UNIVERSIDAD NACIONAL DE COLOMBIA

Estudio comparativo de los métodos de diagnóstico para modelos lineales mixtos y modelos lineales generalizados

Trabajo presentado como requisito para optar al título de:
Magíster en Matemática Aplicada

Presentado por:
Andrés Felipe Morales Foronda

Directora:
PhD. Nubia Esteban Duarte
Profesora Departamento de Matemáticas y Estadística
Universidad Nacional de Colombia

Universidad Nacional de Colombia
Facultad de Ciencias Exactas y Naturales
Departamento de Matemáticas y Estadística
Manizales, Colombia
2022

Agradecimientos

En primera instancia agradezco a los miembros de mi familia, mi padre, mi tía y mi hermana, quienes me han acompañado y apoyado a lo largo de mi vida académica y profesional, gracias a ellos conseguí la motivación necesaria para cumplir y luchar por mis metas. Agradezco a mi madre, quien ya no se encuentra presente, por darme la oportunidad de alcanzar este momento.

Agradezco de forma especial a mi directora de tesis, la profesora Nubia Esteban Duarte, por su constante apoyo, disposición y dedicación durante el trayecto de mi maestría. Muchas gracias profesora, sin su acompañamiento no hubiese podido realizar este trabajo.

Por último agradezco a la Universidad Nacional de Colombia por su contribución a mi formación profesional, a los profesores que me han impartido su conocimiento en muchas áreas diversas, y a mis amigos y compañeros de estudio por su apoyo constante a lo largo de este proceso.

Resumen

Muchos fenómenos de la naturaleza pueden ser representados por medio de modelos estadísticos de forma satisfactoria y, para validar estos modelos, los métodos de diagnóstico resultan ser herramientas muy útiles para la verificación de un buen ajuste. La aplicación de los métodos de diagnóstico es relativamente sencilla para modelos de regresión lineal clásicos, sin embargo el proceso es más complicado cuando se consideran modelos más generales y con fuentes adicionales de variabilidad, como es el modelo lineal mixto o modelos con respuesta binaria o de conteo, como es el caso de modelos lineales generalizados y modelos lineales generalizados mixtos, que en general requieren el uso de técnicas de análisis de residuales y de sensibilidad más complejas.

En este trabajo se presentan diferentes estrategias relacionadas con el diagnóstico de modelos, introduciendo tanto los enfoques clásicos, que son habitualmente utilizados así como los enfoques más recientes. Las metodologías derivadas serán estudiadas para modelos lineales mixtos, modelos lineales generalizados y modelos lineales generalizados mixtos, enfatizando su utilización en diferentes aplicaciones.

Palabras Clave: Modelo Lineal Mixto, Modelo Lineal Generalizado, Modelo Lineal Generalizado Mixto, Residuales, Diagnóstico, Análisis de Influencia, Datos Longitudinales, Medidas Repetidas.

Abstract

Comparative study of diagnostic methods for linear mixed models and generalized linear models.

Many natural phenomena can be represented by means of statistical models in a satisfactory way and, to validate such models, diagnostic methods are very useful tools for the verification of a good fit. The application of diagnostic methods is relatively simple for classical linear regression models, however the process becomes more complicated when considering more general models with additional sources of variability, such as the linear mixed model or models with a binary or counting response as in the case of generalized linear models and generalized linear mixed models, which in general require the use of more complex residual and sensitivity analysis techniques.

In this paper different strategies related to model diagnostics are presented, introducing both classical approaches, which are commonly used as well as more recent approaches. The derived methodologies will be studied for linear mixed models, generalized linear models and generalized linear mixed models emphasizing their use in different applications.

Key words: Linear Mixed Model, Generalized Linear Model, Generalized Linear Mixed Model, Residuals, Diagnostic, Influence Analysis, Longitudinal Data, Repeated Measures.

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

1.1. Introducción

En los últimos años la cantidad de datos generados y disponibles ha crecido significativamente. Según la revista Forbes, en el año 2018 se produjeron más de 2.5 quintillones de bytes de datos a diario alrededor del mundo (Marr, 2018), esta cifra sólo va a seguir aumentando debido a la etapa de transformación digital en la que se encuentra el mundo moderno. De forma específica, se pueden mencionar investigaciones en Agronomía, Bioinformática, Clasificación de patrones, Procesamiento de señales, Ciencias Humanas, Genética, Econometría, entre muchas otras áreas que generan una gran cantidad de datos cuya interpretación y análisis hace necesaria y relevante la utilización de técnicas estadísticas adecuadas. En los grandes avances de la ciencia, la Estadística ha jugado un papel primordial en el desarrollo científico y tecnológico ya que provee herramientas metodológicas generales para analizar la variabilidad, determinar las relaciones entre variables, diseñar de forma óptima experimentos, mejorar las predicciones, calcular probabilidades y en la toma de decisiones en situaciones de incertidumbre. Del mismo modo, el desarrollo computacional también ha influido incrementando la capacidad para manejar información numérica.

En relación a los análisis estadísticos de datos en las diversas áreas, como las mencionadas anteriormente, existe un interés particular en los procedimientos estadísticos que modelan los datos en los que la respuesta se mide repetidamente en diferentes momentos de tiempo para los mismos individuos, este tipo de datos se conocen como datos longitudinales, los cuales son muy utilizados en estudios médicos y en las ciencias del comportamiento. Los datos longitudinales son un caso particular de los datos de medidas

repetidas, en donde cada grupo de observaciones se define de acuerdo a alguna característica, en particular que las medidas estén correlacionadas entre sí. De forma general, para modelar datos teniendo en cuenta su estructura de correlación y cuya variable respuesta es continua se utiliza el **modelo lineal mixto o MLM**.

Para modelar la estructura de correlación de cada grupo, un modelo lineal mixto incluye términos adicionales, llamados **efectos aleatorios**. Por otro lado, los **efectos fijos** son los parámetros desconocidos que describen a la población y se estiman a partir de los datos. En relación a la validación, para determinar si el modelo estimado es adecuado es muy importante realizar un análisis de diagnóstico, este comprende técnicas de análisis de residuales y de sensibilidad, las cuales ayudan a verificar las suposiciones del modelo y también validar las inferencias estadísticas que se hayan obtenido durante el estudio.

En un modelo lineal mixto clásico, los datos se pueden clasificar en dos niveles, el **nivel 1** hace referencia a cada observación registrada, mientras que el **nivel 2** representa a los grupos de medidas repetidas, cada uno de estos grupos se denomina como **unidad**, las unidades pueden tener un gran impacto sobre los resultados del modelo, así tengan sólo una observación influyente dentro, la unidad completa se puede identificar como atípica. Por otro lado, el análisis de residuales en los modelos mixtos es mucho más complejo que en los modelos lineales clásicos, las fuentes adicionales de variabilidad dan lugar a 3 tipos diferentes de residuales:

- **Residuales marginales**: Son los que sólo consideran el componente sistemático del modelo.
- **Residuales condicionales**: Son los que consideran ambos tipos de efectos, condicionados en los componentes de varianza estimados.
- **Residuales de efectos aleatorios**: Los cuales son inducidos por las estimaciones de los efectos aleatorios del modelo, y se usan para predecir los efectos aleatorios presentes en la población.

A partir de los residuales definidos, cada clase de residual cuenta con un conjunto de conceptos y técnicas de diagnóstico que los caracterizan y que pueden contribuir sobre las inferencias del modelo. Sin embargo, es necesario tener un conocimiento extensivo de la teoría que los fundamenta para aplicarlos correctamente. En el presente trabajo se presentará la formalización matemática de los métodos de diagnóstico más relevantes y prometedores en los modelos lineales mixtos.

Los modelos mixtos mencionados anteriormente también son conocidos como **modelos mixtos Gaussianos**, debido a la suposición de normalidad que está presente en la distribución de sus componentes. Esta distinción sugiere que los modelos mixtos también se pueden formular en casos donde no es apropiado que la respuesta y los efectos del modelo sigan una distribución normal. La generalización de los supuestos para estimar un modelo lineal es uno de los papeles que cumple el **modelo lineal generalizado o MLG**, el cual fue introducido por Nelder y Wedderburn (1972), y desarrollado posteriormente por varios autores. Según Tellez y Morales (2016), los modelos lineales generalizados proporcionan una aproximación unificada a la mayoría de los procedimientos usados en estadística aplicada y los autores mencionan que es desafortunado que la mayoría de textos y cursos de estadística se concentran en el modelo lineal clásico, pues muchos estudiantes terminan con una visión muy restringida de las aplicaciones estadísticas. Por otro lado, Vallejo, Ato García et al. (2012) mencionan que los modelos lineales generalizados suponen "una auténtica revolución estadística", y recomiendan el uso de estos modelos para trabajar con cualquier tipo de variable, en especial las que no sigan una distribución normal. De hecho, el modelo lineal clásico es un caso particular del MLG. Dado que el MLG considera distribuciones de probabilidad que no son normales, es importante estudiar los métodos de diagnóstico disponibles que permitan validar sus resultados. En este trabajo se busca estudiar y representar las técnicas de diagnóstico en modelos mixtos y en MLGs, y también estudiar técnicas de diagnóstico que se pueden aplicar en los modelos lineales generalizados mixtos.

1.2. Antecedentes

El fundamento de la mayoría de las pruebas estadísticas se encuentra en la formulación del modelo de regresión lineal clásico ya que su estructura refleja los elementos explicativos de un fenómeno mediante las relaciones funcionales probabilísticas entre variables. Muchos autores han abarcado esta área en particular, como ejemplo se pueden consultar los textos de Weisberg (2005), Tellez y Morales (2016), y Montgomery et al. (2021). Sin embargo, a medida que se desarrolla más la teoría y los problemas prácticos se vuelven más complejos, el modelo lineal clásico empieza a presentar muchas limitaciones en su aplicación, e incluso en algunos casos (por ejemplo, cuando se considera una respuesta categórica) no es el adecuado para modelar el problema. Por lo tanto, surgió la necesidad de extender el modelo lineal con el objetivo de considerar muchos más tipos de respuestas, además de

que las suposiciones que se hacen sean más generales. El modelo que mejor cumple este cometido es el modelo lineal generalizado o MLG, el cual fue introducido por Nelder y Wedderburn (1972), y desarrollado posteriormente por varios autores.

Por otro lado, el modelo clásico de regresión presenta numerosas restricciones en relación a la distribución de los datos, en particular, las medidas de la variable respuesta deben ser independientes entre sí, además deben tener varianza homogénea y constante. Faraway (2016c) expresa que el supuesto de independencia es una de las suposiciones más importantes que debe cumplir un modelo de regresión, y este supuesto no se cumple en los estudios donde la variable respuesta presenta medidas repetidas para una cantidad determinada de individuos.

Los modelos lineales mixtos conforman la aproximación por excelencia a los problemas con datos de medidas repetidas. Su teoría se empezó a desarrollar con el trabajo de Laird y Ware (1982), en donde presentaron una extensión del modelo lineal clásico en el cual se estiman efectos aleatorios, los cuales describen la variabilidad causada por la correlación entre observaciones, la cual es producto de la estructura innata de los datos. Los primeros modelos mixtos considerados fueron los modelos de intercepto aleatorio, y de intercepto y pendiente aleatorios. Para su estimación se utiliza el método de máxima verosimilitud restringida o REML, el cual fue introducido inicialmente por Patterson y Thompson (1971) como una forma de estimar los componentes de varianza en un MLG.

A pesar de que los modelos lineales mixtos son muy flexibles para los casos de medidas repetidas y datos longitudinales, estos también necesitan satisfacer ciertas condiciones sobre la distribución de los datos que se están utilizando. La suposición de normalidad en los errores y en los efectos aleatorios es indispensable para que la estimación por máxima verosimilitud, el cual es el método comúnmente utilizado para estimar los parámetros y los efectos de un modelo mixto, tenga sentido. Por este motivo es importante contar con un conjunto de herramientas de diagnóstico que permitan validar los resultados obtenidos por el analista, y en caso de que los supuestos del modelo no se cumplan, recolectar información suficiente para seleccionar el modelo que mejor se ajuste a las necesidades del estudio realizado.

Debido a que el componente aleatorio de un modelo lineal determina su distribución, la mayoría de técnicas desarrolladas para análisis de diagnóstico se han centrado en el análisis de los residuos del modelo. A diferencia del modelo lineal clásico, un modelo lineal mixto tiene 3 tipos de residuales distintos, por lo tanto, ha sido necesario desarrollar nuevas técnicas y conceptos que permitan generalizar y complementar las técnicas de diagnóstico clásicas

para modelos lineales, y que sean apropiadas en el contexto de los modelos mixtos. Unos componentes muy importantes del diagnóstico son el análisis de sensibilidad y de influencia, los cuales se concentran respectivamente en medir los cambios que presenta el modelo al introducir perturbaciones en los datos y en detectar observaciones que pueden cambiar en gran medida la estimación de los parámetros y efectos aleatorios del modelo.

La investigación referente a las técnicas de diagnóstico en modelos mixtos ha sido extensa, uno de los primeros registros del desarrollo de estos métodos lo presentan Beckman et al. (1987), en donde desarrollaron un método de evaluación de supuestos para modelos mixtos que se enfoca en los efectos de las perturbaciones en los datos que modifican las suposiciones de varianza constante y normalidad en los efectos aleatorios. Las perturbaciones más estudiadas en la época para este tipo de modelos fueron la eliminación de casos, pues los métodos estándar de estimación eran muy sensibles a observaciones atípicas. Christensen et al. (1992) estudiaron las técnicas de diagnóstico en modelos mixtos enfocados principalmente en la eliminación de casos para la detección de observaciones influyentes, en particular obtuvieron una generalización de la distancia de Cook, una medida de influencia muy utilizada en regresión clásica. Los autores además indicaron que la matriz de covarianza en un modelo mixto es lineal en sus parámetros y desarrollaron métodos que tomaban ventaja de este hecho para obtener técnicas de diagnóstico útiles en la práctica. Un estudio similar fue realizado por Banerjee y Frees (1997), en donde analizaron y extendieron los diagnósticos de influencia obtenidos anteriormente, y se enfocaron principalmente en la eliminación de unidades de los datos.

Alrededor del año 2000, empezaron a aparecer rutinas computacionales que facilitaban el ajuste de modelos mixtos. Uno de los primeros paquetes que se desarrollaron fue el procedimiento MIXED del programa de análisis estadístico SAS (Littell et al., 2006). Un par de años antes, Lesaffre y Verbeke (1998) hicieron uso de este procedimiento para aplicar los métodos de análisis de influencia obtenidos hasta el momento en un conjunto de datos relacionado al cáncer de próstata. Tan et al. (2001) mostraron que la distancia de Cook puede fallar al detectar observaciones influyentes debido a las varianzas y covarianzas de los efectos aleatorios, por lo tanto, los autores propusieron una distancia de Cook condicional, la cual buscaba remediar este problema. Años después, Demidenko y Stukel (2005) extendieron muchas de las técnicas de diagnóstico en regresión como los conceptos de leverage, influencia local e infinitesimal y distancias de Cook a los modelos lineales mixtos. Cada medida fue presentada con una interpretación directa en términos de los efectos de los parámetros de interés, y su definición contribuyó

a que el modelo necesitara menos re-estimaciones, una ventaja significativa en el tiempo de cómputo para los diagnósticos.

Los diferentes residuales del modelo lineal mixto se empezaron a estudiar un tiempo después con el trabajo de Nobre y Singer (2007). Los autores exploraron los 3 tipos de residuales, e hicieron una revisión de las técnicas de análisis residual conocidas hasta el momento para proponer una estandarización de los residuales condicionales, con el objetivo de encontrar observaciones atípicas y cúmulos de observaciones que pueden influir en las estimaciones del modelo mixto. Schützenmeister y Piepho (2012) realizaron un estudio de simulación en el que evaluaron el análisis residual de los modelos mixtos en diferentes bancos de datos, los autores introdujeron el uso de gráficos de residuales para verificar las suposiciones del modelo mediante la comparación de distribuciones residuales empíricas con distribuciones nulas apropiadas, apoyados en la técnica del bootstrap paramétrico. Los gráficos propuestos pueden ayudar a revisar la normalidad y la homocedasticidad de los residuos, además de que pueden encontrar posibles observaciones atípicas.

En contraste con el análisis de influencia en modelos mixtos, son pocos los trabajos que exponen y desarrollan el análisis de residuales para este tipo de modelos. Sin embargo se han realizado algunas revisiones de la literatura que buscan unificar la notación usada para definir los modelos y sus residuos y recopilar todos los métodos disponibles hasta el momento, para facilitar su uso en la práctica. Como ejemplo se puede revisar el trabajo de Loy y Hofmann (2013).

1.3. Justificación

Los modelos lineales mixtos constituyen una poderosa herramienta inferencial que se utiliza para el análisis de datos que tienen estructura de correlación. La complejidad de estos modelos, por la presencia de efectos aleatorios, hace que la elección del mismo haya sido objeto de investigación en los últimos tiempos. La elección de un modelo adecuado para los datos involucra la utilización de criterios de bondad de ajuste y del uso de herramientas de diagnóstico para evaluar la existencia de alguna deficiencia en el modelo estimado.

Se justifica su uso ya que en muchos casos hay poca o ninguna base teórica disponible para sugerir la forma específica de cómo las variables se relacionan. Estos métodos pueden ayudar a mejorar la elección del modelo y a identificar valores atípicos o unidades influyentes que merecen una mayor

atención por parte del investigador. Aunque el modelo mixto ofrece gran flexibilidad para modelar la correlación dentro de la unidad, frecuentemente presente en datos con medidas repetidas, sufre de la misma falta de robustez frente a observaciones atípicas que otros modelos estadísticos basados en la distribución Normal. En este trabajo se presentan diferentes estrategias que ayudan al ajuste de modelos, introduciendo tanto los enfoques clásicos, que son habitualmente utilizados, como los de reciente aparición. Las metodologías descritas derivadas para modelos lineales mixtos serán estudiadas, así como las metodologías asociadas a modelos lineales generalizados enfatizando su utilización en diferentes aplicaciones.

Respecto a la viabilidad del proyecto, la mayoría de software disponible para la estimación de modelos mixtos es de libre acceso, se hace énfasis en el lenguaje R (R Core Team, 2021), el cual es un lenguaje de programación gratuito orientado a la estadística. La gran cantidad de rutinas y métodos implementados en el programa hacen muy conveniente el ajuste y evaluación de los modelos mixtos que se van a estudiar.

1.4. Objetivos

1.4.1. Objetivo principal

Relacionar matemáticamente las técnicas de diagnóstico en modelos lineales mixtos y modelos lineales generalizados ilustrando la teoría mediante ejemplos prácticos.

1.4.2. Objetivos específicos

- Establecer de forma analítica los métodos de diagnóstico para modelos lineales mixtos en general y de forma específica para modelos con intercepto aleatorio o modelos con pendiente aleatoria.
- Presentar una base teórica sobre los métodos de diagnóstico, para los modelos lineales mixtos, así como también para modelos lineales generalizados y modelos lineales generalizados mixtos.
- Utilizar el programa R para realizar las aplicaciones que ilustran la teoría en un escenario de interés práctico.

Capítulo 2

Métodos de diagnóstico en modelos lineales mixtos

Las estructuras de datos anidados ocurren naturalmente a partir de las numerosas formas de recolección de datos que existen. Algunas veces aparece una agrupación porque se realizaron medidas repetidas en el mismo individuo, particularmente, estas medidas se pueden registrar a lo largo de diferentes instancias de tiempo; este tipo de medidas son las que se consideran en los estudios de datos longitudinales, los cuales se aplican en una gran variedad de áreas de las ciencias. Un interés común es el de trabajar este tipo de datos con un modelo de regresión.

En un estudio de medidas repetidas, los datos comparten la característica común de correlación de las observaciones dentro de un mismo grupo, por lo que los modelos que asumen independencia de las observaciones no van a ser apropiados. Es necesario proveer un modelo adecuado para la estructura correlacional de las medidas repetidas con el fin de evitar inferencias engañosas sobre los parámetros de interés al investigador.

Para modelar esta correlación han surgido diferentes técnicas a lo largo de los años como las series de tiempo, o los modelos ANOVA y VARCOMP (componentes de varianza), pero estos métodos se quedan cortos cuando se incluyen variables independientes adicionales en el modelo y se necesita realizar las interpretaciones e inferencias correspondientes. Los modelos de efectos mixtos se pueden utilizar para este propósito pues incluyen la correlación dentro del modelo utilizando componentes aleatorios para cada unidad a partir de una descomposición del término de error.

Los términos adicionales del modelo que describen la estructura correlacional de los datos se conocen como **efectos aleatorios**, los cuales son

variables aleatorias que no se pueden estimar directamente, sin embargo, sí se pueden estimar los parámetros que describen su distribución. Por otro lado, los parámetros desconocidos que describen las características de la población de interés, y que se tratan de estimar a partir de los datos son los **efectos fijos**. Un modelo mixto combina elegantemente los efectos fijos y los aleatorios para describir el problema en cuestión.

Faraway (2016a) expone un ejemplo que permite entender cuándo se deben utilizar efectos aleatorios: Se considera un experimento en el que se investiga el efecto de varios tratamientos de una droga particular en una muestra de pacientes, usualmente se presenta interés en tratamientos de droga específicos, por lo que los efectos de la droga se pueden tratar como fijos. Sin embargo, tiene más sentido considerar los efectos de los pacientes como aleatorios, pues es razonable tratar a los pacientes como si se hubiesen seleccionado al azar de una colección más grande de pacientes, de los cuales se necesita hacer estimaciones de sus características. El interés del estudio no recae en los pacientes específicos, sino en la población entera de pacientes, por lo que el enfoque de efectos aleatorios intenta decir algo acerca de la población más grande, más allá de la muestra particular.

En el presente capítulo se define el modelo lineal mixto clásico o modelo multinivel de 2 niveles. En la sección 2.2 se resumen los métodos de estimación más comunes, y en la sección 2.3 se expone un resumen detallado de la mayoría de métodos de diagnóstico que han sido propuestos para esta clase de modelos en los últimos años.

2.1. Definición del modelo lineal mixto

Dado un conjunto de n observaciones, un modelo de regresión lineal simple se puede representar de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \forall i = 1, \ldots, n$$

Al considerar un efecto aleatorio, se debe introducir un nuevo subíndice en la ecuación del modelo, el cual debe recorrer los N grupos o **unidades** que se obtuvieron a partir de la recolección de los datos. El primer modelo mixto que se obtiene es el **modelo de intercepto aleatorio**. Si sólo se considera una variable regresora x, se puede modelar la j-ésima respuesta de la i-ésima unidad como sigue:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \varepsilon_{ij} \qquad \text{para } i = 1, \ldots, N \text{ y } j = 1, \ldots, n_i,$$

donde n_i es el número de observaciones en la unidad i, para un total de $N_T = \sum_{i=1}^{N} n_i$ observaciones y β_0, β_1 son los efectos fijos del modelo. En este caso, se particiona el residual total en un componente aleatorio específico para cada unidad, denotado por b_i, más un término de error aleatorio ε_{ij}. Es importante notar que ε_{ij} es diferente al término de error del modelo de regresión lineal simple, a saber ε_i. Por otro lado, b_i describe el intercepto aleatorio para la recta de regresión ajustada a los datos de la unidad i, y es constante sobre los valores de la variable regresora x.

Además se asume que los interceptos aleatorios b_i y los términos de error tienen media cero y no están correlacionados. Si $Var(b_i) = \sigma_b^2$ y $Var(\varepsilon_{ij}) = \sigma^2$ se obtiene que la varianza de cada medida repetida está dada por:

$$Var(b_i + \varepsilon_{ij}) = \sigma_b^2 + \sigma^2$$

El primer sumando es el componente entre unidades, mientras que el segundo es el componente dentro de cada unidad, por lo tanto, un modelo mixto también se conoce como un **modelo de componentes de varianza**. Las varianzas anteriores inducen una correlación entre observaciones que se encuentran en el mismo nivel como:

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$$

el cual se conoce como **coeficiente de correlación intraclase**, y describe la proporción de la varianza residual total que se debe a la variabilidad residual entre unidades. Si $\rho = 0$, no habría variabilidad entre unidades, en contraste si $\rho = 1$, los predictores y las unidades describen completamente a la variable respuesta. Luego, un modelo de intercepto aleatorio restringe a la varianza de cada medida repetida para que sea la misma, y la covarianza entre cualquier par de medidas repetidas para que sea igual. Este supuesto se conoce como **estructura de simetría compuesta**.

El modelo de intercepto aleatorio asume que no se presentan diferencias en las pendientes de las rectas de regresión de las unidades del modelo, esta suposición puede llegar a ser poco realista, por lo que se puede agregar un efecto aleatorio específico a cada unidad el cual permita diferencias en las pendientes de las rectas regresoras. El siguiente modelo obtenido es el **modelo de intercepto y pendiente aleatorios**, y se denota de la siguiente forma:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + b_{1i} + b_{2i} x_{ij} + \varepsilon_{ij} \qquad \text{para } i = 1, \ldots, N \text{ y } j = 1, \ldots, n_i,$$

donde b_{1i} y b_{2i} representan el intercepto y la pendiente aleatoria respectivas, correspondientes a las observaciones que pertenecen a la unidad i. En este

caso, la varianza del vector aleatorio $b_i = (b_{1i}, b_{2i})^\top$ se representa mediante la matriz:

$$Var(b_i) = \begin{pmatrix} \sigma_{b_{11}} & \sigma_{b_{12}} \\ \sigma_{b_{12}} & \sigma_{b_{22}} \end{pmatrix},$$

donde $\sigma_{b_{11}}$ representa la varianza asociada al intercepto aleatorio de las unidades, $\sigma_{b_{22}}$ representa la varianza asociada a la pendiente aleatoria de la variable regresora y $\sigma_{b_{12}}$ es la covarianza asociada al intercepto aleatorio y la pendiente aleatoria.

El modelo de intercepto y pendiente aleatorios se puede escribir en forma matricial como sigue:

$$y_i = \begin{bmatrix} 1_{n_i} & x_i \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 1_{n_i} & x_i \end{bmatrix} \begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix} + \varepsilon_i \qquad \text{para } i = 1, \ldots, N,$$

donde y_i es el vector $n_i \times 1$ de respuestas en la unidad i, $[\beta_0\ \beta_1]^\top$ es el vector 2×1 de efectos fijos, $[b_{1i}\ b_{2i}]^\top$ es el vector 2×1 de efectos aleatorios, 1_i es un vector de unos de tamaño $n_1 \times 1$, $[1_i\ x_i]$ es la matriz $n_i \times 2$ de diseño para los efectos fijos y aleatorios y ε_i es el vector $n_i \times 1$ de términos de error para las observaciones de la unidad i.

El modelo anterior se puede generalizar para un número de p efectos fijos y q efectos aleatorios correspondientes, de donde se obtiene el **modelo lineal mixto o MLM**:

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i \qquad \text{para } i = 1, \ldots, N,$$

donde y_i es el vector $n_i \times 1$ de respuestas en la unidad i, X_i es la matriz $n_i \times p$ de diseño de los efectos fijos, β es el vector $p \times 1$ de parámetros de la población o efectos fijos, Z_i es la matriz de diseño de los efectos aleatorios, b_i es el vector $q \times 1$ de efectos aleatorios y ε_i es el vector $n_i \times 1$ de términos de error para la unidad i. Se supone que $E(\varepsilon_i) = 0$, $E(b_i) = 0$, y que $Var(\varepsilon_i) = \sigma^2 R_i$, $Var(b_i) = \sigma^2 D$, donde $R_i^* = \sigma^2 R_i$ y $D^* = \sigma^2 D$ son las matrices de covarianza de tamaño $n_i \times n_i$ y $q \times q$ de los errores y los efectos aleatorios respectivamente. Adicionalmente se supone que los vectores b_i y ε_i son independientes para todo $i = 1, \ldots, N$.

Luego para cada unidad, $E(y_i) = X_i\beta$ y $Var(y_i) = \Omega_i = R + Z_i D Z_i^\top$. En la práctica usualmente se considera que la varianza de los errores es constante dentro de cada grupo, es decir, $Var(\varepsilon_i) = \sigma^2 I_{n_i}$, además, se considera una distribución normal multivariada para los efectos aleatorios y los errores, es decir:

$$\varepsilon_i \sim N(0, \sigma^2 I) \qquad \text{y} \qquad b_i \sim N(0, \sigma^2 D)$$

Por lo tanto, el modelo lineal mixto se puede expresar de la siguiente manera:

$$y_i \sim N(X_i\beta,\ \sigma^2(I + Z_i D Z_i^\top)) \qquad \text{para } i = 1, \ldots, N \tag{2.1}$$

El modelo lineal mixto se puede escribir de forma aún más compacta considerando todas las unidades disponibles. Sean $y = (y_1^\top, \ldots, y_N^\top)^\top$, $\varepsilon = (\varepsilon_1^\top, \ldots, \varepsilon_N^\top)^\top$ dos vectores de tamaño $N_T \times 1$, $b = (b_1^\top, \ldots, b_n^\top)^\top$ un vector de tamaño $Nq \times 1$, y $X = (X_1^\top, \ldots, X_N^\top)^\top$, $Z = \oplus_{i=1}^N Z_i$ dos matrices de tamaños $N_T \times p$ y $N_T \times q$ respectivamente. Luego, las N ecuaciones en (2.1) se pueden escribir como:

$$y = X\beta + Zb + \varepsilon \tag{2.2}$$

Por lo tanto, $E(y) = X\beta$ y $Var(y) = V = Z\Gamma Z^\top + R$, donde $\Gamma = I_n \otimes D^*$ y $R = \oplus_{i=1}^n R_i^*$.

