APLICACION DE UN MODELO LINEAL GENERALIZADO UTILIZANDO DATOS LONGITUDINALES

Trabajo de titulación presentado por:

Joselin Fernanda Díaz Jorquera

para optar al grado de:

Licenciado en Estadística

y al título profesional de:

Ingeniero Estadístico

Profesora guía:

Dra. Claudia Navarro Villarroel

Valparaíso, Chile, 2017

Agradecimientos

Este nuevo logro de culminar mi carrera profesional es gran parte a mi familia y personas especiales en mi vida, por el apoyo constante que me han brindado, por el valor manifestado para salir adelante y por confiar en mis decisiones durante los años de estudio.

Resumen

Los modelos lineales generalizados se utilizan con el propósito de medir el efecto de las variables sobre una respuesta, realizar estimaciones o mediciones cuando sea necesario. Una de sus características es el supuesto de independencia de las observaciones, pero pueden suceder casos en que se presente la dependencia entre las observaciones, lo que induce a que los datos se encuentren correlacionados, esto ocurre cuando se trabaja con datos longitudinales, ya que este tipo de datos son mediciones repetidas realizadas a unidades de análisis en el tiempo. El objetivo de este trabajo es aplicar un modelo lineal generalizado a datos reales de carácter longitudinal que incluya la correlación. Para esto, se emplea el modelo lineal generalizado mixto. El propósito es modelar el comportamiento de la cantidad de producción de minerales metálicos de las empresas mineras, determinar si las ventas y la exportación influyen en la producción. Por consiguiente, se estiman los parámetros del modelo a través del método de estimación de máxima verosimilitud y el método de cuadratura de Gauss-Hermite, el cual es útil ocupar cuando resulta complejo evaluar la función log-verosimilitud como una integral múltiple en la derivación de estimación de parámetros. Al ajustar el modelo se puede determinar que las ventas y la exportación afectan significativamente en el tiempo.

Lista de abreviaturas

AIC	Criterio de información Akaike
GEE	Ecuaciones de estimación generalizadas
ML	Modelo lineal
MLG	Modelo lineal generalizado
MLGM	Modelo lineal generalizado mixto

Índice general

Índice de figuras

Índice de tablas

Preliminares

El modelo de regresión lineal utiliza la linealidad para describir la relación entre la media de la variable respuesta y un conjunto de covariables, suponiendo que la distribución de la respuesta es normal. Los MLG amplían los modelos de regresión lineal para incluir distribuciones de respuesta no normales, estos modelos fueron propuestos por Nelder y Wedderburn (1972), son una generalización de los modelos lineales y consta de tres elementos importantes:

1. Distribución: consiste en una variable respuesta Y de observaciones $(y_1, \ldots, y_N)$ independientes que sigue una función de distribución de probabilidad perteneciente a la familia exponencial. Entonces la función de densidad de probabilidad general de Y puede expresarse en la forma

$$f(y_i; \theta_i, \phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}, \tag{1}$$

   donde $a(\cdot)$, $b(\cdot)$ y $c(\cdot)$ son funciones conocidas, θ_i se llama parámetro natural o canónico y ϕ se denomina parámetro de dispersión que representa la ubicación.

   La varianza se relaciona con la media mediante una función de varianza.

$$var(y_i) = v(\mu_i)\phi \tag{2}$$

2. Predictor lineal: es una combinación lineal de las covariables $X_1, \ldots, X_p$ con parámetros desconocidos β. Las variables del predictor lineal pueden ser numéricas o categóricas.

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, \tag{3}$$

con $i = 1, ..., N$.

De forma matricial el predictor lineal queda expresado de la siguiente manera:

$$\eta = X\beta, \tag{4}$$

donde $\beta = (\beta_1, ..., \beta_p)^T$ es el vector columna $(p \times 1)$ de los parámetros y X es la matriz $(n \times p)$ de las covariables x_{ij} del modelo.

3. Función enlace: relaciona la media de la respuesta $E(y_i) = \mu_i$ con el predictor lineal η, de tal manera que

$$g(\mu_i) = \eta_i = x_i^{'}\beta, \tag{5}$$

$x_i = (x_{i1}, ..., x_{ip})^{'}$ es el vector de covariables y β es el vector $p \times 1$ de parámetros no conocidos $\beta = (\beta_1, ..., \beta_p)^{'}$.

Donde la función g es conocida y monótona llamada función enlace.

CAPÍTULO 1

Introducción

En este capítulo se presentan los antecedentes sobre el estudio longitudinal en la estadística, se explican los modelos empleados en este tipo de estudio y se dan conocer los objetivos junto con la hipótesis que sustenta el presente trabajo de titulación. Para comenzar se expone una breve explicación sobre los datos longitudinales.

1.1. Datos longitudinales

En la estadística aplicada, pueden surgir investigaciones que se enfocan en estudiar el comportamiento de variables directamente asociadas con el tiempo. En estos estudios, el análisis de datos puede resultar complejo, debido a la presencia de dependencia de las observaciones repetidas sobre cada unidad de análisis o de estudio, por lo que es trascendental analizar este tipo de datos denominados longitudinales.

Los datos longitudinales o también llamados datos de panel consisten en mediciones repetidas registradas en las unidades de análisis (individuos, ciudades, empresas, países, etc.) a través del tiempo, presenta más de dos mediciones y permite diferenciar la variación entre unidades (inter-individuo) y la variación dentro de las unidades (intra-individuo). De esta manera, las repeticiones constan de uno o más grupos de unidades de análisis medidas

en una o más variables a lo largo del tiempo. La variable de interés o la variable respuesta puede ser categórica, discreta o continua, univariante o multivariante.

El objetivo fundamental de un estudio longitudinal, es conocer no sólo los cambios o perfiles individuales, sino determinar si el cambio es significativo y si se dan diferencias entre las distintas unidades de la muestra. En cuanto al análisis de los datos, se pueden emplear diversos procedimientos; por ejemplo, cuando la variable respuesta se distribuye normal, las técnicas utilizadas son de análisis multivariante, análisis de la varianza de medidas repetidas, análisis de curvas de crecimiento, modelos de efectos mixtos y los modelos de ecuaciones de estimación generalizada (Liang y Zeger, 1986).

A principios de los años ochenta, Laird y Ware (1982), basándose en una clase general de modelos mixtos, propusieron el modelo lineal de efectos mixtos incluyendo los modelos de ANOVA univariante de medidas repetidas y de curva de crecimiento como casos especiales a datos longitudinales. Estos modelos podrían manejar las complicaciones de las mediciones erróneas e incompletas de una manera muy natural.

Luego, a mediados de los años ochenta, se hicieron notables avances en la metodología para analizar datos discretos cuando Liang y Zeger (1986) propusieron el enfoque de las ecuaciones de estimación generalizadas (GEE), fue una extensión natural del enfoque de cuasi-verosimilitud para MLG a la respuesta multivariante. Liang y Zeger demostraron la versatilidad del método GEE en el manejo de datos desequilibrados, mezclas de covariables discretas y continuas. Tiempo después, se estaban aplicando ampliamente modelos marginales para abordar temas característicos sobre el cambio longitudinal. Su trabajo también generó mucha más investigación teórica y aplicada sobre el uso de esta metodología para el análisis de datos longitudinales.

En cuanto a los estudios estadísticos que se han realizado para el análisis longitudinal, los primeros métodos propuestos se basaron en el modelo de análisis de varianza (ANOVA) desarrollado originalmente por Ronald Fisher, siendo un ANOVA de efectos mixtos o también llamado ANOVA de medidas repetidas univariadas, este es un modelo para una variable respuesta única,

lo que admite una correlación positiva entre las medidas repetidas sobre la misma unidad. Otro método utilizado en el análisis de datos longitudinales, pero con cálculos más avanzados, es el análisis de regresión multivariado de varianza (MANOVA) de medidas repetidas. MANOVA es un modelo para respuestas multivariables. Sin embargo, en estas técnicas la variable respuesta debe cumplir con el supuesto de normalidad (Distribución de Gauss), no se permite la pérdida de mediciones y no consideran la existencia de correlación entre las observaciones (Fitzmaurice *et al.*, 2012).

Por otro lado, el modelo lineal de efectos mixtos es probablemente el método más utilizado para analizar datos longitudinales, su utilidad especialmente en las ciencias de la vida, se destacó en los años ochenta en un documento ampliamente citado por Laird y Ware (1982). La idea de permitir que ciertos coeficientes de regresión varían aleatoriamente entre las unidades de análisis, también fue un tema recurrente en las primeras contribuciones al análisis de la curva de crecimiento por wishart (1938), Box (1950), Potthoff y Roy (1964).

