BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

POSTGRADO EN CIENCIAS MATEMÁTICAS

Modelos Mixtos en la determinación del Carbono Orgánico en la Hojarasca en una zona de Teziutlán, Puebla.

T E S I S

Que para obtener el Grado de:

MAESTRO EN CIENCIAS MATEMÁTICAS

Presenta:

Ana Aleyda Oroza Hernández

Directores de Tesis:

Dra. Gladys Linares Fleites
Dra. Hortensia Josefina Reyes Cervantes

Puebla, Pue. Enero 2015

Dedicatoria

A mis padres José Mauro Macario Oroza Hernández
y María Lucia Hernández Primero
quienes son mi ejemplo de vida.

Agradecimientos

A mis Padres...
A ustedes, papi José Mauro Oroza Hernández y mami María Lucia Hernández Primero, por haber hecho la persona que hoy soy, y que nunca me dejaron sin un consejo para educarme.
A ustedes que siempre han demostrado que el amor lo puede todo y siempre me han apoyado en lo que he necesitado. A ustedes, papá y mamá, porque los amo.

A Germán Vazquez...
A ti, Germán Antonio Vazquez Romero, gracias por todo tu amor y apoyo que me has brindado.

A mis Amigos...
Por el apoyo recibido durante la maestría, por la amistad que surgió y por los momentos especiales que vivimos. Gracias.

A mis Asesoras...
A ustedes que con su dedicación me brindaron sus conocimientos en aquellos momentos que fui afortunada en tenerlas como profesoras. A ustedes que me dieron la oportunidad de trabajar a su lado en el desarrollo de este proyecto. Mis sinceros y profundos agradecimientos.

A mis Sinodales...
A mis sinodales: Dr Hugo Adán Cruz Suárez, Dr. Francisco Solano Tajonar Sanabria, Dra. Maribel Castillo y Dr. Fernando Velasco Luna; a quienes agradezco inmensamente por haber aceptado formar parte de mi jurado y por el tiempo dedicado a la revisión de este trabajo.

A Conacyt...
Por haberme otorgado una beca durante el período de dos años que sin duda fue de gran ayuda para mi sustento. Mis más sinceros agradecimientos.

Introducción

La práctica de la modelación estadística ha estado en constante cambio como resultado del desarrollo de diferentes enfoques metodológicos de la Estadística y el progreso de las Ciencias Computacionales. En los últimos decenios se han alcanzado enormes desarrollos en los resultados analíticos del Modelo Lineal General [McCulloch, 2001], que incluye los modelos de Regresión, de Análisis de Varianza ($ANOVA$) y de Análisis de Covarianza. También, dentro del supuesto de normalidad de los errores, pero permitiendo la heteroscedasticidad de la varianza, ha habido considerables trabajos sobre los Modelos Lineales Mixtos, donde la estructura de la varianza está basada sobre efectos aleatorios.

Estos desarrollos están permitiendo modelar muchos aspectos locales, regionales y globales de la problemática ambiental del cambio climático, que se considera el problema ambiental más importante al que se está enfrentado la humanidad en la actualidad. Se define como cambio climático el posible aumento en la temperatura superficial del planeta que se produciría como consecuencia de un aumento importante y rápido de las concentraciones de gases de invernadero en la atmósfera. La causa fundamental de este incremento es la emisión de estos gases provocados por actividades humanas (antropogénicas) que alteran la composición original de la atmósfera.

La comunidad internacional ha emprendido acciones para enfrentar el cambio climático, siendo las más notables la firma de la Convención Marco de las Naciones Unidas sobre el Cambio Climático (CMNUCC), y de su Protocolo de Kyoto, que estableció metas concretas de reducción de emisiones de gases de efecto invernadero para los países desarrollados. México firmó la Convención en 1992, y la ratificó en 1993, entrando en vigor en 1994 [Masera, 2004].

El propósito de la CMNUCC y el Protocolo de Kyoto es lograr estabilizar las concentraciones de gases efecto invernadero, en la atmósfera, a niveles que eviten la interferencia antropogénicas en el sistema climático. Una de las dos formas de provocar las reducciones de emisiones de gases de efecto invernadero, según el Protocolo de Kyoto, es la de incrementar la creación de sumideros de carbono (transferencia neta de carbono orgánico (CO_2) atmosférico a la vegetación y al suelo para su almacenamiento, proceso conocido como secuestro de carbono). Idealmente, debe llegarse a tales niveles dentro de un período suficiente para permitir que los ecosistemas se adapten naturalmente al cambio climático, que no se

amenace la producción de alimentos y que se permita el desarrollo económico de manera sustentable.

Entre los diferentes tipos de sumideros (reservorios, almacenes o depósitos) de carbono que pueden ser medidos se encuentra la hojarasca. La hojarasca es toda la biomasa no viva sobre el suelo (hojas, ramas y cáscaras de frutos) en diferentes estados de descomposición. Por la importancia de la hojarasca en la estabilidad y funcionamiento de los ecosistemas, al constituir el eslabón que garantiza la circulación de materias orgánicas y nutrientes entre las plantas y el suelo, es necesario profundizar en las relaciones funcionales de este reservorio de carbono orgánico.

Debido a la no existencia de información sobre secuestro de carbono en sistemas forestales en el Estado de Puebla, México, y en particular, en la zona de Teziutlán, se hizo necesario iniciar los estudios investigativos pertinentes, desarrollándose una tesis de doctorado [Castillo, 2014] que ha brindado la información, entre otros aspectos, sobre el carbono orgánico almacenado en la hojarasca en esa zona.

El establecimiento y desarrollo de modelos estadísticos que relacionen el carbono orgánico de la hojarasca con propiedades físicos-químicas de la misma, con regímenes de temperatura y precipitación y tipos de suelos de la región, se requiere de la utilización del entorno R, que es una herramienta libre, gratuita, flexible, asequible y accesible.

R es un entorno de programación, análisis estadístico y gráfico, que se inscribe dentro del programa GNU General Public Licence (Licencia Pública General, GNU), que se ha convertido en una necesidad en los tiempos actuales [R, 2007]. Los elementos anteriores nos dirigen a formular los siguientes objetivos:

Objetivo General

Profundizar en el establecimiento de modelos lineales mixtos y las suposiciones del mismo. Estudiar los métodos de estimación de los parámetros desconocidos y los criterios de bondad de ajuste, con el propósito de seleccionar buenos modelos sobre el carbono orgánico en la hojarasca en regiones del estado de Puebla, México.

Objetivos Específicos

1. Reafirmar los conceptos del Modelo Lineal General.
2. Describir en qué consiste el Modelo Lineal Mixto y sus aspectos estadísticos, así como, investigar diferentes comandos y paqueterías del entorno R desarrolladas sobre estos modelos.

3. Determinar en la Región Terrestre Prioritaria (RTP -105), Teziutlán, Puebla, el porcentaje de carbono orgánico en la hojarasca en función de diferentes propiedades físico químicas de la misma.

La estructura de los siguientes capítulos de la Tesis es: en el Capítulo 1 se presenta el Modelo Lineal General y algunos de los procedimientos asociados a estos modelos; en el Capítulo 2 se hace énfasis en los Modelos Lineales Mixtos y se exponen diferentes metodologías que esclarecen la utilización de los modelos Mixtos; en el Capítulo 3 se brindan los resultados obtenidos con los diferentes modelos considerados y se hace una valoración de la conveniencia de alguno de ellos. Finalmente, se arriba a las Conclusiones, se dan las referencias y se desarrollan dos Apéndices.

Índice general

Capítulo 1

Modelo Lineal General

Ya hemos mencionado que en las últimas décadas se han alcanzado enormes desarrollos en los resultados analíticos del Modelo Lineal General, que incluye los modelos de Regresión, de Análisis de Varianza (*ANOVA*) y de Análisis de Covarianza (*ANCOVA*). En este Capítulo se presentan los métodos de estimación y pruebas de hipótesis para estos modelos.

1.1. Modelo de Regresión Lineal

Sea Y la variable dependiente que está relacionada con las p variables independientes $X_1, X_2, \ldots, X_p$ por una función f. Tanto la variable dependiente como las independientes son continuas. No siempre esta relación es exacta por lo que se escribe de la siguiente forma

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon \tag{1.1}$$

donde ϵ es un error aleatorio.
Las Y y las X's se observan sobre n individuos. Cuando f es lineal, la ecuación (1.1) observada en un individuo, se escribe como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \tag{1.2}$$

y se llama **modelo de regresión lineal**; los parámetros $\beta_j, j = 0, 1, \ldots, p$, se llaman **coeficientes de regresión**, los cuales estamos interesados en estimar.
Escribiendo el modelo lineal (1.2) en forma matricial, tenemos que

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & \vdots & \vdots & \dots & \vdots \\ 1 & x_{i1} & x_{i2} & x_{ij} & x_{ip} \\ 1 & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$i = 1, 2, ..., n$
$j = 0, 1, ..., p$
o simplemente

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.3}$$

donde
$\mathbf{Y}$ es un vector aleatorio de dimensión $n\times$ 1,
$\mathbf{X}$ es llamada la *matriz de diseño* de $n \times (p+1)$ no aleatoria,
$\boldsymbol{\beta}$ es un vector de parámetros desconocidos $(p+1) \times 1$,
$\boldsymbol{\epsilon}$ es el vector de errores aleatorio de dimensión $n\times 1$.
Los supuestos sobre el vector de errores se expresan como

$$\begin{aligned} E(\boldsymbol{\epsilon}) &= 0, \\ Var(\boldsymbol{\epsilon}) &= \sigma^2 I. \end{aligned}$$

donde I es la matriz identidad de dimensión $(n \times n)$ (se supone que los errores tienen varianza constante y están incorrelacionadas). Obsérvese que la esperanza de $\mathbf{Y}$ es

$$E(\mathbf{Y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = E(\mathbf{X}\boldsymbol{\beta}) + E(\boldsymbol{\epsilon}) = E(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$$

y la varianza de $\mathbf{Y}$ es

$$Var(\mathbf{Y}) = Var(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = Var(\mathbf{X}\boldsymbol{\beta}) + Var(\boldsymbol{\epsilon}) = Var(\boldsymbol{\epsilon}) = \sigma^2 I.$$

La variable dependiente Y del modelo de regresión se suele nombrar como variable de respuesta y las variables independientes como explicativas o predictores.

1.2. Análisis de Varianza (*ANOVA*)

En la modelación estadística se desea conocer el efecto de una o más variables independientes sobre una respuesta (continua). Las variables independientes que pueden ser controladas en un experimento reciben el nombre de ***factores*** y el nivel de intensidad de un factor se le denomina ***nivel*** del factor. Si existe un solo factor, a los niveles de ese factor se le llaman tratamientos (T) [Crawley, 2008]. Cuando la(s) variable(s) independientes son categóricas en vez de continuas y se desea hacer comparaciones entre los niveles del factor entonces nos enfrentamos ante un caso típico de análisis de varianza ($ANOVA$).

El $ANOVA$ implica el cálculo de la variación total en la variable de respuesta y la partición de ella en componentes informativos [Crawley, 2008]. En el caso más simple, dividimos la variación total en sólo dos componentes: la variación explicada y la variación no explicada:

$$SSY \begin{cases} \nearrow SSA \\ \searrow SSE \end{cases}$$

donde SSY, SSA y SSE son la suma de cuadrados de la varianza. El Cuadro 1.1 brinda más detalle.

Analicemos los grados de libertad que por sus siglas en inglés se escribe df (degrees of freedom). Supongamos que hay m réplicas en cada tratamiento y supongamos que hay k niveles del factor. Al estimar k parámetros a partir de los datos, antes de poder calcular SSE se han perdido k grados de libertad en el proceso. Como cada uno de los k niveles del factor tiene m repeticiones, debe haber $k \times m$ números en todo el experimento. Así que los df asociados con SSE son $km - k = k(m - 1)$. Otra forma de ver esto, es decir, que hay m réplicas en cada tratamiento, y por lo tanto, $(m - 1)df$ para el error en cada tratamiento (1 df porque se pierde en la estimación de cada media del tratamiento). Hay k tratamientos (es decir, k niveles del factor) y por lo tanto, hay $k\times$(m-1) df para el error en el experimento.

El componente de la variación que se explica por las diferencias entre las medias de los tratamientos, la suma de los cuadrados de tratamiento, tradicionalmente se denota por SSA. Cuando se tiene en el análisis dos o más variables independientes categóricas diferentes, SSB se utiliza para denotar la suma de cuadrados

atribuibles a las diferencias entre las medias del segundo factor, SSC se denota la suma de cuadrados medios atribuibles al tercer factor, y así sucesivamente.

El cálculo de estas sumas de cuadrados se presenta tradicionalmente en la Tabla 1.2. Hay seis columnas que indican, de izquierda a derecha, la fuente de variación, la suma de cuadrados atribuibles a esa fuente, los grados de libertad para esa fuente, la varianza para esa fuente (tradicionalmente llamados el cuadrado medio en lugar de la varianza), la proporción F (que sirve para probar la hipótesis nula de que esta fuente de variación no es significativamente distinta de cero) y el valor p asociado con ese valor F (si $p < 0.05$ entonces rechazamos la hipótesis nula). Los cuadrados medios se obtienen simplemente dividiendo cada suma de los cuadrados de sus respectivos grados de libertad (en la misma fila). La varianza del error, s^2, es el cuadrado medio residual (el cuadrado medio de la variación no explicada); esto es a veces llamado la "varianza de error agrupado", ya que se calcula a través de todos los tratamientos; la alternativa sería tener k varianzas separadas, una para cada tratamiento.

Cuadro 1.1: **Sumas de Cuadrados Corregidas *ANOVA* unifactorial**

La definición de la suma total de cuadrados, SSY, es la suma de los cuadrados de las diferencias entre los puntos de datos, y_{ij}, y la media general, $\bar{\bar{y}}$.

$$SSY = \sum_{i=1}^{k} \sum_{j=1}^{m} (y_{ij} - \bar{\bar{y}})^2$$

donde $\sum_{j=1}^{m} y_{ij}$ significa la suma sobre las m réplicas dentro de cada uno de los k niveles de factor . La suma de cuadrados del error, SSE, es la suma de los cuadrados de las diferencias entre los puntos de datos, y_{ij}, y sus medias de los tratamiento individuales, $\bar{y}_i$

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2.$$

La suma de los cuadrados de tratamiento, SSA, es la suma de los cuadrados de las diferencias entre el tratamiento individual, $\bar{y}_i$ y la media general, $\bar{\bar{y}}$

$$SSA = \sum_{i=1}^{k} \sum_{j=1}^{m} (\bar{y}_i - \bar{\bar{y}}_i)^2 = m \sum_{i=1}^{k} (\bar{y}_i - \bar{\bar{y}})^2.$$

Elevando el término al cuadrado en paréntesis y aplicando la suma nos da

$$m \sum_{i=1}^{k} (\bar{y}_i - \bar{\bar{y}}_i)^2 = \sum \bar{y}_i{}^2 - 2\bar{\bar{y}} \sum \bar{y}_i + k\bar{\bar{y}}^2.$$

Denotando el total de todos los valores de la variable de respuesta $\sum_{i=1}^{k} \sum_{j=1}^{m} y_{ij}$ como $\sum y$. Ahora reemplazamos $\bar{y}_i$ por T_i/m (donde T es el nombre convencional para los k tratamientos totales individuales) y reemplazando $\bar{\bar{y}}$ por $\sum y/km$ se obtiene

$$\frac{\sum_{i=1}^{k} T_i^2}{m^2} - 2\frac{\sum y \sum_{i=1}^{k} T_i}{mkm} + k\frac{\sum y \sum y}{kmkm}.$$

Note que $\sum_{i=1}^{k} T_i = \sum_{j=1}^{m} y_{ij}$, Por lo que los términos de la derecha positivos y negativos ambos tienen la forma $(\sum y)^2/km^2$. Finalmente, multiplicando por m da

$$SSA = \frac{\sum_{i=1}^{k} T_i^2}{m} - \frac{(\sum y)^2}{km}.$$

Se puede probar que $SSY = SSA + SSE$.

[Crawley, 2008]

Cuadro 1.2: ANOVA

Fuente de Variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	p valor $Pr(>F)$
Tratamiento	SSA	k	SSA/k	$\frac{SSA/k}{SSE/m-k-1}$	p<0.05
Error	SSE	$m-k-1$	$SSE/m-k-1$		
Total	SSY	$m-1$			

[Montgomery, 2011]

1.3. Análisis de Covarianza (*ANCOVA*)

El análisis de covarianza (*ANCOVA*) combina elementos de la regresión y el análisis de varianza. La variable dependiente es continua, y hay al menos una variable independiente continua (covariables) y con al menos una variable independiente categórica. Según [Crawley, 2008] el *ANCOVA* se resumen con las siguientes relaciones:

- Hacer una regresión lineal de $\boldsymbol{Y}$ contra $\boldsymbol{X}$ para cada nivel del factor.
- Estimar diferentes pendientes e interceptos para cada nivel.
- Usar la simplificación del modelo (las pruebas de eliminación) para los parámetros innecesarios.

Por ejemplo, supongamos que estamos modelando peso (variable dependiente) en función del sexo y la edad. El sexo es un factor con dos niveles (hombres y mujeres) y la edad es una variable continua. Por tanto, el modelo máximo tiene cuatro parámetros: dos pendientes (una pendiente para los hombres y una pendiente para las mujeres, es decir, β_1 y β_2) y dos interceptos (una para hombres y otra para mujeres).

La simplificación del modelo es una parte esencial del análisis de covarianza, porque el principio de la parsimonia requiere que mantengamos el menor número de parámetros posibles en el modelo.
Hay seis modelos posibles en este caso, y el proceso de simplificación del modelo comienza por preguntar si necesitamos todos los cuatro parámetros. Quizás podríamos conformarnos con los interceptos y una pendiente común, o un intercepto en común y dos pendientes diferentes. Una vez más, la edad puede tener

un efecto significativo en la respuesta, por lo que sólo necesita dos parámetros para describir los principales efectos del sexo en el peso; esto aparecería como dos líneas separadas. Alternativamente, puede haber ningún efecto del sexo en absoluto, en cuyo caso sólo tenemos dos parámetros (una pendiente y un intercepto) para describir el efecto de la edad sobre el peso.

1.4. Métodos de Estimación

A continuación se explican distintos métodos de estimación que ayudan a estimar los parámetros $\boldsymbol{\beta}$ y σ^2. Inicialmente se explica el método de Mínimos Cuadrados Ordinarios. Posteriormente se introducen los métodos de Máxima Verosimilitud y Máxima Verosimilitud Restringida.