2.2. Estimación e Inferencia

2.2.1. Estimación del modelo

En general, si Γ y R son conocidas, se pueden obtener los mejores estimadores lineales insesgados (BLUE) de β y el mejor predictor lineal (BLUP) de b al resolver las ecuaciones del modelo mixto de Henderson (Henderson, 1975), las cuales están dadas por:

$$\begin{pmatrix} X^\top R^{-1} X & X^\top R^{-1} Z \\ Z^\top R^{-1} X & Z^\top R^{-1} Z + \Gamma^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} X^\top R^{-1} y \\ Z^\top R^{-1} y \end{pmatrix}$$

A partir de las cuales se obtienen los estimadores,

$$\hat{\beta} = (X^\top V^{-1} X)^{-1} X^\top V^{-1} y$$
$$\hat{b} = \Gamma Z^\top V^{-1} (y - X\hat{\beta}) = \Gamma Z^\top Q y$$

donde la matriz Q está dada por:

$$Q = V^{-1} - V^{-1} X (X^\top V^{-1} X)^{-1} X^\top V^{-1}$$

En este caso, $\hat{\beta}$ coincide con el estimador de mínimos cuadrados generalizados (GLS) de β, sin embargo, en la práctica la matriz D, y por lo tanto la matriz Γ son desconocidas, y en muchas ocasiones el propósito del análisis es la estimación de los componentes de varianza (Faraway, 2016a). Por tanto es necesario estimar simultáneamente los elementos de β y θ, donde

θ es el vector de componentes de varianza, el cual contiene a σ^2 y a todos los términos de la matriz D. El método de máxima verosimilitud (ML) se puede utilizar para este propósito.

El MLM está dado por:

$$y_i \sim N(X_i\beta,\ \sigma^2(I + Z_i D Z_i^\top)) \qquad \text{para } i = 1, \ldots, N,$$

donde $\varepsilon_i \sim N(0, \sigma^2 I)$, $b_i \sim N(0, \sigma^2 D)$ y los vectores $\varepsilon_1, \ldots, \varepsilon_N$, $b_1, \ldots, b_q$ son independientes.

Sea $V_i = I + Z_i D Z_i^\top$ la matriz de covarianza escalada de y_i. Luego la función de verosimilitud de los datos está dada por:

$$\begin{aligned} L(\beta, \theta) &= \prod_{i=1}^{N} (2\pi\sigma^2)^{-n_i/2} |\sigma^2 V_i|^{-1/2} e^{-\frac{1}{2\sigma^2}(y_i - X_i\beta)^\top V_i^{-1}(y_i - X_i\beta)} \\ &= (2\pi\sigma^2)^{-N_T/2} |\sigma^2 V|^{-N/2} e^{-\frac{1}{2\sigma^2}(y - X\beta)^\top V^{-1}(y - X\beta)} \end{aligned}$$

Por tanto, al tomar logaritmo natural se obtiene la función de log-verosimilitud:

$$\ell(\beta, \theta) = -\frac{N_T}{2}\log(2\pi\sigma^2) - \frac{N}{2}\log|\sigma^2 V| - \frac{1}{2\sigma^2}(y - X\beta)^\top V^{-1}(y - X\beta)$$

La función $\ell(\beta, \theta)$ se puede maximizar para obtener los estimadores de máxima verosimilitud de β, σ^2 y D. Optimizar la función de log-verosimilitud es difícil en la práctica, ya que debido a la presencia de los efectos aleatorios la función resulta ser no lineal, y se puede complicar la estimación de los parámetros de modelos que tengan un gran número de efectos aleatorios. Con la excepción de algunos casos particulares, es necesario usar procedimientos iterativos como los que están basados en los **algoritmos EM, Newton-Raphson o Fisher Scoring** para obtener los EMV de θ.

Los efectos aleatorios en el modelo (2.1) son variables aleatorias, por tanto es natural estimarlos con técnicas Bayesianas, en este caso se usan las esperanzas condicionales de los efectos aleatorios, dadas las respuestas observadas y_i:

$$\hat{b}_i = \hat{D} Z_i^\top \hat{V}_i^{-1}(y_i - X_i\hat{\beta}) \tag{2.3}$$

Las esperanzas condicionales se conocen como *estimaciones de Bayes Empíricas*, sin embargo, como los efectos aleatorios no son parámetros, en realidad se están prediciendo los valores de b_i. Por otro lado, se puede mostrar que las esperanzas condicionales son BLUPs de b_i, luego los valores en (2.3) también se conocen como *EBLUPS*, o mejores predictores lineales insesgados empíricos.

Adicionalmente, algunas veces el EMV de un parámetro de varianza puede ser cero (o muy cercano a cero) lo cual ocurre en la frontera de su dominio. Puede que la derivada de la verosimilitud no sea cero en este estado de frontera lo cual causa problemas para los métodos de optimización mencionados anteriormente. Según Bates et al. (2015), en la documentación del paquete *lme4* del programa R, los modelos singulares están bien definidos estadísticamente, pero existen preocupaciones de que:

1. Los ajustes singulares correspondan a modelos sobreajustados que tengan poco poder.
2. Pueden ocurrir problemas numéricos y de no convergencia para los modelos singulares.
3. Los procedimientos estándar de inferencia, como los estadísticos de Wald y los tests de razón de verosimilitud pueden ser inapropiados.

Dadas las anteriores observaciones, es necesario tomar precauciones cuando se obtiene un ajuste singular del modelo. Por otro lado, los estimadores de MV de los componentes de varianza tienden a ser sesgados (Diggle et al., 2002). Para reducir el sesgo en el modelo de componentes de varianza, Patterson y Thompson (1971) propusieron el método de **Máxima verosimilitud restringida (residual) o REML** (por sus siglas en inglés), el cual consiste en encontrar todas las combinaciones lineales independientes de la respuesta, k, tales que $k^\top X = 0$. Luego se forma la matriz K con columnas k, tales que

$$K^\top y \sim N(0, K^\top V K),$$

después se maximiza la verosimilitud basada en $K^\top y$, la cual no depende de los efectos fijos. Este método generalmente produce estimadores menos sesgados de los componentes de varianza, por lo tanto es uno de los métodos más usados en la práctica.

Para reducir la cantidad de cálculos realizados en la estimación de ML o REML, se han desarrollado formas adicionales de parametrizar la función de verosimilitud $\ell(\beta, \theta)$, como las parametrizaciones de **reducción de dimensión**, de **verosimilitud de perfil** y de D **inversa**, las cuales son presentadas en Demidenko (2013).

Además de los métodos basados en máxima verosimilitud, se han desarrollado técnicas de estimación adicionales como el método de los momentos, el MINQUE (estimación insesgada cuadrática de norma mínima), y los estimadores de ANOVA. En los últimos años también se han desarrollado métodos de estimación Bayesiana, de los cuales se pueden encontrar más detalles en Dey et al. (2000), Demidenko (2013), y Correa y Salazar (2016).

2.2.2. Inferencia en el modelo lineal mixto

Dado que los parámetros del modelo se estiman por máxima verosimilitud, uno de los procedimientos más usados para probar los efectos fijos es el **test de razón de verosimilitud o LRT**, el cual se usa para comparar modelos encajados. Sea H_0 la hipótesis que representa al modelo que no contiene el componente de interés, y sea H_1 la hipótesis que representa al modelo que sí lo incluye, y que deja igual al resto de componentes, en este caso, el estadístico de razón de verosimilitud toma la forma:

$$\hat{\Lambda} = 2(\ell(\hat{\beta}_1, \hat{\theta}_1 \mid y) - \ell(\hat{\beta}_0, \hat{\theta}_0 \mid y)),$$

donde $\hat{\beta}_0$, $\hat{\theta}_0$ son los vectores de parámetros de máxima verosimilitud del modelo bajo la hipótesis nula, y $\hat{\beta}_1$, $\hat{\theta}_1$ son los vectores de parámetros correspondientes a la hipótesis alternativa.

Si los modelos a comparar fueron ajustados con el método de REML, entonces sólo se puede usar el test de razón de verosimilitud en el caso que ambos tengan los mismos efectos fijos (Longford, 1995). El procedimiento de REML estima los efectos fijos considerando combinaciones lineales de los datos que remueven los efectos fijos, si los efectos fijos cambian, las verosimilitudes de los dos modelos no van a ser comparables. Por este motivo, Faraway (2016a) recomienda usar el método de máxima versomilitud estándar si es necesario investigar la significancia de los efectos fijos con el LRT.

La distribución del LRT es aproximadamente una chi-cuadrada, sin embargo, este test requiere que se cumpla un conjunto de suposiciones estrictas, además en este caso la prueba tiende a ser conservativa, por lo que los p-valores van a ser más grandes de lo que deberían. El p-valor generado por la aproximación de la chi-cuadrada no es confiable, por lo que Faraway (2016a) sugiere no utilizar esta aproximación. Sin embargo, debido a que se supone una distribución normal para el modelo, se puede utilizar el procedimiento de **bootstrap paramétrico** para hallar p-valores más precisos para el LRT. En este método se generan datos bajo la hipótesis nula usando las estimaciones de los parámetros ajustados, se calcula el estadístico de la razón de verosimilitud para estos datos generados, y el procedimiento se repite muchas veces para juzgar la significancia del estadístico observado.

Al ajustar un modelo lineal mixto en R, el paquete *nlme* va a entregar p-valores para cada efecto fijo del modelo, estos están basados en el test F para efectos fijos que se utiliza en los modelos estándar de regresión lineal: Si Ω representa al modelo con dimensión p, y si ω representa al modelo con

dimensión $q < p$, el siguiente estadístico tiene una distribución F con $p - q$ y $N_T - p$ grados de libertad bajo la hipótesis nula:

$$F_0 = \frac{(RSS_\omega - RSS_\Omega)/(p-q)}{RSS_\Omega/(N_T - p)} \sim F_{p-q, N_T - p}$$

El p-valor se calcula como la probabilidad de encontrar un valor mayor que F_0 con la distribución F. Sin embargo, este método presenta complicaciones cuando se traslada a los modelos mixtos. Según Faraway (2016a), la presencia de efectos aleatorios opacan la definición de los grados de libertad, además el estadístico F_0 no necesariamente va a tener una distribución F, por lo tanto los p-valores tienden a ser incorrectos. Por este motivo, el paquete *lme4* (Bates et al., 2015) no muestra los p-valores al ajustar un MLM.

Es importante determinar si es necesaria la inclusión de los componentes de varianza, Demidenko (2013) describe una prueba de hipótesis para determinar si todos los efectos aleatorios son relevantes, en este caso la hipótesis nula toma la forma:

$$H_0 \colon D = 0$$

Si la hipótesis nula es verdadera, la diferencia entre la suma de cuadrados mínima con los efectos aleatorios, S_{min}, y la suma de cuadrados mínima sin los efectos aleatorios, S_{OLS} debería ser pequeña.

Sea $r = rank(W)$, donde

$$W = [X, Z] = \begin{bmatrix} X_1 & Z_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_N & 0 & 0 & \cdots & Z_N \end{bmatrix}$$

entonces si $D = 0$, Demidenko (2013) demuestra que el siguiente cociente de formas cuadráticas tiene una distribución F:

$$F_0 = \frac{(S_{OLS} - S_{min})/(r-p)}{S_{min}/(N_T - r)} \sim F(r - p, N_T - r) \tag{2.4}$$

Cuando los efectos aleatorios están presentes en el modelo dado por (2.1), S_{min} debería ser relativamente pequeño, y por lo tanto el cociente en (2.4) se vuelve grande. Luego se rechaza $H_0 \colon D = 0$ si $F_0 > F_{\alpha, r-p, N_T - r}$.

También se pueden realizar inferencias sobre cada uno de los componentes de varianza. Si $\sigma_{b_{ij}}$ con $i, j = 1, \ldots, q$ denota a un elemento en D, se puede probar la hipótesis

$$H_0 \colon \sigma_{b_{ij}} = 0$$
$$H_1 \colon \sigma_{b_{ij}} > 0$$

El espacio de parámetros para un componente de varianza es el intervalo abierto $(0, \infty)$, por tanto, las hipótesis se prueban en la frontera del espacio. Liu (2015) recomienda una mezcla 50:50 de dos distribuciones chi-cuadradas para el estadístico de la razón de verosimilitud o LRS, luego el p-valor ajustado del LRT para un componente de varianza ($q = 1$) está dado por:

$$pv = \begin{cases} 1 & \text{si } \hat{\Lambda} = 0 \\ 0.5P(\chi_1^2 > \hat{\Lambda}) & \text{si } \hat{\Lambda} > 0 \end{cases} \tag{2.5}$$

donde $\hat{\Lambda}$ es el valor observado del LRS. Esta prueba se puede generalizar para cualquier número de componentes de varianza, si hay 2 parámetros de efectos aleatorios bajo H_0, el p-valor correspondiente es:

$$pv = 0.5P(\chi_1^2 > \hat{\Lambda}) + 0.5P(\chi_2^2 \geq \hat{\Lambda})$$

En general, para probar la significancia del q-ésimo elemento de b_i, el p-valor del LRS está dado por:

$$pv = 0.5P(\chi_s^2 > \hat{\Lambda}) + 0.5P(\chi_{s+1}^2 \geq \hat{\Lambda}) \qquad s = 1, 2, \ldots, q-1,$$

y la hipótesis se rechaza a favor de la alternativa si el p-valor es menor a un nivel dado de significancia α.

2.3. Diagnóstico

Es fundamental revisar las suposiciones hechas en el ajuste del modelo para asegurar que se cumplan las condiciones necesarias para que las inferencias sean válidas. Debido a que las suposiciones de distribución de los residuos en los modelos mixtos son similares a las de los modelos lineales, varias de las técnicas usuales, como los Q-Q plots y los gráficos de residuales contra valores ajustados se pueden usar en el contexto de los modelos mixtos, sin embargo, existen muchas más variaciones para los métodos de diagnóstico de estos modelos.

Los modelos de efectos mixtos son particularmente sensibles a datos atípicos, pues ellos dependen de los componentes de varianza que pueden ser inflacionados sustancialmente por puntos inusuales. Por tanto, el análisis de influencia es muy importante para determinar si hay observaciones que modifiquen significativamente las estimaciones del modelo. En un modelo mixto, además de las observaciones, las unidades también se pueden identificar como influyentes, por lo que también se requieren técnicas que se apliquen al nivel de las unidades.

Por otro lado, el análisis de residuales en los modelos mixtos es mucho más complejo que en los modelos lineales estándar, pues las fuentes de variabilidad adicionales dan lugar a diferentes tipos de residuos. A continuación, se presenta una clasificación de los tipos de residuos en los modelos mixtos realizada por Hilden-Minton (1995) para modelos lineales jerárquicos (que son equivalentes a los modelos lineales mixtos):

1) Residuales de nivel 1 (condicionales):

Se definen como $\hat{\varepsilon}_i = y_i - X_i\hat{\beta} - Z_i\hat{b}_i$, y se usan para predecir los errores condicionales $\varepsilon_i = y_i - E(y_i \mid b_i) = y_i - X_i\beta - Z_ib_i$.

2) Residuales de nivel 2 (efectos aleatorios):

Se expresan de la forma $Z_i\hat{b}_i$ o simplemente $\hat{b}_i$, y son inducidos por las estimaciones de los efectos aleatorios del modelo. Se utilizan para predecir los efectos aleatorios en la población $Z_ib_i = E(y \mid b) - E(y)$.

3) Residuales marginales (compuestos):

Se expresan como $\hat{\zeta}_i = y_i - X_i\hat{\beta} = Z_i\hat{b}_i + \hat{\varepsilon}_i$. Se utilizan para predecir los errores marginales, $\zeta_i = y_i - E(y_i)$, y se pueden escribir como la suma de los residuales de nivel 1 y los de nivel 2.

Un problema adicional se presenta con el concepto de **confundido** de los residuos. Según Hilden-Minton (1995), se considera que un residuo está confundido para un tipo de error específico si este depende de otros errores diferentes a los que se supone que debe predecir. Claramente los residuales marginales están confundidos con los residuales de nivel 1 y nivel 2, adicionalmente, $\hat{\varepsilon} = RD^*\varepsilon + RD^*Zb$, por lo que los residuos de nivel 1 y de nivel 2 también pueden estar confundidos.

El confundido de los residuos puede llevar a complicaciones en el diagnóstico de las deficiencias del modelo, pues una violación en un tipo de residuo se puede manifestar como una supuesta violación en un residuo diferente, por lo que un analista debe ser muy cuidadoso.

En los últimos años se han propuesto diferentes técnicas de análisis para cada uno de estos residuos y es importante notar que la mayoría de técnicas descritas en las siguientes secciones son gráficas. En el análisis de diagnóstico de modelos lineales, se ha preferido el uso de los métodos de diagnóstico con gráficas en lugar de los tests de hipótesis convencionales. Según Faraway (2016a), los gráficos de residuales son más versátiles que los tests formales, pues pueden observar problemas no anticipados. Las técnicas gráficas son más efectivas para revelar la estructura subyacente de los datos. Loy et al.

(2017) argumentan que los diagnósticos gráficos permiten detectar no sólo cuando hay un problema con el modelo y donde ocurre, sino que también dan algunas indicaciones de cuál puede ser la causa del problema, lo cual es casi imposible de conseguir usando exclusivamente métodos numéricos.

A continuación se presenta un resumen de las diferentes técnicas de diagnóstico sugeridas por varios autores para el diagnóstico de cada uno de los tipos de residuos definidos anteriormente.

2.3.1. Residuales de nivel 1 (condicionales)

Los residuales de nivel 1 o residuales condicionales se definen como:

$$\hat{\varepsilon}_i = y_i - X_i\hat{\beta} - Z_i\hat{b}_i$$

Esta definición lleva a diferentes cantidades dependiendo de la forma en la que se estiman β y b_i.

Los residuales de nivel 1 estandarizados (Loy y Hofmann, 2014) están dados por:

$$\hat{\varepsilon}_i^* = \hat{\Delta}_i^{-1/2}\hat{\varepsilon}_i, \tag{2.6}$$

donde $\hat{\Delta}_i$ es una matriz diagonal con elementos iguales a la diagonal de $Var(\hat{\varepsilon}_i)$. Para cada grupo $i = 1, \ldots, N$, los residuos de nivel 1, $\hat{\varepsilon}_i$, son tales que $Var(\hat{\varepsilon}_i) = \sigma_i^2(1 - h_i)$ donde h_i es un vector que contiene los elementos diagonales de la matriz *hat*, $H_i = X_i(X_i^\top X_i)^{-1}X_i$, también conocida como matriz de "sombrero", que es la matriz usada en el ajuste con mínimos cuadrados.

Un gráfico que puede detectar problemas de identificación del modelo en el nivel 1 es el **gráfico de $\hat{\varepsilon}_{ij}^*$ contra los valores ajustados** $\hat{y}_{ij}$. Según Singer et al. (2017), este gráfico se puede utilizar para revisar si se cumple el supuesto de homocedasticidad en los errores condicionales (si se supone que $Var(\varepsilon_i) = \sigma^2 I_{n_i}$).

De forma similar también se pueden analizar los **gráficos de $\hat{\varepsilon}_{ij}$ contra cada variable regresora**. Adicionalmente se puede utilizar el **gráfico de residuales condicionales contra los índices de las observaciones** para encontrar observaciones atípicas.

Hilden-Minton (1995) indica que la habilidad de revisar si hay normalidad de los errores condicionales incrementa a medida que se minimiza la fracción de confundido para el k-ésimo residual condicional. El autor sugiere el uso de los **residuales condicionales mínimamente confundidos**, estos son una transformación lineal de los residuales condicionales la cual

minimiza la fracción de confundido, y están definidos por:

$$c_k^\top \hat{\varepsilon} = \lambda_k^{-1/2} \ell_k^\top R^{-1/2} \hat{\varepsilon} = \lambda_k^{-1/2} \ell_k^\top R^{-1/2} y, \qquad k = 1, \ldots, N_T - p$$

donde $1 \geq \lambda_1 \geq \ldots \geq \lambda_{N_T - p} > 0$ son los valores propios ordenados de Λ, la cual se obtiene de la descomposición espectral dada por:

$$R^{1/2} Q R^{1/2} = L \Lambda L^\top, \qquad L^\top L = I_{N_T - p}$$

Estos residuales se pueden estandarizar dividiendo $c_k^\top \hat{\varepsilon}$ por la raíz cuadrada del elemento correspondiente en $C \hat{R} \hat{Q} \hat{R} C^\top$, donde $C = [c_1, \ldots, c_{N_T - p}]^\top$. Singer et al. (2017) proponen el uso de **Q-Q plots de los residuales condicionales mínimamente confundidos estandarizados** $c_k^\top \hat{\varepsilon^*}$ para revisar el supuesto de normalidad en el nivel 1.

Por otro lado, Loy y Hofmann (2014) proponen usar los **residuales semi-estandarizados** definidos por Snijders y Berkhof (2008) para revisar la suposición de homocedasticidad de los residuos de nivel 1, estos están dados por:

$$\check{\varepsilon}_i = \hat{\sigma}_i \hat{\varepsilon}_i^* \sim N(0, \sigma^2 I) \tag{2.7}$$

Además de revisar varianza constante, el **gráfico de residuales semi-estandarizados contra las variables regresoras** también se puede usar para evaluar la suposición de linealidad. Los autores también proponen usar boxplots de los residuos de nivel 1 para cada grupo con el objetivo de evaluar la homocedasticidad dentro de cada grupo. En este caso, si la suposicion de homocedasticidad dentro de los grupos es correcta, los boxplots deberían exhibir rangos intercuartiles aproximadamente constantes.

También se puede realizar un **Q-Q plot de los residuales semi-estandarizados** para revisar el supuesto de normalidad en los errores condicionales.

2.3.2. Residuales de nivel 2 (efectos aleatorios)

Los residuos de nivel 2 son las estimaciones de los efectos aleatorios, los cuales están definidos por $Z\hat{b}_i$ o también $\hat{b}_i$. Los residuos de efectos aleatorios se pueden usar para identificar variables explicativas adicionales que contribuyan significativamente al modelo y para revisar la linealidad de las variables explicativas de nivel 2.

Según Singer et al. (2017), la distancia de Mahalanobis $\mathcal{M}_i$ entre $\hat{b}_i$ y $E(b_i) = 0$,

$$\mathcal{M}_i = \hat{b}_i^\top (Var(\hat{b}_i - b_i))^{-1} \hat{b}_i \tag{2.8}$$

debería seguir una distribución chi-cuadrada con q grados de libertad cuando no hay confundido y los efectos aleatorios siguen una distribución Gaussiana de q dimensiones. Por tanto se puede emplear un **Q-Q plot χ^2_q de** $\mathcal{M}_i$ para verificar el supuesto de normalidad de los efectos aleatorios. Adicionalmente los autores recomiendan **gráficos de índice de unidad de** $\mathcal{M}_i$ para detectar datos atípicos.

2.3.3. Residuales marginales (compuestos)

Los residuales marginales se obtienen a partir de la siguiente definición:

$$\hat{\zeta}_i = y_i - X_i\hat{\beta}$$

y se pueden usar para hacer los diagnósticos usuales de los modelos lineales de un sólo nivel, sin embargo, cualquier problema que se presente debe estar debidamente acompañado por el análisis de los otros tipos de residuos con el objetivo de encontrar la fuente del problema.

Según Loy y Hofmann (2014), los residuos marginales son especialmente valiosos cuando se evalúa la estructura de covarianza marginal, como por ejemplo las medidas repetidas y los datos longitudinales, pues los residuos marginales, ζ_i, y los valores observados, y_i, tienen la misma estructura de covarianza.

Lesaffre y Verbeke (1998) comentan que cuando la estructura de covarianza dentro de las unidades es adecuada,

$$\mathcal{V}_i = ||I_{n_i} - e_i e_i^\top||^2$$

donde $e_i = \hat{\Omega}_i^{-1/2}\hat{\zeta}_i$, debería ser cercano a 0. Las unidades cuya estructura de covarianza propuesta no podría ser adecuada son las que tienen valores grandes asociados de $\mathcal{V}_i$. Para evitar que las unidades con muchas observaciones tengan más peso en la definición de $\mathcal{V}_i$, Singer et al. (2017) recomiendan usar $\mathcal{V}_i^* = \sqrt{\mathcal{V}_i}/n_i$ como una medida estandarizada de la adecuación de la estructura de covarianza dentro de cada unidad. Los autores sugieren el uso de **gráficos de $\mathcal{V}_i^*$ contra los índices de las unidades** i, para identificar las unidades con estructura de covarianza no adecuada.

Los elementos de los residuales marginales estandarizados están dados por

$$\hat{\zeta}^*_{ij} = \hat{\zeta}_{ij}/diag_j(Var(\hat{\zeta}_i))^{1/2}$$

donde $diag_j(Var(\hat{\zeta}_i))$ es el j-ésimo elemento de la diagonal principal de $Var(\hat{\zeta}_i)$. Se puede realizar un **gráfico de $\hat{\zeta}^*_{ij}$ contra los valores de cada**

variable explicativa, y también contra los valores ajustados para evaluar la linealidad de los efectos fijos en (2.1).

Singer et al. (2017) también recomiendan **graficar $\hat{\zeta}_{ij}^{*}$ contra los índices de observación** como una herramienta para detectar observaciones atípicas.

2.3.4. Análisis de influencia

Un paso importante en el diagnóstico de un modelo de regresión es la identificación de observaciones **influyentes**, las cuales pueden tener un efecto muy significativo sobre la estimación de los parámetros, las pruebas de hipótesis y los análisis hechos a partir de los resultados del modelo. Los modelos lineales mixtos no son una excepción, en este caso también se deben examinar los grupos o unidades que se utilizaron en la inclusión de los efectos aleatorios. Una sola observación influyente puede hacer que el grupo completo al que esta pertenece sea identificado como influyente, por lo cual es necesario tener a disposición herramientas que permitan identificar no sólo observaciones influyentes, sino también las unidades que tengan un gran impacto sobre el modelo.

El análisis de influencia presenta diversas complicaciones en el contexto de los modelos de efectos aleatorios, por lo cual ha sido una de las áreas de estudio más representativas del análisis de diagnóstico de modelos mixtos ver Schabenberger (2005), Demidenko y Stukel (2005) o Pan et al. (2014). Correa y Salazar (2016) mencionan que los modelos mixtos son muy sensibles a respuestas extrañas y a puntos extremos en los espacios de diseño de los efectos fijos y efectos aleatorios, por otro lado, Loy y Hofmann (2014) afirman que, debido a la estructura jerárquica de los datos, es necesario realizar eliminaciones múltiples para revisar la influencia de las observaciones individuales y los cúmulos de observaciones.

Existen diferentes maneras de revisar la influencia de las observaciones y las unidades pertenecientes a un MLM, la más sencilla es el análisis de **influencia global**, en donde se calculan los diagnósticos de eliminación, los cuales cuantifican el cambio en la estimación de un parámetro cuando se borra algún subconjunto de los datos utilizados. Estos diagnósticos se aplican en diferentes aspectos del modelo, como los componentes de varianza y los efectos fijos en orden secuencial. Una alternativa que considera los efectos fijos y los efectos aleatorios simultáneamente es el análisis de **influencia local**, concepto que fue introducido en un principio por Cook (1986), y desarrollado posteriormente por distintos autores como Lesaffre y Verbeke (1998) o Demidenko y Stukel (2005).

Puntos de leverage (o de apalancamiento)

Antes de realizar un análisis de influencia, vale la pena averiguar cuáles son las observaciones con un valor inusual con respecto a los valores ajustados y a las variables regresoras. La medida más común que se utiliza para este propósito es el **leverage**. Al igual que en la regresión lineal estándar, el leverage describe la razón de cambio en los valores ajustados respecto a los valores observados. Según Loy y Hofmann (2014), si $Var(y_i) = \sigma^2 V_i$ es conocida, el leverage en el nivel i se puede definir como:

$$\begin{aligned} H_i &= \frac{\partial \hat{y}_i^*}{\partial y_i} \\ &= X_i(X_i^\top V_i^{-1} X_i)^{-1} X_i^\top V_i^{-1} + Z_i D Z_i^\top V_i^{-1}(I - H_{1i}) \\ &= H_{1i} + H_{2i} \end{aligned} \tag{2.9}$$

donde $\hat{y}_i^* = X_i\hat{\beta} + Z_i\hat{b}_i$, $H_{1i} = X_i(X_i^\top V_i^{-1} X_i)^{-1} X_i^\top V_i^{-1}$ y
$H_{2i} = Z_i D Z_i^\top V_i^{-1}(I - H_{1i})$.

El leverage se puede descomponer en dos partes, H_{1i} describe el leverage asociado a los efectos fijos, y H_{2i} hace referencia al leverage asociado a los efectos aleatorios. Sin embargo H_{2i} depende de H_{1i}, es decir, el leverage de los efectos aleatorios está confundido por el de los efectos fijos. Para solucionar este problema, se define de forma alternativa el leverage de los efectos aleatorios de la siguiente manera:

$$H_{2i}^* = Z_i D Z_i^\top \tag{2.10}$$

Las observaciones o unidades que tengan un valor alto de H_{1i} o de H_{2i} van a tener un leverage alto en los efectos fijos o en los efectos aleatorios respectivamente. Se puede elegir un valor umbral para las observaciones con leverage alto como el típico $2p/N$, o se pueden graficar los valores de cada componente de leverage y visualizar las observaciones que tengan un valor muy diferente al resto.

