HUMANIDADES, CIENCIA, TECNOLOGÍA E INNOVACIÓN EN PUEBLA

ACADEMIA JOURNALS
OPUS PRO SCIENTIA ET STUDIUM

ISSN 2644-0903 online
VOL. 2, NO. 1, 2020
WWW.ACADEMIAJOURNALS.COM

TRABAJO DE INVESTIGACIÓN AUSPICIADO POR EL CONVENIO CONCYTEP-ACADEMIA JOURNALS

ANA ALEYDA OROZA HERNÁNDEZ

MODELOS MIXTOS EN EL ANÁLSIS ESTADÍSTICO DE IMÁGENES

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

SUPERVISOR PRINCIPAL:
DRA. GLADYS LINARES FLEITES (DICA)
COMITÉ SUPERVISOR:
DRA. HORTENSIA JOSEFINA REYES CERVANTES (FCFM)
DR. BULMARO JUÁREZ HERNÁNDEZ (FCFM)

NÚMERO DE SECUENCIA 2-51

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

Facultad de Ciencias Físico Matemáticas
Posgrado en Ciencias Matemáticas

MODELOS MIXTOS EN EL ANÁLSIS ESTADÍSTICO DE IMÁGENES

Tesis que para obtener el grado de

DOCTOR EN CIENCIAS MATEMÁTICAS

PRESENTA:
M.C. ANA ALEYDA OROZA HERNÁNDEZ

Supervisor principal:
Dra. Gladys Linares Fleites (DICA)

Comité supervisor:
Dra. Hortensia Josefina Reyes Cervantes (FCFM)
Dr. Bulmaro Juárez Hernández (FCFM)

13 Diciembre 2019

Modelos mixtos en el análisis estadístico de imágenes

Oroza Hernández Ana Aleyda

Resumen: Se desarrollan aplicaciones donde se requiere la utilización de la modelación estadística en el tratamiento de imágenes no satelitales y satelitales. La primera con imágenes del pez cebra usando la prueba Kolmogorv Smirnov para comparar imágenes del pez sometido a distintas soluciones de aguas residuales textiles para ver los cambios físicos-morfológicos que ha sufrido el pez cebra.

Los estudios con imágenes satelitales son realizados en dos zonas geográficas del estado de Puebla, México. Una, es la zona sur de la presa de Valsequillo y, la otra, es la Región Terrestre Prioritaria (RTP) 105 En ambas zonas se desarrollan, primero los elementos de localización y uso de las imágenes satelitales y, segundo, se discuten diferentes modelos lineales mixtos generalizados que ayudan a determinar la cantidad de carbono orgánico . En la modelación se desarrollan los problemas de estimación, bondad de ajuste diagnóstico de las suposiciones y selección de modelos utilizando el entorno R..

Dedicatoria

A mis padres

Mauro Oroza y Lucia Hernández

Agradecimientos

Agradezco de manera muy especial a mis padres, Lucia y Mauro por quererme mucho.

A mi hermano Edgar, por ser un compañero de aventuras y proporcionarme otro tipo de felicidad por medio de mis sobrinos maravillosos.

Agradezco a mis directoras de tesis, la Dra. Gladys Lineares Fleites, Dra. Hortensia Josefina Reyes Cervantes y Dr. Bulmaro Juárez Hernández, por el tiempo y el apoyo que me brindaron durante la realización de la presente tesis.

A mis sinodales, Dr. Hugo Adán Cruz Suárez, Dra. María de Lourdes Sandoval Solís, Dr. Víctor Hugo Vázquez Guevara y Dra. Sara Rodríguez Rodríguez, por el tiempo invertido en la revisión del presente trabajo y por las valiosas observaciones hechas al mismo.

Al Consejo Nacional de Ciencia y Tecnología por la beca otorgada para la realización de mis estudios de doctorado.

Ana Aleyda Oroza Hernández
FCFM-BUAP
Puebla, Puebla. 2019

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 (LM) y del Modelo Lineal Generalizado (GLM) que es una extensión natural del LM [29] y [30]. El GLM ha llegado a suponer "una auténtica revolución estadística" [3], convirtiéndose en una solución especialmente adecuada para modelos de dependencia con datos no métricos. También, bajo el supuesto de normalidad de los errores, pero permitiendo la heteroscedasticidad de la varianza, ha habido considerables trabajos sobre los Modelos Lineales Mixtos (LMM), donde la estructura de la varianza está basada sobre efectos aleatorios [17]. Más recientemente, se han introducido los Modelos Lineales Generalizados Mixtos (GLMM), que constituyen una fusión entre el LMM y el GLM, y representan una herramienta analítica poderosa, ya que permiten considerar diferentes distribuciones de errores, derivadas del ajuste de los modelos al ser analizadas, a la vez que permiten la inclusión de observaciones no independientes [10]. Dentro de estos modelos se considera un análisis adecuado cuando las observaciones son independientes pero las suposiciones sobre la Normalidad de la variable dependiente (y los errores) son incumplidas.
Por otro lado, dentro de la búsqueda de información para modelar, se están ideando nuevas metodologías de investigación y se están utilizando nuevas herramientas. Los instrumentos de teledetección son útiles ya que permiten obtener información, en particular sobre la vegetación de alguna zona de interés. Esa información se puede relacionar posteriormente con las mediciones tradicionales *in situ* (estudios que realizan en el mismo lugar donde se encuentra el objeto de análisis). Hace diez años se necesitaba equipamiento especial para tomar imágenes digitales y cuantificarlas, actualmente el procesamiento de imágenes es una disciplina bien establecida y existen diversos textos sobre la materia [17]. El procesamiento de imágenes trata sobre el mejoramiento, restauración y clasificación de la imagen. En la actualidad todavía hay que admitir que el análisis estadístico y comparaciones entre imágenes no son completamente satisfactorios. La idea es distinguir dos tipos de variaciones en las imágenes, variacion dentro de la imagen (para cada imagen) y variación entre las imágenes, conduce inmediatamente a los modelos mixtos, de manera que estos modelos se constituyen en la clave del análisis estadístico de las imágenes digitales. Para ejemplificar lo anterior imágenes del mismo suelo en una región geográfica tomadas en diferentes localizaciones o en diferentes tiempos pueden verse como mediciones repetidas y, por lo tanto, pueden analizarse con la metodología estadística relevante de los modelos mixtos para medidas repetidas.

Actualmente el procesamiento de imágenes digitales ha tomado especial importancia en los estudios relacionados con el cambio climático, ya que como es ampliamente conocido, existen evidencias científicas que sugieren que el clima global se verá altera-

do en este siglo, como en [12] y [60].
Las estimaciones espaciales de la biomasa y el volumen de los árboles son cada vez más importantes para el manejo y la planificación forestal [32]. La mayor incertidumbre para entender el papel de los bosques tropicales en el ciclo del Carbono está relacionada con la biomasa forestal [21] y [24]. Dentro de los estudios más actuales sobre modelación estadística dentro de la agronomía usando los modelos LMM y GLMM, podemos ver el trabajo realizado por Monterubbianesi en [28], mientras que una guía para el modelado con GLMM podemos citar a [7].
Los datos aplicados para el manejo de la biomasa forestal se pueden dividir en dos tipos diferentes. Uno es datos de observación de campo [55] y el segundo tipo de datos se adquiere mediante sensores remotos. Otro uso es la comparación ya que los datos de detección remota pueden cubrir grandes áreas y la información forestal se puede extraer fácilmente en poco tiempo con el software Sistema de Información Geográfica (GIS por sus siglas en inglés), por lo que es eficiente en la recolección y procesamiento de datos. Estudios más actuales que trabajan información de masa forestal e imágenes de satélite podemos citar a [12], [52] y [60].

Hasta donde conocemos son pocas las investigaciones dedicadas a estudiar el porcentaje de Carbono orgánico almacenado en zonas terrestres prioritarias en la República Mexicana. En particular en la zona RTP-105, proponiendo buenos modelos que se ajusten y cumplan con los supuestos de los modelos planteados. Cabe mencionar que los modelos que se ajustan en base a información real se buscan que cumplan la mayor parte se los supuestos. Mostrando que al trabajar con datos reales es una parte dificil al momento de ajustar modelos que cumplan la mayoría de los supuestos.
Dentro de la modelación estadística el Software R es uno de los paquetes estadísticos más potentes y flexibles que permite a los usuarios aplicar muchas técnicas estadísticas que en otros software sería imposible [15], además de ser gratuito, es fácil de entender y usarlo para realizar investigaciones, en esta tesis se usa el software R para el ajuste de modelos lineales, mixtos y generalizados; en particular se utiliza para la modelación el entorno R [54]. Entre los modelos que se ajustan están los Modelos Lineales Generales, Modelos Lineales Mixtos y Modelos Lineales Generalizados Mixtos mejorando el ajuste dentro de cada una de las aplicaciones.

A continuación, se presentan los objetivos de la tesis.

OBJETIVO GENERAL

Establecer los fundamentos del análisis estadístico de las imágenes usando las técnicas de modelación de efectos mixtos para medidas repetidas y, en dependencia del nivel de complejidad de las imágenes, desarrollar modelos lineales de efectos mixtos y modelos lineales generalizados de efectos mixtos.

OBJETIVOS ESPECÍFICOS

- Profundizar en la teoría de los modelos mixtos y sus aspectos estadísticos, en lo que se refiere, en particular, a la metodología de medidas repetidas y cómo aplicarla al análisis estadístico de imágenes.
- Desarrollar los problemas de estimación, bondad de ajuste, diagnóstico de las suposiciones y selección de modelos en la modelación estadística.
- Utilizar el entorno R en las aplicaciones planeadas, investigando diferentes comandos y paqueterías disponibles de modelos de efectos mixtos.
- Desarrollar aplicaciones donde se requiere la utilización de la modelación estadística en el tratamiento de imágenes no satelitales y satelitales.

La tesis mantiene la siguiente estructura: En el **Capítulo 1** se tratan aspectos de los modelos estadísticos básicos que permiten tener una base de la modelación estadística. En el **Capítulo 2** se describen modelos más flexibles como Modelos Lineal Generalizado (GLM) y Modelo Lineal Generalizado Mixto (GLMM), los cuales permiten otro tipo de distribución en la variable respuesta. En el **Capítulo 3** se presenta una aplicación con imágenes del pez cebra, usando la prueba de Kolmogorov-Smirnov, para comparar imágenes. El **Capítulo 4** se desarrollan los aspectos conceptuales sobre teledetección y cómo estos se pueden utilizar en la modelación estadística. Se desarrollan aplicaciones en dos zonas geográficas del estado de Puebla, México. Una, es la zona sur de la presa de Valsequillo y, la otra, es la Región Terrestre Prioritaria (RTP) 105. En ambas zonas se desarrollan, primero, los elementos de localización y uso de las imágenes satelitales y, segundo, se discuten diferentes modelos lineales mixtos generalizados, según los objetivos que se planteen en los estudios particulares. Parte de los resultados presentados en esta tesis fueron publicados en [37] y [38].

Índice

Capítulo 1

La modelación estadística

En este capítulo se describen de manera detallada los Modelos Lineales Generales, métodos de estimación, Bondad de ajuste, diagnósticos de las suposiciones y la selección de los modelos.

1.1. Modelo Lineal General (LM)

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*) [15] y [30]. Los Modelos Lineales Generales (LM) se basan en una serie de supuestos, algunos de los cuales pueden y deben comprobarse una vez ajustado el modelo. Estos son:

(i) Independencia en las observaciones y_i, ya que los errores del modelo aleatorio se suponen independientes entre sí.

(ii) Linealidad, esto es, se considera que la variable respuesta es lineal con respecto a la(s) variable(s) explicativa(s).

(iii) Normalidad de los errores aletaorio, puesto que se supone que los errores siguen una distribución normal con $\mu = 0$ y σ^2 constante, esto se denota como:

$$\epsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, \ldots, n. \tag{1.1}$$

(iv) Homocedasticidad, esto es, las varianzas deben ser homogéneas en los diferentes niveles del factor, y se denota como σ^2 constante.

En esta sección se presentan los métodos de estimación y pruebas de hipótesis para estos modelos.

1.1.1. Establecimiento del modelo LM

Modelo de Regresión Lineal

Sea Y la variable respuesta (variable dependiente) que está relacionada con las p variables explicativas (variables independientes) $X_1, X_2, \ldots, X_p$ por una función f. Tanto la variable respuesta como las explicativas son continuas y debido a que esta relación no es exacta, se escribe de la siguiente forma.

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon \tag{1.2}$$

donde ϵ es un error aleatorio.
Las Y y las X's se observan sobre n individuos. Cuando f es lineal, la ecuación (1.2) observada en un individuo, se escribe como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \tag{1.3}$$

$$i = 1, 2, \ldots, n.$$

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 (1.3) en forma de matricial, tenemos que

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

con $i = 1, 2, ..., n.$ y $j = 0, 1, ..., p.$
o en forma resumida

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.4}$$

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

$$E(\boldsymbol{\epsilon}) = 0$$
$$Var(\boldsymbol{\epsilon}) = \sigma^2 I,$$

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 función para llevar a cabo una regresión lineal con el Software R es lm (que significa "modelo lineal"), en [15] se detalla el uso de esa función, donde se especifica como es su estructura.

Análisis de Varianza (*ANOVA*)

En la modelación estadística se desea conocer el efecto de una o más variables explicativas sobre una respuesta (continua). Las variables explicativas 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) [15].
Cuando la(s) variable(s) explicativa(s) 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 modelo de un solo factor es:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon, \tag{1.5}$$

que se puede escribir como:

$$Y = \mu + \beta factor + \epsilon. \tag{1.6}$$

Los valores tanto de Y como del factor se observan sobre n individuos.
El $ANOVA$ implica el cálculo de la variación total en la variable respuesta y la partición de ella en componentes informativos [15]. En el caso más simple, dividimos la variación total (SSY) en sólo dos componentes: la variación explicada (SSR) y la variación no explicada (SSE):

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

donde $SSY = \sum (y - \bar{y})^2$. El Cuadro 1.1 brinda más detalle.

En el análisis de los grados de libertad, que por sus siglas en inglés se escribe df, se consideran que se tienen 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 pierden 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 $km - k = k(m - 1)$.

El cálculo de las sumas de cuadrados se presenta de manera tradicional como el Cuadro 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

Cuadro 1.1: Sumas de Cuadrados 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, SSR, es la suma de los cuadrados de las diferencias entre el tratamiento individual, $\bar{y}_i$ y la media general, $\bar{\bar{y}}$

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

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 respuesta $\sum_{i=1}^{k} \sum_{j=1}^{m} y_{ij} = \sum y$; 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 se tiene

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

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

Elaborado por Montgomery, Peck and Vinning en [33].

varianza), el estadístico F y el valor p asociado con el valor F. Los cuadrados medios se obtienen simplemente dividiendo cada suma de los cuadrados por 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).

Cuadro 1.2: ANOVA.

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

Análisis de Covarianza (*ANCOVA*)

El análisis de covarianza (*ANCOVA*) combina elementos de la regresión y el análisis de varianza. La variable de respuesta es continua, y hay al menos una variable explicativa continua (covariables) y con al menos una variable explicativa categórica. Según [15] el *ANCOVA* se resumen con las siguientes relaciones:

- Colocar dos o más regresiones lineales de $\boldsymbol{Y}$ contra $\boldsymbol{X}$ (uno 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.

Los modelos antes descritos, se pueden resumir en el Cuadro 1.3.

Cuadro 1.3: Variables explicatorias de los modelos de Regresión, ANOVA y ANCOVA.

MODELO	VARIABLE RESPUESTA	VARIABLES EXPLICATIVAS
Regresión	Continua	Continua
ANOVA	Continua	Categórica
ANCOVA	Continua	Continuas y Categóricas

1.1.2. Métodos de estimación

A continuación se explican distintos métodos de estimación que ayudan a estimar los parámetros $\boldsymbol{\beta}$ y σ^2 de los modelos (1.4) y (1.5). En el contexto de los LM clásicos, el método de estimación más común es el método de Mínimos Cuadrados Ordinarios.

Mínimos Cuadrados Ordinarios (OLS)

Los valores de $\boldsymbol{\beta}$ son desconocidos, pero se pueden estimar, utilizando los datos de la muestra. Para estimarlos se usa el método **"Mínimos Cuadrados Ordinarios"** [33], 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 residuo 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:

$$e_{(n\times 1)} \quad = \quad \mathbf{Y} - \hat{\mathbf{Y}}. \tag{1.7}$$

La suma de cuadrados de los residuos es:

$$\begin{aligned} S(\hat{\boldsymbol{\beta}}) &= \sum_{i=1}^{n} e_i^2 \\ &= \boldsymbol{e}'\boldsymbol{e} \\ &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y}'\mathbf{Y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}. \end{aligned} \tag{1.8}$$

Observemos que $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$ es una matriz de dimensión 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

$$\frac{\partial S}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0, \tag{1.9}$$

que se simplifica en

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}. \tag{1.10}$$

El sistema lineal de (1.10) se denominan **ecuaciones normales de mínimos cuadrados** [45]. Para resolver las ecuaciones normales se multiplican ambos lados de

(1.10) 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

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{1.11}$$

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

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_{OLS}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned} \tag{1.12}$$

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}_{OLS}) &= Cov[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Cov[\mathbf{Y}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2 I(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{1.13}$$

Se puede demostrar que el estimador *OLS* es el mejor estimador lineal insesgado, que en inglés se escribe Best Linear Unbiased Estimaton (*BLUE*) de los parámetros del modelo (1.4). 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, [45]). Sin embargo, esto es sólo cuando los supuestos (varianza constante y no correlación) en los residuos se mantengan.

Máxima Verosimilitud (ML)

El método de Máxima Verosimilitud, que en inglés se escribe Maximum Likelihood (*ML*) es un método alternativo para estimar los parámetros en (1.4), suponiendo que los errores son independientes e idénticamente distribuidos según la normal con varianza constante igual a σ^2, esto es $N(0, \sigma^2 I)$ [33] y [18].
El método *ML* inicia con la función de densidad de los errores

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

La función de verosimilitud es:

$$L(\epsilon, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f(\epsilon_i) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\epsilon_i'\epsilon}, \tag{1.14}$$

de (1.4) tenemos que $\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$. Así se transforma en

$$L(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}. \tag{1.15}$$

Ahora la función log-verosimilitud es:

$$\begin{aligned} l &= \log L(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \sigma^2). \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{1.16}$$

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 ML de β bajo los errores normales equivale al estimador de mínimos cuadrados $\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ [45]. El estimador de máxima verosimilitud de σ^2 es

$$\hat{\sigma}^2_{ML} = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})}{n}. \tag{1.17}$$

donde $\hat{\sigma}^2_{ML}$ es un estimador sesgado, pero asintóticamente insesgado (consistencia, ver más en Apéndice A.1).

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.
El *REML*, consiste en aplicar el método de máxima verosimilitud a un vector $K'\boldsymbol{Y}$ en vez de aplicarlo al 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 $n-p$ [31]. 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 la 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 $i = 1, 2, 3, \ldots, n-p$. En [51] hay un resultado matricial 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 de modelo (la matriz $\mathbf{X}$ tiene rango completo). Por lo tanto, solo necesitamos encontrar un número máximo de tales vectores linealmente independientes y apilarlos en la matriz K. Una forma de encontrar al vector k es utilizar la propuesta en [30].

$$k' = c'[I - XX^{-}], \tag{1.18}$$

donde c' es un vector arbitrario de longitud n y X^{-} es una matriz inversa generalizada de $\mathbf{X}$.
Hay que tener en cuenta que puede haber varios valores posibles de X^{-}. Sin embargo, las estimaciones 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, se obtiene

$$K'\boldsymbol{Y} = K'\boldsymbol{X\beta} + K'\boldsymbol{\epsilon}. \tag{1.19}$$

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 $\mathbf{Y} \sim N(\mathbf{X\beta}, V)$,
entonces

$$K'\mathbf{Y} \sim N(0, K'VK). \tag{1.20}$$

Así

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{V}\mathbf{X})^{-1}\mathbf{X}'\hat{V}\mathbf{Y}. \tag{1.21}$$

Para (1.20), la función de verosimilitud se transforma en

$$L(\sigma^2) = \frac{1}{\sigma^n(2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X\beta})'(\mathbf{Y}-\mathbf{X\beta})}. \tag{1.22}$$

Ahora la función log-verosimilitud es:

$$\begin{aligned} l &= \log L(\sigma^2) && (1.23)\\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log K'VK - \frac{1}{2K'VK}(K'\boldsymbol{Y} - 0)'(K'\boldsymbol{Y} - 0). && (1.24) \end{aligned}$$

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

$$l = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\sigma^{2(n-p)}(K'K) - \tag{1.25}$$

$$\frac{1}{2(n-p)}(K'K)\boldsymbol{Y}'K(K'K)^{-1}K'\boldsymbol{Y}. \tag{1.26}$$

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

$$\frac{\partial l}{\partial \sigma^2} = \frac{n-p}{2\sigma^2} + \frac{1}{2\sigma^4}\boldsymbol{Y}'K(K'K)^{-1}K'\boldsymbol{Y}. \tag{1.27}$$

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

$$\hat{\sigma}^2_{REML} = \frac{1}{n-p}\boldsymbol{Y}'K(K'K)^{-1}K'\boldsymbol{Y}. \tag{1.28}$$

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. Mientras que el vector de parámetros de β puede ser estimado usando el estimador GLS que se explica a continuación.

Mínimos Cuadrados Generalizados ($\boldsymbol{GLS}$)

Si las suposiciones sobre la varianza constante y correlación cero entre los residuales no se cumplen, el estimador OLS del LM sigue siendo insesgado. Sin embargo, ya no es un estimador de mínima varianza. Ahora se puede utilizar el estimador por mínimos cuadrados generalizados, que en inglés se escribe Generalized Least Squares (GLS). Como $\mathbf{V}$ es la matriz de varianza y covarianza del error, es decir, en forma más general $Var(\varepsilon) = \sigma^2\mathbf{V}$, el estimador GLS de $\boldsymbol{\beta}$ minimiza la suma de cuadrados $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Así el estimador GLS es

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}. \tag{1.29}$$

Para el modelo lineal con $Var(\varepsilon) = \sigma^2\mathbf{V}$, 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 métodos de estimación anteriores, sustituyendo $\sigma^2\mathbf{V}$ por $\sigma^2 I$ [31].

1.1.3. Bondad de ajuste del modelo

En esta sección se presentan distintas maneras de evaluar la bondad de ajuste del LM, lo que se refiere al grado en que éste es conveniente como modelo que representa a las variables implicadas en él. Con el uso del software R, anteriormente se comentó que el ajuste de los Modelos Lineales se realizan usando la función lm(), mientras que para ver la bondad del ajuste se usa la función summary() [15]. Esta función da el resumen del modelo ajustado, permitiendo observar las distintas pruebas de la bondad de ajuste del modelo, estas pruebas se describen a continuación.