1.4.1. Mínimos Cuadrados Ordinarios

Los valores de $\boldsymbol{\beta}$ son desconocidos, pero se pueden estimar, utilizando los datos de la muestra. Para estimarlos se usa el **"método de mínimos cuadrados"** [Montgomery, 2011], que en inglés es Ordinary Least Squares (OLS).
Cuando la matriz $\mathbf{X}$ tiene rango completo p, el estimador OLS se obtiene minimizando la suma de cuadrados de los residuos, donde el i-ésimo residual es la diferencia entre el valor observado y_i y el ajustado $\hat{y}_i$.
Los residuos o residuales se pueden escribir en forma matricial como sigue:

$$\mathbf{e}_{(n\times 1)} = \mathbf{Y} - \hat{\mathbf{Y}}. \tag{1.4}$$

La suma de cuadrados de los residuos es:

$$\begin{aligned} S(\hat{\boldsymbol{\beta}}) &= \sum_{i=1}^{n} e_i^2 \\ &= \boldsymbol{e}'\boldsymbol{e} \\ &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y}'\mathbf{Y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}. \end{aligned}$$

Observemos que $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$ es una matriz de 1×1, es decir, un escalar, y que su transpuesta $(\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}$, es el mismo escalar. Los estimadores de mínimos cuadrados deben satisfacer

$$\frac{\partial S}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0, \tag{1.5}$$

que se simplifica en

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}. \tag{1.6}$$

El sistema lineal de (1.6) se denomina **ecuaciones normales de mínimos cuadrados** [Rao, 2007]. Para resolver las ecuaciones normales se multiplican ambos lados de (1.6) por la inversa de $\mathbf{X}'\mathbf{X}$. El estimador $\hat{\boldsymbol{\beta}}$ por mínimos cuadrados es un vector de dimensión $p \times 1$ cuya expresión es

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{1.7}$$

Los estimadores mínimos cuadrados son insesgados y tienen matriz de varianza y covarianza $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_{OLS}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned}$$

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}_{OLS}) &= Cov[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Cov[\mathbf{Y}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2 I(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Se puede demostrar que el estimador *OLS* es el mejor estimador lineal insesgado, que en inglés se escribe Best Linear Unbiased Estimaton (*BLUE*) [Rao, 2007] de los parámetros del modelo (1.3). Esto significa que, de entre los estimadores que son insesgados y lineales con respecto a las observaciones, el estimador *OLS* tiene la menor varianza (teorema Gauss-Markov , [Rao, 2007]). Sin embargo, esto es sólo cuando los supuestos (varianza constante y no correlación) en los residuos se mantengan. Además, para la estimación de $\boldsymbol{\beta}$, no se necesita la condición de que la varianza sea σ^2; es suficiente que la varianza sea constante para todas las observaciones.

1.4.2. Máxima Verosimilitud

El método de máxima verosimilitud, que en inglés se escribe Maximum Likelihood (ML) es un método alternativo para estimar los parámetros en (1.3), suponiendo que los errores son independientes e idénticamente distribuidos según la normal con varianza constante igual a σ^2, esto es ($N(0, \sigma^2 I)$).
El método ML inicia con la función de densidad de los errores

$$f(\epsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\epsilon_i^2}, \qquad i = 1, \ldots, n.$$

La función de verosimilitud es:

$$L(\epsilon, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f(\epsilon_i) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\epsilon_i'\epsilon}, \tag{1.8}$$

de (1.3) tenemos que $\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$, así (1.8) se transforma en

$$L(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}. \tag{1.9}$$

Ahora la función log-verosimilitud es:

$$l = \log L(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{\dot{n}}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{1.10}$$

Para un valor fijo de σ, la función log−verosimilitud se maximiza cuando se minimiza el término $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$.
Por lo tanto, el estimador de máxima verosimilitud de $\boldsymbol{\beta}$ bajo los errores normales equivale al estimador de mínimos cuadrados $\hat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ [Rao, 2007] y el estimador de máxima verosimilitud de σ^2 es

$$\hat{\sigma}^2_{ML} = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n}. \tag{1.11}$$

1.4.3. Máxima Verosimilitud Restringida (*REML*)

El método de Máxima Verosimilitud Restringida, que en inglés se escribe Restricted maximum Likelihood ($REML$) soluciona el problema del sesgo en la estimación de σ^2.

La idea de *REML* es aplicar el método de máxima verosimilitud a un vector $K'\boldsymbol{Y}$ en vez del vector de observaciones originales $\mathbf{Y}$. La matriz K se define de manera que se elimina del vector $\mathbf{Y}$ toda la variación que se explica por la matriz $\mathbf{X}$ del modelo. Una diferencia importante entre los vectores $\mathbf{Y}$ y $K'\mathbf{Y}$ es que la longitud de $K'\mathbf{Y}$ es $n-p$ [Mehtätalo, 2013]. Por lo tanto, un ajuste ML de un modelo lineal de n observaciones ofrece un estimador de la varianza residual con n en el denominador, mientras que el estimador correspondiente para el vector $K'\boldsymbol{Y}$ ofrece un estimador con $(n-p)$ en el denominador.
Surge una cuestión de cómo encontrar una matriz K tal que elimine toda esa variación de $\mathbf{Y}$ que puede ser explicada por $\mathbf{X}$. La condición clave para la eliminación de toda la variación explicada por $\mathbf{X}$ es definir cada columna de la matriz K, denotada por $k_1, ..., k_{n-p}$, tal que $k_i'\mathbf{X} = 0$ para $i = 1, 2, 3, \ldots, n-p$. Hay un resultado matricial (Searle 1982, sección 9.7 a), que establece que el número máximo de columnas linealmente independientes que cumplan la condición anterior es $n-p$, esto es la diferencia entre el número de filas y columnas de la matriz del modelo (la matriz $\mathbf{X}$ tiene rango completo). Una forma de encontrar al vector k es utilizar [McCulloch, 2001],

$$k' = c'[I - XX^{-}] \tag{1.12}$$

donde c' es un vector arbitrario de longitud n y X^{-} es una inversa generalizada de la matriz $\mathbf{X}$.
Hay que tener en cuenta que puede haber varios valores posibles de X^{-}, (debido a que un sistema de ecuaciones con n variables y p ecuaciones puede tener muchas soluciones). Sin embargo, los estimadores finales $REML$ no se ven afectados por la elección de X^{-}.
Una vez que la matriz K se encuentra, vemos que multiplicando el modelo

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

por K', por la izquierda

$$K'\boldsymbol{Y} = K'\boldsymbol{X\beta} + K'\boldsymbol{\epsilon}. \tag{1.13}$$

Por construcción de la matriz K, es decir, $K'\mathbf{X} = 0$, obtenemos

$$K'\mathbf{Y} = K'\boldsymbol{\epsilon}.$$

Si $Var(\boldsymbol{\epsilon}) = V$, entonces $Var(K'\boldsymbol{\epsilon}) = K'VK$.
En forma más general, si $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, V)$,
entonces

$$K'\mathbf{Y} \sim N(0, K'VK). \tag{1.14}$$

Por lo tanto, el modelo se puede ajustar utilizando una verosimilitud restringida basado en la normalidad de $K'\boldsymbol{Y}$. Una diferencia esencial entre estas dos verosimilitudes, además de las longitudes de $\boldsymbol{Y}$ y $K'\boldsymbol{Y}$, es que la verosimilitud *REML* no involucra a $\boldsymbol{X\beta}$. Por lo tanto, REML se puede utilizar solo para la estimación de parámetros relacionados con V. Así

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{V}\mathbf{X})^{-1}\mathbf{X}'\hat{V}\mathbf{Y}. \tag{1.15}$$

Para el modelo (1.14), la función de verosimilitud se transforma en

$$L(\sigma^2) = \frac{1}{\sigma^n(2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}. \tag{1.16}$$

Ahora la función log-verosimilitud es:

$$\begin{aligned} l = \log L(\sigma^2) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log K'VK \\ &\quad -\frac{1}{2K'VK}(K'\boldsymbol{Y} - 0)'(K'\boldsymbol{Y} - 0). \end{aligned}$$

tomando $V = \sigma^2 I$, la función log-verosimilitud se transforma en :

$$\begin{aligned} \log L(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log \sigma^{2(n-p)}|K'K| \\ &\quad -\frac{1}{2\sigma^2}\boldsymbol{Y}'K(K'K)^{-1}K'\boldsymbol{Y}. \end{aligned}$$

Para estimar σ^2, diferenciamos el log- verosimilitud con respecto a σ^2

$$\frac{\partial l}{\partial \sigma^2} = \frac{n-p}{2\sigma^2} + \frac{1}{2\sigma^4}\boldsymbol{Y}'K(K'K)^{-1}K'\boldsymbol{Y}, \tag{1.17}$$

igualando a cero y despejando σ^2. Así tenemos que el estimador *REML* de σ^2 es:

$$\hat{\sigma}^2_{REML} = \frac{1}{n-p}\boldsymbol{Y}'K(K'K)^{-1}K'\boldsymbol{Y}. \tag{1.18}$$

1.5. Bondad de Ajuste del Modelo

En este epígrafe se presentan varias maneras de evaluar la bondad de ajuste del modelo lineal general, lo que se refiere al grado en que esté modelo se conveniente debido a que representa a las variables implicadas en el modelo.

1.5.1. Prueba F para el Ajuste del Modelo

La prueba de significancia del modelo nos permite determinar estadísticamente si las variables independientes (en conjunto) tienen efecto o no sobre la variable dependiente. Este procedimiento suele considerarse como una prueba general o global del ajuste del modelo. Las hipótesis son:

$$H_0 : \beta_0 = \beta_1 = \ldots = \beta_p = 0 \quad \text{vs} \quad H_a : \beta_j \neq 0 \quad \text{al menos para un } j = 1, \ldots, p.$$

El rechazo de la hipótesis nula implica que al menos una de las variables independientes contribuye al modelo en forma significativa.
Aunque en el epígrafe 1.2 desarrollamos como elaborar una tabla ANOVA que muestra el cálculo del estadístico F para esta prueba, creemos conveniente repetirlo con la notación comúnmente explicada en los modelos de regresión.

Este método consiste en una partición de la variabilidad total de la variable y de respuesta. Para obtener esta partición se comienza con la identidad

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i). \tag{1.19}$$

Ahora procedemos a elevar al cuadrado en ambos lados de la ecuación (1.19), y se suman para todas las n observaciones. Así se obtiene

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\begin{aligned} \sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) &= 2\sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i) - 2\sum_{i=1}^{n}\bar{y}_i(y_i - \hat{y}_i) && (1.20) \\ &= 2\sum_{i=1}^{n}\hat{y}_i e_i - 2\bar{y}_i\sum_{i=1}^{n}e_i = 0. && (1.21) \end{aligned}$$

Ya que la suma de los residuales siempre es igual a cero y la suma de los residuales ponderados por el valor ajustado correspondiente también es igual a cero [Montgomery, 2011]. Por lo anterior,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{1.22}$$

Lo anterior nos dice que la **suma total de cuadrados** SS_T se divide en una **suma de cuadrados debidos a la regresión**, SS_R, y una **suma de cuadrados de residuales**, SS_{Res}.

$$SS_T = SS_R + SS_{Res}, \tag{1.23}$$

en [Montgomery, 2011] se demuestra que $\frac{SS_R}{\sigma_2}$ tiene una distribución χ^2_p, con el mismo número de grados de libertad que la cantidad de variables regresoras, $\frac{SS_{Res}}{\sigma_2} \sim \chi^2_{n-p-1}$ y que SS_R y SS_{Res} son independientes.
Tomando a

$$F_0 = \frac{SS_R/p}{SS_{Res}/(n-p-1)} = \frac{MS_R}{MS_{Res}} \tag{1.24}$$

este tiene una distribución $F_{p,n-p-1}$ y rechazamos H_0 si

$$F_0 > F_{\alpha,p,n-p-1}. \tag{1.25}$$

El procedimiento se resume en el Cuadro (1.3) de análisis de varianza de la regresión.

Cuadro 1.3: Análisis de varianza de la regresión

Fuente de Variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_O	P valor $0 < P < 1$
Regresión	SS_R	p	MS_R	$\frac{MS_R}{MS_{Res}}$	$P < \alpha$
Residuales	SS_{Res}	$n-p-1$	MS_{Res}		
Total	SS_T	$n-1$			

[Montgomery, 2011]

1.5.2. Coeficiente de Determinación (R^2) y Ajustado (R_{Adj})

El coeficiente de determinación nos permite expresar la cantidad de la variabilidad presente en las observaciones de $\mathbf{Y}$, que se explica mediante el modelo de regresión lineal múltiple, cuando se utilizan las variables independientes, en conjunto, como variables regresoras. La cantidad

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}, \tag{1.26}$$

se llama **coeficiente de determinación**. Nótese que R^2 indica la proporción de variabilidad explicada por la regresión. Ya que $0 \leq SS_{Res} \leq SS_T$, entonces $0 \leq R^2 \leq 1$. Los valores de R^2 cercanos a 1 implican que la mayor parte de la variabilidad de $\mathbf{Y}$ esta explicada por el modelo de regresión. A medida que el coeficiente se aproxime a cero el modelo deja de ser adecuado, ya que la cantidad de la variabilidad explicada mediante el modelo es pobre [Montgomery, 2011].
En general, R^2 aumenta siempre que se agrega un regresor al modelo, independientemente del valor de la contribución de esa variable. En consecuencia, es difícil juzgar si un aumento de R^2 dice en realidad algo importante.
Algunos investigadores que trabajan con modelos de regresión prefieren usar el estadístico R^2_{Adj} , que se define como sigue:

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-(k+1))}{SS_T/(n-1)}. \tag{1.27}$$

En vista de que $\frac{SS_{Res}}{(n-(k+1))}$ es el cuadrado medio de los residuales, $\frac{SS_T}{(n-1)}$ es constante e independiente de cuántas variables hay en el modelo, R^2_{Adj} sólo aumentará al agregar una variable al modelo si esa variable reduce el cuadrado medio residual.

El R^2_{Adj} penaliza el aumento de términos que no son útiles, además sirve como procedimiento para evaluar y comparar los posibles modelos de regresión.

1.5.3. Prueba t de Student sobre Coeficientes Individuales

Una vez determinado que al menos una de las variables independientes es importante, la siguiente pregunta es: ¿Cuál(es) variable(s) es (son) importante(s)? Si agregamos una variable al modelo de regresión, la suma de cuadrados de la regresión aumenta y la suma de cuadrados residuales disminuye. Se debe decidir si el aumento de la suma de cuadrados de la regresión es suficiente para garantizar el uso del regresor adicional en el modelo.

La adición de un regresor también aumenta la varianza del valor ajustado $\hat{\mathbf{Y}}$, por lo que se debe tener cuidado de incluir sólo variables independientes que tengan valor para explicar la respuesta [Montgomery, 2011]. Además, si agregamos una variable independiente que no es importante se puede aumentar el cuadrado medio de residuales y con eso disminuye la utilidad del modelo.

Las hipótesis para probar la significancia de cualquier coeficiente $\beta_j, j = 1, 2, \ldots, p$, está dado por:

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_a : \beta_j \neq 0 \quad \text{para} \quad j = 1, 2, \ldots, p.$$

El estadístico de prueba para esta hipótesis es,

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\sigma^2 C_{jj}}}, \tag{1.28}$$

donde C_{jj} es el elemento diagonal de $(X'X)^{-1}$ que corresponde a $\hat{\beta}_j$. Se rechaza la hipótesis nula $H_0 : \beta_j = 0$ si

$$|t_0| > t_{\alpha/2, n-p-1}.$$

Notemos que ésta es una prueba parcial, porque el coeficiente de regresión $\hat{\beta}_j$ depende de todas las demás variables regresoras x_j, que hay en el modelo. Esto es una prueba de la contribución de x_j dadas las demás variables del modelo.
En general, el cuadrado de una variable aleatoria t con f grados de libertad es una variable aleatoria F con 1 y f *df* en el numerador y en el denominador, respectivamente. Aunque la prueba t para $H_0 : \beta_1 = 0$ equivale a la prueba F en la regresión lineal simple, la prueba t es algo más adaptable, porque se podría usar para probar hipótesis alternativas unilaterales (Sea $H_1 : \beta_1 < 0$ o $H_1 : \beta_1 > 0$), mientras que la prueba F sólo considera la alternativa bilateral [Montgomery, 2011].

1.6. Suposiciones del Modelo: su Diagnóstico

Las principales suposiciones sobre el modelo lineal general son las siguientes:

- $\mathbf{Y}$ está relacionada con $\boldsymbol{X}$ mediante una función lineal, donde $\boldsymbol{\beta}$ es el vector de parámetros desconocidos que determinan el modelo:

$$E(\mathbf{Y}|\boldsymbol{X} = X_j) = \beta_0 + \beta X_j, \quad j = 1, 2, \ldots, p. \tag{1.29}$$

- Las variables regresoras $X_1, X_2, \ldots, X_p$ se consideran fijas, i.e., no son variables aleatorias.
- $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ son variable aleatorias no observables, distribuidas normalmente con $\mu = 0$ y varianza constante σ^2. Esto se puede escribir como:

$$\epsilon_i \sim N(0, \sigma^2), i = 1, 2, \ldots, n. \tag{1.30}$$

Las grandes violaciones a las suposiciones pueden dar como resultado un modelo inestable, en el sentido que una muestra distinta puede conducir a un modelo totalmente diferente y por tanto, obtener conclusiones opuestas. En general, no se pueden detectar desviaciones respecto a las premisas básicas examinando los estadísticos estándar de resumen (t, F o R^2). Éstas propiedades son globales del modelo y como tal no aseguran la adecuación del mismo.
A continuación se exponen algunos métodos para checar las suposiciones.

1.6.1. Normalidad, Homogeneidad de Varianza e Independencia

En este epígrafe se presentarán algunos métodos para diagnosticar violaciones a las suposiciones básicas del modelo. Estos métodos se basan principalmente en el estudio de los residuos del modelo.
Ya vimos antes que los residuos (o residuales) se definen como:

$$e_i = y_i - \hat{y}_i, \quad i = 1, ..., n. \ , \tag{1.31}$$

siendo y_i una observación y $\hat{y}_i$ su valor ajustado correspondiente. Podemos considerar que un residual es la desviación entre los datos y el ajuste, pero también es una medida de la variabilidad de la variable de respuesta que no explica el modelo de regresión. También es conveniente imaginar que los residuales son los valores realizados (observados), de los errores del modelo (no observados), por lo que toda desviación de las suposiciones sobre los errores se debe reflejar en los residuales [Montgomery, 2011].
Los residuales tienen varias propiedades importantes. Tienen media cero y su varianza promedio se estima con

$$\frac{\sum_{i=1}^{n}(e_i - \overline{e})^2}{n-(p+1)} = \frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)} = \frac{SS_{Res}}{n-(p+1)} = MS_{Res}. \tag{1.32}$$

Sin embargo, los residuales no son independientes, ya que los n residuales sólo tienen $n-(p+1)$ grados de libertad asociados a ellos. Se ha observado que cuando no hay independencia en los residuales se tiene poco efecto para comprobar la

adecuación del modelo, siempre y cuando, n no sea pequeña en relación con la cantidad de parámetros.

