecuaciones (3-1) y (3-2) para evaluar la bondad de ajuste en los modelos lineales mixtos. Estos estadísticos también son extensiones del estadístico R^2 tradicional y por sus definiciones poseen propiedades e interpretación similar.

Los estadísticos R_{DGP}^2 y R_{DG}^2 buscan mejorar el rendimiento de R_1^2 al ser más específicos en el cálculo de la suma de cuadrados total, pues el R_{DGP}^2 y el R_{DG}^2 calculan la suma de cuadrados total para cada individuo en forma particular, tomando la suma de cuadrados de la diferencia entre los valores observados y la media individual en lugar de medir las distancias de los valores observados a la media general.

Los estadísticos propuestos se definen de esta manera con el objeto de obtener un porcentaje de la variación explicada más acorde la variabilidad individual, y por ende ser unas medidas más confiables de bondad de ajuste y herramientas útiles para seleccionar el mejor modelo.

En comparación con los estadísticos R_1^2 y R_c, los estadísticos R_{DG}^2 y R_{DGP}^2 cuando son estimados en los diferentes modelos ajustados, presentan valores menores de la proporción de la variación explicada por las covariables en el modelo. Esto se debe básicamente a que los denominadores de los estadísticos propuestos son menores.

Los estadísticos R_1^2 y R_{DG}^2 son medidas del porcentaje de variación explicada que se diferencian solamente en sus denominadores. El denominador del estadístico R_{DG}^2 mide las desviaciones de las observaciones de cada individuo a su respectiva media y luego las suma, de esta manera estima la variabilidad total (a través de la suma de las estimaciones de las variabilidades totales de cada individuo). El denominador del estadístico R_1^2 mide las desviaciones de las observaciones de cada individuo a la media general, de esta forma estima la variabilidad total. En esta estimación no se tiene en cuenta la clasificación individual y por tanto la variabilidad total queda sobreestimada, pues las observaciones de cada individuo oscilan alrededor de su media y no necesariamente alrededor de la media general.

Por lo anterior, el estadístico R_{DG}^2 resulta ser una medida más razonable de la proporción de la variabilidad explicada por el modelo ajustado.

El estadístico R_{DGP}^2 es una versión ponderada del estadístico R_{DG}^2 que da pesos a los individuos de acuerdo con el número de observaciones, lo cual se fundamenta *intuitivamente* en el hecho de que el individuo con mayor cantidad de datos debe ser el que aporta más

información al modelo.

3.3. Propiedades

Los estadísticos R^2 propuestos en las ecuaciones (3-1) y (3-2) poseen las siguientes propiedades.

1. *Son adimensionales.*
2. *Fáciles de interpretar como la cantidad de variación explicada por el modelo ajustado.*
3. *Son menores o iguales a uno.*
4. *Si alcanzan la cota superior, uno, se tiene perfecto ajuste.*
5. *Un valor cero o negativo indica completa falta de ajuste.*
6. *Si los datos son balanceados entonces* R^2_{DGP}=R^2_{DG}.
7. *El estadístico* R^2_{DG} *es menor o igual que el estadístico* R^2_1

Demostración.

1. Es inmediato que R^2_{DG} es adimencional, pues es el cociente de dos cantidades con las mismas unidades.

2. Veamos que el estadístico R^2_{DG} se puede interpretar como la proporción de la variabilidad de la variable respuesta que es explicada por el modelo.

   $\sum_{i=1}^{m}(y_i - \hat{y}_i)'(y_i - \hat{y}_i)$ es un estimador de la variabilidad que no es explicada por el modelo.

   En tanto que, $\sum_{i=1}^{m}(y_i - \bar{y}_i 1_{n_i})'(y_i - \bar{y}_i 1_{n_i})$ es un estimador de la variabilidad total de la variable respuesta.

   Luego, $\frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)'(y_i-\hat{y}_i)}{\sum_{i=1}^{m}(y_i-\bar{y}_i 1_{n_i})'(y_i-\bar{y}_i 1_{n_i})}$ es la proporción de la variabilidad de la variable respuesta que no es explicada por el modelo.

   Entonces, $R^2_{DG} = 1 - \frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)'(y_i-\hat{y}_i)}{\sum_{i=1}^{m}(y_i-\bar{y}_i 1_{n_i})'(y_i-\bar{y}_i 1_{n_i})}$ representa la proporción de la variabilidad de la variable respuesta que es explicada por el modelo.

3. Tenemos que $R^2_{DG} = 1 - \frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)'(y_i-\hat{y}_i)}{\sum_{i=1}^{m}(y_i-\bar{y}_i 1_{n_i})'(y_i-\bar{y}_i 1_{n_i})}$

   es claro que $\sum_{i=1}^{m}(y_i - \hat{y}_i)'(y_i - \hat{y}_i) \geq 0$ y $\sum_{i=1}^{m}(y_i - \bar{y}_i 1_{n_i})'(y_i - \bar{y}_i 1_{n_i}) > 0$,

   si esta última cantidad fuera cero no sería necesario aplicar el modelo, debido a que los valores de cada individuo permanecerían constantes.

   Por tanto, $\frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)'(y_i-\hat{y}_i)}{\sum_{i=1}^{m}(y_i-\bar{y}_i 1_{n_i})'(y_i-\bar{y}_i 1_{n_i})} \geq 0$

   como $R^2_{DG} = 1 - \frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)'(y_i-\hat{y}_i)}{\sum_{i=1}^{m}(y_i-\bar{y}_i 1_{n_i})'(y_i-\bar{y}_i 1_{n_i})}$,

   entonces $R^2_{DG} \leq 1$.

4. Veamos que si $R^2_{DG} = 1$ entonces el modelo lineal mixto ajusta perfectamente:

   Si se tiene que $R^2_{DG} = 1$,

   entonces se cumple que $\sum_{i=1}^{m}(y_i - \hat{y}_i)'(y_i - \hat{y}_i) = 0$

   esto equivale a $y_i = \hat{y}_i$ para cada $i = 1, 2, \ldots, m$

   Por lo anterior hay perfecto ajuste.

5. Veamos que si $R^2_{DG} \leq 0$ entonces hay completa falta de ajuste:

   Si se tiene que $R^2_{DG} \leq 0$,

   entonces $\frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)'(y_i-\hat{y}_i)}{\sum_{i=1}^{m}(y_i-\bar{y}_i 1_{n_i})'(y_i-\bar{y}_i 1_{n_i})} \geq 1$

   Luego $\sum_{i=1}^{m}(y_i - \hat{y}_i)'(y_i - \hat{y}_i) \geq \sum_{i=1}^{m}(y_i - \bar{y}_i 1_{n_i})'(y_i - \bar{y}_i 1_{n_i})$

   Esta última desigualdad indica que la suma de cuadrados de las distancias entre los valores observados y la recta ajustada no es menor que la suma de cuadrados de las desviaciones de cada valor observado a su respectiva media individual. Es decir, la variabilidad que no explica el modelo no es menor que la variabilidad total.
   Por lo anterior, hay total falta de ajuste.

El R^2_{DGP} posee las mismas propiedades que el R^2_{DG} y las pruebas son análogas, simplemente en el numerador se ponderan las desviaciones de los valores observados a la recta de ajuste.

6. Decir que los datos son balanceados, quiere decir que todos los individuos tienen el mismo número de observaciones.

   Es decir, $n_1 = n_2 = \ldots = n_m$.

   Además, se tiene que $N = \sum_{i=1}^{m} n_i = \sum_{i=1}^{m} n_1 = mn_1$. Entonces:

$$
\begin{aligned}
R^2_{DGP} &= 1 - \frac{m}{N}\frac{\sum_{i=1}^{m} n_i(y_i - \hat{y}_i)'(y_i - \hat{y}_i)}{\sum_{i=1}^{m}(y_i - \bar{y}_i 1_{n_i})'(y_i - \bar{y}_i 1_{n_i})} \\
&= 1 - \frac{m}{N}\frac{\sum_{i=1}^{m} n_1(y_i - \hat{y}_i)'(y_i - \hat{y}_i)}{\sum_{i=1}^{m}(y_i - \bar{y}_i 1_{n_i})'(y_i - \bar{y}_i 1_{n_i})} \\
&= 1 - \frac{mn_1}{N}\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)'(y_i - \hat{y}_i)}{\sum_{i=1}^{m}(y_i - \bar{y}_i 1_{n_i})'(y_i - \bar{y}_i 1_{n_i})} \\
&= 1 - \frac{N}{N}\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)'(y_i - \hat{y}_i)}{\sum_{i=1}^{m}(y_i - \bar{y}_i 1_{n_i})'(y_i - \bar{y}_i 1_{n_i})} \\
&= 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)'(y_i - \hat{y}_i)}{\sum_{i=1}^{m}(y_i - \bar{y}_i 1_{n_i})'(y_i - \bar{y}_i 1_{n_i})} \\
&= R^2_{DG}
\end{aligned}
$$

7. Veamos que $\sum_{i=1}^{m}(y_i - \bar{y}_i 1_{n_i})'(y_i - \bar{y}_i 1_{n_i}) \leq \sum_{i=1}^{m}(y_i - \bar{y} 1_{n_i})'(y_i - \bar{y} 1_{n_i})$

Suponga que se tienen n valores reales $x_1, x_2, \ldots, x_n$ encontremos el valor $x \in \Re$ tal que $\sum_{i=1}^{n}(x_i - x)^2$ sea mínimo.

Sea $f(x) = \sum_{i=1}^{n}(x_i - x)^2$, entonces:

$$\begin{aligned} f^{'}(x) &= \sum_{i=1}^{n} 2(x_i - x)(-1) \\ &= -2\sum_{i=1}^{n}(x_i - x) \\ &= -2(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x) \\ &= -2(\sum_{i=1}^{n} x_i - nx) \\ &= -2n(\frac{1}{n}\sum_{i=1}^{n} x_i - x) \\ &= -2n(\bar{x} - x) \end{aligned}$$

Por tanto, $f^{'}(x) = 0$ si y sólo sí $x = \bar{x}$ Como $f^{''}(x) = 2n$, entonces $f^{''}(\bar{x}) = 2n > 0$ Por tanto en el valor $x = \bar{x}$, la función $f(x)$ alcanza su mínimo.

Ahora se aplica el resultado anterior.