Prueba F para el ajuste del modelo

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

$$\begin{gathered} H_0 : \beta_0 = \beta_1 = \ldots = \beta_p = 0 \\ \text{vs} \\ H_a : \beta_j \neq 0 \;\; al \;\; menos \;\; para \;\; un \;\; j \;\; (j = 1, \ldots, p). \end{gathered}$$

El rechazo de la hipótesis nula implica que al menos uno de los regresores contribuye al modelo en forma significativa.
Aunque en el Cuadro 1.2 desarrollamos como elaborar una ANOVA que muestra el

cálculo del estadístico F para esta prueba, creemos conveniente repetirlo con la notación comúnmente utilizada 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

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i). \tag{1.30}$$

Ahora procedemos a elevar al cuadrado en ambos lados de la ecuación (1.30), y se suman para todas las n observaciones. Así se obtiene la siguiente ecuación:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{1.31}$$

Por otro lado,

$$\begin{aligned}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) &= 2\sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i) - 2\sum_{i=1}^{n}\bar{y}_i(y_i - \hat{y}_i) \\ &= 2\sum_{i=1}^{n}\hat{y}_i e_i - 2\hat{y}_i\sum_{i=1}^{n}e_i = 0 \qquad .\end{aligned} \tag{1.32}$$

Ya que la suma de los residuales siempre es igual a cero y la suma de los residuales ponderados por el valor ajustado $(\bar{y} - y_i)$ correspondiente también es igual a cero, la ecuación (1.31), se reduce a

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{1.33}$$

Esta igualdad nos dice que **suma total de cuadrados** que se denota por SS_T se puede escribir como una **suma de cuadrados debido a la regresión**, SS_R, y una **suma de cuadrados de residuales**, SS_{Res}.

$$SS_T = SS_R + SS_{Res}. \tag{1.34}$$

En [33] 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

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

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

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

El procedimiento se resume en el Cuadro 1.4 de análisis de varianza de la regresión.

Cuadro 1.4: 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
Residuales	SS_{Res}	$n-p-1$	MS_{Res}		
Total	SS_T	$n-1$			

Elaborado por Montgomery, Peck y Vining en [33].

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 $\boldsymbol{Y}$, que se explica mediante el modelo LM. La cantidad

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}, \tag{1.35}$$

se llama **coeficiente de determinación**, 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 Y está 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 [33].
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.
Es posible usar el estadístico R^2_{Adj} , que se define como sigue:

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-(k+1))}{SS_T/(n-1)}. \tag{1.36}$$

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. La R^2_{Adj} penaliza el aumento de términos que no son útiles. Tanto R^2 como R^2_{Adj}, suelen utilizarse como procedimientos para evaluar y comparar los posibles modelos de regresión.

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.
En [33] se sugiere tener cuidado al agregar una variable explicativa, ya que también aumenta la varianza del valor ajustado $\hat{Y}$, por lo que se debe tener cuidado de incluir sólo variables explicativas que tengan valor para explicar la respuesta. Además, si agregamos una variable explicativa 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 $\beta_i, i = 1, 2, \ldots, p$, está dado por:

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

El estadístico de prueba para esta hipótesis es,

$$t_0 = \frac{\hat{\beta}_i}{\sqrt{\sigma^2 C_{ii}}}, \tag{1.37}$$

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

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

Se observa que ésta es una prueba parcial, porque el coeficiente de regresión $\hat{\beta}_i$ depende de todas las demás variables explicativas x_j, que hay en el modelo. Esto es una prueba de la contribución de x_i 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 grados de libertad, respectivamente [33].

1.1.4. Diagnóstico de las suposiciones

Al principio del capítulo se establecieron las principales suposiciones sobre el modelo lineal general, las cuales son:

(i) Independencia.

(ii) Linealidad.

(iii) Normalidad.

(iv) Homocedasticidad.

Las 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). Estas propiedades son globales y como tal no aseguran la adecuación de éste.
A continuación, se exponen algunos métodos para checar las suposiciones.

Estos métodos se basan principalmente en el estudio de los residuos del modelo. Con el Software R, es fácil verificar la comprobación de los supuestos con el uso de la función plot() que dibuja los gráficos, de los residuos cuando el argumento principal es un objeto del tipo lm(función que ajusta modelos lineales en R).
Ya vimos antes que los residuos (o residuales) se definen como:

$$e_i = y_i - \hat{y}_i, \quad i = 1, ..., n. \,, \tag{1.38}$$

siendo y_i la observación $i-$ésima y $\hat{y}_i$ su valor ajustado correspondiente. Podemos considerar que un residual es la desviación entre el valor observado 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 [33].
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 comprobar las suposiciones básicas.

Gráfica de probabilidad normal

Las pequeñas desviaciones respecto a la suposición 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. En [33] se comentó que 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.

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 $e_1 < e_2 < \ldots < e_n$ 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 a la función de distribución acumulada de la distribución normal estándar. Esto es consecuencia de $E(e_i) \simeq \phi^{-1}[(i - \frac{1}{2})/n]$ [33].

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 línea recta que representa la normal acumulad. 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 del modelo, 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 grupos (clusters).

1.1.5. Selección de modelos

Imaginemos que tenemos dos modelos ajustados con la misma variable respuesta, pero con diferente cantidad de variables explicativas. Queremos seleccionar sólo un modelo, para ello surgen los Criterios de Selección de Modelos. Uno es el primer Criterio de Información de Akaike (AIC, por sus siglas en inglés de Akaike's An Information Criterion) y otro el Criterio de información Bayesiano (BIC por sus siglas en inglés de Bayesian Information Criterion). Siguiendo la notación del software R en [54], con las siguientes funciones en R calculamos los criterios AIC()y BIC(). El mejor modelo es aquel que cuente con el valor más pequeño de los valores AIC y BIC. Estás funciones están basadas con las siguientes ecuaciones:

$$AIC = -2\log(L) + 2(n_{par}), \tag{1.39}$$

$$BIC = -2\log(L) + n_{par}\log(n_{obs}), \tag{1.40}$$

donde L representa la verosimilitud, n_{par} es el número de parámetros en el ajuste del modelo y n_{obs} el número de observaciones en el modelo ajustado.

1.2. Modelo Lineal Mixto (LMM)

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 [15]. Fue Eisenhart en 1947 quien 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, lo que 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). Hay muchos paquetes en R, que contienen funciones que permiten ajustar los LMM. La lista incluye, entre otros, paquetes amer, arm, gamm, gamm4, GLMMarp, glmmAK, glmmBUGS, heavy, HGLMM, lme4.0, lmec, lmm, longRPart, MASS, MCMCglmm, nlme, PSM y pedigreemm. En este trabajo nos limitaremos a ajustar los modelos LMM con los paquetes lme4 y nlme, por ser los más usados dentro de los LMM. Los modelos que se ajustan con las funciones *lmer*() y *lme*() de sus paquetes lme4 [5] y nlme [53].
En esta sección se describe a detalle en que consiste un LMM, métodos de estimación, bondad de ajuste del modelo, diagnósticos de las suposiciones y los criterios de selección de modelos.

1.2.1. Establecimiento del modelo LMM

En el establecimiento del modelo deben considerarse las estructuras que se explican a continuación.

Estructura matricial

La siguiente notación corresponde a la estructura Matricial para la $i-$ésima observación de un *LMM*:

$$Y_i = X_i\beta + Z_i u_i + \varepsilon_i. \tag{1.41}$$

En (1.41) Y_i representa un vector de respuesta continua para la $i-$ésima observación, siguiendo la notación de [58]. 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 (1.41) 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 (1.41) 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 (1.41) 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 en la matriz X_i y la matriz Z_i [58].

El vector u_i para la $i-$ésima observación en la ecuación (1.41) 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:

$$u_i \sim N(0, D). \tag{1.42}$$

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:

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

Finalmente, los vectores ε_i en la ecuación (1.41) 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 [58]. 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:

$$\varepsilon_i \sim N(0, R_i). \tag{1.43}$$

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:

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

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 1.2.1), 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 [58].
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.

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

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 la 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:

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

En este caso, el vector θ_D contiene dos parámetros:

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

Estructura de covarianzas para la matriz R_i

La matriz de covarianza más simple para R_i es la **estructura diagonal**, en la que se supone que los residuos asociados a las observaciones sobre el mismo sujeto

se asumen que están correlacionadas y tienen igualdad de varianzas [58]. La matriz diagonal R_i para la $i-$ésima observación tiene la siguiente estructura:

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

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:

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

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 residuales asociados con los valores de respuesta observado para el $i-$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 residuales es plausible (por ejemplo, los ensayos repetidos en las mismas condiciones en un experimento).

Estructura matricial general

Una especificación alternativa, basada en todos los sujetos de estudio, se presenta en la ecuación (1.44)

$$\mathbf{Y} = \underbrace{\mathbf{X}\boldsymbol{\beta}}_{Fijos} + \underbrace{\mathbf{Zu} + \boldsymbol{\varepsilon}}_{Aleatorios} \tag{1.44}$$

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

$$\boldsymbol{u} \sim N(0, \mathbf{D}) \quad y \quad \boldsymbol{\varepsilon} \sim N(0, \mathbf{R}). \tag{1.45}$$

- $\mathbf{Y}$ 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 llamadas "efectos aleatorios".
- $\boldsymbol{\varepsilon}$ es un vector $n \times 1$ de errores aleatorios.
- $\mathbf{D}$ es una matriz diagonal por bloques, representando la matriz de varianzas-covarianzas para todos los efectos aleatorios, con bloques en la diagonal definida por la matriz D.
- $\mathbf{R}$ es una matriz de dimensión $n \times n$ diagonal por bloques, representando la matriz de varianzas-covarianzas para todos los errores, con bloques en la diagonal definida por las matrices R_i.

Propiedades básicas

Una propiedad dentro de un LM es que $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ donde $\boldsymbol{\beta}$ son los efectos fijos; para un LMM también usamos la misma notación $\mathbf{X}\boldsymbol{\beta}$ para efectos fijos, pero añadimos $\mathbf{Zu}$, donde los elementos de $\mathbf{u}$ son variables aleatorias. Así, según [30] el valor esperado de Y es:

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

Se asume que:

$$\mathbf{u} \sim (0, \mathbf{D}). \tag{1.46}$$

Al calcular la $Var(\mathbf{Y})$, necesitamos la $Var(\mathbf{u}) = \mathbf{D}$ de (1.46) y aplicando esta propiedad tenemos

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

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

por lo tanto,

$$Var(\mathbf{Y}) = \mathbf{ZDZ}' + \mathbf{R}. \tag{1.47}$$

Entonces por (1.46) y (1.47) tenemos

$$Y \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{ZDZ}' + \mathbf{R}). \tag{1.48}$$

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

1.2.2. Métodos de estimación

Trataremos ahora los métodos para estimar los parámetros de los *LMM*. Empezamos definiendo el modelo como

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon} && (1.49) \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, && (1.50) \end{aligned}$$

donde la estructura de las matrices es como se define anteriormente, y los supuestos sobre los efectos aleatorios y los residuos son

$$\mathbf{u} \sim N(0, \mathbf{D} = \sigma^2\mathfrak{D}) \;\; y \;\; \boldsymbol{\varepsilon} \sim N(0, \mathbf{R} = \sigma^2\mathfrak{R}). \tag{1.51}$$

Ahora, separamos el factor de escala σ^2, de las matrices de varianza-covarianza $\mathbf{D}$ y $\mathbf{R}$ previamente definidas en 1.45. Las nuevas matrices $\mathfrak{D}$ y $\mathfrak{R}$ especifican la estructura de los efectos aleatorios y los residuales 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 residuos $\boldsymbol{\varepsilon}_i$, entonces $\mathfrak{R} = I_{(n\times n)}$ con $n = \sum n_k$.

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.

Máxima verosimilitud

El método de máxima verosimilitud es el mismo proceso como en la sección 1.1.2, pero basado en un modelo marginal como se muestra a continuación:

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V}(\boldsymbol{\theta})), \tag{1.52}$$

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

$$\mathbf{V}(\boldsymbol{\theta}) = \mathbf{Z}(\theta_D)\mathbf{Z}' + (\theta_R). \tag{1.53}$$

Su función de verosimilitud es:

$$\begin{aligned} &L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = \\ &\prod_{i=1}^{n} (2\pi)^{-n_i/2} |\sigma^2 V_i(\theta)|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{Y_i} - \mathbf{X_i}\beta)'(\sigma^2 V_i(\theta))^{-1}(\mathbf{Y_i} - \mathbf{X_i}\beta)}. \end{aligned} \tag{1.54}$$

Además, su función log-verosimilitud es:

$$\begin{aligned} &\log L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2\mathbf{V}(\boldsymbol{\theta})) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})(\sigma^2\mathbf{V}(\boldsymbol{\theta}))^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{k=1}^{K} log|V_i(\theta)| \\ &- \frac{1}{2\sigma^2}\sum_{i=1}^{K}(\mathbf{Y_i} - \mathbf{X_i}\beta)V_i(\theta)^{-1}(\mathbf{Y_i} - \mathbf{X_i}\beta). \end{aligned} \tag{1.55}$$

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 (1.55). 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 [45].

$$\hat{\beta}(\boldsymbol{\theta}) = (\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{V}(\boldsymbol{\theta}))^{-1}\mathbf{Y}. \tag{1.56}$$

Sustituyendo $\hat{\beta}(\boldsymbol{\theta})$ en la función de log-verosimilitud en (1.55) tenemos

$$\begin{aligned} &\log L(\hat{\beta}, \sigma^2, \boldsymbol{\theta}) = \\ &-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\sigma^2\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\hat{e})(\sigma^2\mathbf{V}(\boldsymbol{\theta}))^{-1}(\mathbf{Y}-\mathbf{X}\hat{e}), \end{aligned} \tag{1.57}$$

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 (1.57) con respecto a σ^2.

$$\hat{\sigma}^2 = \frac{\hat{e}'\mathbf{V}(\boldsymbol{\theta})^{-1}\hat{e}}{n}. \tag{1.58}$$

Esto muestra que $\hat{\sigma}^2$ puede ser una función de $\boldsymbol{\theta}$ (y los datos).
Sustituyendo $\hat{\sigma}^2$ en (1.57) y reduciendo el log- verosimilitud obtenemos:

$$= -\frac{1}{2}(n)log(\hat{e}'\mathbf{V}(\boldsymbol{\theta})^{-1}\hat{e}) + \sum_{i=1}^{n} log|V_i(\theta)|. \tag{1.59}$$

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 $\hat{\boldsymbol{\theta}}$.

Máxima verosimilitud restringida

La idea de máxima verosimilitud restringida es la misma de los ML, como se presentó en el capítulo anterior en la sección 1.1.2. 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 1.1.2.
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 residuos 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 [31]. 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:

$$\begin{aligned} l(\sigma^2, \boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\log|\sigma^2 K'\mathbf{V}(\boldsymbol{\theta})K| \\ &\quad - \frac{1}{2\sigma^2}(K'\mathbf{Y})'(K'\mathbf{V}(\boldsymbol{\theta})K)^{-1}K'\mathbf{Y} \\ &= -\frac{n-p}{2}\log(2\pi) - \frac{n-p}{2}log(\sigma^2) - \frac{1}{2}\log|K'\mathbf{V}(\boldsymbol{\theta})K| \\ &\quad - \frac{1}{2\sigma^2}\mathbf{Y}'K(K'\mathbf{V}(\boldsymbol{\theta})K)^{-1}K'\mathbf{Y}. \end{aligned} \tag{1.60}$$

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 1.1.2.

1.2.3. Bondad de ajuste del modelo

Las inferencias y pruebas de los *LMM* se pueden basar en los procedimientos utilizados en el LM. La prueba dé hipótesis siguiente ayuda a comparar dos modelos anidados; a continuación, se muestra la hipótesis nula y alternativa como:

$$H_0 : \text{El modelo restringido es suficiente.}$$
$$\text{vs}$$
$$H_a : \text{El modelo completo es significativamente mejor que el modelo restringido}$$

El modelo de la hipótesis nula 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 [31].

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 [40]. Con frecuencia se desea comprobar si una muestra puede provenir de una distribución con ciertos parámetros desconocido. Se supone que se desea contrastar la hipótesis nula:

$$H_0 : \theta \in \Theta_0 \text{ ; } \theta \text{ está contenido en una región } \Theta_0 \text{ del espacio paramétrico } \Theta,$$
$$\text{vs}$$
$$H_a : \theta \in \Theta - \Theta_0 \text{ que supone que } \theta \text{ no está restringida a la región } \Theta_0 \text{ .}$$

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 $LRT \leq \alpha$.

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 ([41], p. 83-87). Cualquier prueba LR, utilizando *REML*, debe 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 [41], pág 87-92). Estas pruebas pueden ser severamente anti-conservadoras (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 verdadera matriz [31].

Deviance

La "deviance" es una estadística de bondad de ajuste, definida como:

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

donde modelo saturado corresponde al modelo que incluye más términos que el otro modelo. En caso de que el tamaño de la muestra n es grande y se toma el modelo bajo las suposiciones de que $\lambda(\beta)$ se distribuye

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

Las suiguientes ideas pueden utilizarse para la interpretación:

- 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 [31].

1.2.4. Diagnóstico de las suposiciones

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 las inferencias.

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 LM. El gráfico de residuos en todos los predictores y sobre el valor predicho es un buen punto de partida para ver si el modelo se ajusta a los datos [10]. Para corregir la forma del modelo, las mismas reglas se aplican como en el caso de los LM (véase la sección (1.1.4))

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 a 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 [31].

Los efectos aleatorios se supone que son realizaciones *i.i.d.* (independientes e idénticamente distribuidos) 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, [31]. 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 diferentes componentes son lineales.

Los errores

Una gráfica de los errores estandarizados (condicionales) en el valor predicho debería expresar una varianza constante sin tendencias. Una función de la varianza se puede utilizar para homogeneizar los errores o, alternativamente, una transformación se puede hacer a la respuesta. La normalidad de los errores 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 [31].

1.2.5. Selección de modelos: AIC y BIC

Según [31] 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 calculadas por el softwares estadísticos son el criterio de información de Akaike (AIC) y el criterio de información de Schwartz o Bayesiano (BIC).

$$AIC = log(L) - p, \tag{1.61}$$

$$BIC = log(L) - \frac{1}{2} p \ln n^*, \tag{1.62}$$

donde $n^* = n$ si L es la verosimilitud convencional y $n^* = n - p$ si L es la verosimilitud. Teniendo en cuenta que otros autores usan los criterios utilizando el negativo de las diferencias ([18], 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 en [54].

Capítulo 2

Modelo Lineal Generalizado y Modelo Lineal Generalizado Mixto

Los Modelos Lineales Generalizados son una extensión de los LM que permiten utilizar distribuciones no normales en la variable respuesta (binomiales, Poisson, gamma, exponencial, etc.).
En este capítulo se muestran dos modelos Lineales Generalizados, que se utilizan actualmente: el primero consiste en el Modelo Lineal Generalizado (GLM) y el segundo el Modelo Lineal Generalizado Mixto (GLMM).

2.1. Establecimiento del modelo GLM

McCullagh y Nelder en [29] propusieron una extensión del LM, llamado Modelo Lineal Generalizado (GLM, por sus siglas en inglés de Generalized Linear Model). Dentro de la literatura de los GLM, por mencionar algunos libros en particular en [1], [15], [29] y [35], se especifica de forma más concreta que un GLM puede ser explicando por sólo tres componentes:

- Un **componente aleatorio**.
- Un **componente sistemático**.
- y Una **función de enlace**.

A continuación, se describe a detalle cada uno de los componentes que constituyen un GLM.

1. Un **componente aleatorio** que identifica la variable respuesta Y y su distribución de probabilidad. Se asume que cada componente de la variable respuesta $Y = \{y_1, y_2, \ldots, y_n\}$ siguen una distribución de la familia exponencial. Esta familia tiene función de densidad de probabilidad o también llamada función de masa de la siguiente forma

$$f_{Y_i}(y_i|\theta_i, \phi) = exp\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - c(y_i, \phi)\}. \tag{2.1}$$

   Está representa una parametrización general con un parámetro de escala adicional ϕ (llamado parámetro de dispersión), donde θ_i es llamado el parámetro natural. Si ϕ es conocida, (2.1) representa una familia exponencial lineal. Mientras que si ϕ es desconocida es llamada un modelo de dispersión exponencial [45].

2. El **Componente sistemático** relaciona un vector $\eta = (\eta_1, \eta_2, \ldots, \eta_n)$ a un conjunto de variables explicativas a través de un predictor lineal.

$$\eta = X\beta. \tag{2.2}$$

   Donde η es llamado el predictor lineal, X es una matriz de dimensión $n \times p$ y β es el vector de p-parámetros.

3. La **función enlace** conecta el componente sistemático y el valor esperado del componente aleatorio.

   $E[y_i] = \mu_i$; entonces μ_i es conectado a η_i como:

$$g(\mu_i) = \eta_i, \tag{2.3}$$

   donde g es una *función monótona* y *diferenciable*, con función inversa denotada por g^{-1}. Otra forma de ver (2.33) se muestra a continuación

$$g(\mu_i) = \Sigma_{j=1}^{p}\beta_j x_{ij} \quad i = 1, 2, \ldots, n. \tag{2.4}$$

   Los siguientes son casos especiales de la función g

   i) Cuando $g(\mu) = \mu$ es llamada la función enlace identidad. Otra forma de expresarlo es: $\eta = \mu_i$.

   ii) $g(\mu) = Q(\theta)$ es llamada la función enlace canónica(natural). Otra forma de expresarlo es: $Q(\theta_i) = \Sigma_{j=1}^{p}\beta_j x_{ij}$.

Para ilustrar la forma de un GLM, dentro de las diferentes distribuciones, el caso más importante es la distribución normal, cuya función de densidad es:

$$\begin{aligned} f(y;\mu,\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-[y-\mu]^2/2\sigma^2\} \\ &= exp\{\ln(\frac{1}{\sqrt{2\pi\sigma^2}})\}\exp\{-[y-\mu]^2/2\sigma^2\} \\ &= exp\{\ln(1)-\ln((2\pi\sigma^2)^{1/2})-[y^2-2y\mu+\mu^2]/2\sigma^2\} \\ &= exp\{-\frac{1}{2}\ln(2\pi\sigma^2)-y^2/2\sigma^2+2y\mu/2\sigma^2-\mu^2/2\sigma^2\} \\ &= exp\{-\frac{1}{2}\ln(2\pi\sigma^2)-y^2/2\sigma^2+y\mu/\sigma^2-\mu^2/2\sigma^2\} \\ &= exp\{(y\mu-\mu^2)/2\sigma^2-\frac{1}{2}[y^2/2\sigma^2-\ln(2\pi\sigma^2)]\}. \end{aligned} \tag{2.5}$$

Con la última expresión en (2.5), se observa que puede expresarse como la ecuación (2.1). Por lo tanto,

$$\theta=\mu, \quad b(\theta)=\mu^2/2, \quad a(\phi)=\phi, \quad \phi=\sigma^2 \quad y$$

$$c(y,\phi)=-\frac{1}{2}[y^2/2\sigma^2-\ln(2\pi\sigma^2)].$$

En caso de tener otro tipo de distribución, se puede ver el Cuadro 2.1, la estructura que debe tener su respectivo modelo GLM. El ajuste de los modelos GLM con el software

Cuadro 2.1: Enlaces canónicos para distintas distribuciones de McCullagh y Nelder [29].