Por otro lado, Singer et al. (2017) consideran un enfoque más general para definir el leverage, en este caso, se define **la matriz de leverage marginal generalizada** como:

$$\begin{aligned} L_1 &= \frac{\partial \hat{y}}{\partial y^\top} \\ &= \frac{\partial X(X^\top V^{-1} X)^{-1} X^\top V^{-1} y}{\partial y^\top} \\ &= X(X^\top V^{-1} X)^{-1} X^\top V^{-1} \end{aligned} \tag{2.11}$$

De aquí,

$$tr(L_1) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} L_{1i(jj)} = p$$

donde L_{1i} es el i-ésimo bloque diagonal de L_1, y $L_{1i(jj)}$ es el j-ésimo elemento de la diagonal principal de L_{1i}. Los autores consideran que la unidad i tiene leverage alto con respecto a los valores marginales ajustados $\hat{y}$ si $tr(L_{1i})/n_i > 2p/n$, aunque también sugieren una inspección visual de los gráficos de leverage para obtener la conclusión correspondiente.

Sea $\hat{y}^* = X\hat{\beta} + Z\hat{b}$, luego se define la **matriz de leverage conjunto generalizada** como

$$L = \frac{\partial \hat{y}^*}{\partial y^\top} = \frac{\partial \hat{y}}{\partial y^\top} + \frac{\partial Z\hat{b}}{\partial y^\top} = L_1 + Z\Gamma Z^\top Q = L_1 + L_2 Q \tag{2.12}$$

donde L_1 está dada por (2.11), y $L_2 = Z\Gamma Z^\top$. La matriz L_2 representa a la porción de la variabilidad dentro de cada unidad explicada por la presencia de los efectos aleatorios.

Singer et al. (2017) afirman que el componente de efectos aleatorios en (2.12) se puede utilizar para evaluar la influencia de las observaciones sobre los componentes de varianza estimados. Por lo tanto, se puede usar la matriz $H_2 = Z\Gamma Z^\top Q$ como medida de leverage generalizado para tal propósito. Sin embargo, los autores notan que la matriz H_2 se puede reescribir de la siguiente forma:

$$H_2 = L_2 V^{-1}[I_N - L_1]$$

así, H_2 depende de L_1, por lo que el leverage con respecto a los efectos aleatorios de los valores ajustados condicionales puede estar confundido por el leverage con respecto a los valores ajustados marginales. Por lo tanto, **es preferible usar la matriz L_2 en lugar de H_2** para medir el leverage de las observaciones y unidades que pueden tener un efecto importante sobre la estimación del vector θ.

Influencia global (eliminación de casos)

Uno de los estadísticos de eliminación de casos más utilizados para evaluar el impacto de cada observación en las estimaciones de un modelo de regresión es la **distancia de Cook**, la cual es una medida de la distancia entre las estimaciones del modelo completo a las obtenidas a partir del modelo ajustado con los datos reducidos. En el caso de los MLMs, es importante

identificar primero si alguna unidad es influyente, por tanto la definición de distancia de Cook se puede generalizar de la siguiente manera:

Sea I el conjunto de observaciones a evaluar para el modelo mixto, y sea $\hat{\beta}_{(I)}$ la estimación del vector de efectos fijos al eliminar las observaciones dentro de I, luego la distancia de Cook se define como (Singer et al., 2017):

$$C_I(\hat{\beta}) = \frac{1}{p}(\hat{\beta} - \hat{\beta}_{(I)})^\top \widehat{Var(\hat{\beta})}^{-1}(\hat{\beta} - \hat{\beta}_{(I)}) \tag{2.13}$$

El conjunto I puede contener sólo una observación, en este caso se realiza una **eliminación de nivel 1**. Ahora, si I contiene todas las observaciones dentro de una unidad, se realiza una **eliminación de nivel 2**. Mientras más grande sea el valor de C_I, más grande va a ser el nivel de influencia del conjunto de observaciones sobre la estimación de β.

En la práctica no se conocen los componentes de varianza, por lo tanto, no existe una distribución de referencia para la distancia de Cook con la que se puedan realizar pruebas de hipótesis (Loy y Hofmann, 2014). Luego, los autores recomiendan utilizar **métodos gráficos, de posición relativa o de bootstrap paramétrico** para evaluar la influencia de las observaciones y unidades por medio de la distancia de Cook.

Uno de estos métodos es el gráfico de índices o **index plot**, otra alternativa, la cual es preferida por Loy y Hofmann, es el **dotplot**, el cual compara los valores ordenados de menor a mayor de las distancias de Cook contra sus índices respectivos. Las observaciones o unidades que presenten un valor de C_I relativamente grande o muy diferente al resto se pueden considerar como influyentes. Una ventaja que presenta el dotplot, es que le permite al investigador visualizar brechas, las cuales proporcionan información adicional sobre las posibles observaciones influyentes del modelo.

Un estadístico alternativo que también se puede utilizar para medir la influencia de los datos sobre la estimación de los efectos fijos en los modelos mixtos es el **MDFFITS**, el cual es una versión multivariada del DFFITS que se utiliza en regresión lineal. Si I es el conjunto de las observaciones que se van a evaluar, el MDFFITS se define como:

$$MDFFITS_I(\hat{\beta}) = \frac{1}{p}(\hat{\beta} - \hat{\beta}_{(I)})^\top \widehat{Var(\hat{\beta}_{(I)})}^{-1}(\hat{\beta} - \hat{\beta}_{(I)}) \tag{2.14}$$

Sin embargo la definición de este estadístico presenta una desventaja sobre la distancia de Cook, para calcular el MDFFITS se requiere que la estructura de covarianza sea estimada de nuevo cuando se elimina la i-ésima unidad, y por tanto se debe volver a calcular la matriz inversa correspondiente (Loy y Hofmann, 2014).

La distancia de Cook no es eficiente para realizar diagnósticos de influencia en los modelos mixtos. Singer et al. (2017), y Tan et al. (2001), indican que la posición relativa de las observaciones dentro y a lo largo de las unidades presenta inconvenientes al usar medidas de detección de unidades influyentes. Por lo tanto, una propuesta alternativa es la **distancia condicional de Cook**, la cual es una aproximación condicional basada en las medidas de influencia que se aplican a las observaciones.

Singer et al. (2017) generalizan la definición dada por Tan et al. (2001) para el caso cuando R_i es cualquier matriz de covarianza además de $R_i = \sigma^2 I_{n_i}$. Se considera el modelo condicional dado por $y = X^*\beta^* + \varepsilon$, donde $X^* = [X\ Z]$ y $\beta^* = (\beta^\top, b^\top)^\top$. La distancia condicional de Cook se define como:

$$\begin{aligned} D_{i(j)}^{cond} &= \sum_{i=1}^{N} \frac{(\hat{y}_i^* - \hat{y}_{i(j)}^*)^\top Var(y_i \mid b_i)^{-1}(\hat{y}_i^* - \hat{y}_{i(j)}^*)}{(N-1)q + p} \\ &= \sum_{i=1}^{N} \frac{(\hat{y}_i^* - \hat{y}_{i(j)}^*)^\top \hat{R}_i^{-1}(\hat{y}_i^* - \hat{y}_{i(j)}^*)}{(N-1)q + p} \end{aligned} \tag{2.15}$$

donde $\hat{y}_i^* = X_i\hat{\beta} + Z_i\hat{b}$, $\hat{y}_{i(j)}^* = X_i\hat{\beta}_{i(j)} + Z_i\hat{b}_{i(j)}$. En este caso, $\hat{\beta}_{i(j)}$ y $\hat{b}_{i(j)}$ son los mejores estimadores lineales insesgados (BLUE) de β y de b obtenidos al eliminar la j-ésima observación de la i-ésima unidad.

La distancia condicional de Cook se puede separar en 3 componentes de la siguiente manera (Singer et al., 2017):

$$D_{i(j)}^{cond} = D_{1i(j)}^{cond} + D_{2i(j)}^{cond} + D_{3i(j)}^{cond} \tag{2.16}$$

donde,

$$\begin{aligned} D_{1i(j)}^{cond} &= ((N-1)q + p)^{-1}(\hat{\beta} - \hat{\beta}_{i(j)})^\top X^\top \hat{R}^{-1} X(\hat{\beta} - \hat{\beta}_{i(j)}) \\ D_{2i(j)}^{cond} &= ((N-1)q + p)^{-1} \sum_{i=1}^{N} (\hat{b} - \hat{b}_{i(j)})^\top Z_i^\top \hat{R}_i^{-1} Z_i(\hat{b} - \hat{b}_{i(j)}) \\ D_{3i(j)}^{cond} &= 2((N-1)q + p)^{-1}(\hat{\beta} - \hat{\beta}_{i(j)})^\top \sum_{i=1}^{N} X_i^\top \hat{R}_i^{-1} Z_i(\hat{b} - \hat{b}_{i(j)}) \end{aligned}$$

Cada componente de la suma en (2.16) se puede revisar para evaluar la influencia en diferentes aspectos del modelo:

- $D_{1i(j)}^{cond}$ se puede utilizar para medir la influencia de la j-ésima observación de la i-ésima unidad en la estimación de β.

- $D^{cond}_{2i(j)}$ es útil para evaluar la influencia de la j-ésima observación de la i-ésima unidad en la estimación de los efectos aleatorios, b.
- $D^{cond}_{3i(j)}$ mide la covariación entre un cambio en el perfil promedio y un cambio en la posición de los perfiles específicos a cada unidad relativos al perfil promedio cuando se elimina la j-ésima observación de la i-ésima unidad.
- Si el interés es medir el impacto de una unidad en las estimaciones de los parámetros, basta sumar los componentes de (2.15) sobre todas las observaciones correspondientes a esa unidad.

Existen enfoques adicionales para los diagnósticos de influencia global en los MLMs, se puede considerar el uso de el **cociente de los elipsoides de varianza** (Singer et al., 2017 , Hilden-Minton, 1995), el **análisis de influencia parcial** (Loy y Hofmann, 2013) o los gráficos de sumas de cuadrados residuales estudentizadas **TRSS** (Mun y Lindstrom, 2013).

El paquete estadístico de R, *influence.ME* (Nieuwenhuis et al., 2012), permite calcular una medida adicional de influencia sobre cada uno de los efectos fijos del modelo, los **DFBETAS**. Los autores generalizan los DFBETAS para los modelos mixtos, y proveen un punto de corte para identificar unidades influyentes, dado por $2/\sqrt{N}$.

Por otro lado, se puede revisar la influencia de las observaciones y las unidades sobre la precisión de la estimación de los efectos fijos, la cual se puede evaluar a partir de la matriz de covarianza $Var(\hat{\beta})$. Las observaciones que cambien significativamente la estimación de $Var(\hat{\beta})$ se pueden considerar como influyentes sobre la precisión de $\hat{\beta}$. El **COVTRACE** y el **COVRATIO** son medidas del cambio en la precisión cuando se eliminan las observaciones dentro del conjunto de interés. Se definen de forma general para el modelo mixto de la siguiente manera (Loy y Hofmann, 2014):

$$COVTRACE_I(\hat{\beta}) = \left| tr(\widehat{Var(\hat{\beta})}^{-1} \widehat{Var(\hat{\beta}_{(I)})}) - p \right| \tag{2.17}$$

$$COVRATIO_I(\hat{\beta}) = \det\left(\widehat{Var(\hat{\beta}_{(I)})}\right) \det\left(\widehat{Var(\hat{\beta})}\right)^{-1} \tag{2.18}$$

Ambos estadísticos comparan las matrices de covarianza de $\hat{\beta}$ cuando β se estima con y sin el conjunto de observaciones I. En el caso de que el conjunto I no es influyente, el COVTRACE va a ser cercano a 0, mientras que el COVRATIO va a ser cercano a 1. De igual manera a los estadísticos descritos anteriormente, la precisión mediante el COVTRACE y el COVRATIO se

les puede evaluar utilizando métodos gráficos, de posición relativa, o de bootstrap paramétrico.

Los estadísticos definidos en (2.13), (2.14), (2.17) y (2.18) se pueden utilizar para evaluar la influencia sobre los componentes de varianza estimados, en este caso se reemplazan los valores de $\hat{\beta}$ y $\hat{\beta}_{(I)}$ por $\hat{\theta}$ y $\hat{\theta}_{(I)}$ respectivamente. Sin embargo, recalcular el valor de los estadísticos para los componentes de varianza para cada conjunto a examinar es poco accesible computacionalmente, por lo tanto, Loy y Hofmann (2014) sugieren utilizar un diagnóstico que no requiera que se vuelva a estimar la estructura de covarianza. Los autores recomiendan como medida al **cambio relativo en la varianza** o **RVC**, el cual mide el cambio en las estimaciones del k-ésimo componente de varianza, θ_k, con y sin las observaciones en I. El RVC se define de la siguiente forma:

$$RVC_I(\hat{\theta}_k) = \frac{\hat{\theta}_{k(I)}}{\hat{\theta}_k} - 1 \tag{2.19}$$

donde $\hat{\theta}_{k(I)}$ es la estimación del componente de varianza θ_k cuando se elimina el conjunto I, y $\hat{\theta}_k$ es la estimación del componente de varianza obtenido con todos los datos. Si el conjunto de observaciones I no tiene una influencia significativa en el componente de varianza, entonces el RVC va a ser cercano a 0.

Influencia local

En el análisis de influencia local el interés radica en evaluar el cambio en el análisis que resulta de perturbaciones infinitesimales en los datos del modelo. La medida que se considera para investigar los datos influyentes es el cambio en el **desplazamiento de verosimilitud**, el cual está definido como:

$$LD(\omega) = 2(\mathcal{L}(\hat{\psi}) - \mathcal{L}(\hat{\psi}_\omega \mid \omega))$$

donde $\mathcal{L}$ es la función de verosimilitud del modelo, ψ es un vector de parámetros p-dimensional, $\omega \in \Omega \subset \mathbb{R}^q$ es un vector q-dimensional que contiene las perturbaciones, y $\hat{\psi}$, $\hat{\psi}_\omega$ son los estimadores de máxima verosimilitud de ψ obtenidos a partir de $\mathcal{L}(\psi)$ y $\mathcal{L}(\psi \mid \omega)$ respectivamente. El efecto de las perturbaciones realizadas se mide a través de la curvatura del desplazamiento de verosimilitud.

Singer et al. (2017) describen el proceso: Se asume que existe $\omega_0 \in \Omega$ tal que $\psi_{\omega_0} = \psi$, y se supone que $\mathcal{L}(\psi \mid \omega)$ es de clase C^2 en una vecindad de ω_0. Luego, la curvatura normal del gráfico $G(\omega) = [\omega^\top, LD(\omega)]$ en la dirección

del vector q-dimensional $\mathbf{d}$ de norma unitaria en el punto ω_0 está dada por:

$$C_d = 2\left|d^\top H^\top \ddot{F}^{-1} H d\right|$$

donde

$$\ddot{F} = \left[\frac{\partial^2 \mathcal{L}(\psi)}{\partial\psi\partial\psi^\top}\right]_{\psi=\hat{\psi}} \qquad \text{y} \qquad H = \left[\frac{\partial^2 \mathcal{L}(\psi \mid \omega)}{\partial\psi\partial\omega^\top}\right]_{\psi=\hat{\psi}}$$

Sea $C_{\text{mín}}$ el valor propio más pequeño de $-H\ddot{F}^{-1}H$, y sea $C_{\text{máx}}$ su valor propio más grande, se puede mostrar que $C_{\text{mín}} \leq C_d \leq C_{\text{máx}}$. El vector propio $d_{\text{máx}}$ correspondiente a $C_{\text{máx}}$ se puede usar para identificar cual combinación lineal de los elementos de ω es más influyente en la curvatura de $LD(\omega)$. El vector $d_{\text{máx}}$ se puede emplear para evaluar el efecto de las perturbaciones en distintos componentes del modelo:

- La variable respuesta, $y_i(\omega_i) = y_i + \omega_i$.
- Las variables regresoras, $X_i(W_i) = X_i + W_i$, donde $W_i = [\omega_{i1}, \ldots, \omega_{ip}]$ con $\omega_{ij} = (\omega_{ij1}, \ldots, \omega_{ijn_i})^\top$.
- La varianza de los efectos aleatorios, $D^*(\omega) = \omega D^*$.
- La varianza de los errores, $R_i(\omega_i) = \omega_i R_i$.

En cada uno de los casos anteriores, se pueden realizar **gráficos de los valores absolutos de los elementos de $d_{\text{máx}}$ contra los índices de las observaciones o de las unidades** para identificar los datos que tengan un gran impacto sobre el desplazamiento de la verosimilitud.

Debido a que no es sencillo elegir un esquema de perturbaciones adecuado, Singer et al. (2017) describen otro método, el cual fue desarrollado por Lesaffre y Verbeke (1998). En este caso se considera la verosimilitud marginal de $\psi = (\beta^\top, \theta^\top)^\top$:

$$\mathcal{L}(\psi) = \sum_{i=1}^{N} \mathcal{L}_i(\psi) = -\frac{1}{2}\sum_{i=1}^{N}(\log(V_i) + (y_i - X_i\beta)^\top V_i^{-1}(y_i - X_i\beta))$$

y se incluyen las perturbaciones de sus términos individuales tomando:

$$\mathcal{L}_i(\psi_\omega) = \sum_{i=1}^{N} \omega_i \mathcal{L}_i(\psi)$$

donde $\mathcal{L}_i$ denota la verosimilitud de la i-ésima unidad y $\omega = (\omega_1, \ldots, \omega_n)^\top$. Si u_i denota la i-ésima columna de una matriz identidad de orden N, la curvatura normal calculada en la dirección u_i es:

$$C_{u_i} = 2 \left| H_i^\top \ddot{F}^{-1} H_i \right| \tag{2.20}$$

donde H_i denota la i-ésima columna de H. La i-ésima unidad va a tener un gran impacto en el desplazamiento de verosimilitud si el valor de (2.20) es grande. Lesaffre y Verbeke (1998) muestran que C_{u_i} está relacionada al impacto de la i-ésima unidad en el estimador del vector de parámetros ψ, y por tanto proponen una descomposición de C_{u_i} que se puede usar para identificar cual es la parte del modelo que se ve mayormente afectada por una unidad específica. Los autores recomiendan realizar **gráficos de índices de** $||\mathbf{I_{n_i}} - \mathcal{E_i}\mathcal{E_i}^\top||^\mathbf{2}$, $||\mathcal{E_i}\mathcal{E_i}^\top||^\mathbf{2}$, $||\mathcal{X_i}\mathcal{X_i}^\top||^\mathbf{2}$, $||\mathcal{Z_i}\mathcal{Z_i}^\top||^\mathbf{2}$ y $||\mathbf{V_i}^\top||^\mathbf{2}$, donde $\mathcal{E}_i = \hat{V}_i^{-1/2}\hat{\zeta}_i$, $\mathcal{X}_i = \hat{V}_i^{-1/2}X_i$ y $\mathcal{Z}_i = \hat{V}_i^{-1/2}Z_i$, como medidas de diagnóstico adicionales cuando se identifican unidades influyentes con C_{u_i}.

2.3.5. Medidas remediales

Si el análisis de diagnóstico detecta inadecuaciones en el modelo, es necesario aplicar medidas correctivas para que las inferencias e interpretaciones realizadas sean correctas. Según Loy y Hofmann (2014), los problemas de no linealidad se pueden corregir con una transformación apropiada de los predictores, mientras que para corregir heterocedasticidad y no normalidad se puede transformar la respuesta, sin embargo, se pueden presentar dificultades con la elección de la transformación. Las transformaciones de Box-Cox para modelos mixtos (Gurka et al., 2006) pueden ser útiles en este caso.

El problema de no normalidad en la distribución del modelo o en los efectos aleatorios también puede ser abordado sin necesidad de transformaciones con métodos de estimación robusta (Koller, 2013, Yau y Kuk, 2002 o Dueck y Lohr, 2005), o estimadores de *sandwich* para los errores estándar (Verbeke y Lesaffre, 1997, Yuan y Bentler, 2002). Ghidey et al. (2010) presentan una revisión de métodos adicionales que generalizan la distribución de los efectos aleatorios. Por otro lado, Singer et al. (2017) presentan un resumen de una clase de modelos mixtos que se pueden utilizar en casos de no normalidad, los *modelos mixtos elípticamente simétricos y asimétricos*, sin embargo, los autores comentan que todavía no se han desarrollado métodos de diagnóstico para estos modelos, por lo que se requiere investigación adicional en este campo. Por otro lado, también se pueden considerar modelos que tienen en cuenta la heterocedasticidad en el nivel 1 (Kasim y Raudenbush, 1998), o se

pueden debilitar las suposiciones del modelo para que la varianza residual dependa de algún predictor (Snijders y Bosker, 2011).

Si se encuentran datos atípicos o puntos influyentes, estos deben ser examinados más a fondo para descartar la posibilidad de que haya un error de registro. Según Loy y Hofmann (2014), si se identifica una unidad atípica diferente respecto a algún predictor, entonces se puede incluir una variable dummy en el modelo que explique la diferencia. Las unidades atípicas e influyentes se pueden eliminar del modelo si se determina que provienen de una población de interés diferente. Snijders y Bosker (2011) indican que se pueden postular distribuciones para los residuales con colas más pesadas que las gaussianas, como la distribución t, para reducir la influencia de los datos atípicos. Los métodos robustos también pueden corregir problemas de unidades y observaciones atípicas.

2.3.6. Inferencia visual

Una metodología estadística relativamente nueva, pero que puede llegar a tener un gran potencial para el desarrollo del análisis y diagnóstico de modelos mixtos es la **inferencia visual**. Loy et al. (2017) realizan una revisión del procedimiento y los métodos que se pueden utilizar para realizar inferencias y validar supuestos en un modelo lineal mixto utilizando inferencia visual.

En este contexto, los estadísticos de prueba son gráficos los cuales muestran un aspecto de la suposición del modelo. Primero se generan gráficos a partir de datos que son consistentes con la hipótesis nula llamados **gráficos nulos**, el conjunto de gráficos nulos va a constituir la distribución de referencia. Después se genera un gráfico a partir de los datos observados, este último se asigna a una posición al azar en una alineación que contiene a los gráficos nulos, luego, a un grupo de observadores independientes, que no tienen sesgos ni conocimiento previo de la información contextual del modelo, se les asigna la tarea de identificar cual es el gráfico más diferente en la alineación. El gráfico de los datos observados va a ser indistinguible de los gráficos nulos si la suposición del modelo se cumple.

El número de observadores que identificaron correctamente el gráfico de los datos originales como el más diferente se va a utilizar para definir un **p-valor visual**, con el cual se puede probar la hipótesis planteada en la investigación. El diseño de los gráficos a realizar en la alineación varía dependiendo del aspecto del modelo que se quiere revisar. La figura 2.1 presenta un ejemplo de alineación generada con los datos del objeto *Dialyzer* de la librería *MEMSS* (Bates et al., 2019). La alineación contiene gráficos

de los residuos condicionales contra un predictor para probar el supuesto de homogeneidad de varianza en el nivel 1. El gráfico de los datos, presente en el panel 18, es lo suficientemente diferente del resto de gráficos nulos, por tanto puede ser identificado por un número significativo de observadores independientes.

Según Loy et al. (2017), la ventaja de este método sobre los tests de hipótesis convencionales es que no requiere reglas diferentes basadas en el método de estimación o de la ubicación de un parámetro en el espacio de parámetros. Esta metodología evita el problema de definir los grados de libertad para los tests de inferencia, y permite el testeo de cualquier subconjunto de efectos aleatorios. Gracias a la inferencia visual se puede obtener información adicional sobre la razón del rechazo de la hipótesis nula; también se puede usar para realizar diagnósticos de los supuestos del modelo; al elegir el tipo de gráfico adecuado, se pueden obtener conclusiones claras y muy similares a las que se pueden formular con los métodos convencionales resumidos a lo largo de este capítulo.

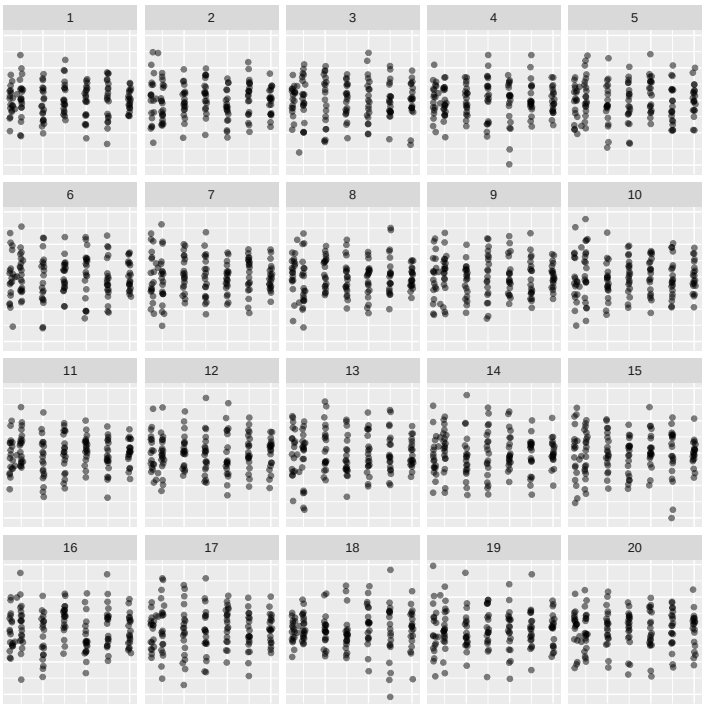

Figura 2.1: Ejemplo de alineación para evaluar el supuesto de homogeneidad de varianza en el nivel 1.

Sin embargo, la inferencia visual depende en gran medida del proceso de simulación para generar los gráficos nulos, por lo que se pueden presentar complicaciones en el proceso computacional. Elegir un diseño apropiado para los gráficos a realizar es muy importante para obtener conclusiones válidas, por último, este método involucra trabajar con observadores humanos, **la dificultad reside en conseguir un grupo de personas dispuestas a concluir adecuadamente y sin ningún sesgo previo sobre los gráficos presentes en las alineaciones.** En cualquier caso, la inferencia visual es un área prometedora en el entorno de investigación actual, en donde es necesario trabajar con modelos estadísticos cada vez mas complejos.

Capítulo 3

Métodos de diagnóstico en modelos lineales generalizados

3.1. Definición del modelo lineal generalizado

El modelo lineal mixto se puede considerar como una generalización del componente aleatorio del modelo lineal estándar dado por:

$$Y = X\beta + \epsilon$$

Sin embargo, en la formulación de ambos modelos se asume que la variable respuesta sigue una distribución normal. En muchas ocasiones la respuesta puede ser binaria, categórica, puede representar un conteo, una proporción que sólo toma valores entre 0 y 1, o puede proceder de alguna configuración en donde no es razonable la suposición de normalidad. Los **modelos lineales generalizados** o **MLGs**, los cuales fueron introducidos por Nelder y Wedderburn (1972), son una clase de modelos lineales que generalizan la distribución de la variable respuesta, y además permiten modelar la relación, no necesariamente lineal, de los predictores con la respuesta.

En este capítulo se define la familia exponencial y el modelo lineal generalizado. En la sección 3.2 se presentan los métodos de estimación y de inferencia más utilizados para este tipo de modelos, en la sección 3.3 se exponen diversas técnicas de diagnóstico para evaluar los supuestos del modelo, y en la sección 3.4 se presenta un corto resumen de los modelos lineales gene-

ralizados mixtos, una extensión del modelo lineal generalizado que considera efectos aleatorios.

Los modelos lineales generalizados proveen un marco unificador para muchas técnicas estadísticas de uso común, por tanto, su teoría y aplicaciones se ha desarrollado en gran medida en los últimos años (ver Myers et al., 2012, Lee et al., 2018 o Dobson y Barnett, 2018). En principio, el MLG fue desarrollado para modelos que sólo tienen efectos fijos, sin embargo estos modelos se pueden generalizar para ajustar modelos mixtos que incluyan efectos aleatorios. En este caso tales modelos se denominan como **Modelos lineales generalizados mixtos** o **MLGMs**.

Muchas de las "buenas" propiedades con las que cuenta la distribución normal están presentes en una clase de distribuciones más amplia conocida como la familia exponencial de distribuciones. La teoría de los MLGs se fundamenta en el supuesto de que la distribución de la variable respuesta pertenece a esta familia. Sin embargo, algunas densidades que no pertenecen a la familia exponencial se pueden usar para ajustar MLGs al realizar las modificaciones apropiadas (Faraway, 2016a).