El análisis de residuales es una forma muy eficaz de descubrir diversos tipos de inadecuación del modelo. Como veremos, el análisis gráfico de los residuales es una forma muy efectiva de investigar la adecuación del ajuste de un modelo de regresión y para comprobar las suposiciones básicas.

Gráfica de Probabilidad Normal

Las pequeñas desviaciones respecto a la hipótesis de normalidad no afectan mucho al modelo, pero tener desviaciones grandes de no normalidad es potencialmente peligroso, porque los estadísticos t o F dependen de la suposición de normalidad. Además, si los errores provienen de una distribución con colas más gruesas (cuando la frecuencia de ocurrencia de eventos que están situados en los extremos de la distribución no es muy baja) que la normal, el ajuste por mínimos cuadrados será sensible a un subconjunto menor de datos. Las distribuciones de los errores con colas gruesas generan con frecuencia valores atípicos que jalan demasiado en su dirección el ajuste por mínimos cuadrados. En esos casos se deben considerar otras técnicas de estimación [Montgomery, 2011].

Un método muy sencillo de comprobar la suposición de **normalidad** es trazar una gráfica de **probabilidad normal** de los residuales. Esta es una gráfica diseñada para que se dibuje una línea recta, que representa a una normal acumulada. Sea $e_1 < e_2 < \cdots < e_n$ los residuales ordenados en orden creciente. Si se grafica e_i en función de la probabilidad acumulada $P_i = (i - \frac{1}{2})/1, i =, 2, \ldots, n$, los puntos que resulten deberían estar aproximadamente sobre una línea recta.

La recta se suele determinar en forma visual, con énfasis en los valores centrales (por ejemplo, los puntos de probabilidad acumulada 0.33 y 0.67) y no en los extremos. Las diferencias apreciables en distancia respecto a la recta indican que la distribución no es normal.

A veces, las gráficas de probabilidad normal se trazan graficando el residual clasificando $\boldsymbol{e}_i$ en función del valor normal esperado, $\phi^{-1}[(i - \frac{1}{2})/n]$, donde ϕ representa la distribución normal estándar acumulada. Esto es consecuencia de $E(e_i) \simeq \phi^{-1}[(i - \frac{1}{2})/n]$ [Montgomery, 2011].

El estudio de las gráficas ayuda a adquirir un grado de percepción de cuánta desviación de la recta es aceptable. Con frecuencia, los tamaños pequeños de

muestra ($n \leq 16$) producen gráficas de probabilidad normal que se desvían bastante de la linealidad. Para muestras mayores ($n \geq 32$), las gráficas se comportan mucho mejor. Por lo general, se requieren alrededor de 20 puntos para producir gráficas de probabilidad suficientemente estables como para poder interpretarse con facilidad.

Gráfica de Residuales en Función de los Valores Ajustados $\hat{y}_i$

Para poder detectar algunas inadecuaciones en nuestro modelo de regresión, es útil tener una gráfica de los residuales en función de los valores ajustados correspondientes $\hat{y}_i$. Está gráfica permite detectar diferentes problemas, tales como:

- **Heterocedasticidad**, la varianza no es constante y se deben de transformar los datos (la variable $\boldsymbol{Y}$) o aplicar otros métodos de estimación.
- **Error en el análisis**, se ha realizado mal el ajuste y se verifica que los residuos negativos se corresponden con los valores pequeños $\hat{y}_i$ y los errores positivos se corresponden con los valores grandes de y_i, o al revés.
- El modelo es inadecuado por **falta de linealidad** (no lineal) y se deben transformar los datos o introducir nuevas variables que pueden ser cuadrados de las existentes o productos de las mismas, o bien se deben introducir nuevas variables explicativas.
- Existencia de **observaciones atípicas** o puntos extremos.
- **Falta de independencia**, los residuales se presentan formando grafos (clusters).

Capítulo 2

Modelos Lineales Mixtos

Entre los años 1920 y 1930, cuando R Fisher ideó el Análisis de Varianza tenía en mente tomar a todas las variables independientes categóricas como si fueran lo mismo [Crawley, 2008]. Fue Eisenhart en 1947 que se dio cuenta de que había en realidad dos tipos diferentes de variables independientes categóricas y las llamó efectos fijos y efectos aleatorios. Se debe tener en cuenta que los efectos fijos sólo influyen en la media de la variable respuesta Y del modelo, mientras que los efectos aleatorios influyen sólo en la varianza de Y (se demostrará más adelante). A los modelos que tienen estos dos tipos de efectos, se les llama *modelos de efectos mixtos*.

Un efecto aleatorio debe ser considerado como proveniente de una población de efectos: la existencia de esta población es una suposición adicional. Se estiman los efectos fijos a partir de los datos, pero se tiene la intención de hacer predicciones sobre la población de la que los efectos fijos fueron muestreados.

A continuación desarrollamos los aspectos fundamentales de estos modelos que denominaremos LMM por sus siglas en inglés (Lineal Mixed Models).

2.1. Estructura Matricial para la $i-$ ésima Observación

Ahora consideremos la especificación general matricial de un LMM para la $i-$ésima observación:

$$Y_i = X_i\beta + Z_i u_i + \varepsilon_i. \tag{2.1}$$

En (2.1) Y_i representa un vector de respuesta continua para la $i-$ ésima observación, siguiendo la notación de [West, 2007]. Los elementos del vector Y_i son:

$$Y_i = \begin{pmatrix} Y_{1i} \\ Y_{2i} \\ \vdots \\ Y_{n_i i} \end{pmatrix}.$$

Notemos que el número n_i en el vector Y_i puede variar de una observación a otra. La matriz de diseño X_i en la ecuación (2.1) es una matriz de dimensión $(n_i \times (p+1))$, la cual representa los valores conocidos de las p variables $X^{(1)}, \ldots, X^{(p)}$, para cada uno de los n_i elementos recogidos en la $i-$ ésima observación:

$$X_i = \begin{pmatrix} 1 & X_{1i}^{(1)} & X_{1i}^{(2)} & \ldots & X_{1i}^{(p)} \\ 1 & X_{2i}^{(1)} & X_{2i}^{(2)} & \ldots & X_{2i}^{(p)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n_i i}^{(1)} & X_{n_i i}^{(2)} & \ldots & X_{n_i i}^{(p)} \end{pmatrix}.$$

Suponemos que las matrices X_i son de rango completo, es decir, ninguna de las columnas (o filas) es una combinación lineal de las restantes.

El vector β en la ecuación (2.1) es un vector de $p+1$ coeficientes de regresión desconocidos (o parámetros de efectos fijos) asociado con las p variables en la construcción de la matriz:

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

La matriz Z_i de tamaño $n_i \times q$ en la ecuación (2.1) es una matriz de diseño que representa los valores conocidos de las q covariables, $Z^{(1)}, \ldots, Z^{(q)}$, para la $i-$ ésima observación. Esta matriz es muy parecida a la matriz X_i ya que representa los valores observados de variables; sin embargo, por lo general tiene menos columnas que la matriz X_i:

$$Z_i = \begin{pmatrix} Z_{1i}^{(1)} & Z_{1i}^{(2)} & \ldots & Z_{1i}^{(q)} \\ Z_{2i}^{(1)} & Z_{2i}^{(2)} & \ldots & Z_{2i}^{(q)} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n_i i}^{(1)} & Z_{n_i i}^{(2)} & \ldots & Z_{n_i i}^{(q)} \end{pmatrix}.$$

En muchos casos, las variables independientes con efectos que varían aleatoriamente entre los sujetos, están representados tanto en la matriz X_i y la matriz Z_i [West, 2007].

El vector u_i para la $i-$ ésima observación en la ecuación (2.1) representa un vector de q efectos aleatorios asociados con las q covariables en la matriz Z_i.

$$u_i = \begin{pmatrix} u_{1i} \\ u_{2i} \\ \vdots \\ u_{qi} \end{pmatrix}.$$

Se asume que los q efectos aleatorios en el vector u_i siguen una distribución normal multivariada, con vector de media 0 y una matriz de varianzas-covarianzas denotada por D, es decir:

$$u_i \sim N(0, D). \tag{2.2}$$

Los elementos a lo largo de la diagonal principal de la matriz D representan las varianzas de cada efecto aleatorio en u_i, y los elementos fuera de la diagonal representa las covarianzas entre los efectos aleatorios correspondientes. Debido a que hay q efectos aleatorios en el modelo asociado con el $i-$ ésimo elemento, D es una matriz de $q \times q$ simétrica y definida positiva (su determinante es positivo). Los elementos de esta matriz se muestran de la siguiente manera:

$$D = Var(u_i) = \begin{pmatrix} Var(u_{1i}) & Cov(u_{1i}, u_{2i}) & \dots & Cov(u_{1i}, u_{qi}) \\ Cov(u_{1i}, u_{2i}) & Var(u_{2i}) & \dots & Cov(u_{2i}, u_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(u_{1i}, u_{qi}) & Cov(u_{1i}, u_{qi}) & \dots & Var(u_{qi}) \end{pmatrix}.$$

Finalmente, los vectores ε_i en la ecuación (2.1) es un vector de n_i errores, donde cada elemento de ε_i denota el error asociado con una respuesta observada para la $i-$ ésima observación.

$$\varepsilon_i = \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{n_i i} \end{pmatrix}.$$

En contraste al modelo lineal, los errores asociados con observaciones repetidas en el mismo sujeto en un LMM pueden estar correlacionadas [West, 2007]. Asumimos

que los n_i errores en el vector ε_i para la $i-$ ésima observación son variables aleatorias que siguen una distribución normal multivariada con un vector de media cero y una matriz de covarianza simétrica definida positiva R_i:

$$\varepsilon_i \sim N(0, R_i). \tag{2.3}$$

También se asume que los errores asociados con diferentes observaciones son independientes uno de otro. Además, se asume que los vectores de los errores $\varepsilon_1, \cdots, \varepsilon_n$, y los efectos aleatorios $u_1, \cdots, u_m$ son independientes uno del otro. Representamos la forma general de la matriz R_i como se muestra a continuación:

$$R_i = Var(\varepsilon_i) = \begin{pmatrix} Var(\varepsilon_{1i}) & Cov(\varepsilon_{1i}, \varepsilon_{2i}) & \dots & Cov(\varepsilon_{1i}, \varepsilon_{n_i i}) \\ Cov(\varepsilon_{1i}, \varepsilon_{2i}) & Var(\varepsilon_{2i}) & \dots & Cov(\varepsilon_{2i}, \varepsilon_{n_i i}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\varepsilon_{1i}, \varepsilon_{n_i i}) & Cov(\varepsilon_{2i}, \varepsilon_{n_i i}) & \dots & Var(\varepsilon_{n_i i}) \end{pmatrix}.$$

2.1.1. Estructura de Covarianzas para la Matriz D

La matriz D que corresponde a la matriz de varianzas-covarianza de los efectos aleatorios **u** (que se verá en la sección 2.2), se conoce como matriz ***no estructurada***. La simetría en la matriz D $(q \times q)$ implica que el vector θ_D tiene $(q \times (q+1))/2$ parámetros [West, 2007].
La matriz siguiente es un ejemplo de una matriz D no estructurada, en el caso de un LMM debe tener dos efectos aleatorios asociados con la $i-$ésima observación.

$$D = Var(u_i) = \begin{pmatrix} \sigma^2_{u1} & \sigma_{u1,u2} \\ \sigma_{u1,u2} & \sigma^2_{u2} \end{pmatrix}.$$

En este caso, definimos un vector θ_D, el cual contiene tres parámetros de covarianza:

$$\theta_D = \begin{pmatrix} \sigma^2_{u1} \\ \sigma_{u1,u2} \\ \sigma^2_{u2} \end{pmatrix}.$$

Una estructura comúnmente utilizada es el de componentes de la varianza, en la que cada efecto aleatorio u_i tiene su propia varianza y todas las covarianzas en D se definen como 0. En general, el vector θ_D requiere q parámetros de covarianza, donde sus elementos corresponden a la diagonal de la matriz D. Por ejemplo, en un *LMM* que tiene dos efectos aleatorios asociados con la $i-$ ésima observación, una matriz D de componentes de la varianza tiene la siguiente forma:

$$D = Var(u_i) = \begin{pmatrix} \sigma^2_{u1} & 0 \\ 0 & \sigma^2_{u2} \end{pmatrix}.$$

En este caso, el vector θ_D contiene dos parámetros:

$$\theta_D = \begin{pmatrix} \sigma^2_{u1} \\ \sigma^2_{u2} \end{pmatrix}.$$

2.1.2. Estructura de Covarianzas para la Matriz R_i

En esta sección, se discuten algunas de las estructuras más utilizadas en [West, 2007] para la matriz de covarianza R_i. La matriz de covarianza más simple para R_i es la **estructura diagonal**, en la que se supone que los errores asociados a las observaciones sobre el mismo sujeto se asumen que están correlacionadas y tienen igualdad de varianzas. La matriz diagonal R_i para la $i-$ ésima observación tiene la siguiente estructura:

$$R_i = Var(\varepsilon_i) = \sigma^2 I = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}.$$

La estructura diagonal requiere un parámetro en θ_R, que define la varianza constante :

$$\theta_R = \left(\sigma^2\right).$$

Otra estructura de R_i es la **simetría compuesta**, cuya forma general para la $i-$ ésima observación es la siguiente:

$$R_i = Var(\varepsilon_i) = \begin{pmatrix} \sigma^2 + \sigma_1 & \sigma_1 & \dots & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1 & \dots & \sigma_1 \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_1 & \sigma_1 & \dots & \sigma^2 + \sigma_1 \end{pmatrix}.$$

En la estructura de covarianza **simetría compuesta**, hay dos parámetros en el vector R que definen las varianzas y covarianzas en la matriz R_i:

$$\theta_R = \begin{pmatrix} \sigma^2 \\ \sigma_1 \end{pmatrix}.$$

Nota: los n_i errores asociados con los valores de respuesta observado para la $i-$ ésima observación se supone que tienen una covarianza constante, σ_1, y una varianza constante, $\sigma^2 + \sigma_1$, en la estructura de simetría compuesta. Esta estructura

se utiliza a menudo cuando un supuesto de igualdad de correlación de los errores es plausible (por ejemplo, los ensayos repetidos en las mismas condiciones en un experimento).

2.2. Estructura Matricial General

Una especificación alternativa, basada en todos los sujetos de estudio, se presenta en la ecuación (2.4)

$$\mathbf{Y} = \underbrace{\mathbf{X}\boldsymbol{\beta}}_{Fijos} + \underbrace{\mathbf{Zu} + \boldsymbol{\varepsilon}}_{Aleatorios} \tag{2.4}$$

y los supuestos sobre los efectos aleatorios y los errores son:

$$\boldsymbol{u} \sim N(0, \mathbf{D}) \quad y \quad \boldsymbol{\varepsilon} \sim N(0, \mathbf{R}).$$

- $\mathbf{Y}$ es un vector $(n \times 1)$, donde $n = \sum n_i$, es el resultado de "ordenar" los Y_i vectores para todos las observaciones verticalmente.
- $\mathbf{X}$ es una matriz de diseño de dimensión $n \times (p+1)$ de constantes conocidas.
- $\boldsymbol{\beta}$ es un vector $(p+1) \times 1$ de parámetros desconocidos no aleatorios y son llamados "efectos fijos".
- $\mathbf{Z}$ es una matriz $n \times q$ de constantes conocidas.
- $\boldsymbol{u}$ es un vector aleatorio de $q \times 1$ y son llamados "efectos aleatorios".
- $\boldsymbol{\varepsilon}$ es un vector $n \times 1$ de errores aleatorios.

2.2.1. Propiedades Básicas

Una propiedad de un Modelo Lineal (ML) es que $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ con $\boldsymbol{\beta}$ siendo los efectos fijos; para un LMM todavía usamos $\mathbf{X}\boldsymbol{\beta}$ para efectos fijos, pero añadimos $\mathbf{Zu}$. Aunque los elementos de $\mathbf{u}$ son variables aleatorias, es conveniente especificar el modelo condicional en sus valores no observados, pero valores realizados. Así, por [McCulloch, 2001] el valor esperado es:

$$\begin{aligned} E(\mathbf{Y}) &= E(\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}) && (2.5)\\ &= E(\mathbf{X}\boldsymbol{\beta}) + E(\mathbf{Zu}) + E(\boldsymbol{\varepsilon}) && (2.6)\\ &= \mathbf{X}\boldsymbol{\beta}. && (2.7) \end{aligned}$$

Se asume que:

$$\mathbf{u} \sim (0, \mathbf{D}), \tag{2.8}$$

Al calcular la $Var(\mathbf{Y})$, necesitamos la $Var(\mathbf{u}) = \mathbf{D}$ de (2.8) y aplicando esta propiedad tenemos

$$Var(\mathbf{Y}) = Var(E[\mathbf{Y} \mid \mathbf{u}]) + E[Var(\mathbf{Y} \mid \mathbf{u})],$$

$$\begin{aligned} Var(\mathbf{Y}) &= Var(E[\mathbf{Y} \mid \mathbf{u}]) + E[Var(\mathbf{Y} \mid \mathbf{u})] \\ &= Var(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) + E[\mathbf{R}] \\ &= Var(\mathbf{X}\boldsymbol{\beta}) + Var(\mathbf{Z}\mathbf{u}) + E[\mathbf{R}] \\ &= Var(\mathbf{Z}\mathbf{u}) + E[\mathbf{R}] \\ &= \mathbf{Z}Var(\mathbf{u})\mathbf{Z}' + E[\mathbf{R}] \\ &= \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}, \end{aligned}$$

por lo tanto,

$$Var(\mathbf{Y}) = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}, \tag{2.9}$$

Entonces por (2.7) y (2.9) tenemos

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}). \tag{2.10}$$

Mostrando así que los efectos fijos solo influyen en la media de $\mathbf{Y}$, mientras que los efectos aleatorios influyen sobre la varianza de $\mathbf{Y}$.

2.3. Métodos de Estimación

Trataremos ahora los métodos para estimar los parámetros de los *LMM*. Para la presentación de los métodos de estimación, vamos a definir el modelo como

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} && (2.11) \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. && (2.12) \end{aligned}$$

donde la estructura de las matrices son como se define anteriormente, y los supuestos sobre los efectos aleatorios y los errores son $\mathbf{u} \sim N(0, \mathbf{D} = \sigma^2\mathfrak{D})$ y

$\boldsymbol{\varepsilon} \sim N(0, \mathbf{R} = \sigma^2 \mathfrak{R})$.