Distribución	Nombre	Función enlace	Función media
Normal	identidad	$\eta=\mu$	$\mu=\eta$
Exponencial	inversa	$\eta=\mu^{-1}$	$\mu=(\eta)^{-1}$
Gamma	inversa	$\eta=\mu^{-1}$	$\mu=(\eta)^{-1}$
Inverse Gaussian	inversa cuadrada	$\eta=\mu^{-2}$	$\mu=(\eta)^{-1/2}$
Poisson	Log	$\eta=\ln(\mu)$	$\mu=\exp(\eta)$
Binomial	Logit	$\eta=\ln(\frac{\mu}{1-\mu})$	$\mu=\frac{\exp(\eta)}{1+\exp(\eta)}$
Multinomial	Logit	$\eta=\ln(\frac{\mu}{1-\mu})$	$\mu=\frac{\exp(\eta)}{1+\exp(\eta)}$

R se hace usando la función $glm()$, donde los argumentos de la función $glm()$ es similar a los de la función $lm()$. El ajuste con el Softwre R se denota de la siguiente forma

```
glm(formula, family=familytype (link=linkfunction), data=),
```

donde

- formula: El modelo ajustado que corresponde a la variable respuesta y variables explicativas.
- data: Datos, subconjuntos en el cual el modelo es justado.
- family: Familia a la que pertenece la variable respuesta.
- link: Función link disponible para cada familia.

Propiedades de la función densidad

Dentro de las propiedades de la función de densidad en la ecuación (2.1), podemos denotar su esperanza y varianza. Primero denotemos a $L_i = logf(y_i; \theta_i, \phi)$ como el logaritmo de función de verosimilitud para una variable, así la función log-verosimilitud para todas las observaciones queda de la forma $L = \Sigma_i L_i$. De la ecuación (2.1),

$$L_i = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - c(y_i, \phi). \tag{2.6}$$

Por lo tanto,

$$\frac{\partial L_i}{\partial \theta_i} = \frac{\partial[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - c(y_i, \phi)]}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)}, \tag{2.7}$$

$$\frac{\partial^2 L_i}{\partial \theta_i^2} = \frac{\partial}{\partial \theta_i}\frac{\partial L_i}{\partial \theta_i} = \frac{\partial \frac{y_i - b'(\theta_i)}{a(\phi)}}{\partial \theta_i} = -\frac{b''(\theta_i)}{a(\phi)}, \tag{2.8}$$

donde $b'(\theta_i)$ y $b''(\theta_i)$ denotan la primera y segunda derivada de $b(\cdot)$ evaluada en θ_i. Aplicando los resultados generales de la verosimilitud

$$E\left[\frac{\partial L}{\partial \theta}\right] = 0 \quad y \quad -E\left[\frac{\partial^2 L}{\partial \theta^2}\right] = E\left[\frac{\partial L}{\partial \theta}\right]^2, \tag{2.9}$$

que se cumplen bajo condiciones de regularidad satisfechas por la familia exponencial que se ven en [14]. De la primera igualdad en (2.9) y considerando una sola observación obtenemos,

$$E\left[\frac{\partial L_i}{\partial \theta_i}\right] = E\left[\frac{y_i - b'(\theta_i)}{a(\phi)}\right] = \frac{1}{a(\phi)}\left[E[y_i] - E[b'(\theta_i)]\right] = 0, \tag{2.10}$$

despejando $E(y_i)$ obtenemos

$$\mu_i = E(y_i) = b'(\theta_i). \tag{2.11}$$

De la segunda igualdad en (2.9), podemos calcular ambas desigualdades

$$-E\left[\frac{\partial^2 L}{\partial\theta^2}\right] = E\left[\frac{\partial L}{\partial\theta}\right]^2, \tag{2.12}$$

$$-E\left[-\frac{b''(\theta_i)}{a(\phi)}\right] = E[\frac{y_i - b'(\theta_i)}{a(\phi)}]^2, \tag{2.13}$$

$$\frac{b''(\theta_i)}{a(\phi)} = \frac{Var(y_i)}{a(\phi)^2}, \tag{2.14}$$

$$b''(\theta_i)a(\phi) = Var(y_i). \tag{2.15}$$

Por lo tanto, tenemos

$$E[y_i] = \mu_i = b'(\theta_i) \quad Var[y_i] = \sigma^2 b''(\theta_i).$$

En la siguiente sección nos enfocamos en estimar los parámetros de β.

2.1.1. Métodos de estimación

Es esta sección se muestra el método de máxima verosimilitud dentro de los modelos GLM.

Estimación por máxima verosimilitud

Sea $\eta_i = x_i'\beta = \Sigma_{j=1}^{p} x_{ij}\beta_j$ el predictor de la $i-$ésima observación de la variable respuesta($i = 1, \ldots, n$), mientras que en representación matricial es de la siguiente forma

$$\eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix} = \begin{pmatrix} x_1'\beta \\ \vdots \\ x_n'\beta \end{pmatrix} = X\beta. \tag{2.16}$$

Considerando que los predictores están enlazados por medio de la función (2.1) monótona y diferenciable, expresando como:

$$g(\mu_i) = \eta_i, \tag{2.17}$$

o, en representación matricial,

$$g(\mu) = \begin{pmatrix} g(\mu_i) \\ \vdots \\ g(\mu_N) \end{pmatrix} = \eta. \tag{2.18}$$

Por las expresiones anteriores observamos que los parámetros θ y β están enlazados por la relación (2.1), que se expresa de la siguiente manera

$$\mu_i = b'(\theta_i) \quad y \quad g(\mu_i) = x_i\beta. \tag{2.19}$$

Nosotros estamos interesados sólo en estimar los parámetros $\beta_1, ..., \beta_p$ que pueden ser estimados por el método de máxima verosimilitud. Partimos de la función log de verosimilitud que depende de β

$$L(\beta) = \Sigma l_i(\beta). \tag{2.20}$$

Usando la regla de la cadena derivamos con respecto a los β's obteniendo la siguiente ecuación:

$$\frac{\partial L_i(\beta)}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}. \tag{2.21}$$

Los resultados de las derivadas parciales son:

$$\frac{\partial L_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} \tag{2.22}$$

$$= \frac{y_i - \mu_i}{a(\phi,)} \tag{2.23}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) \tag{2.24}$$

$$= \frac{Var(y_i)}{a(\phi),} \tag{2.25}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \sum_{k=1}^{p} x_{ik}\beta_k}{\partial \beta_j x_{ij}}. \tag{2.26}$$

Debido a que $\eta_i = g(\mu_i)$, la derivada $\frac{\partial \mu_i}{\partial \eta_i}$ depende de la función de enlace $g(\cdot)$ (más precisos depende de g^{-1}). Pero, no podemos especificar hasta tener la función de enlace definida.
Considerando la siguiente propiedad

$$\frac{\partial \theta_i}{\partial \mu_i} = \left[\frac{\partial \mu_i}{\partial \theta_i}\right]^{-1}. \tag{2.27}$$

Multiplicando lo obtenido en (2.21) y sustituyendo las propiedades obtenidas en las ecuaciones (2.22)-(2.27) determinamos las ecuaciones de verosimilitud.

$$\begin{aligned}\frac{\partial L_i(\beta)}{\partial \beta_j} &= \frac{\partial L_i}{\partial \theta_i}\frac{\partial(\theta_i)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \\ &= \frac{y_i - b'(\theta_i)}{a(\phi)}(\frac{\partial \mu_i}{\partial \theta_i})^{-1}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} \\ &= \frac{y_i - b'(\theta_i)}{a(\phi)}(\frac{Var(y_i)}{a(\phi)})^{-1}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} \\ &= \frac{(y_i - \mu_i)x_{ij}}{Var(y_i)}\frac{\partial \mu_i}{\partial \eta_i}.\end{aligned} \tag{2.28}$$

Finalmente las ecuaciones para todas las observaciones, se muestran en (2.29). Como la log-verosimilitud no es lineal en β se requiere aplicar métodos iterativos para encontrar los valores de β , por ejemplo el algoritmo Fisher [45].

$$\Sigma_{i=1}^{n}\frac{(y_i - \mu_i)x_{ij}}{Var(y_i)}\frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, \ldots, p. \tag{2.29}$$

2.1.2. Bondad de ajuste del modelo

La prueba de la razón de verosimilitud mantiene la idea de comparar los valores maximizados del log-verosimilitud dentro de H_0 y H_0 no restringido. Sea β_1, β_2, y β_3 el correspondiente vector de parámetros para ser estimado. Sea $g(\hat{\mu_1}) = \hat{\eta_1} = X_1\hat{\beta_1}$ y $g(\hat{\mu_2}) = \tilde{\eta_2} = X_2\tilde{\beta_2} = X_1\tilde{\beta_1} + X_3\tilde{\beta_3}$, donde $\hat{\beta_1}$ y $\tilde{\beta_2} = (\tilde{\beta_1}, \tilde{\beta_3})'$ son estimadores LM de los dos modelos, y rango$(X_1) = r_1$, rango$(X_2) = r_2$, y $r_2 - r_1 = r = df$.
La estadística de razón de verosimilitud, la cual compara el modelo más grande X_2 con un submodelo (más pequeño) X_1, se define como sigue (donde L denota la función log de verosimilitud ver (2.1)).

$$\wedge = \frac{max_{\beta_1}L(\beta_1)}{max_{\beta_2}L(\beta_2)}. \tag{2.30}$$

Entonces el estimador de máxima verosimilitud dado por

$$-2\ln\wedge = -2[L(\hat{\beta_1}) - L(\tilde{\beta_2})]. \tag{2.31}$$

Rao en [45] describe la prueba de hipótesis como: $H_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$ si se tiene H_0, entonces $-2\ln\wedge \sim \chi_r^2$ y se rechaza H_0 si la log- verosimilitud es significativamente más grande dentro del modelo que usa X_2. Si la diferencia es grande (el ajuste del modelo no restringido es mucho mejor) entonces H_0 es restringido.

2.1.3. Diagnóstico de las suposiciones

Como ya se ha expresado anteriormnete, los residuos son las diferencias entre los valores estimados por el modelo y los valores observados, muchas veces se utilizan los residuos estandarizados. Conviene analizar los siguientes gráficos, como sugiere Cayuela en [10]:

1. Histograma de los residuos.
2. Gráfico de residuos frente a valores estimados. Estos gráficos puede indicar falta de linealidad, heterocedasticidad (varianza no constante) y valores atípicos.

2.1.4. Selección de modelos

La "deviance" se utiliza como medida del ajuste del modelo a los datos, así como también las pruebas de significancia para los parámetros del modelo. Sin embargo, los criterios AIC y BIC vistos en los capítulos anteriores en las secciones 1.1.5 y 1.2.5 nos ayudan a ver el ajuste del modelo a los datos y la complejidad del modelo. Considerando que cuanto más pequeño es el valor de AIC y BIC mejor es el ajuste, Galecki en [18] recomienda usar AIC para comparar modelos similares con distintos grados de complejidad o modelos iguales (mismas variables) pero con funciones de vínculo distintas. Mientras que las funciones *stepAIC*(), *addterm*() y *dropterm*() del paquete MASS (ver más detalles en [57]) permiten comparar modelos con distintos grados de complejidad en función del AIC.

2.2. Establecimiento del modelo GLMM

El modelo Lineal Generalizado Mixto (GLMM) es una extensión del GLM incluyendo efectos aleatorios. En Demidenko [17] se explica a detalle una parte importante sobre el Modelo Mixto Marginal que juega un papel en la metodología de efectos aleatorios no lineales porque, después de la aproximación, el último modelo se reduce al primero. El GLMM toma una posición intermedia entre el modelo marginal y el modelo de efectos mixtos no lineales.
Estos modelos han ganado una gran popularidad en el modelado, debido a que se admite datos correlacionados dentro del contexto de los GLM y extiende en gran medida su amplitud de aplicabilidad.
En el ajuste de estos modelos con el uso del Software R se requieren las funciones *glmer*() del paquete *lme4* en [5] y *glmm* dentro de [25].
Una discusión amplia de GLMM se puede encontrar en los libros [17] y [30].

Cuadro 2.2: Tabla de distribuciones con sus respectivas funciones de elnace y varianzas que usa el Software R.

Famlilia	Enlace	Rasgo	Varianza
binomial	logit	binary	$\mu(1-\mu)$
gaussian	identity	continuous	ϕ
Gamma	inverse	continuous	$\phi\mu^2$
inverse.gaussian	$\frac{1}{\mu^2}$	continuous	$\phi\mu^3$
poisson	log	count	μ
quasi	identity	continuous	ϕ
quasibinomial	logit	binary	$\phi\mu(1-\mu)$
quasipoisson	log	count	$\phi\mu$

Los GLMM se resumen en:

$$\eta_i = x_i^{'}\beta + z_i^{'}u, \tag{2.32}$$

donde x_i y z_i son vectores conocidos y β un vector de parámetros desconocidos (los efectos fijos), a través de una función de enlace conocida $g(\cdot)$ tal que

$$g(\mu_i) = \eta_i. \tag{2.33}$$

Además, se asume que $u \sim N(O, G)$, donde la matriz de covarianza G puede depender de un vector θ de componentes de varianza desconocidos. Según las propiedades de la familia exponencial, se tiene $b^{'}(x_i) = \mu_i$. En particular, con enlace canónico, se tiene

$$\xi_i = \eta_i. \tag{2.34}$$

2.2.1. Distribución condicional de la variable respuesta

Para especificar el modelo, usamos la distribución condicional de Y dado u. Como en el caso de GLM en las ecuaciones (2.2) y (2.1), la variable respuesta Y se asume que elementos son independientes condicionalmente, cada uno con una distribución de densidad que pertenece a la familia exponencial como en (2.1). Todo lo dicho anteriormente se denota como:

$$y_i \mid u \sim indep.\ f_{Y_i|u}(y_i \mid u), \tag{2.35}$$

$$f_{Y_i|u}(y_i \mid u) = exp\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - c(y_i, \phi)\}, \tag{2.36}$$

Se sabe que la media condicional de y_i está relacionada con θ vía la identidad [30]

$$\mu_i = \frac{\partial b(\theta_i)}{\partial \theta_i}. \tag{2.37}$$

La ecuación (2.37), representa una transformación de la media que se desea modelar. Como en los GLM:

$$E[y_i \mid u] = \mu_i, \tag{2.38}$$

$$g[\mu_i] = x_i^{'}\beta + z_i^{'}u, \tag{2.39}$$

$g(.)$ es una función conocida, llamada función enlace (función link),
$x_i^{'}$ es la $i-$ésima fila de la matriz del modelo de efectos fijos,
β es el vector de los parámetros asociados a los efectos fijos,
$z_i^{'}$ es la $i-$ésima fila de la matriz del modelo de efectos aleatorios,
u es el vector de los efectos aleatorios,
μ_i representa la media condicional de y_i dado u.
Suponiendo

$$u \sim f_U(u). \tag{2.40}$$

2.2.2. Distribución marginal

En esta sección se describe los aspectos de la distribución marginal de y.

- Medias de y puede ser obtenida de forma usual como:

$$E[y_i] = E[E[y_i \mid u]] = E[\mu_i] = E[g^{-1}(x_i^{'}\beta + z_i^{'}u)] \tag{2.41}$$

  Para ilustrar la ecuación anterior, consideramos un caso particular $g(\mu) = log(\mu)$ y $g^{-1}(x) = exp(x)$.

$$\begin{aligned} E[y_i] &= E[exp(x_i^{'}\beta + z_i^{'}u)] \\ &= exp(x_i^{'}\beta)E[exp(z_i^{'}u)] \\ &= exp(x_i^{'}\beta)M_u(z_i), \end{aligned}$$

donde $M_u(z_i)$ es la función generadora de momentos de u evaluada en z_i (ver Apéndice A.4).
Supongamos además que $u_i \sim N(0, \sigma_u^2)$ y que cada fila de Z tiene una entrada igual a 1 y las demás igual a cero. Entonces $M_u(z_i) = exp(\sigma_u^2/2)$.

$$E[y_i] = exp(x_i^{'}\beta)exp(\sigma_u^2/2), \tag{2.42}$$

o

$$log(E[y_i]) = x_i^{'}\beta + \sigma_u^2/2. \tag{2.43}$$

- Varianzas
  Considerando la propiedad en el Apéndice A.2.

$$\begin{aligned} var(y_i) =& var(E[y_i|u]) + E[var(y_i|u)] \\ =& var(\mu_i) + E[var(y_i|u)] \\ =& var(\mu_i) + E[var(\sigma^2 * b''(\theta_i))] \end{aligned} \tag{2.44}$$

2.2.3. Métodos de estimación

Se han propuesto varias formas de aproximar la verosimilitud para estimar los parámetros de los GLMM, entre estos métodos se encuentran: cuasi verosimilitud penalizado(Penalized quasilikelihood PQL [8]), Laplace([46]), Cuadratura de Hermite Gauss(Gauss Hermite quadrature [43]), cadena de Markov monte carlo(MCMC [19]). En todos estos enfoques, uno debe distinguir entre la estimación ML y REML.

Función de verosimilitud

Usando las ecuaciones (2.36), (2.39) y (2.40), se escribe de manera más fácil la función de verosimilitud

$$L = \int \prod_i f_{Y_i|u}(y_i|u) f_U(u) du, \tag{2.45}$$

donde se integra sobre la distribución u q-dimensional. Mediante la función 2.45, estan basados distintos métodos para determinar los parámetros de los modelos GLMM.

La función de verosimilitud tiene función de verosimilitud para cada una de las observaciones y_i (independientes , condicionadas a los efectos aleatorios b_i), con función de verosimilitud dada por

$$L_i(\beta, b_i) \propto expy_i(x_i\beta + b_i) - B(x_i\beta + b_i) \tag{2.46}$$

En base a la notación anterior, la ecuación 2.45, se reescribe de la siguiente forma

$$L(\beta, \sigma) \propto \int_{-\infty}^{\infty} L_i(t)\phi(t; x_i\beta, \sigma)dt \tag{2.47}$$

donde
$L_i(\beta, \sigma) = exp\{y_i t - B(t)\}$, y
$\phi(t, x_i\beta, \sigma)$ es el caso particular de la función de densidad normal con media $x_i\beta$ y varianza σ^2

2.2.4. Aproximación de Laplace

En esta sección explicamos de manera general en que consiste el método de Laplace, que se toma como método de estimación para el ajuste de los modelos GLMM con la función de verosimilitud 2.45 y determinando el máximo sin calcular la integral. Este método es un aproximación principal para los modelos de efectos mixtos no lineales, la idea de la aproximación de Laplace (LA, por sus siglas en ingles Laplace Approximation) es usar una aproximación cuadrática, pero en el punto donde el integrando toma su valor máximo [17]. Recordando que una interpretación de la integral es el área bajo la curva, el mejor cubrimiento será en la vecindad del valor máximo. Por lo tanto, en lugar de tomar $x = 0$ como en la aproximación cuadrática, se aproxima como:

Comunmente la integral 2.45 no puede ser maximizada, supongamos que $\hat{b}_i$ es el máximo de la integral,

$$q_i(b_i; \beta, \sigma) = L_i(\beta, b_i))\phi(b_i; 0, \sigma) \tag{2.48}$$

para dados (β, σ), entonces aplicando el método de Laplace a un factor en 2.45 se obtiene la aproximación. Si aplicamos el método de Laplace a un factor de la ecuación 2.48, la aproximación se escribe como:

$$\int_{-\infty}^{\infty} L_i(\beta, b_i)\phi(b_i; 0, \sigma)db_i \propto q_i(\hat{b_i}; \beta, \sigma)\hat{j_i}^{-1/2}(\beta, \sigma) \tag{2.49}$$

donde $j_i(\hat{\beta}, \sigma) = -\frac{\partial^2}{\partial {b_i}^2} logq_i(b_i; \beta, \sigma)|_{b_i = \hat{b_i}}$

2.2.5. Bondad de ajuste del modelo

En esta sección se describe el coeficiente de determinación y la "deviance" dentro de los GLMM.

Coeficiente de determinación R^2

Siguiendo la notación de Nakagawa en [34], en su trabajo da una generalización del R^2 para a todas las demás distribuciones no Normales, en particular a las distribuciones binomiales y gamma negativas que se usan comúnmente para modelar datos biológicos. En particular para la distribución gamma.

$$y_{ij} \sim gamma(\lambda_{ij}, \nu) \tag{2.50}$$

con función enlace log, se denota por

$$\ln \lambda_{ij} = \beta_0 + \Sigma_{h=1}^{k} \beta_h X_{h_{ij}} + \alpha_i \tag{2.51}$$

y

$$\alpha_i \sim Gaussian(0, {\sigma^2}_{\alpha}) \tag{2.52}$$

donde y_{ij} es la $j-$ésima observación, que sigue una distribución gamma con dos parámetros, λ_{ij} y ν(para ver otra parámetrización ver Apéndice A.7). En el caso de tener una distribución con media λ y varianza λ^2/ν los valores del R^2 marginal y condicional se calculan como:

$${R^2}_{gamma-\ln(m)} = \frac{{\sigma_f}^2}{{\sigma_f}^2 + {\sigma_\alpha}^2 + \ln(1 + 1/\nu)}, \tag{2.53}$$

$${R^2}_{gamma-\ln(c)} = \frac{{\sigma_\alpha}^2}{{\sigma_\alpha}^2 + \ln(1 + 1/\nu)}, \tag{2.54}$$

donde
σ_f^2 es la varianza explicada por los efectos fijos.

Deviance

El paso para evaluar el ajuste de un modelo GLMM es comparar el ajuste modelo con el modelo saturado, realizandolo como en los modelos anteriores, definido en 2.1.2.

Críterios AIC y BIC

Para más detalle ver sección 2.1.4.

Capítulo 3

Aplicación con imágenes no satelitales

En este capítulo se presenta la prueba de Kolmogorov-Smirnov, la cuál es poco usada para comparar imágenes donde el mayor desarrollo sobre este tema lo ha propuesto Demidenko en [17], como parte inicial para entender los Modelos Lineales Mixtos como alternativa de análisis en experimentos con medidas repetidas, una referencia actual que hace uso de los modelos desde este punto de vista es [20]. En esta sección, también se analizan los LMM con un ejemplo aplicado realizando todo el análisis con el software R. Esta prueba, genera una alternativa para analizar los resultados dentro de cada experimento usando imágenes digitales.

Los histogramas son parte importante que se consideran en esta prueba, donde el concepto de histograma de una imagen digital es una herramienta útil para determinar distintas características, ya sea contraste de colores o en escala de grises. Cada imagen tiene su propio histograma, obtenido mediante la información de sus pixeles.

Esta prueba se aplica a la problemática de los peces cebra sometidos a distintas soluciones de agua residual Textil, problema que actualmente enfrentan las empresas al descargar sus residuos de aguas en los ríos. A pesar de existir distintos tratamientos de agua, para tratar esta problemática, se han llevado a cabo tratamiento bajo fotocatálisis solar [11].

El uso de imágenes digitales ha ganado gran desarrollo para el uso y aplicación en todas las áreas, en particular en ciencias ambientales. Actualmente se están presentando investigaciones usando estas herramientas que aportan información rápida y concisa, sin tener un gran conocimiento en diferentes disciplinas.

A nivel mundial sabemos que más de 2 millones de toneladas de aguas residuales, desechos industriales y agrícolas se vierten en las aguas del mundo. No hay un manejo adecuado de las aguas residuales (artículos) y en México no es la excepción en su hidrología, en particular hay un río llamado Atoyac que pasa por diferentes estados

(Veracruz, Tlaxcala, Puebla, etc). Este río pasa por zonas textileras y afecta la vida ecológica de sus habitantes. Dentro de los estudios que hasta ahora se han llevado a cabo en esta zona, uno se enfocan en un *análisis molecular* que da como resultado si el pez cebra ha sufrido algún cambio por el agua residual textil(cambios fisiológicos y/o morfológicos).