La distribución de una variable aleatoria Y pertenece a la **familia exponencial lineal**, si su función de masa o densidad se puede escribir de la siguiente forma:

$$f(y, \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}, \tag{3.1}$$

donde θ es un parámetro de ubicación llamado **parámetro canónico**, ϕ es un parámetro de escala, llamado **parámetro de dispersión**, y $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ son funciones conocidas. Muchas de las distribuciones que se utilizan comúnmente en la práctica, como la Normal, la Binomial, la Poisson, o la Gamma, pertenecen a la familia exponencial. Si la distribución de la variable aleatoria Y pertenece a la familia exponencial, su valor esperado y su varianza se pueden calcular en términos de las funciones a y b:

$$E(Y) = b'(\theta) \tag{3.2}$$

$$Var(Y) = a(\phi)b''(\theta) \tag{3.3}$$

Se puede notar que la media sólo depende del parámetro canónico, mientras que la varianza depende de dos funciones. La función $b''(\theta)$ describe cómo se relaciona la varianza de Y con su media al usar la relación conocida entre θ y $E(Y) = \mu$. Por ejemplo, en el caso de la distribución normal, $b''(\theta) = 1$, por tanto la media es independiente de la varianza. Por otro lado, la función $a(\phi)$ toma el valor de 1 en el caso de algunas distribuciones de uso común como la

Binomial y la Poisson, sin embargo, en la mayoría de los casos esta función va a ser de la forma $a(\phi) = \phi/w$, donde w es un *peso a priori* conocido, el cual varía de una observación a otra. De manera similar se define la función $V(\mu) = b''(\theta)$, la cual se conoce como **función de varianza**.

Otro componente fundamental de los MLG es la **función de enlace**, la cual describe cómo se relaciona la media de la respuesta con los predictores. Esta relación no es necesariamente lineal como en el caso del modelo de regresión Gaussiano, por tanto se supone que las variables independientes producen un predictor lineal η, dado por:

$$\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k = x^\top \beta$$

La función que describe cómo se relaciona la respuesta media μ con los predictores a través del predictor lineal es la función de enlace g, la cual es monótona, diferenciable, y es tal que:

$$\eta = g(\mu)$$

La elección de la función de enlace depende de las condiciones del problema a considerar, por ejemplo, si los datos son de conteo, se necesita que $\mu > 0$, luego una elección viable es $\mu = e^\eta$, de donde $\eta = \log(\mu)$, la función $g(\mu) = \log(\mu)$ garantiza que la media es positiva. El **enlace canónico** es la función g que cumple $\eta = g(\mu) = \theta$, el cual es el parámetro canónico de la distribución. El enlace canónico es una elección conveniente desde el punto de vista teórico y computacional, además hace que $X^\top Y$ sea suficiente para β (Faraway, 2016a).

Tabla 3.1: Enlaces canónicos para los MLGs más comunes.

Familia	**Enlace canónico**	**Función de varianza**
Normal	$\eta = \mu$	$V(\mu) = 1$
Poisson	$\eta = \log(\mu)$	$V(\mu) = \mu$
Binomial	$\eta = \log(\mu/(1-\mu))$	$V(\mu) = \mu(1-\mu)$
Gamma	$\eta = \mu^{-1}$	$V(\mu) = \mu^2$
Normal inversa	$\eta = \mu^{-2}$	$V(\mu) = \mu^3$

Como ejemplo para una respuesta binaria, sea $Y_1, Y_2, \ldots, Y_N$ una muestra aleatoria de N observaciones, para cualquier i, Y_i es una variable aleatoria de Bernoulli con 2 posibilidades, éxito ($Y_i = 1$) y fracaso ($Y_i = 0$). Si

p_i es la probabilidad de éxito para la i-ésima observación, entonces,

$$\begin{aligned} f(y_i, p_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \exp\{\log(p_i^{y_i})\}\exp\{\log((1-p_i)^{1-y_i})\} \\ &= \exp\{y_i \log(p_i) + (1-y_i)\log(1-p_i)\} \\ &= \exp\{y_i \log(p_i/(1-p_i)) + \log(1-p_i)\} \end{aligned}$$

Ahora, si $\theta = \log(p_i/(1-p_i))$, se tiene que:

$$\log(1+e^{\theta}) = \log(1+p_i/(1-p_i)) = \log(1/(1-p_i)) = \log(1) - \log(1-p_i) = -\log(1-p_i)$$

Luego,

$$f(y_i, p_i) = f(y_i, \theta) = \exp\{y_i\theta - \log(1+e^{\theta})\}$$

Por lo tanto, la distribución de Y_i toma la forma general dada en (3.1), donde $\phi \equiv 1$, $a(\phi) = 1$, $b(\theta) = \log(1+e^{\theta})$ y $c(y_i, \phi) = 1$. Es decir, la distribución de Y_i pertenece a la familia exponencial lineal con parámetro canónico $\theta = \log(p_i/(1-p_i))$.
Así, el enlace canónico está dado por la función *logit*, $g(p_i) = \log(p_i/(1-p_i))$, por lo que el MLG correspondiente con k predictores, llamado **modelo de regresión logística**, está dado por:

$$\log\left(\frac{p_i}{1-p_i}\right) = \eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} \qquad i = 1, \ldots, N$$

En este caso para cualquier i, $E(Y_i) = p_i$, luego la probabilidad de que la i-ésima observación presente un éxito es:

$$p_i = \mu_i = b'(\theta) = \frac{e^{\theta}}{1+e^{\theta}} = \frac{e^{x_i^{\top}\beta}}{1+e^{x_i^{\top}\beta}}$$

En general, la distribución utilizada para modelar la respuesta Y varía de acuerdo a la naturaleza de los datos, por ejemplo, para modelar conteos se puede usar una distribución de Poisson, para proporciones comúnmente se utiliza una distribución Binomial, y para datos continuos positivos se pueden utilizar las distribuciónes Gamma o la Normal inversa.

3.2. Estimación e Inferencia

3.2.1. Estimación del modelo

El vector de efectos fijos en un MLG se estima con el método de máxima verosimilitud, sin embargo, debido a la complejidad de la definición del

modelo, a menudo es necesario utilizar métodos númericos para la estimación de β. Sea $Y_1, Y_2, \ldots, Y_N$ una muestra aleatoria obtenida a partir de una población cuya distribución pertenece a la familia exponencial, luego la distribución de cada Y_i es de la forma dada en (3.1). La log-verosimilitud para una sola observación y_i es:

$$\ell_i = \log(L(\beta, y_i)) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)$$

donde β es un vector de tamaño $p \times 1$ de efectos fijos. Por tanto, la log-verosimilitud del modelo está dada por:

$$\ell = \sum_{i=1}^{N} \ell_i = \sum_{i=1}^{N} \left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right)$$

Al aplicar la regla de la cadena, y las propiedades de las distribuciones pertenecientes a la familia exponencial, entre ellas las descritas en (3.2) y (3.3), es posible mostrar (Tellez y Morales, 2016) que la derivada parcial de ℓ respecto a cada efecto fijo se puede escribir como:

$$U_j := \frac{\partial \ell}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{N} w_i \left(\frac{y_i - \mu_i}{V(\mu_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \right) \qquad \forall j = 1, 2, \ldots, p$$

El vector de **score U**, está dado por $U = [U_1, U_2, \ldots, U_p]^\top$, luego, se debe solucionar el sistema de ecuaciones $U = 0$ para obtener las estimaciones de máxima verosimilitud de β. En general, la solución del sistema se obtiene iterativamente con métodos numéricos, en particular, se utilizan los algoritmos de *Newton-Raphson* y de *Fisher Scoring* (los cuales son equivalentes cuando se usa el enlace canónico). En este caso, $\hat{\beta}$ es la solución de un proceso de **mínimos cuadrados ponderados iterativos** o **MCPI**, en el cual se solucionan las ecuaciones dadas por:

$$X^\top W X \beta^{(m)} = X^\top W z \tag{3.4}$$

donde,

- W es una matriz diagonal de dimensión $N \times N$ con elementos

$$w_{ii} = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

- z es un vector de tamaño $N \times 1$ con componentes de la forma

$$z_i = \eta_i + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i}$$

- y $\beta^{(m)}$ es la estimación de β en la iteración m.

El algoritmo utilizado se resume en los siguientes pasos:

1. Se realiza una estimación inicial de β, $\hat{\beta}^{(0)}$, a partir de la cual se calculan $\hat{\mu}^{(0)}$ y $\hat{\eta}^{(0)}$.

2. Se calcula el vector $z^{(0)}$, con componentes

$$z_i^{(0)} = \hat{\eta}_i^{(0)} + (y_i - \hat{\mu}_i^{(0)}) \left.\frac{\partial \eta_i}{\partial \mu_i}\right|_{\hat{\eta}_i^{(0)}}$$

3. Se forma la matriz diagonal de pesos $W^{(0)}$ con elementos

$$w_i^{(0)} = \frac{1}{Var(\hat{\mu}_i^{(0)})} \left(\left.\frac{\partial \mu_i}{\partial \eta_i}\right|_{\hat{\eta}_i^{(0)}} \right)^2$$

4. Se soluciona el sistema en (3.4) para obtener la nueva estimación de β, $\hat{\beta}^{(1)}$, a partir de la cual se obtienen $\hat{\eta}^{(1)}$ y $\hat{\mu}^{(1)}$. Luego se repiten los pasos anteriores para la siguiente iteración.

5. Se siguen realizando iteraciones hasta llegar a la convergencia.

3.2.2. Inferencia en el modelo lineal generalizado

Existen diferentes enfoques para realizar inferencias sobre los parámetros de un MLG. Una primera aproximación es el uso del test de razón de verosimilitud, si n es el número de parámetros del modelo saturado, es decir, el modelo que incluye todos los predictores que se pueden considerar, y p es el número de parámetros del modelo de interés, se puede evaluar la diferencia entre las log-verosimilitudes del modelo saturado y el modelo de interés, esta cantidad se conoce como el **desvío**, y se escribe como:

$$D = 2(\log(L(\hat{\beta}_{\text{máx}}, y)) - \log(L(\hat{\beta}, y))) \tag{3.5}$$

Según Dobson y Barnett (2018), si el modelo de interés tiene tan buen ajuste como el modelo saturado, entonces D tiene una distribución chi-cuadrada aproximada con grados de libertad iguales a la diferencia entre el número de parámetros de los modelos, es decir,

$$D \sim \chi^2_{n-p}$$

Además, la distribución chi-cuadrada es exacta cuando la distribución de los Y_i es normal. Se pueden obtener las expresiones para los desvíos de las distribuciones más utilizadas en la práctica:

Tabla 3.2: Desvíos de distribuciones comunes de la familia exponencial.

Familia	**Desvío**
Normal	$\sum_i (y_i - \hat{\mu}_i)^2$
Poisson	$2\sum_i (y_i \log(y_i/\hat{\mu}_i) - (yi - \hat{\mu}_i))$
Binomial	$2\sum_i (y_i \log(y_i/\hat{\mu}_i) + (m - yi)\log(\frac{m-y_i}{m-\hat{\mu}_i})$
Gamma	$2\sum_i (-\log(y_i/\hat{\mu}_i) + (yi - \hat{\mu}_i)/\hat{\mu}_i)$
Normal inversa	$2\sum_i (y_i - \hat{\mu}_i)^2/(\hat{\mu}_i^2 y_i)$

El valor esperado de una distribución chi-cuadrada es igual a sus grados de libertad, luego se puede determinar la bondad de ajuste del modelo empíricamente al comparar el desvío dado en (3.5) con el valor de $n - p$, por tanto, el modelo va a tener un buen ajuste si el cociente $D/(n - p)$ es cercano a 1.

Comúnmente en la práctica, la varianza de los datos es mucho mayor que la esperada por el MLG elegido, por tanto se van a subestimar los errores estándar, lo que a su vez va a llevar a que los test de significancia de los predictores sean rechazados con mayor frecuencia, es decir, se infla la probabilidad de error tipo I. Este fenómeno se conoce como **sobredispersión**, y se presenta cuando existen indicios de falta de ajuste en el modelo. Para incluir la sobredispersión en el modelo, se han desarrollado MLGs que ajustan la variación del parámetro ϕ de acuerdo al tipo de respuesta utilizada. En particular, para corregir la sobredispersión en modelos binomiales se usa la *regresión Beta Binomial*, y en el caso de la Poisson se usa la *Regresión Binomial Negativa*.

En general, si el objetivo es comparar dos modelos encajados se puede usar la diferencia entre los desvíos como estadístico de prueba para el test de la razón de verosimilitud. Sea β_1 el vector de p efectos fijos del modelo mayor M_1, y sea β_0 el vector de q efectos fijos del modelo menor M_0, luego se considera el estadístico del cambio en los desvíos:

$$\Delta D = D_0 - D_1 = 2(\log(L(\hat{\beta}_0, y)) - \log(L(\hat{\beta}_1, y)))$$

el cual tiene aproximadamente una distribución chi-cuadrada con $p-q$ grados de libertad cuando el ajuste de M_1 no presenta una diferencia significativa

al ajuste de M_0. Si se considera la hipótesis dada por:

$$H_0: \ \beta = \beta_0$$
$$H_1: \ \beta = \beta_1$$

entonces se rechaza H_0 a favor del modelo más grande si $\Delta D > \chi^2_{p-q}$.

Otra medida de bondad de ajuste que se puede utilizar en lugar del desvío es el **$\mathbf{X^2}$ de Pearson**, dado por:

$$X^2 = \sum_{i=1}^{N} \frac{w_i(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \tag{3.6}$$

el cual también tiene una distribución chi-cuadrada aproximada con $n - p$ grados de libertad.

Según Faraway (2016a) el test de razón de verosimilitud no se puede realizar directamente con MLGs que tienen un parámetro de dispersión ϕ, en este caso, si se obtiene una estimación de ϕ, se puede definir un estadístico basado en la diferencia de desvíos:

$$F = \frac{(D_0 - D_1)/(p - q)}{\hat{\phi}} \tag{3.7}$$

donde una buena estimación de ϕ es $\hat{\phi} = X^2/(n - p)$, con X^2 dado en (3.6). En general, el estadístico en (3.7) tiene una distribución F aproximada con $p - q$ grados de libertad en el numerador y $n - p$ en el denominador cuando la hipótesis nula es verdadera. Esta distribución es exacta en el caso de que el modelo sea Gaussiano. Luego se rechaza H_0 a favor del modelo mayor si $F > F_{p-q,n-p}$.

Para probar la hipótesis sobre el vector de p efectos fijos β también se puede usar el **estadístico de Wald**. Si U es el vector de score, la matriz $\mathfrak{J} = Var(U)$, se conoce como *matriz de información*, luego se denota como $\mathfrak{J}(\hat{\beta})$ a la matriz de información evaluada en $\hat{\beta}$. Según Dobson y Barnett (2018), el estadístico de Wald, el cual está dado por $(\hat{\beta} - \beta)^\top \mathfrak{J}(\hat{\beta})(\hat{\beta} - \beta)$, sigue una distribución chi-cuadrada con p grados de libertad, es decir,

$$(\hat{\beta} - \beta)^\top \mathfrak{J}(\hat{\beta})(\hat{\beta} - \beta) \sim \chi^2_p$$

Los autores demuestran que $\hat{\beta} \sim N(\beta, \mathfrak{J}^{-1})$, por tanto se puede probar la hipótesis

$$H_0: \ \beta_i = \beta_{i0}$$
$$H_1: \ \beta_i \neq \beta_{i0}$$

Si $\mathfrak{J}_{ii}$ denota el i-ésimo elemento en la diagonal de $\mathfrak{J}^{-1}$, se puede usar el siguiente estadístico:

$$Z_i = \frac{\hat{\beta}_i - \beta_{i0}}{\sqrt{\mathfrak{J}_{ii}}} \tag{3.8}$$

el cual va a tener una distribución Normal estándar cuando H_0 es verdadera. Es de notar que esta prueba es la que se realiza en el resumen de un MLG ajustado con la función *glm*(). En lugar del estadístico de Wald también se puede usar el **estadístico de score**, el cual también sigue una distribución chi-cuadrada con p grados de libertad:

$$U^\top \mathfrak{J}^{-1} U \sim \chi^2_p$$

La distribución normal en la prueba de Wald es asintótica, excepto en el caso de una respuesta Gaussiana, por otro lado, Hauck Jr y Donner (1977) mostraron que en algunos casos, el estadístico dado en (3.8) decrece hacia 0 cuando las estimaciones de los parámetros son extremas, pues los errores estándar pueden estar sobrestimados y en consecuencia se obtienen p-valores grandes que pueden pasar por alto la significancia de algún predictor. Esta situación se conoce como el *efecto o problema de Hauck y Donner*, y se presenta especialmente cuando los datos son escasos, por lo tanto es preferible utilizar el test de la razón de verosimilitud en lugar de la prueba de Wald para realizar inferencias sobre β.

3.3. Diagnóstico

Los métodos de diagnóstico que se han desarrollado para los modelos lineales generalizados no son muy diferentes a las técnicas que se utilizan en los modelos de regresión estándar. De nuevo, se resalta la ventaja de los métodos gráficos sobre los tests formales en el momento de obtener conclusiones acerca de la adecuación y la validación del modelo. Las suposiciones a revisar para un MLG son las siguientes:

1. Las respuestas y_i deben ser independientes, y deben proceder de la misma familia exponencial.
2. La función de enlace $g(\cdot)$ utilizada es correcta.
3. La función de varianza $V(\mu)$ utilizada es correcta.
4. La relación de los predictores con η es de carácter lineal.

5. El parámetro de dispersión ϕ se asume constante.

6. El modelo no contiene datos atípicos.

Los residuos son la base de la verificación de supuestos en un modelo de regresión, para los MLGs, unos posibles candidatos serían los *residuos netos o de respuesta*, dados por $y_i - \hat{\mu}_i$, sin embargo, estos últimos no son útiles para realizar diagnósticos, pues generalmente, dependiendo de la distribución considerada, los errores aleatorios no necesariamente van a presentar una distribución normal o una varianza constante. Para considerar un tipo de residuo más intuitivo, se definen los **residuales de Pearson $\mathbf{r_{p_i}}$** como:

$$r_{p_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)/w_i}}$$

Los residuales de Pearson cumplen que $\sum_{i=1}^{N} r_{p_i} = X^2$, donde X^2 está dado en (3.6). También se puede usar el desvío dado en (3.5) para definir residuales, luego los **residuales de desvío $\mathbf{r_{d_i}}$** se definen tales que $\sum_{i=1}^{N} r_{d_i} = D = \sum_{i=1}^{N} d_i$, donde los d_i se conocen como *desvíos unitarios*, por tanto:

$$r_{d_i} = sign(y_i - \hat{\mu}_i)\sqrt{w_i d_i}$$

Según Dunn, Smyth et al. (2018), bajo las condiciones adecuadas, los residuales de Pearson y los de desvío van a seguir una distribución normal aproximada. Los autores indican que existen casos en los que la distribución de los residuos de desvío y de Pearson claramente no es normal, especialmente cuando las distribuciones son discretas, por tanto recomiendan el uso de los **residuales de cuantil**, los cuales siguen una distribución normal exacta, y evalúan mejor la tendencia presente en los gráficos de diagnóstico en comparación con los residuos de desvío y de Pearson.

Sea $F(y_i, \mu_i, \phi)$ la función de distribución acumulada (o FDA) de una variable aleatoria Y, luego los residuales de cuantil están dados por:

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

donde $\Phi(\cdot)$ es la FDA de la distribución normal estándar.

En el contexto de los MLGs, la matriz *hat* H depende en parte de los pesos utilizados en el algoritmo de MPCI, por tanto si W es la matriz de pesos dada en (3.4), entonces

$$H = W^{1/2} X (X^\top W X)^{-1} X^\top W^{1/2}$$

Los elementos de la diagonal de H, h_{ii} son los *leverages* de cada observación, y si su valor es grande, pueden dar un indicio de los puntos que pueden ser influyentes sobre la respuesta. Los leverages también se pueden usar para estandarizar los residuos de un MLG, luego los **residuales estandarizados** de Pearson, de desvío y de cuantil son respectivamente:

$$r_{sp_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{(\hat{\phi}/w_i)V(\hat{\mu}_i)(1-h_{ii})}}$$

$$r_{sd_i} = \frac{sign(y_i - \hat{\mu}_i)\sqrt{w_i d_i}}{\sqrt{\hat{\phi}(1-h_{ii})}}$$

$$r_{sq_i} = \frac{r_{q_i}}{\sqrt{1-h_{ii}}}$$

Los residuales definidos se pueden utilizar para identificar puntos atípicos. También se pueden usar los residuos de la **aproximación de Williams** (Williams, 1984), $\mathbf{r_{G_i}}$, los cuales son una media ponderada de los residuos de desvío y los de Pearson, y se pueden considerar como análogos a los residuos estudentizados en regresión lineal estándar:

$$r_{G_i} = sign(y_i - \hat{\mu}_i)\sqrt{(1-h_{ii})r_{sd_i}^2 + h_{ii}r_{sp_i}^2}$$

Para realizar la revisión de supuestos se puede usar cualquiera de los residuos definidos anteriormente, sin embargo, Pierce y Schafer (1986) determinan que los residuales de desvío son los que mejor se aproximan a la normalidad entre todos los tipos de residuos, por tanto los autores recomiendan usarlos para los diagnósticos gráficos. Por otro lado, Dunn, Smyth et al. (2018) sugieren el uso de residuos de cuantil cuando la variable respuesta sigue una distribución discreta como la Poisson o la Binomial.

Uno de los gráficos mas útiles para el diagnóstico de MLGs, y de modelos de regresión en general, es el de residuales contra valores ajustados $\hat{\mu}$, sin embargo, según Faraway (2016a), usualmente es mejor utilizar el predictor lineal η en lugar de los valores ajustados, por tanto se puede usar el **gráfico de $\mathbf{r_{sd_i}}$ o $\mathbf{r_{sq_i}}$ contra $\hat{\eta}_\mathbf{i}$** para verificar el ajuste y los supuestos distribucionales del modelo. Si el modelo es correcto, se espera ver una varianza constante en el gráfico, cualquier tendencia identificada indica que se puede mejorar el componente sistemático del modelo. También se puede obtener información adicional al graficar los residuos contra cada predictor.

La preferencia de $\hat{\eta}$ sobre $\hat{\mu}$ para el gráfico de diagnóstico se debe a la necesidad de escoger una escala apropiada para los valores ajustados. Un

método alternativo consiste en estabilizar la varianza de $\hat{\mu}$ con una transformación adecuada, conocida como la *transformación de escala de información constante*. Según Dunn, Smyth et al. (2018), las transformaciones de escala más usadas en los distintos MLGs son las siguientes:

Tabla 3.3: transformaciones de escala de información constante para los MLGs más comunes.

Familia	**Transformación**
Binomial	$\sin^{-1}(\sqrt{\hat{\mu}})$
Poisson	$\sqrt{\hat{\mu}}$
Gamma	$\log(\hat{\mu})$
Normal Inversa	$1/\sqrt{\hat{\mu}}$

McCullagh y Nelder (1989) sugieren usar un **gráfico de $|\mathbf{r_{sd_i}}|$ contra $\hat{\mu}_\mathbf{i}$**, escalado con su transformación correspondiente, para determinar si la elección de función de varianza es correcta, en este caso, el gráfico no debería mostrar ninguna tendencia.

Para estudiar la relación entre la respuesta y los predictores se pueden realizar gráficos de dispersión, sin embargo hay que tener en cuenta la función de enlace utilizada, por lo que Faraway (2016a) recomienda utilizar en lugar de la respuesta a la **respuesta linealizada** dada por:

$$z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\frac{d\eta_i}{d\mu_i}$$

Los gráficos de dispersión que usan la respuesta linealizada deberían presentar una tendencia lineal si la relación entre las variables es correcta. Por otro lado, se puede realizar un **gráfico de z_i contra $\hat{\eta}_i$** para evaluar si la función de enlace fue elegida correctamente, en este caso, el gráfico debería presentar una tendencia lineal. La presencia de curvatura en este gráfico indica que se debe considerar otra función de enlace. La respuesta linealizada se puede escribir como $z_i = \hat{\eta}_i + e_i$, donde los valores de e_i se conocen como *residuos linealizados*, y están dados por $e_i = (y_i - \hat{\mu}_i)\frac{d\eta_i}{d\mu_i}$.

A partir de los residuos linealizados se pueden calcular los **residuales parciales** respecto a cada covariable, los cuales se definen como:

$$u_j = e_i + \hat{\beta}_j x_j$$

Faraway (2016a) indica que se pueden realizar **gráficos de $\mathbf{u_j}$ contra $\mathbf{x_j}$** también conocidos como *gráficos de residuales parciales*, para considerar el

efecto adicional del resto de predictores cuando se examina la variable de interés. Dunn, Smyth et al. (2018) recomiendan usar estos gráficos para determinar si la variable predictora x_j está incluida en la escala correcta, si este es el caso, los gráficos deben presentar tendencias aproximadamente lineales.

En general, no se espera que los residuos de un MLG sigan una distribución normal (a excepción del caso Gaussiano), sin embargo, Dunn, Smyth et al. (2018) indican que se puede realizar un **Q-Q plot de los residuales de cuantil** para determinar si la elección de la distribución y_i es apropiada. El Q-Q plot también se puede utilizar para detectar datos atípicos, pero Faraway (2016a) recomienda usar en su lugar un **gráfico medio Normal de los residuos**, mientras que los residuales que más se utilizan para detectar datos atípicos son los residuales estudentizados, en este caso r_{G_i}. El gráfico medio normal compara el valor absoluto de los residuos ordenados con los cuantiles de la distribución *medio Normal* (si $X \sim N(0, \sigma^2)$, la distribución de la variable $Y = |X|$ se denomina **Medio Normal**), y está implementado en el paquete *faraway* en R (Faraway, 2016b). Una observación se considera atípica si se aleja mucho de la tendencia del resto de puntos en el gráfico.

Para identificar observaciones influyentes se pueden usar los estadísticos como el DFFITS, los DFBETAS, el COVRATIO o la distancia de Cook, siendo este último el que más se utiliza en la práctica. La distancia de Cook, se define en un MLG a partir del algoritmo de MCPI como sigue:

$$C_i(\hat{\beta}) = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^\top (X^\top W X)(\hat{\beta}_{(i)} - \hat{\beta})}{p\hat{\phi}}$$

Y se interpreta de la misma manera que en un modelo lineal estándar. Por lo que se puede realizar un **Gráfico de índices de $\mathbf{C_i}(\hat{\beta})$, o un gráfico medio Normal de $\mathbf{C_i}(\hat{\beta})$** para detectar observaciones influyentes sobre la estimación de los parámetros del modelo.

En el caso de regresión logística, los gráficos de residuos no proveen mucha información pues la respuesta sólo puede tomar dos valores, por tanto Dobson y Barnett (2018) recomiendan utilizar estadísticos de bondad de ajuste como X^2 y D para diagnosticar el modelo. También se puede usar el **test de Brown (C. C. Brown, 1982)** para seleccionar el enlace de logit apropiado para el modelo.

Si se identifican problemas en los gráficos de diagnóstico se pueden realizar transformaciones de las variables independientes, no es útil transformar la respuesta, pues su distribución puede cambiar, lo que hace que el modelo sea inválido. Otros métodos que pueden corregir violaciones en los supuestos del MLG son la estimación de quasi-verosimilitud (McCullagh y Nelder,

1989), la estimación robusta (Zeileis, 2004) o la estimación de *sandwich* (Künsch et al., 1989).

3.4. Definición del modelo lineal generalizado mixto (MLGM)

En el modelo lineal generalizado se asume que las observaciones son independientes entre ellas. Este supuesto no es razonable en un estudio de medidas repetidas debido a la correlación presente entre las observaciones dentro de cada unidad. Cuando la respuesta no sigue una distribución normal, se puede modelar la correlación entre observaciones agregando efectos aleatorios al MLG. En este caso, la distribución de la respuesta también pertenece a la familia exponencial, por tanto, para la j-ésima observación de la unidad i, la distribución de Y_{ij}, la cual está condicionada por el vector de efectos aleatorios b_i, se puede escribir como:

$$f(y_{ij}, \theta, \phi \mid b_i) = \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi)} + c(y_{ij}, \phi)\right\} \tag{3.9}$$

donde θ_{ij} es el parámetro canónico, ϕ es el parámetro de dispersión, y las funciones $a(\cdot)$, $b(\cdot)$, y $c(\cdot)$ son las mismas dadas en (3.1). Si se utiliza el enlace canónico, entonces el **modelo lineal generalizado mixto o MLGM** se puede representar mediante la ecuación:

$$\theta_i = \eta_i = X_i^\top \beta + Z_i^\top b_i \qquad \forall i = 1, \ldots, N \tag{3.10}$$

donde β es el vector de efectos fijos, $b_i \sim N(0, D^*)$ es el vector de efectos aleatorios, y X_i, Z_i son sus matrices de diseño correspondientes. Según Liu (2015), la matriz R que se utiliza en el MLM no se puede especificar libremente para el MLGM pues la distribución y la varianza de los errores aleatorios entre sujetos no se pueden representar explícitamente. Sin embargo, se puede incluir un término de error aleatorio para una mayor conveniencia analítica sobre la incertidumbre en los MLGMs:

$$\eta_i = X_i^\top \beta + Z_i^\top b_i + \varepsilon_i \qquad \forall i = 1, \ldots, N$$

donde $\varepsilon_i = [\varepsilon_{i1}, \ldots, \varepsilon_{in_i}]^\top$ es el vector de errores aleatorios, el cual se asume con valor esperado condicional en b_i igual a 0, es decir, $E(\varepsilon_i \mid b_i) = 0$. La especificación de la varianza de ε_i es mucho más compleja en el caso de los MLGMs, y depende de la función de enlace específica seleccionada.