Ahora, por razones técnicas, separamos el factor de escala σ^2 [Mehtätalo, 2013], de las matrices de varianza-covarianza $\mathbf{D}$ y $\mathbf{R}$ previamente definidas en (2.1.1) y (2.1.2). Las nuevas matrices $\mathfrak{D}$ y $\mathfrak{R}$ especifican la estructura de los efectos aleatorios y los errores hasta una constante escalar σ^2. Además, se deduce que

$$Var(\mathbf{e}) = \sigma^2(\mathbf{Z}\mathfrak{D}\mathbf{Z}' + \mathfrak{R}).$$

También notemos que si se supone independencia y varianza constante para los errores $\boldsymbol{\varepsilon}_i$, entonces $\mathfrak{R} = I_{(n\times n)}$ con $n = \sum n_i$.
La estructura de $\mathbf{D}$ y $\mathbf{R}$ especifican un conjunto parsimonioso de parámetros θ_D y θ_R, que se agrupan en $\boldsymbol{\theta} = (\theta_D, \theta_R)'$. La estimación de los parámetros involucrados en el modelo implica encontrar las estimaciones de los parámetros $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ y σ^2.

2.3.1. Mínimos Cuadrados Generalizados (*GLS*)

Si las hipótesis sobre la varianza constante y correlación cero entre los errores no se cumplen, el estimador *OLS* sigue siendo insesgado. Sin embargo, ya no es estimador de mínima varianza. Ahora se puede utilizar el estimador por mínimos cuadrados generalizados, que en inglés se escribe Generalized Least Squares (*GLS*).

En la matriz $\mathbf{V}$ se parametrizan las hipótesis de heterocedasticidad residual y la correlación entre los errores, donde $\mathbf{V}$ es la matriz de varianza y covarianza del error, es decir, en forma más general $Var(\mathbf{e}) = \sigma^2\mathbf{V}$, el estimador *GLS* de $\boldsymbol{\beta}$ minimiza la suma de cuadrados $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ [Mehtätalo, 2013]. Así el estimador *GLS* es

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}. \tag{2.13}$$

Para el modelo lineal con $Var(\mathbf{e}) = \sigma^2\mathbf{V}$, este estimador es el *BLUE*, es decir, tiene la variación más pequeña de entre todos las posibles estimadores insesgados (generalización del teorema de Gauss-Markov).
El estimador *GLS* es insesgado y tiene matriz de varianza-covarianza $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}^{-1}$

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}E[\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned}$$

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}) &= Cov((\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}Cov(\mathbf{Y})((\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})' \\ &= \sigma^2((\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}) \\ &= \sigma^2((\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}) \\ &= \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \end{aligned}$$

El estimador *OLS* se puede derivar de los resultados anteriores, sustituyendo $\sigma^2\mathbf{V}$ por $\sigma^2 I$ [Mehtätalo, 2013].

2.3.2. Máxima Verosimilitud

El método de máxima verosimilitud está basado en un modelo marginal

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V}(\boldsymbol{\theta})), \tag{2.14}$$

donde la matriz de varianza-covarianza, $\mathbf{V}(\boldsymbol{\theta})$, se define como

$$\mathbf{V}(\boldsymbol{\theta}) = \mathbf{Z}\mathfrak{D}(\theta_D)\mathbf{Z}' + \mathfrak{R}(\theta_R).$$

Su función de verosimilitud es:

$$L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = \prod_{i=1}^{n}(2\pi)^{-n_i/2}|\sigma^2 V_i(\theta)|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{Y_i}-\mathbf{X_i}\beta)'(\sigma^2 V_i(\theta))^{-1}(\mathbf{Y_i}-\mathbf{X_i}\beta)}. \tag{2.15}$$

Además, su función log-verosimilitud es:

$$\begin{aligned} \log L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2\mathbf{V}(\boldsymbol{\theta})) - \frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})(\sigma^2\mathbf{V}(\boldsymbol{\theta}))^{-1} \\ &\quad (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{k=1}^{K} log|V_k(\theta)| \\ &\quad -\frac{1}{2\sigma^2}\sum_{k=1}^{K}(Y_k - X_k\beta)V_K(\theta)^{-1}(Y_k - X_k\beta). \end{aligned} \tag{2.16}$$

En el paso siguiente, el log- verosimilitud se puede reducir a una verosimilitud que depende únicamente de los parámetros de componentes de la varianza $(\sigma^2, \boldsymbol{\theta})$. Este se puede alcanzar mediante la sustitución del valor de $\boldsymbol{\beta}$ en (2.16). Especialmente, se pueden eliminar utilizando el estimador de *GLS* (o *ML*).
Por lo tanto, el estimador de máxima verosimilitud de β bajo los errores normales equivale al estimador de mínimos cuadrados [Rao, 2007]

$$\hat{\beta}(\boldsymbol{\theta}) = (\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{V}(\boldsymbol{\theta}))^{-1}\mathbf{Y}. \tag{2.17}$$

Sustituyendo $\hat{\beta}(\boldsymbol{\theta})$ en la función de log-verosimilitud en (2.16) tenemos

$$\log L(\hat{\beta}, \sigma^2, \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\sigma^2\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\hat{e})$$

$$(\sigma^2\mathbf{V}(\boldsymbol{\theta}))^{-1}(\mathbf{Y} - \mathbf{X}\hat{e}), \tag{2.18}$$

donde $\hat{e} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{Y})$.

El estimador de máxima verosimilitud de σ^2 se obtiene diferenciando (2.18) con respecto a σ^2.

$$\hat{\sigma}^2 = \frac{\hat{e}'\mathbf{V}(\boldsymbol{\theta})^{-1}\hat{e}}{n}. \tag{2.19}$$

Esto muestra que $\hat{\sigma}^2$ puede ser una función de $\boldsymbol{\theta}$ (y los datos).
Sustituyendo $\hat{\sigma}^2$ en (2.18) y reduciendo el log- verosimilitud obtenemos:

$$= -\frac{1}{2}(n)log(\hat{e}'\mathbf{V}(\boldsymbol{\theta})^{-1}\hat{e}) + \sum_{i=1}^{K} log|V_i(\theta)|. \tag{2.20}$$

El problema de estimación es ahora maximizar el log-verosimilitud anterior, con respecto al parámetro $\boldsymbol{\theta}$, igualar a cero y resolver el valor de θ, para obtener el valor $\hat{\boldsymbol{\theta}}$.

2.3.3. Máxima Verosimilitud Restringida

La idea de máxima verosimilitud restringida se presentó en la sección 1.4.3. En esta sección se continúa la discusión en el contexto de modelos lineales de efectos mixtos.
El modelo asumido se especifica como

$$K'\mathbf{Y} \sim N[0, K'(\sigma^2\mathbf{V}(\boldsymbol{\theta}))^{-1}],$$

donde K es una matriz de rango completo de tamaño $n \times (n-p)$ y cumple la condición $K'\mathbf{X} = 0$. El método para encontrar la matriz K fue discutida en la sección 1.4.3.
El nuevo vector de respuesta, $K'\mathbf{Y}$, tiene solo $n-p$ elementos en vez de los elementos del $\mathbf{Y}$ original. Sin embargo, sí incluye toda la información contenida en los errores originales. Por lo tanto, para adaptarse al nuevo modelo en los datos transformados conduce a tales estimaciones de la parte aleatoria que tengan en cuenta los grados de libertad que se utilizan para la estimación de la parte fija [Mehtätalo, 2013]. Además, debido a las condiciones $K'\mathbf{X} = 0$ y $K'\mathbf{X}\boldsymbol{\beta} = 0$, los datos $K'\mathbf{Y}$ no incluye cualquier variación que podría ser explicado por β's que corresponde a la matriz del modelo especificado de la parte fija, $\mathbf{X}$.
La función log-verosimilitud de REML está dada por:

$$\begin{aligned} l(\sigma^2, \boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\log|\sigma^2 K'\mathbf{V}(\boldsymbol{\theta})K| - \frac{1}{2\sigma^2}(K'\mathbf{Y})'(K'\mathbf{V}(\boldsymbol{\theta})K)^{-1}K'\mathbf{Y} \\ &= -\frac{n-p}{2}\log(2\pi) - \frac{n-p}{2}log(\sigma^2) - \frac{1}{2}\log|K'\mathbf{V}(\boldsymbol{\theta})K| \end{aligned}$$

$$-\frac{1}{2\sigma^2}\mathbf{Y}'K(K'\mathbf{V}(\boldsymbol{\theta})K)^{-1}K'\mathbf{Y}. \tag{2.21}$$

Debido a que K se define de forma que elimina los efectos de los parámetros fijos, la verosimilitud es una función sólo de $\boldsymbol{\theta}$ y σ^2. Por lo tanto, el log- verosimilitud restringido de los modelos lineales mixtos se vuelve más simple que en el contexto de *ML*. Sobre todo, sólo se necesita σ^2, usando el estimador *REML*

$$\hat{\sigma}^2 = \tfrac{1}{n-p}(K'\mathbf{Y})'(K'\mathbf{V}(\boldsymbol{\theta})K)^{-1}(K'\mathbf{Y}).$$

Sustituyendo $\hat{\sigma}^2$ en la función log- verosimilitud del LMM nos queda una función como sigue:

$$l_R(\boldsymbol{\theta}) = l_R(\hat{\sigma}^2, \boldsymbol{\theta}),$$

que es una función solo de $\boldsymbol{\theta}$. Maximizando esta función con respecto a $\boldsymbol{\theta}$ da la estimación de $\boldsymbol{\theta}$. Esto permite el cálculo de una estimación de la matriz $\mathbf{V}(\boldsymbol{\theta})$ y, además, la estimación de σ^2 utilizando el estimador *REML* antes estimado. Por último, la estimación numérica de $\boldsymbol{\beta}$ se calcula utilizando el estimador *GLS* en la sección 2.3.1.

2.4. Ajuste del Modelo

Inferencias y pruebas de los *LMM* se pueden basar en los procedimientos utilizados en el modelo lineal general. Como se discutió anteriormente en el Capítulo 1, las pruebas sobre los coeficientes separados y sobre el ajuste global del modelo son todos los casos especiales de una situación en la que se comparan dos modelos anidados uno contra el otro. Por lo tanto, es suficiente presentar un procedimiento para tal situación, formulamos la hipótesis nula y alternativa como:

H_0 : El modelo restringido es suficiente. vs
H_a : El modelo completo es significativamente mejor que el modelo restringido

El modelo nulo se obtiene al hacer restricciones a los parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\theta}$ del modelo completo. La prueba se realiza mediante el cálculo de la probabilidad de obtener un valor tan extremo o incluso más extremo de una estadística de prueba, cuando la hipótesis nula es verdadera. Si la probabilidad es baja, entonces lo tomamos como evidencia en contra de la hipótesis nula y aceptamos la alternativa [Mehtätalo, 2013].

2.4.1. Pruebas de Razón de Verosimilitud

Una de las alternativas de pruebas estadísticas del modelo con una estructura general de $\mathbf{V}$ es la razón de verosimilitud estadística, que se desarrolla a continuación.
La razón de verosimilitud es una metodología general para construir pruebas de hipótesis [Peña, 2002]. Con frecuencia se desea comprobar si una muestra puede provenir de una distribución con ciertos parámetros conocidos. Se supone que se desea contrastar la hipótesis nula:

$H_o : \theta \in \Theta_0$
θ está contenido en una región Θ_0 del espacio paramétrico Θ, vs
$H_a : \theta \in \Theta - \Theta_0$
que supone que θ no está restringida a la región Θ_0 .

Para comparar estas hipótesis se analiza su capacidad de prever los datos observados, y, para ello, compararemos las probabilidades de obtenerlos bajo ambas hipótesis. El método de razón de verosimilitudes resuelve este problema tomando el valor que hace más probable obtener la muestra observada y que es compatible con las hipótesis.
El estadístico de prueba es el cociente de las verosimilitudes:

$$LRT \quad = \quad 2\ln(\frac{L_2}{L_1}) = 2[\ln(L_2) - \ln(L_1)].$$

Bajo la hipótesis nula, tenemos (al menos asintóticamente)

$$LRT \sim \chi^2(p-q)$$

donde $f(H_o)$ y $f(H_a)$ son el máximo valor de las verosimilitudes compatibles con H_o y H_a, respectivamente.

Por construcción $LRT \leq 1$ y rechazaremos H_o cuando LRT sea suficientemente grande. La región de rechazo de H_o vendrá en consecuencia, definida por $LRT \leq \alpha$, donde α se determinará imponiendo que el nivel de significación de la prueba, sea α.

Al llevar a cabo las pruebas en *LMM*, las siguientes cuestiones deben ser reconocidas.

1. Si hay restricción en $\boldsymbol{\theta}$, se comparan los modelos anidados con diferente estructura de $\mathbf{V}$, por lo general, se recomienda que la verosimilitud utilizada debe basarse en *REML*. Estas pruebas pueden ser conservadoras, es decir, pueden fallar con demasiada facilidad para rechazar la hipótesis nula ([Pinnheiro, 2000], p. 83-87). Cualquier prueba LR utilizando *REML* deben basarse en modelos con la misma parte fija.

2. Si se comparan los modelos anidados con diferentes partes fijas, se debe preferir una prueba condicional t o F. Si se utilizan pruebas de razón de verosimilitud, la verosimilitud utilizada debe basarse en *ML*. Sin embargo, en algunas situaciones (especialmente si el número de observaciones por grupo es baja; ver [Pinnheiro, 2000], pág. 87- 92). Estas pruebas pueden ser severamente anti-conservadora (es decir, que con demasiada facilidad rechazan la hipótesis nula).

3. La prueba de razón de verosimilitud es asintótica, lo que significa que tiene la distribución χ^2 sólo con muestras grandes. La prueba condicional F es aproximada, ya que supone que la matriz estimada $\mathbf{V}$ es la matriz verdadera $\mathbf{V}$. [Mehtätalo, 2013].

2.4.2. Deviance

También se brinda el valor de la "deviance" residual con sus correspondientes grados de libertad, que en el caso de suponer el error normal no es más que el error de la suma de cuadrados residuales dividido entre la varianza del error. [Wackerly, 2010]. Con los cuartiles de la "deviance" de los residuos, que también se muestran, se puede revisar el supuesto de normalidad.
La "deviance" del modelo se define como:

$$\lambda(\beta) \quad = \quad 2\ln L(modelo \quad saturado) - 2\ln L(\beta).$$

Si el tamaño de la muestra n es grande y el modelo es correcto $\lambda(\beta)$ se distribuye

$$\chi^2_{n-p}.$$

Algunas ideas que pueden utilizarse para la interpretación son:

- Valores grandes de la "deviance": el modelo NO es correcto.
- Valores pequeños de la "deviance": el modelo se ajusta a los datos tan bien como el modelo saturado.

Una regla fácil que suele utilizarse en la práctica es que si el cociente $[\lambda(\beta)]/(n-p)]$ es aproximadamente 1, se considera el modelo adecuado.

2.5. Métodos de Selección de Modelos: AIC Y BIC

Según [Mehtätalo, 2013] no están disponibles pruebas formales sobre modelos no anidados. Sin embargo, si dos modelos se basan en el mismo conjunto de datos y la misma respuesta, la comparación de modelos puede estar basada en los criterios de información. Los dos criterios comúnmente calculados por los softwares estadísticos son el criterio de información de Akaike (AIC) y el criterio de información de Schwartz o Bayesiano (BIC).

$$AIC = \log(L) - p, \tag{2.22}$$

$$BIC = \log(L) - \frac{1}{2}p\ln n^*, \tag{2.23}$$

donde $n^* = n$ si L es la verosimilitud convencional y $n^* = n - p$, si L es la verosimilitud restringida.

Esencialmente, los criterios tienen dos componentes: verosimilitud y una penalización asociada con el número de parámetros que intervienen en el modelo. Como el ajuste del modelo se basa en la máxima verosimilitud, se selecciona el modelo con mayor *AIC* o *BIC*. Tenga en cuenta que a veces los criterios se definen utilizando el negativo de las diferencias (ver más [Galecki, 2013], pág. 87). En este caso, el modelo con el valor más pequeño del criterio se considera mejor, y este convenio es adoptado en el software R.

2.6. Revisión de las Suposiciones del Modelo

Diferentes gráficos de diagnóstico se utilizan para evaluar qué tan bien se cumplen las suposiciones que se hicieron en la formulación del modelo en el conjunto de datos. Los gráficos de diagnóstico proporcionan información (i) para mejorar la formulación del modelo y (ii) para evaluar la validez de la inferencia y pruebas.

2.6.1. Gráficos de Diagnóstico

Con el modelo mixto, la comprobación de que el modelo se ajusta a los datos es también tan importante como con los modelos simples. El gráfico de los residuales en todos los predictores y sobre el valor predicho es un buen punto de partida para ver si el modelo tiene una buena forma [Cayuela, 2012]. Para corregir la forma del modelo, las mismas reglas se aplican como en el caso de los modelos lineales (véase la sección 1.6.1)

Efectos Aleatorios u

La evaluación gráfica de los supuestos sobre los efectos aleatorios es posible, especialmente si el número de parámetros aleatorios es ≤ 2. En un modelo con constante aleatoria, sólo el nivel del modelo asumido se asume que variar entre los grupos, mientras que la pendiente se supone que es el mismo en todos los grupos. El gráfico de los datos y ajustes de los grupos específicos se puede usar para ver si esto es una hipótesis realista [Mehtätalo, 2013].

Los efectos aleatorios se supone que son realizaciones i.i.d. de una distribución normal (multivariante). La suposición de que la distribución es i.i.d. puede ser parcialmente evaluada explorando si la varianza es constante sobre el rango de los grupos, [Mehtätalo, 2013]. Estas gráficas deben mostrar la variabilidad homoscedastica en el rango de predicciones. La normalidad de los efectos aleatorios

puede ser parcialmente evaluada mediante la exploración de que si la distribución marginal de los efectos aleatorios es un gráfico de una normal (usando q-q) y si la correlación de todos los pares de efectos aleatorios es lineal. La base en estas evaluaciones se encuentra en las propiedades de la distribución normal multivariante, donde todas las distribuciones marginales son también normales y las correlaciones de las componentes diferentes son lineales.

Los Residuales e

Una gráfica de los residuales estandarizados (condicionales) en el valor predicho debería expresar una varianza constante sin tendencias. Una función de la varianza se puede utilizar para homogeneizar los residuos o, alternativamente, una transformación se puede hacer a la respuesta. La normalidad de los residuos se debe comprobar, por ejemplo, mediante el uso de gráficos q-q, como los métodos de *ML* y *REML* se basan en la normalidad. Normalmente, se permite una ligera discrepancia de la normalidad [Mehtätalo, 2013].

Capítulo 3

Aplicación de Modelos Mixtos

La hojarasca es el término que se emplea para definir la mezcla de hojas, flores, frutos y parte lignificadas (ramitas no mayores de 1 cm de diámetro, corteza, etc.), que caen al suelo proveniente del estrato arbóreo y constituye la fuente principal de incorporación de materia orgánica, la cual posee composición y características diferentes en dependencia de la especie o el tipo de bosque de que proceda. Por la importancia de esta temática en la estabilidad y funcionamiento de los ecosistemas, al constituir el eslabón que garantiza la circulación de materia orgánica y nutrientes entre las plantas y el suelo, así como las potencialidades que posee para atenuar los cambios climáticos, protegiendo al suelo contra agentes erosivos que degraden su fertilidad, es que se incursiona en la modelación de la hojarasca en función de las propiedades físico-químicas del suelo, el tipo de suelo y el clima de la región. Dado que los posibles predictores de los modelos de regresión que se pueden establecer son mezclas de variables cuantitativas y cualitativas, estas últimas aleatorias, es necesario utilizar los modelos Mixtos. En el Apéndice B se desarrollan los elementos obtenidos en la Tesis Doctoral de [Castillo, 2014], que son el punto de partida de nuestra investigación.

Planteamiento del problema

En la Sierra Norte del Estado Puebla en la Región Terrestre Prioritaria (RTP-105), se obtuvieron 10 muestras de hojarasca, en cada uno de los cinco tipos de suelos P30F, P35F, P36F, Piñonero y Yucca (PERFILES) junto con sus regímenes de temperatura (Térmico, Isomésico y Isotérmico). En la tesis doctoral de [Castillo, 2014] se explica a detalle el análisis de laboratorio de las muestras, para obtener, el porcentaje de carbono Orgánico (PorcientodeCorg), porcentaje de nitrógeno (PorcientodeNT), pH de agua(pHdeagua), pH de Cloruro de potasio (pHdeKCL) y Deltha pH de la hojarasca [Oroza, 2012]. El interés del investigador

es ver la relación que existe en estas variables, suponiendo que tanto los PERFILES del suelo como los Regímenes de temperatura, son una muestra de todos los posibles PERFILES que existen en la Región Terrestre Prioritaria (RTP-105), así como también los regímenes de temperatura.

3.1. Modelos Lineales Mixtos en la Determinación de Carbono Orgánico en la Hojarasca

3.1.1. Formulación de un Modelo con un Efecto Fijo y un Efecto Aleatorio

Para entender un poco los modelos mixtos un primer modelo que sería interesante conocer para el investigador del Departamento de Investigación de Ciencias Agrícolas (DICA) de la Benemérita Universidad Autónoma de Puebla (BUAP) es del porcentaje de carbono orgánico en función del porcentaje de nitrógeno, como variable fija y como variable aleatoria los PERFILES, así el modelo mixto para la $i-$ ésima observación de esta relación quedaría de la siguiente forma:

$$PorcientoCorg_{ij} = \beta_0 + \beta_1 PorcientodeNT_{ij} + u_j + \varepsilon_{ij}. \tag{3.1}$$

$$i = 1,\ldots,10.$$
$$j = 1,2,3,4,5.$$

$$u_j \sim N(0,\sigma_u^2) \quad \varepsilon_{ij} \sim N(0,\sigma^2).$$

El índice j (representa a los PERFILES del suelo) toma valores de 1 a 5 (P30F, P35F, P36F, Piñonero y Yucca). Además, hay una variable aleatoria, u_j, que añade cierta cantidad de variación al modelo general en cada uno de los PERFILES. Se asume que esta variable aleatoria sigue una distribución normal con media 0 y varianza σ_u^2. Por lo tanto, los parámetros que se deben estimar en el modelo son cuatro, β_0, β_1, la varianza del error σ^2, y la varianza σ_u^2. El modelo para todos los datos está especificado de la siguiente forma

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon} \tag{3.2}$$

definiendo

$$\mathbf{Y} = \begin{pmatrix} 82 \\ 82 \\ 80 \\ 80 \\ 82 \\ \vdots \\ 82 \end{pmatrix}_{(50\times1)} \qquad \mathbf{X} = \begin{pmatrix} 1 & 2{,}61 \\ 1 & 2{,}55 \\ 1 & 2{,}80 \\ 1 & 2{,}83 \\ 1 & 2{,}53 \\ \vdots & \vdots \\ 1 & 0{,}54 \end{pmatrix}_{(50\times2)} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{(2\times1)}$$

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}_{(50\times5)} \qquad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{pmatrix}_{(5\times1)}$$

$Var(\boldsymbol{\varepsilon}) = \mathbf{R} = \sigma^2 I_{(50\times 50)}$.

$Var(\mathbf{u}) = \mathbf{D} = \sigma_u^2 I_{(5\times 5)}$.

$Var(e) = Var(\mathbf{Zu} + \boldsymbol{\varepsilon}) = \mathbf{ZDZ}' + \mathbf{R} =$

$$\left(\begin{array}{cccc|c|cccc} \sigma_u^2+\sigma^2 & \sigma_u^2 & \dots & \sigma_u^2 & \dots & 0 & 0 & \dots & 0 \\ \sigma_u^2 & \sigma_u^2+\sigma^2 & \dots & \sigma_u^2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & \dots & 0 & 0 & \dots & 0 \\ \sigma_u^2 & \sigma_u^2 & \dots & \sigma_u^2+\sigma^2 & \dots & 0 & 0 & \dots & 0 \\ \hline \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \hline 0 & 0 & 0 & 0 & \dots & \sigma_u^2+\sigma^2 & \sigma_u^2 & \dots & \sigma_u^2 \\ 0 & 0 & 0 & 0 & \dots & \sigma_u^2 & \sigma_u^2+\sigma^2 & \dots & \sigma_u^2 \\ 0 & 0 & 0 & 0 & \dots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \sigma_u^2 & \sigma_u^2 & \dots & \sigma_u^2+\sigma^2 \end{array}\right)_{(50\times 50)}.$$

3.2. Metodologías para la Formulación de Modelos Lineales Mixtos

3.2.1. Metodología Crawley

Al iniciar el proceso de modelación, una pregunta engorrosa que se hace un investigador es: ¿Emplear efectos fijos o emplear efectos aleatorios? Es difícil, sin mucha experiencia, saber cuándo usar variables independientes categóricas como efectos fijos y, cuándo usarlas como efectos aleatorios.

Según Crawley J. M. [Crawley, 2008] deben considerarse dos aspectos. Uno se refiere a la utilización de gráficos que destaquen las relaciones entre la variable dependiente y la posible variable cualitativa con efecto aleatorio. El otro se refiere a hacer comparaciones de diferentes modelos. Recomienda hacer diferentes procedimientos de comparación, tales como:

* Tipo de Comparación 1:

  La clave para entender es ajustar modelos de regresión lineal para cada categoría de la variable de efectos aleatorios y ajustar un modelo de efectos mixtos, tomando en cuenta las diferencias entre las categorías de la variable en términos de su contribución a la varianza en la respuesta medida por

una desviación estándar en el intercepto y una desviación estándar en la pendiente.

* Tipo de Comparación 2:

  Podemos ajustar diferentes modelos mixtos. Debido a que intentamos comparar los modelos con estructuras de efectos fijos diferentes necesitamos especificar el método de máxima verosimilitud. Después de correr diferentes modelos se hace la comparación de los modelos a través de un ANOVA.

* Tipo de Comparación 3:

  Hacer el modelo lineal tradicional. Para probar si uno debiera usar un modelo con efectos mixtos o hacer el modelo lineal tradicional, D. Bates escribió en el archivo de ayuda del programa lm en R: "Yo recomendaría la prueba de razón de verosimilitud contra un modelo lineal ajustado por lm. El valor de p que retorna este test será conservativo porque se está probando en la vecindad del espacio paramétrico" [Pinnheiro, 2000].

3.2.2. Metodología Zuur

La metodología de [Zuur, 2009], es diferente a la metodología de Crawley, pues en la de Crawley queremos decidir, si modelar con mixtos o no, sin embargo, Zuur ya sabe que es mejor modelar con modelos mixtos, solo que se desea obtener el mejor modelo mixto, basado en los criterios de selección de modelos AIC, BIC y Deviance. Obsérvese que en los *LMM*, tenemos dos tipos de efectos, los efectos fijos y los aleatorios. Aunque generalmente vamos a estar más interesados en los efectos fijos, si tenemos una estructura de efectos aleatorios mal definida es posible que afecte a la estimación de los efectos fijos. Por ello es importante seleccionar la mejor estructura para cada uno de estas dos componentes, utilizando la siguiente metodología [Zuur, 2009]:

- Empezar ajustando un modelo en donde la componente fija contenga todas las variables independientes e interacciones posibles. Esto es lo que se conoce como modelo más allá del óptimo, que en inglés se escribe beyond optimal model. Si no es posible ajustar este modelo porque hay demasiados parámetros, hay que intentar seleccionar las variables e interacciones que se cree que influyen más sobre la variable dependiente.

- Utilizando el modelo más allá del óptimo, encontrar la estructura de la componente aleatoria óptima. Para ello se proponen distintos modelos alternativos con la misma estructura en la componente fija pero que varían

en su componente aleatoria. Estos modelos tienen que estar estimados utilizando *REML*. La comparación entre distintos modelos podemos hacerla utilizando la función anova() del sofware R.

- Una vez definida la estructura óptima de la componente aleatoria, buscar la estructura óptima de la componente fija del modelo. Para ello podemos utilizar el estadístico F o el estadístico t obtenido mediante el estimador *REML* con la función lme() o comparar modelos anidados. Para comparar modelos que tienen la misma estructura en la componente aleatoria pero difieren en la componente fija se debe de utilizar un estimador *LM* y no un estimador de *REML*.

- Cuando se ha seleccionado la estructura de la componente fija, presentar el modelo final utilizando un estimador *REML* y analizar las suposiciones establecidas en el mismo.

A continuación, se aplica esta metodología al estudio del carbono orgánico en la hojarasca.

3.3. Aplicación en la Determinación de Carbono Orgánico en la Hojarasca: formulación

3.3.1. Metodología de Crawley

Como se planteó anteriormente el propósito de la investigación es modelar, el PorcientodeCorg en función de sus propiedades físico-químicas y del tipo de suelo, que en los comandos de R lo nombraremos como "PERFIL". Para alcanzar este propósito debemos analizar qué tipo de modelo es el más adecuado: el de efectos fijos o el de efectos aleatorios.

Siguiendo la metodología apuntada por [Crawley, 2008], primero haremos uso de gráficos para comprobar algunas ideas, una de ellas es que se espera que el porcentaje de carbono orgánico y porcentaje de nitrógeno de la hojarasca se correlacionen positivamente.

Leemos nuestra base de datos y los guardamos en datos.

```
> datos<-read.table(file="clipboard",head=T)
> attach(datos)
> names(datos)
```

```
 [1] "RegHumed"          "RegTemp"           "EstaciónClimatica"
 [4] "PERFIL"            "Muestra"           "PorcientoNT"
 [7] "pHdeagua"          "pHdeKCl"           "DeltadepH"
[10] "PorcientoCorg"
```

```
> plot(PorcientoNT,PorcientoCorg,pch=16,col=PERFIL)
```

A continuación graficamos las respectivas muestras, con diferentes colores. Las instrucciones en R para obtener este gráfico son:

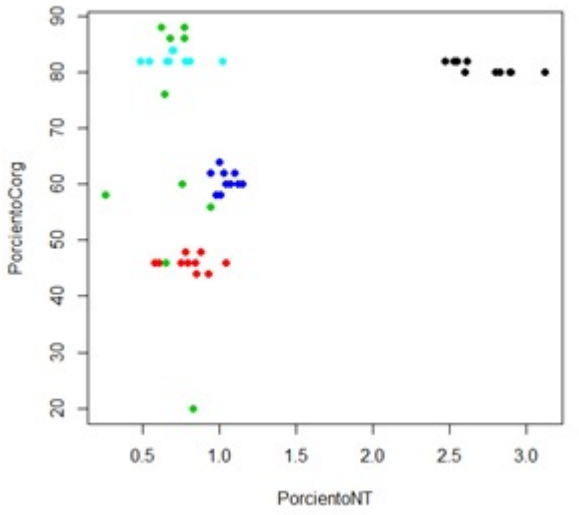

Figura 3.1: Diagrama de PorcientoCorg vs PorcientoNT

Se puede observar en el Gráfico (3.1) el PorcientodeCorg y el PorcientodeNT para cada uno de los distintos PERFILES y, aunque no se aprecia en conjunto que existe correlación entre las variables cuantitativas PorcientoCorg y PorcientoNT, se destaca que los tipos de suelos (PERFIL) tienen comportamientos particulares.

Como segundo aspecto de la metodología [Crawley, 2008] se hacen comparaciones entre diferentes modelos.

Iniciamos con la comparación de tipo 1. La distinción clave para entender es, entre varios ajustes de modelos de regresión lineal (uno para cada tipo de suelo)

y el modelo de ajuste de efectos, teniendo en cuenta las diferencias entre los distintos tipos de suelo, en términos de su contribución a la variación en la variable dependiente, medida por una desviación estándar en intercepto y una desviación estándar en la pendiente. Estas diferencias se investigan contrastando las dos funciones de ajuste de R, lmList y lme.

Comenzamos por la colocación de cinco modelos lineales separados, uno para cada tipo de suelo (PERFIL). Los comandos en R y la salida correspondiente se muestran debajo:

```
> linear.models<-lmList(PorcientoCorg~PorcientoNT|PERFIL,datos)
> coef(linear.models)
         (Intercept) PorcientoNT
P30F        91.36415  -3.8696538
P35F        47.12163  -1.3933237
P36F        75.84150 -13.6437908
Piñonero    63.38115  -2.6639344
Yucca       82.50092  -0.1437667
\end{frame}
```

Observamos variaciones sustanciales en el valor del intercepto 91.36415 en el PERFIL $P30F$ con respecto al intercepto 47.12163 en el PERFIL $P35F$. Ambas pendientes son negativas. Este es un problema clásico en el análisis de regresión, cuando el intercepto es muy grande con respecto del valor medio de X: grandes valores del intercepto están casi obligados a estar correlacionadas con bajos valores de la pendiente.

Las pendientes y los interceptos del modelo especificado por completo en términos de efectos aleatorios: una población de regresión con pendientes predichas dentro de cada PERFIL con nitrógeno como variable independiente:

```
> random.model<-lme(PorcientoCorg~1,random=~PorcientoNT|PERFIL)
> coef(random.model)
         (Intercept)   PorcientoNT
P30F        80.18426  1.997411e-07
P35F        46.96447  5.960822e-07
P36F        66.43814 -2.019024e-07
Piñonero    60.90151  1.436161e-07
Yucca       81.71161 -4.772319e-07
```

Este resultado muestra una variación similar de los interceptos, pero las pendientes tienen signos opuestos.

Los Gráficos (3.2) y (3.3), pueden ayudar a esclarecer la situación. Los comandos de R utilizados y las salidas gráficas resultantes se muestran debajo; la primera se refiere a los interceptos y las segundas a las pendientes de los modelos respectivos.

```
> mm<-coef(random.model)
> barra<-coef(linear.models)
> par(mfrow=c(2,2))
> plot(barra[,1],mm[,1],pch=16,xlab="linear",ylab="random effects")
> abline(0,1)
> plot(barra[,2],mm[,2],pch=16,xlab="linear",ylab="random effects")
> abline(0,1)
```

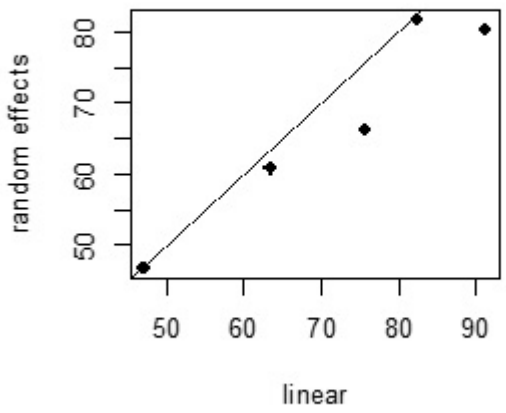

Figura 3.2: Interceptos de modelos de regresión y del Modelo Mixto

En el Gráfico (3.2) los interceptos de los modelos de efectos aleatorios son menores que sus equivalentes de modelos lineales, que están por debajo de la línea de 45 grados.

Mientras que en el Gráfico (3.3) la mayoría de las pendientes de efectos aleatorios (derecha de la salida de random.model) son más grandes que sus equivalentes al modelo lineal (es decir, por la izquierda de la línea). Por ejemplo, para el modelo lineal el valor del intercepto es de 91.36415 y el valor del intercepto en el modelo

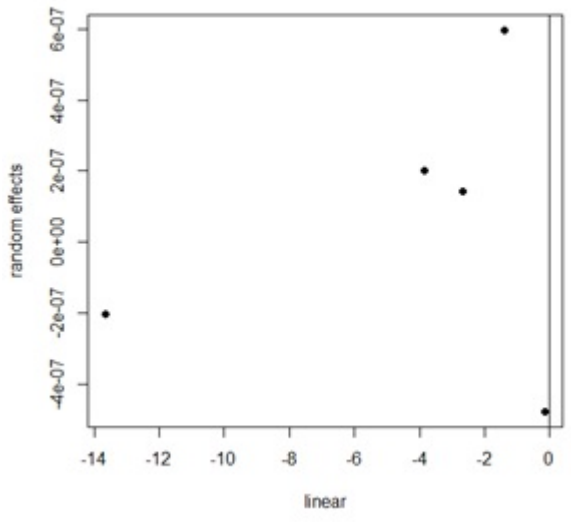

Figura 3.3: Pendientes del Modelo Lineal y del Modelo aleatorio

de efectos aleatorios es de 80.18426 en el PERFIL $P30F$. Del mismo modo para el modelo lineal el valor de la pendientes es de -3.8696538 y el valor del intercepto del modelo de efectos aleatorios es de 1.997411e-07 (0.0000001997411) en el PERFIL $P30F$.
Ahora, hacemos las comparaciones de tipo 2. Podemos ajustar modelos mixtos con efectos fijos y aleatorios. Uno de esos modelos es cuando el PorcientodeCorg se modela como una función de PorcientodeNT como efecto fijo, y el tipo de suelo (PERFIL) como efecto aleatorio. Tenemos la intención de comparar los modelos con diferentes estructuras de efectos fijos y por esa razón necesitamos especificar el método *ML* en lugar de *REML*. Los comandos R requeridos se muestran debajo.

```
> PERFIL<-factor(PERFIL)
> mixed.model1<-lme(PorcientoCorg~PorcientoNT*PERFIL,random=~1|PERFIL,
method="ML")
> mixed.model2<-lme(PorcientoCorg~PorcientoNT+PERFIL,random=~1|PERFIL,
method="ML")
> mixed.model3<-lme(PorcientoCorg~PorcientoNT,random=~1|PERFIL,
method="ML")
> mixed.model4<-lme(PorcientoCorg~1,random=~1|PERFIL,method="ML")
```

Se muestra el resultado obtenido para la comparación:

```
> anova(mixed.model1,mixed.model2,mixed.model3,mixed.model4)
           Model df   AIC      BIC   logLik   Test L.Ratio p-value
mixed.model1  1  12 391.8267 414.77 -183.91
mixed.model2  2  8  384.1454 399.44 -184.07 1 vs 2  0.3187 0.9886
mixed.model3  3  4  395.8869 403.53 -193.94 2 vs 3 19.7414 0.0006
mixed.model4  4  3  394.1020 399.83 -194.05 3 vs 4  0.2151 0.6428
```

El modelo mixed.model1 no muestra diferencia significativa con el modelo mixed.model2. Tomando en cuenta el criterio de selección de modelo AIC, puede considerarse mixed.model2 el mejor modelo por tomar el valor mínimo en este criterio ($AIC = 384.1454$), sin embargo el modelo mixed.model3 muestra diferencias significativas con el modelo mixed.model2 ($p <.006$), es decir se ajusta a los datos
Las comparaciones de tipo 3, referida al modelo tradicional de análisis de covarianza, también se requieren para decidir la conveniencia de utilizar modelos mixtos o no. La interpretación del análisis de covarianza es exactamente la misma que la interpretación del modelo mixto cuando hay una estructura equilibrada y replicaciones iguales, pero cuando hay replicación desigual, los modelos lineales mixtos brinda mejores posibilidades de interpretación que el modelo lineal tradicional.
Para obtener el modelo tradicional usamos los siguientes comandos R:

```
> model<-lm(PorcientoCorg~PorcientoNT*factor(PERFIL))
```

La salida, que aparece a continuación, muestra que el modelo se ajusta a los datos ya que el $R^2 = 66.63$ % y el, R^2 ajustado es solo ligeramente menor. En cuanto al estadístico F es significativo, por lo que podemos aceptar que el modelo se ajusta a los datos. El problema de este modelo es la cantidad de parámetros a estimar, pues estimaríamos más parámetros a comparación de los modelos anteriores.

```
> summary(model)

Call:
lm(formula = PorcientoCorg ~ PorcientoNT * factor(PERFIL))

Residuals:
    Min      1Q  Median      3Q     Max
-44.517  -0.504  -0.293   1.026  22.664

Coefficients:
```

```
                                Estimate Std.Error t value Pr(>|t|)
(Intercept)                       91.364   46.761   1.954   0.0577 .
PorcientoNT                       -3.870   17.084  -0.227   0.8220
factor(PERFIL)P35F               -44.243   51.277  -0.863   0.3934
factor(PERFIL)P36F               -15.523   48.836  -0.318   0.7523
factor(PERFIL)Piñonero           -27.983   73.476  -0.381   0.7053
factor(PERFIL)Yucca               -8.863   49.882  -0.178   0.8599
PorcientoNT:factor(PERFIL)P35F     2.476   30.942   0.080   0.9366
PorcientoNT:factor(PERFIL)P36F    -9.774   26.117  -0.374   0.7102
PorcientoNT:factor(PERFIL)Piñonero  1.206   56.818   0.021   0.9832
PorcientoNT:factor(PERFIL)Yucca    3.726   29.673   0.126   0.9007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.71 on 40 degrees of freedom
Multiple R-squared:  0.6663,    Adjusted R-squared:  0.5913
F-statistic: 8.876 on 9 and 40 DF,  p-value: 3.424e-07
```

Ahora procedemos a ajustar un modelo muy simplificado con una pendiente común, pero diferentes interceptos para los diferentes tipos de suelo(PERFILES).

```
> model2<-lm(PorcientoCorg~PorcientoNT+factor(PERFIL))
> anova(model,model2)
Analysis of Variance Table

Model 1: PorcientoCorg ~ PorcientoNT * factor(PERFIL)
Model 2: PorcientoCorg ~ PorcientoNT + factor(PERFIL)
  Res.Df    RSS    Df Sum of Sq    F   Pr(>F)
1     40    4585.6
2     44    4614.9 -4   -29.322 0.0639  0.9922
```

Obsérvese que este análisis no proporciona apoyo a las diferencias significativas entre las pendientes. Pero, ¿y qué pasa con los interceptos? Para ello hacemos la siguiente comparación, cuyos comandos en R y salida de la tabla ANOVA se muestran a continuación.

```
> model3<-lm(PorcientoCorg~PorcientoNT)
> anova(model2,model3)
```

```
Analysis of Variance Table

Model 1: PorcientoCorg ~ PorcientoNT + factor(PERFIL)
Model 2: PorcientoCorg ~ PorcientoNT
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     44  4614.9
2     48 12134.8 -4   -7519.9 17.924 8.477e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Esto demuestra que existen diferencias altamente significativas en los interceptos entre los diferentes tipos de suelo ($P30F$, $P35F$, $P36F$, Piñonero y Yucca) considerados en este problema, lo que apunta hacia la conveniencia de modelar el PorcientodeCorg con un modelo de efectos mixtos.
La salida del modelo mixto seleccionado por la métodologia Crawley se muestra a continuación.

```
 summary(mixed.model2)
Linear mixed-effects model fit by maximum likelihood
 Data: NULL
       AIC      BIC    logLik
  384.1454 399.4416 -184.0727

Random effects:
 Formula: ~1 | PERFIL
         (Intercept) Residual
StdDev: 0.0002551959 9.607227

Fixed effects: PorcientoCorg ~ PorcientoNT + PERFIL
                   Value Std.Error DF   t-value p-value
(Intercept)     95.54177  26.94307 44  3.546061  0.0009
PorcientoNT     -5.39991   9.79770 44 -0.551141  0.5843
PERFILP35F     -45.19483  19.40871  0 -2.328585     NaN
PERFILP36F     -25.40502  20.48625  0 -1.240101     NaN
PERFILPiñonero -29.30425  17.14211  0 -1.709490     NaN
PERFILYucca     -9.35103  20.39077  0 -0.458591     NaN
 Correlation:
               (Intr) PrcnNT PERFILP35 PERFILP36 PERFILPñ
PorcientoNT    -0.993
```

```
PERFILP35F     -0.985  0.972
PERFILP36F     -0.987  0.975  0.974
PERFILPiñonero -0.979  0.964  0.968     0.969
PERFILYucca    -0.986  0.974  0.973     0.975     0.969

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-4.752132118 -0.122249770 -0.005652403  0.136810663  2.292148749

Number of Observations: 50
Number of Groups: 5
```

3.3.2. Metodología Zuur

Como se planteó anteriormente el propósito de la investigación es modelar el porcentaje de carbono orgánico de la hojarasca en función de sus propiedades físico-químicas, de los perfiles y regímenes de temperatura de la región. Para alcanzar este propósito debemos analizar qué modelo es el más adecuado.
Siguiendo la metodología apuntada por [Zuur, 2009] y utilizando el software R, primero ajustaremos un modelo en donde la componente fija contenga todas las variables independientes posibles y los efectos aleatorios posibles, después seleccionaremos cual o cuales variables influyen más sobre la variable de respuesta. A continuación se muestran los comandos y las correspondientes salidas del paquete *lme*4 del software R, manteniendo el idioma inglés.

```
> regresion1 <- lm(PorcientoCorg ~ PorcientoNT+pHdeagua+pHdeKCl+
DeltadepH,
+ data = datos)
> summary(regresion1)

Call:
lm(formula = PorcientoCorg ~ PorcientoNT+pHdeagua+pHdeKCl+DeltadepH,
+ data = datos)

Residuals:
    Min          1Q        Median      3Q        Max
-35.272     -7.417     -1.538      4.133     32.998

Coefficients: (1 not defined because of singularities)
```

```
              Estimate     Std.      Error t   value Pr(>|t|)
(Intercept)    39.640      16.014      2.475     0.01706 *
PorcientoNT     3.041       2.503      1.215     0.23060
pHdeagua      -47.447      13.982     -3.393     0.00143 **
pHdeKCl        52.057      12.459      4.178     0.00013 ***
DeltadepH         NA          NA         NA         NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.29 on 46 degrees of freedom
Multiple R-squared:  0.4941,    Adjusted R-squared:  0.4611
F-statistic: 14.98 on 3 and 46 DF,  p-value: 6.164e-07
```

Analizando la salida del software R, vemos que el pHdeagua y pHdeKCl, son variables significativas y que DeltadepH no es considerada ya que, por definición, es la diferencia entre pHdeagua y pHdeKCl que está altamente correlacionada con ambas, así que el mejor modelo de efectos fijos nos queda de la siguiente manera.

```
> regresion3 <- lm(PorcientoCorg ~  pHdeagua + pHdeKCl, data = datos)
> summary(regresion3)

Call:
lm(formula = PorcientoCorg ~ pHdeagua + pHdeKCl, data = datos)
Residuals:
    Min      1Q     Median    3Q     Max
-35.937   -7.818   -1.924    5.748   31.774
Coefficients:
           Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  48.59      14.29      3.400   0.00138 **
pHdeagua    -55.61      12.33     -4.511   4.29e-05 ***
pHdeKCl      59.22      11.03      5.370   2.39e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.36 on 47 degrees of freedom
Multiple R-squared:  0.4779,    Adjusted R-squared:  0.4556
F-statistic: 21.51 on 2 and 47 DF,  p-value: 2.334e-07
```

Este modelo explica aproximadamente el 48 % de la variabilidad, pero la prueba F es significativa luego puede considerarse que se ajusta a los datos.

Como segundo aspecto de la metodología hay que encontrar la estructura de la componente aleatoria óptima. Para ello propondremos distintos modelos alternativos con la misma estructura en la componente fija pero que varían en su componente aleatoria. La comparación entre distintos modelos podemos hacerla utilizando la función anova(). Utilizamos la paquetería lme4, para ajustar los modelos con estimadores ($REML$). Realizamos distintos modelos para efectos aleatorios.

```
> library(lme4)
> regresion4 <- lmer(PorcientoCorg~ pHdeagua+pHdeKCl +(1|RegTemp ),
+ data = datos)
> regresion5 <- lmer(PorcientoCorg~ pHdeagua+pHdeKCl +(1|PERFIL),
+ data = datos)
>regresion6 <- lmer(PorcientoCorg~ pHdeagua+pHdeKCl +(1|RegTemp)+
(1|PERFIL), data = datos)
>regresion7 <- lmer(PorcientoCorg~ pHdeagua+pHdeKCl +
(1|RegTemp)*(1|PERFIL), data = datos)
```

Ajustando todos los modelos anteriores usandoLM y después comparándolos con la función anova(), seleccionamos el mejor modelo.

```
> regresion4.1 <- update(regresion4,REML=FALSE)
> regresion5.1 <- update(regresion5,REML=FALSE)
> regresion6.1 <- update(regresion6,REML=FALSE)
> regresion7.1 <- update(regresion7,REML=FALSE)
> anova(regresion4.1 ,regresion5.1,regresion6.1,regresion7.1)
Data: datos
Models:
regresion4.1: PorcientoCorg ~ pHdeagua + pHdeKCl + (1 | RegTemp)
regresion5.1: PorcientoCorg ~ pHdeagua + pHdeKCl + (1 | PERFIL)
regresion6.1: PorcientoCorg ~ pHdeagua + pHdeKCl + (1 | RegTemp)
 (1 | PERFIL)
regresion7.1: PorcientoCorg~pHdeagua+pHdeKCl+(1|RegTemp)*(1|PERFIL)
             Df   AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
regresion4.1  5 400.22 409.78 -195.11   390.22
regresion5.1  5 393.38 402.94 -191.69   383.38 6.8412  0  <2e-16 ***
regresion6.1  6 394.88 406.36 -191.44   382.88 0.4912  1   0.4834
regresion7.1  6 394.88 406.36 -191.44   382.88 0.0000  0   1.0000
```

Los procedimientos de selección de modelos utilizando criterios de información, permiten comparar muchos modelos, no necesariamente anidados, de forma simultánea. Tomando en cuenta el criterio de selección de modelo AIC, puede considerarse que el modelo regresión5.1 es el mejor, por tomar el valor mínimo en este criterio ($AIC = 393.38$); de igual manera el criterio de selección de modelos BIC señalan que el mejor modelo es el de regresion5.1. La estadística de prueba deviance nos ayuda a decidir que el mejor modelo es regresion5.1, pues es significativo.
Finalmente, se ajusta el modelo utilizando la función lme para después poder hacer los gráficos de los residuales, necesarios para el análisis de las suposiciones.

```
> regresion5lme <- lme(PorcientoCorg~  pHdeagua + pHdeKCl ,
data = datos, random=~1|factor(PERFIL))
>  summary(regresion5lme)
Linear mixed-effects model fit by REML
 Data: datos
       AIC           BIC        logLik
  375.9163     385.167        -182.9581
Random effects:
 Formula: ~1 | factor(PERFIL)
         (Intercept)        Residual
StdDev:    11.48457       10.19486
Fixed effects: PorcientoCorg ~ pHdeagua + pHdeKCl
              Value   Std.Error  DF   t-value   p-value
(Intercept) 18.13091  37.32212   43  0.4857952   0.6296
pHdeagua   -16.91178  16.49596   43 -1.0252074   0.3110
pHdeKCl     26.22723  15.75875   43  1.6642968   0.1033
 Correlation:
           (Intr)   pHdeag
pHdeagua  -0.315
pHdeKCl   -0.109    -0.907
Standardized Within-Group Residuals:
     Min          Q1           Med           Q3          Max
-4.386987759  -0.193212451 -0.009326507  0.166663151  2.209938652
Number of Observations: 50
Number of Groups: 5
```

Observamos que la salida del modelo regresion5lme muestra que las variables pHdeagua y pHdeKCl no son significativas con un nivel de $\alpha = .05$ según la prueba t; sin embargo, como seguimos la metodología de Zuur y en el primer paso las

tomamos como las variables fijas a considerar, estas deben mantenerse.
El mejor modelo mixto según Zuur quedo de la siguiente forma:

$$PorcientoCorg_{ij} = \beta_0 + \beta_1 pHdeagua_{ij} + \beta_2 pHdeKCl_{ij} + u_j + \varepsilon_{ij}. \qquad (3.3)$$

$$u_j \sim N(0, \sigma_u^2) \quad \varepsilon_{ij} \sim N(0, \sigma^2).$$

donde pHdeagua y pHdeKCl son las variables fijas y u_j corresponde al factor aleatorio "PERFILES", $j = 1, 2, 3, 4, 5$. que corresponde a los niveles (P30F, P36F, P36F, Piñonero y Yucca). Los parámetros desconocidos corresponden a β_0, β_1, β_2, la varianza de error σ^2, y σ_u^2. La salida anterior nos muestra el valor de las estimaciones de los parámetros desconocidos, corresponden a $\hat{\beta}_0$ =18.13091, $\hat{\beta}_1$ = -16.91178, $\hat{\beta}_2$ =26.22723, $\hat{\sigma}^2 = 10{,}19486^2$ y $\hat{\sigma}_u^2 = 11{,}48457^2$.

Finalmente este modelo nos muestra una relación entre el porcentaje de carbono orgánico que está en función de pHdeagua, pHdeKCL y los PERFILES, donde el porcentaje de carbono orgánico es el porcentaje en peso de materia orgánica , el pHdeagua y el pHdeKCL es una medida de acidez o alcalinidad de las muestras en agua y en solución de KCL. Esto ayuda al investigador a conocer una relación entre estas variables en todas la zonas de la Región Terrestre Prioritaria(RTP-105), Teziutlán, Puebla.

3.3.3. Revisión de las Suposiciones del Mejor Modelo

Para saber si el modelo es adecuado debemos analizar los residuales normalizados del modelo estimado a partir de *REML*.

```
> Res <- residuals(regresion5lme, type="normalized")
> Fit <- fitted(regresion5lme)
> plot(Res ~ Fit,xlab="Fitted values", ylab="Residuals",
 main="Residuals vs. fitted")
> abline(h=0)
```

Para analizar la normalidad usamos los siguientes comandos.

```
> hist(Res, main="Histogram of residuals", xlab="Residuals")
```

```
> qqnorm(Res)
> qqline(Res)
```

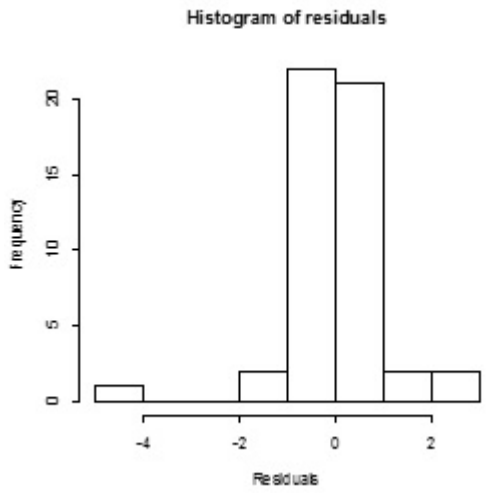

Figura 3.4: Histograma de normalidad

Los dos Gráficos (3.4) y (3.5) nos permiten analizar el supuesto de normalidad de los residuales. Existen dudas sobre la normalidad de los residuales; en los valores centrales están cercanos a la recta, pero sin embargo, los valores de los extremos se separan de la recta.

Como vemos en el Gráfico (3.6) los datos que manejamos en nuestro caso son datos agrupados. Al analizar el gráfico vemos que existen datos atípicos. Para corregir este problema se podría optar por un modelo no lineal y detectar cuales son los datos atípicos.

En este último Gráfico (3.7) se señala que el P36F muestra una gran variabilidad y mostrando un punto muy lejano cuyo valor corresponde a

$$PorcientodeCorg_{P36F} = 20.$$

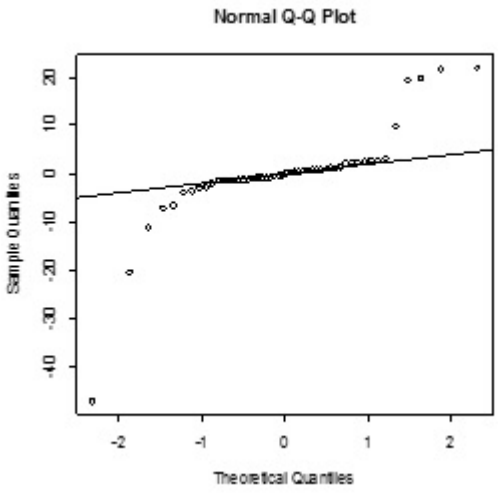

Figura 3.5: Normalidad

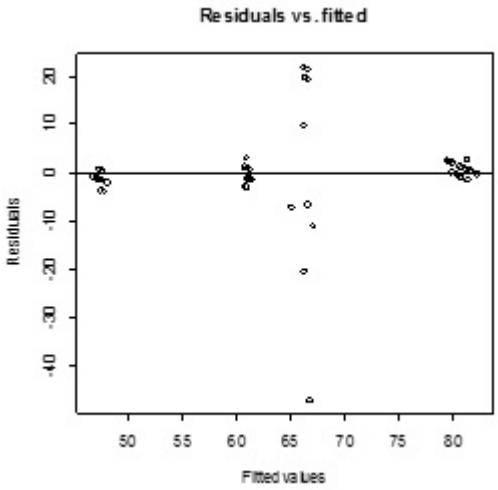

Figura 3.6: Homogeneidad, linealidad e independencia.

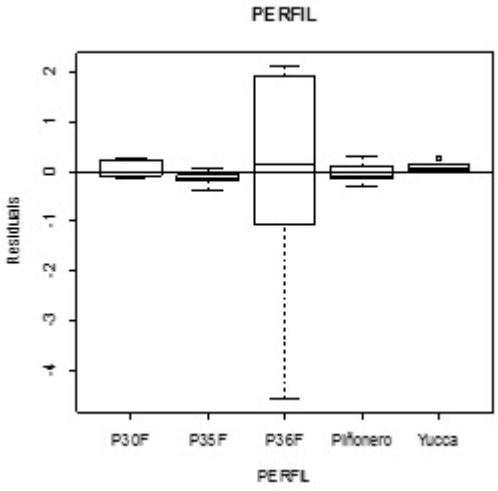

Figura 3.7: PERFILES vs Residuales.

Conclusiones

El modelo Mixto es un modelo muy flexible, que permite tener en su formulación tanto variables fijas como variables aleatorias. Decidir si en un problema deberían incorporarse efectos aleatorios, requiere un análisis estadístico exhaustivo. En la tesis describimos una metodología que nos ayuda a decidir si es conveniente o no trabajar con modelos mixtos.

Por otra parte, en presencia de problemas complejos con variables cuantitativas y cualitativas, seleccionar un buen modelo es un trabajo arduo. En la tesis también se describe la metodología propuesta por [Zuur, 2009] y se aplica al proponer un modelo que sea útil para explicar o predecir el porciento de carbono orgánico de la hojarasca en la región de Teziutlán, Puebla. En un primer intento, tomamos como mejor modelo, el modelo lineal mixto que está en función de pHdeagua y pHdeKCl como factores fijos y el perfil del suelo como factor aleatorio. Sin embargo, el análisis de las suposiciones a través de los gráficos de residuos sugiere que debiera seguir profundizándose. Se debiera profundizar en Modelos Lineales Mixtos Generalizados debido a que la variable de respuesta es un porcentaje u otra manera es adentrarnos en los modelos mixtos no lineales.

Los resultados de esta tesis son de suma importancia en la Región Terrestre Prioritaria (RTP- 105) para incrementar el nivel de conocimientos sobre la dinámica del carbono orgánico en la hojarasca, considerado uno de los principales reservorios de materia orgánica de la naturaleza. El modelo que se propone ayuda a conocer el porcentaje de carbono orgánico de la hojarasca a partir de las variables que mayor influyen en el incremento de este porcentaje: pH de agua y pH de KCl . Por otra parte, se constató que la modelación estadística puede dar resultados confiables sobre las direcciones y cantidades de cómo se comporta la materia orgánica en la hojarasca con las variables más influyentes en la zona de estudio y, puede recomendarse como enfoque metodológico para futuras investigaciones que se realicen para encontrar mecanismos de mitigación del cambio climático propuestos por el Protocolo de Kyoto.

Apéndice A

Software R

Existen distintos paquetes de software estadístico como : SPSS, Statgraphics, Statistica, Minitab, SAS, S-Plus, etc., que cubren todas las necesidades de cualquier usuario de técnicas estadísticas, básicas o avanzadas, sin embargo, R ha surgido con fuerza como alternativa de software libre en muy distintos ambientes docentes y de investigación.
R es un lenguaje de programación especialmente indicado para el análisis estadístico. A diferencia de la mayoría de los programas que solemos utilizar en nuestros ordenadores, que tienen interfaces tipo ventana, R es manejado a través de una consola en la que se introduce código propio de su lenguaje para obtener los resultados deseados.
El código de R está disponible como software libre bajo las condiciones de la licencia GNU-GPL, y puede ser instalado tanto en sistemas operativos tipo Windows como en Linux o MacOS X. La página principal desde la que se puede acceder tanto a los archivos necesarios para su instalación como al resto de recursos del proyecto R es http://www.r-project.org.
Después de instalar R lo primero que nos aparece es una ventana, también llamada consola, donde podemos manejar R mediante la introducción de código. Sin embargo, esta no es la manera más eficiente de trabajar en R, para ello debemos acceder a un documento en blanco del editor, llamado script. La utilidad de un script o guion de trabajo radica en que podemos modificar nuestras líneas de código con comodidad y guardarlas para el futuro. Parte de la vasta información disponible sobre R es accesible a través de la web CRAN (Comprehensive R Archive network; http://cran.r-project.org/) sitio oficial de R.
La bibliografía sobre R es amplia y muchas obras publicadas recientemente aumentan sus posibilidades de utilización. La obra de [Crawley, 2008] ilustra el proceso de modelado estadístico con R y ha sido de gran utilidad en el desarrollo de la presente Tesis.

A.1. Paqueterías nlme y lme4 del Sofware R

Las paqueterías nlme y lme4, pertenecen al software R, las cuales se utilizan para estimar los parámetros de los modelos lineales y modelos lineales mixtos.

paquetería lme4	se utiliza para Modelos Lineales, Lineal Generalizado, y los Modelos Mixtos no lineales
paquetería nlme	Modelos Lineales y Modelos de efectos mixtos no lineales

Descripción

paquetería lme4	Ofrece funciones para ajustar y analizar modelos lineales mixtos: (lmer) lineal (lmer), lineal generalizado (glmer) y no lineales (nlmer).
paquetería nlme	es más completo que lme4, ver más [Pinheiro, 2009].

Según [Bates, 2014] algunas de las diferencias entre nlme y lme4 son:

- lme4 cubre aproximadamente los mismos aspectos que el paquete nlme.
- lme4 utiliza métodos modernos y eficientes de álgebra lineal, por lo tanto es probable que sea más rápido y más eficiente que nlme
- lme4 incluye funciones para modelar modelos lineales mixtos generalizados (GLMM).
- lme4 actualmente no implementa características para el modelado de heterocedasticidad y correlación de los residuales.
- lme4 no ofrece actualmente la misma flexibilidad que nlme para componer estructuras complejas de varianza-covarianza, pero no implementa que los efectos aleatorios de una manera que sea a la vez más fácil para el usuario y mucho más rápido.
- lme4 está diseñado para ser más modular que nlme, por lo que es más fácil para los desarrolladores de paquetes y los usuarios finales para volver a utilizar sus componentes para modelo mixto. También permite una mayor flexibilidad para especificar diferentes funciones para la optimización en los parámetros de varianza-covarianza de efectos aleatorios.
- lme4 (aún) no está tan bien documentada como nlme.

Apéndice B

Hojarasca

La hojarasca constituye la vía de entrada principal de los nutrientes en el suelo y es uno de los puntos clave del reciclado de la materia orgánica y de los nutrientes. Se entiende por hojarasca la acumulación de los residuos vegetales (hojas, tallos, etc.) sobre la superficie del suelo y ello contribuye, de forma significativa, al flujo de los nutrientes y la energía, así como, a la constitución de las reservas húmicas del suelo.

La caída de la hojarasca representa el mayor proceso de transferencia de nutrientes de las partes aéreas de la planta hacia el suelo. La hojarasca que cae al suelo forma un estrato orgánico conocido como mantillo, el cual cubre el suelo y lo protege de los cambios de temperatura y de humedad, y también permite que retornen elementos nutritivos en una cantidad importante. Los residuos vegetales depositados (hojas, ramas, flores y frutos) son una fuente valiosa de materia orgánica que después de sufrir procesos de descomposición liberan elementos nutritivos que se incorporan al suelo para ser nuevamente utilizados por las plantas.

En la investigación llevada a cabo por [Castillo, 2014] se establecieron algunas mediciones sobre la hojarasca en la zona Teziutlán, Puebla, en el año 2009, con el propósito de contar con esta información para futuros trabajos que permitan un conocimiento más completo de la dinámica del carbono orgánico. La Figura 1 brinda la localización de la zona.

Varios autores han estudiado con detalle la dinámica de la descomposición de la hojarasca de las plantas leñosas, tanto en climas templados como en el mediterráneo. Sin embargo, hay pocos estudios sobre la dinámica de la descomposición de la hojarasca en los pastizales a pesar de su importancia en la producción primaria y secundaria, sobre todo en los sistemas donde los nutrientes disponibles para la vegetación escasean, como ocurre en los ecosistemas de pastizales [Saray, 2008].

Los principales factores que controlan la dinámica de la descomposición de la hojarasca son:

Efecto de la composición química de la hojarasca. La cantidad de material vegetal, su composición y sus propiedades son esenciales, dado que controlan los procesos de descomposición, mineralización y humificación y actúan como la fase de transición entre la biomasa viva y el suelo.

Efecto de los factores climáticos. Numerosos autores coinciden en señalar que los factores climáticos influyen en el proceso de descomposición de la hojarasca de las diferentes especies vegetales y, en especial, identifican a la temperatura y a las precipitaciones como los indicadores de mayor importancia.

Efecto de los organismos del suelo. El proceso de descomposición de la materia orgánica en los suelos del trópico es controlado por los factores biológicos. La descomposición que realizan los organismos se caracteriza por una compleja comunidad de biota, que incluye la microflora y la fauna del suelo. Los hongos y las bacterias son, fundamentalmente, los responsables de que se efectúen los procesos bioquímicos en la descomposición de los residuos orgánicos.

La base de datos con la que se desarrolló la presente tesis, se obtuvo por medio de los siguientes pasos:

Trabajo de campo: Consistió en la selección de los sitios de muestreo, así como, el muestreo de suelos y recolección de hojarasca. Se tomaron las muestras de hojarasca con un bastidor de aproximadamente 40 x 40 cm; dichas muestras se etiquetaron adecuadamente y fueron colocadas en bolsas de papel.

Preparación de las muestras de hojarasca para el análisis en el laboratorio: Incluyó las etapas de

- *Secado*
  Las muestras se secaron lo más pronto posible en una estufa a 60^oC (a temperatura constante) durante 24 a 48 horas.

- *Molienda*
  Este proceso se llevó a cabo para facilitar el manejo del material, además de homogeneizar su composición.

- *Tamizado*
  Este proceso se realizó para homogeneizar la composición del material.

Determinaciones de las propiedades químicas en las muestras de hojarasca: Se analizaron:

- pH (Relación 1:5)
  El potencial de hodrógeno mide la condición llamada acidez y se determinó por el método potencio métrico en agua y en solución de KCl 1N, en relación hojarasca/agua y hojarasca/solución de 1:5.

- *Porcentaje de materia orgánica*
  La diferencia de los pesos (peso seco inicial - peso seco después de la ignición) es igual a la cantidad de carbón orgánico incinerado. Es decir, el resultado es el porcentaje en peso de materia orgánica presente en la muestra. (Dean, 1974).
  Método por ignición Se eligió la técnica de pérdida de peso por ignición, toda vez que, este método acepta la combustión de la totalidad del material contenido en la muestra original. Esta técnica, modificada por Galle y Runiiels en 1960, propone que la incineración de la materia orgánica se alcanza a la temperatura de 550^oC. Normalmente cuando se calientan los sedimentos que contienen materia orgánica y carbonato de calcio en mufla, la materia orgánica se descompone alrededor de los 200^oC y se incinera por completo cuando alcanza la temperatura de 550^oC, aproximadamente.
  El desprendimiento de bióxido de carbono (CO_2) que se deriva del carbonato de calcio, empieza alrededor de 800^oC y termina a 850^oC. El peso que pierde la muestra durante el proceso (referido como porcentaje de materia orgánica total), puede calcularse de la diferencia del peso inicial menos el final. Es necesario tener presente que trabajar con esta técnica puede ser inconveniente. Cuando la muestra contiene cantidades significativas de arcillas, la mayoría de estas contienen cerca de 5 % de agua que se desprende cuando se alcanza la temperatura comprendida entre 550^o y 1000^oC, por lo que, en los resultados puede estar incluida el agua intersticial, así como el CO_2. La suposición de que la pérdida de peso por ignición representa el peso faltante en forma de CO_2 derivado del carbonato de calcio, puede ser un error, si se considera que la cantidad de agua es directamente proporcional a la cantidad de carbonato presente.
  Procedimiento:

  - Secar la cuarta parte de la muestra original a temperatura de 90 a 100^0C por una hora.
  - Moler de 6 a 10 gramos de sedimento seco en mortero de porcelana.

- Mantener en peso constante los crisoles de porcelana.
- Agregar 2 gramos de muestra en cada crisol y pesarlos con aproximación de 0.0001 de gramo. Esta es la medida del peso seco.
- Calentar la muestra a 550^{o}C por una hora.
- Dejar enfriar los crisoles en desecador provisto con sílice gel hasta alcanzar la temperatura ambiente. Pesarlos nuevamente (2da medida del peso seco).
  La diferencia de los pesos es igual a la cantidad de carbón orgánico incinerado. Es decir, el resultado es el porcentaje en peso de materia orgánica presente en la muestra [Dean, 1974].

- *% Nitrógeno Total.*
  Se determinó por el método Semimicro - Kjeldahl. Se colocó la muestra seca, molida y tamizada (malla 40), se agregó la mezcla de catalizadores, H_2SO_4 concentrado y se pusieron a calentar en la unidad digestora a temperatura media alta hasta que el digestado se tornó claro. Se procedió a ebullir la muestra por espacio de 4 a 6 horas a partir de este momento, regulando la temperatura para mantener los vapores de ácido sulfúrico condensados en el tercio inferior del cuello del tubo de digestión. Una vez completada esta etapa, se dejó el frasco y se agregó suficiente agua para colocar el digestado en suspensión, mediante agitación y se dejó decantar. Se destiló la solución digerida con agua destilada e NaOH - $Na_2S_2O_3$ y se recibió el destilado en un matraz Erlenmeyer que contenía ácido bórico e indicadores. El destilado se valoró con H_2SO_4 0.01 N.

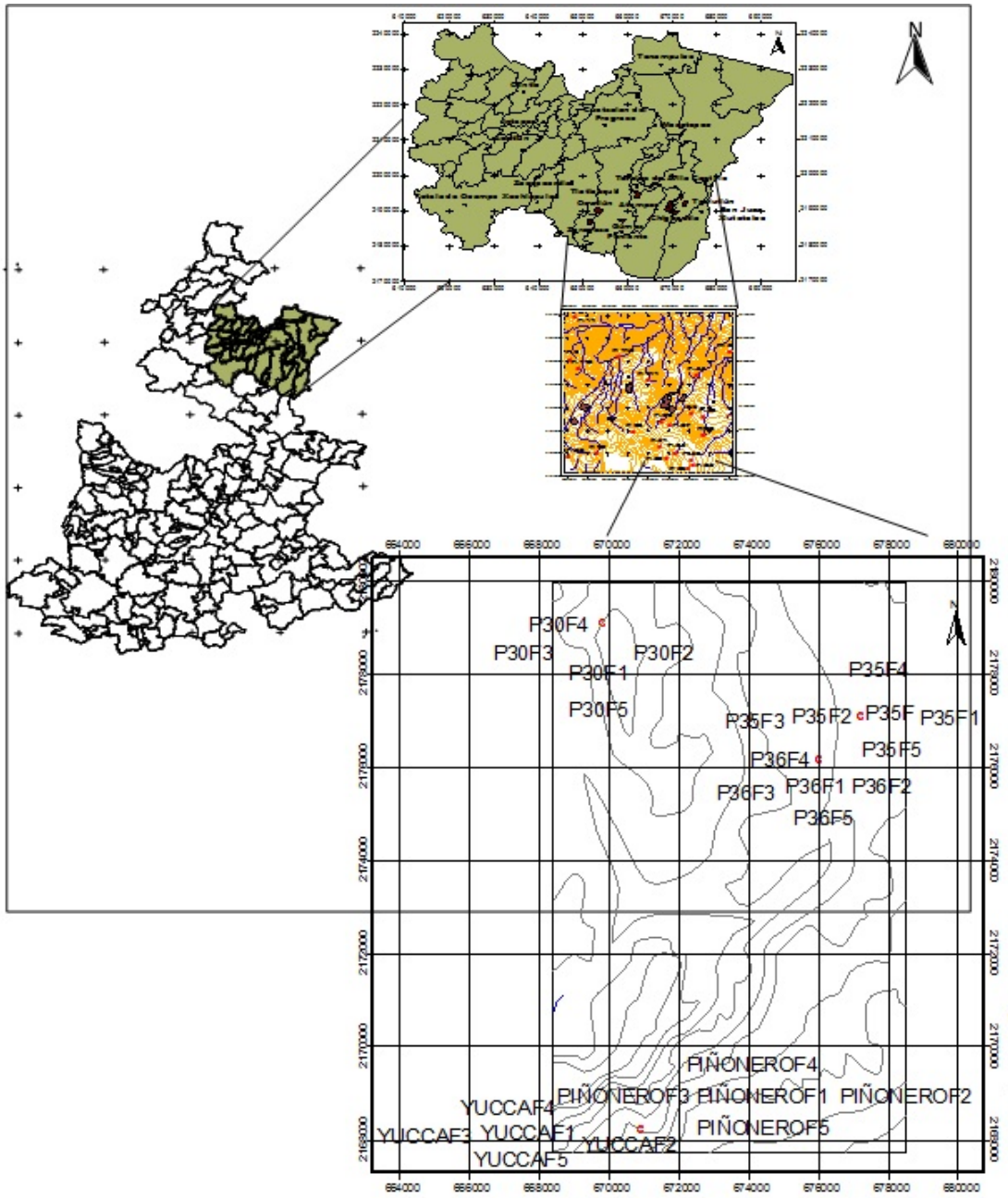

Figura B.1: Mapa de la localización de la zona de estudio ubicada en la Región Terrestre Prioritaria (RTP – 105), Teziutlán, Puebla.

B.1. Base de Datos

Las siguientes variables fueron obtenidas en la tesis doctoral [Castillo, 2014]. Los nombres de las variables utilizadas en la modelación son:
Porciento de Carbono Orgánico= PorcientoCorg
Porciento de Nitrógeno = PorcientoNT
pH de agua= pHdeagua
pH de Cloruro de potación= pHdeKCl
DeltadepH = Delta de Ph
Tipos de suelo= PERFILES
Régimenes de Temperatura=RegTemp

RegHumed	RegTemp	EstaciónClim	PERFIL	PorcientoNT	pHdeagua	pHdeKCl	DeltadepH	PorcientoCorg
Perúdico	Térmico	Teziutlán	P30F	2.61	5.3	5.43	0.13	82
Perúdico	Térmico	Teziutlán	P30F	2.55	5.33	5.41	0.08	82
Údico	Isomésico	Teziutlán	P30F	2.80	5.28	5.39	0.11	80
Údico	Isomésico	Teziutlán	P30F	2.83	5.33	5.41	0.08	80
Údico	Isomésico	Teziutlán	P30F	2.53	5.44	5.43	-0.01	82
Údico	Isomésico	Teziutlán	P30F	3.12	5.34	5.38	0.04	80
Perúdico	Térmico	Teziutlán	P30F	2.47	5.18	5.38	0.20	82
Perúdico	Térmico	Teziutlán	P30F	2.89	5.17	5.46	0.29	80
Perúdico	Térmico	Teziutlán	P30F	2.90	5.19	5.38	0.19	80
Perúdico	Térmico	Teziutlán	P30F	2.60	5.18	5.42	0.24	80
Perúdico	Térmico	Jalacingo	P35F	0.79	5.6	5.15	-0.45	46
Perúdico	Térmico	Jalacingo	P35F	0.93	5.43	5.09	-0.34	44
Perúdico	Térmico	Jalacingo	P35F	0.85	5.34	5.13	-0.21	44
Perúdico	Térmico	Jalacingo	P35F	0.58	5.32	5.08	-0.24	46
Údico	Isomésico	Jalacingo	P35F	1.04	5.29	5.09	-0.20	46
Údico	Isomésico	Jalacingo	P35F	0.88	5.29	5.09	-0.20	48
Perúdico	Térmico	Jalacingo	P35F	0.75	5.25	5.10	-0.15	46
Perúdico	Térmico	Jalacingo	P35F	0.61	5.29	5.08	-0.21	46
Perúdico	Térmico	Jalacingo	P35F	0.78	5.24	5.07	-0.17	48
Perúdico	Térmico	Jalacingo	P35F	0.84	5.24	5.07	-0.17	46
Údico	Isomésico	Jalacingo	P36F	0.62	4.73	4.57	-0.16	88
Údico	Isomésico	Jalacingo	P36F	0.77	4.68	4.55	-0.13	88
Perúdico	Isotérmico	Jalacingo	P36F	0.77	4.58	4.56	-0.02	86
Perúdico	Isotérmico	Jalacingo	P36F	0.68	4.61	4.55	-0.06	86
Údico	Isotérmico	Jalacingo	P36F	0.65	4.66	4.58	-0.08	46
Údico	Isotérmico	Jalacingo	P36F	0.83	4.65	4.49	-0.16	20
Údico	Isomésico	Jalacingo	P36F	0.76	4.66	4.52	-0.14	60
Údico	Isomésico	Jalacingo	P36F	0.64	4.69	4.52	-0.17	76
Údico	Isotérmico	Jalacingo	P36F	0.26	4.63	4.47	-0.16	58
Údico	Isotérmico	Jalacingo	P36F	0.94	4.65	4.53	-0.12	56
Perúdico	Isotérmico	Oyameles	Piñonero	1.01	4.92	4.68	-0.24	58
Perúdico	Isotérmico	Oyameles	Piñonero	1.00	4.77	4.68	-0.09	64
Údico	Isomésico	Oyameles	Piñonero	1.03	4.66	4.68	0.02	62
Údico	Isomésico	Oyameles	Piñonero	1.15	4.63	4.65	0.02	60
Údico	Isomésico	Oyameles	Piñonero	1.10	4.59	4.63	0.04	62
Údico	Isomésico	Oyameles	Piñonero	1.12	4.58	4.63	0.05	60
Údico	Isomésico	Oyameles	Piñonero	1.07	4.56	4.64	0.08	60
Údico	Isomésico	Oyameles	Piñonero	0.98	4.58	4.65	0.07	58
Údico	Isomésico	Oyameles	Piñonero	1.04	4.53	4.63	0.10	60
Údico	Isomésico	Oyameles	Piñonero	0.94	4.59	4.65	0.06	62
Údico	Isomésico	Oyameles	Yucca	1.02	6.64	6.65	0.01	82
Perúdico	Térmico	Oyameles	Yucca	0.78	6.62	6.7	0.08	82
Perúdico	Térmico	Oyameles	Yucca	0.81	6.54	6.67	0.13	82
Perúdico	Térmico	Oyameles	Yucca	0.67	6.55	6.82	0.27	82
Perúdico	Térmico	Oyameles	Yucca	0.66	6.58	6.85	0.27	82
Perúdico	Térmico	Oyameles	Yucca	0.70	6.55	6.77	0.22	84
Perúdico	Térmico	Oyameles	Yucca	0.69	6.6	6.70	0.10	84
Perúdico	Térmico	Oyameles	Yucca	0.66	6.57	6.79	0.22	82
Perúdico	Térmico	Oyameles	Yucca	0.49	6.62	6.76	0.14	82
Perúdico	Térmico	Oyameles	Yucca	0.54	6.55	6.76	0.21	82

Figura B.2: Base de Datos

B.2. Análisis Descriptivo de los Datos

En este apéndice se muestra un análisis descriptivo de la base de datos. Dicho análisis se llevó acabo para las variables porcentaje de carbono orgánico (porcientoCorg), Régimen de Temperatura (RegTem), porcentaje de nitrógeno (PorcientoNT, Potencial de Hidrógeno en agua (pHdeagua), Potencial de Hidrógeno en cloruro de potasio (pHdeKCl) y los tipos de suelo (PERFIL). La captura de datos y el análisis descriptivo se hace en el software R.

En la siguiente salida se muestra la captura de datos de las variables correspondientes al estudio.

```
> datos<-read.table(file="clipboard",head=T)
> attach(datos)
> names(datos)
 [1] "RegHumed"          "RegTemp"
 [4] "PERFIL"            "PorcientoNT"
 [7] "pHdeagua"          "pHdeKCl"
[10] "DeltadepH"         "PorcientoCorg"
```

A continuación se muestra estadística descriptiva para cada una de las variables.

```
> summary(RegHumed)
Perúdico    Údico
      27       23

> summary(RegTemp)
 Isomésico Isotérmico    Térmico
        19          8         23

 > summary(PERFIL)
    P30F     P35F     P36F Piñonero    Yucca
      10       10       10       10       10

> summary(PorcientoNT)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2600  0.6925  0.8650  1.1950  1.0920  3.1200

> summary(pHdeagua)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.530   4.660   5.240   5.296   5.408   6.640
```

```
> summary(pHdeKCl)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.470   4.632   5.090   5.287   5.428   6.850

> summary(DeltadepH)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.4500 -0.1600  0.0200 -0.0086  0.1075  0.2900

> summary(PorcientoCorg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   56.50   70.00   67.24   82.00   88.00
```

En la salida anterior del software R se observa que la variable categórica RegHumed tiene dos niveles que corresponden a perúdico y údico, mientras que PERFIL tiene 5 niveles. La variable continua PorcientoCorg contiene valores de porcentaje entre 20 y 80.
Como la variable de interés es el PorcientoCorg, se realizan gráficas de puntos que ayuda a ver la relación de esta variable con las demás. En la Gráfica (B.3)

```
> plot(PorcientoCorg~PorcientoNT)
```

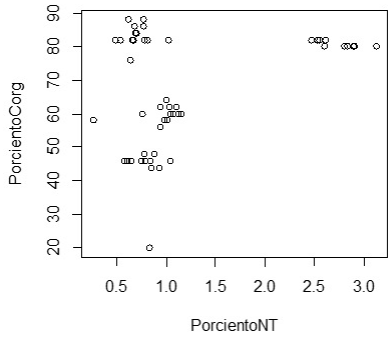

Figura B.3: PorcientodeCorg vs PorcientoNT

se observa que el porcientodeNT tiene comportamientos particulares, es decir que varía la cantidad de porcentajedeCorg, pero esta se ve reflejada por grupos, pues se ve que un grupo tiene cantidad de PorcentajedeCorg que varía de 40 a 50.
Ahora se muestran las respectivas gráficas de cajas para cada una de las variables

cualitativas. En la Gráfica (B.4) se observa que la mediana de los datos corres-

```
> boxplot(PorcientoNT)
```

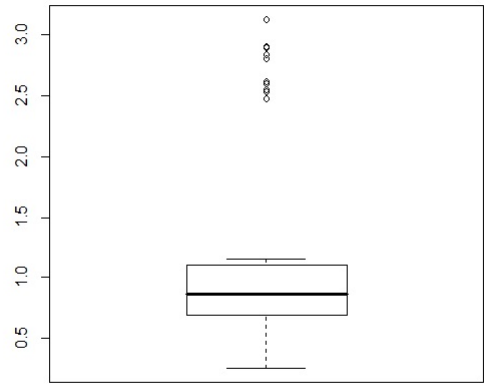

Figura B.4: Gráfica de cajas del porcentaje de Nitrógeno

ponde a .8650, su promedio es 1.1950, pero lo que hay que resaltar es que existen datos atípicos, los cuales se encuentran en la parte superior de la caja. A continuación se muestran los resultados referentes a la cantidad de carbono orgánico en cada uno de los niveles de las respectivas variables categóricas. En la Gráfica (B.8) podemos observar que la cantidad de carbono orgánico varía en cada uno de los tipos de suelo (PERFIL), pues las medianas de cada PERFIL varían de 54 a 80. Respecto al carbono orgánico en los diferentes Regímenes de temperatura, en la Gráfica (B.9) podemos observar que el carbono orgánico varía tanto del isomésico, isotérmico, y del Térmico. Este tipo de análisis descriptivo, nos ayuda a conocer las características de cada una de las variables.

```
> boxplot(pHdeagua)
```

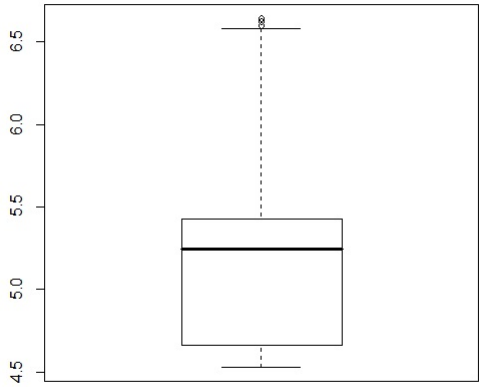

Figura B.5: Gráfica de cajas del pH de agua

```
> boxplot(pHdeKCl)
```

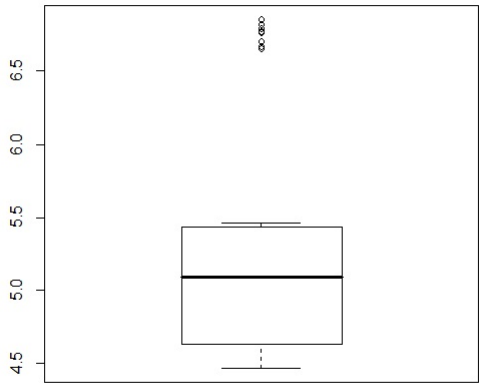

Figura B.6: Gráfica de cajas del pHdeKCl

```
> boxplot(PorcientoCorg)
```

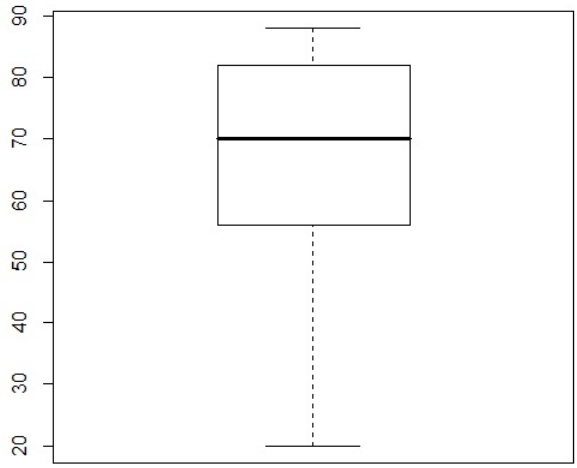

Figura B.7: Gráfica de cajas del porcentaje de carbono orgánico

```
> plot(PorcientoCorg~PERFIL)
```

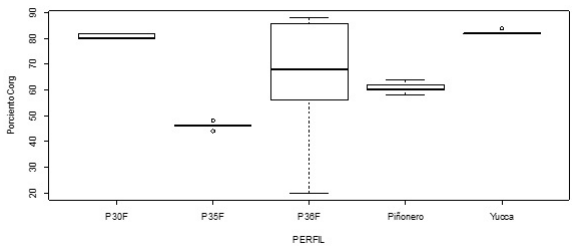

Figura B.8: Gráfica de cajas de los PERFILES

```
> plot(PorcientoCorg~RegTemp)
```

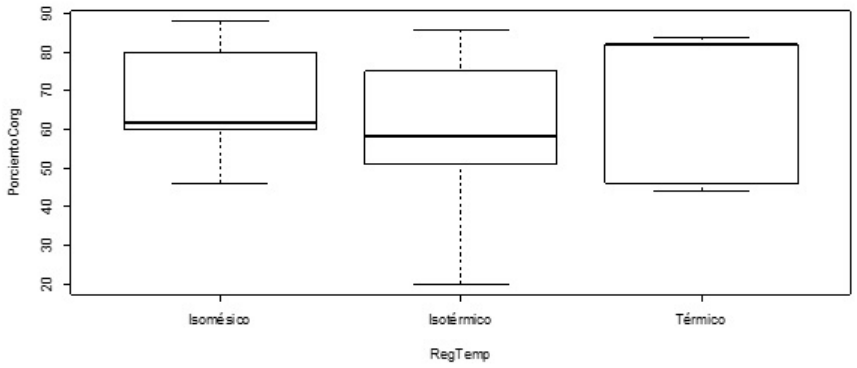

Figura B.9: cantidad de carbono orgánico respecto a los Regímenes de Temperatura

```
> plot(PorcientoCorg~RegHumed)
```

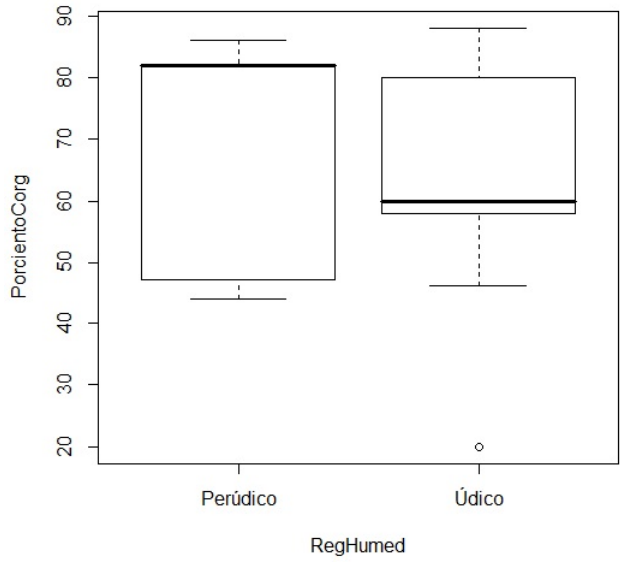

Figura B.10: Regímen de Humedad

Bibliografía

[Bates, 2014] Bates D., Bolker B., Dai B., Haubo R., Maechler M., Singmann H., Walker S. "lme4: Linear mixed-effects models using Eigen and S4". URL http://lme4.r-forge.r-project.org/. Versión 1.1-7, pág 3,4. (2014).

[Castillo, 2014] Castillo M. "Medición de la variabilidad especial y temporal del secuestro de carbono en suelos forestales de la Sierra Norte de Puebla". Tesis de Doctorado en Ciencias Ambientales. Posgrado en Ciencias Ambientales. Instituto de Ciencias. Benemérita Universidad Autónoma de Puebla. México.(2014).

[Cayuela, 2012] Cayuela L. "Modelos lineales mixtos (LMM) y modelos lineales generalizados mixtos (GLMM) en R". Aréa de Biodiversidad y Conservación, Universidad Rey Juan Carlos, versión 2.(2012).

[Crawley, 2008] Crawley J. M. "The R Book". John Wiley and Sons. ISBN-13: 978-0-470-51024-7. pág 449,459,489 y 629.(2007).

[Eisenhart, 1947] Eisenhart C. "The assumptions underlying the analysis of variance- Biometrics". pág 3,1–21.(1947).

[Galecki, 2013] Burzykowski T., Galecki A. "Linear Mixed-Effects Models Using R. A Step - by- Approach". Springer. ISBN-978-1.4614-3900-4. pág 245.(2013).

[Masera, 2004] Masera O., Sheinbaum C. "Mitigación de emisiones de carbono y prioridades de desarrollo nacional". En: Cambio climático: una visión desde México. Compiladores: Martínez J. y Fernández A. Instituto Nacional de Ecología. Secretaría del Medio Ambiente y Recursos Naturales. México, pág. 255-368.(2004).

[McCulloch, 2001] McCulloch E. C., Searle R. S., "Generalized, Linear and Mixed Models". JHON WILEY & SONS, ISBN 0-471-19364. pág 156. (2001).

[Mehtätalo, 2013] Mehtätalo L. "Linear mixed- effects models with exmaples in R". University of Eastern Finland, School of Computing. pág 13-20,32,61-75.(2013).

[Montgomery, 2011] Montgomery C. D., Peck A. E., Vining G.G. "Introducción al Análisis de Regresión Lineal". Patria. ISBN 978-970-24-0327-2. 3ra edición.(2011).

[Oroza, 2012] Linares F. G., Oroza H. A. "Modelos Mixtos en el estudio del Carbono Orgánico de la Hojarasca en una zona de Teziutlán, Puebla", BUAP. En memorias:"5ta Semana Internacional de la Estadística y la Probabilidad Facultad de Ciencias Físico Matemáticas". (2012).

[Peña, 2002] Peña D. "Análisis de Datos Multivariantes". McGRAW-HILL. ISBN: 84-481-3610-1. (2002).

[Pinnheiro, 2000] Bates M. D., Pinheiro C. J. " Mixed-effects Models in S and S-PLUS". New York:Springer-Verlag.(2000).

[Pinheiro, 2009] Pinheiro J.C., Bates D.M. "Linear and Nonlinear Mixed Effects Models". New York:Springer-Verlag.(2009).

[R, 2007] "R Development Core Team. R: A language and environmental for statistical computing". Vienna. Austria. R Foundation for Statistical Computing.(2007).

[Rao, 2007] Heumann, C., Rao R. C., Shalab, Toutenburg H. " Linear Models and Generalizations: Least Squares and Alternatives". Springer,ISBN- 978-3-540-74226-5. Tercera edicción. (2008).

[Saray, 2008] Crespo G. , Hernández G.,Sánchez S. García. Factores Bioticos y Abioticos que influyen en la descomposición de la hojarasca en pastizales.Pastos y Forrajes, Vol. 31, No. 2, pág 99- 118.(2008).

[Sheather, 2009] Sheather J. S. "A Modern Approach to Regression with R". Springer. ISBN 978-0-387-09607-0. (2009).

[Wackerly, 2010] Mendenhall III W., Scheaffer R L., Wackerly D." Estadística Matemática con aplicaciones". 7ma. Edición. CENGAGE Learning. México. pág 915.(2010).

[West, 2007] Gatecki A. T., Welch K. B., West B. T. "Linear Mixed Models. A Practical Guide Using Statistical Software". Chapman & Hall. ISBN 978-1-58488-480-4. (2007).

[Zuur, 2009] Ieno N. E., Saveliev A. A., Smith M. G., Walker J. N., Zuur F. A.“ Mixed Effects Models and Extensions in Ecology with R”. Springer. ISBN 978-0-387-87457-9. (2009).

Internet:

[1] http://CRAN.R-project.org/package=lme4