En esta sección se utilizan las imágenes digitales del pez cebra que se obtienen en los estudios de análisis moleculares y la prueba de Kolmogorov-Smirnov, con el objetivo de comparar las imágenes y evaluar si existe efecto sobre el pez cebra sometido a distintas soluciones de agua residual textil en el Río Atoyac, el análisis de las imágenes se hacen con el software R.

3.1. Prueba estadística de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov (K-S) es una prueba de hipótesis no paramétrica y se usa para contrastar si dos distribuciones son iguales. En el uso de las imágenes se ha trabajado con intensidades de gris de 256 niveles. Según Demidenko en [17] muestra que la prueba K-S puede ser exitosa para comparar imágenes microscópicas, donde se obtiene el histograma de la imagen y se usa para construir la distribución de la intensidad de niveles de gris, donde se toma la información de los pixeles de cada imagen.

En general la función de distribución empirica F_n para n observaciones i.i.d, $X_1, X_2, \ldots, X_n$ se define como:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty,x]}(X_i).$$

Donde
$I_{[-\infty,x]}(X_i)$, corresponde a la función indicadora, asignando 1 si $X_i \leq x$ y cero en otro caso.
$\frac{1}{n}$, es el valor de paso.
A continuación esta idea es implementada para el caso de la función de distribución empírica para escala de grises, primero definiendo lo que es el histograma de una imagen.

El histograma de imágenes es una técnica de procesamiento de imágenes frecuentemente utilizada, siguiendo la notación de Demidenko la denotamos como $h_g, g = 0, 1, \ldots, 255$. La siguiente función representa la función de distribución de niveles de grises acumulativa empírica

$$F_g = \sum_{g'=0}^{g} h_{g'}. \tag{3.1}$$

La función de paso F_g no decreciente con paso $\frac{1}{256}$ en $g = 0, 1, \ldots, 255$. Sean $F^1 = \{F_g^1, g = 0, \ldots, 255\}$ y $F^2 = \{F_g^2, g = 0, \ldots, 255\}$ dos distribuciones de nivel de gris. De las imágenes M_1 y M_2, con dimensiones $P_1 \times Q_1$ y $P_2 \times Q_2$ respectivamente. El máximo valor, $\hat{D} = max_g \mid F_g^1 - F_g^2 \mid$, corresponde a la distancia de una distribución a otra. Kolmogorov y Smirnov demostraron que, si las distribuciones teóricas son iguales, entonces la probabilidad de que la distancia observada, $\hat{D}$, sea mayor que D es

$$Q_{KS}(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}exp(-2j^2\lambda^2), \tag{3.2}$$

donde $\lambda_{KS} = D[\sqrt{J} + \frac{0.11}{\sqrt{J}} + 0.12]$ y $J = \frac{P_1 \times Q_1 \times P_2 \times Q_2}{P_1 \times Q_1 + P_2 \times Q_2}$.
$Q_{KS}(\lambda)$ se toma como el valor p de la prueba. Cuanto mayor sea la distancia entre distribuciones, menor será la probabilidad $Q_{KS}(\lambda_{KS})$.
En el caso de que dos imágenes produzcan la distancia D y la probabilidad calculada $Q_{KS}(\lambda_{KS}) < 0.05$, rechazamos la hipótesis de que las dos imágenes son iguales con un 5 % de error. Podemos encontrar λ_{KS} tales que $Q_{KS}(\lambda_{KS}) = 0.05$, dando el umbral $\lambda_{KS} = 1.358$. Recordando que la hipótesis es de la siguiente manera,

$$H_0 : F_1(x) = F_2(x) \;\; vs \;\; H_A : F_1(x) \neq F_2(x) \; para \; al \; menos \; una \; x. \tag{3.3}$$

3.2. Aplicación usando imágenes del pez cebra

Los datos utilizados en este trabajo provienen del proyecto de maestría de Rivera [47], cuyo objetivo fue evaluar la toxicidad del agua residual textil (ART) vertida al Río Atoyac y su efecto en la morfología, fisiología e histología del pez cebra (Danio rerio). Se observó que una de las principales fuentes de contaminación en el lugar son las descargas de agua residual vertidas a la cuenca del Río Atoyac provenientes de las industrias textiles productoras de mezclilla (proceso Denim) y los talleres de lavandería de acabado de las prendas confeccionadas con dicha tela [48], en la población de villa Alta, municipio de Tepetitla de Lardizábal.

El agua residual Textil colectada directamente en la zona de estudio se le denominó agua residual compuesta (ARTC) o solución madre (100 %). A partir de esta solución se realizaron distintas soluciones con agua de garrafón: para el 75 % se obtuvo (75 % ARTC +25 % agua de garrafón), también para las disoluciones de 50, 35, 25 y 13.5 % (ver Figura 3.1). En este trabajo sólo se utilizó la disolución al 75 % y los peces sometidos al agua de garrafón. Las imágenes originales Figuras 3.2, 3.3 y 3.4 tienen formato JPG, las cuáles fueron convertidas a formato PGM que se muestran en las Figuras 3.5, 3.6 y 3.7. La imagen en formato PGM es la que se usa en este método, junto con la prueba K-S.

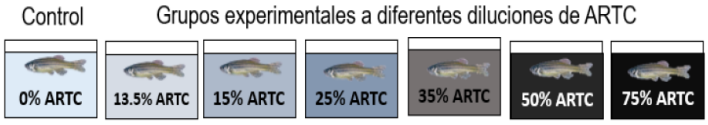
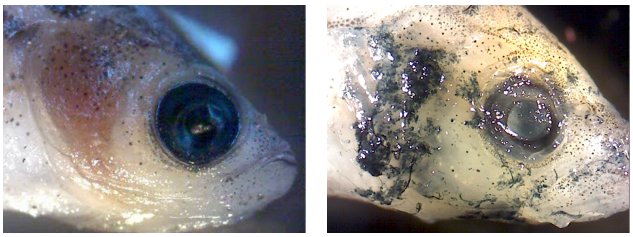


Figura 3.1: Grupos experimentales a diferentes diluciones de ARTC.

Las comparaciones que se hacen con imágenes del pez cebra (cabeza, arco branquial y branquiespinas filamentos) del grupo control (agua de garrafón) y la segunda del grupo experimental (75 % de agua residual textil).

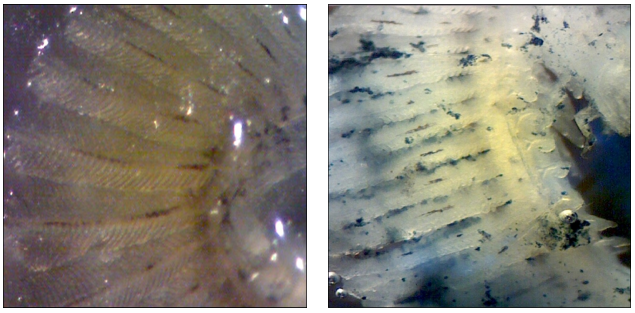


(a) Grupo control (b) Grupo experimental

Figura 3.2: Imagen Grupo Control en agua de garrafón y Grupo experimental sometido a 75 % ART. Cabeza, pertenece a la tesis de maestría de Rivera 2018.

(a) Grupo control (b) Grupo experimental

Figura 3.3: Imagen Grupo Control en agua de garrafón y Grupo experimental sometido a 75 %. Aparato (indigo), Arco braquial, pertenece a la tesis de maestría de Rivera 2018.



(a) Grupo control (b) Grupo experimental

Figura 3.4: Imagen Grupo Control en agua de garrafón y Grupo experimental sometido a 75 % ART branquiespinas Filamentos y Lamelas, pertenece a la tesis de maestría de Rivera 2018.

3.2.1. Comparación de imágenes usando prueba Kolmogorov-Smirnov

La imágenes que se muestran en las Figuras 3.8, 3.9 y 3.10, tienen formato PGM donde las matrices que se obtuvieron de cada imagen en escala de grises fueron de orden 1600×1200, información importante para la construcción de sus respectivos histogramas.

- Se usa el código en R que propone Demidenko en [16], el cual se modificó para implementarlo con las imágenes de este trabajo. El código completo que se generó para la lectura de las imágenes se muestra en el Apéndice B.1, para obtener los valores de la prueba K-S.

  Comparación de las cabezas de los peces cebra

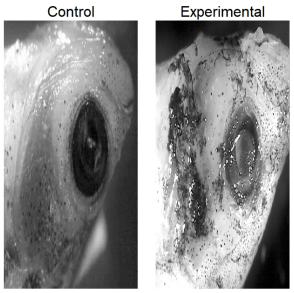


Figura 3.5: Imágenes de la cabeza de los peces en formato PGM (Comparación 1).

Comparación del Aparato Branquial

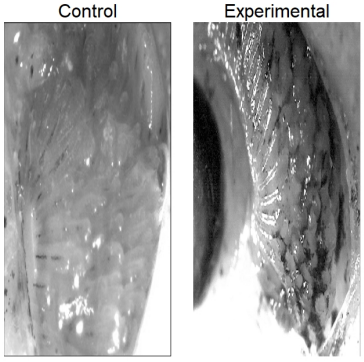


Figura 3.6: Imágenes del aparato branquial de los peces en formato PGM (Comparación 2).

Comparación de los Filamentos

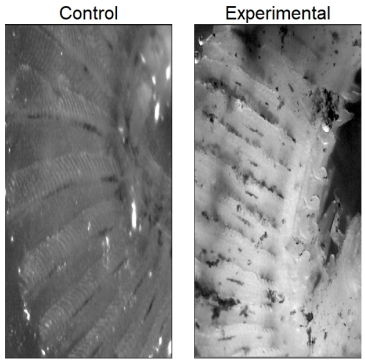


Figura 3.7: Imágenes de Branquiespinas y Filamentos de los peces en formato PGM (Comparación 3).

- Se muestra el respectivo histograma para cada imagen, correspondiente a las comparaciones anteriores. Todo es generado por el algoritmo que se muestra en el Apéndice B.1.

  Histograma y Función de distribución (Comparación 1).

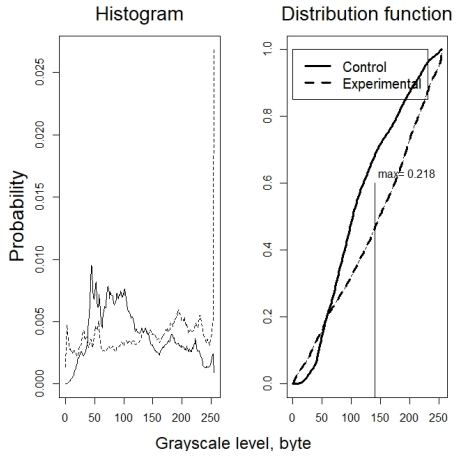


Figura 3.8: Histograma y Función de distribución(Comparación 1).

Histograma y Función de distribución (Comparación 2).

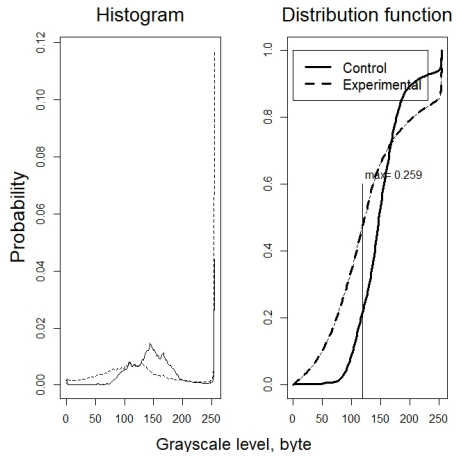


Figura 3.9: Histograma y Función de distribución (Comparación 2).

Histograma y Función de distribución (Comparación 3).

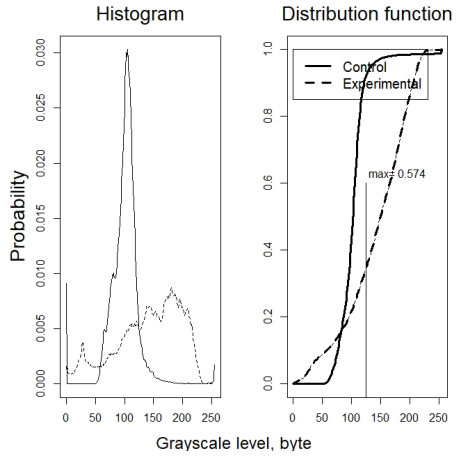


Figura 3.10: Histograma y Función de distribución (Comparación 3).

- Por útilo el algoritmo da como resultado el nivel máximo de pixel (byte.máx), la distancia máxima (D), el valor de λ y el p valor Q_{KS}, todo generado con el Software R. Estos resultados son resumidos en el Cuadro 3.1

Cuadro 3.1: Resultados de la prueba K-S

IMÁGENES	byte.max	D	λ	Q_{KS}
Comparación 1[a]	141	.218	213.308	0
Comparación 2[b]	119	.259	253.344	0
Comparación 3[c]	125	.574	357.809	0

Las comparaciones del pez cebra dan como resultado $Q_{KS} < .05$, lo cual genera rechazar las tres pruebas de hipótesis que son de la forma 3.3. Esto genera una forma numérica de mostrar que las imágenes que se comparan son diferentes, siendo una alternativa más confiable que solo tener la parte visual al comparar las imágenes. Al tener imágenes microscópicas el ojo humano es más difícil de percibir diferencia entre ellas, por lo cual se recomienda usar la prueba K-S, para contrastar imágenes.

En este capítulo se usó la prueba K-S con el objetivo de comparar imágenes del pez cebra que ha sido sometido a distintas soluciones de agua residual Textil. Esta prueba hace uso de los pixeles de cada imagen con el objetivo de compararla dos de ellas y concluir si las imágenes son diferentes. Se usa la métrica del valor máximo de las distancias de los pixeles de cada imagen, para las decisiones de rechazar o no prueba de hipótesis. El resultado obtenido de rechazar la prueba implicaría que no existe cambio de una imagen a otra, es decir, son la misma imagen.
Se realizaron tres comparaciones con imágenes del pez cebra, para cada una de las imágenes se construyó su respectivo histograma y su función de distribución. La prueba K-S fue parte importante que dio como resultado que las imágenes son distintas. Las funciones de distribuciones muestran que hubo cambios rotundos en los peces, a tal grado que sufrieron cambios físicos y morfológicos drásticos sobre cada uno de los peces.
Los análisis fisicoquímicos del agua residual textil demostraron que siete parámetros se encuentran fuera de Norma 001-SEMARNAT-1996: nitrógeno total, sólidos sedimentables, grasas y aceites, DBO5, fósforo, DQO, sólidos suspendidos totales. Dichos parámetros inducen: respuestas fisiológicas, morfológicas, al estrés, adaptaciones bioquímicas y ajustes del comportamiento del pez cebra. Cuando se producen cambios en los hábitats naturales las especies pueden lograr sobrevivir con mucho esfuerzo, adaptándose a los cambios generados en su entorno. Toda esta información coincide con los resultados obtenidos usando la prueba K-S.
Finalmente, la prueba K-S resultó otra forma de analizar los resultados de un experimento con imágenes que un resultado numérico y que indica sí existe un cambio en los peces sometidos a diferentes soluciones de agua. Esta prueba es poco usada y

se recomienda implementar en más áreas para obtener resultados de comparación de imágenes.

Capítulo 4

Aplicaciones con imágenes satelitales

En este capítulo se presentan dos aplicaciones de la modelación medioambiental construyendo una variable NDVI (índice de Vegetación Normalizado, en inglés Normalized Diference Vegetation Index) por medio de imágenes de satélite que es añadida en cada uno de los modelos de regresión como variable independiente de efectos fijos. La primera aplicación tiene como objetivo modelar la cantidad de Formación Vegetal existente en el sur de la presa de Valsequillo en Estado de Puebla, México.
La segunda aplicación modela la cantidad de Carbono orgánico en la zona terrestre prioritaria en Puebla RTP-105, en dos enfoques diferentes, el primero incluyendo la variable NDVI como variable fija y el segundo como variable aleatoria. En los modelos ajustados se trato de incremente el R^2, en esta segunda aplicación los modelos mejoran el ajuste y dado que se tienen clasificaciones, variables de efectos fijos y aleatorios los modelos GLMM resultaron ser más apropiados. En ambos modelos se busca que las variables sean significativas y se cumpla los supuestos de los modelos.

Antes de adentrarnos en las aplicaciones mencionadas anteriormente, definimos conceptos importantes que nos llevan a la información obtenida por medio de las imágenes. Entre los conceptos que mencionamos son Teledetección y la obtención de las imágenes de satélite.

4.0.1. Teledetección

Esta sección se inicia con el concepto de teledetección, que sirve como base para la obtención de información, para determinar la variable NDVI que es parte de la construcción de los modelos. La teledetección es una técnica que nos permite obtener información a distancia de objetos que se encuentran situados sobre la superficie te-

rrestre [22]. El fenómeno de la Teledetección es posible gracias a la interacción de la energía electromagnética con las cubiertas terrestres. Estas tienen un comportamiento reactivo variable, condicionado tanto por los factores externos (ambientales) como por sus propias características fisicoquímicas en el momento de la toma de la imagen [13]. La primera experiencia de teledetección se retoma en 1859, donde Gaspar Félix de Tournachon obtuvo las primeras fotografías aéreas desde un globo, después se hizo uso del avión hasta llegar en 1960 al uso de los satélites. Actualmente son numerosos los centros de producción, enseñanza e investigación que trabajan activamente en este campo con el uso de satélites (Teledetección espacial). Para entender un poco más sobre la Teledetección espacial, se definen conceptos básicos que nos llevan a entender el concepto de Procesamiento de Imagen de Satélite, en la siguiente sección se explica a detalle cada uno de los conceptos, iniciando con el concepto de Ecuación de una onda, hasta llegar a la obtención del NDVI.

4.0.2. Procesamiento de imágenes de satélite

Para entender un poco más sobre el procesamiento de imágenes, iniciamos con el concepto de una onda. En general la decripción matemática de una Onda más utilizada dentro de la literatura se describe por la siguiente función:

$$f(x,t) = A\sin(wt - kx). \tag{4.1}$$

Donde,
A es la amplitud de una onda.
k es un número de onda angular puede ser asociado con la longitud de onda por la relación:

$$k = \frac{2\pi}{\lambda}, \tag{4.2}$$

T es el tiempo.
La frecuencia f es el número de ciclos completos transcurridos en la unidad de tiempo

$$f = \frac{1}{T}. \tag{4.3}$$

w representa la frecuencia en radianes por segundo.

Según [13], el tipo de Ondas que se usa para la Teledetección corresponden a las Ondas Electromagneticas.

Onda Electromagnética:Las ondas electromagnéticas abarcan un espectro extremadamente amplio de longitud de onda y frecuencia. Este espectro electromagnético comprende la luz visible, la radiación infrarroja y ultravioleta, los rayos x, los rayos gamma y la transmisión de radio y televisión.

Bandas Del Espectro Electromagnético: Para su estudio, el espectro electromagnético se divide en segmentos o bandas, aunque esta división es inexacta. Existen

ondas que tienen una frecuencia, pero varios usos, por lo que algunas frecuencias pueden quedar en ocasiones incluidas en dos rangos.

El Sistema de Referencia Mundial (WRS) es un sistema de notación global para los datos del Landsat. Permitiendo a los usuarios encontrar imágenes digitales sobre cualquier parte del mundo, mediante la especificando de los números Path y Row (longitud/latitud)(ver Apéndice F.1, para determinar el Path and Row de cualquier zona de estudio).

Para obtener información mediante teledetección espacial es importante que los objetos y el sensor tengan algún tipo de interacción. Para que la observación sea posible, según Chuvieco en [13] se necesitan los siguientes elementos:

- Sensor: Instrumentos de grabación, instrumentos de escaneo, aviones, satélites, astronave, boyas o barcos.
- Objeto observado: arboles, suelos.
- Flujo energético que permite poner a ambos en relación.

Considerando los puntos anteriores, dentro de los sensores podemos contemplar a constelación de satélites LANDSAT (LAND=tierra y SAT=satélite), que inicialmente se llamaron ERTS (por sus siglas en inglés, Earth Resources Technology Satellites), fue la primera misión de los Estados Unidos para el monitoreo de los recursos terrestres, el primer satélite se puso en órbita el 23 de julio de 1972.

Índices de Vegetación

Se han desarrollado muchos índices de vegetación basados en el hecho de que las plantas reflejan menos en luz roja visible, pero más en la radiación infrarroja cercana en comparación con la superficie sin vegetación [4] y [22].
Por lo tanto, los índices de vegetación pueden mejorar o extraer algunas características específicas que las bandas espectrales individuales no pueden. Los índices de vegetación que más se utilizan son: el índice de vegetación de diferencia normalizada (NDVI), el índice de vegetación ajustado al suelo (SAVI), el índice de vegetación de diferencia re normalizada (RDVI), el NDVI transformado (TNDVI), índice de vegetación simple (SVI) y la proporción simple (RVI).
El NDVI mide la cantidad de vegetación verde, tomando un cociente de las diferencias de reflectancia espectral entre infrarrojo cercano (NIR) y rojo (RED) para calcularlo, se ha utilizado ampliamente en estudios de teledetección cuyo rango de valores es de -1.0 a 1.0, donde los valores más altos son para vegetación verde y los valores bajos para otro tipo de superficie. Así, el suelo desnudo o rocas se representan con valores de NDVI más cercanos a 0 y los valores negativos corresponden principalmente a las nubes, el agua y la nieve.

La ecuación para calcular el NDVI queda de la siguiente forma:

$$NDVI = \frac{(NIR - RED)}{(NIR + RED)}. \tag{4.4}$$

Índice de Vegetación Relativo (FVC)
El Índice de Vegetación Relativo (en inglés "Fractional Vegetation Cover") es la estimación del porcentaje de vegetación [6]; se basa en el supuesto de que el NDVI de la vegetación se distribuye gradualmente entre el NDVI del suelo desnudo y el NDVI de la vegetación totalmente verde. Para calcular el FVC se utiliza la ecuación:

$$FVC = \frac{NDVI - NDVI_{sd}}{(NDVI_{VV} - NDVI_{sd})^2} * 100. \tag{4.5}$$

donde $NDVI_{sd}$ es el NDVI del suelo desnudo y $NDVI_{VV}$ es el NDVI del pixel con la mayor cobertura de vegetación verde en el área.