Para estimar el modelo se puede maximizar la función de verosimilitud, la cual está dada por la siguiente expresión:

$$\begin{aligned} L(\beta, \phi, D^*) &= \prod_{i=1}^{N} f_i(y_i, \beta, \phi \mid b_i) \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}, \beta, \phi \mid b_i) f(b_i \mid D^*)\, db_i \end{aligned} \tag{3.11}$$

donde la integración se hace sobre la distribución de b_i. A excepción del caso cuando la respuesta es Gaussiana (MLM), la verosimilitud se vuelve muy difícil de calcular pues la integral resultante no puede evaluarse en forma cerrada, por tanto es necesario el uso de métodos numéricos para optimizar $L(\beta, \phi, D^*)$. En los últimos años se ha desarrollado una variedad de métodos para aproximar la verosimilitud, no existe un método que se pueda utilizar en cualquier situación, pues todos cuentan con ventajas y desventajas, de hecho, Liu (2015) afirma que los resultados analíticos de los diferentes métodos de aproximación tienden a ser diferentes. A continuación se presenta un resumen de las técnicas más usadas:

Linealización

Una primera aproximación para estimar el MLGM consiste en reformular la ecuación dada en (3.10) para que el modelo pueda ser ajustado como un MLM estándar. El procedimiento es iterativo, y estima los efectos fijos y los efectos aleatorios de forma separada hasta llegar a la convergencia.

Primero se realiza una estimación inicial de β, D^* y de ϕ, a partir de los cuales se puede obtener una estimación de b, luego se produce una versión linealizada de la respuesta, y^*,

$$y^* = \hat{\eta}^{(i)} + (y - \hat{\mu}^{(i)}) \left.\frac{\partial \eta}{\partial \mu}\right|_{\hat{\eta}^{(i)}}$$

después se ajusta el modelo lineal mixto correspondiente con y^* como respuesta, de donde se obtienen nuevas estimaciones de β, D^* y de ϕ las cuales pueden ser usadas en la siguiente iteración. Se puede mostrar que $E(y^* \mid b)$ y $Var(y^* \mid b)$ son consistentes con la definición del MLM (Liu, 2015), por lo que el modelo ajustado es válido para reformular el MLGM. El procedimiento anterior fue denominado por Breslow y Clayton (1993) como **quasi-verosimilitud penalizada o PQL**.

El método de PQL es popular en la práctica, pues el algoritmo converge fácilmente, es sencillo de implementar, y se puede utilizar para modelos con estructuras complejas de componentes de varianza. Según Faraway (2016a), cuando se utiliza este método, la inferencia sólo es asintóticamente correcta y los p-valores obtenidos son muy inestables, pero el mayor problema de esta metodología es que las estimaciones por PQL tienden a ser sesgadas hacia abajo, Stroup y Claassen (2020) muestran varios escenarios de simulación en donde se verifica el sesgo del PQL en comparación con otros métodos de estimación.

Aproximación integral

Si el modelo tiene pocos efectos aleatorios, se puede obtener una estimación más precisa de los parámetros del MLGM al aproximar la función de verosimilitud con métodos de integración numérica. Un método popular en el análisis de datos longitudinales es la **aproximación de Laplace**, el cual aproxima integrales de la forma

$$\int_{\hat{a}}^{\hat{b}} \exp\{Nf(z)\}\, dz$$

donde se asume que $f(z)$ es una función continua y unimodal con máximo en z_0. El punto z_0 es raíz del gradiente de f ($\Delta f(z_0) = 0$), y la matriz Hessiana de $f(z)$ evaluada en z_0 es definida positiva. La función $f(z)$ se expande en series de Taylor alrededor de z_0, y después se aproxima la log-verosimilitud del modelo de la siguiente manera:

$$\begin{aligned}\ell(\beta, \theta, \hat{b}, y) = \sum_{i=1}^{N} (\sum_{j=1}^{n_i} \log(f(y_{ij} \mid b_i)) + n_i \log(f(b_i)) \\ + \frac{1}{2} n_{b_i} \log(2\pi) - \log \left| -\frac{1}{2} n_i f''(\beta, \theta, \hat{b}_i) \right|)\end{aligned}$$

donde n_{b_i} es la dimensión de b_i. Al maximizar $\ell(\beta, \theta, \hat{b}, y)$ se obtienen los parámetros del método de Laplace. Liu (2015) indica que este método se ajusta mejor a datos longitudinales no lineales, especialmente para muestras grandes.

Otro método que también se usa regularmente en la práctica, y que puede mejorar la precisión de la estimación de los parámetros comparado a la aproximación de Laplace es el método de **cuadratura Gaussiana**, el cual aproxima integrales mediante una suma ponderada de evaluaciones de

la función dada en diferentes abscisas. Si $f(z)$ es una función continua con densidad de probabilidad dada por $p(z)$, la cuadratura Gaussiana aproxima integrales de la siguiente forma:

$$\int_{-\infty}^{\infty} f(z)p(z)\ dz \approx \sum_{q=1}^{Q} w_q f(z_q)$$

donde Q es el número de abscisas o *puntos de cuadratura* y w_q es el peso correspondiente a la evaluación en la abscisa z_q, $q = 1, \ldots, Q$. Las abscisas también se conocen como *nodos*. El método de cuadratura Gaussiana se usa para aproximar la integral dada en (3.11) después de estandarizar el vector de efectos aleatorios b para que tenga varianza igual a la identidad.

La aproximación va a ser mejor a medida que se vayan agregando más puntos de cuadratura, sin embargo, al agregar más nodos, el costo computacional va a aumentar en gran medida, especialmente para modelos con una estructura compleja de componentes de varianza. Dado el número de puntos de cuadratura, uno de los algoritmos más usado para determinar los nodos y los pesos adecuados es la **cuadratura de Gauss-Hermite**, la cual se usa para aproximar integrales de la forma:

$$\int_{-\infty}^{\infty} \exp\{-z^2\}f(z)\ dz \approx \sum_{q=1}^{Q} \tilde{w}_q f(z_q)$$

donde los puntos z_q son las raíces de los *polinomios de Hermite*, $H_Q(z)$ y los pesos están dados por $\tilde{w}_q = \frac{2^{Q-1}Q!\sqrt{\pi}}{Q^2(H_{Q-1}(z_q))^2}$.

La aproximación de Laplace es equivalente al método de cuadratura de Gauss Hermite con 1 nodo, por tanto este último requiere más recursos para su computación. Según Faraway (2016a), los métodos de aproximación integral resultan ser más precisos que los métodos de linealización, además las inferencias realizadas van a ser más confiables, dado que se está usando una aproximación a la verdadera verosimilitud en lugar de una verosimilitud basada en pseudo-variables como en el caso del PQL.

Existen otros métodos de estimación de modelos mixtos con respuesta no Gaussiana, un ejemplo son las **ecuaciones de estimación generalizadas** o **GEEs** por sus siglas en inglés, introducidas por Liang y Zeger (1986), y estudiadas por varios autores como Hardin y Hilbe (2002). Como opción adicional también se pueden utilizar los **modelos aditivos generalizados mixtos o GAMMs** (Wood, 2006). Por otro lado, también se han desarrollado métodos bayesianos como el **MCMC o cadenas de Markov Monte**

Carlo, para más información sobre los métodos bayesianos se pueden consultar los textos de Liu (2015), Gelman et al. (1995) o Correa y Salazar (2016). Los métodos bayesianos en particular son muy útiles para la estimación de los parámetros del MLGM, pues la formalización del modelo es condicional en los efectos aleatorios no observables, por lo que se puede incorporar información útil *a priori*. Faraway (2016a) indica que los métodos bayesianos cuentan con un alto grado de precisión y se pueden ajustar a modelos complejos, sin embargo requieren muchos recursos de computación, y hay que tener cuidado con los detalles de la evaluación de la bondad de ajuste.

El MLGM se define de forma similar al modelo lineal generalizado, por tanto, para probar hipótesis sobre el vector de efectos fijos β se pueden usar las mismas pruebas de la sección (3.2.2), como el test de razón de verosimilitud, la prueba de Wald o el test de Score. Similarmente, para probar hipótesis sobre la significancia de los componentes de varianza, Liu (2015) indica que se puede usar la prueba de la mezcla 50:50 de distribuciones chi-cuadradas mencionada en la sección (2.2.2) del capítulo 2.

Debido a la presencia de efectos aleatorios, y a los métodos de estimación utilizados, el análisis de diagnóstico en los MLGMs es más complejo en comparación a los MLGs. Similarmente en este caso, la varianza de los residuos en el primer nivel no necesariamente es constante, por tanto se pueden calcular los residuales de Pearson r_p al dividir los residuos netos $y_{ij} - \hat{\mu}_{ij}$ por sus errores estándar predichos. Según Jiang y Nguyen (2021), las herramientas de diagnóstico, y en particular las técnicas gráficas son muy escasas para los MLGMs. Sin embargo, se pueden realizar algunos gráficos para evaluar el ajuste del modelo, Wood (2006) emplea gráficos de los residuales de Pearson y los residuales netos contra los valores ajustados (escalados con la transformación de varianza correspondiente) para evaluar el ajuste general del modelo y determinar si existe un efecto de la variable de grupo para la inclusión de efectos aleatorios. Es tentativo usar Q-Q plots de los efectos aleatorios estimados para evaluar la suposición de normalidad, sin embargo H. Brown y Prescott (2015) indican que los gráficos de efectos aleatorios pueden ser útiles para identificar unidades atípicas pero puede que no siempre detecten una falta de normalidad. Por otro lado, Jiang y Nguyen (2021) describen varios tests formales que se pueden utilizar para evaluar la bondad de ajuste del MLGM.

Capítulo 4

Aplicaciones

En el presente capítulo se realizan los análisis de diagnóstico de tres bases de datos aplicando los métodos descritos a lo largo de este trabajo. El primer ejemplo se presenta en la sección 4.1, y en éste se ilustra el uso de varias de las técnicas de diagnóstico para modelos mixtos expuestas en la sección 2.3. En el segundo ejemplo, dado en la sección 4.2, se ajusta un modelo lineal generalizado, y se exploran los métodos de diagnóstico de la sección 3.3. Por último, en la sección 4.3 se presenta un ejemplo corto del ajuste de un modelo lineal generalizado mixto.

4.1. Ultrafiltración de dializadores

El primer ejemplo de aplicación utiliza los datos descritos por Vonesh y Carter (1992) en donde se analizaron las características del transporte de agua de 20 dializadores de membrana de alto flujo (un dializador es un filtro a través del que se bombea la sangre durante el proceso de hemodiálisis), con el objetivo de evaluar sus características de ultrafiltración *in vivo* para reducir el tiempo que un paciente se gasta en hemodiálisis. Se estudiaron 20 dializadores *in vitro* usando sangre bovina con dos velocidades de flujo diferentes, 200 y 300 dL/min. La variable respuesta es la razón de ultrafiltración o RUF en ml/hr, la cual fue medida en 7 presiones diferentes de transmembrana (en dmHg). Los datos, los cuales se presentan en la tabla 4.1, están disponibles en el objeto *Dialyzer* de la librería *MEMSS* (Bates et al., 2019).

Sea Y_{ij} las j-ésima razón de ultrafiltración para el dializador i, $i = 1, \ldots, 20$, $j = 1, \ldots, 7$. Sea X_{1ij} su respectiva medida de presión, y sea X_{2ij} la variable indicadora que clasifica a cada observación de acuerdo a la razón

de flujo usada QB ($X_2 = 1$ si el flujo es de 300 dL/min, 0 en caso contrario). La velocidad de 200 dL/min se aplicó a los primeros 10 dializadores, mientras que la velocidad de 300 dL/min fue aplicada para el resto. En la tabla 4.1 se puede observar la estructura de los datos utilizados:

Tabla 4.1: Razones de ultrafiltración en los 20 dializadores, en total se obtuvieron 140 medidas.

	X_2	X_1	Y
Dializador	QB	presión	RUF
1	200	0.24	0.64
1	200	0.50	20.11
1	200	0.99	38.46
1	200	1.49	44.98
1	200	2.02	51.77
1	200	2.50	46.58
1	200	2.97	40.81
2	200	0.24	3.72
2	200	0.54	18.89
2	200	0.99	34.70
⋮	⋮	⋮	⋮
19	300	2.02	60.44
19	300	2.50	64.83
19	300	2.98	63.83
20	300	0.40	10.94
20	300	0.47	13.47
20	300	1.01	35.35
20	300	1.51	45.34
20	300	1.98	49.44
20	300	2.51	53.62
20	300	3.00	56.43

La figura 4.1 muestra los perfiles de cada dializador respecto a la presión utilizada para cada una de las razones de flujo. A partir del gráfico se observa que la RUF aumenta junto a la presión usada siguiendo una tendencia polinomial, mientras que a su vez también aumenta la variabilidad de los perfiles. Por otro lado, se presenta una diferencia clara en la respuesta para cada una de las razones de flujo utilizadas.

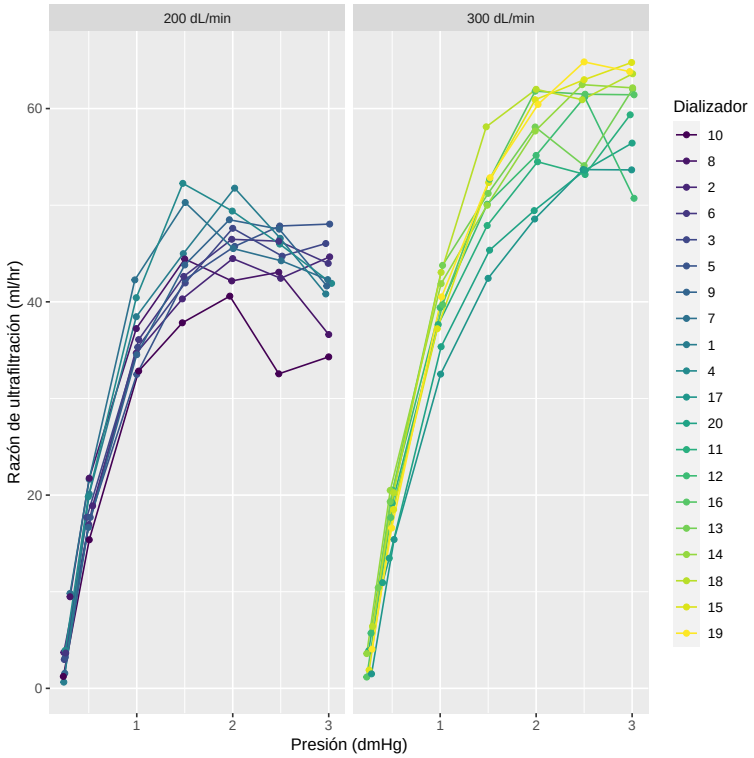

Figura 4.1: Razón de ultrafiltración contra la presión de transmembrana diferenciando por las velocidades de flujo para cada uno de los dializadores.

En este caso se asume que las medidas de RUF dentro de cada dializador no son independientes entre sí, luego se puede formular un modelo inicial de intercepto aleatorio como sigue:

$$Y_{ij} = (\beta_0 + b_{0i}) + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \varepsilon_{ij}, \tag{4.1}$$

donde b_{0i} representa el intercepto aleatorio correspondiente al dializador i. De manera similar se ajustó un modelo de mínimos cuadrados generalizados sin b_{0i}, el cual está dado por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \tag{4.2}$$

La tabla 4.2 muestra la estimación de los modelos (4.1) y (4.2), los cuales fueron ajustados con las funciones *gls()* y *lme()* (Pinheiro et al., 2020) respectivamente.

Tabla 4.2: Estimación de los modelos de mínimos cuadrados generalizados y de intercepto aleatorio.

	Variable dependiente:	
	Razón de ultrafiltración	
	Modelo gls	Modelo mixto
	(1)	(2)
presión	16.151***	16.151***
	(0.884)	(0.884)
QB300	7.383***	7.383***
	(1.667)	(1.667)
Intercepto	8.874***	8.874***
	(1.797)	(1.797)
Número de unidades		20
sd(Intercepto aleatorio)		0.0004
N_T	140	140
Nota:	*p<0.1; **p<0.05; ***p<0.01	

Las estimaciones de los efectos fijos de ambos modelos son iguales, por otro lado, la estimación de la varianza del intercepto aleatorio en el modelo (4.1) es muy pequeña, por lo que es posible que los componentes de varianza incluidos no son significativos. El test de razón de verosimilitud presente en la tabla 4.3 muestra que el modelo no mejora al incluir el intercepto aleatorio.

Tabla 4.3: Test de razón de verosimilitud para los modelos de mínimos cuadrados generalizados (mod_gls) y de intercepto aleatorio (mod1)

	Model	df	AIC	BIC	logLik	L.Ratio	p.value
mod_gls	1	4.00	1043.12	1054.89	-517.56		
mod1	2	5.00	1045.12	1059.83	-517.56	0.00	1.00

Como alternativa para probar si los componentes de varianza de los efectos aleatorios son significativos se puede utilizar la función *exactRLRT()*

del paquete *RLRsim* (Scheipl et al., 2008). Se determinó que el intercepto aleatorio no es significativo, sin embargo, la figura 4.1 sugiere la inclusión de términos polinomiales para la presión, por lo que puede que la especificación del modelo en (4.1) y (4.2) no sea adecuada. Por lo tanto se considera un nuevo modelo mixto con términos cuadráticos y cúbicos para la presión:

$$Y_{ij} = (\beta_0 + b_{0i}) + \beta_1 X_{1ij} + \beta_2 X_{1ij}^2 + \beta_3 X_{1ij}^3 + \beta_4 X_{2ij} + \varepsilon_{ij} \tag{4.3}$$

Similarmente se define el modelo equivalente ajustado con mínimos cuadrados generalizados sin intercepto aleatorio:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \beta_3 X_{1i}^3 + \beta_4 X_{2i} + \varepsilon_i \tag{4.4}$$

Los resultados del test de razón de verosimilitud presentes en la tabla 4.4 revelan que el modelo (4.4) mejora en comparación al modelo en (4.2) sin términos polinomiales. Adicionalmente, se determina que la inclusión del intercepto aleatorio si es significativa para el modelo en (4.3).

Tabla 4.4: Test de razón de verosimilitud para los modelos dados en (4.2) (mod_gls), (4.4) (mod_gls2) y (4.3) (mod2)

	df	AIC	BIC	logLik	Test	L.Ratio	p.value
mod_gls	4.00	1043.12	1054.89	-517.56			
mod_gls2	6.00	853.75	871.40	-420.88	1 vs 2	193.37	0.00
mod2	7.00	848.78	869.37	-417.39	2 vs 3	6.97	0.01

Al realizar de nuevo el LRT, se determinó que el modelo de intercepto aleatorio mejora aún más al agregar una pendiente aleatoria b_{1i} para la presión. El nuevo modelo está dado por:

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) X_{1ij} + \beta_2 X_{1ij}^2 + \beta_3 X_{1ij}^3 + \beta_4 X_{2ij} + \varepsilon_{ij} \tag{4.5}$$

Por otro lado, se determinó que la inclusión de un término de grado 4 al modelo en (4.5) no mejora el ajuste. Los resultados se muestran en la tabla 4.5.

Por lo tanto, el modelo seleccionado para realizar el análisis de diagnóstico es dado en la ecuación (4.5), el cual tiene un intercepto aleatorio para cada dializador considerado y una pendiente aleatoria para la presión de transmembrana. Se puede visualizar el ajuste del modelo para cada dializador en la figura 4.2.

Tabla 4.5: Test de razón de verosimilitud para los modelos mixtos con términos polinomiales de intercepto aleatorio (mod2), intercepto aleatorio y pendiente para la presión (mod3) y de grado 4 (mod4).

	df	AIC	BIC	logLik	Test	L.Ratio	p.value
mod2	7.00	848.78	869.37	-417.39			
mod3	9.00	734.40	760.87	-358.20	1 vs 2	118.39	0.00
mod4	10.00	734.25	763.66	-357.12	2 vs 3	2.15	0.14

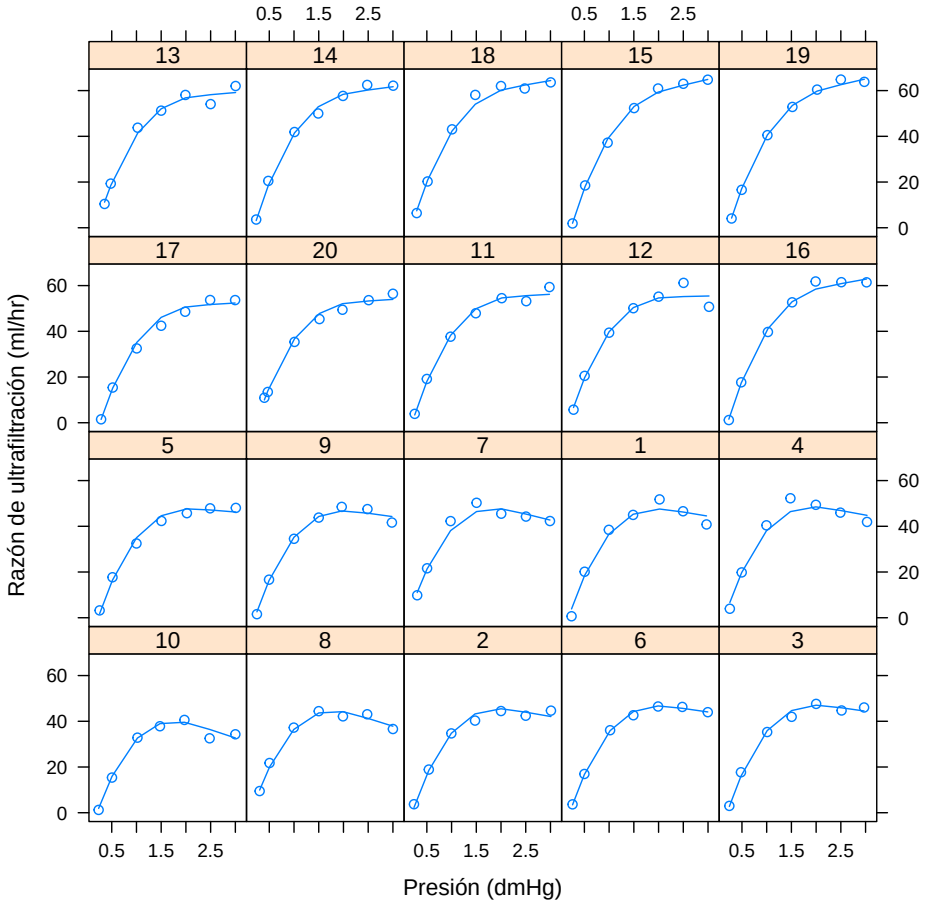

Figura 4.2: Ajuste del modelo de intercepto y pendiente aleatorios con términos polinomiales definido en (4.5).

La tabla 4.6 muestra una comparación de las estimaciones del modelo mixto en (4.5), y de su análogo sin efectos aleatorios dado en (4.4). En este caso los componentes de varianza estimados son significativos, y las estimaciones de sus desviaciones estándar son relativamente grandes en comparación con la desviación residual estimada.

Tabla 4.6: Estimaciones del modelo mixto de términos polinomiales seleccionado y de su análogo ajustado con mínimos cuadrados generalizados.

	Variable dependiente:	
	Razón de ultrafiltración	
	Modelo gls	Modelo mixto
	(1)	(2)
presión	79.097***	79.948***
	(5.845)	(2.753)
presión2	−31.145***	−31.713***
	(4.169)	(1.860)
presión3	4.016***	4.122***
	(0.840)	(0.375)
QB300	7.254***	2.288**
	(0.842)	(1.042)
Intercepto	−18.051***	−15.871***
	(2.119)	(1.360)
Número de unidades		20
sd(Intercepto aleatorio)		3.737
sd(Pend. aleatoria presión)		3.868
sd Residual	4.89	2.175
N_T	140	140
Nota:	*p<0.1; **p<0.05; ***p<0.01	

4.1.1. Análisis de diagnóstico

Para evaluar las suposiciones del modelo, se empezó con el análisis de los residuos en el primer nivel de agrupamiento. La figura 4.3 muestra los gráficos obtenidos al comparar los valores ajustados contra los residuales condicionales estandarizados $\hat{\varepsilon}^*_{ij}$ y los residuales semi-estandarizados $\check{\varepsilon}_{ij}$ dados en las ecuaciones (2.6) y (2.7) respectivamente. A partir de los gráficos se concluye que no se presentan problemas con el supuesto de linealidad de los efectos fijos, sin embargo, en ambos gráficos se presenta un caso de heterogeneidad en la varianza de las observaciones, la cual no es tan notable en el caso de $\check{\varepsilon}_{ij}$. Luego, es razonable asumir que la varianza de los datos incrementa a medida que lo hace la presión de transmembrana, y por tanto, $var(\varepsilon_i) \neq \sigma^2 I$.

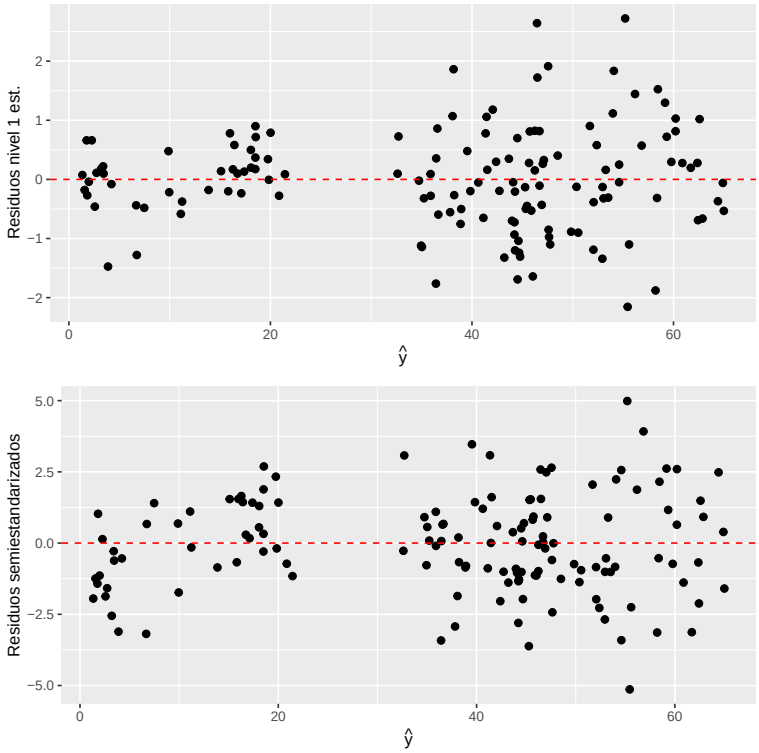

Figura 4.3: Gráficos de $\hat{\varepsilon}^*_{ij}$ y $\check{\varepsilon}_{ij}$ contra los valores ajustados.

Por otro lado, la figura 4.4 presenta los gráficos de los residuales semi-estandarizados contra cada uno de los predictores continuos, y en este caso,

el problema de heterogeneidad de varianza es menos notable, mientras que tampoco se presentan problemas de linealidad.

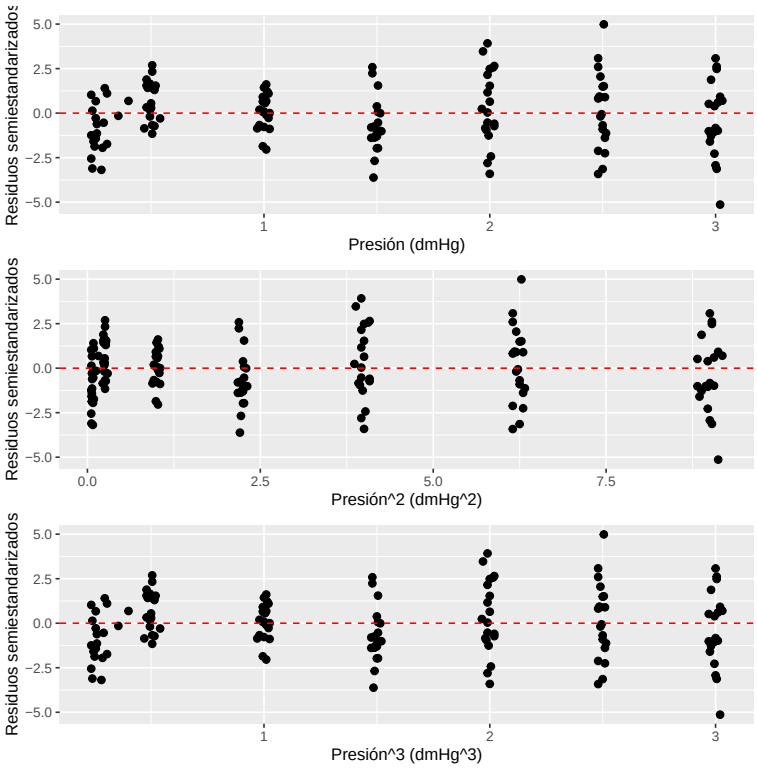

Figura 4.4: Comparación de $\check{\varepsilon}_{ij}$ contra cada uno de los predictores continuos del modelo.

Continuando con el diagnóstico, la figura 4.5 muestra el Q-Q plot de los residuales semi-estandarizados para evaluar el supuesto de normalidad de los errores condicionales. En este caso, no se presentan desvíos de linealidad en la tendencia de los puntos, por lo que no hay problemas con la distribución de ε_{ij}.

El siguiente paso es revisar la especificación del modelo en el nivel de los dializadores (nivel 2). La figura 4.6 muestra el gráfico de índices de la medida de Lesaffre-Verbeke $\mathcal{V}_i^*$, el cual se obtiene mediante la función *residdiag.nlme* creada por Singer et al. (2017). Cada punto del gráfico representa un dializador diferente, y en este caso, no se presenta ningún valor extremo, por lo que no se identificaron unidades con estructura de covarianza inadecuada.