Tenemos que:
m es el número de individuos.
y_i es el vector de valores observados del individuo i, $i = 1, 2, \ldots, m$.
$\hat{y}_i$ es el vector de los valores ajustados o predichos del individuo i, $i = 1, 2, \ldots, m$.
$\bar{y}_i$ es la media de los valores observados del individuo i, $i = 1, 2, \ldots, m$.
n_i es el número de observaciones del individuo i, $i = 1, 2, \ldots, m$.
1_{n_i} es un vector de unos de tamaño $n_i \times 1$, $i = 1, 2, \ldots, m$.
$\bar{y}$ es la media de todos los valores observados.

Sea $i = 1, 2, \ldots, m$ entonces:

$$\begin{aligned}(y_i - \bar{y}_i 1_{n_i})'(y_i - \bar{y}_i 1_{n_i}) &= \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \\ &\leq \sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 \\ &= (y_i - \bar{y} 1_{n_i})'(y_i - \bar{y} 1_{n_i})\end{aligned}$$

Por tanto, $\sum_{i=1}^{m}(y_i - \bar{y}_i 1_{n_i})'(y_i - \bar{y}_i 1_{n_i}) \leq \sum_{i=1}^{m}(y_i - \bar{y} 1_{n_i})'(y_i - \bar{y} 1_{n_i})$

Como los estadísticos R^2_{DG} y R^2_1 son de la forma $\left(1 - \frac{Numerador}{Denominador}\right)$, tienen el mismo numerador y el denominador del estadístico R^2_{DG} es menor que el denominador del estadístico R^2_1 como acabamos de mostrar. Entonces $R^2_{DG} \leq R^2_1$ y también se tiene que el estadístico R^2_{DG} es más sensible a cambios que el estadístico R^2_1 al tener un denominador menor. Se debe notar que el numerador depende del modelo ajustado.

Además, el denominador del estadístico R^2_{DG} es una medida más coherente de la variabilidad total de la variable respuesta que el denominador del estadístico R^2_1. Pues el primero mide las desviaciones de las observaciones de cada individuo a su respectiva media, en tanto que el segundo mide las desviaciones de las observaciones de cada individuo a la media general de todas las observaciones. Por tanto el estadístico R^2_{DG} es una medida más razonable de la proporción de la variabilidad explicada por el modelo ajustado en comparación del estadístico R^2_1.

4 Estudio de simulación

4.1. Generación de datos

En este capítulo se presenta la metodología utilizada para evaluar el desempeño de algunos estadísticos R^2 encontrados en la literatura y los dos estadísticos propuestos. Se muestran dos estudios de simulación: uno con datos balanceados y el otro con datos desbalanceados controlando los siguientes parámetros:

- Tamaño de muestra
- El vector de parámetros fijos del modelo.
- La distribución del efecto aleatorio y del error aleatorio.
- La covarianza de los efectos aleatorios.

Para el conjunto de datos balanceados, los datos consisten de 100 individuos y 7 repeticiones por sujeto, por tanto se generó una variable respuesta longitudinal continua con tres términos de efectos: Una covariable de tiempo y dos variables dicotómicas. Las covariables de efectos aleatorios consisten de un término de intercepto aleatorio y término lineal para el tiempo.
Para formar los datos desbalanceados se tomó el conjunto de datos balanceados descrito anteriormente y de éste se extrajeron algunos valores en forma aleatoria, de forma tal que el número de repeticiones por individuo no sea fijo. Lo anterior, manteniendo como mínimo una observación por individuo. El número de valores retirados o extraidos de la muestra inicial es aleatorio pero garantizando que a lo sumo halla un 10 % de datos faltantes.

Se simularon en total seis conjuntos de datos, tres considerando los datos balanceados y los tres restantes para los datos desbalanceados asumiendo que la covarianza de los efectos aleatorios no tiene una estructura determinada y la covarianza de los errores intrasujeto es $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i}$. Los tres conjuntos con datos balanceados y los tres conjuntos con datos desbalanceados difieren en los valores de correlación intrasujeto, los respectivos valores de σ^2 son: $14, 58, 220$, estos valores fueron seleccionados para que la correlación intrasujeto estuviera alrededor de 0.8, 0.5, 0.2, de tal forma que se pueda evaluar el desempeño de los estadísticos R^2 cuando el grado de asociación entre las observaciones de un mismo individuo es baja, media y alta. Los parámetros utilizados en la simulación se presentan a continuación:

$\mathbf{X} = [\mathbf{1}_7, (9,9.5,9.75,10,10.5,10.75,11)', 1_7 I_1, 1_7 I_2]$; $\mathbf{Z} = [\mathbf{1}_7, (9,9.5,9.75,10,10.5,10.75,11)']$ donde $\mathbf{X}$ es una matriz diseño que contiene las covariables asociadas a los efectos fijos, $\mathbf{Z}$ es la matriz de diseño que contiene las covariables para los efectos aleatorios y $\mathbf{1}_7$ es un vector de unos 7×1, $\mathbf{I}_1$ e $\mathbf{I}_2$ son funciones indicadoras que toman el valor 0 ó 1, donde 1 indica que el individuo posee una característica determinada y 0 lo contrario. $\mathbf{1}_7$ en $\mathbf{X}$ y $\mathbf{Z}$ es una columna de unos necesaria para incluir en el modelo el intercepto fijo y el intercepto aleatorio respectivamente, $\mathbf{1}_7\mathbf{I}_1$ y $\mathbf{1}_7\mathbf{I}_2$ corresponden a dos variables dicotómicas, que según sea el caso se tendrá, para $k = 1$ o $k = 2$, $\mathbf{1}_7\mathbf{I}_k = \mathbf{0}_7$ cuando el individuo no posee la característica, ó, $\mathbf{1}_7\mathbf{I}_k = \mathbf{1}_7$ cuando el individuo posee la característica. $\mathbf{1}_7\mathbf{I}_1$ y $\mathbf{1}_7\mathbf{I}_2$ se generaron en forma aleatoria para cada individuo. Se tomó

$$\beta = \begin{pmatrix} 16 \\ 9 \\ 17 \\ 17 \end{pmatrix} \text{ y } D = \begin{pmatrix} 2 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Los valores de los parámetros para dos de los efectos fijos (las variables dicotómicas) tiene el mismo valor (17) con el objetivo que tenga el mismo peso dentro de los modelo ajustados. Para la matriz de varianzas y covarianzas de los efectos aleatorios, los valores fueron seleccionados para que la correlación entre el intercepto aleatorio y la pendiente sea 0.5.

Para el conjunto de datos balanceados al igual que para el conjunto de datos desbalanceados se probaron tres escenarios de simulación, uno para cada valor de σ^2. Para cada conjunto de datos simulados, se ajusto el modelo verdadero que consiste del tiempo y las dos variables dicotómicas como efectos fijos y del intercepto y el tiempo como efectos aleatorios. En el modelo verdadero se tomó el tiempo como una covariable tanto de efecto fijo como de efecto aleatorio debido a que en muchas ocasiones es interés para el investigador que trabaja con datos longitudinales ver como la variable respuesta es afectada por el tiempo; esto se logra incluyendo el tiempo como efecto fijo en la estructura de la media. Sin embargo, el tiempo puede afectar a los individuos de forma diferente de tal manera que los individuos pueden presentar un comportamiento superior o inferior a la media dadas sus particularidades, por tanto se incluye el tiempo como efecto aleatorio. Aparte del modelo verdadero se ajustaron tres diferentes modelos. En cada uno de ellos se toman los mismos efectos fijos: El tiempo, las dos variables dicotómicas y una covariable no relacionada con la variable respuesta (dicha variable se generó a partir de una distribución uniforme en forma independiente de la variable respuesta); en tanto que los efectos aleatorios varían de modelo a modelo. En el primer modelo se ajustan los efectos fijos ya mencionados y un intercepto aleatorio solamente; en el segundo modelo se agrega el tiempo como efecto aleatorio con respecto al primer modelo; y por último, en el tercer modelo se agrega la covariable no relacionada con la variable respuesta como efecto aleatorio con respecto al segundo modelo.

En resumen los modelos que se ajustaron para cada valor de σ^2 son los siguientes:

Tabla 4-1: Modelos Ajustados.

σ^2	**Modelo**	**Efectos Fijos**	**Efectos Aleatorios**
14	*Modelo Verdadero*	*Efectos Fijos*	*Intercepto + Pendiente*
	Modelo 1	*Efectos fijos +*	*Sólo intercepto*
	Modelo 2	*una covariable*	*Intercepto + Pendiente*
	Modelo 3	*independiente*	*Intecepto + Pendiente + una variable independiente*
58	*Modelo Verdadero*	*Efectos Fijos*	*Intercepto + Pendiente*
	Modelo 1	*Efectos Fijos +*	*Sólo intercepto*
	Modelo 2	*una covariable*	*Intercepto + Pendiente*
	Modelo 3	*independiente*	*Intecepto + Pendiente + una variable independiente*
220	*Modelo Verdadero*	*Efectos Fijos*	*Intercepto + Pendiente*
	Modelo 1	*Efectos Fijos +*	*Sólo intercepto*
	Modelo 2	*una covariable*	*Intecepto + Pendiente*
	Modelo 3	*independiente*	*Intecepto + Pendiente + una variable independiente*

En comparación con el modelo verdadero el modelo 1 es un modelo reducido en donde se elimina el tiempo en los efectos aleatorios, similarmente, en comparación con el modelo verdadero el modelo 2 y el modelo 3 son modelos sobre ajustados. En el modelo 2 se adiciona una variable no relacionada con la respuesta como efecto fijo y en el modelo 3 se adiciona tanto para los efectos fijos como para los efectos aleatorios una variable no relacionada con la respuesta.

- El sofware estadístico **R** versión 2.9.2.
- Se simularon 1000 muestras dentro de cada conjunto de datos (Cada valor de σ^2).
- Para cada muestra se ajustaron los diferentes modelos usando **REML** para estimar los parámetros.
- Para cada modelo se calcularon los estadísticos R^2 descritos en el capítulo 2 y 3, resumidos en la tabla 4-2.

Tabla 4-2: Resumen de los diferentes estadísticos R^2 condicionales para modelos lineales mixtos.