4.1. Análisis de la cobertura edáfica en el sureste de la presa Valsequillo, Puebla.

El uso y manejo no sostenible de la tierra está llevando a una mayor degradación del suelo y la pérdida de un recurso clave que es fundamental para la vida en el planeta [23]. En el Plan Estratégico Decenal 2008-2018 aprobado en 2007 por la 8ª Conferencia de las Partes de la Convención de las Naciones Unidas de Lucha contra la Desertificación, se estableció el objetivo de proteger los suelos contra la erosión y la contaminación. En México, en el Programa Nacional Manejo Sustentable de Tierras a favor de un manejo y uso sostenible, publicado en 2008, se señaló que la erosión y el declive de la fertilidad del suelo afectan a la viabilidad de los terrenos agrícolas.
Desde hace décadas, en la región sur de la presa de Valsequillo, el manejo de los recursos naturales se ha llevado de manera inadecuada provocando la degradación del suelo y la pérdida de la vegetación. La escasa investigación en los diferentes ecosistemas, el empleo de las tecnologías de explotación sin tener en cuenta las condiciones del entorno y el desconocimiento de los principios físicos, químicos y biológicos que determinan la continuidad de los sistemas de producción, son los principales factores que han afectado gravemente la estabilidad de los sistemas naturales en la Región Sur de la Presa de Valsequillo. La Figura 4.1 muestra la localización de la zona de estudio.
La zona de estudio corresponde a la parte norte del municipio de Tzicatlacoyan donde se sitúan las localidades San Miguel Acuexcomac, San Bernardino Tepenene, San José Texaluca y San Martín los Tételes.
Identificar mediante imágenes de satélite las afectaciones de la cobertura edáfica de esta zona es importante para la posterior toma de decisiones en el manejo de estos suelos y es el objetivo que se persigue en este estudio.

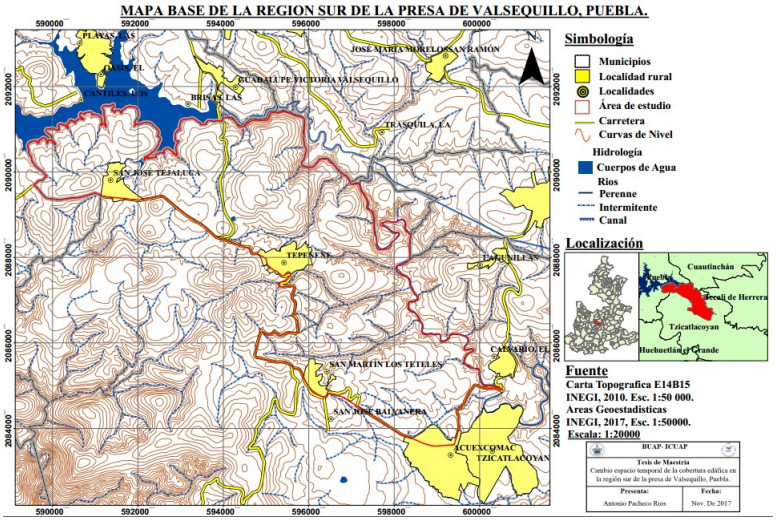


Figura 4.1: Localización de la zona de estudio. Fuente: Pacheco Ríos, A., 2018.

4.1.1. Factores de manejo de cobertura de suelo.

A continuación, se resumen algunos de los indicadores que se utilizan para estos estudios y que constituyen factores a tener en cuenta en el manejo de la cobertura del suelo.

- Los valores de este factor se calculan tomando el logaritmo del porcentaje de FVC usando la función de regresión

$$C\_FVC = 0.6508 - 0.343 \log(FVC). \tag{4.6}$$

- Es uno de los factores de la Ecuación Universal de Pérdida de Suelo, (USLE, por sus siglas en inglés). Este factor refleja el efecto de las prácticas de cultivo y manejo sobre las tasas de erosión. Sus valores indican cuánto protege la vegetación la erosión del suelo [59]. El factor C_USLE corresponde a la pérdida de suelo bajo condiciones específicas de cultivo, en relación con la que ocurre en un suelo desnudo. Toma valores de 0 (cubierta vegetal alto) a 1 (suelo desnudo). Esto puede ser representado como

$$C\_USLE = \frac{Acrop}{Afallow} * 100. \tag{4.7}$$

  Este factor refleja el efecto de las prácticas de cultivo ($Acrop$) y manejo sobre las tasas de erosión (A fallow). Los valores de C_USLE indican cuánto protege de

la erosión la vegetación del suelo. Es uno de los factores de la Ecuación Universal de Pérdida de Suelo, [59].

- El factor C_exp puede estimarse aplicando la relación utilizada por Van der Knijff en [56]:

$$C\_exp = \exp(-\alpha * \frac{NDVI)}{(\beta - NDVI)}). \tag{4.8}$$

Se considera como un factor de gestión de cobertura de cultivo calculado. En [56] se sugiere que al aplicar esta relación, se obtienen mejores resultados que usando una relación lineal. Los valores para los dos factores de escala α y β son 2 y 1, respectivamente.

4.1.2. Obtención de la imagen satelital y análisis exploratorio.

La imagen de satélite fue obtenida del Servicio Geológico de los Estados Unidos (USGS, por sus siglas en inglés) y corresponde al año 2015. Dicha imagen se obtuvo sin cobertura de nubes en la temporada de sequía, lo que permitió reducir el efecto espectral de la vegetación y de los cultivos en el procesamiento de la imagen. El análisis e integración de la información se realizó con el Sistema de Información Geográfica de uso libre Quantum GIS [44]. Con la información obtenida se elaboró una tabla de 34552 filas (puntos georreferenciados) y 5 columnas con las variables cuantitativas NDVI, FVC, C_FVC, C_USLE y C_exp, realizándose el análisis exploratorio de los datos. Se detectaron 5 puntos aberrantes o "outliers" que fueron eliminados, dado que el análisis gráfico realizado y la gran cantidad de datos disponibles indicaron que no se afectaría la modelación estadística posterior. La matriz de correlaciones, que se muestra en el Cuadro 4.1, indica que las variables NDVI, FVC, C FVC y C exp se encuentra altamente correlacionadas entre ellas, por lo cual es importante considerar su relación para la construcción de modelos.

Cuadro 4.1: Correlación entre las variables.

	NDVI	FVC	C_FVC	C_USLE	C_exp
NDVI	1.00	0.98	-0.97	-0.32	-0.96
FVC		1.00	-0.93	-0.32	-0.90
C_FVC			1.00	0.30	0.99
C_USLE				1.00	0.29
C_exp					1.00

4.1.3. Modelos de regresión ajustados

Se ajustaron tres modelos lineales, uno para cada variable FVC, C_FVC y C_exp, tomadas como variables dependientes y como variable independiente el NDVI. Sus

estimaciones fueron ajustadas usando la función lm() del software R. Las estimaciones de intercepto y pendientes de cada modelo se muestran en el Cuadro 4.2.

Cuadro 4.2: Estimaciones de los parámetros en cada modelo lineal.

MODELO		Estimate	Pr(>|t|)
Modelo 1	(Intercept)	-20.2621	0.0000
FVC	NDVI	95.7324	0.0000
Modelo 2	(Intercept)	0.5324	0.0000
C_FVC	NDVI	-0.6875	0.0000
Modelo 3	(Intercept)	0.7403	0.0000
C_exp	NDVI	-1.1064	0.0000

A pesar de que las estimaciones para cada modelo, tanto del intercepto como de la pendiente, son significativas, una parte importante dentro de la modelación que no debe ser olvidado, es la validez de los supuestos del modelo, para ello hicimos uso de los gráficos de residuales que proporciona el comando plot() en el software R. Los gráficos de valores ajustados y residuales mostraron un comportamiento particular que sugirieron el uso de otro tipo de modelo, por ejemplo, un modelo polinomial para ajustar los datos.
Las estimaciones de cada modelo usando modelos polinomiales se muestran en en Cuadro 4.3. Estas estimaciones son de igual forma significativas, con excepción de la pendiente del modelo 1 no lineal, por lo cual sólo se realiza el ajuste del modelo sin intercepto para el primer modelo polinomial. El ajuste y las predicciones dentro de cada modelo se muestran en los gráficos correspondientes en las Figuras 4.2, 4.3 y 4.4 Las estimaciones de los parámetros de los modelos no lineales son significativas y también poseen un buen ajuste ya que el coeficiente de determinación R^2 es igual 0.98. Con estas ecuaciones se puede calcular la cantidad de cobertura vegetal, en función del índice de vegetación NDVI que es obtenido por medio de la imagen de satélite considerada.

Finalmente, los modelos estimados se muestran a continuación:

$$FCV = 100 * NDVI^2 + \varepsilon \tag{4.9}$$

$$C_FCV = .82 - 2.44 * NDVI + 3.01 * NDVI^2 - 1.52 * NDVI^3 + \varepsilon \tag{4.10}$$

$$C_exp = 1.1 - 2.8 * NDVI + 1.8 * NDVI^2 + \varepsilon \tag{4.11}$$

Cuadro 4.3: Estimaciones de los parámetros en cada modelo no lineal.

MODELOS		Estimate	Pr(>|t|)
Modelo 1. FVC	(Intercept)	0.0000	0.5800
	NDVI	-0.0000	0.2917
	I(NDVI^2)	100.0000	0.0000
Modelo 2. C_FVC	(Intercept)	0.8272616	0.0000
	NDVI	-2.4437965	0.0000
	I(NDVI^2)	3.0196644	0.0000
	I(NDVI^3)	-1.5291057	0.0000
Modelo 3. C_exp	(Intercept)	1.1108	0.0000
	NDVI	-2.8568	0.0000
	I(NDVI^2)	1.8285	0.0000

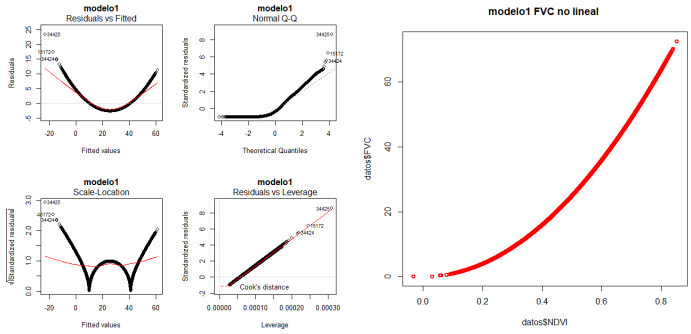


Figura 4.2: Gráfico del modelo ajustado: Modelo 1.

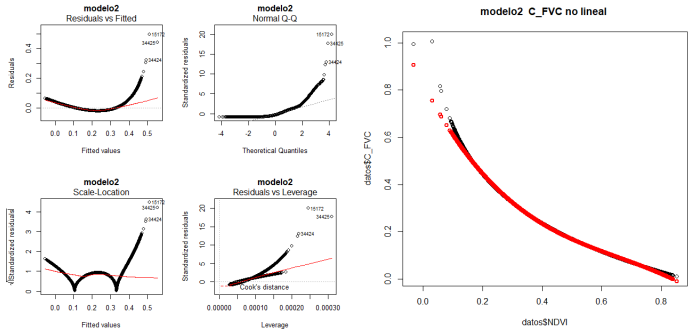


Figura 4.3: Gráfico del modelo ajustado: Modelo 2.

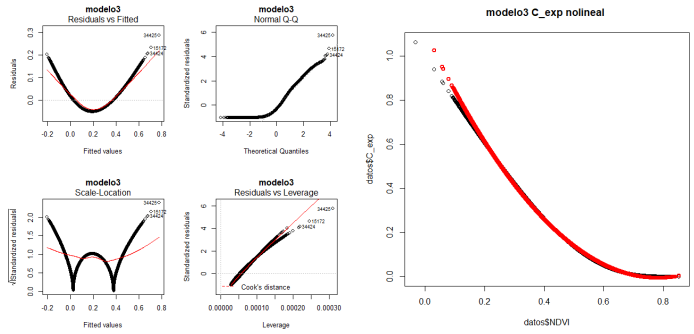


Figura 4.4: Gráfico del modelo ajustado: Modelo 3.

4.2. Secuestro de Carbono en la RTP 105

Se presenta la localización de la Región Terrestre Prioritaria 105 (RPT 105) que se ubica en las coordenadas extremas: $19^0 46' 23''$ y $20^0 11' 55''$ de latitud norte y $97^0 09' 17''$ a $97^0 38' 36''$ de longitud oeste y está conformada por 28 municipios de los que 4 pertenecen al estado de Veracruz y 24 en la Sierra Norte de Puebla. La Figura 4.5 muestra en color rojo la localización de la zona de estudio. Para el estudio del secuestro de

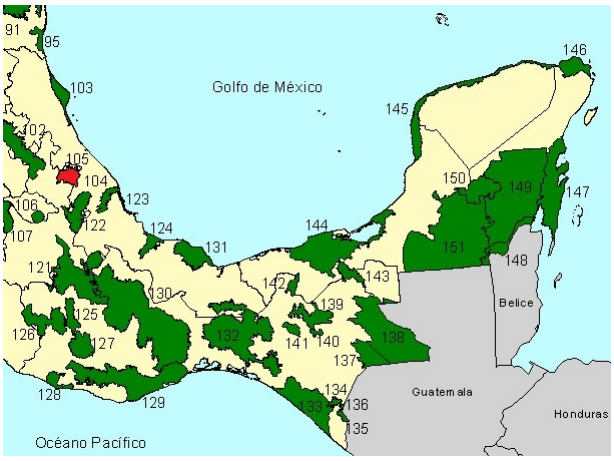


Figura 4.5: Localización de la Región Terrestre Prioritaria 105, por Arriaga en [2].

Carbono en suelos en la RTP-105, se buscaron imágenes con sensor ETM+ y TM. Las muestras de suelo (propiedades físico y químicas del suelo) habían sido tomadas en el año 2005, por lo que era de interés tomar imágenes de ese año y el satélite que se encontraba en órbita correspondía al satélite Landsat 7. Se descargaron las imágenes

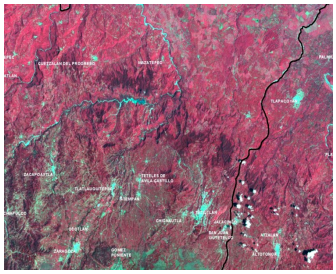


Figura 4.6: Imagen satelital de la zona de estudio.

del satélite Landsat 7 (2 de noviembre de 2005), que se obtuvo a través de USGS (Earth Resource Observation Systems Data Center). Se consideraron también las condiciones atmosféricas para que la información no se viera afectada por las condiciones climáticas.
Ya obtenidas las imágenes de satélite que corresponden a la banda 3 y 4 del satélite Landsat 7, se hizo uso del software GIS (Geographic Information System), para poder obtener el índice de vegetación NDVI.
Con la información disponible se realizaron dos análisis diferentes en la RPT 105, que se muestran a continuación.

4.2.1. Modelos LMM

Se utilizó la metodología de Zuur en [62] que está enfocada a obtener el mejor modelo mixto, basado en los criterios de selección de modelos AIC y BIC. A continuación, se muestra la metodología aplicada al estudio de la RTP-105:

PASO 1 Ajustar un modelo donde la componente fija contenga todas las variables explicativas e interacciones posibles. Dentro de las variables de la zona RTP-105, se realizó una selección de variables que aportaran más información al modelo. Los distintos modelos de regresión se realizarón usando la función lm() del software R, llegando a que el mejor modelo de efectos fijos es el que se muestra en el Cuadro 4.4.

Cuadro 4.4: Estimaciones del modelo modelo1.lm.

	Estimate	Std. Error	t value	Pr(>| t|)
Ntot	3.9109	0.3168	12.35	1.25e-11 ***

Tomando como base el modelo anterior, se introduce la variable NDVI mostrando sus estimaciones en el Cuadro 4.5.

Cuadro 4.5: Estimaciones del modelo modelo2.lm.

	Estimate	Std. Error	t value	Pr(>| t|)
Ntot	9.545	1.950	4.894	6.81e-05 ***
NDVI	5.382	2.204	2.442	0.0231 *

PASO 2 Teniendo como base el mejor modelo con las componentes fijas, ahora continuamos con el segundo paso de la metodología de Zuur, sin embargo, en este estudio sólo se requiere ver las variaciones en F_V (formación vegetal), como variable aleatoria que queremos introducir dentro del modelo, por lo cual el modelo en este paso de la metodología de de Zuur, para el caso de la Zona RTP-105 queda como: parte fija las variables del modelo `modelo2.lm` y parte aleatoria F_V.

PASO 3 Se busca 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.

Se construyen distintos modelos, donde se varía la componente aleatoria, introduciéndola como una intercepto, pendiente, que afecta a los distintos modelos, para buscar la mejor estructura de modelo con componente fija y componente aleatoria.

Para realizar los distintos modelos mixtos, en este trabajo se usan la librería library(lme4), library(Matrix), library(Rcpp)y library(nlme). Estas librerías, contienen distintas funciones que nos ayudan a poder estimar los efectos de las variables fijas y aleatorias en modelos mixtos. Los siguientes modelos mixtos contienen la misma componente aleatoria, pero buscando la mejor componente fija. Para seleccionar el mejor modelo en esta parte, usamos los criterios de selección de modelos AIC y BIC.

Cuadro 4.6: Críterios de selección de modelos.

	df	AIC	BIC
mixed.model1	4.00	38.34	42.70463
mixed.model2	3.00	50.49	53.89353
mixed.model3	2.00	55.51	57.86206

En el Cuadro 4.6, se observan los criterios para seleccionar el mejor modelo. El modelo *mixed.model*1 es seleccionado por tener los valores más pequeños de los criterios, con un valor de $AIC = 38.34$ y $BIC = 42.70463$.

PASO 4 El último paso consiste en presentar el modelo final utilizando un estimador REML y analizar las suposiciones establecidas en el mismo.

Modelo lineal Mixto seleccionado

```
mixed.model1<- lme(log(Corg)~-1+Ntot+NDVI, random = ~1| F_V)
```

Como parte final de la metodología de Zuur se muestra la salida completa de las estimaciones del modelo, usando el método REML.

```
Linear mixed model fit by REML t-tests use Satterthwaite
approximations to degrees of
  freedom [lmerMod]
Formula: log(Corg) ~ -1 + Ntot + NDVI + (1 | F_V)
Random effects:
 Groups   Name        Variance Std.Dev.
 F_V      (Intercept) 0.1654   0.4067
 Residual             0.1528   0.3909
Number of obs: 24, groups:  F_V, 6

Fixed effects:
     Estimate Std. Error      df t value Pr(>|t|)
Ntot   1.9082     0.4780 21.7410   3.992 0.000627 ***
NDVI   2.2151     0.5194 21.9890   4.264 0.000317 ***

Correlation of Fixed Effects:
     Ntot
NDVI -0.569
```

Con los siguientes comandos en R, se realizan los gráficos, para analizar las suposiciones del modelo `mixed.model1`.

```
Res <- residuals(mixed.model1, type="normalized")
Fit <- fitted(mixed.model1)
par(mfrow=c(2,2))
plot(Res ~ Fit, xlab="Fitted values", ylab="Residuals",
     main="Residuals vs. fitted")
abline(h=0)
plot(Res ~ datos$F_V, xlab="FORMACIÓN VEGETAL", ylab="Residuals",
     main = "FORMACIÓN VEGETAL")
abline(h=0)
hist(Res, main="Histogram of residuals", xlab="Residuals")
qqnorm(Res)
qqline(Res)
```

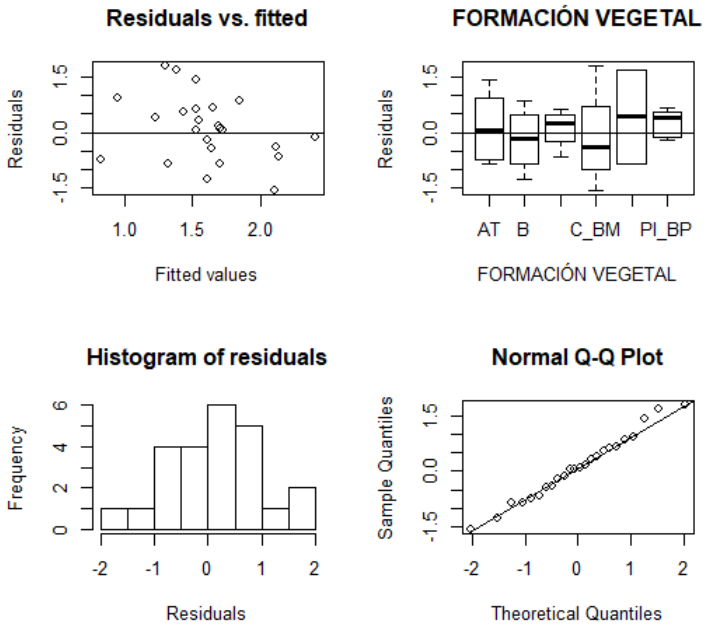


Figura 4.7: Suposiciones del modelo mixed.model1.

La forma del modelo mixto ajustado se muestra a continuación:

$$log(Corg_{ij}) = \beta_1 * Ntot_{ij} + \beta_2 * NDVI_{ij} + a_j + \epsilon_{ij}. \tag{4.12}$$

El índice j corresponde a las distintas Formaciones vegetales (F_V) que toma valores de 1 a 6, e i representa las muestras dentro de cada F_V.

La parte de los resultados que se refiere a los efectos aleatorios nos muestra que la varianza residual es $\sigma^2 = .39^2$ y la varianza de la constante $\sigma_a^2 = .40^2$. Para la parte de los efectos fijos del modelo, $\beta_1 * Ntot + \beta_2 * NDVI_{ij}$, la constante se estima con el valor $\beta_1 = 1.90$ y la segunda en $\beta_2 = 2.21$. Ambos parámetros son significativamente distintos de 0. Una observación importante es sobre la correlación entre las constantes es de $-.50$.

El Ntot y NDVI tienen efecto significativo sobre el logaritmo del Carbono orgánico, observando en la salida del Software R que en hay en promedio más cantidad de Carbono orgánico por el índice de vegetación, mientras que el Ntot hay ligeramente menos. El modelo es bastante adecuado y se cumplen los supuestos que se observan en la Figura 4.7.

A continuación mostramos en el Cuadro 4.7, los valores correspondientes a los datos reales y valores ajustados con el modelo mixed.model1. En la Figura 4.8 se observa el correspondiente gráfico, mostrando que el ajuste de los datos es aproximado a los datos reales , donde el modelo seleccionado cumple con los supuestos del modelo.

Cuadro 4.7: Porcentaje de Carbono Orgánico y valores ajustados.

Valores Reales	Valores Ajustados
10.50	11.048003
4.00	3.366455
5.20	4.125301
4.60	4.920010
3.70	2.571466
1.70	2.258541
6.70	5.109541
3.90	5.417217
5.30	4.655223
5.80	5.413445
7.0	8.305574
7.40	3.710805
5.90	4.595697
4.46	8.257774
7.71	3.895440
2.70	3.709559
6.53	8.416806
8.00	4.569030
5.70	5.576224
4.70	4.598780
5.70	5.547538
8.87	6.405834
4.38	5.214449
3.02	5.056493

En la Figura 4.8 se muestra que el modelo ajustado que resulta ser apropiado debido a que cumple los supuestos del modelo y los valores ajustados son aproximados a los valores reales. Así estos modelos pueden tomarse como referencia para otras zonas RTP-105, con el objetivo de determinar la cantidad de Carbono orgánico almacenada dentro de la zona de estudio.