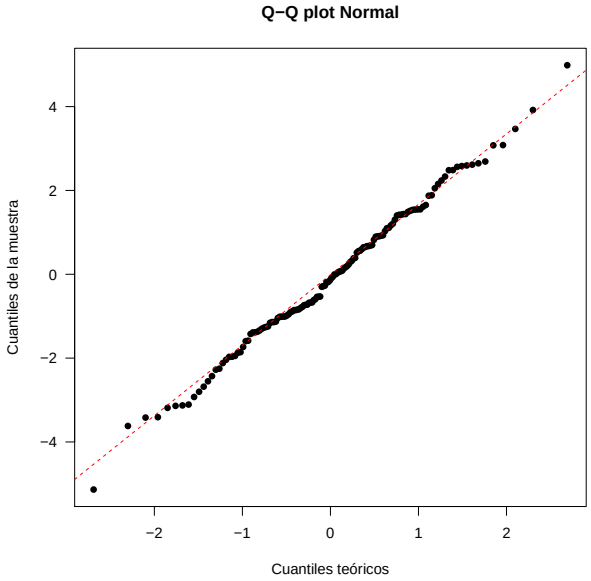

Figura 4.5: Q-Q plot de los residuos semiestandarizados.

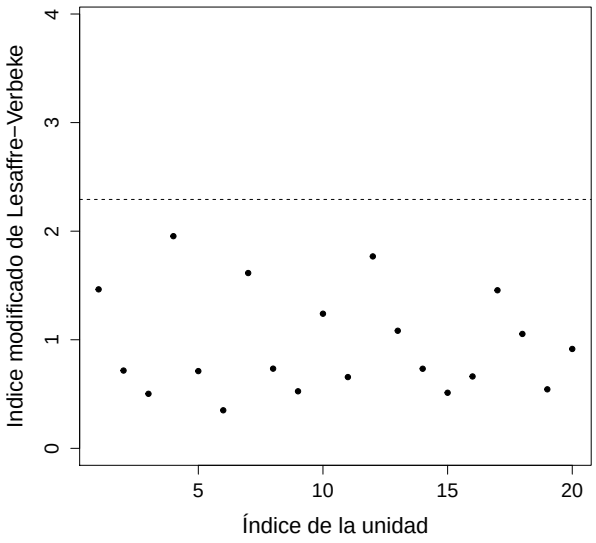

Figura 4.6: Gráfico de índices de $\mathcal{V}_i^*$.

Con el objetivo de verificar el supuesto de normalidad de los efectos aleatorios, la figura 4.7 presenta el Q-Q plot chi-cuadrado de la distancia de Mahalanobis dada en la ecuación (2.8). El gráfico no presenta problemas notables, pues la mayoría de los puntos caen cerca de la recta, luego en este caso también se puede concluir que el vector de efectos aleatorios sigue una distribución normal.

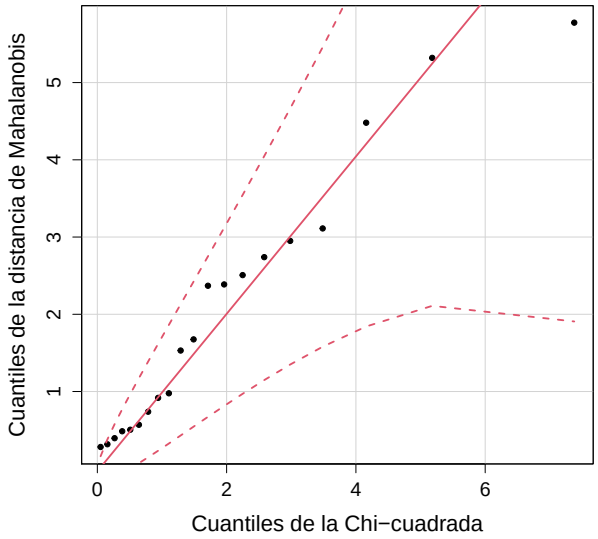

Figura 4.7: Q-Q plot chi-cuadrado de $\mathcal{M}_i$.

Antes de realizar el análisis de influencia sobre los efectos fijos estimados, es muy importante revisar si existen unidades que tengan un efecto significativo sobre la estimación de los componentes de varianza. La figura 4.8 muestra los dotplots del cambio relativo en la varianza definido en la ecuación (2.19) para la varianza residual σ^2, la varianza asociada al intercepto aleatorio de los dializadores σ_{00}, la varianza asociada a la pendiente aleatoria para la presión de transmembrana σ_{11} y para la covarianza asociada con el intercepto aleatorio y la pendiente aleatoria σ_{01}. En este caso, ningún gráfico presenta puntos extremos, el valor umbral de cada dotplot es calculado con una medida de escala interna generada por la función *dotplot_diag* del paquete *HLMdiag* (Loy y Hofmann, 2014). Se puede concluir que no hay unidades influyentes sobre la estimación de los componentes de varianza.

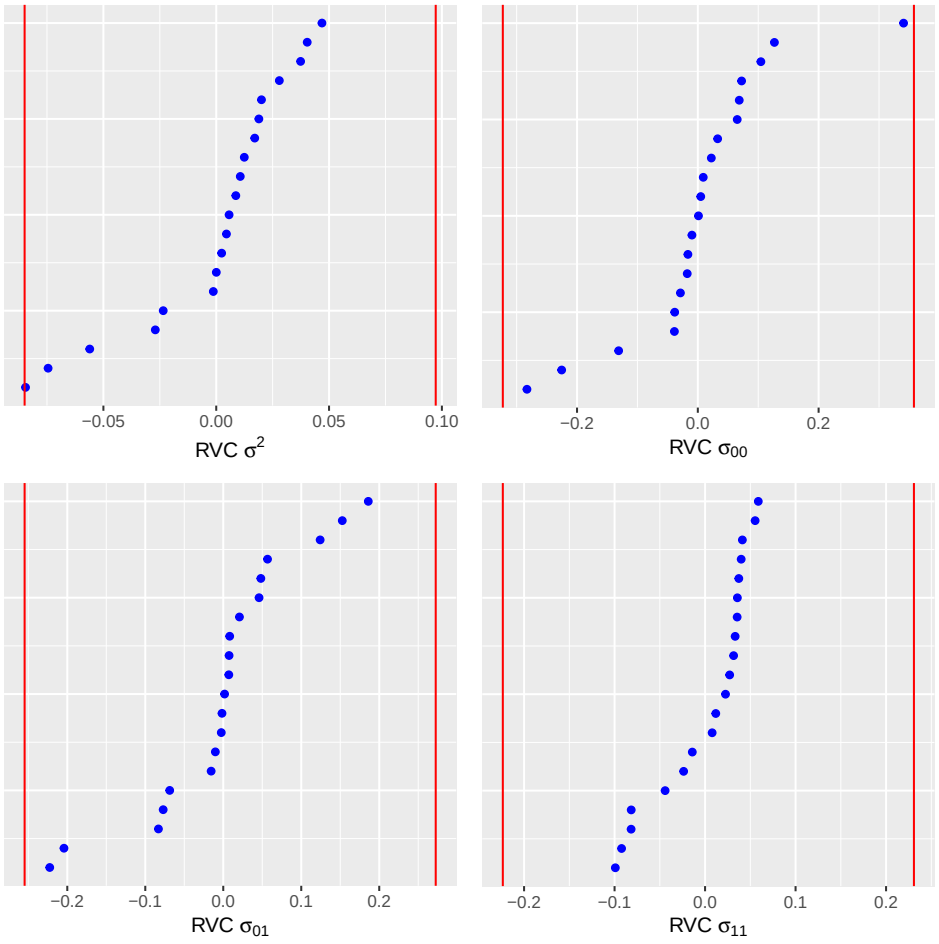

Figura 4.8: Cambio relativo en la varianza para los componentes de varianza del modelo para cada unidad, según la ecuación (2.19). El eje vertical de cada gráfico representa los índices de las unidades.

La figura 4.9 presenta el gráfico de índices de los residuales condicionales estandarizados con el objetivo de encontrar observaciones atípicas. En este caso, se destacan 3 puntos, que se identifican como la observación 25, que pertenece a la unidad 4, y las observaciones 83 y 84, las cuales pertenecen a la unidad 12. Por otro lado, el gráfico de índices de la distancia de Mahalanobis en la figura 4.10 no indica ninguna unidad como un grupo atípico potencial.

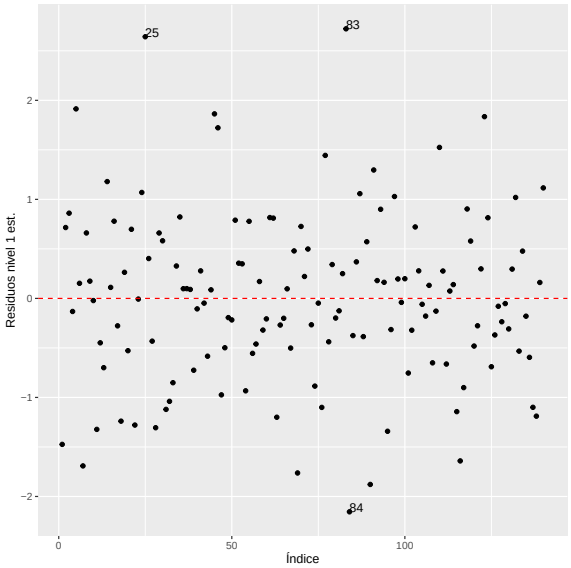

Figura 4.9: Gráfico de índices de los residuos condicionales estandarizados.

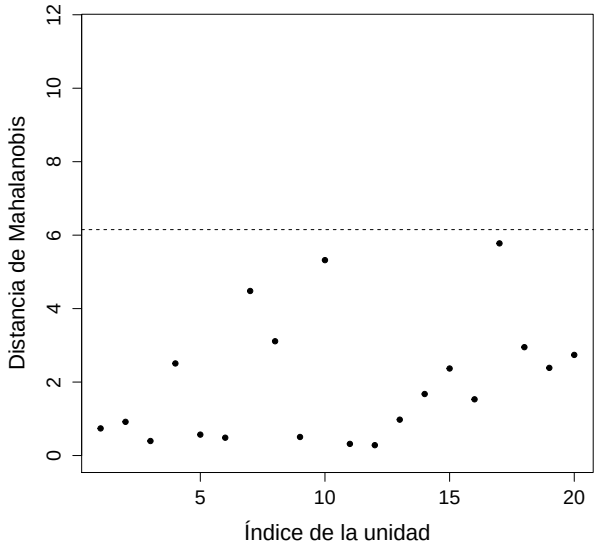

Figura 4.10: Gráfico de índices de la distancia de Mahalanobis dada en (2.8).

Por otro lado, se pueden calcular las medidas de leverage para identificar puntos atípicos adicionales, la figura 4.11 presenta dotplots de las matrices H_{1i} y H_{2i}^{*} dadas en las ecuaciones (2.9) y (2.10) respectivamente. Ninguno de los gráficos presenta unidades con leverage elevado, un análisis similar se realizó para el leverage de las observaciones en el primer nivel, y tampoco se presentaron puntos problemáticos.

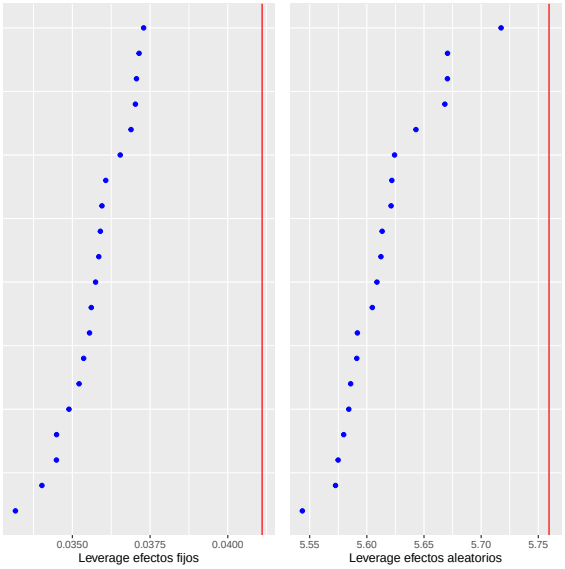

Figura 4.11: Gráficos de descomposición del leverage para efectos fijos (izquierda) y efectos aleatorios (derecha) en el nivel de los dializadores.

Para determinar cuáles unidades son influyentes sobre la estimación de los parámetros del modelo, en la figura 4.12 se realizaron dotplots de la distancia de Cook dada en (2.13) para los dializadores (arriba), a su vez se generó un dotplot para las observaciones en el primer nivel (abajo). El primer gráfico no identifica ningún grupo influyente, por otro lado, el gráfico del nivel 1 identifica varias observaciones, en particular la observación 84 se separa mucho de la tendencia del resto de puntos. También se destacan en menor medida los puntos 7 y 83. Por otro lado, la figura 4.13 muestra el gráfico de la distancia condicional de cook $D_{i(j)}^{cond}$ dada en (2.16), donde se aprecian los puntos 84, 7 y 83 (los cuales se indican en el gráfico con las notaciones 12.7, 1.7 y 12.6 respectivamente) como potenciales observaciones influyentes. Estos resultados son más confiables, sin embargo, en este caso los resultados de ambas distancias son similares.

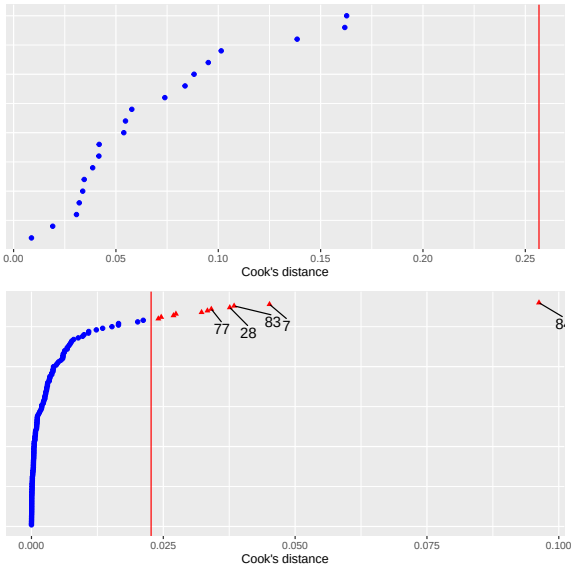

Figura 4.12: Gráficos de las distancias de cook calculadas para las unidades (arriba) y para las observaciones (abajo).

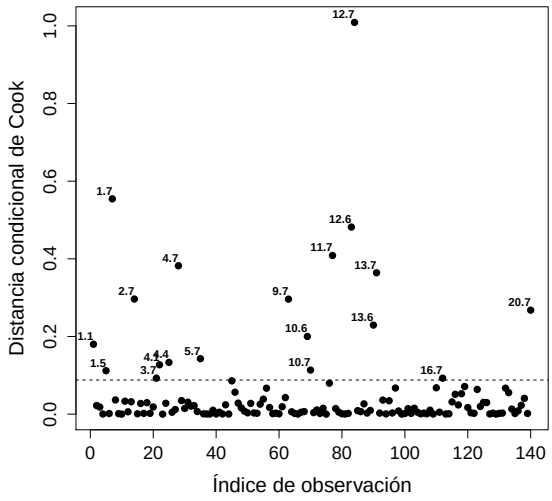

Figura 4.13: Gráfico de índices de la distancia condicional de Cook $D^{cond}_{i(j)}$.

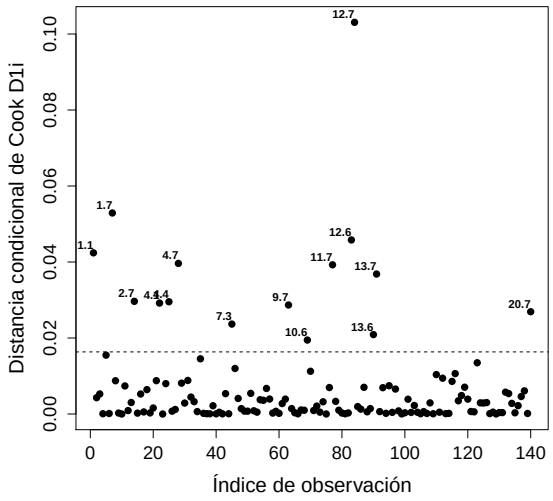

Figura 4.14: Gráfico de índices de la distancia condicional de Cook $D^{cond}_{1i(j)}$ para efectos fijos.

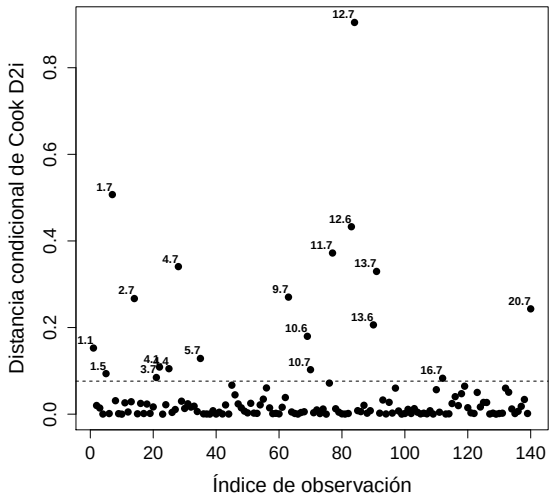

Figura 4.15: Gráfico de índices de la distancia condicional de Cook $D^{cond}_{2i(j)}$ para efectos aleatorios.

De forma similar, En las figuras 4.14 y 4.15 se realizaron los Gráficos de índices de los componentes de la distancia de cook condicional para efectos fijos $D^{cond}_{1i(j)}$ y para efectos aleatorios $D^{cond}_{2i(j)}$ respectivamente. Los resultados son similares, se concluye que los puntos identificados anteriormente pueden ser influyentes en la estimación de todos los parámetros del modelo.

En una examinación más a fondo de los puntos identificados, se encuentra que las observaciones 83 y 84 tienen valores relativamente grandes de presión y RUF, en contraste, el punto 7 cuenta con valores muy pequeños para las mismas variables. Para evaluar el cambio en la estimación, se volvió a ajustar el modelo sin incluir a estos puntos. Los resultados se presentan en la tabla 4.7 (La estimación de ambos modelos se hizo con REML).

A partir de los resultados de la tabla 4.7 se concluye que los cambios en las estimaciones de los efectos fijos son relativamente pequeños, por tanto los puntos no influyen de forma considerable en el ajuste del modelo, es de notar que los puntos tampoco modifican en gran medida los componentes de varianza estimados, por lo que se pueden conservar en el modelo.

Por último, la figura 4.16 presenta los dotplot del COVTRACE dado en (2.17) en los niveles 1 y 2, para evaluar si hay unidades u observaciones que influyen en gran medida en la estimación de $Var(\beta)$. En este caso, la precisión de los efectos fijos no se ve afectada por la presencia de puntos influyentes.

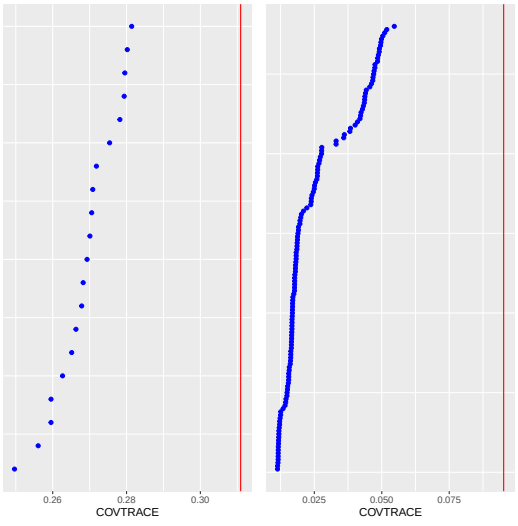

Figura 4.16: Dotplot de las medidas de COVTRACE para las unidades (izquierda) y las observaciones (derecha).

Tabla 4.7: Comparación de las estimaciónes del modelo completo con el modelo sin las observaciones 7, 83 y 84

	Variable dependiente:	
	Razón de ultrafiltración	
	Modelo completo	Modelo sin las obs 7,83,84
	(1)	(2)
presión	79.950***	81.374***
	(2.735)	(2.550)
presión2	−31.714***	−32.979***
	(1.845)	(1.712)
presión3	4.122***	4.421***
	(0.372)	(0.347)
QB300	2.290**	1.969*
	(1.079)	(1.075)
Intercepto	−15.873***	−16.078***
	(1.373)	(1.307)
Número de unidades	20	20
sd(Intercepto aleatorio)	3.886	3.763
sd(Pend. aleatoria presión)	3.972	3.958
$cor(b_0, b_1)$	-0.828	-0.812
sd Residual	2.197	2.014
N_T	140	137
Nota:		*p<0.1; **p<0.05; ***p<0.01

En resumen, la mayoría de diagnósticos utilizados no encontraron problemas en el modelo, no se identificaron unidades problemáticas, la estructura de covarianza supuesta parece razonable, no hay problemas con la estimación de los componentes de varianza, y las suposiciones distribucionales de los errores y los efectos aleatorios también se cumplen. El único problema notable es la heterogeneidad de los errores en el nivel de las observaciones, en este caso, se pueden utilizar los métodos mencionados en la sección (2.3.5) para corregir la varianza no constante en el nivel 1.

4.2. Incidencia de cáncer de piel en mujeres

La base de datos utilizada en este ejemplo está disponible en el objeto *skincancer* del paquete *glmtoolbox* (Vanegas et al., 2022), y contiene información acerca del número de casos de cáncer de piel no melanoma de una muestra de mujeres estratificadas por edad en dos ciudades de EE. UU., St. Paul y Forth Worth. Los datos se presentan en la tabla 4.8:

Tabla 4.8: Incidencia de cáncer de piel no melanoma.

casos	**ciudad**	**edad**	**población**
1	St.Paul	15-24	172675
16	St.Paul	25-34	123065
30	St.Paul	35-44	96216
71	St.Paul	45-54	92051
102	St.Paul	55-64	72159
130	St.Paul	65-74	54722
133	St.Paul	75-84	32185
40	St.Paul	85+	8328
4	Ft.Worth	15-24	181343
38	Ft.Worth	25-34	146207
119	Ft.Worth	35-44	121374
221	Ft.Worth	45-54	111353
259	Ft.Worth	55-64	83004
310	Ft.Worth	65-74	55932
226	Ft.Worth	75-84	29007
65	Ft.Worth	85+	7583

En este conjunto de datos, las 16 observaciones son representadas por cada uno de los grupos de mujeres clasificadas por rango de edad en ambas ciudades. Sea $Y_i \sim Poisson(\mu_i)$ la variable aleatoria que denota el número de casos de cáncer no melanoma en el grupo i, y sea P_i su respectiva población. El número de casos es relativo respecto al tamaño de cada grupo, por tanto la respuesta es la razón de casos Y_i/P_i. En este caso, $E(Y_i/P_i) = \mu_i/P_i$, luego si se utiliza el enlace canónico, el predictor lineal está dado por $\eta_i = \log(\mu_i/P_i)$. Así, al utilizar propiedades del logaritmo, el modelo para la razón de casos está dado por:

$$\begin{cases} Y_i \sim Poisson(\mu_i) \\ \log(\mu_i) = \beta_0 + \sum_{k=1}^{p} \beta_k x_k + \log(P_i) \end{cases} \tag{4.6}$$

donde las variables explicativas dadas por x_j son las variables dummy necesarias para representar las variables categóricas de edad y ciudad. Se puede notar que $\log(P_i)$ es un predictor que no tiene un coeficiente para estimar, estos términos se conocen como *offsets*, y son muy comunes en los modelos de razón de conteos de Poisson.

La figura 4.17 presenta la relación entre la incidencia de casos por cada 10000 habitantes respecto al grupo de edad y a la ciudad de cada individuo:

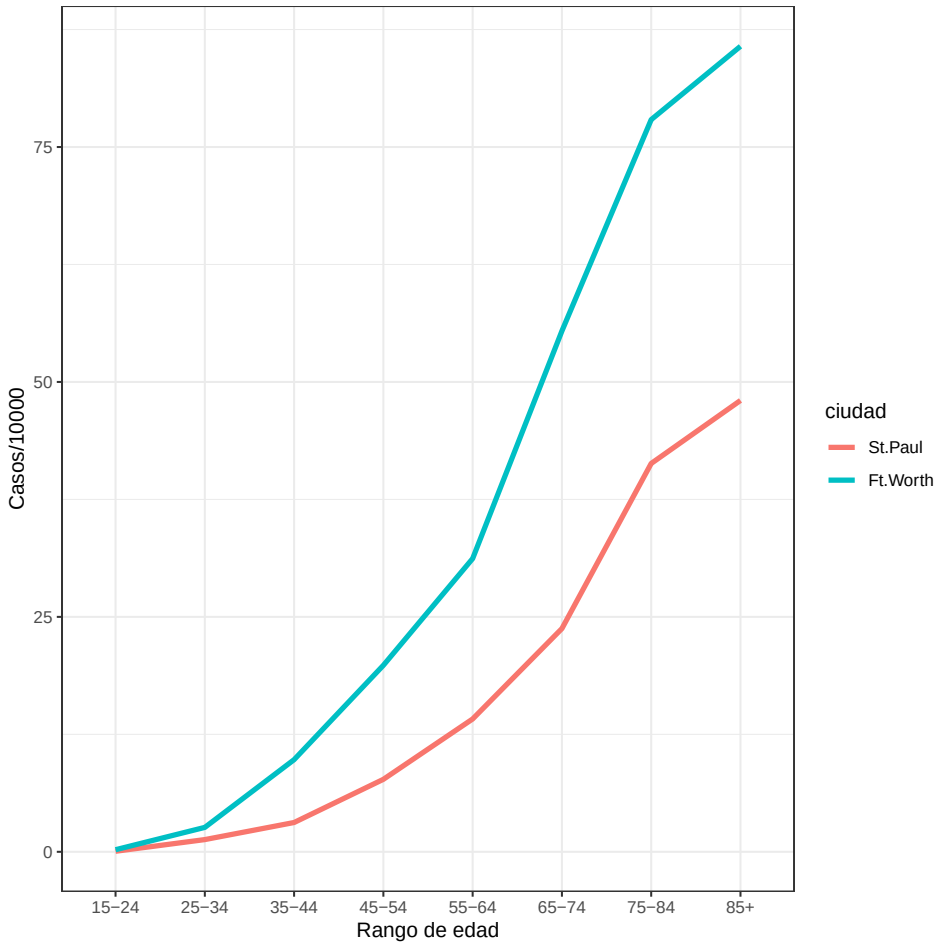

Figura 4.17: Razón de cáncer no melanoma vs rango de edad diferenciada por ciudad.

A partir del gráfico se puede observar que la incidencia del cáncer de piel aumenta proporcionalmente al rango de edad, también existe una diferencia notable entre las 2 ciudades.

La tabla 4.9 presenta los resultados del test de razón de verosimilitud con diferentes combinaciones de los predictores, *mod0*, el cual tiene como predictor a la ciudad, *mod1*, el cual tiene como predictores a la ciudad y el rango de edad, y *mod2*, quien contiene el término de interacción entre ciudad y rango de edad. Adicionalmente, se incluyó $\log(P)$ como offset en todos los modelos. A partir de la tabla, se determinó que el modelo que incluye a los predictores de edad y ciudad no mejora si se agrega el término de interacción correspondiente.

Tabla 4.9: Test de razón de verosimilitud para comparar los modelos con diferentes combinaciones de predictores.

	Resid..Df	Resid..Dev	Df	Deviance	Pr..Chi.
mod0	14.00	2569.15			
mod1	7.00	8.26	7.00	2560.89	0.00
mod2	0.00	0.00	7.00	8.26	0.31

Por otro lado, la tabla 4.10 presenta la estimación de los parámetros del modelo dado en (4.6), el cual contiene las variables de ciudad y rango de edad sin incluir el término de interacción. Según los resultados, todos los predictores resultaron significativos, además, al comparar el valor del desvío (8.2585) con sus grados de libertad (7), se concluye que el modelo cuenta con un buen ajuste, por tanto no es necesario cambiar la estructura interna del modelo para lidiar con la sobredispersión.

Tabla 4.10: Estimación de los parámetros del modelo lineal generalizado con offset definido en (4.6).

	Variable dependiente:
	Número de casos
ciudadFt.Worth	0.804^{***}
	(0.052)
edad25-34	2.630^{***}
	(0.467)
edad35-44	3.847^{***}
	(0.455)
edad45-54	4.595^{***}
	(0.451)
edad55-64	5.087^{***}
	(0.450)
edad65-74	5.645^{***}
	(0.450)
edad75-84	6.059^{***}
	(0.450)
edad85+	6.174^{***}
	(0.458)
Intercepto	-11.658^{***}
	(0.449)
Desvío nulo	2789.68
gl desvío nulo	15
Desvío residual	8.2585
gl desvío residual	7
N	16
Note:	*p<0.1; **p<0.05; ***p<0.01

4.2.1. Análisis de diagnóstico

Para evaluar el componente sistemático del modelo se realizaron los gráficos de residuales contra el predictor lineal. La distribución de Poisson es discreta, por tanto se consideran los residuales de cuantil. El gráfico se presenta en la figura 4.18, en donde no se destaca una tendencia de dispersión clara, sin embargo el gráfico se distorsiona por los 2 puntos más extremos a la izquierda. Por otro lado, el gráfico de los residuos de cuantil contra las estimaciones escaladas por varianza en la figura 4.19 no presenta datos atípicos y muestra que no hay problemas de ajuste en el modelo.