Estadístico	Formula	Autor(es)
R_1^2	$R_1^2 = 1 - \frac{\sum_{i=1}^{m}(y-\hat{y}_i)'(y_i-\hat{y}_i)}{\sum_{i=1}^{m}(y_i-\bar{y}1_{n_i})'(y_i-\bar{y}1_{n_i})}$	Vonesh y Chinchilli.(1997)
R_c	$R_c = 1 - \frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)'(y_i-\hat{y}_i)}{\sum_{i=1}^{m}(y_i-\bar{y}_i1_{ni})'(y_i-\bar{y}_i1_{ni})+\sum_{i=1}^{m}(\hat{y}_i-\hat{y}1_{ni})'(\hat{y}_i-\hat{y}1_{ni})+N(\bar{y}-\bar{y})^2}$	Vonesh et al.(1996)
$\hat{\Omega}^2$	$\hat{\Omega}^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}$	Xu(2003)
R_2^2	$R_2^2 = 1 - \frac{RSS}{RSS_0}$	Xu(2003)
$\hat{\rho}^2$	$\hat{\rho}^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\exp(\frac{RSS}{\hat{\sigma}^2} - \frac{RRS_0}{\hat{\sigma}_0^2})$	Xu(2003)
R_{DG}^2	$R_{DG}^2 = 1 - \frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)'(y_i-\hat{y}_i)}{\sum_{i=1}^{m}(y_i-\bar{y}_i1_{n_i})'(y_i-\bar{y}_i1_{n_i})}$	Guzman y Salazar (VII Coloquio Regional de Estadística, 2010)
R_{DGP}^2	$R_{DGP}^2 = 1 - \frac{m}{N}\frac{\sum_{i=1}^{m}n_i(y_i-\hat{y}_i)'(y_i-\hat{y}_i)}{\sum_{i=1}^{m}(y_i-\bar{y}_i1_{n_i})'(y_i-\bar{y}_i1_{n_i})}$	Guzman y Salazar (VII Coloquio Regional de Estadística, 2010)

[1] $\hat{\mathbf{Y}}_i = \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\hat{\mathbf{b}}_i$.

[2] RSS es la suma de los residuales al cuadrado del modelo especificado.

[3] RSS_0 es la suma de los residuales al cuadrado bajo un modelo nulo especificado en la ecuación (2-9).

Con el propósito de estimar la previsibilidad de las covariables en el modelo se utilizó la fórmula propuesta por Xu[18] para calcular Ω^2 de la siguiente manera:

$$\Omega^2 = 1 - \frac{\sigma^2}{(\beta_{tiempo}^2 + \gamma^2)var(tiempo) + \sigma^2}, \tag{4-1}$$

γ es el coeficiente de correlación intrasujeto y *var(tiempo)* es la varianza de la variable tiempo. A partir de los valores dados anteriormente se obtienen los siguientes valores para Ω^2 para el modelo verdadero

Tabla 4-3: Valores Ω^2 para el modelo verdadero.

β	var(Tiempo)	σ^2	γ	Ω^2
9	0.51	14	0.8	0.75
		58	0.5	0.42
		220	0.2	0.16

4.2. Resultados

A cada una de las mil muestras simuladas que se describieron en la sección 4.1, se ajustaron modelos lineales mixtos, los modelos ajustados se describieron en la tabla 4-1. A partir de los modelos ajustados se obtuvieron estimaciones de los siete estadísticos R^2 mostrados en la tabla 4-2 para el conjunto de datos balanceados y el conjunto de datos desbalanceados.

4.2.1. Datos balanceados

La tabla 4-4 muestra la media, el mínimo y el máximo de los valores obtenidos de los estadísticos R^2 ajustados para cada valor de σ^2 en el conjunto de datos balanceado.

Todos los estadísticos están entre cero y uno, aunque cuando se especifica valores muy grandes de σ^2 los estadísticos propuestos pueden tomar valores negativos. Sin embargo, esto refleja total falta de ajuste tal como se demuestra en la sección 3.3. Los estadísticos propuestos por Xu [18] y los estadísticos propuestos en este trabajo están cercanos al valor de Ω^2, el porcentaje de previsibilidad de las covariables definido en la ecuación (4-1). En conjuntos de datos balanceados tal como se mencionó anteriormente los estadísticos R^2_{DG} y R^2_{DGP} son iguales, lo cual se verifica en la tabla 4-4.

De la tabla 4-4 y de la figura 4-1 se observa que los resultados son consistentes con los resultados arrojados en el estudio realizado por Orelien y Edwards [11], en cuanto a que los estadísticos R^2 condicionales, son invariantes o presentan leves cambios al evaluarlos entre los diferentes modelos. Es de anotar que en dicho estudio se mantienen fijos los efectos aleatorios y se ajustan diversos modelos adicionando o suprimiendo efectos fijos. En tanto que en este estudio se modifican los efectos aleatorios.
Si se consideran los leves cambios que presentaron los estadísticos R^2 para discriminar entre los modelos, se obtienen los siguientes resultados:

Caso $\sigma^2 = 14$

- Los valores medios de los estadísticos R^2_1, R_c y los propuestos por Xu [18], se mantuvieron constantes entre el modelo verdadero y el modelo uno que tiene sobreajuste de un efecto fijo para una covariable no significativa y subajuste al suprimir el efecto aleatorio para la tiempo. Entonces estos estadísticos se inclinan por el modelo uno porque tiene menos parámetros. Por otro lado, los estadísticos propuestos en este trabajo, R^2_{DG} y R^2_{DGP}, presentaron una leve reducción, lo que permite descartar el modelo uno a favor del modelo verdadero.

- Ninguno de los estadísticos R^2 presentó cambios entre el modelo completo y el modelo dos, sobreajustado con un efecto fijo para una covariable independiente (no significativa), por tanto, todos permitieron descartar el modelo dos que tiene sobreajuste de efectos fijos. Lo que permite identificar esa covariable independiente como no importante para el modelo.

- Los estadísticos R^2_1, R_c y el $\hat{\Omega}^2$ no presentan cambios en su valores medios en el modelo 3 en comparación con el modelo verdadero, descartando el modelo 3 que tiene mayor cantidad de parámetros. Los demás estadísticos presentan un incremento del modelo

3 con respecto al verdadero modelo, sin embargo, el incremento en el valor de estos estadísticos obtenidos no fue muy grande teniendo en cuenta que el modelo 3 que está sobreajustado implica la adición de 101 parámetros. No obstante, con ninguno de estos estadísticos es posible descartar el modelo 3 en favor del verdadero.

Caso $\sigma^2 = 58$

- Los valores medios de los estadísticos R_1^2, R_c y el $\hat{\Omega}^2$, permanecen constantes entre el modelo verdadero y el modelo uno inclinándose a favor del modelo uno de acuerdo al principio de parsimonia. Los estadísticos R_2^2, $\hat{\rho}^2$, R_{DG}^2 y R_{DGP}^2, presentaron un cambio leve con respecto al modelo verdadero, lo que permite descartar el modelo uno a favor del modelo verdadero.

- Todos los estadísticos permiten descartar el modelo dos que tiene sobreajuste de efectos fijos a favor del verdadero modelo. Esto se debe a que los estadísticos R_1^2, R_c y los propuestos por Xu [18] no presentaron cambios y los estadísticos R_{DG}^2 y R_{DGP}^2 fueron mayores en el modelo verdadero que en el modelo dos.

- Al comparar el modelo verdadero versus el modelo tres se encontró que el estadístico R_c no presentó cambios, los demás estadísticos obtuvieron su mayor valor en el modelo tres, que está sobreajustado tanto con un efecto fijo como con un efecto aleatorio. Sólo el estadístico R_c elige el verdedero modelo, con los demás no es posible descartar el modelo tres.

Caso $\sigma^2 = 220$

- Al comparar el modelo uno contra el verdadero, todos los estadísticos, exceptuando el R_1^2, presentaron leves cambios, sugiriendo descartar el modelo uno, mientras que con el valor medio obtenido del R_1^2 se sugiere descartar el modelo verdadero.

- Al elegir entre el modelo dos y el verdadero, todos los estadísticos dan evidencia para descartar el modelo dos.

- Cuando se enfrentan el modelo tres y el verdadero, ningún estadístico permite descartar el modelo tres, pues todos toman su máximo valor al evaluarlos en este modelo.

4.2.2. Datos desbalanceados

La tabla 4-5 muestra la media, el mínimo y el máximo de los valores obtenidos de los estadísticos R^2 ajustados para cada valor de σ^2 en el conjunto de datos desbalanceados. Se obtuvieron los siguientes resultados:

- Todos los estadísticos están entre cero y uno.

- Los estadísticos R^2_{DG} y R^2_{DGP} no son iguales.

- Los estadísticos propuestos por Xu [18] y los estadísticos propuestos en este trabajo están cercanos al valor de Ω^2, el coeficiente de previsibilidad, por tanto son buenos estimadores de la proporción de la variabilidad explicada por las covariables en el modelo.

- Al igual que el conjunto de datos balanceados se observa que los resultados R^2 condicionales son invariantes o presentan leves cambios al evaluarlos entre los diferentes modelos.

Caso $\sigma^2 = 14$

- Los estadísticos R^2_1, R_c, los propuestos por Xu [18] y el R^2_{DG} se mantuvieron constantes entre el modelo uno y el verdadero, no permiten descartar el modelo uno a favor del modelo verdadero, en tanto que el R^2_{DGP} si lo hace.

- Al seleccionar entre el modelo dos y el verdadero, todos los estadísticos permanecen constantes, de esta manera se inclinan por el verdadero modelo.

- Cuando se selecciona entre el modelo tres y el verdadero, los estadísticos R^2_{DGP}, R^2_1, R_c y $\hat{\Omega}^2$ permanecen constantes y así descartan el modelo tres, los estadísticos R^2_2, $\hat{\rho^2}$ y R^2_{DG} toman su valor máximo en el modelo tres y de esta forma no descartan el modelo tres.

Caso $\sigma^2 = 58$

- Los R^2_1, R_c y $\hat{\Omega}^2$ se mantuvieron constantes entre el modelo uno y el verdadero, por tanto, no permiten descartar el modelo uno a favor del modelo verdadero, los demás estadísticos si lo hacen.

- Al seleccionar entre el modelo dos y el verdadero, el estadístico R^2_{DGP} toma un valor mayor en el modelo verdadero, en tanto que el resto de estadísticos permanecen constantes, en cualquier caso se descarta el modelo dos.

- Cuando se selecciona entre el modelo tres y el verdadero, el único estadístico que elije el verdadero modelo es el estadístico R_c pues permanece constante, el resto de estadísticos presentan variaciones y no descartan el modelo tres.

Caso $\sigma^2 = 220$

- Al comparar el modelo uno con el verdadero, solamente el estadístico R_1^2 permanece constante y no descarta el modelo uno, el resto de estadísticos toman valores diferentes entre los dos modelos y dan evidencia para elegir el modelo verdadero.

- Al seleccionar entre el modelo dos y el verdadero, todos los estadísticos permiten descartar el modelo dos, se destaca que el R_{DGP}^2 varía entre el par de modelos y los otros permanecen constantes.

- Cuando se selecciona entre el modelo tres y el verdadero, todos los estadísticos eligen el modelo tres.

De los resultados anteriores obtenidos mediante este estudio de simulación, si se consideran los leves cambios como significativos en los valores de los estadísticos en los modelos que no tienen sobreajuste de efectos aleatorios, los estadísticos R_{DG}^2 y R_{DGP}^2 alcanzaron a identificar adecuadamente los efectos fijos que eran importantes para el modelo en el caso de datos balanceados; sólo el R_{DGP}^2 logró identificar los efectos fijos importantes en el conjunto de datos desbalanceados. Pero en general, tanto para el conjunto de datos balanceados como para el conjunto de datos desbalanceados, si se consideran los leves cambios que presentan los estadísticos R^2 como significativos se termina seleccionando el modelo tres que tiene sobreajuste. Pero si se considera que los leves cambios no son significativos, los estadísticos R^2 conducen a la elección del modelo uno que tiene menos parámetros, un modelo subajustado. En cualquier caso los estadísticos R^2 no permiten elegir el verdadero modelo y conducen a la elección de un modelo mal especificado.

Las simulaciones también muestran que los estadísticos R^2 condicionales analizados no son adecuados como medidas de bondad de ajuste cuando se trabaja con modelos lineales mixtos, esto se debe principalmente a que los estadísticos R^2 en los diferentes modelos ajustados no presentan mayores variaciones en relación al modelo verdadero y en consecuencia no logran identificar cuando las covariables son importantes. Lo anterior porque se supone que una propiedad deseada en un R^2 es que pueda disminuir su valor cuando se eliminan una o más covariables importantes para el modelo y la disminución en el valor debe ser proporcional a la cantidad de la variación de la variable respuesta que es explicada por la variable eliminada. Por otra parte, al adicionar una covariable no relacionada con la variable respuesta se espera que el valor del estadístico R^2 no aumente.

En las simulaciones se nota que los estadísticos R_{DG}^2 y R_{DGP}^2 no son cercanos a uno en el modelo verdadero la razón es porque estos son de la forma $1 - \frac{Numerador}{Denominador} = 1 - \frac{S\hat{S}E}{S\hat{S}T}$. Cuando se calcula el denominador se supone que los individuos son independientes dos a dos, es decir, que las observaciones de un individuo son independientes de las observaciones de cualquier

otro individuo, por tanto, al estimar la variabilidad total, SST, se suman las variabilidades de cada individuo, de esta forma el denominador no resulta ser una cantidad grande con respecto al numerador y por ende $1 - \frac{Numerador}{Denominador}$ no es cercano a uno. Se debe tener en cuenta que al generar el conjunto de datos se consideró un componente de error sistemático, ϵ_i, que distancia los puntos generados de la estructura de la media, esto hace que el $Numerador \neq 0$ y en consecuencia $1 - \frac{Numerador}{Denominador} \neq 1$. Los otros estadísticos son más cercanos a uno porque son de la misma forma $1 - \frac{Numerador}{Denominador}$, pero en el cálculo del $Denominador$ toman todos los valores de todos los individuos y por tanto obtienen estimadores del $Denominador$, $\hat{SST}$, que producen valores mayores. En consecuencia, $1 - \frac{Numerador}{Denominador}$ resultan más cercanos a uno en comparación de R^2_{DG} y R^2_{DGP}.

Cabe señalar que a medida que la variabilidad intraindividual (σ^2) aumenta los valores de los estadísticos R^2 disminuyen.

Para determinar si la falta de cambio de los estadísticos R^2 entre los modelos subajustados y los sobreajustados se debe a que la estructura de covarianzas de los errores intrasujeto que se tomó es de la forma: $R_i = \sigma^2 I_{n_i}$ para cada individuo, la cual asume que las correlaciones entre las observaciones del mismo individuo son nulas, se realizaron simulaciones bajo las misma condiciones descritas anteriormente pero tomando las siguientes estructuras para R_i: AR(1) y matriz de varianzas y covarianzas desestructurada (UN). Los valores medios de los estadísticos R^2 no varían de modelo a modelo aún asumiendo que hay correlación entre las observaciones del mismo individuo. Por tanto, los resultados no se muestran debido a que las conclusiones obtenidas son las mismas que para el caso de $R_i = \sigma^2 I_{n_i}$.

Las figuras 4-1 a 4-6 muestran los valores medios de los estadísticos R^2 para cada uno de los modelos considerados para cada de valor de σ^2, tanto para el conjunto de datos balanceados como para el conjunto de datos desbalanceados. Las figuras muestran la falta de cambios de los estadísticos R^2 al ser evaluados en los diferentes modelos en comparación del modelo verdadero. Para las figuras 4-1 a 4-6, el 1 corresponde a R^2_1, 2 a R_c, 3 a $\hat{\Omega}^2$, 4 a R^2_2, 5 a $\hat{\rho}^2$, 6 a R^2_{DG} y 7 a R^2_{DGP}.

Tabla 4-4: Medias, máximos y mínimos de los estadísticos R^2 estudiados para datos balanceados

σ^2	Modelo	Ω^2	R_1^2	R_c	$\hat{\Omega}^2$	R_2^2	$\hat{\rho}^2$	R_{DG}^2	R_{DGP}^2
14	***Modelo Verdadero.*** Efectos fijos: Tiempo y dos variables dicotómicas. Efectos aleatorios: Intercepto y el tiempo como Pendiente aleatoria.	0.75	0.95(0.93,0.97)	0.98(0.96,0.98)	0.75(0.69,0.8)	0.75(0.69,0.81)	0.75(0.69,0.81)	0.75(0.69,0.81)	0.75(0.69,0.81)
	Modelo 1. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto.		0.95(0.93,0.97)	0.98(0.96,0.98)	0.75(0.69,0.78)	0.75(0.69,0.78)	0.75(0.69,0.78)	0.74(0.69,0.78)	0.74(0.69,0.78)
	Modelo 2. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto y el tiempo como Pendiente.		0.95(0.93,0.97)	0.98(0.96,0.98)	0.75(0.69,0.79)	0.75(0.69,0.79)	0.75(0.69,0.79)	0.75(0.69,0.79)	0.75(0.69,0.79)
	Modelo 3. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto, el tiempo como Pendiente aleatoria más una covariable independiente.		0.95(0.93,0.97)	0.98(0.96,0.98)	0.75(0.7,0.8)	0.76(0.7,0.81)	0.76(0.70,0.81)	0.76(0.7,0.81)	0.76(0.7,0.81)
58	***Modelo Verdadero***. Efectos fijos: Tiempo y dos variables dicotómicas. Efectos aleatorios: Intercepto y el tiempo como Pendiente aleatoria.	0.42	0.83(0.76,0.88)	0.91(0.86,0.94)	0.42(0.33,0.51)	0.42(0.32,0.52)	0.42(0.32,0.52)	0.42(0.32,0.51)	0.42(0.32,0.51)
	Modelo 1. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto		0.83(0.75,0.87)	0.91(0.86,0.93)	0.42(0.33,0.49)	0.41(0.32,0.49)	0.41(0.32,0.49)	0.41(0.31,0.49)	0.41(0.31,0.49)
	Modelo 2. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto y el tiempo como Pendiente aleatoria.		0.83(0.76,0.88)	0.91(0.86,0.93)	0.42(0.33,0.50)	0.42(0.33,0.50)	0.42(0.33,0.50)	0.41(0.32,0.49)	0.41(0.32,0.49)
	Modelo 3. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto, el tiempo como Pendiente aleatoria más una covariable independiente.		0.84(0.76,0.89)	0.91(0.86,0.94)	0.43(0.33,0.53)	0.44(0.32,0.55)	0.44(0.32,0.54)	0.43(0.32,0.54)	0.43(0.32,0.54)
220	***Modelo Verdadero***. Efectos fijos: Tiempo y dos variables dicotómicas. Efectos aleatorios: Intercepto y el tiempo como Pendiente aleatoria.	0.16	0.57(0.47,0.69)	0.73(0.64,0.82)	0.17(0.07,0.28)	0.15(0.05,0.30)	0.15(0.05,0.29)	0.13(0.03,0.28)	0.13(0.03,0.28)
	Modelo 1. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto.		0.57(0.45,0.66)	0.72(0.62,0.80)	0.16(0.07,0.24)	0.14(0.05,0.23)	0.14(0.05,0.23)	0.12(0.02,0.20)	0.12(0.02,0.20)
	Modelo 2. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto y el tiempo como Pendiente aleatoria.		0.57(0.46,0.66)	0.73(0.63,0.80)	0.16(0.07,0.26)	0.15(0.05,0.26)	0.15(0.06,0.26)	0.12(0.03,0.24)	0.12(0.03,0.24)
	Modelo 3. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto , el tiempo como Pendiente aleatoria más una covariable independiente.		0.59(0.47,0.70)	0.74(0.64,0.83)	0.18(0.09,0.28)	0.17(0.05,0.32)	0.17(0.06,0.31)	0.15(0.03,0.30)	0.15(0.03,0.30)

Los estadísticos R^2 fueron calculados usando en el numerador $\hat{Y}_i = \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\hat{b}_i$

Los modelos son especificados en la tabla 4-1

Tabla 4-5: Medias, máximos y mínimos de los estadísticos R^2 estudiados para datos desbalanceados

σ^2	Modelo	Ω^2	R_1^2	R_c	$\hat{\Omega}^2$	R_2^2	$\hat{\rho}^2$	R_{DG}^2	R_{DGP}^2
14	***Modelo Verdadero.*** Efectos fijos: Tiempo y dos variables dicotómicas. Efectos aleatorios: Intercepto y el tiempo como Pendiente aleatoria.	0.75	0.95(0.93,0.97)	0.98(0.97,0.98)	0.75(0.7,0.8)	0.75(0.7,0.81)	0.75(0.7,0.81)	0.75(0.7,0.81)	0.75(0.69,0.8)
	Modelo 1. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto.		0.95(0.93,0.97)	0.98(0.96,0.98)	0.75(0.7,0.79)	0.75(0.7,0.79)	0.75(0.7,0.79)	0.75(0.7,0.79)	0.74(0.69,0.78)
	Modelo 2. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto y el tiempo como Pendiente.		0.95(0.93,0.97)	0.98(0.97,0.98)	0.75(0.7,0.79)	0.75(0.7,0.79)	0.75(0.7,0.79)	0.75(0.7,0.79)	0.75(0.69,0.79)
	Modelo 3. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto, el tiempo como Pendiente aleatoria más una covariable independiente.		0.95(0.93,0.97)	0.98(0.97,0.99)	0.75(0.7,0.8)	0.76(0.7,0.81)	0.76(0.7,0.81)	0.76(0.7,0.81)	0.75(0.69,0.81)
58	***Modelo Verdadero***. Efectos fijos: Tiempo y dos variables dicotómicas. Efectos aleatorios: Intercepto y el tiempo como Pendiente aleatoria.	0.42	0.83(0.77,0.88)	0.91(0.87,0.94)	0.42(0.32,0.52)	0.42(0.32,0.53)	0.42(0.32,0.53)	0.42(0.31,0.53)	0.41(0.3,0.52)
	Modelo 1. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto		0.83(0.76,0.88)	0.91(0.86,0.94)	0.42(0.32,0.51)	0.41(0.31,0.51)	0.41(0.31,0.51)	0.41(0.31,0.51)	0.4(0.3,0.5)
	Modelo 2. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto y el tiempo como Pendiente aleatoria.		0.83(0.77,0.88)	0.91(0.87,0.94)	0.42(0.32,0.52)	0.42(0.32,0.52)	0.42(0.32,0.52)	0.42(0.31,0.51)	0.4(0.3,0.51)
	Modelo 3. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto, el tiempo como Pendiente aleatoria más una covariable independiente.		0.84(0.78,0.89)	0.91(0.88,0.94)	0.43(0.32,0.54)	0.44(0.32,0.56)	0.44(0.32,0.56)	0.43(0.31,0.55)	0.42(0.3,0.55)
220	***Modelo Verdadero***. Efectos fijos: Tiempo y dos variables dicotómicas. Efectos aleatorios: Intercepto y el tiempo como Pendiente aleatoria.	0.16	0.57(0.45,0.69)	0.73(0.62,0.82)	0.17(0.09,0.27)	0.15(0.06,0.27)	0.15(0.07,0.27)	0.13(0.05,0.26)	0.12(0.02,0.24)
	Modelo 1. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto.		0.57(0.46,0.66)	0.72(0.63,0.8)	0.16(0.08,0.24)	0.14(0.05,0.22)	0.14(0.05,0.23)	0.12(0.03,0.2)	0.1(0,0.18)
	Modelo 2. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto y el tiempo como Pendiente aleatoria.		0.57(0.46,0.66)	0.73(0.63,0.8)	0.17(0.09,0.25)	0.15(0.06,0.26)	0.15(0.06,0.26)	0.13(0.04,0.24)	0.11(0.01,0.21)
	Modelo 3. Efectos fijos: Tiempo, dos variables dicotómicas y una covariable independiente Efectos aleatorios: Intercepto , el tiempo como Pendiente aleatoria más una covariable independiente.		0.59(0.47,0.69)	0.74(0.64,0.82)	0.18(0.1,0.3)	0.17(0.07,0.32)	0.17(0.08,0.32)	0.16(0.05,0.31)	0.14(0.02,0.3)

Los estadísticos R^2 fueron calculados usando en el numerador $\hat{Y}_i = \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\hat{b}_i$

Los modelos son especificados en la tabla 4-1

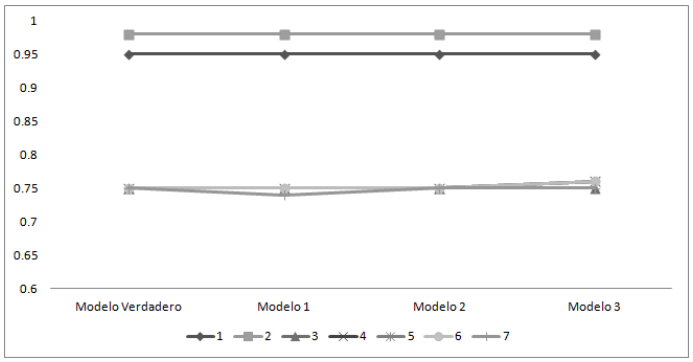

Figura 4-1: Perfiles individuales de los estadísticos R^2 entre modelos cuando $\sigma^2 = 14$ para el conjunto de datos balanceado

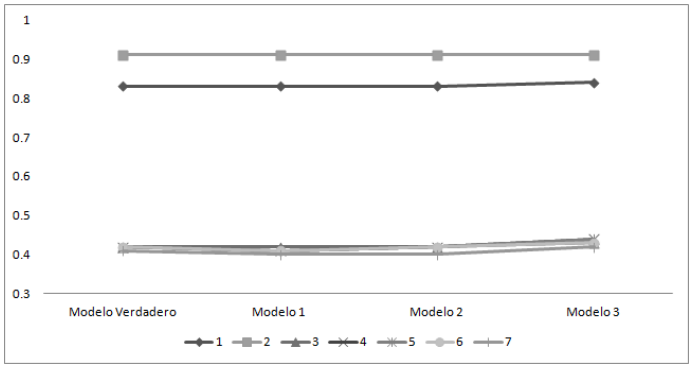

Figura 4-2: Perfiles individuales de los estadísticos R^2 entre modelos cuando $\sigma^2 = 58$ para el conjunto de datos balanceado

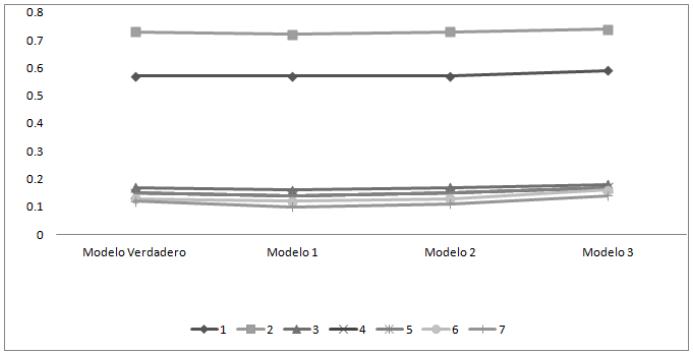

Figura 4-3: Perfiles individuales de los estadísticos R^2 entre modelos cuando $\sigma^2 = 220$ para el conjunto de datos balanceado

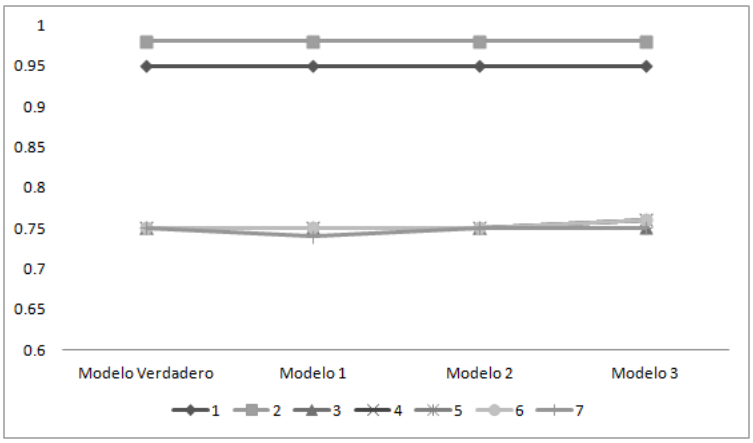

Figura 4-4: Perfiles individuales de los estadísticos R^2 entre modelos cuando $\sigma^2 = 14$ para el conjunto de datos desbalanceado

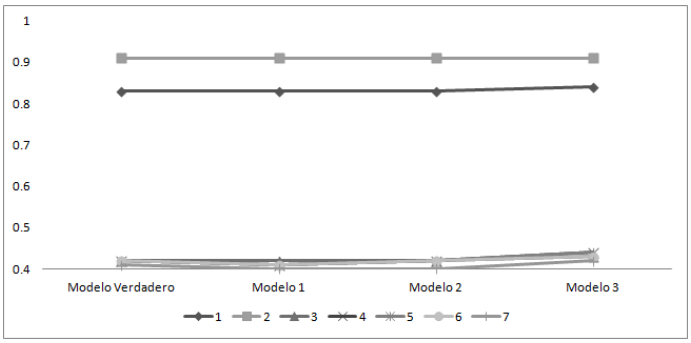

Figura 4-5: Perfiles individuales de los estadísticos R^2 entre modelos cuando $\sigma^2 = 58$ para el conjunto de datos desbalanceado

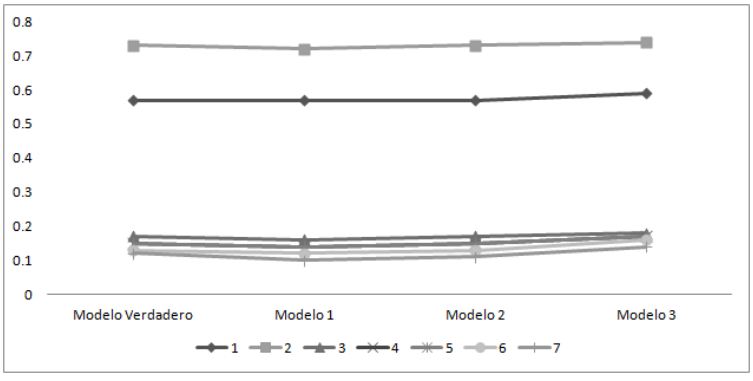

Figura 4-6: Perfiles individuales de los estadísticos R^2 entre modelos cuando $\sigma^2 = 220$ para el conjunto de datos desbalanceado

5 Aplicación

En este capítulo se hace una breve descripción de dos conjuntos de datos reales y se estiman los estadísticos R^2 expuestos en capítulos anteriores a partir de modelos lineales mixtos ajustados, con el objetivo de ilustrar la utilidad de estos estadísticos R^2 como medidas de bondad ajuste y criterios de selección.

5.1. Secado de madera ciprés

Botero[1] diseño un experimento para analizar el secado al aire de unos tablones de ciprés en función del tiempo y su grosor para usos industriales. El experimento consistió en colocar piezas de madera de cinco diferentes grosores en un recinto protegido de la lluvia para que fueran perdiendo humedad hasta alcanzar un secado natural. Para cada grosor se tomaron veinte piezas de madera manteniendo el mismo ancho y longitud. Las especificaciones se dan en la tabla 5-1.

Tabla 5-1: Especificaciones madera ciprés

Numero de Tablones	**Espesor(mm)**	**Ancho(mm)**	**Largo(m)**
20	50	180	2.54
20	40	180	2.54
20	30	180	2.54
20	25	180	2.54
20	20	180	2.54

Cada pieza de madera se pesó el primer día una vez se dio inicio al experimento, luego para evaluar la pérdida de humedad de cada pieza se pesó cada tablón durante los días 7, 14, 21, 28, 35, 42, 49, 56, 62, 70, y 92. La base de datos tomada de Botero[1] contiene el número que identifica el tablón, el grosor, el peso inicial junto con los pesos tomados durante los once días correspondientes a las mediciones. Se cuenta entonces con un total de cien tablones y doce repeticiones por tablón más la variable grosor. Estos datos corresponden al tipo de datos longitudinales debido a que cada tablón posee mediciones sucesivas en el tiempo tal como se muestra en la figura 5-1. En este gráfico se muestra los perfiles individuales (tablones) clasificados por grosor.

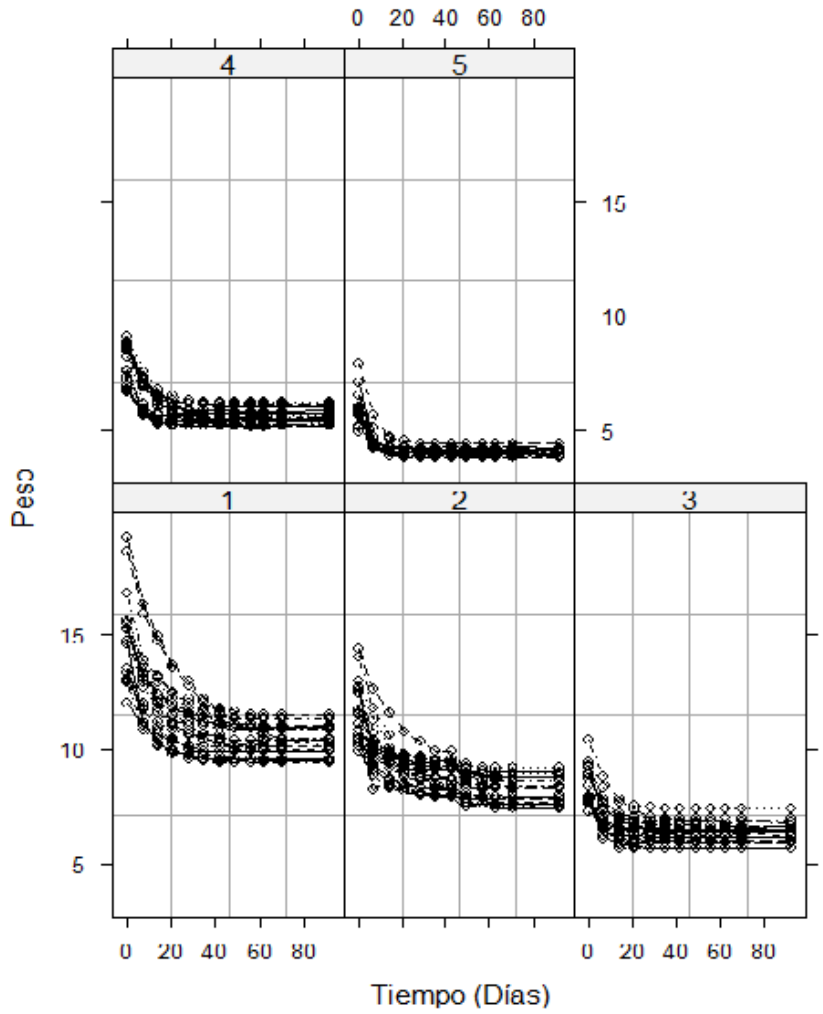

Figura 5-1: Perfiles Individuales Clasificado por Grosor

Del gráfico 5-1 se nota que cada tablón de madera ciprés va perdiendo peso a medida que transcurre el tiempo hasta que se seca, a partir de ese instante su peso permanece constante. Por tanto, en las últimas observaciones del peso de cada tablón no hay cambios, esto independientemente del grosor. Dadas las consideraciones anteriores sobre los datos relacionados con los tablones de maderas ciprés mostrados en la figura 5-1, se sugiere un modelo no lineal con efectos mixtos. Para efectos de este estudio, se tomaron únicamente las mediciones correspondientes al primer día y los días 7, 14 y 21 a las cuales se les ajustan diversos modelos lineales mixtos para ilustrar el desempeño de los estadísticos R^2. El gráfico 5-2 muestra los perfiles individuales para cada tablón clasificados por grosor tomando las mediciones del primer día y los días 7, 14 y 21.

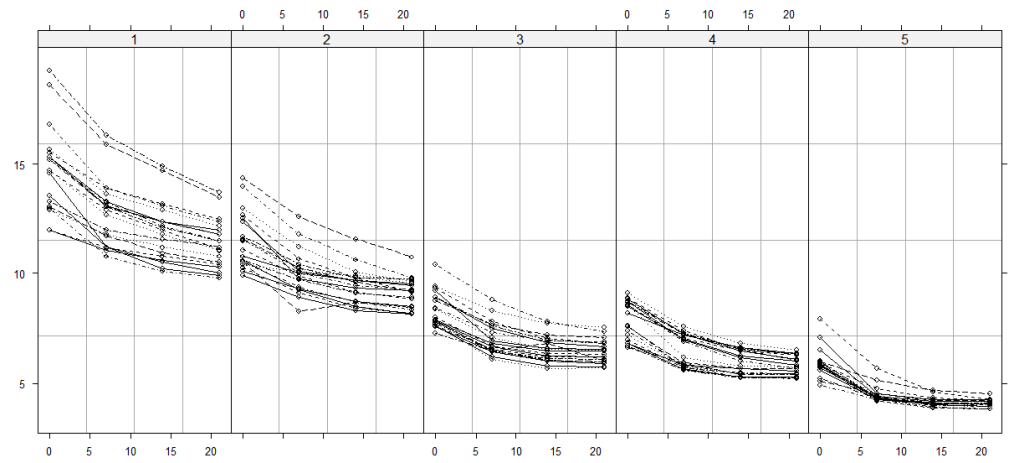

Figura 5-2: Perfiles Individuales Clasificado por Glosor Recortado

Al conjunto de datos de secado de madera ciprés se ajustaron varios modelos lineales con efectos mixtos de la forma $y_i = X_i\beta + Z_i b_i + e_i$, $i = 1, 2, \ldots, m$; donde y_i denota el peso de la i-ésima pieza de madera, $m = 100$ es el número de piezas de madera, X_i contiene las covariables de efectos fijo, β es el vector de parámetros de efecto fijo, Z_i es la matriz que contiene las variables explicativas de efecto aleatorio, b_i es el vector de parámetros de efecto aleatorio, con $b_i \sim N(0, D)$ y e_i es un vector de tamaño $m \times 1$ de términos de error (ó desviación), $e_i \sim N(0, \sigma^2 I)$.

Para la implementación se utilizó el paquete R versión 2.9.2, librería nmle mediante la función REML. Los modelos ajustados difieren tanto en términos de efecto fijo como aleatorio. A continuación se precisan los modelos ajustados.

- **Modelo 1**: Consiste del tiempo y el espesor como efectos fijos, un intercepto aleatorio y el tiempo como efecto aleatorio.
- **Modelo 2**: Tiene como efectos fijos el tiempo y el espesor y un intercepto aleatorio.
- **Modelo 3**: El espesor es la única covariable de efecto fijo, en tanto que como efectos aleatorios considera el intercepto y el tiempo.
- **Modelo 4**: Considera el tiempo como efecto fijo y un intercepto aleatorio junto con el tiempo como efecto aleatorio.
- **Modelo 5**: Solamente considera el tiempo como efecto fijo y un intercepto aleatorio.
- **Modelo 6**: No tiene ninguna covariable como efecto fijo, ajusta un intercepto fijo uno aleatorio y el tiempo como efecto aleatorio.

En la tabla 5-3 se observan los valores que toman los estadísticos R^2 condicionales, obtenidos a través del ajuste de diferentes modelos descritos al conjunto de datos de secado de madera ciprés.

En general los estadísticos R^2 presentan pequeñas variaciones en sus valores entre los modelos ajustados, sin embargo los estadísticos propuestos Xu [18] y los estadísticos propuestos en este trabajo tienen mayores variaciones y además tienden a disminuir cuando no se consideran el tiempo con una covariable de efecto aleatorio en relación a los demás modelos. Los valores de los estadísticos R_1^2 y el R_c permanecen prácticamente constantes desde el modelo 1 hasta el modelo 6.

Tabla 5-2: Estadísticos R^2 condicionales evaluados en los datos balanceados de maderas ciprés

modelo	Efectos fijos	Efectos aleatorios	R_1^2	R_c	$\widehat{\Omega}^2$	R_2^2	$\widehat{\rho}^2$	R_{DG}^2	R_{DGP}^2
1	Tiempo + Espesor	Intercepto+ Tiempo	0.9823298	0.9910862	0.8182904	0.8202038	0.8197365	0.8180971	0.8180971
2	Tiempo + Espesor	Intercepto	0.9747244	0.9872004	0.7449945	0.7428176	0.7433373	0.7398042	0.7398042
3	Espesor	Intercepto + Tiempo	0.981482	0.9906545	0.802492	0.811577	0.8092667	0.8093693	0.8093693
4	Tiempo	Intercepto + Tiempo	0.9824615	0.9911532	0.8124347	0.8215435	0.8192206	0.8194526	0.8194526
5	Tiempo	Intercepto	0.9752371	0.9874633	0.7449945	0.7480349	0.7472911	0.7450826	0.7450826
6	Intercepto	Intercepto + Tiempo	0.9827345	0.991292	0.7975106	0.8243218	0.816866	0.8222634	0.8222634

Los estadísticos R^2 fueron calculados usando en el numerador $\hat{y}_i = X_i\hat{\beta} + Z_i\hat{b}_i$

5.2. Datos de crecimiento dental

Para aplicar los estadísticos R^2 a datos reales se utilizan datos de crecimiento dental que fueron introducidos por Potthoff y Roy[12] estos contiene medidas del crecimiento de 27 niños 16 hombres y 11 mujeres. A cada niño con rayos X se le registró la distancia desde el centro de la pituitaria hasta la fisura del maxilar en las edades de 8, 10 ,11 y 14 años.

La base de datos contiene 27 individuos (niños) con 4 repeticiones cada uno, para un total de 108 observaciones. Las especificaciones se dan en la tabla 5-3. Estos datos corresponden al tipo de datos longitudinales balanceados debido a que cada niño posee mediciones sucesivas en el tiempo tal como se muestra la figura 5-3. Esta figura muestra los perfiles individuales (niños) clasificados por género.

Con el objetivo de ilustrar el desempeño de los estadísticos R^2 incluidos en este trabajo en la sección 5-3 se ajustaron diferentes modelos al conjunto de datos completo descrito anteriormente. Además, de la tabla 5-2 se suprimieron nueve observaciones dejando nueve individuos incompletos a la edad de 10 años tal como lo ilustran Little y Rubin[7] quienes describen el mecanismo por el que los individuos pequeños a los 8 años tienen a perderse en el seguimiento a los 10 años de edad. Las medidas retiradas se presentan en la tabla 5-2 marcadas con asterisco. Este conjunto de datos se utiliza para observar el desempeño de los estadísticos R^2 en datos desbalanceados.

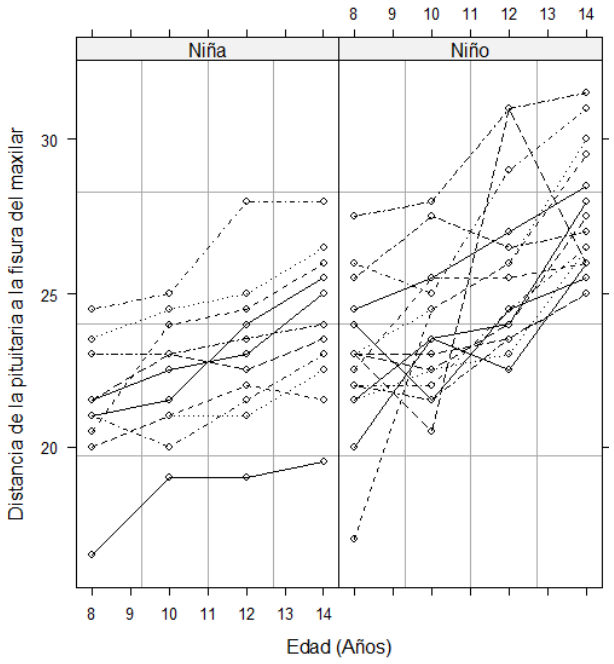

Figura 5-3: Perfiles Individuales Clasificado por genero

Tabla 5-3: Datos de crecimiento dental para 11 niñas y 16 niños Verbeke[15].

	Edad (Años)					Edad(Años)			
Niña	**8**	**10**	**12**	**14**	Niño	**8**	**10**	**12**	**14**
1	21	20	21.5	23	1	26	25	29	31
2	21	21.5	24	25.5	2	21.5	22.5*	23	26.5
3	20.5	24*	24.5	26	3	23	22.5	24	27.5
4	23.5	24.5	25	26.5	4	25.5	27.5	26.5	27
5	21.5	23	22.5	23.5	5	20	23.5*	22.5	26
6	20	21*	21	22.5	6	24.5	25.5	27	28.5
7	21.5	22.5	23	25	7	22	22	24.5	26.5
8	23	23	23.5	24	8	24	21.5	24.5	25.5
9	20	21*	22	21.5	9	23	20.5	31	26
10	16.5	19	19	19.5	10	27.5	28	31	31.5
11	24.5	25	28	28	11	23	23	23.5	25
					12	21.5	23.5*	24	28
					13	17	24.5	26	29.5
					14	22.5	25.5	25.5	26
					15	23	24.5	26	30
					16	22	21.5*	23.5	25

Se modeló, utilizando un modelo de efectos mixtos con datos longitudinales estimando los efectos fijos y los aleatorios a través del modelo $y_i = X_i\beta + Z_i b_i + e_i$ para $i = 1, 2 \ldots, m$ donde y_i denota la distancia desde el centro de la pituitaria hasta la fisura del maxilar del i-ésimo niño, m=27 es el número de niños que participaron en el estudio; X_i contiene los predictores para los efectos fijos, género, y la variable edad medida en años; β es el vector de parámetros de efectos fijos, Z_i es una matriz de covarianzas de los efectos aletorios en el cual se incluye una columna de unos y una columna para la edad (años); b_i es el vector de parámetros aleatorios y $b_i \sim N(0, D)$; y e_i es un vector de tamaño $(m \times 1)$ de errores, $e_i \sim N(0, \sigma^2 I)$. Para evaluar el desempeño de los estadísticos R^2 se ajustaron diferentes modelos en el paquete R versión 2.9.2 con la librería nlme utilizando REML. En los modelos ajustados se modificaron tanto términos de efectos fijos como términos de efectos aleatorios.

- **Modelo 1:** Consiste de las covariables *edad*, *género* y la interacción *edad* × *género* como efectos fijos, además posee un intercepto aleatorio y la covariable *edad* como efecto aleatorio.
- **Modelo 2:** Los efectos fijos están dados por las covariables *edad*, *género* y la interacción *edad* × *género*, en tanto que como efecto aleatorio se ajusta un intercepto aleatorio.
- **Modelo 3:** Se ajustan efectos fijos para la *edad* y el *género* en efectos aleatorios se incluye un intercepto aleatorio y una pendiente medida a través de la *edad* como efecto aleatorio.
- **Modelo 4:** En este se ajustan efectos fijos para la *edad* y el *género* junto con un intercepto aleatorio.
- **Modelo 5:** Este modelo explica la distancia desde el centro de la pituitaria hasta la fisura del maxilar a través de la covariable *edad* como efecto fijo, un interceptoto aleatorio y la *edad* como efecto aleatorio.
- **Modelo 6:** Este modelo explica la distancia desde el centro de la pituitaria hasta la fisura del maxilar a través de la covariable *edad* como efecto fijo y un interceptoto aleatorio.
- **Modelo 7:** Posee solamente un efecto fijo, el intercepto, un intercepto aleatorio y la covariable *edad* como efecto aleatorio.

Los modelos anteriores se ajustaron a los datos de crecimiento de Pathoff y Roy[12] . En la tabla 5-4 se muestran los resultados de los estadísticos R^2 condicionales para los datos balanceados. En la tabla 5-5 se presentan los valores de los estadísticos R^2 condicionales en el caso de los datos desbalanceados.

Todos los estadísticos R^2 presentaron diferencias pequeñas al ser calculados en los diferentes modelos desde el modelo uno hasta el modelo siete para ambos conjuntos de datos. Los estadísticos R_1^2 y R_c toman valores más altos que los otros estadísticos dando a entender que los modelos se ajustan muy bien a los datos, sin embargo, en el estudio de simulación, se mostró que estos sobreestimaban el valor del coeficiente de previsibilidad y que los otros son más cercanos a dicho valor, por tanto, son más confiables los valores de los otros estadísticos y dan una mejor estimación del porcentaje de la variación explicada por el modelo.

Tabla 5-4: Estadísticos R^2 condicionales evaluados en los datos balanceados de crecimiento Pathoff y Roy[12].

modelo	efectos fijos	efectos aleatorios	R_1^2	R_c	$\widehat{\Omega}^2$	R_2^2	$\widehat{\rho}^2$	R_{DG}^2	R_{DGP}^2
1	Edad + género +edad*género	Intercepto+ edad	0.8631646	0.9265576	0.6577502	0.7086512	0.696572	0.704056	0.704056
2	Edad + género +edad*género	Intercepto	0.8278813	0.905837	0.6101137	0.6335262	0.6286134	0.6277461	0.6277461
3	Edad + género	Intercepto + Edad	0.862878	0.9263924	0.6519404	0.708041	0.6945171	0.7034361	0.7034361
4	Edad + género	Intercepto	0.8137688	0.8973236	0.5842706	0.6034779	0.5995323	0.5972239	0.5972239
5	Edad	Intercepto + Edad	0.863402	0.9266943	0.6542705	0.7091567	0.6959657	0.7045695	0.7045695
6	Edad	Intercepto	0.8153696	0.898296	0.5842706	0.6068863	0.6021813	0.600686	0.600686
7	Intercepto	Intercepto + Edad	0.8839817	0.9384186	0.6521794	0.7529748	0.7249113	0.7490786	0.7490786

Los estadísticos R^2 fueron calculados usando en el numerador $\hat{Y}_i = \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\hat{b}_i$

Tabla 5-5: Estadísticos R^2 condicionales evaluados en los datos desbalanceados de crecimiento Pathoff y Roy[12].

modelo	efectos fijos	efectos aleatorios	R_1^2	R_c	$\widehat{\Omega}^2$	R_2^2	$\widehat{\rho}^2$	R_{DG}^2	R_{DGP}^2
1	Edad + género +edad*género	Intercepto+ edad	0.8681996	0.9294506	0.6806271	0.7403232	0.7253565	0.7202088	0.6947732
2	Edad + género +edad*género	Intercepto	0.8227952	0.902784	0.6211161	0.6508662	0.6443861	0.6238224	0.5896244
3	Edad + género	Intercepto + Edad	0.8694504	0.9301668	0.6770288	0.7427875	0.7259786	0.7228639	0.6976697
4	Edad + género	Intercepto	0.8087004	0.8942337	0.5976335	0.6230962	0.6176716	0.5939013	0.5569833
5	Edad	Intercepto + Edad	0.868738	0.929759	0.6770043	0.741384	0.7250083	0.7213517	0.69602
6	Edad	Intercepto	0.810769	0.8954969	0.5975584	0.6271718	0.6207672	0.5982927	0.5617738
7	Intercepto	Intercepto + Edad	0.8924633	0.9431763	0.6776356	0.7881281	0.7555531	0.7717166	0.7509636

Los estadísticos R^2 fueron calculados usando en el numerador $\hat{Y}_i = \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\hat{b}_i$

6 Conclusiones y recomendaciones

6.1. Conclusiones

Cuando se evalúa el desempeño de los estadísticos R^2 como medidas de bondad de ajuste y como criterios de selección del mejor modelo que explique adecuadamente la variabilidad presente se obtuvieron los siguientes resultados:

- A través de las simulaciones mostradas en el capítulo 4 queda claro que el desempeño de los estadísticos condicionales propuestos por Xu[18], los propuestos por Vonesh y otros [17] y Vonesh y Chinchilli [16] y los propuestos en este trabajo no presentaron cambios en los valores medios al ser evaluados en un modelo subajustado o sobreajustado en relación al modelo verdadero cuando se adicionan o se quitan covariables de efectos aleatorios. Lo que pone en tela de juicio su utilidad como criterios de selección y medidas de bondad de ajuste en modelos lineales mixtos, cuando están mal especificadas las covariables de efecto aleatorio en el contexto de los datos longitudinales con variable respuesta continua.

- Orelien y Edwards [11] obtuvieron resultados similares. Ellos evalúan la utilidad de estos estadísticos asumiendo que los efectos aleatorios están bien especificados y ajustan modelos adicionado y quitando covariables de efectos fijos.

- El hecho de que los estadísticos R^2 estudiados no logren discriminar entre un modelo verdadero y uno mal especificado se debe básicamente, a que en el ajuste de un efecto aleatorio se incluyen muchos parámetros en el modelo, uno por cada individuo, por tanto es difícil medir su efecto. Teniendo en cuenta que la correlación intrasujeto dificulta esta labor, al existir correlación entre los efectos aleatorios. Por tanto, al incluir el intercepto aleatorio en un modelo, este contiene información de los otros efectos aleatorios importantes para el modelo, así la inclusión de un segundo efecto aleatorio no es tan significativa como se pudiera esperar, su efecto se ve reducido y es más difícil determinar si el segundo efecto aleatorio es importante para el modelo. Similarmente, la inclusión de los dos primeros efectos aleatorios contiene información de los demás efectos aleatorios importantes para el modelo, lo que reduce aún más el efecto de la inclusión del próximo efecto aleatorio; y así ocurre sucesivamente. Entonces el ajuste de los primeros efectos aleatorios puede incrementar tanto el valor del R^2 que al incluir

los demás no se alcanza a notar un cambio significativo en el R^2 lo que conduce a descartar efectos aleatorios importantes para el modelo.

- Si bien todos los estadísticos R^2 no cambian de modelo a modelo en relación al verdadero, una condición no deseable, se resalta el hecho que los estadísticos propuestos en este trabajo, poseen las propiedades principales que fueron descritas por Kvalseth [5] para medidas tipo R^2, es decir, poseen buenas propiedades estadísticas. Por ejemplo: son fáciles de interpretar, intuitivamente dan el porcentaje de la variación explicada por el modelo; tienen por cota superior el uno que es alcanzada cuando hay perfecto ajuste, si toman valores negativos o cero se tiene total falta de ajuste.

- Los estadísticos propuestos por Xu[18] y los estadísticos R^2_{DG} y R^2_{DGP} son cercanos al verdadero valor de Ω^2, el coeficiente de previsibilidad, que mide la proporción de la variación explicada por las covariables en un modelo. Es decir, que estos estadísticos son buenos estimadores de la proporción de la variabilidad de la variable respuesta que es explicada por el modelo.

- Cuando se consideraron las pequeñas variaciones de los estadísticos R^2 entre los modelos como significativas y sólo se analizan los modelos que no tienen sobreajuste de efectos aleatorios, los estadísticos propuestos en este trabajo alcanzaron a identificar adecuadamente los efectos fijos que eran importantes para el modelo en el caso de datos balanceados; y el R^2_{DGP} también logró identificar los efectos fijos importantes en el conjunto de datos desbalanceados.

- Los estadísticos R^2_{DG} y R^2_{DGP} son adecuados para aplicarlos en el marco de los modelos lineles con efectos mixtos, pero al igual que los otros estadísticos analizados en este trabajo no presentan ninguna ventaja como técnica para seleccionar el mejor modelo.

6.2. Recomendaciones

En futuros trabajos se deben desarrollar estadísticos que permitan identificar con mayor facilidad las covariables que debe incluir un modelo lineal con efectos mixtos para datos longitudinales. También puede resultar útil un procedimiento en el que a medida que se agregan covariables a un modelo, a estás se les vaya quitando el efecto lineal de las primeras que se incluyeron para evitar los problemas de correlación intraindividual en el ajuste del modelo.

Los estudios futuros pueden investigar el desempeño de los estadísticos R^2 para elegir la estructura de la media en modelos donde se fija una estructura de efectos aleatorios mal especificada.

A Algoritmo para el cálculo de los estadísticos R^2

Lectura de datos y graficos de perfiles

```
maderas=read.table('C:/Users/tosh/Desktop/Tesis formato harvard/madera12.txt',h=T)
attach(maderas)
grosor.f=factor(grosor)
tiempo.f=factor(tiempo)
library(lattice)
trellis.device(color=F)
library(nlme)
IR=length(tablon)
Nindividuos=100
matrizDatos.updated=data.frame(maderas,tiempo.f,grosor.f)
matrizDatosPI=cbind(matrizDatos.updated)

austism.g1=groupedData(peso~tiempo | tablon, outer=~ grosor.f,data=matrizDatos.updated)
plot(austism.g1, display = 'Tablón', outer = TRUE, aspect = 2, key = F, xlab = 'Tiempo',
ylab = 'Peso')
```

Modelos ajustados

```
library(nlme)
Datos.agrupados=groupedData(peso~tiempo+grosor | tablon, data=matrizDatos.updated,
order.groups = F)
lme.model.fit1=lme(peso~tiempo+grosor,random = ~ 1+tiempo, data =Datos.agrupados,
method = 'REML',control= list(opt='optim'))
Modelo1=summary(lme.model.fit1)

lme.model.fitnulo=lme(peso ~ 1,random = ~1, data =Datos.agrupados, method = 'REML',
control= list(opt='optim'))
```

```
Modelonulo=summary(lme.model.fitnulo)
```

Calculo de Estadísticos R2

```
Rcuadrados=function(modelo,modelonulo,IR,Nindividuos,k)
Nindividuos=Nindividuos
modelo=lme.model.fit1
modelonulo=lme.model.fitnulo
p1nulo=t(getResponse(lme.model.fitnulo)-predict(lme.model.fitnulo))
p2nulo=getResponse(lme.model.fitnulo)-predict(lme.model.fitnulo)
denominador3=sum(p1nulo %* %p2nulo)
p1p=(t(getResponse(lme.model.fit1)-predict(lme.model.fit1))*matrizDatosPI[,6])
p1=t(getResponse(lme.model.fit1)-predict(lme.model.fit1))
p2=getResponse(lme.model.fit1)-predict(lme.model.fit1)
numerador=sum(p1 %* %p2)
numeradorp=(Nindividuos/IR)*sum(p1p %* %p2)
p3=t(getResponse(lme.model.fit1)-mean(getResponse(lme.model.fit1))*matrix(1,IR,1))
p33=t(getResponse(lme.model.fit1)-matrizDatosPI[,5])
p4=getResponse(lme.model.fit1)-mean(getResponse(lme.model.fit1))*matrix(1,IR,1)
p44=getResponse(lme.model.fit1)-as.matrix(matrizDatosPI[,5],1)
denominador=sum(p3 %* %p4)
denominador1=sum(p33 %* %p44)
R2=1-(numerador/denominador)
Rcc=(2*R2)/(R2+1)
R_DG=1-(numerador/denominador1)
R_DGP=1-(numeradorp/denominador1)
R22=1-(numerador/denominador3)
Varmodel1=as.matrix(VarCorr(lme.model.fit1))
Varmodelnulo=as.matrix(VarCorr(lme.model.fitnulo))
SigmacuadradoEs=as.numeric(Varmodel1[k,1])
SigmacuadradoEsnulo=as.numeric(Varmodelnulo[2,1])
Romega=1-SigmacuadradoEs/SigmacuadradoEsnulo
rho=1-(SigmacuadradoEs/SigmacuadradoEsnulo)*exp((numerador/(IR*SigmacuadradoEs))-
(denominador3/(IR*SigmacuadradoEsnulo)))
Rtruemodel=cbind(R2,Rcc,Romega,R_DG,R_DGP,R22,rho)
return(Rtruemodel,deparse.level = 0)

Rcuadrados(lme.model.fit1,lme.model.fitnulo,IR,100,4)
```

Bibliografía

[1] Botero, J.G.: *Secado del ciprés (Cupressus lusitanica Mill) para usos industriales:estibas, Molduras y Muebles.* Medellín : Tesis Universidad Nacional, 1993.

[2] Cameron, A. C. ; Windmeijer, F.: An R^2 measure goodness-of-fit for some common nonlinear regression models.

[3] Cameron, A. C. ; Windmeijer, F.: R^2 measures for count data regression models with applications to health-care utilization. En: *American Statistical Association.* 14 (1996), p. 209–220

[4] Edwards, L. J. ; Others.: An R^2 statistics for fixed effects in the linear mixed models. En: *Statistics In Medicine.* 27 (2008), p. 6137–6157

[5] Kvalseth, T. O.: Cautionary note about R^2. En: *American Statistical Association.* 39 (1985), p. 279–285

[6] Laird, N. M. ; Ware, J. H.: Random-effects models for longitudinal data. En: *Biometrics.* 38 (1982), p. 963–974

[7] Little, R.J.A ; Rubin, B.D.: Statistical Analysis with Missing Data. En: *New York:Jonh Wiley Sons* (1987)

[8] Liu, Zheng Y. ; Shen, Jie.: Goodness-of-fit measures of R^2 for repeated measures mixed effect models. En: *Journal of Applied Statistics.* 35 (2008), p. 1081–1092

[9] Magee, Lonie.: R^2 measures based on Wald and likelihood ratio joint significance tests. En: *The Statistical Association.* 44 (1990), p. 250–253

[10] Nagelkerke, N. J. D.: A note on a general definition of the coefficient of determination. En: *Biometrika.* 78 (1991), p. 691–692

[11] Orelien, J. G. ; J., Edwards L.: Fixed-effect variable selection in linear mixed models using R^2 statistics. En: *Computational Statistics Data Analysis.* 52 (2007), p. 1896–1907

[12] Potthoff, S.N.: A generalized multivariate analysis of variance model useful especially for growth curve problems. En: *Biometrika.* 85 (1964), p. 313–326

[13] R Development Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2009. – ISBN 3-900051-07-0

[14] Thompson, Jr: The problem of negative estimates of variance components. En: *Annals of Mathematical Statistics.* 33 (1962), p. 273–289

[15] Verbeke, G. ; Molenberghs, G.: *Linear Mixed Models for Longitudinal Data*. London : New York, Springer-Verlag, 2000

[16] Vonesh, Chinchilli V. M.: Goodness-of-fit in generalized nonlinear models for the analysis of repeated measurements. En: *Marcel Dekker,New York.* (1997), p. 419–424

[17] Vonesh, Chinchilli V. M. ; Pu, K.: Goodness-of-fit in generalized nonlinear mixed-effects models. 52 (1996), p. 572–587

[18] Xu, Ronghui.: Measuring explained variation in linear mixed effects models. En: *Statistics In Medicine.* 22 (2003), p. 3527–3541.