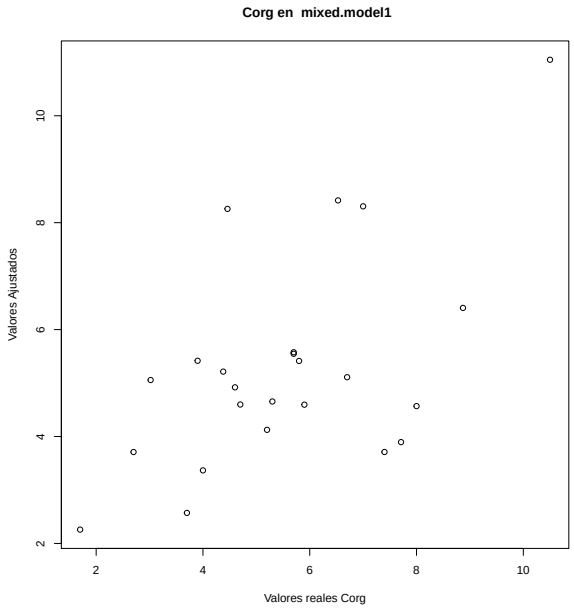


Figura 4.8: Comparación de valores reales y ajustados.

4.2.2. Estimación del porcentaje de Carbono orgánico en suelos de la RTP 105, Cuetzalan, México

Con el propósito de hacer estudios de secuestro de Carbono en suelo, especialistas del Departamento de Investigación en Ciencias Agrícolas (DICA) de la Benemérita Universidad Autónoma de Puebla (BUAP) vienen realizando investigaciones desde años anteriores(ver [9], [26], [27], [36] y [55]). En la década de los noventa obtuvieron 51 muestras de suelo. Estas muestras, fueron procesadas en el laboratorio para obtener diferentes propiedades físico y químicas, entre las que se tienen el porcentaje de Carbono orgánico (Corg), el porcentaje de Nitrógeno total (Ntot) y la Densidad Aparente (DA). El trabajo más reciente en esta zona de estudio corrresponde a [38], que se detalla en esta sección.

Por otra parte, se obtuvieron imágenes satelitales de la zona, lo que permitió, a través del índice de vegetación NDVI, estimar la densidad en la cobertura vegetal del año

1994 (ver Figura 4.9). A través de este índice fue posible identificar la presencia de vegetación verde en la superficie y caracterizar su distribución espacial. La actividad fotosintética permitió clasificar la densidad de vegetación de la zona en tres clases: baja, media y alta. Estas clases concuerdan con la cobertura vegetal observada en campo.

Modelación bajo el supuesto de normalidad de los errores

Para la estimación del secuestro de Carbono en la zona bajo estudio, se especificaron varios modelos lineales mixtos. Como señalamos anteriormente, los modelos mixtos son usados cuando los datos tienen algún tipo de agrupación, como es el caso que nos ocupa, donde se tienen tres niveles de densidades de vegetación (baja, media y alta), determinados por el índice de vegetación NDVI. Los modelos mixtos permiten tener coeficientes fijos y aleatorios (cuyos niveles son sólo una realización de todos los posibles niveles procedentes de una población) y varios términos de error. Para los efectos fijos se utilizaron, según el modelo especificado, las variables DA y Ntot, elegidas según criterios de especialistas en edafología y estudio exploratorios previos [26]. Se utilizó como variable dependiente o respuesta para los modelos el Corg. El

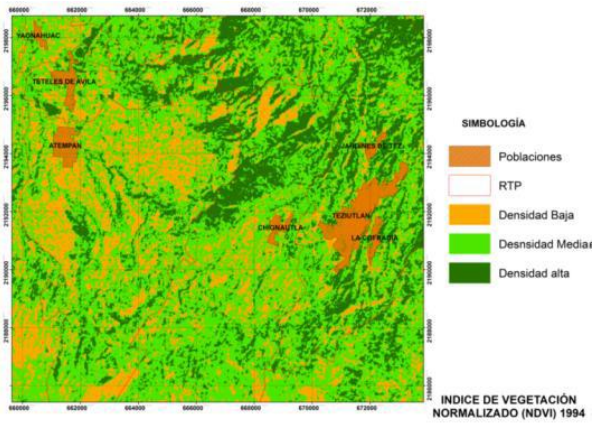


Figura 4.9: Imágenes con el cálculo de NDVI para el año 1994.

procesamiento estadístico se realizó en lenguaje R [54]. El método de estimación que brinda el paquete estadístico lme4 en el lenguaje R, y que se ha usado aquí, es: primero estimar los parámetros de varianza y covarianza usando REML y, posteriormente, utilizar estimadores mínimos cuadrados generalizados para los parámetros de la parte sistemática. Debe destacarse que el método ML puro es favorable en situaciones donde se requieran pruebas de hipótesis para comparar modelos, como en este estudio.
Se establecieron dos modelos diferentes bajo el supuesto de normalidad de los errores

y sus parámetros se estimaron con los métodos REML y ML, haciendo uso de la paquetería lme4 [5]. Los resultados se muestran en los Cuadros 4.8 y 4.9.

Cuadro 4.8: Métodos de estimación bajo el supuesto de distribución normal de los errores (Modelo 1).

Método REML			Método ML		
Modelo 1:			Modelo 1:		
Corg $\sim$ DA +(1 | NDVI)			Corg $\sim$ DA +(1 | NDVI)		
Efectos aleatorios:			Efectos aleatorios:		
Grupos	Nombre	Varianza	Grupos	Nombre	Varianza
NDVI	(Intercepto)	3.695e-05	NDVI	(Intercepto)	1.067e-05
Residuo		8.182e-04	Residuo		8.000e-04
Número de observaciones:			Número de observaciones:		
51, Grupos: NDVI, 3			51, Grupos: NDVI, 3		
Efectos fijos:			Efectos fijos:		
	Estimador			Estimador	
(Intercepto)	0.13556		(Intercepto)	0.13465	
DA	-0.12389		DA	-0.12316	
Correlación de efectos fijos:			Correlación de efectos fijos:		
	(Intercepto)			(Intercepto)	
DA	-0.974		DA	-0.982	

También se utilizaron los Criterios de Información para la comparación de modelos que brinda el paquete estadístico lme4. Los más comúnmente utilizados son el criterio AIC y el Criterio BIC. En la literatura consultada para estos y otros criterios de selección de modelos que aparecen en las salidas del programa lme4, se elige como mejor modelo aquel con los valores más pequeños del criterio. En el Cuadro 4.10, se muestran los resultados del procedimiento para la comparación de ambos modelos, reajustándolos con ML en vez de REML. El resultado de la prueba χ^2 es significativo al nivel de significancia .05, lo que apunta a considerar que el modelo 2 es mejor. Los criterios AIC y BIC también señalan que este modelo es ligeramente mejor. El modelo 2 fue ajustado usando el estimador REML, que es recomendado por tener un algoritmo más eficiente, puede ser expresado como

$$y_{ij} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + b_{0i} + \varepsilon_{ij}. \tag{4.13}$$

donde y_{ij} es el porcentaje de Carbono orgánico de la $i-$ésima muestra en el j$-$ésimo nivel del índice de vegetación NDVI. Las variables independientes de efectos fijos son: X_1 que representa DA, y X_2 Ntot y la variable de efecto aleatorio es NDVI agrupada en tres niveles. El error ε_{ij} se supone que se distribuye normal con varianza σ^2.

Cuadro 4.9: Métodos de estimación bajo el supuesto de distribución normal de los errores (Modelo 2).

Método REML			Método ML		
Modelo 2:			Modelo 2:		
COrg ∼ DA +Ntot+ (1 | NDVI)			COrg ∼ DA +Ntot +(1 | NDVI)		
Efectos aleatorios:			Efectos aleatorios:		
Grupos	Nombre	Varianza	Grupos	<Nombre	Varianza
NDVI	(Intercepto)	5.683e-05	NDVI	(Intercepto)	2.579e-05
Residual		7.531e-04	Residuo		7.210e-04
Número de observaciones:			Número de observaciones:		
51, Grupos: NDVI, 3			51, Grupos: NDVI, 3		
Efectos fijos:			Efectos fijos:		
	Estimador			Estimador	
(Intercepto)	0.128136		(Intercepto)	0.12762	
DA	-0.120814		DA	-0.12038	
Ntot	0.014284		Ntot	0.01404	
Correlación de efectos fijos:			Correlación de efectos fijos:		
	(Intercepto)	DA		(Intercepto)	DA
DA	-0.963		DA	-0.982	
Ntot	-0.155	0.054	Ntot	-0.156	0.053

Los estimadores de los parámetros son β_0= 0.1281364, β_1= -0.120814, β_2= 0.014284 y σ^2=7.531e-04. Estos valores destacan, como era de esperarse, la importancia de la densidad aparente para la explicación del porcentaje de Carbono orgánico en estos suelos, lo que se manifiesta también al observar la alta correlación negativa con el intercepto. Por otra parte, los residuos estandarizados apuntan a que su distribución no es simétrica ya que el mínimo es aproximadamente -2 mientras que el máximo es aproximadamente 4.
Dado que en el método de estimación ML se asume la normalidad en la etapa de ajuste del modelo, se requiere de la comprobación de las suposiciones de este modelo. En la Figura 4.10 se muestran los gráficos de residuos y prueba de normalidad. Puede apreciarse que los datos muestran una estructura no normal y presencia de heteroscedasticidad.

Como es conocido, en el pasado, las únicas herramientas disponibles para tratar la ausencia de normalidad eran la transformación de la variable respuesta o la adopción de métodos no paramétricos. Hoy en día, existen otras alternativas, que son los GLMM. Estos modelos permiten especificar distintos tipos de distribución de los errores como

Cuadro 4.10: : Selección de modelos.

	gl	AIC	BIC	logLik	χ^2	$gl - \chi^2$	Pr($>\chi^2$)
Modelo1	4	-210.3	-202.6	109.1			
Modelo2	5	-212.8	-203.1	111.4	4.5	1	0.03

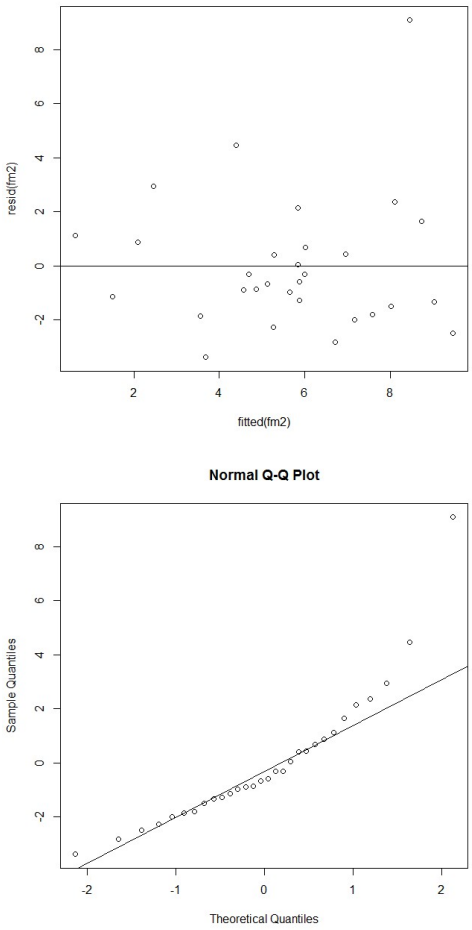


Figura 4.10: Gráficos de residuos bajo el supuesto de distribución normal.

puede ser la distribución gamma. Esta distribución es útil con datos que muestran un coeficiente de variación constante, esto es, en donde la varianza aumenta según aumenta la media de la muestra de manera constante, lo que se cumple en este caso [10].

Modelación bajo el supuesto de distribución gamma de los errores

En [38] se aborda un poco sobre los modelos con variable respuesta acotada sobre el intervalo (0,1), considerando que una de las distribuciones recomendable a considerar corresponde a una distribución beta (ver más en [50]). En este caso de estudio se obtuvieron mejores resultados considerando una distribución gamma. Bajo este supuesto y considerando los enlaces "log" y "probit" que brinda la función glmer del paquete lme4 del software R, se obtuvieron otros modelos para explicar Corg en función de la DA y del Ntot como efectos fijos y la densidad de vegetación como efecto aleatorio (NDVIA, NDVIB, NDVIM). Utilizando la notación del paquete lme4 este modelo puede expresarse como el modelo 2 antes considerado:

$$Corg \sim 1 + DA + Ntot + NDVI. \tag{4.14}$$

Donde,
DA y $Ntot$, son las variables fijas y
$NDVI$, es la variable aleatoria que se quiere relacionar con el Carbono orgánico. En el Cuadro 4.11 se muestran los resultados obtenidos. Con el enlace "log" se obtuvo la convergencia en 7 iteraciones y las pruebas t se basaron en 47 grados de libertad (gl), mientras que el enlace "probit" convergió en 8 iteraciones y las pruebas se basaron en 48 gl. Todos los estimadores de los efectos fijos fueron significativos a diferentes niveles de significación en ambos enlaces, lo que corrobora la importancia de la DA y del Ntot en la explicación del Corg. Estos modelos pueden considerarse más adecuados que los obtenidos bajo el supuesto de normalidad de los errores del modelo, ya que no se presentan incumplimientos de las suposiciones.

Se presentaron dos casos de estudios, donde el NDVI, que se obtiene por medio de imágenes de satélite, resultó ser significativo en ambos modelos. En el primer caso se determinaron modelos de regresión polinómicos para mejorar el ajuste de los modelos, debido a que los datos tienen comportamientos curvilíneos. En el segundo caso los modelos lineales mixto resultaron ser mejor para ajustar y poder determinar la cantidad de Carbono orgánico dentro de la zona RTP-105.
Existe reconocimiento internacional del papel de la modelación estadística en la investigación medioambiental, la cuantificación de la incertidumbre presente en los problemas ambientales es uno de los retos actuales de la Ciencia Estadística, que junto al desarrollo de la Teledetección y la Ciencia de la Computación, se enfocan a brindar mejores explicaciones a tan complejos problemas, no viéndolos como un fin meramente técnico, sino como un medio necesario para mejorar las condiciones de vida de millones

Cuadro 4.11: Método de estimación bajo el supuesto de distribución gamma de los errores.

Familia =Gamma("log")			Familia= Gamma("probit")		
MODELO DE MEDIA			MODELO DE MEDIA		
Estimadores de efectos fijos:			Estimadores de efectos fijos:		
	Estimador	Pr(>|t|)		Estimador	Pr(>|t|)
Interc	-1.4442	0.031017 *	Interc	-1.3805	1.75e-05***
DA	-2.7823	0.001657 **	DA	-1.0156	0.00491 **
Ntot	0.6766	0.000645 ***	Ntot	1.1379	2.41e-05 ***
Estimadores de efectos aleatorios:			Estimadores de efectos aleatorios:		
	Estimador	Error Estándar		Estimador	Error Estándar
NDVIA	0.0254	0.1110	NDVIA	.0003	0.0437
NDVIB	0.0455	0.1090	NDVIB	0.0222	0.0428
NDVIM	-0.0709	0.1078	NDVIM	-0.0224	0.0424
MODELO DE DISPERSIÓN			MODELO DE DISPERSIÓN		
Parámetro de dispersión :			Parámetro de dispersión para		
para el modelo de media:			el modelo de media:		
0.6057283			0.5107561		
Efectos:			Efectos:		
Estimador	Error Estándar		Estimador	Error Estándar	
-0.5013	0.2053		-0.6719	0.2052	
Parámetro de dispersión para:			Parámetro de dispersión para		
el efecto aleatorio:			el efecto aleatorio:		
0.01408			0.002061		
Efectos:			Efectos:		
.|Aleatorio1			.|Aleatorio1		
Estimador	Error Estándar		Estimador	Error Estándar	
-4.2631	1.9066		-6.1846	2.0375	

de personas.

Conclusiones

El campo de la modelación es una herramienta útil en distintas áreas que permite adentrarnos a la estadística inferencial, con el objetivo de estimar parámetros que desconocemos y que estamos interesados en conocer. Esta disciplina cada vez se va ampliando permitiendo considerar más características con datos reales.

Como primera aplicación, se usa la propuesta de Demidenko que proporciona una alternativa para adentrarte a la modelación estadística con el uso de imágenes no satelitales, donde se considera una parte importantes la comparación entre y dentro de cada imagen.
El primer resultado que se obtuvo es por medio de la prueba de Kolmogorov-Smirnov, que permitió comparar imágenes no satelitales de un experimento con peces cebra, concluyendo que sí existe diferencias entre las imágenes. Este resultado nos permite concluir que los peces sometidos a distintas soluciones de agua que se descarga en el Río Atoyac, afecta la parte física y morfológica de los peces cebra.
Una segunda aplicación se refiere a el uso de información de la presa de Valsequillo. Se ajustaron tres modelos lineales, uno para cada variable: índice de vegetación relativo, factor del índice de vegetación relativo y factor de las prácticas de cultivo, tomadas como variables dependientes y como variable independiente el índice de vegetación de diferencia normalizada.
El análisis de índices de vegetación en la Región Terrestre Prioritaria 105: Cuetzalan, Puebla, México, mediante el uso de sensores remotos, permitió visualizar el contraste entre tres clases de densidad de vegetación, baja, media y alta, mostrando que las observaciones tienen un comportamiento de agrupación y, dado que la variable respuesta es una proporción, los Modelos Lineales Generalizados Mixtos permitieron una modelación más adecuada que la práctica tradicional de considerar normalidad en los errores. Se obtuvieron modelos de regresión gamma de efectos mixtos que estiman el porcentaje de Carbono orgánico en función de propiedades fisicoquímicas del suelo (efectos fijos) y del índice de vegetación Normalizado considerado como efecto aleatorio, con los enlaces "log" y "probit".
Se llegaron a tres conclusiones principales:
(1) Se ajustaron modelos lineales Generalizados mixtos para determinar el porcentaje de Carbono orgánico y el tipo de vegetación dentro de una zona llamada RTP-105. La mayoría de los modelos ajustados dentro de la zona de estudio se limitan a Modelos Lineales y el modelo más actual usa modelos lineales mixtos, pero los modelos ajustados hasta ahora solo consideran variables como: propiedades físico, químicas y vegetación de forma fija y no consideran los efectos aleatorios como la formación vegetal. Muchos trabajos que hasta ahora se tienen dentro de la zona de estudio, podrían considera los factores específicos de la región como efectos aleatorios que podrían mejorar el ajuste de los modelos, las variables que pueden incluirse en este tipo son geología, clima, efectos del suelo y sitios de muestreo.

(2) Se añadió una variable índice de vegetación de diferencia normalizada que es obtenida por medio de imágenes de satélites basada en los píxeles de las imágenes Landsat, y se ajustaron modelos que incluyeron esta variable. La obtención de la variable indica la cantidad de vegetación en la zona de estudio y proporciona información de vegetación que no sólo se limita a zonas pequeñas, sino también a obtener información de zonas más grandes que requieren menos inversión de tiempo y dinero que dentro de la literatura se denomina Teledetección, a comparación de muestreos y análisis de información de campo. Esto proporciona otro enfoque para estimar la cantidad de Carbono orgánico en función de información de campo, el índice de vegetación de diferencia normalizada (variable fija que es obtenida por medio de imágenes satelitales) y formación vegetal (variable aleatoria).

(3) Se ajustaron Modelos Lineales Mixtos y Modelos Lineales Generalizados Mixtos, seleccionando el mejor modelo dependiendo el tipo de información que se tiene en la variable respuesta en la zona Terrestre Prioritaria-105, en función de información de campo, índice de vegetación de diferencia normalizada y formación vegetal . Estos modelos ayudan a determinar cantidades de Carbono orgánico almacenado en la zona errestre Prioritaria-105, incluyendo información que es parte importante dentro del ciclo del carbono orgánico.

Dentro de los trabajos a futuros se pretende escribir en forma de metodología el Modelo Lineal Generalizado Mixto usando una distribución beta, que se considera una parte muy usada en distintas áreas al considerar una variable que se encuentra dentro del intervalo unitario. En particular se pretende modelar el porcentaje de Carbono orgánico considerando dicha distribución y esperando tener mejores resultados en los ajustes. Esto se debe al tipo de variable respuesta y considerando la información de la variable que muchas veces está dada dentro de un intervalo (0,1).
En cuanto a las imágenes no satelitales existen muchos campos de aplicación, en particular se desea trabajar en campo la medicina, donde los modelos Lineales Mixtos y Modelos Lineales Generalizados Mixtos generan una herramienta para la investigación. También, para dar a conocer la utilidad y la mejoría de ajustes al considerar estos modelos en otras áreas.

Índice de figuras

Índice de cuadros

Apéndices

Apéndice A

Propiedades

A.1. Consistencia

La consistencia es una propiedad de ciertos estimadores, se dice que un estimador es consistente cuando éste converge a su valor verdadero cuando el número de datos de la muestra tiende a infinito.

A.2. Propiedades de Varianza

Para el cálculo de la varianza de y, se usa con frecuencia la ecuación de la varianza y el valor esperado condicional, se usa la siguiente propiedad para particionar la variabilidad:

$$var(y) = var(E[y|u]) + E[var(y|u)] \tag{A.1}$$

A.3. Teorema de Gauss Markov

Lo que buscamos en demostrar que $\hat{\beta}$ minimiza la varianza para cualquier combinación lineal de los coeficientes estimados, $\ell'\hat{\beta}$.
Calculamos la varianza de la combinación lineal $\ell'\hat{\beta}$.

$$Var(\ell\hat{\beta}) = \ell' Var(\hat{\beta})\ell = \ell'\sigma^2(X'X)^{-1} = \sigma^2\ell'(X'X)^{-1}\ell \tag{A.2}$$

donde el resultado anterior es un escalar ([58], pág. 540).

A.4. Función generadora de momentos

Sea X una variable aleatoria. El valor esperado

$$m_X(t) = E[\exp(tX)] \quad -c \leq t \leq c \tag{A.3}$$

recibe el nombre de función generadora de momentos.
Si X es una v.a. discreta

$$m_X(t) = E[\exp(tX)] = \Sigma_x exp(tx) * p(x) \tag{A.4}$$

Si X es una v.a. continua

$$m_X(t) = E[\exp(tX)] = \int_{-\infty}^{\infty} exp(tx) * p(x) \tag{A.5}$$

A.5. Distribuciones relacionadas con la distribución normal

A.5.1. Distribución normal

1. Si la variable aleatoria Y tiene la distribución normal con media μ y varianza σ^2, su función de densidad de probabilidad es que se denota por $Y \sim N(\mu, \sigma^2)$

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-[y-\mu]^2/2\sigma^2\}$$

2. La distribución normal con $\mu = 0$ y $\sigma^2 = 1$, $Y \sim N(0, 1)$, es llamada la **distribución normal estandar**.

3. Sea $Y_1, Y_2, \ldots, Y_n$ denota las variables aleatorias distribuidas normalmente con $Y \sim N(\mu_i, \sigma 2)$ para $i = 1, \ldots, n$ y sea las covariables de Y_i y Y_j que se denota por

$$Cov(Y_i, Y_j) = \rho_{ij}\sigma_i\sigma_j,$$

donde ρ_{ij} es el coeficiente de correlación para Y_i y Y_j

4. Suponga que las variables aleatorias $Y_1, Y_2, \ldots, Y_n$ son independientes y distribuidos normalmente con las distribuciones $Y \sim N(\mu_i, \sigma_i^2)$ para $i = 1, \ldots, n$. Si

$$W = a_1Y_1 + a_2Y_2 + \ldots + a_nY_n$$

donde los $a_i's$ son constantes. Entonces W es también distribuida normalmente, así que

$$W = \Sigma_{i=1}^n a_iY_i \sim N(\Sigma_{i=1}^n a_i\mu_i, \Sigma_{i=1}^n a_i^2\sigma_i^2)$$

A.5.2. Distribución Chi cuadrada

1. La distribución chi cuadrada centra con n grados de libertad está definida como la suma de cuadrados de n variables aleatorias independientes $Z_1, \ldots, Z_n$ cada una con distribución normal estándar. Se denota por
$$X^2 = \Sigma_{i=1}^n Z_i^2 \sim \chi^2(n).$$
2. Si X^2 tiene la distribución $\chi^2(n)$, entonces el valor esperado es $E(X^2) = n$ y su varianza es $var(X^2) = 2n$.
3. Si $Y_1, Y_2, \ldots, Y_n$ son variables aleatorias independientes distribuidas normalmente cada una con la distribución $Y_i \sim N(\mu_i, \sigma_i^2)$ entonces
$$\Sigma_{i=1}^n X^2 = \Sigma_{i=1}^n (\frac{Y_i - \mu_i}{\sigma_i})^2 \sim \chi^2(n)$$
porque cada variable $Z_i = \frac{Y_i - \mu_i}{\sigma_i}$ tiene una distribución normal $N(0,1)$.

A.5.3. Distribución t

La distribución t con n grados de libertad está definida como el radio de dos variables aleatorias independientes. El numerador tiene distribución estándar y el denominador es la raíz de una v. a. chi- cuadrada central dividida por sus grados de libertad: esto es,
$$T = \frac{Z}{(X^2/n)^{1/2}} \quad denotadoporT \sim t(n).$$
donde $Z \sim N(0,1)$, $X^2 \sim \chi^2(n)$, tanto Z y X^2 son independientes.

A.5.4. Distribución F

La distribución central con n y m grados de libertad está definida como el radio de dos variables aleatorias independientes chi cuadradas centrales cada una dividida por sus grados de libertad
$$F = \frac{X_1^2/n}{X_2^2/m} \quad denotadoporF \sim F(n,m) \tag{A.6}$$
donde $X_1^2 \sim \chi^2(n)$, $X_2^2 \sim \chi^2(m)$ y X_1^2 y X_2^2 son independientes.

A.6. Otras distribuciones.

A.6.1. Distribución Gamma

La distribución gamma está definida por dos parámetros, α y θ, y su función de densidad se puede escribir como:

$$f(u;\alpha,\theta) = \frac{u^{\alpha-1}\exp -u/\theta}{(\theta^{\alpha} * \Gamma(\alpha))}, \quad para \alpha > 0. \tag{A.7}$$

Para α un entero positivo, se tiene $\Gamma(\alpha) = (\alpha - 1)!$ Una referencia completa de los GLMM considerando una distribución gamma en la variable respuesta, podemos citar a Rossi en [49].

A.6.2. Distribución Beta

La distribución beta está definida por dos parámetros, p y q, y su función de densidad se puede escribir como:

$$f(u;p,q) = \frac{1}{(\beta(p,q))x^{p-1}(1-x)^{q-1}}, \quad para 0 < x < 1 \quad \beta(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx \tag{A.8}$$

Apéndice B

Programa en R con la prueba Kolmogorov-Smirnov.

```
KSimage <-function(job=1)
{
dump("KSimage2","C:/Users/hp-pavilion/Dropbox/PECES/KSimage.r")
if(job==1)
{
par(mfrow = c(1, 2), mar = c(1, 1, 3, 1), omi = c(0, 0, 0, 0))
d <- scan("C:/Users/hp-pavilion/Dropbox/PECES/IMAGE117.pgm",what="")
##lectura de la imagen
d <- as.numeric(d[2:length(d)])
#descartamos el primer valor, ya que interesa  el valor de los pixeles.
nr <- d[1];nc <- d[2];# dimensiones de la imagen
d <- matrix(d[4:length(d)], nrow = nr, ncol = nc)
#convierte la imagen en matriz
image(1:nr, 1:nc, d, xlab = "", ylab = "", axes = F,col=gray(0:255/255))
mtext(side = 3, "Control", line = 0.25, cex = 2)
d <- scan("C:/Users/hp-pavilion/Dropbox/PECES/IMAGE126.pgm",what="")
d <- as.numeric(d[2:length(d)])
nr <- d[1];nc <- d[2];
d <- matrix(d[4:length(d)], nrow = nr, ncol = nc)
image(1:nr, 1:nc, d, xlab = "", ylab = "", axes = F,col=gray(0:255/255))
mtext(side = 3, "Experiment", line = 0.25, cex = 2)
}
if(job==2)
{
par(mfrow = c(1, 2), mar = c(4, 4, 3, 1))
```

```
d <- scan("C:/Users/hp-pavilion/Dropbox/PECES/IMAGE117.pgm",what="")
d <- as.numeric(d[2:length(d)])
nr <- d[1]
nc <- d[2]
J1 <- nr * nc
d1 <- d[1:length(d)]
d <- scan("C:/Users/hp-pavilion/Dropbox/PECES/IMAGE126.pgm",what="")
d <- as.numeric(d[2:length(d)])
nr <- d[1]
nc <- d[2]
print(c(nr, nc))
J2 <- nr * nc
d2 <- d[1:length(d)]
h1 <- f1 <- h2 <- f2 <- rep(0, 256)
for(i in 0:255) {
h1[i + 1] <- length(d1[d1 == i])/length(d1)
h2[i + 1] <- length(d2[d2 == i])/length(d2)
}
for(i in 2:256) {
f1[i] <- f1[i - 1] + h1[i]
f2[i] <- f2[i - 1] + h2[i]
}
f1[256] <- f2[256] <- 1
matplot(cbind(0:255, 0:255), cbind(h1, h2), type = "l",
 col = 1,xlab="",ylab="")
mtext(side = 2, "Probability", line = 2.5, cex = 1.75)
mtext(side = 3, "Histogram", line = 1, cex = 1.75)
matplot(cbind(0:255, 0:255), cbind(f1, f2), type = "l",
 col = 1,xlab="",ylab="")
lines(0:255, f1, lwd = 3)
lines(0:255, f2, lwd = 3, lty = 2)
mf <- max(abs(f1 - f2))
jm <- (0:255)[abs(f1 - f2) == mf]
segments(jm, -0.25, jm, 0.6)
text(jm+5, 0.63, paste("max=", round(mf, 3)),adj=0)
mtext(side = 3, "Distribution function", line = 1, cex = 1.75)
legend(0, 1, c("Control", "Experiment"), lty = 1:2, cex = 1.25,
lwd = 3)
mtext(side = 1, "Grayscale level, byte", line = -1, outer = T,
cex = 1.5)
J <- (J1 * J2)/(J1 + J2)
lambda <- mf * (sqrt(J) + 0.11/sqrt(J) + 0.12)
```

```
j <- 1:10000
js <- rep(1, 10000)
js[seq(from = 2, to = 10000, by = 2)] <- -1
Q <- 2 * sum(js * exp(-2 * j^2 * lambda^2))
cat("\nbyte.max =",jm," max.cdf.distance =",round(mf,3)," lambda =",
round(lambda,3)," QKS =", round(Q,4),"\n")
}
}
```

B.1. Salidas en R.

Primera comparación

>KSimage(job=1)
Read 1920004 items
Read 1920004 items
>KSimage(job=2)
Read 1920004 items
Read 1920004 items
[1]16001200
byte.max = 141 max.cdf.distance = 0.218 lambda = 213.308 QKS = 0

Segunda comparación

>KSimage(job=1)
Read 1920004 items
Read 1920004 items
>KSimage(job=2)
Read 1920004 items
Read 1920004 items
[1]16001200
byte.max = 119 max.cdf.distance = 0.259 lambda = 253.344 QKS = 0

Tercera comparación

>KSimage(job=1)
Read 776934 items
Read 776934 items
>KSimage(job=2)
Read 776934 items
Read 776934 items
[1]1009770

byte.max = 125 max.cdf.distance = 0.574 lambda = 357.809 QKS = 0

Apéndice C

Abreviaciones: Propiedades físico y químicas del suelo en la RTP-105

Número de muestra de suelo	Muestra
Notación que se toma según el experto en cada muestra	PERFIL
Latitud	X
Longitud	Y
Altitud	Altitud
Precipitación	Prec
Formación vegetal que contiene 8 clasificaciones	F_V
Profundidad	Prof
Horizonte	Hor
Densidad Aparente	DA
Arena	Arena
Limo	Limo
Arcilla	Arcilla
Textura(T1,T2,T3,T4,T31,...)	Textura
Matería Orgánica	MO
Carbono orgánico	Corg
Carbono orgánico en Suelos	COS
Nitrógeno Total	Ntot
Carbono Notrógeno	C_N
Capacidad de intercambio catiónico	CIC
Mag	Mg
Potasio	K
Calidad del suelo (Inherente, Dinámico)	Calidad_S

Muestra	F_V	Prof	Hor
1	PI_BP	3	A
2	PI_BP	32	B
3	PI_BP	4	A
4	PI_BP	35	B
5	PI_BP	6	A
6	PI_BP	24	B
7	PI_BP	5	A
8	PI_BP	35	B
9	P_AT	7.5	A
10	P_AT	36	B
11	P_AT	15	A
12	P_AT	46	C
13	PI_BP	10	A
14	PI_BP	31	B
15	M_AT	15	A
16	M_AT	32.5	B
17	PI_BP_E	5	A
18	PI_BP_E	20	B
19	PI_BP_E	21	A
20	PI_BP_E	70	B
21	C_BM	20	A
22	C_BM	73	B
23	C_BM	33.5	A
24	C_BM	97	B
25	BP_E	5	A
26	BP_E	17.5	B
27	C_BM	17.5	A
28	C_BM	55	B
29	PI	13	A
30	PI	51.5	B
31	PI	13	A
32	PI	33	C
33	PI_BP_E	13.5	A
34	PI_BP_E	38	C
35	M_AT	14.5	A
36	M_AT	45	C
37	M_AT	15	A
38	M_AT	17.5	B
39	M_AT	11	A
40	M_AT	33.5	B

Cuadro C.1: Propiedades físico y químicas del suelo en la RTP-105

DA	Corg	COS	Ntot
0.61	10.5	19.215	0.81
0.78	0.7	17.472	0.06
0.78	4	12.48	4.27
0.78	0.9	24.57	0.07
0.62	5.2	19.344	0.32
0.8	1.8	34.56	0.14
0.71	4.6	16.33	0.33
0.69	1.6	38.64	0.12
0.8	3.7	22.2	0.25
0.79	2.6	73.944	0.2
0.87	1.7	22.185	0.11
0.89	0.3	12.282	0.02
0.7	6.7	46.9	0.39
0.63	1.6	31.248	0.13
0.69	3.9	40.365	0.26
0.79	0.2	5.135	0.01
0.71	5.3	18.815	0.41
0.66	3.2	42.24	0.23
0.63	5.8	76.734	0.41
0.57	1.3	51.87	0.07
0.5	7	70	0.54
0.77	0.3	16.863	0.04
0.69	7.4	171.051	0.46
0.72	6.9	481.896	0.53
0.75	5.9	22.125	0.45
0.64	2.5	28	0.19
0.8	4.46	62.44	0.46
0.84	1.3	60.06	0.13
0.53	7.71	53.1219	0
0.66	3.47	117.9453	0
0.86	2.7	30.186	0.36
0.94	0.9	27.918	0.2
0.6	6.53	52.893	0.45
0.67	4.93	125.5178	0.5
0.75	8	87	0.16
0.89	0.8	32.04	0.09
0.75	5.7	64.125	0.35
0.61	3.4	36.295	0.25
0.78	4.7	40.326	0.35

Cuadro C.2: Propiedades físico y químicas del suelo en la RTP-105

C_N	CIC	Ca	Mg	Na	K	Calidad_S
13	19.4	3.3	1.6	0.2	0.6	Inherente
11	15.1	2.5	1.4	0.7	1.5	Inherente
15	13.55	2.9	1.1	0.4	0.6	Inherente
12	14.85	3.3	1.4	0.3	0.4	Inherente
16	19.25	5.4	2.8	0.9	1.4	Inherente
13	13.25	2.8	10	0.8	1.4	Inherente
14	13.2	3	2.3	0.4	0.5	Inherente
13	12.8	2.4	1.5	0.6	0.2	Inherente
15	16.4	3.4	0.6	0.4	0.5	Dinamico
13	15	4	1	0.4	0.5	Dinamico
15	21.4	3.4	0.9	1.4	1	Dinamico
17	14.3	2	0.6	0.9	0.6	Dinamico
17	33	2.3	0.5	0.4	0.3	Inherente
12	30.7	1.5	1	0.5	0.1	Inherente
15	21.3	3	2	0.8	0.6	Dinamico
13	25	5	2	0.6	1.6	Dinamico
13	19.4	2	1	0.35	0.5	Inherente
14	14.3	1.5	1	0.4	0.3	Inherente
14	34.6	3.5	1	0.2	0.1	Inherente
18	29	1.3	0.2	0.3	0.1	Inherente
13	35.8	5.6	1.4	2	0.8	Inherente
16	10.5	2	0.8	1.2	0.3	Inherente
16	16.8	4.1	1.3	2.8	0.3	Inherente
13	13.8	4.3	0.7	0.2	0.2	Inherente
13	34.2	9.7	3.1	2.2	4	Inherente
13	33.7	6.5	1.2	0.7	1.7	Inherente
9.69	30.5	1.06	0.21	1.6	0.3	Inherente
10	31.8	1	0.06	1.4	0.4	Inherente
7.38	29.1	0.84	0.21	0.62	0.08	Dinamico
4.35	38.5	0.42	0.22	1.1	0.23	Dinamico
7.38	11.05	2.7	1.3	0.4	0.31	Inherente
4.35	10.35	1.8	1.7	0.56	0.3	Inherente
14.5	29	5	3	0.6	0.6	Inherente
9.86	18.5	5.6	1.5	0.7	0.45	Inherente
13.3	48.4	3	1.2	0.18	0.14	Dinamico
8.89	53.2	3	1	0.31	0.38	Dinamico
14.72	44.2	4	1	0.2	0.13	Dinamico
13.6	39.9	2.1	1.6	0.2	0.16	Dinamico
13.42	27.3	2	1	0.37	0.31	Dinamico
9.33	29	1.6	0.7	0.37	0.31	Dinamico

Apéndice D

Onda

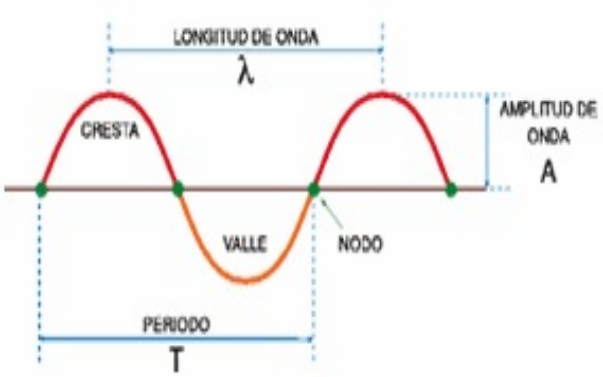


Figura D.1: Elementos de una onda.

Cresta: La cresta es el punto de máxima amplitud de onda.

Período (T): Es el tiempo requerido para que el movimiento de oscilación de la onda describa un ciclo completo.

Amplitud (A): La amplitud es la elongación máxima o altura de la cresta de la onda

Frecuencia (f): Número de veces que es repetida dicha vibración por unidad de tiempo.

$$T = \frac{1}{f}. \tag{D.1}$$

Valle: Es el punto más bajo de una onda.

Longitud de onda (λ): Es la distancia entre dos crestas o valles seguidos

Nodo: Es el punto donde la onda cruza la línea de equilibrio.

Elongación (x): es la distancia que hay, en forma perpendicular, entre un punto de la onda y la línea de equilibrio.

Ciclo: es una oscilación, o viaje completo de ida y vuelta.

Velocidad de propagación (v): es la velocidad a la que se propaga el movimiento ondulatorio. Su valor es el cociente de la longitud de onda y su período.

Apéndice E

Bandas Satélites Landsat 7 y 8

Landsat 7

Landsat 7 es el séptimo de un grupo de satélites lanzados por Estados Unidos fue lanzado el 15 de abril de 1999. El objetivo principal del Landsat 7 es actualizar la base de datos de imágenes de todo el planeta Tierra sin nubes. Aunque el programa Landsat es controlado por la NASA, las imágenes recibidas por el Landsat 7 son procesadas por el Servicio Geológico de los Estados Unidos (USGS por sus siglas en inglés). Las bandas de su sensor ETM+ se muestran en la Figura E.1.

Landsat 8

Landsat 8 fue lanzado el 11 de febrero de 2013. Es el octavo satélite en el programa Landsat; el séptimo para llegar a órbita con éxito. Las bandas de su sensor ETM se muestran en la Figura E.2.

De sus 11 bandas, sólo aquellos en las longitudes de onda más cortas (bandas 1-4 y 8) de luz visible , todos los demás están en las partes del espectro que no podemos ver. La vista de color verdadero del Landsat es menos de la mitad de lo que se ve.

Cuadro E.1: Bandas sensor ETM+, obtenida en la página oficial Landsat Science NASA.

Sensor	Número de banda	Nombre de banda	Longitud de onda (μm)	Resolución (m) Resolución (m) Resolución (m)	Aplicaciones de banda
ETM +	1	Azul	0.45-0.52	30	Mapeo batimétrico, que distingue el suelo de la vegetación y la vegetación caducifolia de la conífera.
ETM +	2	Verde	0.52-0.60	30	Enfatiza la vegetación pico, que es útil para evaluar el vigor de la planta.
ETM +	3	rojo	0.63-0.69	30	Discrimina las pendientes de la vegetación.
ETM +	44	NIR	0.77-0.90	30	Destaca el contenido de biomasa y las costas.
ETM +	55	SWIR 1	1.55-1.75	30	Discrimina el contenido de humedad del suelo y la vegetación; penetra nubes delgadas.
ETM +	66	Térmico	10.40-12.50	60 * (30)	Mapeo térmico y humedad estimada del suelo.
ETM +	77	SWIR 2	2.09-2.35	30	Rocas alteradas hidrotermalmente asociadas con depósitos minerales.
ETM +	8	Pancromático	0.52-0.90	15	Resolución de 15 metros, definición de imagen más nítida.

Cuadro E.2: Bandas del sensor OLI, obtenida en la página oficial Landsat Science NASA.

Sensor	Número de banda	Nombre de banda	Longitud de onda (μm)	Resolución (m)	Aplicaciones de banda
OLI	1	Costero	0.43 - 0.45	30	Estudios costeros y de aerosoles.
OLI	2	Azul	0.45 - 0.51	30	Mapeo batimétrico, que distingue el suelo de la vegetación y la vegetación caducifolia de la conífera.
OLI	3	Verde	0.53 - 0.59	30	Enfatiza la vegetación pico, que es útil para evaluar el vigor de la planta.
OLI	4 4	rojo	0.63 - 0.67	30	Discrimina las pendientes de la vegetación.
OLI	5 5	NIR	0.85 - 0.88	30	Destaca el contenido de biomasa y las costas.
OLI	6 6	SWIR 1	1.57 - 1.65	30	Discrimina el contenido de humedad del suelo y la vegetación; penetra nubes delgadas.
OLI	7 7	SWIR 2	2.11 - 2.29	30	Mejora del contenido de humedad del suelo y la vegetación y penetración de nubes.
OLI	8	Pan	0.50 - 0.68	15	Resolución de 15 metros, definición de imagen más nítida
OLI	9 9	Cirro	1.36 - 1.38	30	Detección mejorada de la contaminación de las nubes cirrus.
Llantas	10	NEUMÁTICOS 1	10.60 - 11.19	30 (100)	Resolución de 100 metros, mapeo térmico y humedad estimada del suelo.
Llantas	11	TIRS 2	11.50 - 12.51	30 (100)	Resolución de 100 metros, mapeo térmico y humedad estimada del suelo.

Apéndice F

¿Cómo descargar una imagen de satélite?

Existen dos formas de descargar imágenes de satélite Landsat:

- **Antiguas imágenes de satélite**
  Se ingresa siguiente página http://glcfapp.glcf.umd.edu:8080/esdi/, apareciendo lo que se muestra en la Figura F.1.

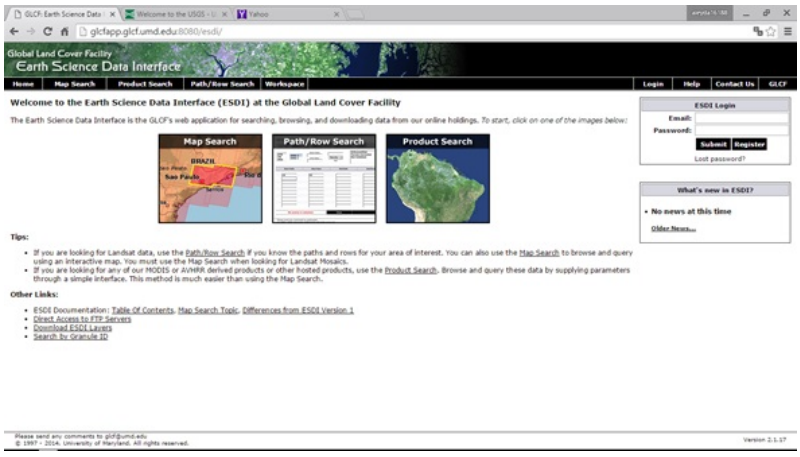


Figura F.1: Ventana de página web.

> Map Search.

Se selecciona en la parte izquierda el tipo de sensor que se requiere utilizar (ETM+,TM o OLI).
> Da clic en la pestaña Path/Row.
Llenar con los datos correspondientes sobre el Path y Row a la zona de interés, en nuestro caso los valores de partida de Path y Rrow corresponden a Start path: 25 y Start Row:46 >Update Map

>Preview and Download, para descargar las imágenes. >Selecciona en la columna ID la imagen que se analiza y en la parte de arriba se selecciona Download.

- **La segunda forma es:** Ingresando a la página : http://www.usgs.gov/ , >Donde primero se ingresa a una cuenta antes ya creada, con su respectivo usuario y contraseña. >Hacer clic en la pestaña **Maps, Imagery, and Publications**, como se muestra en la Figura F.2

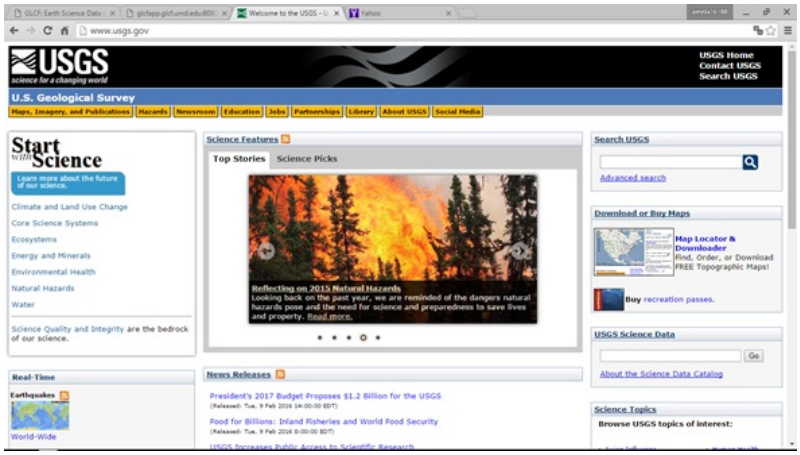


Figura F.2: Página web USGS, mostrando sección de Mapas y Publicaciones.

>

F.1. Definir zona por medio del Path y Row

¿Cómo obtener Path y Row de cualquier punto geográfico en el mapa?

- La siguiente página ayuda a conocer los valores de latitud y longitud de un punto cualquiera, solo haciendo clic sobre el mapa donde sea de interes conocer el Path y Row.

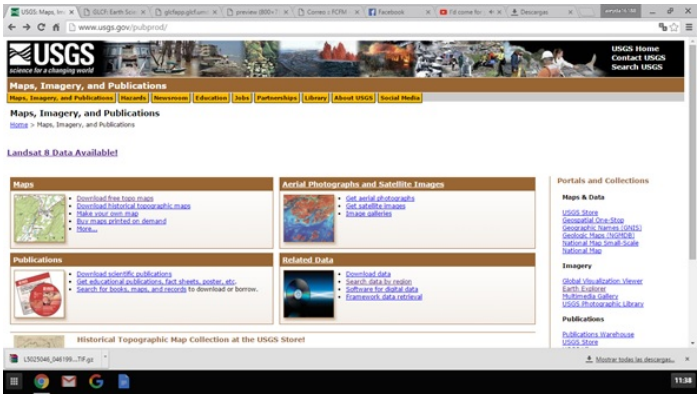


Figura F.3: Página Web USGS.

`http://earthexplorer.usgs.gov/`

Las coordenadas que arroja son de la forma; grados, minutos y segundos
Ejemplo:
$19^0 49' 24'' N$
$97^0 21' 58'' W$

- **Obtener Path y Row de una imagen, conociendo Latitud y Longitud.**

`https://landsat.usgs.gov/tools_latlong.php`

- **Transformar Latitud y Longitud de grados a decimales** Con el convertidor de INEGI, uno realiza una trasformación de convertir grados a decimales, diciendo detalladamente la longitud, latitud y ubicación. La página Web del convertidor es el siguiente:

`http://convertir-grados-minutos-y-segundos-a-decimales.todala.info/`

nota: elegir "Descending Node" ver ejemplo en la imagen.

F.2. Uso del software Arcgis(Arcmap), para el procesamiento de la imagen de satélite

Antes de usar ArcMap, se necesita configurar esta paquetería, con los pasos siguientes:

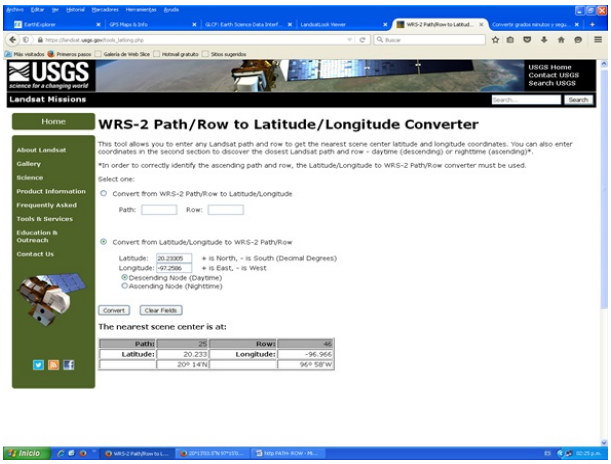


Figura F.4: path/row a latitud/longitud

> Clic derecho en la barra de herramientas.
> Selecciona “3D analyst”, “editor”,“Georeferencing”, “Spacial analyse”.
segundo paso
> “Customize”
> “Extensiones”
> “3D analyse”, “Arc Scan” y “Spacial Analyse”.

F.2.1. Ubicación de los puntos muéstrales en Arcmap

- Agrega un shaperfile en el el icono +
- En file
  > add data
  > add XY data y se carga el archivo en Excel donde este la información de los puntos muestrales, en Edit se debe especificar las propiedades de referencia espacial, las cuales son:
  - Projected Coordinate Systems.
  - UTM.
  - WGS 1984.
  - Northern Hemisphere.
  - WGS 1984 UTM Zone 14 N.

- NOTA: de igual manera se ubican los vértices del rectángulo que contenga los puntos muestrales, estos puntos se ubican tomando el máximo y mínimo de cada una de las coordenadas X y Y (información que se tiene de un estudio de campo).

F.2.2. Creación de Polígonos que contengan todos los puntos muestrales

Se requiere crear un shaperfile para guardar el polígono, que se puede realizar con los comandos siguientes:
> Catalog.
> Hacer clic derecho en la dirección donde se desee guardar este shaperfile que se acaba de generar.
>New.
>Shaperfile

- Escribe el nombre del shaperfile que se desea crear.
- Polygon.
- y en edit escribe el mismo Projected Coordinate Systems anterior. > Editor. > Star Editing y se selecciona el shaperile que se acaba de crear anteriormente.

  Después,

  >Editor. >Editing Windows. >Create Features.
  Se selecciona el archivo que aparece en la parte derecha, > polygon.
  Se crea un polígono cuyos vértices serán los cuatro puntos que se generaron a partir de la muestra, haciendo clic en cada uno de ellos y terminando el polígono con doble clic.
  >Editor.
  >Stop Editing.

F.2.3. Recorte de la imagen de satélite de una zona de estudio con un polígono

>"ArcToolbox".
>"Spacial Analysis Tools".
>"Extraction".
> "Extraction by Mask".
Llenar los campos que se muestran en la siguiente imagen

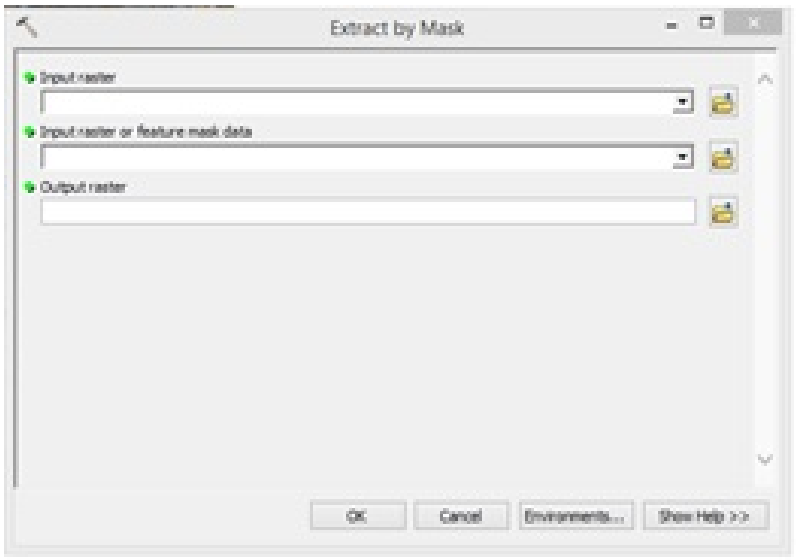


Figura F.5: Extración by mask

F.2.4. Procedimiento para realizar un recorte de una zona de estudio

> Arctoolbox.
>Analysis Tools.
>Extract.
>Clip
. >Llenar el formulario, indicando en

- **Input Features** El archivo en formato shaperfile.
- **Clip Features** El polígono que se fabrico en formato shaperfile.
- **Ouput Features Class** El archivo de salida, del recorte del Estado con polígono, indicando la ruta donde se guardará el archivo.

En la pantalla se muestra el polígono recortado.

F.2.5. Recorte de una Zona de estudio en formato tiff

Para exportar el recorte de la zona de estudio, se realiza los siguientes pasos:

- Export Map.
  > File.

>Export Map.
>selecciona formato jpg y 300 dpi.

- Export Data.
  > Hacer clic izquierdo en el shaperfile que se desee exportar en formato tiff.
  >Data.
  >Export Data.
  Finalmente se capturan los campos correspondientes a la ruta donde se guardó la imagen, así como el nombre de la imagen (ver ejemplo de captura de datos en la Figura F.6).

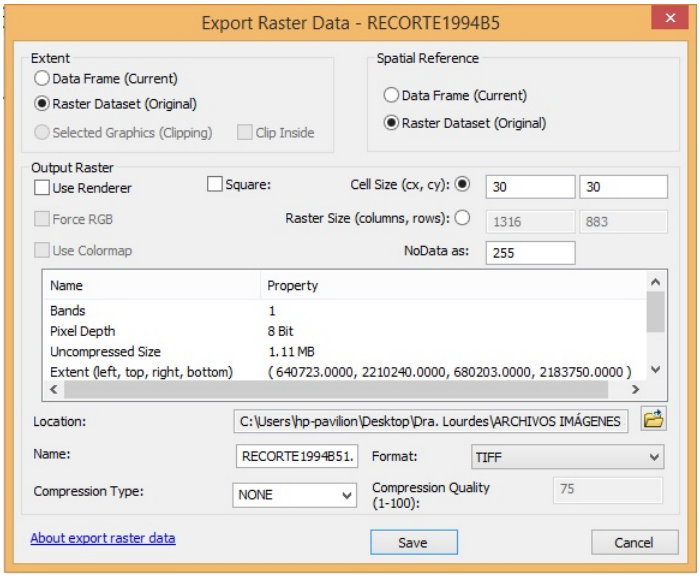


Figura F.6: Ejemplo de captura de datos

F.2.6. Procedimiento para la Obtención del Índice de Vegetación Normalizado a partir de la imagen satélite

Haciendo uso del software ArcGis y teniendo los recortes de las bandas correspondientes de la zona de estudio, se obtiene el Índice de vegetación NDVI con los siguientes pasos:

- Abrir el paquete ArcMap del Software ArcGis.

- Cargar las imágenes de las bandas 3(reflectividad en el rojo) y 4(Reflectividad en el Infrarrojo Cercano) de la imagen de satélite correspondiente al año 1994.
- Abrir "spatial analyst" para usar el "Raster Calculator".
- Ingresar la ecuación para calcular el NDVI, guardándolo con algún nombre. Por ejemplo, el nombre que se seleccionó en este trabajo es NDVI.
- $$X = (B4 - B3)/(B4 + B3)$$
- En "Identify from", seleccione "Visible layers" y haga click en distintas partes de la imagen NDVI (zonas de alta vegetación y zonas de valores bajos de NDVI) y se muestra cómo está la imagen se relaciona con la clasificación, como se muestra en la Figura **??**.

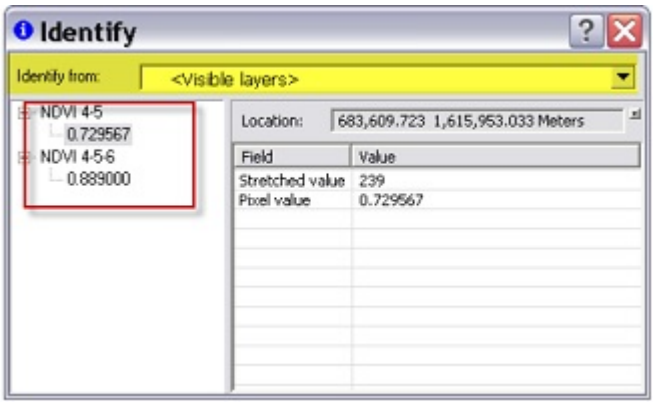


Figura F.7: Ventana identify Arcmap.

Bibliografía

[1] Agresti A., *Foundations of linear and generalized linear models*, John Wiley & Sons Inc, ISBN: 101118730038 , 2015.

[2] Arriaga L., Espinoza J. M., Aguilar C., Martínez E, Gómez L. and Loa E.(coordinadores), *Regiones terrestres prioritarias de México*, Comisión Nacional para el Conocimiento y uso de la Biodiversidad, México, 2000.

[3] Ato M., Losilla J. M., Navarro J. B., Palmer A. y Rodrigo M. F., *Modelo lineal generalizado*, Documenta Universitaria-EAP SL,ISBN: 978-84-96367-19-7, 2005.

[4] Bannari A., Morin D., Bonn F. and Huete A. R., *A review of vegetation indices*, Remote Sensing Reviews, 13, 95-120, 1995.

[5] Bates D., Mächler M., Bolker B. and Walker S., *lme4: Linear mixed-effects models using Eigen and S4*, R package version 1.1-15, 2017.

[6] Boettinger J. L., Ramsey R. D., Bodily J. M., Cole N. J., Kienast-Brown S., Nield S. J., ... ,and Stum A. K., *Landsat spectral data for digital soil mapping. In Digital soil mapping with limited data*, Springer, Dordrecht, ISBN: 978-1-4020-8592-5, 2008.

[7] Bolker B. M., Brooks M. E., Clark C. J., Geange S. W., Poulsen J. R., Stevens M. H. and White J. S., *Generalized linear mixed models: a practical guide for ecology and evolution*, Trends in ecology & evolution, 24(3), 127-135, 2009.

[8] Breslow N. E. and Clayton D. G., *Approximate inference in generalized linear mixed models*, J. Am. Stat. Assoc., 88(421), 9–25, 1993.

[9] Castillo M., *Medición de la variabilidad especial y temporal del secuestro de Carbono en suelos forestales de la Sierra Norte de Puebla*, Posgrado en Ciencias Ambientales (Tesis de Maestría), Instituto de Ciencias, BUAP, 2014.

[10] Cayuela L., *Modelos lineales mixtos (LMM) y modelos lineales generalizados mixtos (GLMM) en R*, Área de Biodiversidad y Conservación, Universidad Rey Juan Carlos, España, 2018.

[11] Chacón J. M., Leal M. T., Sánchez M. and Bandala E. R., *Tratamiento de agua residual proveniente de la industria textil mediante fotocatálisis solar*, In XXVII Congreso Interamericano de Ingeniería Sanitaria y Ambiental, Cancún, México, 2002.

[12] Chaparro D., Duveiller G. P., Cescatti M., Vall-Llossera A. M., Camps A. and Entekhabi D., *Sensitivity of L-band vegetation optical depth to carbon stocks in tropical forests: a comparison to higher frequencies and optical indices*, Remote Sensing of Environment, 232, 111-303. 2019.

[13] Chuvieco E., *Teledetección, SIG y cambio global*, Geographicalia, (29), 33-56, 1992.

[14] Cox D. R. and Hinkley D. V., *Theoretical statistics*, Chapman and Hall/CRC, ISBN: 0412124203, 1974.

[15] Crawley M. J., *The R book*, John Wiley & Sons., ISBN: 978-0-97392-9, 2012.

[16] Demidenko E., *Kolmogorov-Smirnov test for image comparison*, In International Conference on Computational Science and Its Applications, Springer, Berlin, Heidelberg, ISBN: 978-3-540-24768-5, 2004.

[17] Demidenko E., *Mixed models: theory and applications with R*, John Wiley & Sons., ISBN: 978-1-118-09157-9, 607-618, 2013.

[18] Gałecki A. and Burzykowski T., *Linear mixed-effects models using R: A step-by-step approach*, Springer Science & Business Media, ISBN: 978-1-4614-3899-1, 2013.

[19] Gilks W. R., Richardson, S. and Spiegelhalter D. J., *Introducing Markov chain Monte Carlo*, In Markov Chain Monte Carlo in Practice, Chapman and Hall, 1, 1–19, 1996.

[20] Gómez S., Torres V., Medina Y., Rodríguez Y., Sardiñas Y., Herrera M. and Rodríguez R., *Aplicación del Modelo Lineal Mixto y Lineal Generalizado Mixto, como alternativas de análisis en experimentos con medidas repetidas*. Cuban Journal of Agricultural Science, 53(1), 7-12, 2019.

[21] Houghton R. A., *Revised estimates of the annual net flux of carbon to the atmosphere from changes in land use and land management 1850–2000*, Tellus B, 55(2), 378-390, 2003.

[22] Jensen J. R., *Remote sensing of the environment: An earth resource perspective 2/e*. Pearson Education India, New Delhi, ISBN: 13: 978-8131716809, 2009.

[23] Jones, A. P., Panagos S., Barcelo F., Bouraoui C., Bosco O., Dewitte C., *The State of Soil in Europe*, Joint Research Centre Reference Report, Publications Office of the European Union, ISBN 978-92-79-22805-6, 2012.

[24] Keller K. L., *Building customer-based brand equity: A blueprint for creating strong brands*, Cambridgeend, MA: Marketing Science Institute, Report No 01-107, 3-27, 2001.

[25] Knudson C., *glmm: Generalized Linear Mixed Models via Monte Carlo Likelihood Approximation*, R package version 1.2.2, 2017.

[26] Linares F. G., Tenorio A. M. G., Trejo T. E. y Oroza H. A. A., *Estimación del Carbono Orgánico en suelos por teledetección y modelos de regresión*, Artículo Revista Latinoamericana el ambiente y las ciencias. 8(18)-3, 2017.

[27] Linares F. G., Oroza H. A. A. y Paz F. and Torres R., *Modelación de la dinámica del secuestro de Carbono en suelos forestales, Estado Actual del Conocimiento del Ciclo del Carbono y sus Interacciones en México: Síntesis a 2016*, Serie Síntesis Nacionales, Programa Mexicano del Carbono en colaboración con la Universidad Autónoma del Estado de Hidalgo, Texcoco Estado de México., ISBN: 978-607-96490-4-3, 2016.

[28] Monterubbianesi M. G., *Evaluación de alternativas para el análisis estadístico y de aspectos del diseño en ensayos de larga duración para estudios agronómicos (Tesis Doctoral)*, Universitat de Lleida, 2017.

[29] McCullagh P. and Nelder J., *Generalized linear models*, Chapman & Hall CRC,ISBN 0-412-31760-5, 1989.

[30] McCulloch E. C., Searle R. S. and Neuhaus M. J., *Generalized, Linear and Mixed Models*, Jhon Wiley & Sons, ISBN: 1118209966, 2011.

[31] Mehtätalo L., *Linear mixed-effects models with examples in R*, University of Eastern Finland, Addison-Wesley, 2013.

[32] Meng Q., Cieszewski C. J., Madden M., Borders B., *A linear mixed-effects model of biomass and volume of trees using Landsat ETM+ images*, Forest Ecology and Management, 244(1-3), 93-101, 2007.

[33] Montgomery D. C., Peck E. A. and Vining G. G., *Introduction to Linear Regression Analysis*, Jhon Wiley & Sons, ISBN: 978-0-470-54281-1, 2012.

[34] Nakagawa S., Johnson P. C. and Schielzeth H., *The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded*, Journal of the Royal Society Interface, 14(134), 1-11, 2017.

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

[36] Oroza H. A. A., *Modelos mixtos en la determinación del Carbono orgánico en la hojarasca en una zona de Teziutlán, Puebla (Tesis de Maestría)*, BUAP, 2015.

[37] Oroza H. A. A., Linares G. F., Reyes C. H. J., *Modelación estadística con imágenes satelitales en Ciencias Ambientales*, Aplicaciones en Estadística y la Probabilidad, ISBN: 978-607-525-589-7, 2019.

[38] Oroza H. A. A., Linares G. F., Reyes C. H. J. and Juárez B., *Estimación del porcentaje de Carbono Orgánico en suelos utilizando Imágenes Satelitales y Modelos Mixtos*, Ciencias Matemáticas, 2(33), 11-17, 2019.

[39] Pacheco R. A., *Cambio espacio temporal de la cobertura edáfica en el sureste de la Presa Valsequillo, Puebla. Causas y efectos*, Posgrado en Ciencias Ambientales, Instituto de Ciencias, BUAP, 2018.

[40] Peña D., *Análisis de datos multivariantes*, McGraw-Hill, ISBN: 9788448136109, 2002.

[41] Pinheiro J. and Bates D., *Mixed-effects models in S and S-PLUS*, Springer Science & Business Media,91, 2006.

[42] Pinheiro J., Bates D., DebRoy S., Sarkar D. and R Core Team, *nlme: Linear and Nonlinear Mixed Effects Models*, R package version 3.1-142, https://CRAN.R-project.org/package=nlme, 2019.

[43] Pinheiro J. C. and Chao E. C.*Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models*, J. Comput. Graph. Statist., 15(1), 58–81, 2006.

[44] Quantum GIS Development Team, *Quantum GIS Geographic Information System*, Open Source Geospatial Foundation Project, 2017.

[45] Rao C. R. and Toutenburg H., *Linear Models and Generalizations: Least Squares and Alternatives*, Springer, 2007.

[46] Raudenbush S. W., Yang M. L. and Yosef M., *Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation*, J. Comput. Graph. Statist., 9(1), 141–157, 2000.

[47] Rivera A. E., *Evaluación toxicológica del agua residual textil (proceso Denim) vertida al Río Atoyac (Tesis de Maestría)*, BUAP, 2018.

[48] Román J. V., *Talleres de mezclilla y transformaciones sociambientales en un municipio rural*, El caso de Tepetitla de Lardizábal, Tlaxcala, México, Sociedad y Ambiente, 11, 68-91, 2016.

[49] Rossi, R. J., *Mathematical Statistics: An Introduction to Likelihood Based Inference*, John Wiley & Sons, ISBN: 9781118771075, 2018.

[50] Salinas R. A., Pérez N. R. and Ávila B. L., *Modelos de regresión para variables expresadas como una proporción continua*, salud pública de méxico, 48, 395-404, 2006.

[51] Searle S. R. and Khuri A. I., *Matrix algebra useful for statistics*, John Wiley & Sons,ISBN: 978-1-118-93515-6, 2017.

[52] Tao H., Li M., Wang M. and Lü G., *Genetic algorithm-based method for forest type classification using multi-temporal NDVI from Landsat TM imagery*, Annals of GIS, 25(1), 33-43, 2019.

[53] Team, R. C., *nlme: linear and nonlinear mixed effects models*, R package version 3.1–137, 2018.

[54] Team, R. C., *R: A language and environment for statistical computing*, 2013.

[55] Torres E., Linares G., Tenorio M. G., Peña R., Castelán R. and Rodríguez A., *Obtención de índices de vegetación para estimar densidades de vegetación y cambios de Uso de Suelo en la Región Terrestre Prioritaria 105: Cuetzalan, Puebla, México*, Revista Iberoamericana de Ciencias, ISBN: 2334-2501, 2014.

[56] Van der Knijff J M., Jones R. J. A. and Montanarella L., *Soil erosion risk: assessment in Europe*, European Soil Erosion Risk Assessment 2000.

[57] Venables W. N. and Ripley B. D., *Modern applied statistics with S-PLUS*, Springer Science & Business Media, ISBN: 978-1-4757-3123-1, 2013.

[58] West B. T., Welch K. B. and Galecki A. T., *Linear mixed models: a practical guide using statistical software*, Chapman and Hall/CRC, ISBN 13: 978-1-4665-6102-1, 2014.

[59] Wischmeier W. H. and Smith D. D., *Predicting rainfall erosion losses R*, USDA Agricultural Handbook 537, 1978.

[60] Yang C., Huang H., Ni J., Yang D., *Effects of Topographic Normalization on the Relationship Between Tropical Forest Biomass and Landsat TM Images*, Journal of the Indian Society of Remote Sensing, 47(4), 595-601, 2019.

[61] Zheng D., Rademacher J., Chen J., Crow T., Bresee M., Le Moine J. and Ryu S. R., *Estimating aboveground biomass using Landsat 7 ETM+ data across a managed landscape in northern Wisconsin, USA*, Remote sensing of environment, 93(3), 402-411, 2004.

[62] Zuur A., Ieno E. N., Walker N., Saveliev A. A. and Smith G. M., *Mixed effects models and extensions in ecology with R*, Springer Science & Business Media, ISBN: 1431-8776, 2009.