Se determinó que la función de varianza del modelo es correcta, pues el gráfico de la figura 4.20 no presenta ninguna tendencia notable. Adicionalmente se puede concluir que la función de enlace elegida, en este caso el enlace canónico, es correcta, pues la figura 4.21 no presenta ningún problema. La distribución elegida para la respuesta es adecuada, pues no aparecen desvíos de linealidad en el el Q-Q plot de la figura 4.22.

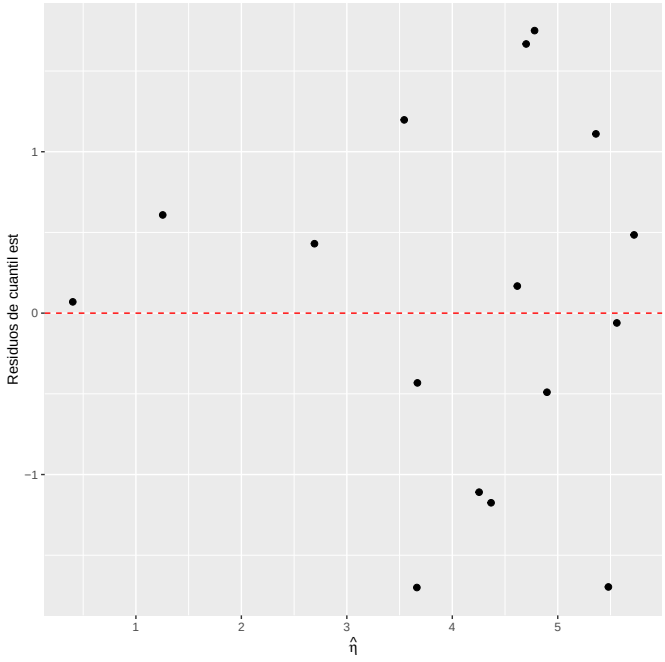

Figura 4.18: Relación entre r_{sq} y $\hat{\eta}$.

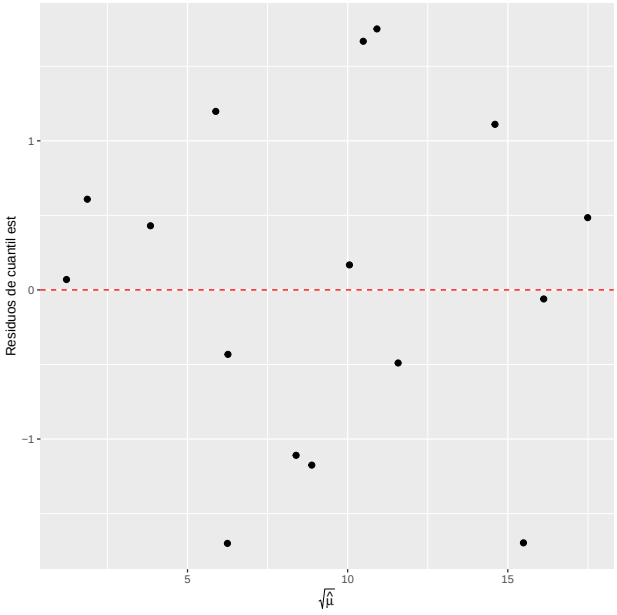

Figura 4.19: Relación entre r_{sq} y $\sqrt{\hat{\mu}}$.

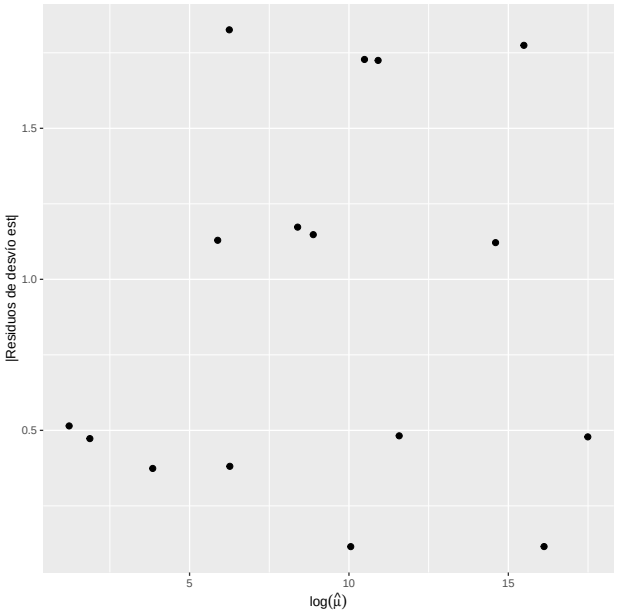

Figura 4.20: Gráfico para la elección de función de varianza.

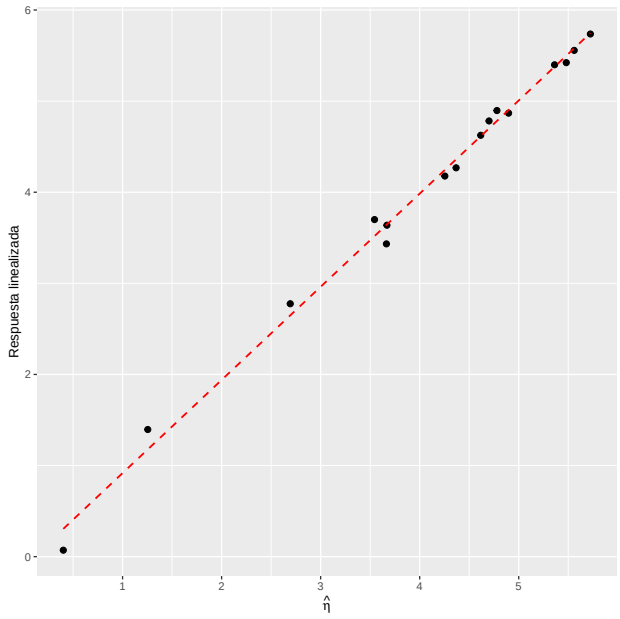

Figura 4.21: Gráfico para confirmar la elección de función de enlace.

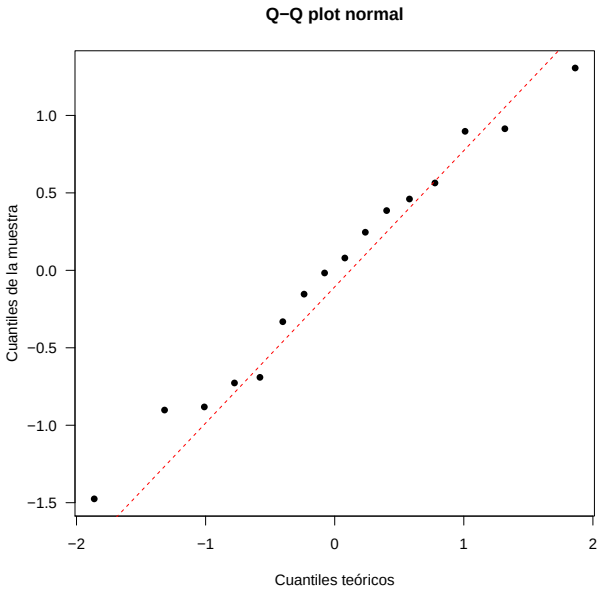

Figura 4.22: Q-Q plot de los residuos de cuantil.

Para el análisis de influencia se realizó el gráfico medio normal y el Gráfico de índices de las distancias de Cook presentes en la figura 4.23, en donde se identificaron a las observaciones 11 y 15 como posibles puntos influyentes.

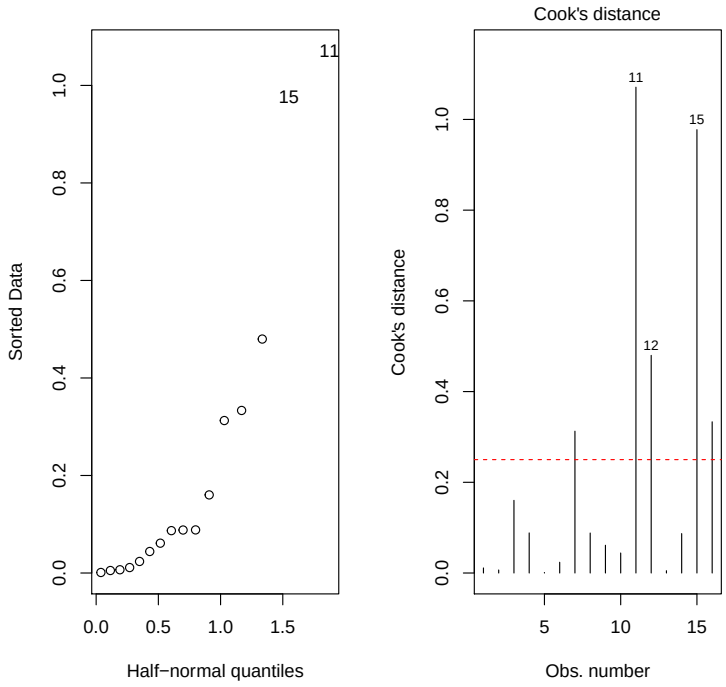

Figura 4.23: Gráficos de influencia para el modelo.

La tabla 4.11 muestra una comparación entre el modelo completo y la reestimación que resulta al quitar los puntos 11 y 15. Sin embargo, el cambio observado en los coeficientes fue muy pequeño, por tanto estas observaciones se encuentran dentro del rango de influencia del resto de puntos.

En resumen, los métodos de diagnóstico utilizados no encontraron problemas significativos, por lo tanto se puede concluir que el modelo propuesto, y el método de estimación utilizado son apropiados para modelar la incidencia de cáncer de piel en las mujeres pertenecientes a las ciudades de Fort Worth y St. Paul de acuerdo a un rango de edad determinado. Luego las inferencias, interpretaciones y predicciones realizadas con el modelo van a tener un alto grado de confiabilidad.

Tabla 4.11: Comparación de las estimaciones del modelo completo y el modelo sin las observaciones 11 y 15.

	Variable dependiente:	
	Número de casos	
	Modelo completo	Modelo sin las obs 11,15
	(1)	(2)
ciudadFt.Worth	0.804***	0.823***
	(0.052)	(0.062)
edad25-34	2.630***	2.630***
	(0.467)	(0.467)
edad35-44	3.847***	3.598***
	(0.455)	(0.485)
edad45-54	4.595***	4.595***
	(0.451)	(0.451)
edad55-64	5.087***	5.087***
	(0.450)	(0.450)
edad65-74	5.645***	5.646***
	(0.450)	(0.450)
edad75-84	6.059***	6.182***
	(0.450)	(0.458)
edad85+	6.174***	6.175***
	(0.458)	(0.458)
Intercepto	−11.658***	−11.671***
	(0.449)	(0.449)
Desvío nulo	2789.68	2324.43
gl desvío nulo	15	13
Desvío residual	8.2585	2.8037
gl desvío residual	7	5
N	16	14
Note:		*p<0.1; **p<0.05; ***p<0.01

4.3. Muertes en Europa por melanoma maligno

Para la última aplicación se consideró un conjunto de datos que describe la razón de mortalidad por melanoma maligno en varios países de Europa asociada a la exposición de radiación ultravioleta. Los datos, disponibles en la tabla 4.12, fueron estudiados por Langford et al. (1998) y están disponibles en el objeto *Mmmec* de la librería *mlmRev* (Bates et al., 2020).

Tabla 4.12: Datos de Muertes en Europa por melanoma maligno.

nación	región	condado	muertes	esperadas	uvb
Belgium	1	1	79	51.22	-2.91
Belgium	2	2	80	79.96	-3.21
Belgium	2	3	51	46.52	-2.80
Belgium	2	4	43	55.05	-3.01
Belgium	2	5	89	67.76	-3.01
Belgium	2	6	19	35.98	-3.42
⋮	⋮	⋮	⋮	⋮	⋮
Netherlands	78	350	32	28.01	-3.94
Netherlands	78	351	107	74.90	-4.21
Netherlands	78	352	150	97.64	-4.19
Netherlands	78	353	15	11.01	-3.88
Netherlands	79	354	64	65.77	-3.68
Netherlands	79	355	31	34.47	-3.64

donde cada observación corresponde a un condado, indexado por la variable *condado*, *muertes* cuenta el número de muertes de hombres debido a melanoma maligno desde el año 1971 hasta 1980, *uvb* representa una medida centrada de la dosis de luz ultravioleta que alcanza a cada condado, *esperadas* denota el número esperado de muertes en cada condado, y las variables *región* y *nación* son factores de agrupación. De forma similar al ejemplo en la sección (4.2), es razonable modelar la incidencia de muertes por melanoma maligno, es decir, *muertes/esperadas*, por lo que inicialmente se puede ajustar un modelo MLG de Poisson con un *offset* para el logaritmo del número esperado de muertes:

$$\begin{cases} Y_i \sim Poisson(\mu_i) \\ \log(\mu_i) = \beta_0 + \beta_1 X_{1i} + \log(X_{2i}) \end{cases}$$

donde Y_i denota al número de muertes, y X_1, X_2 representan la dosis de luz UV y al número esperado de muertes respectivamente.

La tabla 4.13 presenta las estimaciones de los parámetros del modelo.

Tabla 4.13: Estimaciones del MLG de Poisson para el número de muertes.

	Variable dependiente:
	Número de muertes
	Modelo poisson
uvb	-0.057^{***}
	(0.003)
Intercepto	-0.070^{***}
	(0.011)
Desvío nulo	2357.3
gl desvío nulo	353
Desvío residual	1852.5
gl desvío residual	352
N	354
Nota:	*p<0.1; **p<0.05; ***p<0.01

4.3.1. Análisis de diagnóstico

Una revisión inicial del gráfico de residuales contra el predictor lineal presente en la figura 4.24 indica un problema serio de ajuste, por otro lado, el cociente del desvío con sus grados de libertad correspondientes sugiere un posible caso de sobredispersión, por tanto se ajusta un modelo de regresión binomial negativa. La figura 4.25 muestra de nuevo el gráfico de residuales contra el predictor lineal para el modelo nuevo, en este caso el ajuste mejora considerablemente.

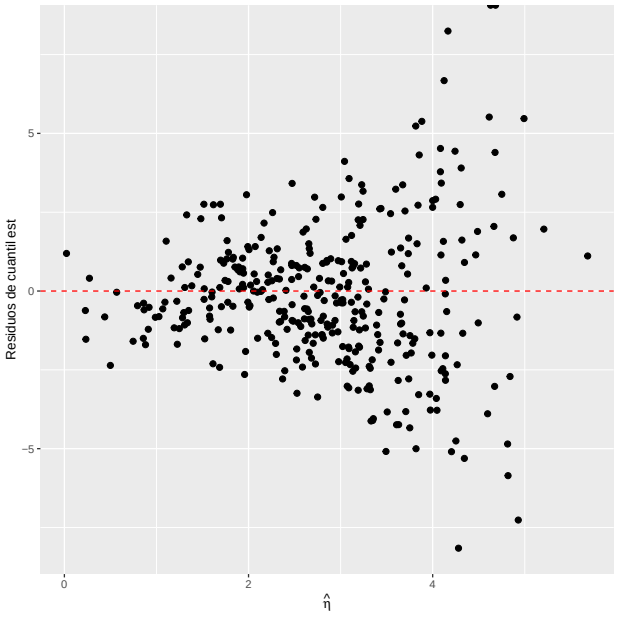

Figura 4.24: Relación entre r_{sq} y $\hat{\eta}$ para el modelo de Poisson.

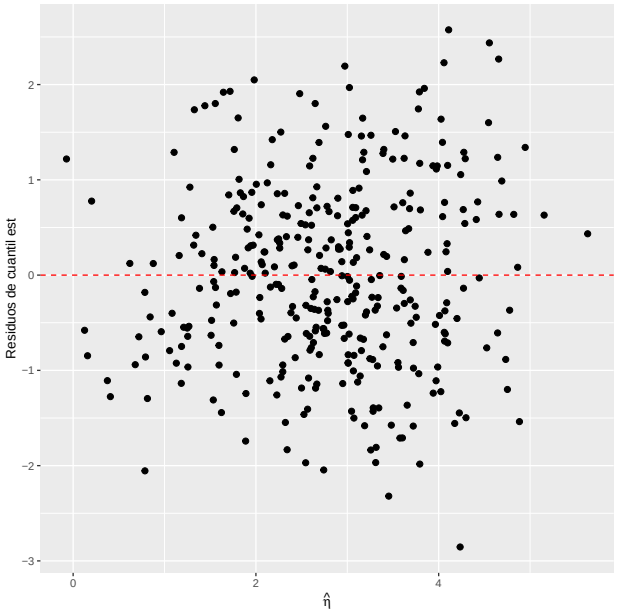

Figura 4.25: Relación entre r_{sq} y $\hat{\eta}$ para el modelo de la binomial negativa.

El modelo de la binomial negativa es válido para explicar la incidencia de melanoma maligno de forma general, sin embargo no se está considerando el efecto de las regiones. Cada región está compuesta por un número determinado de condados, la presencia de rayos UV depende de factores como el clima o la contaminación, los cuales pueden ser diferentes dependiendo de la ubicación, en otras palabras, es probable que haya un componente de varianza para las regiones, por lo que las medidas tomadas en una región cualquiera no van a ser independientes. Por este motivo se ajusta un MLGM con los datos, considerando un intercepto aleatorio para cada región. El MLGM de Poisson correspondiente está dado por:

$$\begin{cases} Y_{ij} \sim Poisson(\mu_{ij}) \\ \log(\mu_{ij}) = (\beta_0 + b_{0i}) + \beta_1 X_{1ij} + \log(X_{2ij}) \end{cases}$$

donde el índice j recorre todas las observaciones dentro de la región i, y b_{oi} representa el efecto aleatorio incluido. Como el número de efectos aleatorios es pequeño, se puede usar el método de cuadratura de Gauss Hermite dado en la sección 3.4 para ajustar el modelo con la función *glmer()* del paquete *lme4*. Los resultados se presentan en la tabla 4.14:

Tabla 4.14: Estimaciones del modelo ajustado con cuadratura de Gauss-Hermite

	Variable dependiente:
	Numero de muertes MLGM Gauss-Hermite
uvb	-0.034^{***} (0.010)
Intercepto	-0.139^{***} (0.049)
Número de unidades (región)	78
sd(Intercepto aleatorio)	0.4122
N	354
Nota:	*p<0.1; **p<0.05; ***p<0.01

Para determinar si la inclusión de efectos aleatorios mejora el modelo en comparación con el de la distribución binomial negativa, se puede realizar el test de razón de verosimilitud con modelos sin predictores que toman como respuesta a los residuales del MLG, y que sólo difieren por la inclusión del intercepto aleatorio. El resultado de la prueba, presente en la tabla 4.15 indica que el efecto de las regiones es significativo.

Tabla 4.15: Test de razón de verosmilitud para verificar si los efectos aleatorios son significativos. Ninguno de los modelos tiene predictores, y en ambos casos la respuesta es el vector de residuos de desvío del modelo de la binomial negativa. El modelo *mod0* sólo tiene al intercepto como efecto fijo, y el modelo *mod1* incluye un intercepto aleatorio para la región.

	df	AIC	BIC	logLik	Test	L.Ratio	p.value
mod0	2.00	1004.52	1012.25	-500.26			
mod1	3.00	847.71	859.31	-420.86	1 vs 2	158.81	0.00

La figura 4.26 presenta el gráfico de valores ajustados contra residuales de Pearson para el MLGM ajustado. En este caso se observa un buen ajuste, a excepción de algunos datos atípicos. Por otro lado, la figura 4.27 presenta un Q-Q plot de los efectos aleatorios estimados, el gráfico sugiere la presencia de normalidad, a excepción de la cola inferior, la cual presenta varios casos de posibles unidades atípicas que merecen una revisión más detallada.

En resumen, el MLGM Poisson parece presentar una mejora al incluir el efecto de cada región, incluso comparado con el modelo de la Binomial negativa que se usó para modelar la sobredispersión, sin embargo los diagnósticos utilizados sólo dan una idea general del ajuste del modelo, se requieren pruebas adicionales para determinar si es necesario realizar modificaciones en el modelo, desafortunadamente todavía no existe un repertorio completamente definido de técnicas gráficas de diagnóstico y análisis de influencia que permitan obtener conclusiones significativas.

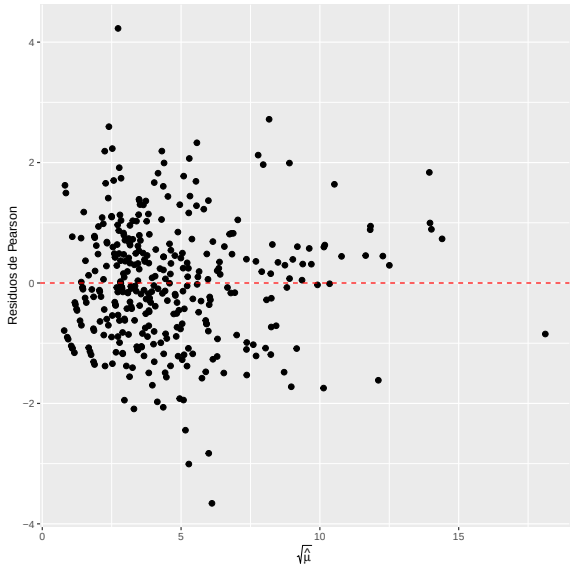

Figura 4.26: Relación entre r_p y $\sqrt{\hat{\mu}}$ para el MLGM Poisson.

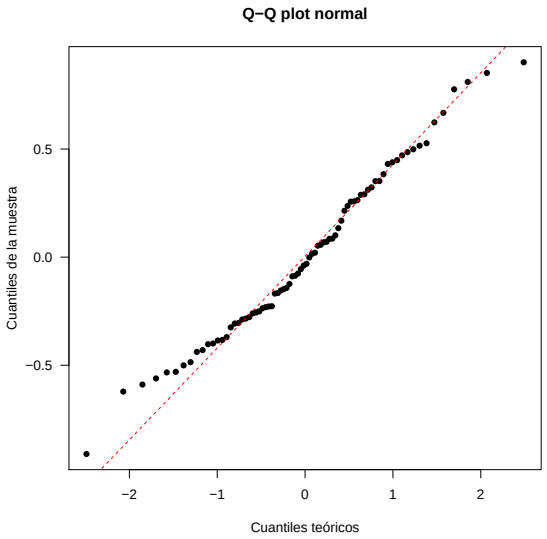

Figura 4.27: Q-Q plot de los interceptos aleatorios estimados para el MLGM Poisson.

Capítulo 5

Conclusiones

Este trabajo se concentró en los métodos de diagnóstico para los modelos lineales mixtos, los modelos lineales generalizados y los modelos lineales generalizados mixtos. Como se evidenció a lo largo del trabajo, se tienen múltiples aplicaciones reales para estos modelos, pero su fundamentación teórica direccionada a métodos de diagnóstico no se ha compilado de forma ordenada, clara y precisa con las ilustraciones y códigos respectivos. De ahí nace el marco en el que se desarrolla este trabajo de tesis.

El objetivo general planteado en el anteproyecto de tesis fue:
"Relacionar matemáticamente las técnicas de diagnóstico en modelos lineales mixtos y modelos lineales generalizados ilustrando la teoría mediante ejemplos prácticos".
Para lograr este objetivo general, se propusieron tres objetivos específicos cuyo cumplimiento se puede observar a lo largo del texto. Inicialmente se hizo una revisión sistemática de los modelos lineales mixtos incluyendo los de pendiente e intercepto aleatorios. Posteriormente, se definió el modelo lineal generalizado así como el modelo lineal generalizado mixto, todo direccionado a los métodos de diagnóstico que se tienen en la literatura. Finalmente, se presentan diversas aplicaciones para ilustrar la teoría de cada modelo en cuestión; estos ejemplos se sistematizaron en el lenguaje de programación R.

El primer objetivo específico es:
"Establecer de forma analítica los métodos de diagnóstico para modelos lineales mixtos en general y de forma específica para modelos con intercepto aleatorio o modelos con pendiente aleatoria".
En el capítulo 2 se definió el modelo lineal mixto general, el cual incluye el modelo de intercepto y de pendiente aleatoria. En las diferentes secciones del capítulo se expone la teoría asociada a estimación, inferencia y diagnóstico enfatizando en los diferentes tipos de residuales, análisis de influencia y medidas remediales, así como una introducción a un tema relativamente nuevo, conocido como **inferencia visual**. El recuento del capítulo dos se hizo secuencial y escrito de forma matemática presentando las diferentes aproximaciones teóricas de diferentes autores que han aportado a este campo, estableciendo así una teoría completa sobre los métodos de diagnóstico para este tipo de modelos.

El segundo objetivo específico propuesto fue:
"Presentar una base teórica sobre los métodos de diagnóstico, para los modelos lineales mixtos, así como también para modelos lineales generalizados y modelos lineales generalizados mixtos".
En el capítulo 3 se presentan los métodos de diagnóstico en modelos lineales generalizados y modelos generalizados mixtos. En la sección 3.1. se presenta la definición del modelo lineal generalizado y en la sección 3.4. se define el modelo lineal generalizado mixto. También se fundamenta la estimación e inferencia para este tipo de modelos, así como el diagnóstico enfatizando los distintos tipos de residuos para evaluar la eficacia del modelo, haciendo énfasis en la forma gráfica. Es importante notar que el capítulo 3 presenta de manera suscinta los tipos de diagnóstico que deben ser aplicados a esta clase de modelos, que en general muchos investigadores omiten por falta de teoría completa, organizada y fundamentada en los principios estadísticos de dichos modelos. Por lo anterior se considera que este capítulo es de interés general de los lectores, ya que al tener esta teoría bien estructurada, se permite una aplicación más aproximada del diagnóstico de modelos estadísticos de lo que normalmente se hace.

El objetivo específico final se asocia a las aplicaciones:
"Utilizar el programa R para realizar las aplicaciones que ilustran la teoría en un escenario de interés práctico".
En el capítulo 4 se muestran varios ejemplos de la aplicación del diagnóstico de los modelos lineales mixtos y tanto generalizados como generalizados

mixtos.

El primer ejemplo está relacionado con el campo médico para la reducción de tiempo de hemodiálisis, en este caso se ajustó un modelo de intercepto y pendiente aleatoria. Se presentaron los diferentes modelos ajustados junto con los respectivos análisis de diagnóstico que fueron descritos de forma teórica en los capítulos anteriores.

El siguiente ejemplo contiene datos sobre el cáncer de piel y su incidencia en mujeres, estratificadas por edad. En este ejemplo se ajusta un modelo de conteo de poisson, y al igual que en la descripción anterior, se presentaron los diferentes residuos para el diagnóstico, de tal forma que este modelado en particular logró un alto grado de confiabilidad, aunque desde luego el objetivo principal de este trabajo es la ilustración de la teoría de diagnóstico para validar el modelo seleccionado, mas no el análisis de selección y de interpretación del modelo para obtener resultados definitivos.

El tercer y último ejemplo se trata de datos donde se presenta la razón de mortalidad por melanoma maligno en diferentes países de Europa, asociados a la exposición a radiación UV. En este caso también se le aplicó la teoría de diagnóstico y se ajustó un modelo generalizado mixto con resultados adecuados para los datos. También se presentan algunas recomendaciones para la aplicación de esta teoría ya que no hay técnicas gráficas de diagnóstico que permitan obtener conclusiones significativas para este tipo de modelo. La programación de estos ejemplos se presenta, de forma detallada, en el Apéndice A, utilizando el lenguaje de programación R. Vale resaltar que los códigos se asocian a cada resultado obtenido, lo cual facilita el uso en problemas prácticos de aplicación.

Por lo tanto, se logró cada uno de los objetivos planteados, llegando a tener un material compilatorio relevante sobre el diagnóstico de los diferentes tipos de modelos propuestos, con toda la fundamentación estadística y además escritos matemáticamente. Se presenta la teoría de manera sistemática y completa y finalmente se llega a unas aplicaciones que ilustran claramente esta teoría. Se considera que este trabajo de investigación cumplió a cabalidad los objetivos propuestos, no sólo en la parte investigativa sino logrando un documento con todo el rigor y claridad necesario para aplicar los métodos de diagnóstico adecuados a diferentes conjuntos de datos, modelados bajo las metodologías estudiadas.

Como trabajos futuros se plantean las siguientes líneas a partir del estudio detallado de la teoría que se logró en este trabajo:

- Profundizar en los métodos alternativos de diagnóstico de modelos mixtos, como lo son la inferencia visual, los métodos de estimación robusta, los métodos de inferencia bayesiana o los modelos mixtos elípticamente simétricos y asimétricos, con el fin de utilizarlos en un rango amplio de aplicaciones.

- Dar continuidad al estudio de diagnóstico para los modelos lineales generalizados mixtos, en especial para el estudio de técnicas gráficas y de análisis de sensibilidad.

- Un aspecto importante del ajuste de modelos mixtos en la práctica es el tratamiento de *datos faltantes* y de *datos censurados o dropouts* (por ejemplo pacientes que murieron en el transcurso de un estudio, lo que imposibilita la obtención de medidas adicionales). Varios autores han estudiado el manejo de datos faltantes, por ejemplo Liu (2015) expone diversos métodos para el manejo de datos faltantes para modelos mixtos en el capítulo 14, sin embargo, es necesario seguir investigando técnicas de ajuste y de diagnóstico en estructuras más complejas de datos faltantes.