Cabe mencionar que las correlaciones entre las observaciones repetidas de la misma unidad quedan formadas en la estructura de covarianza y no todos los modelos estadísticos parten de los mismos supuestos con respecto a esta estructura. Así, los procedimientos clásicos, como el análisis univariante de la varianza (ANOVA) y el análisis multivariante (MANOVA), no toman en cuenta el problema de la correlación. Cuando no se tiene en consideración la estructura de covarianza entre las medidas repetidas, se corre el riesgo de obtener conclusiones incorrectas en los análisis estadísticos.

1.2. Modelos estadísticos

Los métodos de análisis mencionados en la sección anterior, se han basado en modelos lineales para respuestas continuas que pueden estar aproximadamente distribuidas normalmente. Cuando la variable respuesta es discreta, los modelos lineales ya no son apropiados para relacionar los cambios en la respuesta media a las covariables. En cambio, los estadísticos han desarrollado extensiones de modelos lineales generalizados para datos longitudinales. Los modelos lineales generalizados, (MLG), proporcionan una clase unificada de modelos para el análisis de regresión de observaciones independientes de una respuesta discreta o continua. Los estadísticos han ampliado los MLG para manejar las observaciones longitudinales de diferentes maneras considerando tres clases de modelos de regresión: los modelos marginales, modelos de transición y modelos de efectos mixtos. Estos modelos se diferencian no sólo en considerar la correlación entre las medidas repetidas, sino que también tienen parámetros de regresión con interpretaciones distintas.

El análisis de un estudio longitudinal se lleva a cabo dentro del contexto de los MLG y tiene como finalidad admitir las herramientas comunes de regresión, en las que se relaciona el efecto con las diferentes exposiciones y considerar la correlación de las medidas entre las unidades de análisis (Delgado Rodríguez y Llorca Díaz, 2004).

Adicionalmente, Liang y Zeger (1986) detallan las extensiones de los MLG para los datos longitudinales, los definen como aquellas mediciones que se realizan repetidamente a una misma unidad y que se caracterizan por el hecho de que las observaciones repetidas tienden a estar correlacionadas.

En los estudios longitudinales las observaciones dentro de las unidades de análisis (intra-individuo) están típicamente correlacionadas y los modelos requieren explicar esa correlación. Por lo tanto, un modelo de regresión para datos longitudinales debe abordar tanto la relación entre la respuesta y las covariables como la correlación en las mediciones repetidas.

Existen tres enfoques comunes para incorporar la correlación en datos longitudinales (Wu, 2009):

a. Modelo de transición

   También llamado modelo de Markov, consisten en modelar la distribución condicional de la respuesta en cualquier ocasión dado las respuestas anteriores y las covariables. Se cree que la dependencia entre las medidas repetidas se debe a valores pasados de la respuesta que influyen en la observación actual. En estos modelos, la correlación intra-individuo (dentro de cada unidad de análisis) se modela a través de estructuras de Markov.

b. Modelos marginales

   Se basan en modelar la respuesta media en cada momento del tiempo y consideran la falta de independencia entre las observaciones, es decir, la correlación entre las observaciones. En este modelo, la estructura media y la estructura de correlación (covarianza) se modelan por separado sin supuestos distributivos para los datos.

c. Modelos mixtos

   En este tipo de modelo, también conocido como modelo de efectos mixtos o de efectos aleatorios, se modela la variable respuesta en función de covariables y de factores aleatorios, los cuales incorporan la variación entre unidades y la correlación dentro de cada unidad en los datos.

Considerando esta situación, ¿Qué modelo estadístico y método de estimación se debe utilizar para el análisis de los datos propuestos? ¿El modelo aplicado se ajustará adecuadamente o de buena manera a los datos longitudinales propuestos?.

Por esta razón, es de interés estudiar los modelos lineales generalizados utilizando datos longitudinales, determinar un método apropiado para realizar las estimaciones necesarias del modelo y lograr aplicarlo a datos reales.

Los MLG pueden extenderse a los datos longitudinales permitiendo que un subconjunto de los coeficientes de regresión varíe aleatoriamente de una unidad a otra. Estos modelos se conocen como modelo lineal generalizado mixto (MLGM) y amplían de manera natural el enfoque conceptual de los modelos lineales de efectos mixtos.

El presente trabajo se enfocará en el modelo lineal mixto para el análisis de datos longitudinales.

1.3. Objetivos

Los objetivos del presente trabajo se exponen a continuación:

Objetivo general

El objetivo general de este estudio es aplicar un modelo lineal generalizado para analizar la producción de minerales metálicos de las empresas mineras de Chile.

Objetivos específicos

i. Emplear un modelo lineal generalizado para la estimación de la producción de los minerales metálicos.

ii. Determinar los cambios de la producción de minerales metálicos a través del tiempo.

iii. Determinar como influyen las ventas y la exportación sobre la producción de minerales metálicos.

iv. Analizar la producción de minerales metálicos entre las empresas y de cada empresa minera.

1.4. Presentación de los datos

Para llevar a cabo este estudio se escogieron datos sobre la actividad minera realizada en Chile. Esta actividad consiste en la extracción de los minerales de las rocas que concentran uno o más minerales metálicos.

Cabe destacar que la actividad minera es importante para el desarrollo del país, el consejo minero en sus publicaciones destaca el crecimiento que ha tenido la economía chilena en las últimas décadas. Luego de una producción estancada en torno a 1,4 millones de toneladas anuales en la década de los años ochenta, se observó un crecimiento sostenido, alcanzando más de 4 millones de toneladas. Posteriormente ese crecimiento continuó, aunque a menor ritmo, y en los últimos años la producción se está acercando a 6 millones de toneladas anuales. Esto ha permitido que Chile pase a representar desde un 16 % de la producción mundial a un 32 % en años recientes, y un 30 % durante el 2011.

El conjunto de datos estudiados consisten en registros mensuales de la cantidad de producción, ventas y exportación de minerales metálicos en toneladas métricas. Con estos datos se espera explicar la producción en función de las ventas y la exportación (covariables) a través de la información adquirida de las empresas mineras. El conjunto de datos se obtuvieron por medio de la página web de la Comisión Chilena del Cobre (www.cochilco.cl) y el Servicio Nacional de Geología y Minería (www.sernageomin.cl).

La intención de analizar la actividad minera es lograr conocer la evolución de la producción física de los minerales metálicos, modelar el comportamiento de la producción y determinar si las ventas y la exportación afectan en la cantidad producida de minerales metálicos. Asimismo, analizar la variación dentro de las empresas y entre las empresas con el conjunto de covariables que presentan un efecto aleatorio en la respuesta de las diferentes empresas. Para esto, se aplica el MLGM que incluye tanto efectos fijos como efectos aleatorios para analizar la evolución a través del tiempo de la actividad minera.

En resumen, se propone aplicar un modelo lineal generalizado a datos longitudinales, con la finalidad de estimar los parámetros del modelo y ajustarlo a un conjunto de datos reales correspondiente a la actividad minera en Chile.

La hipótesis que sustenta este trabajo de titulación es la siguiente:

El modelo lineal generalizado empleado se ajusta de buena forma al conjunto de datos de la actividad minera, los cuales cumplen con la característica de que las observaciones sean dependientes en el tiempo.

En el siguiente capítulo se da a conocer la metodología, presentando el modelo propuesto para analizar los datos longitudinales, el método de estimación y las pruebas de hipótesis.

CAPÍTULO 2

Metodología

En este capítulo se describe el modelo lineal generalizado mixto, el cual es escogido para la aplicación de los datos propuestos y se presenta el método de estimación que se emplea en estos casos.

2.1. Modelo lineal generalizado mixto

Para empezar, es necesario explicar brevemente el modelo lineal mixto antes de discutir sobre el modelo lineal generalizado mixto con más detalle.

Los modelos lineales mixtos, son modelos de regresión lineal que incluyen efectos aleatorios normalmente distribuidos además de efectos fijos. Una aplicación natural de estos modelos es a datos longitudinales donde los efectos aleatorios varían entre unidades e inducen dentro de la unidad dependencia entre mediciones repetidas (Laird y Ware, 1982). Estos modelos podrían manejar las complicaciones de las mediciones erróneas e incompletas de una manera muy natural y se caracterizan por contener tanto efectos fijos como efectos aleatorios. Los efectos fijos son análogos a los coeficientes de regresión y sus estimaciones se calculan directamente. Los efectos aleatorios no se estiman directamente, pero se resumen de acuerdo con sus varianzas estimadas y covarianzas.

El modelo lineal mixto está dado por (Fitzmaurice *et al.*, 2008):

$$Y_{ij} = x_{ij}^{'}\beta + z_{ij}^{'}b_i + e_i, \tag{2.1}$$

donde Y_{ij} representa las respuestas, el valor $i(i = 1, ..., m)$ indica la i-ésima unidad de análisis y j $(j = 1, ..., n_i)$, indica las observaciones repetidas n_i de la variable respuesta dentro de una unidad o un grupo, $x_{ij}^{'}\beta$ es la parte fija, $z_{ij}^{'}b_i$ es la parte aleatoria del modelo, e_i representa el error aleatorio, se supone que b_i sigue una distribución normal con media 0 y matriz varianza-covarianza G y $e_i \sim N(0, R_i)$.

Siguiendo las mismas ideas básicas que en el MLM, el modelo lineal generalizado (MLG) se puede extender a los datos longitudinales incluyendo efectos fijos y aletorios, permitiendo que un subconjunto de los coeficientes de regresión varíe aleatoriamente de una unidad a otra. Este tipo de modelo se conoce como modelo lineal generalizado mixto (MLGM). A diferencia del MLM, en el MLGM las variables de respuesta puede provenir de diferentes distribuciones además de la Gaussiana (normal), se utiliza una función enlace y los parámetros de regresión tienen interpretaciones específicas de la unidad, en lugar de la población. Es decir, los efectos fijos no describen cambios en la respuesta media de la población, sino que describen los cambios en la respuesta media de una unidad y la relación de estos cambios con las covariables. También, incluye una variedad de modelos que permiten la distribución normal, binomial, Poisson, multinomial, entre otras sobre la variable respuesta. Por lo tanto, es aplicable a casos en los que las observaciones pueden ser continuas o discretas.

El MLGM es un modelo lineal paramétrico para datos agrupados, longitudinales o medidas repetidas y presenta dos clases de parámetros, los parámetros de efectos fijos y los efectos aleatorios, los cuales se exponen a continuación.

a) Efectos fijos: describen la relación entre la variable respuesta y las covariables para toda una población de unidades de análisis o para un número relativamente pequeño de subpoblaciones definidas por niveles de un factor fijo. También, detallan diferencias entre los niveles de un factor fijo en términos de respuestas medias para la variable repuesta. Se supone que los efectos fijos son cantidades fijas desconocidas y su estimación se basa en el análisis de los datos recogidos en un estudio de investigación dado.

b) Efectos aleatorios: son valores aleatorios asociados con los niveles de un factor o factores aleatorios en un MLGM. Estos valores, explican la variación aleatoria en la variable respuesta, no se estiman directamente, pero se resumen de acuerdo a sus varianzas estimadas y covarianzas. Los efectos aleatorios pueden tomar la forma de intercepciones aleatorias o de coeficientes aleatorios, y la estructura de agrupación de los datos puede consistir en múltiples niveles de grupos anidados. En contraste con los efectos fijos, los efectos aleatorios se representan como variables aleatorias (no observadas).

Para analizar medidas repetidas se puede emplear el MLGM de diseño multinivel que considera dos o más dimensiones de análisis y tiene una estructura jerárquica (observaciones agrupadas en bloques). El primer nivel modela la evolución que sigue cada unidad de análisis a lo largo del tiempo y el segundo nivel representa la variación de las trayectorias entre unidades. Dicho modelo estima tanto los valores esperados de las observaciones (efectos fijos) como las varianzas y covarianzas de las observaciones (efectos aleatorios). Lo que distingue, por tanto, al MLGM del modelo lineal general, es el cálculo de los parámetros de covarianza que permiten analizar datos de carácter longitudinal, correlacionados, incompletos y con intervalos entre observaciones no constantes.

A diferencia del MLG, las respuestas bajo un MLGM están (marginalmente) correlacionadas. En los efectos aleatorios se incorpora la correlación entre las mediciones repetidas dentro de la unidad de análisis y facilitan la inferencia individual, ya que representan la influencia de cada unidad de

análisis (grupo) en las observaciones repetidas que no se capturan por las covariables observadas. Por esta razón, MLGM se utiliza a menudo para modelar respuestas correlacionadas.

Cabe agregar, que existen dos fuentes de variaciones en los datos longitudinales: la intra-individuo, que es la variación en las mediciones repetidas dentro de cada unidad de análisis y la entre-individuo, que es la variación en los datos entre diferentes unidades. El modelado dentro de la variación individual permite estudiar el cambio en el tiempo, mientras que el modelado entre las variaciones individuales permite entender las diferencias entre las unidades. Un MLGM incorpora específicamente ambas fuentes de variación, para esto utiliza efectos aleatorios para representar desviaciones de trayectorias longitudinales individuales del promedio poblacional.

El MLGM también es conocido en la literatura como modelo multinivel o modelo jerárquico, este modelo se utiliza cuando el objetivo es hacer inferencias sobre las unidades de análisis más que sobre la población, se asume que existe un conjunto de covariables que tienen un efecto aleatorio en la respuesta de diferentes unidades, el efecto de dichas variables, varía aleatoriamente de una unidad a otra.

La especificación general del MLM en un contexto de MLG se puede detallar en tres partes (Fitzmaurice *et al.*, 2012):

1. Se supone que, dado un vector de efectos aleatorios b, las mediciones repetidas $y_{i1}, y_{i2}, ..., y_{in_i}$ son independientes entre sí de tal manera que la distribución condicional de Y dado b es un miembro de la familia exponencial. Entonces, la función de densidad de probabilidad general de y_{ij} puede expresarse en la forma

$$f(y_{ij} \mid b_i) = exp\left\{\frac{[y_{ij}\theta_{ij} - b(\theta_{ij})]}{a_{ij}(\phi)} + c(y_{ij}, \phi)\right\}, \tag{2.2}$$

   donde $b(\theta_i)$, $a_i(\phi)$, $c(y_i, \phi)$ son funciones conocidas, ϕ parámetro de dispersión que puede o no ser conocido y θ_i se llama parámetro natural o canónico.

2. El MLGM tiene tanto efectos fijos, β, como efectos aleatorios, b_i, en el predictor lineal. La media condicional de Y_{ij}, dependerá de los efectos fijos y aleatorios a través del siguiente predictor lineal:

$$\eta_{ij} = x_{ij}^{'}\beta + z_{ij}^{'}b_i. \tag{2.3}$$

x y z son vectores conocidos, β un vector de parámetros desconocidos (los efectos fijos), y b_i son los parámetros de efectos aleatorios.
Se une el predictor lineal con la media condicional a través de una función enlace conocida $g(\cdot)$ tal que

$$g(\mu_{ij}) = g\{E(Y_{ij}|b_i)\} = \eta_{ij}. \tag{2.4}$$

3. La varianza se relaciona con la media condicional mediante una función de varianza con parámetros de efectos aleatorios $b_i = (b_{i1}, \ldots, b_{iq})$. La varianza condicional es dada por

$$var(Y_{ij} \mid b_i) = var\{E(Y_{ij}|b_i)\}\phi = \phi\, v(\mu_{ij}), \tag{2.5}$$

$v(\mu_{ij})$ es una función de varianza conocida que permite modelar la variabilidad y ϕ es un parámetro de escala que puede ser conocido o puede necesitar ser estimado.

En general el MLGM puede escribirse como

$$g(\mu_{ij}) = x_{ij}^{'}\beta + z_{ij}^{'}b_i, \qquad j = 1, 2, ..., n_i; \;\; i = 1, 2, ..., m, \tag{2.6}$$

donde i es la unidad de análisis, j es el tiempo, μ_{ij} es la media condicional, β es un vector de efectos fijos, x_{ij}, z_{ij} son vectores que contienen covariables y G es la matriz de varianza-covarianza.

Cada parámetro β, representa el efecto fijo de un cambio de una unidad en la covariable del vector x_{ij} correspondiente sobre el valor medio de la variable respuesta, suponiendo que las otras covariables permanecen constantes en algún valor. Estos parámetros β son efectos fijos que se desean estimar, y su combinación lineal con las covariables define la parte fija del modelo.

En definitiva, el MLGM se formula combinando una distribución de respuesta condicional $f(y_{ij} \mid b_i)$, dados los efectos aleatorios b_i asociado con la i-ésima unidad y las covariables del vector z_{ij}. Se asume una distribución de efectos aleatorios normal multivariada con la media de cero y la matriz de covarianza G de dimensión $q \times q$.

$$b_i \sim N(0, G)$$

Ya que, hay q efectos aleatorios en el modelo asociado a la i-ésima unidad, G es una matriz simétrica y cuadrada, la cual se escribe de la siguiente forma:

$$G = Var(b_i) \begin{pmatrix} var(b_{1i}) & cov(b_{1i}, b_{2i}) & \cdots & cov(b_{1i}, b_{qi}) \\ cov(b_{1i}, b_{2i}) & var(b_{2i}) & \cdots & cov(b_{2i}, b_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(b_{1i}, b_{qi}) & cov(b_{2i}, b_{qi}) & \cdots & var(b_{qi}) \end{pmatrix} \tag{2.7}$$

Los elementos a lo largo de la diagonal principal de la matriz G, representan las varianzas de cada efecto aleatorio en b_i, y los elementos fuera de la diagonal, representan las covarianzas entre dos efectos aleatorios correspondientes.

Las covariables y los efectos aleatorios determinan la unidad específica o la media condicional μ_{ij} y, los coeficientes de regresión β pueden por lo tanto ser interpretados como unidades específicas o efectos condicionales de covariables x_{ij}, dados los efectos aleatorios.

2.2. Estimación

La inferencia estadística de los modelos de regresión, tiene por objeto especificar las funciones de verosimilitud para la estimación de los parámetros del modelo. En la construcción del MLGM, el aspecto único de la inferencia estadística es la forma de predecir los efectos aleatorios, además de la estimación de los parámetros de regresión fija. Se basa típicamente en el método de verosimilitud en donde los parámetros de interés a estimar son β para los efectos fijos y b para los efectos aleatorios. Lo principal de la estimación de máxima verosimilitud, es determinar los valores del parámetro del modelo que maximizan la probabilidad. En el contexto de la familia exponencial, se realiza maximizando la función log-verosimilitud con respecto al parámetro canónico θ dada la observación y junto con el parámetro de escala ϕ.

El MLGM se puede ajustar mediante la maximización de la probabilidad de (2.2) por medio de la integración de los efectos aleatorios (Fitzmaurice *et al.*, 2008).

En MLGM la distribución marginal para y_i es

$$f(y_i \mid \beta, G) = \int \prod_{j=1}^{n_i} f(y_{ij} \mid \beta, \phi, b_i) f(b_i \mid G) db_i. \tag{2.8}$$

La probabilidad es dada por

$$L(\beta, G \mid y) = \int \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} \left[f(y_{ij} \mid \beta, \phi, b_i) f(b_i \mid G) db_i \right]. \tag{2.9}$$

Dado que los efectos aleatorios b no son observados, la inferencia sobre β y G se basa en la llamada función de verosimilitud conjunta.

La probabilidad de los datos observados es una probabilidad marginal, pero esta probabilidad no suele tener una expresión de forma cerrada y se debe emplear métodos aproximados de estimación mediante integración numérica. Debido a esto se utiliza el método cuadratura Gauss-Hermite, el cual se detalla más adelante.

La función log-verosimilitud de MLGM se puede formular de la siguiente forma:

$$log\ L(\beta, G, \phi) = \ log\ \left[\prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f(y_{ij} \mid \beta, b_i, \phi) f(b_i|, G) db_i\right]. \tag{2.10}$$

El segundo término en la segunda igualdad de la ecuación (2.10), no es fácil de estimar empíricamente. Hay una variedad de métodos para aproximar esta función log-verosimilitud. El procedimiento general consiste en estimar los efectos fijos y aleatorios, β y b_i, por separado como se indica en las siguientes ecuaciones de verosimilitud.

a) Ecuaciones de verosimilitud en parámetros de efectos fijos.
   Las ecuaciones de verosimilitud se pueden escribir de una forma simple, a pesar de que son numéricamente difíciles. La primera derivada parcial de la función log-verosimilitud con respecto a b puede escribirse como

$$L = log \int f(y \mid b) f(b) db = log f(y), \tag{2.11}$$

   así que

$$\frac{\partial L}{\partial \beta} = \frac{\partial}{\partial \beta} \int f(y \mid b) f(b) db \,/ f(y) \tag{2.12}$$

$$= \int \left[\frac{\partial}{\partial \beta} f(y \mid b)\right] f(b) db \,/ f(y) \tag{2.13}$$

$$\frac{\partial}{\partial \beta}\ L(\beta, G, \phi) = \frac{\partial}{\partial \beta} \int f(y \mid b) f(b \mid G) db \,/ f(y) \tag{2.14}$$

$$= \int \left[\frac{\partial}{\partial \beta}\ log\ f(y \mid b)\right] f(b \mid y) db, \tag{2.15}$$

Ya que $f(b)$ no implica β, la primera derivada parcial de la función log-verosimilitud con respecto a β puede escribirse

$$\frac{\partial}{\partial \beta} f(y \mid b) = \frac{\partial \, log \, f(y \mid b)}{\partial \beta} f(y \mid b) \tag{2.16}$$

Luego, la fórmula (2.12) se reescribe como sigue

$$\begin{aligned} \frac{\partial L}{\partial \beta} &= \int \frac{\partial \, log \, f(y|b)}{\partial \beta} f(y \mid b) f(b) db \, / f(y) && (2.17) \\ &= \int \frac{\partial \, log \, f(y|b)}{\partial \beta} f(b \mid y) db. && (2.18) \end{aligned}$$

b) Ecuaciones de verosimilitud para los parámetros de efectos aleatorios.

En cuanto a los efectos aleatorios, la primera derivada parcial de la función log-verosimilitud con respecto a b es

$$\begin{aligned} \frac{\partial}{\partial b} \, L(\beta, G, \phi) &= \int \frac{\partial \, log \, f(\beta)}{\partial b} f(b \mid y) db && (2.19) \\ &= E\left[\int \frac{\partial \, log \, f(\beta)}{\partial b} \mid y \right] && (2.20) \end{aligned}$$

Dado que no hay expresiones analíticas disponibles para las integrales en la ecuación (2.10), se requieren procedimientos estadísticos para especificar formas cercanas de la función log-verosimilitud empleando aproximaciones numéricas para la estimación de parámetros. Uno de los métodos que se puede utilizar para resolver esta dificultad, es el método cuadratura de Gauss-Hermite descrito en la siguiente sección.

2.2.1. Método cuadratura de Gauss-Hermite

En la derivación de estimación de parámetros, resulta complejo evaluar la función log-verosimilitud como una integral múltiple. En este caso, se utiliza un método de aproximación de la integral en MLGM llamado cuadratura Gauss-Hermite. Este método consiste en aproximar las integrales de funciones dadas por una suma ponderada de evaluaciones funcionales en puntos de cuadratura seleccionados, es decir, aproxima la integración en la verosimilitud por una suma en un número específico de puntos de cuadratura para cada dimensión de la integración, suponiendo que los efectos aleatorios se distribuyen normalmente (Rabe-Hesketh *et al.*, 2002). Dada la aproximación, se puede seleccionar el número de puntos en cuadratura con un nivel de precisión deseado y se logra aproximar la integral a cualquier grado práctico de precisión, por lo que se denomina un método exacto.

Sea una función conocida $f(t)$ y $g(t)$ una función de densidad de probabilidad, en donde la función $f(t)$ puede ser integrada contra $g(t)$.

El método de cuadratura de Gauss-Hermite aproxima la integral por

$$\int f(t)g(t)dt \approx \sum_{r=1}^{R} p_r \ f(t_r). \tag{2.21}$$

En el contexto de MLGM, $f(t)$ puede considerarse como la distribución condicional de datos longitudinales dados los efectos aleatorios, y $g(t)$ representa la distribución de efectos aleatorios, respectivamente, se define como la densidad de la distribución normal estándar. El peso en cuadratura es $p_r, (r = 1, ..., R)$, t_r es la abscisa, estadísticamente denominada nodo y d indica el número de puntos en cuadratura. De una manera simple se puede entender la cuadratura aproximada como un promedio ponderado. Teniendo en cuenta esta explicación el método de cuadratura Gauss-Hermite se ajusta a (2.8) como una función de verosimilitud.

Es conveniente modificar las variables de integración a los efectos aleatorios estándar distribuidos normalmente independientes v_i, utilizando por ejemplo la descomposición de Cholesky Q de la matriz de covarianza G para

que $b_i = Qv_i$. La log-probabilidad se puede escribir como

$$L(\beta, G) = \ log \ \prod_{i=1}^{m} \int_{-\infty}^{\infty} \phi(v_{iq}) \dots \left[\int_{-\infty}^{\infty} \phi(v_{i1}) \int \prod_{j=1}^{n_i} f(y_{ij} \mid v_i) dv_{i1} \right] \dots dv_{iq}. \tag{2.22}$$

Donde $\phi(\cdot)$ es la función de densidad normal estándar univariante. Cada integral unidimensional puede ser aproximada por cuadratura de Gauss-Hermite como sigue

$$\int_{-\infty}^{\infty} \phi(v_{i1}) \prod_{j=1}^{n_i} f(y_{ij} \mid v_i) dv_{i1} \approx \sum_{r=1}^{R} p_r \prod_{j=1}^{n_i} f(y_{ij} \mid a_{ir}). \tag{2.23}$$

Cabe señalar que si el producto de las distribuciones de respuesta condicional en la ecuación está bien aproximado por un polinomio de bajo grado en v_{i1} la cuadratura funcionará bien. Sin embargo, en la práctica se requiere a menudo un gran número de puntos de cuadratura para aproximar la probabilidad. Este caso ocurre en las grandes correlaciones n_i (Rabe-Hesketh *et al.*, 2002). Donde los lugares y pesos específicos de la unidad son dados por

$$a_{ir} \equiv \tau_i a_r + \mu_i, \tag{2.24}$$

y

$$p_{ir} \equiv \sqrt{2\pi} \ \tau_i \ exp \ (a_r^2/2) \ \phi \ (\tau_i a_r + \mu_i) \ p_r. \tag{2.25}$$

Con media μ_i específica del grupo y la varianza τ_i^2.

2.3. Pruebas de hipótesis y criterio de evaluación

La inferencia estadística permite probar hipótesis, seleccionar el mejor modelo y obtener conclusiones a través de las estimaciones de los parámetros del MLGM. En cuanto a las pruebas de hipótesis, se compara las pruebas bajo la hipótesis nula estimando el valor p.

Uno de los intereses habituales es probar hipótesis sobre los parámetros que se han estimado. Pueden formularse en el contexto de dos modelos que tienen una relación de anidación, el modelo de referencia que abarca tanto la hipótesis nula como la alternativa y modelo anidado que satisface la hipótesis nula. Con efectos fijos, se prueba hipótesis de que las diferencias entre los niveles de un factor son cero o que igualan alguna constante predeterminada. En los efectos aleatorios, una hipótesis útil es que un componente de varianza es cero, que es igual a algún valor predeterminado. Las hipótesis sobre los parámetros en un MLGM se especifican proporcionando hipótesis nulas (H_0) y alternativas (H_1) sobre los parámetros en cuestión. Para un modelo escalar parámetro β las pruebas son $H_0 : \beta = \beta_0$ contra $H_1 : \beta \neq \beta_0$. Aquí, β_0 denota un valor nulo particular, típicamente 0, no el parámetro de intercepción. Para las pruebas de hipótesis, se utilizan ampliamente pruebas basadas en la verosimilitud: la prueba de Wald y la prueba de razón de verosimilitud. Estas dos pruebas se describen brevemente como sigue.

- Prueba de Wald. Propuesta por Abraham Wald en 1943, para la prueba de H_0 frente a H_1 está dada por

$$T_W = (\hat{\beta} - \beta_0)' \, \hat{G}^{-1} \, (\hat{\beta} - \beta_0) \; \sim \; \chi^2_p, \tag{2.26}$$

  donde la estadística de prueba T_W asintóticamente en H_0, donde p es la dimensión del parámetro β y $\hat{G} = I(\hat{\beta})^{-1}$ es una estimación de la matriz de covarianza de $\hat{\beta}$. Para probar un componente individual de β, sea $H_0 : \beta_j = \beta_{j0}$ versus $H_1 : \beta_j \neq \beta_{j0}$, se puede considerar la

estadística de prueba de tipo Wald individual.

$$T_W = \frac{(\hat{\beta}_j - \beta_0)^2}{\hat{var}(\hat{\beta}_j)}, \tag{2.27}$$

donde $\hat{var}(\hat{\beta}_j) = (I(\hat{\beta})^{-1})_{jj}$. La estadística de prueba $T_W \sim \chi^2_p$ asintóticamente bajo H_0.

- Prueba de razón de verosimilitud. Propuesto por Neyman y Pearson, se puede emplear para probar hipótesis sobre los parámetros de covarianza o parámetros de efectos fijos en el contexto de MLGM. Se basa en comparar los valores de las funciones de verosimilitud para dos modelos, el modelo anidado (hipótesis nula) y el modelo de referencia correspondiente a una hipótesis especificada, que definen una hipótesis que se está probando. La estadística de prueba se calcula restando dos veces la probabilidad de log para el modelo de referencia de la del modelo anidado. Entonces, la estadística de prueba de H_0 frente a H_1 es

$$T_V = 2\ log\ L(\hat{\beta}) - 2\ log\ L(\beta_0). \tag{2.28}$$

  La estadística de prueba $T_V \sim \chi^2_p$ asintóticamente bajo H_0.

  En (2.28), $\hat{\beta}$ es la maximización de β sobre el rango completo de valores de cada elemento de β, $L(\beta_0)$ se refiere al valor de la función de verosimilitud evaluada en las estimaciones de los parámetros del modelo anidado y $L(\hat{\beta})$ se refiere al valor de la función de verosimilitud en el modelo de referencia.

- Selección del modelo. Para seleccionar el modelo adecuado, es necesario realizar pruebas de un conjunto de modelos estadísticos para determinar las covariables que se incluirán en el modelo final, para esto se utiliza el criterio de información Akaike (AIC) con el propósito de hallar el modelo que presenta el ajuste de muestra más cercano al verdadero ajuste del modelo. El AIC puede calcularse basándose en la probabilidad logarítmica, de un modelo ajustado y se definen como (Wu, 2009):

$$AIC = -2\ell(\hat{\beta}) + 2p, \tag{2.29}$$

donde $\ell(\hat{\beta})$ es la log-verosimilitud maximizada bajo el modelo ajustado y p representa el número total de parámetros que se están estimando en el modelo, tanto para los efectos fijos como aleatorios. Al momento de comparar los modelos, se escoge el modelo con valores AIC más pequeños, ya que este valor indica que tiene la menor pérdida de información.

2.4. Diagnóstico

Después de ajustar un modelo, es importante llevar a cabo un diagnóstico para comprobar las suposiciones del modelo. Para el MLGM el diagnóstico puede basarse en el análisis de los residuos de Pearson o del desvío residual.

Una suposición comúnmente usada con respecto a los efectos aleatorios es que se distribuyen normal. Lange y Ryan (1989) consideraron el modelo longitudinal suponiendo que G desarrolló una gráfica normal ponderada para evaluar la normalidad de los efectos aleatorios en un modelo longitudinal, suponiendo que se conocen los parámetros fijos y el vector de componentes de varianza. Los residuos son las diferencias entre los valores estimados por el modelo y los valores observados. Se utilizan para decidir si existe o no un patrón específico en los residuos, o sea, decidir si un determinado conjunto de residuos representados gráficamente contra los valores ajustados indican un patrón aleatorio o no. En el contexto de MLGM el desvío residual frente a los valores ajustados se utilizan para verificar los supuestos del modelo y para detectar valores atípicos. Este tipo de residuos tienen las mejores propiedades para examinar la bondad de ajuste de un MLGM y están aproximadamente distribuidos normalmente si el modelo está correctamente especificado. Pueden trazarse contra los valores ajustados o contra una covariable para inspeccionar el ajuste del modelo (McCulloch y Neuhaus, 2001).

Se define el desvío residual de la siguiente forma:

$$r_{ij} = signo(y_{ij} - \hat{\mu_{ij}})\sqrt{d_{ij}}, \tag{2.30}$$

donde d_{ij} es el componente de desvío que mide la diferencia de los logaritmos de la función de verosimilitud observada y ajustada y μ_{ij} es la media pronósticada condicionada a los efectos aleatorios.

La elección para diagnosticar efectos aleatorios es considerar los valores ajustados de los efectos aleatorios, debido a sus propiedades. Se recomienda el uso de diagnóstico estándar como el Q-Q plot para comprobar visualmente la normalidad de la distribución de los efectos aleatorios y para investigar los posibles valores atípicos que pueden justificar una investigación adicional (West *et al.*, 2014).

En el capítulo que sigue se presentan las variables estudiadas, un análisis exploratorio de los datos escogidos y la aplicación del MLGM, empleando el software estadístico Stata versión 14.

CAPÍTULO 3

Aplicación

En este capítulo se presenta el MLGM aplicado a los datos obtenidos a partir de las empresas que desarrollan la actividad minera, el cual se utiliza a menudo para el análisis de datos longitudinales cuando el investigador está interesado en modelar los efectos del tiempo y otras covariables que varían en el tiempo.

Para la aplicación del modelo, se consideró el conjunto de datos de la actividad económica de minería metálica obtenidos por la Comisión Chilena del Cobre (COCHILCO) y el Servicio Nacional de Geología y Minería (SERNAGEOMIN).

Los datos que se utilizaron para la aplicación del modelo consisten en mediciones de la producción de minerales metálicos de veinte empresas durante un periodo de doce meses sucesivos, desde enero hasta diciembre de 2011. Los minerales estudiados son el cobre, molibdeno, hierro y zinc, los cuales son analizados en conjunto.

Las unidades de análisis son las empresas que desarrollan la actividad minera y pertenecen a la pequeña, mediana y gran minería. Cabe señalar que todas las empresas presentan registro de todos los meses, lo que significa que los datos son balanceados, es decir, no hay datos faltantes o perdidos.

El propósito de este estudio es analizar la cantidad de producción de los minerales metálicos y determinar si las ventas, el tiempo y la exportación influyen en la cantidad de producción.

Las variables que se utilizaron en el análisis se presentan en la tabla 3.1.

Tabla 3.1: Clasificación de las variables de estudio.

Variable	**Tipo de variable**	**Unidad de medida**
Producción	Respuesta - Continua	Toneladas métricas
Tiempo	Covariable - Discreta	Meses
Ventas	Covariable - Continua	Toneladas métricas
Exportación	Covariable - Continua	Toneladas métricas

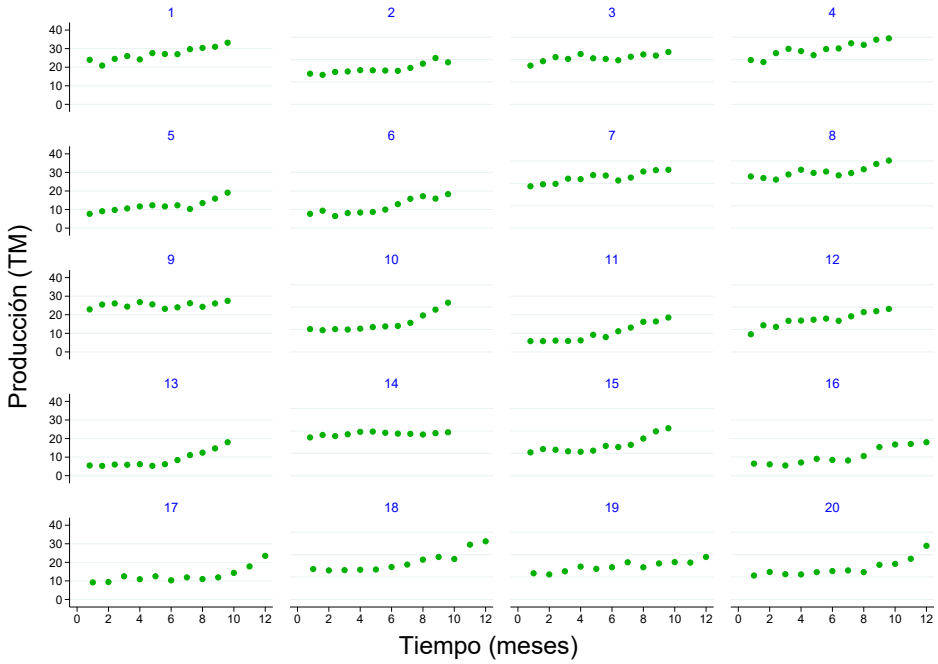
Fuente: Elaboración propia

Las covariables que se desean estudiar son tres; el tiempo que representa los meses en los que se midió la variable producción, la variable ventas que corresponde a la cantidad de minerales metálicos vendidos en cada mes registrado y la exportación que representa la cantidad de minerales metálicos exportados en diferentes países.

3.1. Análisis exploratorio

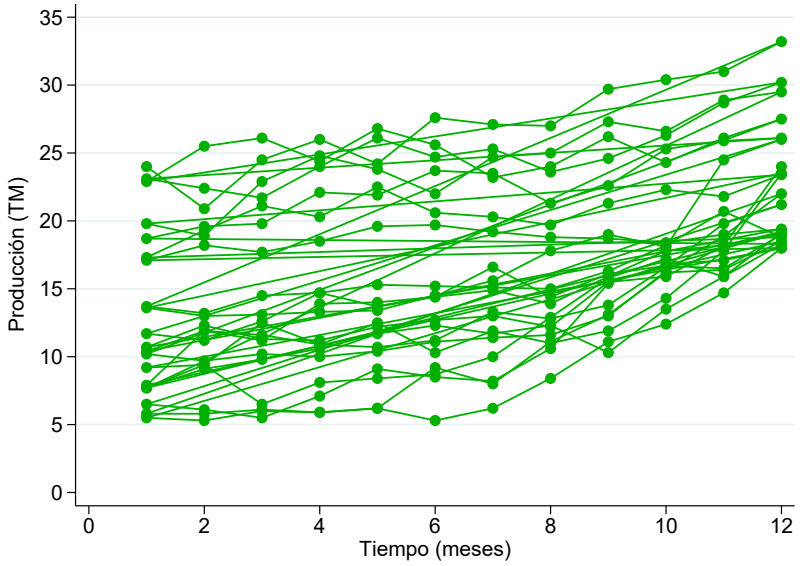
Antes de la aplicación del modelo, es necesario realizar un análisis exploratorio de los datos escogidos.

Figura 3.1: Dispersión de la producción de minerales metálicos de cada empresa en el tiempo.



La Figura 3.1 corresponde a los perfiles individuales de las veinte empresas consideradas en el análisis para la producción de minerales metálicos en función del tiempo en meses. La trayectoria longitudinal de la producción de cada empresa experimenta una posible tendencia lineal entre producción y el tiempo en el periodo estudiado, lo que significa que la cantidad de minerales producidos aumentaría a medida que aumentan los meses y el número de mediciones en cada empresa varía, ya que depende del tiempo de producción registrado.

Figura 3.2: Gráfico de tiempo de la producción de minerales metálicos de cada empresa minera.



Al visualizar las toneladas métricas de los minerales metálicos producidos para las veinte empresas. Se aprecia que existe una variabilidad dentro de cada empresa en la producción. Esto se puede distinguir a partir de la apariencia algo irregular de los segmentos de línea que unen las medidas repetidas sobre cualquier empresa. También, se observa que las trayectorias de producción varían entre las empresas, es decir, existe variabilidad entre empresas (variación entre-individuo), por lo que un modelo con parámetros de regresión individuales puede ser útil para modelar estos datos. Tal modelo se denomina modelo de efectos mixtos, en el que se incluyen efectos aleatorios en el modelo para representar los efectos individuales (empresas).

3.2. Modelo estadístico propuesto

Para analizar las mediciones repetidas en el tiempo de cada empresa, se puede considerar la estructura jerárquica con las observaciones repetidas agrupadas dentro de cada empresa y el tiempo se puede tomar en cuenta como una covariable dentro de cada grupo. El primer nivel modela la evolución que sigue cada empresa a lo largo del tiempo y el segundo nivel representa la variación de las trayectorias entre las empresas.

El conjunto de datos consiste en mediciones de la cantidad de producción de 20 empresas en 12 meses sucesivos, lo que da un total de 240 observaciones. Cada empresa experimenta una tendencia lineal, pero las mediciones de producción general varían de empresa a empresa. Las empresas se tratan como una muestra aleatoria y se modela la variabilidad entre empresas como un efecto aleatorio. Por lo tanto, se desea ajustar el modelo

$$g(\mu_{ij}) = \eta_{ij} = x_{ij}^{'}\beta + z_{ij}^{'}b_i, \tag{3.1}$$

para $i = 1, 2, ..., 20$ empresas y $j = 1, 2, ..., 12$ meses.

Donde x_{ij} es el vector de covariables fijas de la empresa i en el tiempo j, β es el vector de parámetros desconocidos correspondientes a los efectos fijos, b_i es el vector de parámetros de efectos aleatorios y z_{ij} es el vector de covariables correspondientes a los efectos aleatorios.

Sea Y_{ij} la producción de minerales metálicos en toneladas métricas de las diferentes empresas. Se asume que $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$ y una función enlace $g(\mu_{ij}) = \mu_{ij}$, la cual corresponde a la identidad.
Para $z_{ij}b_i$ se asume que b_i tiene una matriz de varianza-covarianza, tal que $Var(b_i) = G$. Los efectos aleatorios no se estiman directamente, sino que se caracterizan por los elementos de G, conocidos como componentes de varianza y se supone que $b \sim N(0, G)$.

Los parámetros del MLGM se estiman utilizando el método de máxima verosimilitud y se emplea un método de integración especial para su desarrollo, presentado en el capítulo 2. Luego de obtener las estimaciones se realizan

pruebas de modelos para saber que variables se incluirán en el modelo final, para esto se considera el criterio de información Akaike (AIC), con el propósito de escoger el modelo más apropiado o adecuado para el conjunto de datos en estudio. En la siguiente tabla se da a conocer los valores de AIC de los diferentes modelos propuestos.

Tabla 3.2: Criterio de información Akaike.

Modelo	**df**	**AIC**
Tiempo + ventas	5	1095,795
Tiempo + exportación	5	1101,532
Ventas + exportación	5	1236,779
Tiempo + ventas + exportación	6	1062,713

Para establecer el modelo más apropiado, se debe seleccionar el menor valor entre todos los valores de AIC. En este caso, el modelo que muestra un mejor ajuste al conjunto de datos acorde a un valor de AIC= 1062,71, es aquel modelo que incluye la variable tiempo, ventas y exportación, ya que este valor es el más pequeño de todos los AIC.

En la tabla 3.3 se presenta los estimadores de los coeficientes de regresión para el modelo (3.1).

Tabla 3.3: Estimaciones de los parámetros de efectos fijos en el modelo lineal generalizado mixto.

Coeficientes	**Estimación**	**Error estándar**	**Valor p**	**Intervalos**
Intercepto	9,3336	1,5213	$< 0,0001$	6,3518 - 12,3154
Tiempo	0,6560	0,0792	$< 0,0001$	0,5009 - 0,8111
Ventas	0,2284	0,0631	0,0010	0,1048 - 0,3520
Exportación	0,0013	0,0032	0,0413	-0,0049 - 0,0075

Observación: en la tabla 3.3 el intervalo de confianza del coeficiente exportación contiene al 0, significa que con la información disponible cabe la posibilidad de que las toneladas métricas de minerales metálicos exportados no afectan en la producción de minerales metálicos a través del tiempo. Sin embargo, se considera el valor p para determinar si la covariable exportación es significativa o no en el modelo. En este caso el valor p señala que los minerales metálicos exportados afectan en la cantidad de minerales metálicos producidos.

Al realizar el ajuste del modelo y observar el valor p, se puede determinar que las covariables tiempo, ventas y exportación son significativas para el modelo. Por lo tanto, las ventas y la exportación están causando algún efecto en la producción de minerales metálicos.

Conociendo los valores de las estimaciones de los parámetros para los efectos fijos, se puede decir lo siguiente:

- A medida que el tiempo avanza, la producción promedio de minerales metálicos aumenta 9,33 toneladas métricas.
- Un incremento en las ventas, la producción aumenta aproximadamente en 0,23 toneladas métricas de minerales metálicos producidos.
- En las empresas mineras, por cada unidad que aumenta la exportación, la producción varía aumentando 0,001 toneladas métricas.
- En cuanto al error estándar, se puede deducir que el valor de la covariable exportación es el más pequeño con un error de 0,003, de manera que el MLGM logro estimar el coeficiente exportación con mayor precisión, en comparación con el resto de las covariables.

Tabla 3.4: Estimaciones de los parámetros de covarianza asociados con los efectos aleatorios en el modelo lineal generalizado mixto.

Componentes de varianza	**Estimación**	**Error estándar**	**Intervalos**
Varianza (intercepto)	32,7228	10,9491	16,9838 - 63,0470
Varianza (tiempo)	0,0853	0,0336	0,0394 - 0,1846
Varianza (residual)	2,6342	0,2667	2,1601 - 3,2124

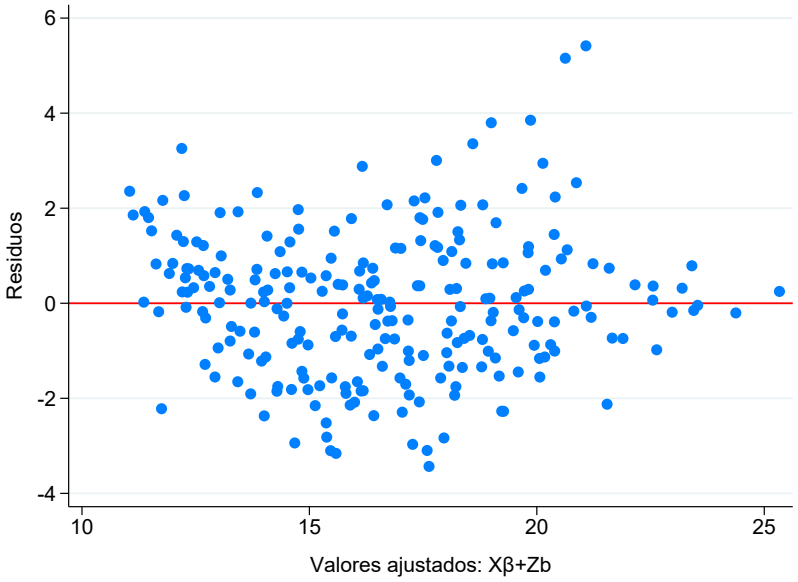
En los efectos aleatorios se logra conocer las estimaciones de los componentes de varianza y por medio de estos componentes se deduce lo siguiente:

- La variabilidad de cada empresa, en cuanto a la producción de minerales metálicos en el tiempo, es de 0,09.
- La variabilidad entre las 20 empresas es 32,72, esto indica que las 20 empresas difieren entre sí como las observaciones dentro de cada una de ellas.
- La variabilidad entre las observaciones dentro de las empresas mineras es 2,63.
- El MLGM logró estimar la variación entre las empresas en el tiempo con mayor precisión, con un error estándar de 0,03.

3.3. Diagnóstico del modelo ajustado

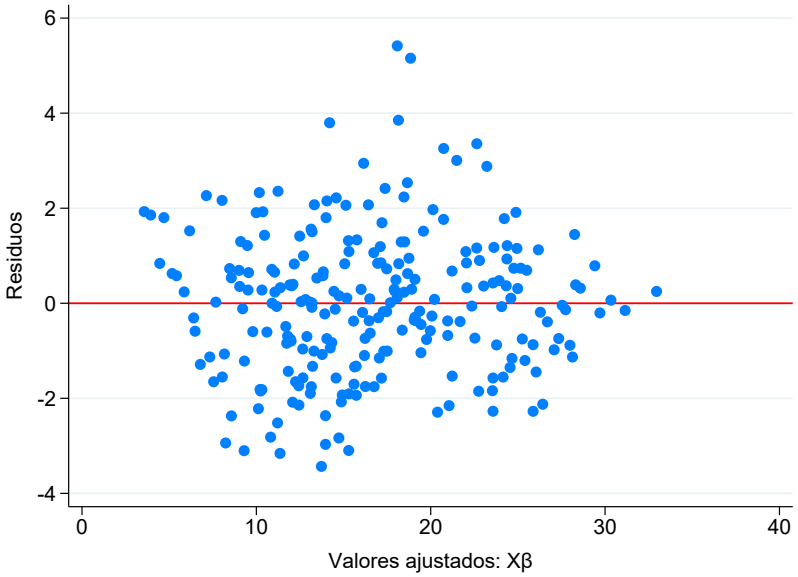
Después de seleccionar el modelo, lo que sigue es realizar un análisis de diagnóstico para verificar el ajuste de los datos en el MLGM. Para llevar a cabo el diagnóstico se deben utilizar los residuos, ya que estos expresan la diferencia entre una observación y su valor ajustado, también permiten indicar la presencia de valores anormales o atípicos y conocer la variabilidad de los datos. Para este diagnóstico se emplea el desvío residual.

Figura 3.3: Gráfico de residuos versus valores ajustados.



En la gráfica de la figura 3.3 la dispersión de los residuos en los valores ajustados es algo irregular, se aprecia heterocedasticidad en los residuos en función de los valores ajustados del modelo, quizás por la presencia de datos atípicos. Se muestra un patrón aleatorio de residuos a ambos lados del cero, es decir, puntos esparcidos aleatoriamente a ambos lados del cero.

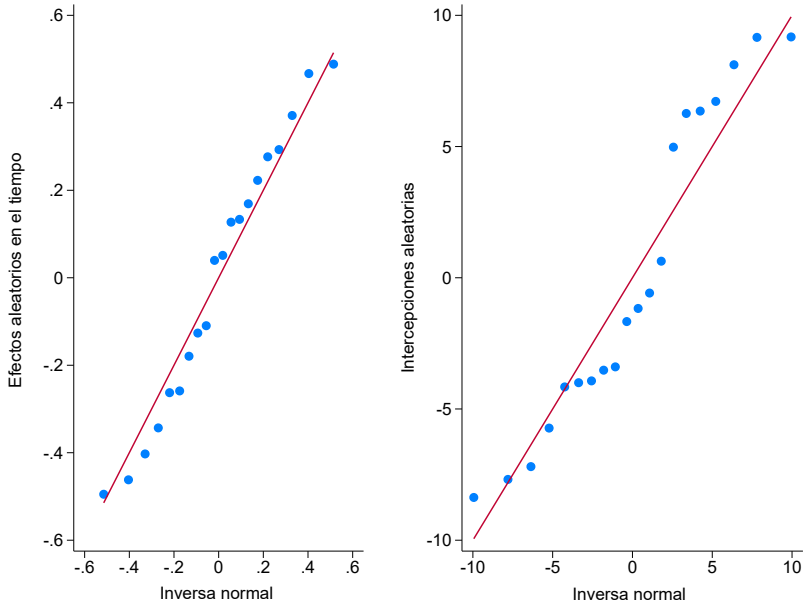
Figura 3.4: Gráfico de residuos versus valores ajustados parte fija del modelo.



En la figura 3.4 se muestra el desvío residual versus valores ajustados de los efectos fijos del modelo. Aparentemente existen posibles datos extremos, ya que hay valores alejados del cero, puntos crecientes y drecientes.

Esto se puede comprender por el hecho de que algunas de las empresas producen toneladas métricas de minerales metálicos constantemente altos, mientras que otras producen toneladas de minerales metálicos constantemente bajos, esto provoca una posible variabilidad no constante.

Figura 3.5: Efectos aleatorios de las empresas.



En la figura 3.5 se presenta el diagnóstico de los efectos aleatorios del modelo ajustado utilizando el gráfico Q-Q plot normal para cada conjunto de valores ajustados específicos de la empresa, para así comprobar la existencia valores atípicos y el supuesto de que los efectos aleatorios siguen una distribución normal. En este caso, no se observan valores atípicos positivos y tampoco negativos en términos de efectos aleatorios asociados con las empresas. Los puntos están situados casi sobre la línea recta, lo cual es un indicio de que los efectos aleatorios de este análisis parecen seguir una distribución normal.

Conclusión

En este estudio se aplicó el MLGM a la producción de minerales metálicos y el método que se utilizó para la estimación de los parámetros fue el método de máxima verosimilitud, integrando sobre los efectos aleatorios. En la práctica el método de estimación resulta complejo, debido a la presencia de integrales intratables en la derivación de estimación de parámetros, al momento de evaluar la función log-verosimilitud. Esta situación conduce a emplear el método de aproximación numérica cuadratura Gauss-Hermite, puesto que permite maximizar la probabilidad marginal con cualquier grado de precisión deseado y es especialmente interesante para los datos longitudinales donde la dimensión de los efectos aleatorios es a menudo relativamente baja.

El objetivo fundamental fue analizar la cantidad de producción de los minerales metálicos en toneladas métricas registrados durante un periodo de doce meses consecutivos. Esto se pudo lograr al ajustar el MLGM, estimando los parámetros fijos y aleatorios del modelo. De esta manera, se confirmó que durante el tiempo que se registró la producción de diferentes empresas mineras, la cantidad de ventas y de exportación en toneladas métricas influyeron de manera positiva en la producción de minerales metálicos. Se determina que un incremento de las ventas, la producción creció aproximadamente en 0,23 toneladas métricas y cuando variaban las toneladas de minerales exportados, la producción de minerales metálicos aumentó en 0,001 toneladas métricas.

También, los cambios que presentó la producción de minerales metálicos a través del tiempo afectaron de manera positiva, a medida que iba aumentando el número de meses, la cantidad de minerales metálicos producidos creció en 0,66 toneladas métricas.

El MLGM resulta apropiado para obtener lo componentes de varianza que se encuentran en los efectos aleatorios, permitiendo conocer la variación entre las diferentes empresas mineras. La variación señala el grado de dispersión de los datos, permitiendo conocer la producción entre empresas y de cada empresa minera. Al realizar el análisis de los efectos aleatorios, se establece que las veinte empresas varían entre sí al igual que las observaciones dentro de cada empresa con una variación de 32,72, en este caso se presenta una variabilidad alta. Esto se debe a la presencia de correlación entre las respuestas dentro de los grupos de observaciones. El valor del error estándar del intercepto aleatorio también es alto, dado que el espacio de inferencia es más amplio. La gran variabilidad observada entre las empresas provoca que la magnitud de los errores estándares obtenidos en el MLGM sean valores altos.

Por otra parte, en el diagnóstico del modelo ajustado no se aprecia una evidencia fuerte de una varianza no constante. Sin embargo, hay evidencia de valores atípicos, dado que se registraron toneladas métricas extremas de minerales metálicos producidos en comparación con otras. Un análisis sin estos valores produjo estimaciones similares para cada uno de los efectos fijos incluidos en el modelo, lo que sugiere que esta observación, no tuvo influencia sobre estas estimaciones. Y, por último, el MLGM aplicado funcionó correctamente para este tipo de datos.

Apéndice

Código en Stata para la implementación de la base de datos.

```
clear all
version 14

/* Importar base de datos */
import excel "C:\ Users\ Joselin\ Documents\ Datos Proyecto\ Datos
01.xlsx", sheet("Hoja1") firstrow

format produccion %10.0g
format ventas %10.0g
format exportacion %9.0g
recast long exportacion, force
format produccion %9.0g
recast long produccion, force
format ventas %9.0g
recast long ventas, force

**** Guardar base de datos modificada ****
save "C:\Users\Joselin\Documents\Datos Proyecto\Produccion
minerales.dta"
clear all
use "C:\Users\Joselin\Documents\Datos Proyecto\Produccion
minerales.dta"
```

```
*** Análisis exploratorio ***
/* Gráfico de dispersión de cada empresa */
scatter produccion tiempo, by(empresa)

**** MLG mixto o multinivel ****
** Prueba de modelos a través de AIC: Criterio de información Akaike ***

/* Modelo 1 */
meglm produccion tiempo ventas || empresa:, family(gaussian) link(identity)
estat ic

/* Modelo 2 */
meglm produccion tiempo exportacion || empresa:, family(gaussian)
link(identity)
estat ic

/* Modelo 3 */
meglm produccion ventas exportacion || empresa:, family(gaussian)
link(identity)
estat ic

/* Modelo 4 */
meglm produccion tiempo ventas exportacion || empresa: tiempo,
family(gaussian) link(identity)
estat ic

/* Modelo Final */
meglm produccion tiempo ventas exportacion || empresa: tiempo,
covariance(unstructured )

*** Diagnóstico para el modelo ***

/* Desvianza residual */
predict residuos, deviance

/* Valores ajustados (parte fija y aleatoria) */
predict rev, fitted

/* Valores ajustados para la parte fija */
predict predictor1, xb
```

```
*** Gráficos del diagnóstico ***
/*Residuos frente a valores ajustados parte fija y aleatoria*/
twoway scatter residuos rev, yline(0)

/*Residuos frente a valores ajustados parte fija*/
twoway scatter residuos predictor1, yline(0)

*** Diagnóstico para los efectos aleatorios ***
/*Creación de valores ajustados para el efecto aleatorio (empresa)*/
predict emva*, reffects ebmodes

/* Datos de cada empresa*/
collapse emva1 emva2, by(empresa)

/* Gráfico efectos aleatorios en el tiempo*/
qnorm emva1, ytitle(Efectos aleatorios en el tiempo)

/* Gráfico intercepciones aleatorias */
qnorm emva2, ytitle(Intercepciones aleatorias)
```

Referencias

Box, G. E. (1950). Problems in the analysis of growth and wear curves. *Biometrics*, 6(4):362–389.

Delgado Rodríguez, M. y Llorca Díaz, J. (2004). Estudios longitudinales: concepto y particularidades. *Revista española de salud pública*, 78(2):141–148.

Diggle, P. (2002). *Analysis of longitudinal data*. Oxford University Press.

Fitzmaurice, G., Davidian, M., Verbeke, G., y Molenberghs, G. (2008). *Longitudinal data analysis*. CRC Press.

Fitzmaurice, G. M., Laird, N. M., y Ware, J. H. (2012). *Applied longitudinal analysis*, volumen 998. John Wiley & Sons.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media.

Kachman, S. D. (2000). An introduction to generalized linear mixed models. pp. 59–73. Citeseer.

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

Lange, N. y Ryan, L. (1989). Assessing normality in random effects models. *The Annals of Statistics*, pp. 624–642.

Liang, K.-Y. y Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, pp. 13–22.

McCulloch, C. E. y Neuhaus, J. M. (2001). *Generalized linear mixed models*. Wiley Online Library.

Nelder, J. A. y Baker, R. J. (1972). Generalized linear models. *Encyclopedia of statistical sciences*.

Potthoff, R. F. y Roy, S. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, pp. 313–326.

Rabe-Hesketh, Sophia and Skrondal, Anders and Pickles, Andrew and others (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1):1–21.

West, B. T., Galecki, A. T., y Welch, K. B. (2014). *Linear mixed models*. CRC Press.

Wishart, J. (1938). Growth-rate determinations in nutrition studies with the bacon pig, and their analysis. *Biometrika*, 30(1/2):16–28.

Wu, L. (2009). *Mixed effects models for complex data*. CRC Press.