Apéndice A

Códigos utilizados en el capítulo 4

A.1. Funciones para extraer residuos

La función *residdiag_nlme()* fue programada por Singer et al. (2017), y está disponible en el material suplementario del artículo. Esta función se aplica en modelos mixtos ajustados con *nlme*. Los códigos fueron ejecutados con la versión 4.0.3 de R, sin embargo, al usar una versión posterior de R, la función *residdiag_nlme()* no se va a ejecutar correctamente. Para resolver este problema, se recomienda eliminar el argumento LINPACK = F dentro de la función.

Las tablas de estimaciones de coeficientes fueron generadas con el paquete *stargazer* (Hlavac, 2022). A continuación se presentan los códigos utilizados para generar las salidas en R del trabajo. Los códigos se pueden encontrar en el siguiente repositorio de github: https://github.com/AndresM0796/Codigos_Tesis_de_Maestria.

```
#### Funcion para extraer residuos de un MLG ####
residuos_glm <- function(mod){
  require(statmod)
  r_p <- resid(mod, type = "pearson") # Pearson
  r_d <- resid(mod) # desvio
  r_q <- qresid(mod) # cuantil

  h_i <- hatvalues(mod) # leverages
  phi.P <- summary(mod)$dispersion # phi estimado con
     Pearson

```

```
11    r_sp <- rstandard(mod, type = "pearson") # Pearson est.
12    r_sd <- rstandard(mod) # desvio est.
13    r_sq <- r_q/sqrt(1-h_i) # cuantil est.
14
15    r_G <- rstudent(mod) # Williams
16
17    mu_hat <- fitted(mod) # mu estimado
18    eta_hat <- predict(mod, type = "link") # eta estimado
19    z <- resid(mod,type = "working") + eta_hat # respuesta
          linealizada
20    c_i <- cooks.distance(mod)
21
22    res.data <-
          data.frame(r_p,r_d,r_q,h_i,r_sp,r_sd,r_sq,r_G,
23                           mu_hat, eta_hat, z, c_i)
24    return(res.data)
25  }
26
27  #### Funcion para extraer residuos en el nivel 1 de un
        MLM ####
28  residuos_LMM_1 <- function(mod.lme4){
29    # mod es un modelo de lme4
30    require(HLMdiag)
31    require(lme4)
32    require(nlme)
33
34    res_fm1 <- hlm_resid(mod.lme4, level = 1, type =
          "LS",standardize = F)
35    res_fm1s <- hlm_resid(mod.lme4, level = 1, type =
          "LS",standardize = T)
36    res_fmsemi <- hlm_resid(mod.lme4, level = 1, type =
          "LS", standardize = "semi")
37
38    r1 <- residuals(mod.lme4)
39    r1.est <- res_fm1s$.std.resid
40    r1.sem <- res_fmsemi$.semi.ls.resid
41    r_m <- res_fm1$.mar.resid
42    r_ms <- res_fm1s$.chol.mar.resid
43    y_hat <- predict(mod.lme4)
44    ci_1 <- cooks.distance(mod.lme4, level = 1)
45    mdf_1 <- mdffits(mod.lme4, level = 1)
46    covr_1 <- covratio(mod.lme4, level = 1)
47    covt_1 <- covtrace(mod.lme4, level = 1)
48    rvc_1 <- rvc(mod.lme4, level = 1)
49    lev_1 <- leverage(mod.lme4, level = 1)
```

```
50
51   res.data.lvl1 <- data.frame(r1, r1.est, r1.sem, r_m,
        r_ms, y_hat, ci_1,
52                               mdf_1, covr_1, covt_1,
                                    rvc_1, lev_1)
53   return(res.data.lvl1)
54 }
55
56
57 #### Funcion para obtener medidas de influencia en el
       nivel 2 de un MLM ####
58 residuos_LMM_2 <- function(mod.lme4){
59   require(HLMdiag)
60   require(lme4)
61   require(nlme)
62   grupo <- names(summary(mod.lme4)$ngrps)
63   ci_2 <- cooks.distance(mod.lme4, level = grupo)
64   mdf_2 <- mdffits(mod.lme4, level = grupo)
65   covr_2 <- covratio(mod.lme4, level = grupo)
66   covt_2 <- covtrace(mod.lme4, level = grupo)
67   rvc_2 <- rvc(mod.lme4, level = grupo)
68   lev_2 <- leverage(mod.lme4, level = grupo)
69
70   res.data.lvl2 <- data.frame(ci_2, mdf_2, covr_2,
        covt_2, rvc_2, lev_2)
71   return(res.data.lvl2)
72 }
```

A.2. Códigos de la aplicación de la sección 4.1

A.2.1. Ajuste del modelo

```
1  #### Paquetes a cargar ####
2  library(MEMSS)
3  library(nlme)
4  library(MVA)
5  library(lattice)
6  library(mlmRev)
7  library(dplyr)
8  library(lme4)
9  library(ggplot2)
10 library(RLRsim)
```

```
library(HLMdiag)
library(plotly)
library(ggpubr)
library(latex2exp)
library(stargazer)
library(xtable)

#### Base de datos ####
data(Dialyzer)
names(Dialyzer)

# Crear una copia de la base de datos para las figuras
Dialyzer2 <- Dialyzer %>%
  dplyr::select(-c("index")) %>%
  rename("Dializador" = "Subject", "Presion" =
     "pressure", "RUF" = "rate")
Dialyzer2$QB = ifelse(Dialyzer2$QB == "200", "200␣
   dL/min", "300␣dL/min")


## Figura 4.1
Dialyzer2 %>%
  group_by(Dializador) %>%
  ggplot(aes(Presion, RUF, group = Dializador, color =
     Dializador)) +
  geom_line() + geom_point() + facet_wrap(~QB) +
  xlab("Presion␣(dmHg)") + ylab("Razon␣de␣
     ultrafiltracion␣(ml/hr)")

#### modelo en (4.1) ####
# el modelo se ajusta con maxima verosimilitud para
   realizar comparaciones
# con el LRT
mod1 <- lme(rate ~ pressure + QB, random = ~ 1|Subject,
            data = Dialyzer, method = "ML")

## estimacion del modelo con gls (ecuacion (4.2))
mod_gls <- gls(rate ~ pressure + QB, data = Dialyzer,
               method = "ML", na.action = "na.omit")

## Test de componentes de varianza con RLRsim
exactRLRT(mod1)

## Comparacion con el LRT (mod_gls vs mod1)
anova(mod_gls, mod1)
```

```
51
52
53 #### modelos con terminos polinomiales #####
54
55 ## modelo mixto ecuacion (4.3)
56 mod2 <- lme(rate ~ pressure + I(pressure^2)+
       I(pressure^3) + QB,
57             random = ~ 1|Subject,
58             data = Dialyzer, method = "ML")
59
60 ## modelo gls ecuacion (4.4)
61 mod_gls2 <- gls(rate ~ pressure + I(pressure^2)+
       I(pressure^3) + QB,
62                 data = Dialyzer, method = "ML",
                       na.action = "na.omit")
63
64
65 ## test de razon de verosimilitud
66 anova(mod_gls, mod_gls2, mod2)
67
68
69
70 ## termino cubico y pendiente aleatoria en pressure,
       ecuacion (4.5)
71 mod3 <- lme(rate ~ pressure + I(pressure^2)+
       I(pressure^3) + QB,
72             random = ~ pressure|Subject,
73             data = Dialyzer, method = "ML")
74
75 anova(mod2, mod3)
76
77
78 ## termino cuartico
79 mod4 <- lme(rate ~ pressure + I(pressure^2)+
       I(pressure^3) + I(pressure^4) + QB,
80             random = ~ pressure|Subject,
81             data = Dialyzer, method = "ML")
82
83 anova(mod3, mod4) # no mejora
84
85 # Comparacion de los 3 modelos
86 anova(mod2, mod3, mod4)
87
88 # Comparacion del modelo mixto y el modelo gls con
       terminos polinomiales
```

```
anova(mod_gls2, mod3)



#### graficos de ajuste ####
# Funcion de panel para el grafico de ajuste
pfun <- function(x,y){
  panel.xyplot(x,y[1:length(x)])
  panel.lines(x,y[1:length(x)+length(x)], lty = 1)
}


## Figura 4.2
Dialyzer$pred <- predict(mod3)
plot(xyplot(cbind(rate,pred) ~ pressure | Subject,
            data = Dialyzer, panel = pfun,
            ylab = "Razon de ultrafiltracion (ml/hr)"))


#### modelo definitivo ####
## en nlme
mod.nlme <- lme(rate ~ pressure + I(pressure^2)+
   I(pressure^3) + QB,
                random = ~ pressure|Subject,
                data = Dialyzer, method = "REML")

## en lme4
rate.mod <- lmer(rate ~ pressure + I(pressure^2) +
   I(pressure^3)+ QB
                 +(pressure|Subject),data = Dialyzer)

# Las estimaciones de ambos modelos son iguales
```

A.2.2. Gráficos de diagnóstico

```
#### Calculo de los residuos ####
res.data.lvl1 <- residuos_LMM_1(rate.mod)
res.data.lvl2 <- residuos_LMM_2(rate.mod)
r_2 <- data.frame(ranef(rate.mod)$Subject) # efectos
   aleatorios

res.data.lvl1$pressure <- Dialyzer$pressure
res.data.lvl1$pressure2 <- Dialyzer$pressure^2
```

```
8  res.data.lvl1$pressure3 <- Dialyzer$pressure^3
9
10 #### graficos ####
11 ## Figura 4.3
12 res.data.lvl1 %>%
13   ggplot(aes(y_hat, r1.est)) + geom_point(cex=2) +
14   geom_hline(yintercept = 0, col = "red", lty = 2) +
       theme_gray()+
15   xlab(expression(hat(y))) + ylab("Residuos nivel 1
       est.")
16 res.data.lvl1 %>%
17   ggplot(aes(y_hat, r1.sem)) + geom_point(cex=2) +
18   geom_hline(yintercept = 0, col = "red", lty = 2) +
       theme_gray()+
19   xlab(expression(hat(y))) + ylab("Residuos
       semiestandarizados")
20
21 ## Figura 4.4
22 res.data.lvl1 %>%
23   ggplot(aes(pressure, r1.sem)) + geom_point(cex=2) +
       theme_gray() +
24   geom_hline(yintercept = 0, col = "red", lty = 2) +
25   ylab("Residuos semiestandarizados") + xlab("Presion
       (dmHg)")
26 res.data.lvl1 %>%
27   ggplot(aes(pressure2, r1.sem)) + geom_point(cex=2) +
       theme_gray() +
28   geom_hline(yintercept = 0, col = "red", lty = 2) +
29   ylab("Residuos semiestandarizados") + xlab("Presion^2
       (dmHg^2)")
30 res.data.lvl1 %>%
31   ggplot(aes(pressure, r1.sem)) + geom_point(cex=2) +
       theme_gray() +
32   geom_hline(yintercept = 0, col = "red", lty = 2) +
33   ylab("Residuos semiestandarizados") + xlab("Presion^3
       (dmHg^3)")
34
35 ## Figura 4.5
36 qqnorm(res.data.lvl1$r1.sem, las=1, pch = 16, xlab =
     "Cuantiles teoricos",
37         ylab = "Cuantiles de la muestra", main = "Q-Q
             plot Normal")
38 qqline(res.data.lvl1$r1.sem, col = "red", lty = 2)
39
40
```

```
41 ## Figura 4.6
42 residdiag_nlme(mod.nlme, limit=2, plotid=1)
43
44 ## Figura 4.7
45 residdiag_nlme(mod.nlme, limit=2, plotid=3)
46
47 ## Figura 4.8
48 res.data.lvl2 %>%
49   dotplot_diag(x = sigma2, cutoff = "internal", name =
       "rvc") +
50   ylab(TeX(r'(RVC␣$\sigma^2$)'))
51 res.data.lvl2 %>%
52   dotplot_diag(x = D11, cutoff = "internal", name =
       "rvc") +
53   ylab(TeX(r'(RVC␣$\sigma_{0␣0}$)'))
54 res.data.lvl2 %>%
55   dotplot_diag(x = D21, cutoff = "internal", name =
       "rvc") +
56   ylab(TeX(r'(RVC␣$\sigma_{0␣1}$)'))
57 res.data.lvl2 %>%
58   dotplot_diag(x = D22, cutoff = "internal", name =
       "rvc") +
59   ylab(TeX(r'(RVC␣$\sigma_{1␣1}$)'))
60
61 ## Figura 4.9
62 res.data.lvl1 %>%
63   ggplot(aes(1:dim(res.data.lvl1)[1], r1.est)) +
       geom_point() + theme_gray() +
64   geom_hline(yintercept = 0, col = "red", lty = 2) +
       ylab("Residuos␣nivel␣1␣est.")+
65   xlab("indice") +
66   geom_text(aes(label=ifelse(abs(r1.est)>2,
67     as.character(1:140),'')),hjust=0,vjust=0)
68
69 ## Figura 4.10
70 residdiag_nlme(mod.nlme, limit=2, plotid=4)
71
72 ## Figura 4.11
73 res.data.lvl2 %>%
74   dotplot_diag(x = fixef, cutoff = "internal", name =
       "leverage") +
75   ylab("Leverage␣efectos␣fijos")
76 res.data.lvl2 %>%
77   dotplot_diag(x = ranef.uc, cutoff = "internal", name =
       "leverage") +
```

```
78    ylab("Leverage␣efectos␣aleatorios")
79
80  ## Figura 4.12
81  res.data.lvl2 %>%
82    dotplot_diag(x = ci_2, cutoff = "internal", name =
          "cooks.distance") +
83    ylab("Cook's␣distance") + xlab("school")
84  res.data.lvl1 %>%
85    dotplot_diag(x = ci_1, cutoff = "internal", name =
          "cooks.distance",
86                 xlim = c(0,0.15))+ylab("Cook's␣distance")
                      + xlab("school")
87
88  ## Figura 4.13
89  residdiag_nlme(mod.nlme, limit=2, plotid=7)
90
91  ## Figura 4.14
92  residdiag_nlme(mod.nlme, limit=2, plotid=8)
93
94  ## Figura 4.15
95  residdiag_nlme(mod.nlme, limit=2, plotid=9)
96
97  ### comparaciones en las estimaciones
98  ## modelo original
99  summary(rate.mod)
100
101 ## modelo sin los puntos 84,83 y 7
102 library(influence.ME)
103 mod.new <- exclude.influence(rate.mod, obs = c(7,83,84))
104 summary(mod.new)
105
106 ## Figura 4.16
107 res.data.lvl2 %>%
108   dotplot_diag(x = covt_2, cutoff = "internal", name =
          "covtrace") +
109   ylab("COVTRACE")
110 res.data.lvl1 %>%
111   dotplot_diag(x = covt_1, cutoff = "internal", name =
          "covtrace") +
112   ylab("COVTRACE")
```

A.3. Códigos de la aplicación de la sección 4.2

A.3.1. Ajuste del modelo

```
#### Paquetes a cargar ####
library(glmtoolbox)
library(statmod)
library(faraway)
library(plotly)
library(ggplot2)
library(dplyr)

#### Base de datos ####
data("skincancer")

#### Grafico relacion incidencia de casos ####
skincancer2 <- skincancer %>%
  rename(ciudad = city) %>%
  mutate(rate = (cases/population)*10000)

## Figura 4.17
ggplot(data = skincancer2, aes(age,rate, group = ciudad,
    color = ciudad)) +
  geom_line(lwd = 1.3) + theme_bw() + xlab("Rango␣de␣
     edad") + ylab("Casos/10000")


#### modelos ajustados ####
## ciudad como predictor
mod0 <- glm(cases ~ city, offset=log(population),
    family=poisson("log"),
            data=skincancer)

## ciudad y edad como predictores
mod1 <- glm(cases ~ city+age, offset=log(population),
    family=poisson("log"),
            data=skincancer)

## interaccion entre ciudad y edad
mod2 <- glm(cases ~ city*age, offset=log(population),
    family=poisson("log"),
            data=skincancer)

## comparacion de los modelos con el LRT
```

```
anova(mod0,mod1,mod2, test = "Chi")

## resumen modelo seleccionado
summary(mod1)
```

A.3.2. Gráficos de diagnóstico

```
#### Calculo de los residuos ####
mod1 <- glm(cases ~ city+age, offset=log(population),
   family=poisson("log"),
          data=skincancer)
res.data <- residuos_glm(mod1)

#### graficos ####
## Figura 4.18
res.data %>%
  ggplot(aes(eta_hat, r_sq)) + geom_point(cex=2) +
     theme_gray() +
  geom_hline(yintercept = 0, col = "red", lty = 2) +
     ylab("Residuos␣de␣cuantil␣est")+
  xlab(expression(hat(eta)))

## Figura 4.19
res.data %>%
  ggplot(aes(sqrt(mu_hat), r_sq)) + geom_point(cex=2) +
     theme_gray() +
  geom_hline(yintercept = 0, col = "red", lty = 2) +
     ylab("Residuos␣de␣cuantil␣est")+
  xlab(expression(sqrt(hat(mu))))

## Figura 4.20
res.data %>%
  ggplot(aes(sqrt(mu_hat), abs(r_sd))) +
     geom_point(cex=2) + theme_gray() +
  ylab("|Residuos␣de␣desvio␣est|") +
     xlab(expression(log(hat(mu))))

## Figura 4.21
res.data %>%
  ggplot(aes(eta_hat, z)) + geom_point(cex=2) +
     theme_gray() +
  geom_smooth(method = "lm", se = F, col = "red", lty =
     2, lwd = 0.7) +
```

```
28    ylab("Respuesta␣linealizada") +
         xlab(expression(hat(eta)))
29
30  ## Figura 4.22
31  qqnorm(res.data$r_q, las=1, pch = 16,cex=1, xlab =
       "Cuantiles␣teoricos",
32          ylab = "Cuantiles␣de␣la␣muestra", main = "Q-Q␣
               plot␣normal")
33  qqline(res.data$r_q, col = "red", lty = 2)
34
35
36  ## Figura 4.23
37  par(mfrow = c(1,2))
38  halfnorm(res.data$c_i) # halfnorm
39  plot(mod1,4); abline(h=4/dim(res.data)[1], col = "red",
       lty=2)
40
41  ## Modelo sin las observaciones 11 y 15
42  infl <- c(11,15)
43  mod.infl <- update(mod1, subset=(-infl))
44
45  ## Comparacion de coeficientes
46  round(coef(mod1),3)
47  round(coef(mod.infl),3)
```

A.4. Códigos de la aplicación de la sección 4.3

```
#### Paquetes a cargar ####
library(mlmRev)
library(faraway)
library(gamair)
library(MASS)
library(lme4)
library(nlme)
library(dplyr)
library(ggplot2)

#### Base de datos ####
data("Mmmec")
datos <- Mmmec

#### Modelos ajustados ####
## modelo de Poisson
mod_glm <- glm(deaths ~ uvb + offset(log(expected)),
   poisson, data = datos)
summary(mod_glm)

## calculo de los residuos
res.data <- residuos_glm(mod_glm)

## Figura 4.24
res.data %>%
  ggplot(aes(eta_hat, r_sq)) + geom_point(cex=2) +
     theme_gray() +
  geom_hline(yintercept = 0, col = "red", lty = 2) +
     ylab("Residuos de cuantil est")+
  xlab(expression(hat(eta)))

## modelo binomial negativa
mod_nb <- glm.nb(deaths ~ uvb +
   offset(log(expected)),datos)
summary(mod_nb)

## calculo de los residuos
res.data <- residuos_glm(mod_nb)

## Figura 4.25
res.data %>%
```

```
  ggplot(aes(eta_hat, r_sq)) + geom_point(cex=2) +
     theme_gray() +
  geom_hline(yintercept = 0, col = "red", lty = 2) +
     ylab("Residuos de cuantil est")+
  xlab(expression(hat(eta)))

## Modelo lineal generalizado mixto, con cuadratura de
   Gauss
modgh <- glmer(deaths ~ uvb + offset(log(expected)) +
   (1|region),
               family = poisson, nAGQ = 25,data = datos)
summary(modgh)

### test para determinar si son necesarios efectos
   aleatorios
## residuos de desvio para el modelo binomial negativa
rfn <- residuals(mod_nb,type="d")
datos$rfn <- rfn

mod0 <- gls(rfn~1, datos)
mod <- lme(rfn~1, random= ~1|region, datos)
anova(mod0, mod)

## Residuos del MLGM
datos$resid <- residuals(modgh, type = "pearson")
datos$fitted <- fitted(modgh)

## Figura 4.26
datos %>%
  ggplot(aes(sqrt(fitted), resid)) + geom_point(cex=2) +
     theme_gray() +
  geom_hline(yintercept = 0, col = "red", lty = 2) +
     ylab("Residuos de Pearson")+
  xlab(expression(sqrt(hat(mu))))

## Efectos aleatorios estimados
r_2 <- data.frame(ranef(modgh)$region)

## Figura 4.27
qqnorm(r_2$X.Intercept., las=1, pch = 16, xlab =
   "Cuantiles teoricos",
       ylab = "Cuantiles de la muestra", main = "Q-Q
          plot normal")
qqline(r_2$X.Intercept., col = "red", lty = 2)
```

Bibliografía

Banerjee, M., & Frees, E. W. (1997). Influence diagnostics for linear longitudinal models. *Journal of the American Statistical Association*, *92*(439), 999-1005.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Bates, D., Maechler, M., & Bolker, B. (2019). *MEMSS: Data Sets from Mixed-Effects Models in S* [R package version 0.9-3]. https://CRAN.R-project.org/package=MEMSS

Bates, D., Maechler, M., & Bolker, B. (2020). *mlmRev: Examples from Multilevel Modelling Software Review* [R package version 1.0-8]. https://CRAN.R-project.org/package=mlmRev

Beckman, R. J., Nachtsheim, C. J., & Cook, R. D. (1987). Diagnostics for mixed–model analysis of variance. *Technometrics*, *29*(4), 413-426.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, *88*(421), 9-25.

Brown, C. C. (1982). On a goodness of fit test for the logistic model based on score statistics. *Communications in Statistics-Theory and Methods*, *11*(10), 1087-1105.

Brown, H., & Prescott, R. (2015). *Applied mixed models in medicine*. John Wiley & Sons.

Christensen, R., Pearson, L. M., & Johnson, W. (1992). Case-deletion diagnostics for mixed models. *Technometrics*, *34*(1), 38-45.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society: Series B (Methodological)*, *48*(2), 133-155.

Correa, J. C., & Salazar, J. C. (2016). Introducción a los modelos mixtos. *Escuela de Estadística Sede Medellín*.

Demidenko, E. (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.

Demidenko, E., & Stukel, T. A. (2005). Influence analysis for linear mixed-effects models. *Statistics in medicine*, *24*(6), 893-909.

Dey, D. K., Ghosh, S. K., & Mallick, B. K. (2000). *Generalized linear models: A Bayesian perspective*. CRC Press.

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Zeger, S., et al. (2002). *Analysis of longitudinal data*. Oxford university press.

Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models*. Chapman; Hall/CRC.

Dueck, A., & Lohr, S. (2005). Robust estimation of multivariate covariance components. *Biometrics*, *61*(1), 162-169.

Dunn, P. K., Smyth, G. K., et al. (2018). *Generalized linear models with examples in R* (Vol. 53). Springer.

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

Faraway, J. J. (2016b). *faraway: Functions and Datasets for Books by Julian Faraway* [R package version 1.0.7]. https://CRAN.R-project.org/package=faraway

Faraway, J. J. (2016c). *Linear models with R*. Taylor & Francis.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman; Hall/CRC.

Ghidey, W., Lesaffre, E., & Verbeke, G. (2010). A comparison of methods for estimating the random effects distribution of a linear mixed model. *Statistical methods in medical research*, *19*(6), 575-600.

Gurka, M. J., Edwards, L. J., Muller, K. E., & Kupper, L. L. (2006). Extending the Box–Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(2), 273-288.

Hardin, J. W., & Hilbe, J. M. (2002). *Generalized estimating equations*. chapman; hall/CRC.

Hauck Jr, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the american statistical association*, *72*(360a), 851-853.

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

Hilden-Minton, J. A. (1995). *Multilevel diagnostics for mixed and hierarchical linear models*. University of California, Los Angeles.

Hlavac, M. (2022). *stargazer: Well-Formatted Regression and Summary Statistics Tables* [R package version 5.2.3]. Social Policy Institute. Bratislava, Slovakia. https://CRAN.R-project.org/package=stargazer

Jiang, J., & Nguyen, T. (2021). *Linear and generalized linear mixed models and their applications. Second edition.* Springer.

Kasim, R. M., & Raudenbush, S. W. (1998). Application of Gibbs sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics*, *23*(2), 93-116.

Koller, M. (2013). *Robust estimation of linear mixed models* (Tesis doctoral). ETH Zurich.

Künsch, H. R., Stefanski, L. A., & Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, *84*(406), 460-466.

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

Langford, I. H., Bentham, G., & McDonald, A.-L. (1998). Multi-level modelling of geographically aggregated health data: a case study on malignant melanoma mortality and UV exposure in the European Community. *Statistics in medicine*, *17*(1), 41-57.

Lee, Y., Nelder, J. A., & Pawitan, Y. (2018). *Generalized linear models with random effects: unified analysis via H-likelihood.* Chapman; Hall/CRC.

Lesaffre, E., & Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, 570-582.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Oliver, S. (2006). *SAS for mixed models.* SAS publishing.

Liu, X. (2015). *Methods and applications of longitudinal data analysis.* Elsevier.

Longford, N. T. (1995). Random coefficient models. En *Handbook of statistical modeling for the social and behavioral sciences* (pp. 519-570). Springer.

Loy, A., & Hofmann, H. (2013). Diagnostic tools for hierarchical linear models. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(1), 48-61.

Loy, A., & Hofmann, H. (2014). HLMdiag: A suite of diagnostics for hierarchical linear models in R. *Journal of Statistical Software*, *56*, 1-28.

Loy, A., Hofmann, H., & Cook, D. (2017). Model choice and diagnostics for linear mixed-effects models using statistics on street corners. *Journal of Computational and Graphical Statistics*, *26*(3), 478-492.

Marr, B. (2018). How much data do we create every day? the mind-blowing stats everyone should read. [Online].

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. Chapman; Hall.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Mun, J., & Lindstrom, M. J. (2013). Diagnostics for repeated measurements in linear mixed effects models. *Statistics in Medicine*, *32*(8), 1361-1375.

Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J. (2012). *Generalized linear models: with applications in engineering and the sciences*. John Wiley & Sons.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370-384.

Nieuwenhuis, R., Te Grotenhuis, H., & Pelzer, B. (2012). Influence. ME: tools for detecting influential data in mixed effects models.

Nobre, J., & Singer, J. (2007). Residual analysis for linear mixed models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *49*(6), 863-875.

Pan, J., Fei, Y., & Foster, P. (2014). Case-deletion diagnostics for linear mixed models. *Technometrics*, *56*(3), 269-281.

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

Pierce, D. A., & Schafer, D. W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*, *81*(396), 977-986.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2020). *nlme: Linear and Nonlinear Mixed Effects Models* [R package version 3.1-149]. https://CRAN.R-project.org/package=nlme

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Schabenberger, O. (2005). Mixed model influence diagnostics. *SUGI*, *29*, 189-29.

Scheipl, F., Greven, S., & Kuechenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, *52*(7), 3283-3299.

Schützenmeister, A., & Piepho, H.-P. (2012). Residual analysis of linear mixed models using a simulation approach. *Computational Statistics & Data Analysis*, *56*(6), 1405-1416.

Singer, J. M., Rocha, F. M., & Nobre, J. S. (2017). Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. *International Statistical Review*, *85*(2), 290-324.

Snijders, T. A., & Berkhof, J. (2008). Diagnostic checks for multilevel models. En *Handbook of multilevel analysis* (pp. 141-175). Springer.

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. sage.

Stroup, W., & Claassen, E. (2020). Pseudo-likelihood or quadrature? What we thought we knew, what we think we know, and what we are still trying to figure out. *Journal of Agricultural, Biological and Environmental Statistics*, *25*(4), 639-656.

Tan, F. E., Ouwens, M. J., & Berger, M. P. (2001). Detection of influential observations in longitudinal mixed effects regression models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *50*(3), 271-284.

Tellez, C. F., & Morales, M. A. (2016). *Modelos Estadísticos lineales con aplicaciones en R*. Ediciones de la U.

Vallejo, G., Ato García, M., et al. (2012). *Diseños experimentales en psicología*. Pirámide.

Vanegas, L. H., Rondón, L. M., & Paula, G. A. (2022). *glmtoolbox: Set of Tools to Data Analysis using Generalized Linear Models* [R package version 0.1.3]. https://CRAN.R-project.org/package=glmtoolbox

Verbeke, G., & Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, *23*(4), 541-556.

Vonesh, E. F., & Carter, R. L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics*, 1-17.

Weisberg, S. (2005). *Applied linear regression* (Vol. 528). John Wiley & Sons.

Williams, D. (1984). Residuals in generalized linear models. *Proceedings of the 12th. International Biometrics Conference*, 59-68.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. chapman; hall/CRC.

Yau, K. K., & Kuk, A. Y. (2002). Robust estimation in generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(1), 101-117.

Yuan, K.-H., & Bentler, P. M. (2002). On normal theory based inference for multilevel models with distributional violations. *Psychometrika*, *67*(4), 539-561.

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators.