BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Análisis de Regresión para la Estimación del Secuestro de Carbono Orgánico en Suelos.

TESIS
que para obtener el título de:

Lic. en Matemáticas Aplicadas.

---

presenta:

Gabriela López Pineda

Directoras de Tesis:

Dra. Gladys Linares Fleites
Dra. Hortensia Josefina Reyes Cervantes

PUEBLA, PUE. SEPTIEMBRE 2013.

Análisis de Regresión para la Estimación del Secuestro de Carbono Orgánico en Suelos.

Gabriela López Pineda

Septiembre 2013

Índice general

Introducción

Como resultado del aumento en concentraciones de gases de efecto invernadero, existen evidencias científicas que sugieren que el clima global se verá alterado en este siglo. El principal responsable del cambio climático global es el CO_2, que tiene entre sus fuentes emisoras la deforestación y la destrucción de los suelos [7].

El carbono orgánico del suelo (COS) es un gran y activo reservorio, y los ecosistemas forestales pueden absorber cantidades significativas de CO_2, por lo cual existe un gran interés en incrementar el contenido de carbono orgánico en estos ecosistemas.

La relevancia que tiene en la actualidad la modelación del Carbono Orgánico del Suelo, se encuentra íntimamente ligado a la propiedad de incidir directamente en el almacenamiento del carbono por medio del suelo y, por ende, en la mitigación de las emisiones de CO_2, que es uno de los Gases de Efecto Invernadero (GEI) de mayor importancia.

En las últimas décadas se ha reconocido que el cambio del uso de suelo constituye uno de los factores plenamente implicados en el cambio climático global, alterando procesos y ciclos. Lo anterior se vuelve trascendental si se considera que es a través de estos cambios donde se materializa la relación entre el hombre y el medio ambiente.

Los ecosistemas terrestres han sufrido grandes transformaciones, la mayoría debido a la conversión de la cobertura del terreno y a la degradación e intensificación del uso del suelo. Estos procesos usualmente englobados en lo que se conoce como deforestación o degradación forestal, se asocian a impactos ecológicos importantes en casi todas las escalas [9].

Se ha comprobado que la destrucción de la biodiversidad y los bosques tropicales y templados puede perturbar el clima global y poner en riesgo una fuente importante de captura de carbono. Cada vez es más evidente la transformación que sufre el territorio. Los cambios de uso del suelo ya sean legales o ilegales son cada día más frecuentes.

Algunos autores como en [2] consideran que la modificación de la cobertura y uso del suelo se debe a la interacción de factores económicos, políticos y ecológicos. Sin embargo, otros piensan que la escasa conexión entre los estudios explícitos del uso del suelo y los aspectos socioeconómicos provoca serias dificultades para integrar realmente aspectos biofísicos y humanos. También se acusa

la falta de trabajos de análisis cuantitativos que permitan explicar las causas y efectos de estos factores, ya que también las interpretaciones de cómo estos interactúan para estimular el cambio varían ampliamente de una región a otra.

Los estudios del caso en México muestran que los métodos utilizados, en general son diferentes e indefinidos en cuanto a los parámetros y variables que se incluyen; incomparables en términos de las categorías que utilizan y con escalas de trabajo incompatibles. De igual forma sólo algunos trabajos revisados hacen uso de técnicas estadísticas para tratar de conocer las causas de los cambios. De acuerdo con lo anterior, se puede decir que aún son escasas las investigaciones en el país que tratan de explicar las causas de estos cambios utilizando variables socioeconómicas y ambientales [2].

En la actualidad existe un gran interés científico por establecer la relación que guarda el Carbono Orgánico del Suelo con otras propiedades del mismo. En los problemas de predicción del COS es bastante frecuente que las variables independientes del modelo sean altamente colineales, y por tanto, los estimadores de la regresión por mínimos cuadrados ordinarios (OLS, por sus siglas en inglés) no son adecuados ya que poseen varianzas muy grandes. Se han propuesto diferentes técnicas para manejar los problemas causados por la multicolinealidad.

En este trabajo se muestran los resultados obtenidos al desarrollar un modelo de regresión y analizar sus supuestos, así mismo se revisan dos metodologías relativamente similares y usadas en la solución a estos problemas: *Regresión con Componentes Principales* (RCP) y *Regresión por Mínimos Cuadrados Parciales* (PLS, por sus siglas en inglés). Ambos métodos transforman las variables predictoras en componentes ortogonales, las cuales representan la solución al problema de multicolinealidad y permiten hacer una reducción de la dimensionalidad del espacio de variables predictoras.

El propósito del presente estudio es comparar estos diferentes métodos de regresión para seleccionar el mejor modelo que prediga el porcentaje de COS en función de otras propiedades del suelo, en la zona de la Caldera de Teziutlán en el estado de Puebla.

En el Capítulo 1 se describe esencialmente el Modelo de Regresión Lineal Múltiple, así como la comprobación de los supuestos del mismo. La parte fuerte de este trabajo se encuentra en el Capítulo 2, donde se explica el problema de *Multicolinealidad* al igual que sus diagnósticos, en esté mismo Capítulo se brindan las características esenciales de las metodologías: RCP y PLS y se explican los criterios utilizados para medir la capacidad predictiva de los modelos.

Posteriormente en el Capítulo 3 se aplican los conceptos al problema antes mencionado con las metodologías descritas y se comparan a través de diferentes estadísticos que muestran la capacidad predictiva de los mismos. Finalmente, se dan las conclusiones.

Capítulo 1

Regresión Lineal Múltiple

El análisis de regresión es la técnica estadística de uso más frecuente para investigar y **modelar la relación entre variables**. Su atractivo y utilidad generalmente son el resultado del proceso conceptualmente lógico de usar una ecuación para expresar la relación entre una variable de interés (la respuesta) y un conjunto de variables predictoras relacionadas [10].

1.1. Modelo de Regresión Lineal Múltiple

Sean k variables predictoras; $x_1, x_2, \ldots, x_k$ que se cree están relacionadas con una variable de respuesta Y.

El modelo clásico de regresión lineal indica que la variable Y se compone de una media, que depende de manera continua de las $x_i's$, y un error aleatorio ε, que representa el error de medición y los efectos de otras variables que no fueron consideradas explicítamente en el modelo.

Los valores de las variables predictoras son tratados como *fijos*, mientras que el error y la respuesta son vistos como variables aleatorias cuyo comportamiento se caracteriza por un conjunto de hipótesis de distribución [15].

Especificamente, el modelo de regresión lineal con una sola respuesta es de la siguiente forma:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon. \tag{1.1}$$

El término lineal se refiere al hecho de que la media es considerada como una función lineal de los parámetros desconocidos $\beta_0, \beta_1, \ldots, \beta_k$, donde dichos parámetros β_j, $j = 0, 1, \ldots, k$ son llamados *coeficientes de regresión*. El parámetro β_j representa el cambio esperado en la respuesta Y por cambio unitario en x_j *cuando todas las demás variables regresoras* $x_i, (i \neq j)$ *se mantienen constantes.*

Para estimar los $\beta's$ en (1.1), necesitamos una muestra de n observaciones de Y y de las $x's$ asociadas. El modelo para la observación i-ésima está dado por:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i. \tag{1.2}$$

Las suposiciones para ε_i son las siguientes:

1. $E(\varepsilon_i) = 0$, para $i = 1, \ldots, n$, o equivalentemente; $E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$.
2. $Var(\varepsilon_i) = \sigma^2$, para $i = 1, \ldots, n$, o equivalentemente; $Var(y_i) = \sigma^2$.
3. $Cov(\varepsilon_i, \varepsilon_j) = 0$, para $i \neq j$, o equivalentemente; $Cov(y_i, y_j) = 0$.

La suposición 1 establece que el modelo es correcto; es decir, toda la información relevante de las $x's$ está incluida y el modelo es lineal. La suposición 2 afirma que la varianza de Y es constante y por lo tanto, no depende de las $x's$. Por otra, parte la suposición 3 establece que las $y's$ no están correlacionadas entre sí y que por lo general son tomadas de una muestra aleatoria. Más adelante vamos a añadir el supuesto de normalidad en el cual las $y's$ serán independientes y no correlacionadas [12].

Con n observaciones independientes en Y y los valores asociados de x_i, el modelo completo se convierte en:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \ldots + \beta_k x_{1k} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \ldots + \beta_k x_{2k} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \ldots + \beta_k x_{nk} + \varepsilon_n. \end{aligned} \tag{1.3}$$

Es más cómodo manejar modelos de regresión múltiple cuando se expresan en notación matricial. Eso permite presentar en forma muy compacta al modelo, los datos y los resultados. La notación matricial del modelo es la siguiente,

$$Y = X\beta + \varepsilon, \tag{1.4}$$

en donde,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

En general, $\mathbf{Y}$ es una matriz de $n \times 1$ de las observaciones, $\mathbf{X}$ es una matriz de $n \times (k+1)$ de variables regresoras, β es un vector de $(k+1) \times 1$ de los coeficientes de regresión desconocidos y ε es una matriz de $n \times 1$ de errores aleatorios.

Podemos observar que las tres suposiciones mencionadas antes, se pueden expresar mediante el modelo dado por (1.4) como:

- $E(\varepsilon) = 0$, o bien; $E(Y) = X\beta$.
- $Cov(\varepsilon) = \sigma^2 I$, o bien; $Cov(Y) = \sigma^2 I$.

Observemos que la suposición $Cov(\varepsilon) = \sigma^2 I$ incluye los supuestos $Var(\varepsilon_i) = \sigma^2$ y $Cov(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$.

La matriz $\mathbf{X}$ en (1.4) es de tamaño $n \times (k+1)$; supondremos que $n > k+1$ y $Rango(X) = k+1$. Si $n < k+1$ o si existe una relación lineal entre las $x's$, entonces X no tendrá rango completo. Si los valores de los x_{ij} se han fijado (escogido por el investigador), entonces la matriz X contiene esencialmente el diseño experimental, por lo cual es llamada *matriz de diseño*.

Modelo de Regresión Lineal

$$Y_{n\times 1} = X_{n\times(k+1)}\beta_{(k+1)\times 1} + \varepsilon_{n\times 1}$$

$$E(\varepsilon) = 0_{n\times 1} \qquad y \qquad Cov(\varepsilon) = \sigma^2 I_{n\times n} \tag{1.5}$$

donde β y σ^2 son los parámetros desconocidos.

1.2. Estimación de β y σ^2

Para obtener estimadores para β, no es necesario que la ecuación de predicción $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}$ este basada en $E(y_i)$. Sólo es necesario postular un modelo empírico que sea lineal en $\hat{\beta}'s$, y el método de mínimos cuadrados se encarga de encontrar el *mejor* ajuste a este modelo.

1.2.1. Estimador de Mínimos Cuadrados para β

En esta sección, se discute la aproximación de mínimos cuadrados para la estimación de los $\beta's$ en el modelo con $X - fijas$ en (1.1) o (1.4), para ello no son necesarias suposiciones sobre alguna distribución para la respuesta Y. Para los parámetros $\beta_0, \beta_1, \ldots, \beta_k$, buscamos estimadores $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ que minimizen:

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik} \right)^2 . \end{aligned} \tag{1.6}$$

Se debe minimizar la función $S(\beta)$ respecto a $\beta_0, \beta_1, \ldots, \beta_k$. Los estimadores de $\beta_0, \beta_1, \ldots, \beta_k$ por mínimos cuadrados deben satisfacer

$$\frac{\partial S}{\partial \beta_0}\Big|_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k} = -2\sum_{i=1}^{n}\Big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik})\Big) = 0, \tag{1.7a}$$

$$\frac{\partial S}{\partial \beta_j}\Big|_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k} = -2\sum_{i=1}^{n}\Big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik})\Big) x_{ij} = 0. \tag{1.7b}$$

Al simplificar las ecuaciones (1.7a) y (1.7b) se obtienen las **ecuaciones normales de mínimos cuadrados**

$$\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i, \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} &= \sum_{i=1}^{n} x_{i1} y_i, \\
&\vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik} y_i.
\end{aligned} \tag{1.8}$$

Notemos que hay $k+1$ ecuaciones normales, una para cada parámetro desconocido de la regresión. La solución de estas ecuaciones serán los estimadores por mínimos cuadrados $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. Es más cómodo manejar un modelo de regresión en forma matricial, de modo que $S(\beta)$ se puede expresar de la siguiente manera,

$$\begin{aligned}
S(\beta) =& Y'Y - \beta' X'Y - Y'X\beta + \beta' X'X\beta \\
=& Y'Y - 2\beta' X'Y + \beta' X'X\beta.
\end{aligned}$$

Ahora, diferenciando con respecto a β e igualando a cero, se obtiene:

$$\frac{\partial S(\beta)}{\partial \beta}\Big|_{\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0, \tag{1.9}$$

de aqui tenemos,

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

La expresión (1.10) son las **ecuaciones normales de mínimos cuadrados**, y son la forma matricial de las ecuaciones dadas por (1.8). Para resolver el sistema de ecuaciones normales multiplicamos ambos lados de (1.10) por la inversa de $X'X$, de esta forma el estimador por mínimos cuadrados para β es,

$$\hat{\beta} = (X'X)^{-1}X'Y, \tag{1.11}$$

siempre y cuando exista la inversa $(X'X)^{-1}$. La matriz $(X'X)^{-1}$ existe sólo si las variables regresoras son **linealmente independientes**.

La matriz de valores ajustados $\hat{Y}$ que corresponde a la matriz de valores observados Y es

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY. \tag{1.12}$$

La matriz $H = X(X'X)^{-1}X'$ es llamada *matriz sombrero* y cumple con ser idempotente[1]. H es útil para las manipulaciones teóricas, pero por lo general no se calcula de forma explícita, ya que es una matriz de $n \times n$ que podría ser muy grande para algunos conjuntos de datos.

La diferencia entre el valor observado y_i y el valor ajustado $\hat{y}_i$ correspondiente es el **residual** $\varepsilon_i = y_i - \hat{y}_i$. Los n residuales pueden escribirse de forma matricial como sigue:

$$\varepsilon = Y - \hat{Y}. \tag{1.13}$$

El vector residual ε estimado en el modelo (1.4) se puede utilizar para comprobar la validez del modelo y los supuestos correspondientes.

1.2.2. Propiedades de los Estimadores de Mínimos Cuadrados

Los estimadores de mínimos cuadrados $\hat{\beta}$ tienen ciertas propiedades, las cuales se muestran con facilidad.

Si $E(Y) = X\beta$, entonces

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

por que $E(\varepsilon) = 0$ y $(X'X)^{-1}X'X = I$, entonces $\hat{\beta}$ es un **estimador insesgado** de β.

La propiedad de varianza de $\hat{\beta}$ se expresa con la **matriz de covarianza**

$$\begin{aligned} Cov(\hat{\beta}) &= E\{[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]'\} \\ &= Cov((X'X)^{-1}X'Y), \end{aligned}$$

[1]Se dice que una matriz $A_{k\times k}$ es idempotente si $A = AA$

que es una matriz simétrica, cuyo j-ésimo elemento de la diagonal es la varianza de $\hat{\beta}_j$ y cuyos elementos fuera de la diagonal son $Cov(\hat{\beta}_i, \hat{\beta}_j)$.

Recordemos que $Cov(Y) = \sigma^2 I$ y $\hat{\beta} = (X'X)^{-1}X'Y$, entonces

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

pues $Cov((X'X)^{-1}X'Y) = (X'X)^{-1}X'Cov(Y)[(X'X)^{-1}X']'$ y $[(X'X)^{-1}X']' = X(X'X)^{-1}$, así la matriz de covarianzas para $\hat{\beta}$ está dada por $\sigma^2(X'X)^{-1}$.

Además de $E(\hat{\beta}) = \beta$ y $Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$, una tercera propiedad importante de $\hat{\beta}$ es que bajo los supuestos estándares, la varianza de cada $\hat{\beta}_j$ es mínima:

Teorema 1.1 (Teorema de Gauss-Markov) *Si $E(Y) = X\beta$ $Cov(Y) = \sigma^2 I$ se tiene que el estimador de mínimos cuadrados $\hat{\beta}$, tiene varianza mínima entre todos los estimadores insesgados para β.*

El teorema de Gauss-Markov en [12] establece que el estimador de mínimos cuadrados para β, es **BLUE**(*Best Linear Unbiased Estimator*), el mejor estimador lineal insesgado. En este caso, se puede decir que $\hat{\beta}$ tiene la varianza mínima, entre la clase de todos los estimadores insesgados que son combinaciones lineales de los datos. Si además se supone que los errores ε_i tienen distribución normal, como se verá más adelante, entonces el estimador de mínimos cuadrados es el mismo que el estimador de maxima verosimilitud y también es un estimador insesgado y de varianza mínima.

1.2.3. Un Estimador para σ

Podemos determinar un estimador para σ^2 a partir de la suma de cuadrados de los residuos,

$$\begin{aligned} SS_{Res} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \varepsilon'\varepsilon. \end{aligned}$$

Si sustituimos $\varepsilon = Y - X\hat{\beta}$ se obtiene,

$$\begin{aligned} SS_{Res} =& (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ =& Y'Y - \beta' X'Y - Y'X\beta + \beta' X'X\beta \\ =& Y'Y - 2\beta' X'Y + \beta' X'X\beta \end{aligned} \tag{1.14}$$

Como $X'X\hat{\beta} = X'Y$, la ultima igualdad se transforma en

$$SS_{Res} = Y'Y - \hat{\beta}' X'Y.$$

Ahora como se deben estimar $k+1$ parámetros, tendremos que la suma de cuadrados de los residuales tienen $n-(k+1)$ grados de libertad asociados al modelo. Así, el cuadrado medio residual esta dado por

$$MS_{Res} = \frac{SS_{Res}}{n-(k+1)}. \tag{1.15}$$

En [12] se muestra que el valor esperado para MS_{Res} es σ^2 y por tanto, un estimador insesgado para σ^2 es

$$\hat{\sigma}^2 = MS_{Res}. \tag{1.16}$$

1.2.4. Estimadores de Máxima Verosimilitud

El método de mínimos cuadrados se aplica para estimar los parámetros en un modelo de regresión lineal, independientemente de la distribución que tengan los errores ε. Con el método de mínimos cuadrados se obtienen los mejores estimadores lineales insesgados $\hat{\beta}$. Otros procedimientos, como prueba de hipótesis, suponen que los errores se distribuyen normalmente, de modo que un método alternativo para estimar parámetros es el **método de máxima verosimilitud**.

El modelo esta dado por $Y = X\beta + \varepsilon$ y los errores estan distribuidos en forma normal e independiente con varianza constante σ^2, de modo que la función de densidad normal para los errores es,

$$f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2}} e^{-\frac{1}{2\sigma^2}\varepsilon_i^2}.$$

La función de máxima verosimilitud está dada por:

$$L(\varepsilon, \beta, \sigma^2) = \prod_{i=1}^{n} f(\varepsilon_i) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\varepsilon_i^2}, \tag{1.17}$$

ahora sustituyendo $\varepsilon = Y - X\beta$ en (1.17) y dado que la transformación logarítmica es monotóna, es apropiado maximizar 1.18 en lugar de 1.17, como el argumento maximización permanece tomando el logaritmo tenemos:

$$\ln L(Y, X, \beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta). \tag{1.18}$$

Si no hay restricciones sobre los parámetros, podemos derivar la función dada por 1.18 respecto a los parametros β y σ, e igualar a cero, de modo que los estimadores de máxima verosimilitud deben satisfacer:

$$\frac{\partial L}{\partial \beta}\Big|_{Y,X,\beta,\sigma^2} = -\frac{1}{2\sigma^2}\left[-2X'Y + 2X'X\hat{\beta}\right] = \frac{X'X\beta - X'Y}{\sigma^2} = 0, \tag{1.19a}$$

$$\frac{\partial L}{\partial \sigma^2}\Big|_{Y,X,\beta,\sigma^2} = -\frac{n}{2}\frac{2\pi}{2\pi\sigma^2} - \frac{1}{2\sigma^4}(Y - X\beta)'(Y - X\beta) = 0. \tag{1.19b}$$

De (1.19a) se tiene

$$X'X\beta = X'Y \implies \hat{\beta} = (X'X)^{-1}X'Y.$$

Y de (1.19b) se tiene

$$-n + \frac{1}{\sigma^2}(Y - X\beta)'(Y - X\beta) = 0 \implies \hat{\sigma}^2 = \frac{(Y - X\beta)'(Y - X\beta)}{n}.$$

Notemos que el estimador $\hat{\beta}$ obtenido por el método de máxima verosimilitud es el mismo que el obtenido por mínimos cuadrados, en cambio $\hat{\sigma}^2$ es un estimador sesgado. Tambien es conveniente señalar que los estimadores de máxima verosimilitud tienen mejores **propiedades estadísticas** que los obtenidos con el método de mínimos cuadrados.

1.3. Pruebas de Bondad del Ajuste

Una vez estimados los parámetros del modelo, surgen de inmediato dos preguntas:

1. ¿Qué tan bien se ajusta el modelo a los datos?
2. ¿Qué regresores específicos son importantes?

Hay varios procedimientos de pruebas de hipótesis que demuestran su utilidad para contestar esas preguntas.

1.3.1. R^2 y R^2_{Adj}

Existen otras maneras de evaluar el ajuste general del modelo, a saber los estadísticos R^2 y R^2_{Adj}.

La cantidad

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}, \tag{1.20}$$

se llama **coeficiente de determinación**. Nótese que R^2 indica la proporción de variabilidad explicada por la regresión. Ya que $0 \leq SS_{Res} \leq SS_T$, entonces $0 \leq R^2 \leq 1$. Los valores de R^2 cercanos a 1 implican que la mayor parte de la

variabilidad de Y esta explicada por el modelo de regresión.

En general, R^2 aumenta siempre cuando se agrega un regresor al modelo, independientemente del valor de la contribución de esa variable. En consecuencia, es difícil juzgar si un aumento de R^2 dice en realidad algo importante.

Algunos investigadores que trabajan con modelos de regresión prefieren usar el estadístico $\boldsymbol{R^2_{Adj}}$, que se define como sigue:

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-(k+1))}{SS_T/(n-1)}. \tag{1.21}$$

En vista de que $SS_{Res}/(n-k-1)$ es el cuadrado medio de residuales, $SS_T/(n-1)$ 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 ajustada penaliza el aumento de términos que no son útiles, además es un procedimiento para evaluar y comparar los modelos posibles de regresión.

1.3.2. Pruebas de Significancia de la Regresión

La prueba de **significancia de la regresión** nos sirve para determinar si existe una relación lineal entre la respuesta Y y cualquiera de las variables regresoras $x_1, x_2, \cdots, x_k$. Este procedimiento suele considerarse como una prueba general o global del ajuste del modelo. Las hipótesis pertinentes son:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad Vs \quad H_1 : \beta_j \neq 0, \text{ al menos para alguna } j.$$

Dado un nivel de significancia α, el rechazo de la hipótesis nula implica que al menos uno de los regresores $x_1, x_2, \cdots, x_k$ contribuye al modelo en forma significativa.

La **suma total de cuadrados**, SS_T se puede dividir en una **suma de cuadrados debido a la regresión**, SS_R y en una **suma de cuadrados de residuales**, SS_{Res}, entonces tenemos:

$$SS_T = SS_R + SS_{Res}.$$

Por otra parte, tenemos que si la hipótesis nula es cierta, entoces SS_R/σ^2 tiene una distribución χ^2_k, con la misma cantidad de grados de libertad que la cantidad de variables regresoras en el modelo. También tenemos que SS_{Res}/σ^2 tiene una distribución $\chi^2_{(n-k-1)}$, y además como los estadísticos SS_R y SS_{Res} son independientes, por consiguiente:

$$\frac{SS_R}{k\sigma^2} = \frac{MS_R}{\sigma^2} \qquad y \qquad \frac{SS_{Res}}{(n-k-1)\sigma^2} = \frac{MS_{Res}}{\sigma^2},$$

son variables aleatorias χ^2, cada una dividida entre sus grados de libertad respectivos, de modo que de acuerdo con la definición de un estadístico F se tiene que

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}} \sim F_{k,n-k-1}. \tag{1.22}$$

También se cumple que,

$$E(MS_{Res}) = \sigma^2.$$
$$E(MS_R) = \sigma^2 + \frac{\beta^{*\prime} X_c' X_c \beta^*}{k\sigma^2}.$$

Donde $\beta^* = (\beta_1, \beta_2, \cdots, \beta_k)$ y X_c es la matriz *centrada* del modelo definida por,

$$X_c = \begin{pmatrix} x_{11} - \overline{x_1} & x_{12} - \overline{x_2} & \cdots & x_{1k} - \overline{x_k} \\ x_{21} - \overline{x_1} & x_{22} - \overline{x_2} & \cdots & x_{2k} - \overline{x_k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i1} - \overline{x_1} & x_{i2} - \overline{x_2} & \cdots & x_{ik} - \overline{x_k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \overline{x_1} & x_{n2} - \overline{x_2} & \cdots & x_{nk} - \overline{x_k} \end{pmatrix}.$$

Estos cuadrados medios esperados indican que si el valor observado de F_0 es grande, lo más probable es que al menos una $\beta_j \neq 0$. Y Si al menos una $\beta_j \neq 0$ $j = 1, 2, \cdots, k$, entonces F_0 tiene una distribución $F_{k,n-k-1}$ no central[2], y con parámetro de no centralidad definido por

$$\lambda = \frac{\beta^{*\prime} X_c' X_c \beta^*}{k\sigma^2}.$$

Este parámetro de no centralidad nos indica que el valor observado de F_0 debe ser grande para que al menos una $\beta_j \neq 0$. Por consiguiente, para probar la hipótesis $H_0 : \beta_1 = \beta_2 = \beta_k = 0$, se calcula el estadístico de prueba F_0 y se rechaza H_0 si

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

El procedimiento de prueba se resume normalmente en una **tabla de análisis de varianza** como en el Cuadro 1.1.

Para calcular SS_R se obtiene partiendo de

$$SS_{Res} = Y'Y - \hat{\beta}' X'Y,$$

[2]Se describen las Distribuciones No Centrales en el Apéndice C.

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	$\mathbf{F}_0$
Regresión	SS_R	k	MS_R	MS_R/MS_{Res}
Residuales	SS_{Res}	$n-k-1$	MS_{Res}	
Total	SS_T	$n-1$		

Cuadro 1.1: Análisis de varianza para determinar el significado de la regresión.

ya que

$$SS_T = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = Y'Y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}.$$

Se puede escribir la ecuación anterior como sigue

$$SS_{Res} = Y'Y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} - \left[\hat{\beta}'X'Y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right].$$

O bien,

$$SS_{Res} = SS_R - SS_T.$$

Por tanto, la **suma de cuadrados de la regresión** está dada por

$$SS_R = \hat{\beta}'X'Y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n},$$

la **suma residual de cuadrados** es

$$SS_{Res} = Y'Y - \hat{\beta}'X'Y,$$

y la **suma total de cuadrados** es

$$SS_T = Y'Y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}.$$

1.3.3. Pruebas para Coeficientes Individuales

Una vez determinado que al menos una de las variables regresoras es importante, la siguiente pregunta es: ¿Cuál(es) variable(s) (es)son importante(s)?

Si agregamos una variable al modelo de regresión, la suma de cuadrados de la regresión aumenta y la suma de cuadrados residuales disminuye. Se debe decidir si el aumento de la suma de cuadrados de la regresión es suficiente para garantizar el uso del regresor adicional en el modelo.

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

El juego de hipótesis para probar la significancia de cualquier coeficiente β_j $j = 1, 2, \cdots, k$, está dado por:

$$H_0 : \beta_j = 0 \qquad Vs \qquad H_1 : \beta_j \neq 0.$$

Si no se rechaza $H_0 : \beta_j = 0$, quiere decir que se puede eliminar el regresor x_j del modelo. El **estadístico de prueba** para esta hipótesis es,

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}, \tag{1.23}$$

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

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

Notemos que ésta es una **prueba parcial**, por que el coeficiente de regresión $\hat{\beta}_j$ depende de todas las demás variables regresoras x_i $(i \neq j)$, que hay en el modelo. Esto es una prueba de la **contribución** de x_j dadas las demás variables del modelo.

1.4. Comprobación de las Suposiciones del Modelo

Las principales **suposiciones** sobre el modelo de regresión son las siguientes:

- Y esta relacionada con X mediante un modelo de regresión lineal, donde β es el vector de parámetros desconocidos que determinan la recta de regresión:
$$E(Y|X = x_i) = \beta_0 + \beta x_i, \;\; i = 1, \cdots, n. \tag{1.24}$$
- Las variables regresoras $x_1, x_2, \cdots, x_k$ se consideran fijas, i.e; no son variables aleatorias.

- $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ son variables aleatorias **no observables**, **independientes**, distribuidas normalmente con $\mu = 0$ y σ^2 constante. Esto puede escribirse como:

$$\varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \cdots, n. \tag{1.25}$$

Siempre se debe tener en cuenta que la validez de estas premisas es dudosa y se deben hacer análisis para examinar la adecuación del modelo que se ha desarrollado en forma tentativa. Las grandes violaciones a las suposiciones pueden dar como resultado un modelo inestable, en el sentido que una muestra distinta puede conducir a un modelo totalmente diferente y por tanto obtener conclusiones opuestas. En general, no se pueden detectar desviaciones respecto a las premisas básicas examinando los estadísticos estándar de resumen (t, F o R^2). Éstas propiedades son globales del modelo y como tal no aseguran la adecuación del mismo.

En este capítulo se presentarán algunos métodos de utilidad para diagnosticar violaciones a las suposiciones básicas de la regresión. Estos métodos se basan principalmente en el estudio de los **residuales** del modelo.

1.4.1. Análisis de Residuales

Como mencionamos antes, los residuales se definen como:

$$\varepsilon = y_i - \hat{y}_i, \quad i = 1, \cdots, n. \tag{1.26}$$

siendo y_i una observación y $\hat{y}_i$ su valor ajustado correspondiente. Podemos considerar que un residual es la **desviación** entre los **datos** y el **ajuste**, 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, por lo que toda desviación de las suposiciones sobre los errores se debe reflejar en los residuales.

Los residuales tienen varias propiedades importantes. Tienen media cero y su varianza promedio se estima con:

$$\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n - (k+1)} = \frac{SS_{Res}}{n - (k+1)} = MS_{Res}. \tag{1.27}$$

Sin embargo los residuales no son independientes, ya que los n residuales sólo tienen $n - (k + 1)$ grados de libertad asociados a ellos. Se ha observado que cuando no hay independencia en los residuales se tiene poco efecto para comprobar la adecuación del modelo, siempre y cuando, n no sea pequeña en relación con la cantidad de parámetros.

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

1.4.2. Gráfica de Probabilidad Normal

Las pequeñas desviaciones respecto a la hipótesis de normalidad no afectan mucho al modelo, pero tener desviaciones grandes de no normalidad es potencialmente peligroso, por que los estadísticos t o F dependen de la suposición de normalidad. Además, si los errores provienen de una distribución con colas más gruesas que la normal, el ajuste por mínimos cuadrados será sensible a un subconjunto menor de datos. Las distribuciones de los errores con *colas* gruesas generan con frecuencia valores atípicos que *jalan* demasiado en su dirección el ajuste por mínimos cuadrados. En esos casos se deben considerar otras técnicas de estimación [10].

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.

Sean $\varepsilon_{[1]} < \varepsilon_{[2]} < \cdots < \varepsilon_{[n]}$ los residuales ordenados en orden creciente. Si se grafican $\varepsilon_{[i]}$ en función de la probabilidad acumulada $P_i = (i - \frac{1}{2})/n$, $i = 1, 2, \cdots, n$, 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 clasificado $\varepsilon_{[i]}$ en función del *valor normal esperado*, $\phi^{-1}[(i - \frac{1}{2})/n]$, donde ϕ representa la distribución normal estándar acumulada. Esto es consecuencia de que $E(\varepsilon_{[i]}) \simeq \phi^{-1}[(i - \frac{1}{2})/n]$ [10].

El estudio de las gráficas ayuda a adquirir un grado de percepción de cuánta desviación de la recta es aceptable. Con frecuencia, los tamaños pequeños de muestra ($n \leq 16$) producen gráficas de probabilidad normal que se desvían bastante de la linealidad. Para muestras mayores ($n \geq 32$), las gráficas se comportan mucho mejor. Por lo general, se requieren alrededor de 20 puntos para producir gráficas de probabilidad suficientemente estables como para poder interpretarse con facilidad.

1.4.3. Gráfica de Residuales en Función de los $\hat{Y}$

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

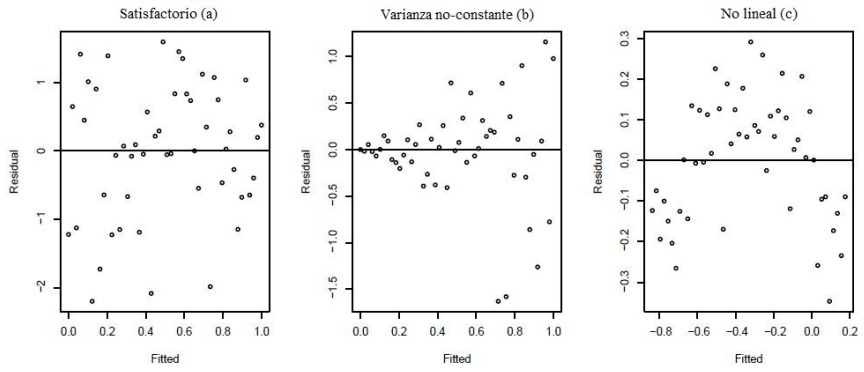

Figura 1.1: Patrones en las gráficas de residuales.

Si la gráfica que obtengamos es parecida a la Figura 1.1 del lado izquierdo, esta indica que los residuales se pueden encerrar en una banda horizontal y por tanto, no hay defectos graves en el modelo.

Las gráficas parecidas a la Figura 1.1 del centro, indican que la varianza de los errores no es constante.

El método más utilizado para manejar la varianza no común es aplicar una transformación adecuada ya sea a las variables regresoras o a la variable de respuesta.

Una gráfica en curva, como la de la Figura 1.1 del lado derecho, señala **no linealidad**. Esto podría indicar que se necesitan otras variables regresoras en el modelo.

Capítulo 2

Multicolinealidad

Los modelos de regresión son utilizados en diversas aplicaciones y un problema serio que puede influir mucho sobre la utilidad de un modelo es la **multicolinealidad** y la alta dimensionalidad de sus variables predictoras.

La multicolinealidad describe la dependencia casi lineal entre las variables predictoras, las cuales son las columnas de la matriz X; por lo cual es claro que una dependencia lineal puede influir en forma importante sobre la capacidad de estimar los coeficientes de regresión. Este también es un problema que hace difícil cuantificar con precisión el efecto que cada variable predictora ejerce sobre la variable dependiente Y [14].

2.1. Fuentes de Multicolinealidad

Definiremos nuestro modelo de regresión lineal con la ecuación (1.4) donde Y es una matriz de $n \times 1$ de las observaciones, $\mathbf{X}$ es una matriz de $n \times k$ de variables regresoras, β es un vector de $k \times 1$ de los coeficientes de regresión y ε es una matriz de $n \times 1$ de errores aleatorios, donde $\varepsilon_i \sim N(0, \sigma^2)$.

Supondremos que las variables regresoras y la respuesta se han centrado y escalado a longitud unitaria[1], en consecuencia, $R = \mathbb{X}'\mathbb{X}$ es una matriz de correlaciones entre variables de orden $k \times k$ y además $\mathbb{X}'Y$ será una matriz de tamaño $k \times 1$, de correlaciones entre los regresores y la respuesta.

Sea $\mathbb{X}_j$ la j-ésima columna de la matriz $\mathbb{X}$, de modo que $\mathbb{X} = [\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_k]$, entonces $\mathbb{X}_j$ contiene los n niveles de la j-ésima variable regresora. Definiremos formalmente la multicolinealidad en términos de la dependencia lineal de las columnas de $\mathbb{X}$.

[1] Ver Apéndice B.2

Recordemos que las columnas de $\mathbb{X}$ serán linealmente dependientes si existe un conjunto de constantes $\alpha_1, \alpha_2, \ldots, \alpha_k$ no todas iguales a cero, tales que

$$\sum_{j=1}^{k} \alpha_j \mathbb{X}_j = 0. \tag{2.1}$$

Si la ecuación (2.1) es exactamente válida para un subconjunto de las columnas de $\mathbb{X}$, el rango de la matriz R es menor que k y no existirá $(\mathbb{X}'\mathbb{X})^{-1}$. Ahora supongamos que la ecuación (2.1) es válida para algún subconjunto de columnas de $\mathbb{X}$, en este caso tendremos una dependencia casi lineal en $R = \mathbb{X}'\mathbb{X}$ y por tanto, diremos que existe un problema de **multicolinealidad** [10].

Como hemos mencionado antes, la presencia de multicolinealidad puede hacer que el análisis del modelo de regresión, por mínimos cuadrados no sea adecuado, esto es principalmente debido a diversas fuentes de multicolinealidad, como las siguientes:

- El método de recolección de los datos que se empleo.
- Restricciones en el modelo o en la población.
- Especificación del modelo.
- Un modelo sobredefinido.

Es importante entender las diferencias entre estas fuentes de multicolinealidad, por que los datos y la interpretación del modelo que resulte, depende en cierto grado de la causa de este problema.

El **método de obtención de los datos** puede originar problemas de multicolinealidad cuando el analista sólo muestrea un subespacio de la región de los regresores definidos. Así mismo, la multicolinealidad causada por la técnica de muestreo no es inherente al modelo o a la población que se muestrea. Por otra parte, las **restricciones en el modelo** o en la población que se muestrea son causas de multicolinealiad.

También se puede inducir la multicolinealidad por la **elección del modelo**. Con frecuencia se encuentran casos como ésos, cuando dos o más regresores tienen dependencia casi lineal y el retener esos regresores puede contribuir a la multicolinealidad, en esos casos suele ser preferible algún subconjunto de regresores. Por último, tenemos un **modelo sobredefinido** donde se tienen más variables regresoras que observaciones, donde el método más común para manejar la multicolinealidad en ese contexto es eliminar algunas variables regresoras.

2.2. Efectos de la Multicolinealidad

La presencia de multicolinealidad tiene una gran cantidad de efectos graves sobre los estimados de coeficientes de regresión por mínimos cuadrados [10].

Supongamos que solo tenemos dos variables regresoras x_1 y x_2, y que se han escalado a longitud unitaria las variables x_1, x_2 y Y, de modo que el modelo está dado por

$$Y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

Así, las ecuaciones normales de mínimos cuadrados son

$$(\mathbb{X}'\mathbb{X})\hat{\beta} = \mathbb{X}'Y.$$

$$\begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} r_{1Y} \\ r_{2Y} \end{pmatrix},$$

en donde r_{ij} es la correlación simple entre x_i y x_j, y r_{jY} es la correlación simple entre x_j y Y, $j = 1, 2$. Ahora bien, la inversa de $R(\mathbb{X}'\mathbb{X})$ es

$$C = (\mathbb{X}'\mathbb{X})^{-1} = R^{-1} \begin{pmatrix} \dfrac{1}{1-r_{12}^2} & \dfrac{-r_{12}}{1-r_{12}^2} \\ \dfrac{-r_{12}}{1-r_{12}^2} & \dfrac{1}{1-r_{12}^2} \end{pmatrix}, \tag{2.2}$$

y los estimadores de los coeficientes de regresión son

$$\hat{\beta}_1 = \frac{r_{1Y} - r_{12}r_{2Y}}{1-r_{12}^2}, \quad \hat{\beta}_2 = \frac{r_{2Y} - r_{12}r_{2Y}}{1-r_{12}^2}.$$

Si hay fuerte multicolinealidad entre x_1 y x_2, el coeficiente de correlación r_{12} será grande. De acuerdo con la ecuación (2.2) se puede observar que cuando $|r_{12}| \to 1$, $Var(\hat{\beta}_j) = C_{jj}\sigma^2 \to \infty$ y $Cov(\hat{\beta}_1, \hat{\beta}_2) = C_{12}\sigma^2 \to \pm\infty$, dependiendo de si $r_{12} \to +1$ o $r_{12} \to -1$. Por consiguiente, la fuerte multicolinealidad entre x_1 y x_2 nos da como resultado **grandes varianzas y covarianzas** de los estimadores de los coeficientes de regresión por mínimos cuadrados. Esto implica que distintas muestras tomadas con los mismos valores de x podrían ocasionar estimaciones muy diferentes de los parámetros del modelo.

Cuando hay más de dos variables regresoras, la multicolinealidad produce efectos parecidos. Se puede demostrar que los elementos diagonales de la matriz $C = (\mathbb{X}'\mathbb{X})^{-1}$ son

$$C_{jj} = \frac{1}{1-R_j^2}; \;\; j = 1, 2, \ldots, k;$$

en donde R_j^2 es el coeficiente de determinación múltiple de la regresión de x_j respecto a las demás variables predictoras. Si hay fuerte multicolinealidad entre x_j y cualquier subconjunto de los demás regresores, el valor de R_j^2 será cercano a 1.

Como la varianza de $\hat{\beta}_j$ es $Var(\hat{\beta}_j) = C_{jj}\sigma^2 = (1 - R_j^2)^{-1}\sigma^2$, una fuerte multicolinealidad implica que la varianza del estimado del coeficiente de regresión β_j por mínimos cuadrados es muy grande, por lo general, la covarianza de $\hat{\beta}_i$ y $\hat{\beta}_j$ también será grande, si los regresores x_i y x_j están correlacionados.

2.3. Diagnósticos de Multicolinealidad

Se han propuesto varias técnicas para **detectar la multicolinealidad**. A continuación se describirán algunas de esas medidas de diagnóstico.

Las características deseables en un método de diagnóstico requieren que este refleje el grado del problema de multicolinealidad, y que proporcione información de utilidad para determinar qué regresores están implicados.

2.3.1. Examen de la Matriz de Correlación

Una medida sencilla para detectar la multicolinealidad es la inspección de los elementos r_{ij} no diagonales en R (la matriz de correlaciones). Si los regresores x_i y x_j son casi linealmente dependientes, entonces $|r_{ij}|$ será proximo a la unidad.

Es útil examinar las correlaciones simples r_{ij} entre los regresores, para detectar la dependencia lineal, sólo entre **pares de regresores**, desafortunadamente, cuando intervienen más de dos regresores en una dependencia casi lineal, no hay la seguridad de que alguna de las correlaciones apareadas r_{ij} sean grandes.

2.3.2. Factores de Inflación a la Varianza

La multicolinealidad puede ser determinada mediante el cálculo del Factor de Inflación de la Varianza VIF *(de variance inflation factor)*. Se puede demostrar que en general, el factor de inflación de la varianza para el j-ésimo coeficiente de regresión se escribe como sigue [14].

$$VIF_j = C_{jj} = \frac{1}{1 - R_j^2}, \quad i = 1, \cdots, n. \tag{2.3}$$

R_j^2 es el coeficiente de determinación obtenido cuando se hace la regresión de x_j respecto a las demás variables predictoras. Es claro que si x_j depende casi linealmente de alguno de los demás regresores, entonces R_j^2 será cercano a uno y *VIF* será grande. Como la varianza de los j-ésimos coeficientes de regresión es $Var(\hat{\beta}_j) = VIF_j\sigma^2$, se puede considerar que *VIF* es el factor en el que aumenta la varianza de $\hat{\beta}_j$ debido a dependencias casi lineales entre los regresores.

El factor *VIF* para cada término del modelo mide el efecto combinado que tienen las dependencias entre los regresores sobre la varianza de ese término. La experiencia indica que si cualquiera de los *VIF* es mayor que 5 o 10, es inicio de que los coeficientes asociados de regresión están mal estimados debido a la multicolinealidad.

2.3.3. Análisis del Eigensistema de R

Los **eigenvalores** o valores propios de R, a saber $\lambda_1, \lambda_2, \ldots, \lambda_k$ se pueden usar para medir el grado de multicolinealidad en los datos. Si hay una o más dependencias casi lineales en los datos, uno o más eigenvalores serán pequeños, lo cual indica que hay dependencias casi lineales entre las columnas de R [10].

Algunos analistas prefieren examinar el **número de condición** de R, el cual se define como:

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}. \tag{2.4}$$

En general, si el número de condición es menor que 100, no hay problema grave de multicolinealidad. Los números de condición de 100 a 1000 implican multicolinealidad de moderada a fuerte, y si κ es mayor que 1000, es indicio de una fuerte multicolinealidad.

Los **índices de condición** de la matriz R son

$$\kappa_j = \frac{\lambda_{max}}{\lambda_j}, \quad j = 1, \cdots, k. \tag{2.5}$$

Es claro que el máximo indice de condición es el número de condición definido por la ecuación (2.4). La cantidad de índices de condición que son grandes (digamos 1000) es una medida útil de la cantidad de dependencias casi lineales en R.

También se puede aplicar el **análisis del eigensistema** para identificar la naturaleza de las dependencias casi lineales en los datos. La matriz R se puede descomponer como sigue,

$$R = T'\Lambda T,$$

donde Λ es una matriz diagonal de tamaño $k \times k$, cuyos elementos en la diagonal principal son los eigenvalores λ_j $(j = 1, 2, \ldots, k)$ de R y T es una matriz ortogonal de $k \times k$, cuyas columnas son los eigenvetores de R. Sean $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_k$ las columnas de T. Si el eigenvalor λ_j es cercano a cero, indicando una dependencia casi lineal en los datos, los elementos del eigenvector t_j son los coeficientes $\alpha_1, \alpha_2, \ldots, \alpha_k$ en la ecuación (2.1).

2.4. Métodos para Corregir la Multicolinealidad

Se han propuesto varias técnicas para manejar los problemas causados por la multicolinealidad. Entre los métodos generales están el reunir más datos, la reespecificación del modelo o bien el uso de métodos de estimación distintos al de mínimos cuadrados ordinarios. Entre ellos, pueden citarse los métodos de selección de variables, que son técnicas que permiten reespecificar el modelo.

Otro grupo de técnicas que se encaminan a buscar estimadores sesgados de los coeficientes de regresión pero con varianzas más pequeñas que los estimadores de mínimos cuadrados, son la regresión con componentes principales (RCP) y la regresión por mínimos cuadrados parciales (PLS, por sus siglas en inglés) [8].

2.4.1. Regresión con Componentes Principales (RCP)

Para estudiar las relaciones que se presentan entre variables correlacionadas (que miden información común) se puede transformar el conjunto original de variables en otro conjunto de nuevas variables *no* correlacionadas entre sí (que no tenga repetición o redundancia en la información) llamado conjunto de *componentes principales* [10].

Las nuevas variables son combinaciones lineales de las originales y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra.

De modo ideal, se buscan $m < k$ variables que sean combinaciones lineales de las k originales y que estén incorrelacionadas, recogiendo la mayor parte de la información o variabilidad de los datos.

La regresión de componentes principales es un método matemático que aplica mínimos cuadrados sobre un nuevo conjunto de variables llamadas componentes principales, obtenidas a partir de la matriz de correlación R.

La forma canónica del modelo es

$$Y = Z\alpha + \varepsilon, \tag{2.6}$$

donde,

$$Z = \mathbb{X}T, \;\; \alpha = T'\beta, \;\; T'RT = Z'Z = \Lambda.$$

Recordemos que $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_k)$ es una matriz diagonal de $k \times k$ de los eigenvalores de R, y T es una matriz ortogonal cuyas columnas son los eigenvectores asociados a cada $\lambda_1, \lambda_2, \ldots, \lambda_k$.

La matriz de componentes principales Z de orden $n \times k$, es obtenida transformando la matriz $\mathbb{X}$, de la siguiente manera,

$$\begin{aligned} Z &= \mathbb{X}T \\ &= \mathbb{X}(\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_k) \\ &= (\mathbb{X}\boldsymbol{t}_1, \mathbb{X}\boldsymbol{t}_2, \ldots, \mathbb{X}\boldsymbol{t}_k) \\ &= (Z_1, Z_2, \ldots, Z_k). \end{aligned}$$

Las columnas de Z definen un nuevo conjunto de regresores ortogonales, como $Z = [Z_1, Z_2, \ldots, Z_k]$ llamados *componentes principales.*

El estimador de $\hat{\alpha}$ por mínimos cuadrados es

$$\hat{\alpha} = (Z'Z)^{-1}Z'Y = \Lambda^{-1}Z'Y, \tag{2.7}$$

y la matriz de covarianza de $\hat{\alpha}$ es,

$$Var(\hat{\alpha}) = \sigma^2(Z'Z)^{-1} = \sigma^2\Lambda^{-1}. \tag{2.8}$$

Así, un pequeño eigenvalor de R indica que la varianza del coeficiente ortogonal de regresión correspondiente será grande. Ya que

$$Z'Z = \sum_{i=1}^{k}\sum_{j=1}^{k} Z_i Z_j' = \Lambda,$$

se llama con frecuencia *varianza de la j-ésima componente principal* al eigenvalor λ_j. Si todas las λ_j son iguales a 1, los regresores **originales** son ortogonales, mientras que si una λ_j es exactamente igual a cero, esto implica una regresión casi perfecta entre los regresores originales. Si hay una o más de las λ_j cercanas a cero, quiere decir que hay multicolinealidad. Notemos también que la matriz de covarianza de los coeficientes de los eigenvalores de regresión $\hat{\beta}$ es

$$Var(\hat{\beta}) = Var(T\hat{\alpha}) = T\Lambda^{-1}T'\sigma^2.$$

Esto implica que la varianza de $\hat{\beta}_j$ es $\sigma^2\sum_{j=1}^{k} t_j^2/\lambda_j$. Por consiguiente, la varianza de $\hat{\beta}_j$ es una combinación lineal de los reciprocos de los eigenvalores. Eso muestra cómo uno o más eigenvalores pequeños pueden destruir la precisión del estimado $\hat{\beta}_j$, por mínimos cuadrados.

Antes se hizo la observación de cómo los eigenvalores y eigenvectores de R proporcionan información específica acerca de la naturaleza de la multicolinealidad. Como $Z = \mathbb{X}T$, entonces

$$Z_i = \sum_{j=1}^{k} t_{ij}\mathbb{X}_j, \tag{2.9}$$

siendo $\mathbb{X}_j$ la j-ésima columna de la matriz $\mathbb{X}$ y t_{ij} son los elementos de la i-ésima columna de T (el i-ésimo eigenvector de R). Si la varianza de la i-ésima componente principal (λ_i) es pequeña, quiere decir que Z_i es casi constante , y de acuerdo con la ecuación (2.9), hay una combinación lineal de los **regresores originales** que es casi constante. La definición de multicolinealidad lo que nos dice es, que las t_{ij} son las constantes de la ecuación (2.1), por consiguiente, la ecuación (2.9) explica por que los elementos del eigenvector asociado con un pequeño eigenvalor de R identifican a los regresores que intervienen en la multicolinealidad.

El método de regresión con componentes principales combate a la multicolinealidad, al usar en el nuevo modelo menos componentes que el conjunto completo de componentes principales. Para obtener el estimador de componentes principales, se supone que los regresores están ordenados por eigenvalores decrecientes $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$. Supongamos que los últimos s eigenvalores de éstos son aproximadamente iguales a cero. En la regresión con componentes principales, se eliminan los componentes pricipales que correspondan a eigenvalores cercanos cero, y se aplica el método de mínimos cuadrados a los componentes restantes. Esto es

$$\hat{\alpha}_{CP} = \boldsymbol{B}\hat{\alpha},$$

en donde $b_1 = b_2 = \cdots = b_{k-s} = 1$, y $b_{k-(s+1)} = b_{k-(s+2)} = \cdots = b_k = 0$. Así el estimador de componentes principales es

$$\hat{\alpha}_{CP} = \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_{k-s} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Uno de los problemas que resuelve la RCP, es que reduce la dimensionalidad del problema, para lo cual debemos poner nuestra atención sobre las primeras componentes principales, pero sin perder mucha variabilidad.

Para elegir el número de componentes existen diferentes criterios:

- Gráfico de los λ_j (screeplot): Se observa en el gráfico a partir de qué valor de λ_j la gráfica se aplana.

- Regla de Kaiser : Se excluyen todos los componentes principales con eigenvalores menores a 1.

2.4.2. Regresión por Mínimos Cuadrados Parciales (*PLS*)

La regresión PLS es un método para analizar un conjunto de datos el cual se caracteriza por tener muchas variables predictoras, con problemas de multicolinealidad, y pocas observaciones [11].

Esta metodología (*PLS*) generaliza y combina características de la RCP y el Análisis de Regresión Múltiple. La popularidad por esta técnica va en aumento debido a la evidencia de sus soluciones en las diferentes ciencias [13].

En general, la regresión PLS consta de dos pasos fundamentales [14] [3]

- Transformar la matriz X, con la ayuda del vector de respuestas Y, en una matriz de componentes ortogonales, $T = (T_1, T_2, \cdots, T_k)$ llamadas componentes *PLS*.
- Realizar un nuevo Análisis de Regresión Múltiple utilizando la misma variable de respuesta Y, y como variables predictoras las componentes PLS.

Para llevar a cabo la regresión de Y con las variables $x_1, x_2, \cdots, x_k$, PLS trata de encontrar nuevos factores que juegan el mismo papel que las $X's$. Estos nuevos factores se llaman variables latentes o componentes. Análogamente como en el método de RCP, cada componente es una combinación lineal de $x_1, x_2, \cdots, x_k$, pero mientras *RCP* usa sólo la variación de X para construir los nuevos factores, PLS usa tanto la variación de X como de Y para construir los nuevos factores que se usarán como variables explicatorias del modelo [5].

La reducción de la dimensionalidad puede ser aplicada directamente sobre los componentes ya que estos son ortogonales. El número de componentes necesarios para el análisis de regresión debe ser mucho menor que el número de predictoras. Existen diferentes algoritmos para obtener los estimadores de la regresión PLS, pero los más usados son el NIPALS y el SIMPLS [16].

2.5. Capacidad Predictiva de los Modelos

Para comparar la capacidad predictiva de los modelos es útil introducir varias medidas del ajuste del modelo a los datos y de la fuerza de predicción de esos modelos. Con frecuencia se usan ecuaciones de regresión para predecir observaciones o para la estimación de la respuesta promedio, en general, se desea seleccionar los regresores de tal modo que el error cuadrático medio de la predicción se reduzca al mínimo, para ello se podría usar la estadística de *PRESS* [8].

Definimos los residuales *PRESS* como $\varepsilon_{(i)} = y_{(i)} - \hat{y}_{(i)}$ para toda $i = 1, \cdots, n$; siendo $\hat{y}_{(i)}$ el valor predicho de la i-ésima respuesta observada, basado en un ajuste del modelo con los $n - 1$ puntos de muestra.

La Suma de Cuadrados de Error de Predicción, conocida como la estadística *PRESS*, se considera una medida de lo bien que funciona un modelo de regresión para predecir nuevos datos y se define como la suma de los residuales *PRESS* al cuadrado que son los residuos que se obtiene entre el valor observado y el valor predicho de la i-ésima respuesta observada, basado en un ajuste de modelo con los puntos restantes de la muestra.

$$\boldsymbol{PRESS} = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2. \tag{2.10}$$

Se considera que el *PRESS* es una medida de lo bien que funciona un modelo de regresión para *predecir nuevos datos*. Lo deseable es tener un modelo con valor pequeño de PRESS.

Con la estadística PRESS se puede calcular otro estadístico parecido a la R^2 para la predicción, que se define como,

$$R_{pred}^2 = 1 - \frac{PRESS}{SS_T}.$$

El estadístico da cierta indicación de la capacidad predictiva del modelo de regresión. Una aplicación muy importante de estos estadísticos es poder comparar los modelos de regresión. En general un modelo con valor pequeño de PRESS es preferible a uno con PRESS grande. El R_{pred}^2 se interpreta de manera similar al estadístico R^2 que mide la bondad del ajuste del modelo con los datos: a mayores valores de estos estadísticos, mayor es la bondad del ajuste y mayor es la capacidad predictiva del modelo [10].

El uso principal de los modelos de regresión es la predicción de futuras observaciones, por ello utilizaremos la estadística PRESS y el R_{pred}^2 para seleccionar el mejor modelo [8].

Capítulo 3

Estudio del Carbono Orgánico en Suelos (COS) en la Caldera de Teziutlán

Los problemas ambientales, en particular los referidos al cambio climático, son muy importantes para el ser humano y es necesario utilizar modelos estadísticos para investigar dichos problemas en el estado de Puebla.

3.1. Planteamiento del Problema

El calentamiento global producto del aumento de las emisiones antropogénicas de carbono a la atmósfera y sus consecuencias sobre el medio ambiente ha generado controversias en todos los ámbitos de la Sociedad; por ello, se han planteado diversas ideas sobre la posibilidad de mitigar el cambio climático [2].

El dióxido de carbono (CO_2) es uno de los llamados gases de efecto invernadero. Estos gases son continuamente emitidos y removidos en la atmósfera mediante procesos naturales sobre la Tierra, pero las actividades antropogénicas causan cantidades adicionales de los mismos, incrementando sus concentraciones en la atmósfera, lo que tiende a sobrecalentarla.

El carbono orgánico del suelo (COS) es un gran y activo reservorio, y los ecosistemas forestales pueden absorber cantidades significativas de CO_2, por lo que hay un gran interés en incrementar el contenido de carbono en estos ecosistemas, lo que se conoce como *secuestro de carbono*.

El secuestro de carbono por el suelo es el proceso de transformación del carbono del aire (dióxido de carbono o CO_2) en carbono almacenado en suelo.

También recibe el nombre de sumidero (transferencia neta de CO_2 atmosférico a la vegetación y al suelo para su almacenamiento). El dióxido de carbono es absorbido por las plantas a través del proceso de *fotosíntesis* y es incorporado al tejido vegetal. Cuando las plantas mueren, el carbono de las hojas, tallos y raíces se descompone en el suelo y se convierte en materia orgánica [7].

A través del secuestro de carbono, los niveles de dióxido de carbono atmosférico son reducidos con el incremento de los niveles de carbono del suelo. Así el carbono, si no es perturbado, puede permanecer por muchos años como materia orgánica estable. Este proceso, como mencionamos antes, puede atenuar el cambio climático global o el calentamiento de la atmósfera. A pesar de la importancia del secuestro de carbono para la mitigación del cambio climático, su evaluación se encuentra muy limitada en muchas zonas del mundo, en particular, en los suelos de la Sierra Norte de Puebla, México.

Los estudios del caso en México muestran que los métodos utilizados en general son diferentes; indefinidos en cuanto a los parámetros y variables que se incluyen; incomparables en términos de las categorías que utilizan y con escalas de trabajo incompatibles.

Aún son escasas las investigaciones en el país que tratan de explicar el almacenamiento de carbono en suelos a través de propiedades de los mismos [2]. Las muestras de suelo en las que se determinaron los diferentes contenidos extraíbles corresponden a perfiles, que fueron caracterizados previamente en el *Departamento de Investigación en Ciencias Agrícolas* (DICA) del *Instituto de Ciencias* de la Benemérita Universidad Autónoma de Puebla ($BUAP$).

El estudio se llevó a cabo en los suelos de la Región de Teziutlán, situada en la porción nororiental del estado de Puebla y que comprende una porción de la Región Terrestre Prioritaria para la Conservación en México (*RPT-105*), la cual se encuentra entre los paralelos $19°43'30''$ y $20°14'54''$ de latitud norte y los meridianos $97°07'42''$ y $97°43'30''$ de longitud occidental. Los suelos son derivados de material piroclástico, los cuales se presentan cubriendo una superficie aproximada de $846Km^2$.

La caracterización física y química de las muestras de suelo se efectuó de acuerdo a la Norma Oficial Mexicana NOM-021-RECNAT-2000 (SEMARNAT, 2000), determinándose 22 propiedades en muestras obtenidas de 21 perfiles de suelo dispuestos en localizaciones no regulares representativas de la zona de estudio Figura 3.1.

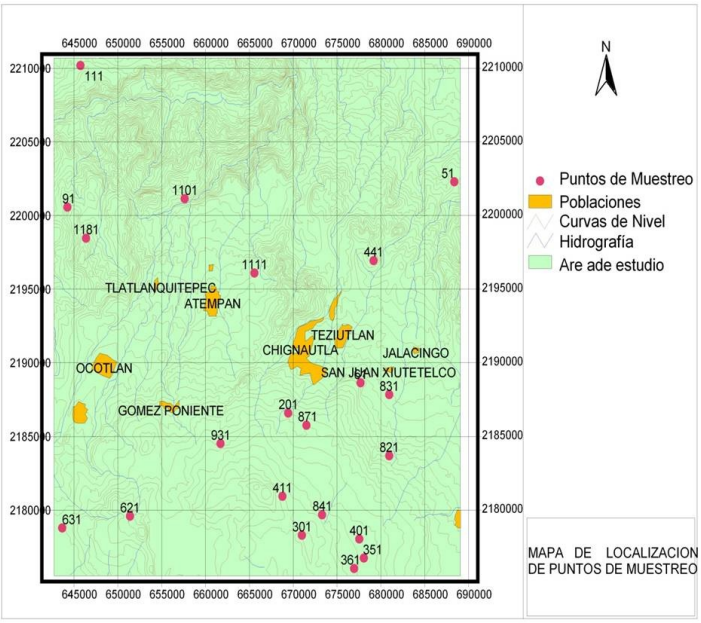

Figura 3.1: Gráfica de la zona de estudio donde se localizan los puntos de muestreo.

3.2. Análisis Estadístico

La información obtenida se analizó con diferentes técnicas de Regresión. Para dicho análisis se utilizó una base de datos que consta de 42 observaciones, 22 variables predictoras y una variable respuesta.

La variable respuesta es $Y = COS$ y las 22 variables predictoras corresponden a diferentes características que componen el suelo: la *Densidad Aparente* (DA), $\%Arena$, $\%Limo$, $\%Arcilla$, *Contenidos de Nitrógeno Total* (%N), *Relación* C/N, contenidos en *Al y Fe extraibles* (%Al y %Fe), *Retención de fosfatos*, *pH en NaF 1N*, *pH en* H_2O, *pH en* KCl, Delta pH, *Capacidad de Intercambio Catiónico Total* (CIC), *Porcentaje de saturación en bases* (%V), *Calcio* (Ca), *Magnesio* (Mg), *Sodio* (Na) y *Potacio* (K) Intercambiables.

A continuación se muestran los resultados obtenidos por cada método de regresión en el problema bajo estudio.

3.2.1. Modelo de Regresión Lineal Múltiple

A continuación se muestran los resultados obtenidos al hacer un análisis de regresión lineal múltiple con las variables descritas anteriormente, este modelo fue determinado con el uso del software estadístico R y la ecuación de regresión obtenida es la siguiente:

```
COS = 29526 + 259 Densidad - 298 Arena - 298 Limo - 300 Arcilla + 250 M.O
      - 413 C.Org - 17.7 N - 4.16 C.N + 1.51 RetenciónFosfatos
      + 199 Al.Extraible + 167 Fe.Extraible - 168 Al.Fe.extr - 1.66 pH.NaF.N
      + 39.0 pH.H2O.S - 42.1 pH.KCl + 68.5 Delta.pH + 2.01 CIC + 4.53 V
      - 29.0 Ca - 5.5 Mg - 18.5 Na - 15.1 K
```

El Cuadro 3.1 muestra los coeficientes del modelo de regresión obtenidos con el método de mínimos cuadrados ordinarios, los errores estándar para cada coeficiente, así como el estadístico *t de Student* con sus correspondientes valores p, donde podemos observar que en este caso los coeficientes obtenidos no contribuyen en forma significativa al modelo, pues los valores p son muy altos.

También puede apreciarse que el modelo de regresión lineal obtenido no tiene un buen ajuste, dado que el coeficiente de determinación es $R^2 = 0.6369$, es decir; el 63.69 % de la variabilidad de los datos queda explicado, por otra parte el coeficiente de determinación ajustado sólo alcanza el 21.65 %; además, la prueba F no es significativa ya que el `p-value` empírico es mucho mayor que algún nivel de significancia α aceptable.

Predictor	Coef	SE Coef	T	P	VIF
Constant	29526	35870	0.82	0.421	
Densidad	258.7	202.8	1.28	0.217	2.5
Arena	-298.0	358.3	-0.83	0.416	**202286.8**
Limo	-298.4	357.5	-0.83	0.414	**81219.9**
Arcilla	-300.2	357.9	-0.84	0.412	**71030.6**
M.O	249.8	433.9	0.58	0.5722	**23469.7**
C.Org	-413.4	746.6	-0.55	0.586	**23453.3**
N	-17.68	24.21	-0.73	0.474	1.5
C.N	-4.162	7.950	-0.52	0.607	3.6
RetenciónFosfatos	1.512	1.983	0.76	0.455	4.0
Al.Extraible	199.4	115.2	1.73	0.100	**93.6**
Fe.Extraible	167.36	92.93	1.80	0.088	**11.3**
Al.Fe.extr	-167.6	114.7	-1.46	0.160	**123.4**
pH.NaF.N	-1.660	2.317	-0.72	0.482	**10.8**
pH.H2O.S	38.96	51.77	0.75	0.461	7.4
pH.KCl	-42.122	9.93	-0.70	0.491	7.7
Delta.pH	68.54	72.81	0.94	0.358	4.4
CIC	2.011	3.905	0.52	0.612	**13.5**
V	4.533	3.890	1.17	0.258	**22.1**
Ca	-28.95	25.53	-1.13	0.271	**13.1**
Mg	-5.46	11.24	-0.49	0.633	1.9
Na	-18.50	41.60	-0.44	0.662	4.0
K	-15.07	41.25	-0.37	0.719	4.9

```
Residual standard error: 81.7 on 19 degrees of freedom
Multiple R-squared: 0.6369, Adjusted R-squared: 0.2165
F-statistic: 1.515 on 22 and 19 DF, p-value: 0.1821
```

Cuadro 3.1: Modelo de regresión: Método de mínimos cuadrados.

El análisis de los supuestos del modelo, referidos a la normalidad y homogeneidad en los residuos, pueden observarse a través de las Figuras 3.2 y 3.3 que se muestran a continuación, donde puede observarse que dichos supuestos no se cumplen.

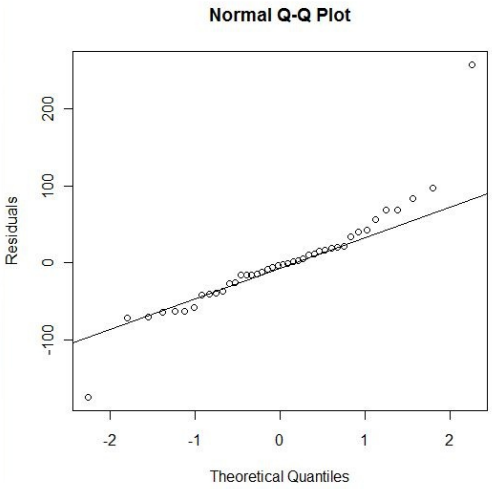

Figura 3.2: Gráfica de probabilidad Normal.

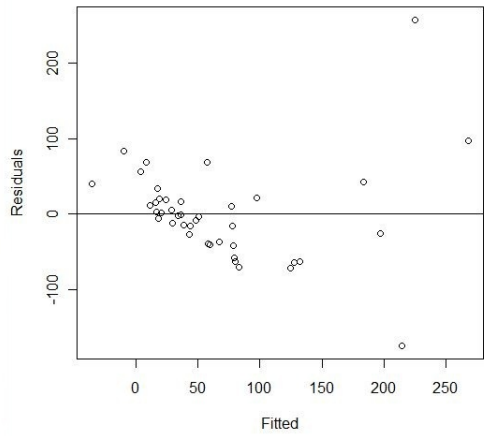

Figura 3.3: Gráfica de Residuos.

Para profundizar en el análisis de los supuestos del modelo requerimos indagar en la existencia de multicolinealidad, como podemos observar en el Cuadro 3.1, se detectaron problemas de multicolinealidad: pues los valores VIF, en **12** de las **22** variables predictoras, están en el rango de **10.8** a **202286.8**, lo cual es señal de que tenemos un problema serio de multicolinealidad. El alto valor del número condición, $\boldsymbol{\kappa = 7138.416}$ confirma también la existencia de multicolinealidad en un grado bastante alto.

3.2.2. Regresión con Componentes Principales

Después del diagnóstico realizado y la evidencia de la existencia de *multicolinealidad*, se decidió llevar a cabo una Regresión con Componentes Principales (RCP), ya que es una de las ténicas recomendada para corregir este problema.

Siguiendo el procedimiento señalado en el Capítulo anterior, la matriz de predictores se transformó en componentes principales para eliminar la multicolinealidad. Posteriormente se calcularon los eigenvalores de la matriz de correlaciones y utilizando la Regla de Kaiser, se excluyeron todos los componentes principales con eigenvalores menores que 1.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Eigenvalue	**5.3085**	**4.1620**	**2.4285**	**2.0625**	**1.4369**	**1.3017**	0.8880	0.8403
Standard deviation	2.3040	2.0400	1.5583	1.4361	1.1987	1.1409	0.9423	0.9166
Proportion of Variance	0.2412	0.1891	0.1103	0.0937	0.0653	0.0591	0.0403	0.0381
Cumulative Proportion	0.2412	0.4304	0.5408	0.6346	0.6999	0.7590	0.7994	0.8376
	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	Comp.16
Eigenvalue	0.7282	0.6002	0.5753	0.4739	0.3578	0.3193	0.1982	0.1370
Standard deviation	0.8533	0.7747	0.7584	0.6884	0.5981	0.5650	0.4451	0.3701
Proportion of Variance	0.0330	0.0272	0.0261	0.0215	0.0162	0.0145	0.0090	0.0062
Cumulative Proportion	0.8707	0.8980	0.9241	0.9457	0.9619	0.9765	0.9855	0.9917
	Comp.17	Comp.18	Comp.19	Comp.20	Comp.21	Comp.22		
Eigenvalue	0.0982	0.0573	0.0213	0.0049	0.0000	0.0000		
Standard deviation	0.3133	0.2394	0.1459	0.0697	0.2297	0.0836		
Proportion of Variance	0.0044	0.0026	0.0009	0.0002	0.0000	0.0000		
Cumulative Proportion	0.9962	0.9988	0.9997	0.9999	1.0000	1.0000		

Cuadro 3.2: Eigenvalores, para elegir las componentes principales.

El Cuadro 3.3 muestra en la primera fila los eigenvectores ordenados en forma decreciente, la segunda fila nos muestra la desviación estandar para cada eigenvalor, la tercera fila describe para cada eigenvalor la proporción de varianza explicada y la cuarta fila describe la proporción acumulada.

Siguiendo la Regla de Kaiser antes mencionada se eligen las 6 primeras componentes ($\lambda_i > 1$) y se puede observar en el Cuadro 3.2 que con estas componentes se explica más del 75 % de la variabilidad total del fenómeno .

Esta elección es respaldada con el gráfico de la pendiente (en inglés, screeplot) que se muestra en la Figura 3.4, en la cual se observa que a partir de la componente 6, es donde el comportamiento de las variables se estabiliza (aplana).

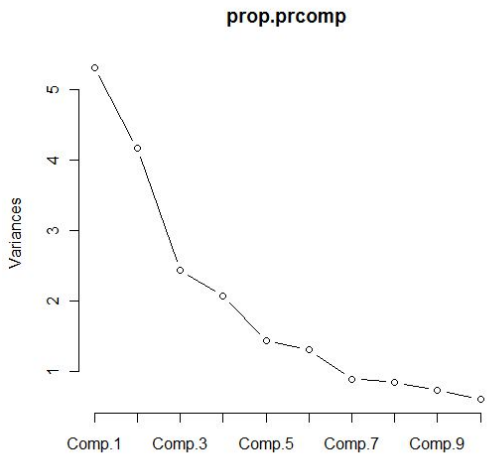

Figura 3.4: Gráfico de varianzas (*Screeplot*)

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Densidad	-0,167	-0,267	-0,073	-0,008	-0,348	0,106
Arena	0,173	**-0,354**	-0,169	0,206	0,104	-0,270
Limo	-0,089	0,245	0,279	-0,239	-0,067	0,325
Arcilla	-0,196	**0,334**	-0,013	-0,091	-0,104	0,109
M.O	0,195	0,271	0,007	0,340	0,276	0,252
C.Org	0,195	0,271	0,006	0,342	0,273	0,254
N	-0,012	0,133	0,099	0,286	-0,146	0,153
C.N	-0,123	0,288	0,137	0,015	0,071	-0,380
Ret.Fos	**0,305**	0,057	-0,216	-0,190	-0,228	0,071
Al.Ext	**0,323**	0,152	-0,137	-0,264	0,103	-0,027
Fe.Ext	0,186	0,208	-0,137	-0,212	-0,374	0,190
Al.Fe.ext	**0,324**	0,167	-0,145	-0,299	0,002	0,030
pH.NaF.N	**0,303**	-0,271	-0,121	0,056	0,074	0,143
pH.H2O.S	-0,190	-0,194	-0,211	-0,252	0,327	0,306
pH.KCl	-0,255	-0,028	-0,115	-0,416	0,302	0,009
Delta.pH	-0,028	0,287	0,138	-0,131	0,135	-0,486
CIC	**0,312**	-0,007	-0,092	-0,102	0,241	-0,098
V	-0,287	0,101	-0,296	0,177	-0,146	0,132
Ca	-0,116	0,225	-0,422	0,159	0,051	-0,068
Mg	-0,208	0,034	-0,268	0,028	0,265	0,106
Na	0,091	0,108	-0,326	0,112	-0,313	-0,229
K	-0,169	0,135	-0,459	0,015	0,025	-0,117

Cuadro 3.3: Eigenvectores correspondientes a las variables seleccionadas.

El Cuadro 3.3 tiene la siguiente interpretación: cada columna, está denotada por $PC1, \cdots, PC6$, que son los eigenvectores asociados a los respectivos eigenvalores ya ordenados de mayor a menor en el Cuadro 3.2.

Por ejemplo, para la primera componente, se puede observar que los coeficientes correspondientes a las variables $Ret.Fos$, $Al.Extraible$, $Al.Fe.extr$, $pH.NaF.N$ y CIC, son las que tienen mayor relevancia y según el Cuadro 3.2 ésta componente logra reunir alrededor de la cuarta parte de la información del fenómeno.

Por otra parte, en la segunda componente las variables más relevantes son *Arena* y *Arcilla* las cuales tienen signos contrarios, lo que significa que estas variables se encuentran en oposición. Este mismo análisis se puede realizar con cada componente para determinar las variables más relevantes en cada componente.

Posteriormente como indica el procedimiento, se llevo a cabo un nuevo análisis de regresión, con las componentes principales elegidas.

Podemos observar en el Cuadro 3.4 que el modelo de RCP obtenido tampoco tiene un buen ajuste, pues el coeficiente de determinación $R^2 = 0.3456$ es bastante bajo, mientras que el coeficiente de determinación ajustado alcanza el $23.35\,\%$ que es ligeramente más alto que en la RLM. Sin embargo, la prueba F es significativa al nivel de significancia $\alpha < 0.05$, ya que el valor de p empírico es 0.01, esto es, $p < \alpha$ y se rechaza la hipótesis nula, aceptándose que algún coeficiente es diferente de cero.

Predictor	Coef	SE Coef	T	P	VIF
Constant	67.43	12.47	5.41	0.000	
Cp1	19.20	5.478	3.51	0.001	1.0
Cp2	9.833	6.186	1.59	0.121	1.0
Cp3	-5.461	8.099	-0.67	0.505	1.0
Cp4	-3.971	8.788	-0.45	0.654	1.0
Cp5	-7.35	10.53	-0.70	0.490	1.0
Cp6	17.59	11.06	1.59	0.121	1.0

```
Residual standard error: 80.8098 on 35 degrees of freedom
Multiple R-squared: 0.3456, Adjusted R-squared: 0.2335
F-statistic: 3.081 on 6 and 35 DF, p-value: 0.01579
```

Cuadro 3.4: Modelo de RCP con las componentes elegidas.

Por otra parte es importante resaltar que aunque con esta nueva regresión se elimina el problema de multicolinealidad; pues los VIF son 1, un análisis adicional nos mostró que los demás supuestos siguen sin cumplirse.

3.2.3. Regresión por Mínimos Cuadrádos Parciales

Como se ha mencionado antes, la regresión PLS es un método para analizar conjuntos de datos los cuales se caracterizan por tener problemas de multicolinealidad, como lo es el caso que se presenta.

El Cuadro 3.5 muestra los resultados de la regresión PLS, así como la selección del mejor modelo. Puede apreciarse que el modelo con cuatro componentes tiene la mayor capacidad predictiva, por lo cual éste fue el modelo seleccionado.

Components	X Variance	Error SS	R-Sq	PRESS	Rs.q(Pred)
1	0.230528	228128	0.346865	267967	23.28%
2	0.325924	188149	0.461323	258144	26.09%
3	0.464035	174069	0.501636	250803	28.19%
4	**0.541830**	**166483**	**0.523356**	**246840**	**29.75%**
5	0.618580	163475	0.531966	252366	27.75%
6	0.655610	158439	0.546384	292275	16.32%
7	0.726539	155911	0.553624	299961	14.12%
8	0.759362	151564	0.566069	301034	13.81%
9	0.800449	148919	0.573641	317340	9.14%
10	0.840156	147212	0.578527	322398	7.70%

Cuadro 3.5: Regresión PLS: Selección del modelo.

El Análisis de varianza de ese modelo Cuadro 3.6 mostró que la prueba F es significativa con un p empírico de 0.000.

Analysis of Variance for COS

Source	DF	SS	MS	F	P
Regression	4	182798	45699.5	10.16	0.000
Residual Error	37	166483	4499.5		
Total	41	349281			

Cuadro 3.6: Análisis de varianza para COS por PLS (4 componentes)

En el Cuadro 3.7 se muestran los coeficientes de regresión del modelo. Puede apreciarse la importancia que tiene la *Densidad Aparente*, ya que toma el valor más alto en valor absoluto. Las variables *Al y Fe extraibles* (%Al y %Fe), *pH en* H_2O *S*, *Sodio* (Na) y *Potacio* (K), aunque con valores más bajos se muestran importantes en la explicación del COS.

	COS	COS standardized
Constant	-407.034	0.000000
DensidadA	**214.482**	0.230915
Arena	-0.009	-0.001501
Limo	-0.232	-0.025537
Arcilla	0.266	0.027396
M.O.	4.415	0.215471
C.Org.	7.583	0.215027
N	-3.892	-0.027243
C.N	1.418	0.047064
RetenciónFosfatos	1.062	0.147400
Al.Extraible	**19.898**	0.230973
Fe.Extraible	**25.124**	0.125432
Al.Fe.extr	**12.483**	0.167098
pH.NaF.N	0.059	0.011492
pH.H2O.S	**14.395**	0.104448
pH.KCl.S	3.305	0.021135
Delta.pH	4.331	0.017227
CIC	-0.758	-0.098558
V	1.196	0.199814
Ca	-6.576	-0.129073
Mg	-3.341	-0.056063
Na	**16.039**	0.107195
K	**-38.056**	-0.283054

Cuadro 3.7: Coeficientes de regresión.

3.2.4. Comparación de los Modelos Ajustados Según su Capacidad Predictiva

Posteriormente, con fines de comparación, se analizaron ambos modelos de regresión *RCP* y *PLS*. En esta parte se utilizo el paquete estadístico MINITAB 14, para calcular los valores del estadístico $PRESS$ y R^2_{Pred}, debido a que en el software R no se contaba con un programa para obtener dichos resultados.

Para hacer una comparación de los valores $PRESS$, se realizaron nuevos modelos de regresión considerando diferente número de componentes en cada modelo (componentes principales y componentes PLS respectivamente), esto con el fin de comparar el valor del $PRESS$ obtenido con las ambas metodologías.

De acuerdo con las observaciones anteriores se sugiere que la regresión ***RCP*** sea con 6 componentes, de modo que; si ambos modelos de regresión fuesen estimados con 6 componentes, los valores *PRESS* serán 379640 y 292275 respectivamente.

RCP	(No. de comp)	PLS	(No. de comp)
631620	(10)	322398	(10)
881279	(9)	317340	(9)
688408	(8)	301034	(8)
683901	(7)	299961	(7)
379640	(**6**)	**292275**	(**6**)
366661	(5)	252366	(5)
344069	(4)	246840	(4)
334969	(3)	250803	(3)
317842	(2)	258144	(2)
303262	(1)	267967	(1)

Cuadro 3.8: Comparación del *PRESS*.

Por otra parte, para la regresión ***PLS*** el mejor modelo fue el estimado con 4 componentes PLS. En el Cuadro 3.9 se muestra la comparación entre los modelos según su capacidad predictiva. Una metodología similar es brindada en [16].

Puede observarse que el modelo con mayor capacidad predictiva para estos datos, es el obtenido por el método PLS con 4 componentes. El Cuadro 3.9 muestra los valores $PRESS$ y R^2_{Pred} correspondientes a los modelos considerados anteriormente para hacer las comparaciones.

Estos resultados confirman que el método *PLS* supera al método *RCP*, ya que si consideramos el mismo número de componentes en para los dos métodos, en ambos casos el valor $PRESS$ es menor para la metodología PLS, de igual forma el R^2_{Pred} es más alto considerando las componentes PLS.

	RCP		PLS	
Components	PRESS	Rs.q(Pred)	PRESS	Rs.q(Pred)
1	303262	13.18 %	267967	23.28 %
2	317842	9.00 %	258144	26.09 %
3	334969	4.10 %	250803	28.19 %
4	**344069**	**1.49 %**	**246840**	**29.33 %**
5	366661	0.00 %	252366	27.75 %
6	379640	0.00 %	292275	16.32 %
7	683901	0.00 %	299961	14.12 %
8	688408	0.00 %	301034	13.81 %
9	881279	0.00 %	317340	9.14 %
10	631620	0.00 %	322398	7.70 %

Cuadro 3.9: Comparación del $PRESS$ y R^2_{pred}.

Resumiendo los resultados anteriores, podemos elegir el modelo que se obtiene con la metodología PLS con 4 componentes, pues éste nos brinda el R^2_{Pred} más alto y tiene un valor $PRESS$ relativamente pequeño. Para el modelo elegido la ecuación de regresión es:

```
COS = 67.4 + 25.2 Comp1 + 27.7 Comp2 + 12.1 Comp3 + 12.1 Comp4
```

Se puede observar en el Cuadro 3.10 que la prueba F es significativa con un p empírico de 0.000 por lo que el ajuste a los datos es bueno, por su parte las pruebas t para algunos de los regresores también son buenas. El R^2 explica en un 50.2 % el ajuste a los datos. Además el valor para R^2Pred nos dice que éste modelo tiene una capacidad predictiva del 29.33 %.

```
Predictor   Coef    SE Coef    T     P       VIF
Constant   67.43    10.44     6.46  0.000
Comp1      25.200   4.900     5.14  0.000   1.0
Comp2      27.725   9.385     2.95  0.005   1.0
Comp3      12.070   6.885     1.75  0.088   1.0
Comp4      12.097   9.316     1.30  0.202   1.0

S = 67.6813 R-Sq = 50.2% R-Sq(adj) = 46.2%
PRESS = 246840 R-Sq(pred) = 29.33%
F-statistic: 12.75 on 2 and 38 DF, p-value: 0.00
```

Cuadro 3.10: Coeficientes de regresión con componentes PLS.

Se debe resaltar que con esta nueva regresión se ha eliminado el problema de *multicolinealidad*, pues los valores VIF son igual a 1.

Conclusión

En la investigación empírica, como el caso que presentamos, es necesario extraer información de los datos numéricos, de ahí la importancia del enfoque estadístico en las mismas. El analista de estadística tiene la responsabilidad de presentar sus resultados de manera transparente facilitando la comprensión de los fenómenos bajo estudio.

En este trabajo se ha llevado a cabo un estudio comparativo entre diferentes métodos de regresión en el estudio de Carbono Orgánico del Suelo (COS) en función de otras propiedades del suelo.

De acuerdo con los resultados obtenidos se pudo observar que en este caso existen diversos problemas en el cumplimiento de los supuestos del modelo de regresión, uno de ellos es la existencia de *multicolinealidad*, la cual fue tratada con dos metodologías relativamente similares: la regresión con componentes principales (RCP) y la regresión por mínimos cuadrados parciales (PLS).

Como sabemos ambos métodos transforman las variables predictoras en componentes ortogonales, las cuales representan la solución al problema de multicolinealidad y permiten hacer una reducción de la dimensionalidad del espacio de variables predictoras, por ello se decidió hacer un análisis comparativo de estas metodologías.

Los resultados obtenidos con el análisis de los datos verifican la eficiencia de la regresión PLS sobre la regresión RCP, pues en este caso la comparación se realizó bajo el criterio de mejor capacidad predictiva. De esta forma si fijamos el mismo número de componentes para estimar ambos modelos de regresión: PLS y RCP, se observa que el valor $PRESS$ siempre es menor para el modelo PLS al igual que su capacidad predictiva.

Después de hacer la comparación de los valores $PRESS$ y R^2_{Pred} con ambos métodos se decidió elegir el modelo considerando 4 componentes PLS, ya que éste es el modelo que nos daba mayor información acerca del caso de estudio, además de tener una capacidad predictiva del 29.33 %, lo cual es de gran utilidad para la investigación.

Un análisis adicional de los residuos en el modelo elegido (PLS con 4 componentes) mostró que siguen sin cumplirse los supuestos de normalidad así como la homogeneidad de varianzas, por lo cual es recomendable utilizar otras técnicas y métodos para corregir estas fallas.

Cabe resaltar que el uso del lenguaje R, resultó muy eficiente para el análisis de los datos que se ilustran en este trabajo y es necesario señalar que para algunos cálculos, se utilizo como alternativa el paquete estadístico MINITAB 14 ya que con el paquete R no se obtenían los resultados requeridos para hacer un análisis completo.

Apéndice A

Base de Datos

$COS(ton/ha)$	$Densidad\,Aparente(g/cm3)$	$\%Arena$	$\%Limo$	$\%Arcilla$	$\%M.O.$	$\%C.Org.$	$\%N.Tot.$
19.22	0.61	30.10	42.70	27.20	18.10	10.50	0.81
17.47	0.78	33.30	41.60	25.10	1.20	0.70	0.06
12.48	0.78	34.80	39.10	26.10	6.90	4.00	4.27
24.57	0.78	43.20	35.60	21.20	1.60	0.90	0.07
19.34	0.62	31.00	39.30	29.70	9.00	5.20	0.32
34.56	0.80	40.50	32.90	26.60	3.10	1.80	0.14
16.33	0.71	28.20	40.20	31.60	7.90	4.60	0.33
38.64	0.69	35.80	42.30	21.90	2.80	1.60	0.12
22.20	0.80	48.30	25.60	26.10	6.40	3.70	0.25
73.94	0.79	43.80	33.60	22.60	4.50	2.60	0.20
22.19	0.87	59.10	25.40	15.50	2.90	1.70	0.11
12.28	0.89	70.70	13.60	15.70	0.50	0.30	0.02
46.90	0.70	35.10	48.60	16.30	11.50	6.70	0.39
31.25	0.63	47.90	26.60	25.50	2.80	1.60	0.13
40.37	0.69	39.10	49.10	11.80	6.80	3.90	0.26
5.14	0.79	53.70	33.90	12.40	0.40	0.20	0.01
18.82	0.71	45.40	35.60	19.00	9.10	5.30	0.41
42.24	0.66	39.00	33.50	27.50	5.50	3.20	0.23
76.73	0.63	42.10	25.60	32.30	10.00	5.80	0.41
51.87	0.57	41.60	38.30	20.10	2.30	1.30	0.07
70.00	0.50	45.70	32.30	22.00	12.00	7.00	0.54
16.86	0.77	30.20	48.50	21.30	0.50	0.30	0.04
171.05	0.69	27.80	34.70	37.50	12.70	7.40	0.46
481.90	0.72	20.30	45.50	34.20	11.90	6.90	0.53
22.13	0.75	35.30	31.80	32.90	10.10	5.90	0.45
28.00	0.64	37.90	22.60	39.50	4.40	2.50	0.19
62.44	0.80	73.80	16.00	10.20	7.69	4.46	0.46
60.06	0.84	61.10	26.00	12.90	2.25	1.30	0.13
225.72	0.76	45.84	37.64	16.52	13.67	7.92	0.60
364.78	0.69	67.84	25.64	6.52	10.25	5.94	0.50
53.12	0.53	45.80	41.64	12.88	13.30	7.71	0.00
117.95	0.66	73.48	23.64	2.88	5.99	3.47	0.00
30.19	0.86	46.20	32.40	21.40	4.60	2.70	0.36
27.92	0.94	66.20	22.70	11.10	1.50	0.90	0.20
52.89	0.60	89.32	9.00	1.68	11.26	6.53	0.45
125.52	0.67	92.80	5.40	1.75	8.50	4.93	0.50
87.00	0.75	53.00	34.64	12.36	13.80	8.00	0.16
32.04	0.89	56.00	28.64	15.36	1.30	0.80	0.09
64.13	0.75	45.22	27.28	27.50	9.90	5.70	0.35
36.30	0.61	50.22	38.28	11.50	5.90	3.40	0.25
40.33	0.78	45.12	30.00	24.88	8.10	4.70	0.35
35.18	0.75	47.12	44.00	8.88	2.50	1.40	0.15

Cuadro A.1: Propiedades obtenidas en la muestra.

C/N	$\%Ret.Fosfatos$	$\%Al\,Ext.$	$\%Fe\,Ext.$	$Al+1/2Fe\,extr.(\%)$	$pH(NaF\,1N)$	$pH(2:1\,H2O:S)$
13.00	46.00	1.17	0.35	1.34	9.20	6.00
11.00	50.40	1.51	0.53	1.77	9.40	6.90
15.00	51.50	1.93	0.60	2.23	9.40	5.10
12.00	63.20	2.05	0.60	2.35	10.50	7.00
16.00	60.90	1.58	0.44	1.80	10.30	6.50
13.00	57.00	1.32	0.32	1.48	10.80	7.00
14.00	50.50	1.42	0.28	1.56	9.80	5.30
13.00	44.00	2.13	0.25	2.26	10.00	5.90
15.00	61.80	1.05	0.37	1.23	9.50	5.00
13.00	65.30	1.23	0.34	1.40	9.80	6.00
15.00	60.20	1.52	0.30	1.67	9.00	6.00
17.00	58.00	1.41	0.22	1.52	9.20	6.00
17.00	34.90	1.93	0.35	2.10	10.20	5.90
12.00	56.70	2.35	0.66	2.68	11.20	6.10
15.00	72.30	1.88	0.65	2.20	10.50	5.00
13.00	63.40	1.89	0.21	1.99	10.00	6.00
13.00	64.70	1.80	0.49	2.04	10.60	5.40
14.00	70.40	2.06	0.27	2.20	10.80	5.60
14.00	70.00	2.03	0.34	2.20	10.80	5.30
18.00	74.00	2.96	0.66	3.30	9.50	6.20
13.00	86.90	3.42	1.80	4.32	11.50	4.80
16.00	70.50	2.09	0.91	2.54	10.8	4.40
16.00	80.60	3.09	1.25	3.72	10.30	5.10
13.00	87.20	3.06	1.96	4.04	10.80	5.40
13.00	76.20	2.98	0.94	3.45	10.70	6.00
13.00	80.90	3.42	1.00	3.92	10.70	6.30
9.69	68.10	1.08	1.02	1.60	48.75	5.40
10.00	76.50	1.01	0.45	1.23	36.75	5.30
13.20	82.20	3.72	0.50	3.97	35.50	5.10
11.88	85.90	5.87	0.90	6.32	58.75	6.30
7.38	87.90	3.58	0.67	3.92	49.50	5.00
4.35	72.5	3.82	0.35	4.00	67.75	5.00
7.38	65.90	1.71	0.25	1.83	46.50	6.80
4.35	67.00	0.86	0.15	0.93	40.25	7.40
14.50	63.00	2.28	0.52	2.54	53.25	5.30
9.86	72.40	2.73	0.40	2.93	54.50	5.90
13.30	76.30	2.15	0.36	2.33	29.50	6.50
8.89	78.10	1.82	0.83	2.24	44.00	6.00
14.72	80.70	4.11	0.55	4.38	23.00	5.50
13.60	79.60	3.85	0.55	4.125	43.25	5.70
13.42	81.20	3.74	1.60	4.54	34.50	5.80
9.33	83.60	3.13	1.90	5.08	41.25	6.10

Cuadro A.2: Propiedades obtenidas en la muestra.

$pH(2:1\,KCl:S)$	$Delta\,pH$	$CIC(cmol(+)/Kg\,S)$	$\%V$	Ca	Mg	Na	K
5.20	-0.80	19.40	29.40	3.30	1.60	0.20	0.60
6.20	-0.70	15.10	40.40	2.50	1.40	0.70	1.50
4.40	-0.70	13.55	36.90	2.90	1.10	0.40	0.60
6.50	-0.50	14.85	36.40	3.30	1.40	0.30	0.40
5.80	-0.70	19.25	54.50	5.40	2.80	0.90	1.40
6.20	-0.80	13.25	45.28	2.80	10.00	0.80	1.40
4.50	-0.80	13.20	47.00	3.00	2.30	0.40	0.50
5.20	-0.70	12.80	36.70	2.40	1.50	0.60	0.20
4.40	-0.60	16.40	28.40	3.40	0.60	0.40	0.50
5.30	-0.70	15.00	29.80	4.00	1.00	0.40	0.50
5.10	-0.90	21.40	31.30	3.40	0.90	1.40	1.00
5.20	-0.80	14.30	29.30	2.00	0.60	0.90	0.60
5.30	-0.60	33.00	10.60	2.30	0.50	0.40	0.30
5.40	-0.70	30.70	10.20	1.50	1.00	0.50	0.10
4.20	-0.80	21.30	30.00	3.00	2.00	0.80	0.60
5.20	-0.80	25.00	32.80	5.00	2.00	0.60	1.60
4.70	-0.70	19.40	19.80	2.00	1.00	0.35	0.50
4.90	-0.70	14.30	22.40	1.50	1.00	0.40	0.30
4.80	-0.50	34.60	14.20	3.50	1.00	0.20	0.10
5.70	-0.50	29.00	6.70	1.30	0.20	0.30	0.10
4.00	-0.80	35.80	27.40	5.60	1.40	2.00	0.80
5.40	-1.00	10.50	40.90	2.00	0.80	1.20	0.30
4.50	-0.60	16.80	50.60	4.10	1.30	2.80	0.30
4.70	-0.70	13.80	38.30	4.30	0.70	0.20	0.20
5.40	-0.60	34.20	55.70	9.70	3.10	2.20	4.00
5.60	-0.70	33.70	30.10	6.50	1.20	0.70	1.70
4.40	-1.00	30.50	10.59	1.06	0.21	1.60	0.30
4.40	-0.90	31.80	8.89	1.00	0.06	1.40	0.40
4.80	-0.30	41.10	17.00	1.90	0.20	1.67	0.30
5.00	-1.30	53.30	8.33	0.28	0.28	1.76	0.30
4.10	-0.90	29.10	4.73	0.84	0.21	0.62	0.08
4.40	-0.60	38.50	5.11	0.42	0.22	1.10	0.23
5.30	-1.50	11.05	43.50	2.70	1.30	0.40	0.31
5.50	-1.90	10.35	47.50	1.80	1.70	0.56	0.30
4.40	-0.90	29.00	32.70	5.00	3.00	0.60	0.60
4.70	-1.20	18.50	48.30	5.60	1.50	0.70	0.45
4.70	-1.80	48.40	9.30	3.00	1.20	0.18	0.14
4.30	-1.70	53.20	8.80	3.00	1.00	0.31	0.38
5.10	-0.40	44.20	12.00	4.00	1.00	0.20	0.13
5.60	-0.10	39.90	10.20	2.10	1.60	0.20	0.16
4.80	-1.00	27.30	13.47	2.00	1.00	0.37	0.31
5.20	-0.90	29.00	10.30	1.60	0.70	0.37	0.31

Cuadro A.3: Propiedades obtenidas en la muestra.

Apéndice B

Coeficientes Normalizados de Regresión

En general, es difícil comparar en forma directa coeficientes de regresión, por que la magnitud de $\hat{\beta}_j$ refleja las unidades de medida del regresor x_j. Por esta razón a veces ayuda trabajar con regresores y variables de respuesta escalados, que produzcan *coeficientes de regresión adimensionales*. A esos coeficientes adimencionales se les suele llamar *coeficientes estandarizados de regresión*. A continuación se describe como se calculan, aplicando dos técnicas frecuentes de escalamiento [10].

B.1. Escalamiento Normal Unitario

El primer método emplea el *escalamiento normal unitario* para los regresores y la variable de respuesta. Esto es;

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{S_j},\ i = 1, 2, \ldots, n,\ j = 1, 2, \cdots, k;$$

y

$$y_i^* = \frac{y_i - \overline{y}_i}{S_y},\ i = 1, 2, \ldots, n;$$

en donde,

$$S_j^2 = \frac{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2}{n-1},$$

es la varianza muestral del regresor x_j, y

$$S_y^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1},$$

es la varianza muestral de la respuesta. Nótese la semejanza con la estandarización de una variable aleatoria normal. Todos los regresores escalados y la respuesta escalada tienen media muestral igual a cero y varianza muestral igual a 1.

Con estas nuevas variables, el modelo de regresión se transforma en

$$y_i^* = b_1 z_{i1} + b_2 z_{i2} + \cdots + b_k z_{ik} + \varepsilon_i,\ i = 1, 2, \cdots, n.$$

Al centrar las variables regresoras y de respuesta restando $\overline{x}_j$ y $\overline{y}$, se elimina la ordenada al origen del modelo (en realidad, el estimado de b_0 por mínimos cuadrados es $\hat{b} = \overline{y}^* = 0$). El estimador de **b** por mínimos cuadrados es

$$\hat{b} = (Z'Z)^{-1} Z'y^*. \tag{B.1}$$

B.2. Escalamiento de Longitud Unitaria

El segundo escalamiento que se usa con frecuencia es el *escalamiento de longitud unitaria*,

$$w_{ij} = \frac{x_{ij} - \overline{x}_j}{S_{jj}^{1/2}},\ i = 1, 2, \ldots, n,\ j = 1, 2, \cdots, k;$$

y

$$y_i^\star = \frac{y_i - \overline{y}_i}{SS_T^{1/2}},\ i = 1, 2, \ldots, n;$$

en donde

$$S_{jj} = \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)^2,$$

es la suma corregida, para el regresor x_j. En este escalamiento, cada nuevo regresor w_j tiene promedio $\overline{w}_j = 0$ y longitud $\sqrt{\sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)^2} = 1$. En función de esas variables, el modelo de regresión es

$$y_i^\star = b_1 w_{i1} + b_2 w_{i2} + \cdots + b_k w_{ik} + \varepsilon_i,\ i = 1, 2, \cdots, n.$$

El vector de los coeficientes de regresión por mínimos cuadrados es

$$\hat{b} = (W'W)^{-1} W'y^\star. \tag{B.2}$$

En el escalamiento de longitud unitaria, la matriz $W'W$ tiene la forma de una **matriz de correlación**, es decir;

$$W'W = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{12} & 1 & r_{23} & \cdots & r_{2k} \\ r_{13} & r_{23} & 1 & \cdots & r_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{1k} & r_{2k} & r_{3k} & \cdots & 1 \end{pmatrix},$$

en donde

$$r_{ij} = \frac{\sum_{u=1}^{n} (x_{ui} - \overline{x}_i)(x_{uj} - \overline{x}_j)}{(S_{ii}S_{jj})^{1/2}} = \frac{S_{ij}}{(S_{ii}S_{jj})^{1/2}},$$

es la correlación simple entre el regresor x_i y el x_j. De igual modo,

$$W'y^{\star} = \begin{pmatrix} r_{1y} \\ r_{2y} \\ r_{3y} \\ \vdots \\ r_{ky} \end{pmatrix},$$

en la que

$$r_{jy} = \frac{\sum_{u=1}^{n} (x_{uj} - \overline{x}_j)(x_u - \overline{y})}{(S_{jj}SS_T)^{1/2}} = \frac{S_{jy}}{(S_{jj}SS_T)^{1/2}},$$

es la correlación simple entre el regresor x_j y la respuesta y. Si se usa escalamiento normal unitario, la matriz $Z'Z$ se relaciona en forma importante con la matriz $W'W$; de hecho,

$$Z'Z = (n-1)W'W.$$

En consecuencia, los coeficientes de regresión estimados en las ecuaciones B.1 y B.2 son idénticos, por lo que no importa que escalamiento se use; ambos producen el mismo conjunto de coeficientes adimencionales de regresión $\hat{\boldsymbol{b}}$.

A los coeficientes de regresión $\hat{b}$ se les suele llamar *coeficientes estandarizados de regresión*. La relación entre los coeficientes originales y los estandarizados es

$$\hat{\beta}_j = \hat{b}_j \left(\frac{SS_T}{S_{jj}}\right)^{1/2}, \; j = 1, 2, \cdots, k,$$

y

$$\hat{\beta}_0 = \overline{y} - \sum_{i=1}^{n} \hat{\beta}_j \overline{x}_j.$$

Apéndice C

Distribuciones No Centrales

Supondremos que Y es una variable aleatoria que sigue una distribución normal con media μ y varianza σ^2, lo cual se denota como $Y \sim N(\mu, \sigma^2)$. [12].

C.1. Distribución Chi Cuadrada no Central

Antes de describir la distribución Chi Cuadrada no central, primero recordemos la Chi Cuadrada central. Sean $z_1, z_2, \cdots, z_n$ una muestra de variables aleatorias con distribución normal estandar $N(0,1)$. Entonces los $z's$ son independientes (por ser una muestra aleatoria), así :

$$\sum_{i=1}^{n} z_i^2 \sim \chi^2(n). \tag{C.1}$$

En resumen tenemos que, la suma de n variables aleatorias normal estandar al cuadrado nos da como resultado una variable Chi Cuadrada (central) con n grados de libertad.

Ahora supongamos que $y_1, y_2, \cdots, y_n$ son independientes y con distribución $N(\mu_i, 1)$, tales que el vector aleatorio Y es $N_n(\boldsymbol{\mu}, \boldsymbol{I})$, donde $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)'$ e $\boldsymbol{I}$ es la matriz identidad. En este caso, $\sum_{i=1}^{n} = y_i^2 = Y'Y$ no tiene una distribución Chi Cuadrada, debido a que cada y_i no está estandarizada, pero $\sum_{i=1}^{n} = (y_i - \mu_i)^2 = (Y - \boldsymbol{\mu})'(Y - \boldsymbol{\mu})$ es $\chi^2(n)$ ya que $y_i - \mu_i$ tiene distribución $N(0,1)$.

La densidad de $\nu = \sum_{i=1}^{n} = y_i^2 = Y'Y$, donde las $y's$ son independientes con distribución $N(\mu_i, 1)$ es llamada distribución *Chi Cuadrada No Central* y es denotada por $\chi^2(n, \lambda)$. El parámetro de no centralidad λ esta definido por

$$\lambda = \frac{1}{2} \sum_{i=1}^{n} \mu_i^2 = \frac{1}{2} \boldsymbol{\mu}' \boldsymbol{\mu}.$$

C.2. Distribución F no central

Recordemos la distribución F (centrada) [12]. Sean υ y ν dos distribuciones independientes $\chi^2(p)$ y $\chi^2(q)$ respectivamente, entonces

$$\omega = \frac{\upsilon/p}{\nu/q} \sim F(p, q); \tag{C.2}$$

es la distribución F (central) con p y q grados de libertad.

Ahora supongamos que υ tiene distribución no central Chi Cuadrada, $\chi^2(p, \lambda)$, mientras que ν tiene distribución no central Chi Cuadrada $\chi^2(q)$, con υ y ν independientes. Entonces

$$z = \frac{\upsilon/p}{\nu/q} \sim F(p, q, \lambda); \tag{C.3}$$

es la distribución F no central con parámetro de no centralidad λ, donde λ es igual al parámetro de no centralidad en la distribución Chi Cuadrada no central.

Cuando una estadística F se utiliza para probar la hipótesis H_0, la distribución será central si la hipótesis (nula) es verdadera y no central si la hipótesis es falsa. Así, la distribución F no central, se puede utilizar con frecuencia para evaluar la potencia de una prueba F. La potencia de una prueba es la probabilidad de rechazar H_0 para un valor λ dado. Si F_α es el punto α de porcentaje superior de la distribución F (central), entonces la potencia, $P(p, q, \alpha, \lambda)$, se puede definir como

$$P(p, q, \alpha, \lambda) = P(z > F_\alpha).$$

Apéndice D

Análisis Estadístico con R

Es sabido que existen en el mercado distintos paquetes de software estadístico propietario, como *SPSS, Statgraphics, Minitab, Statistica, SAS, S-Plus, etc*, que cubren todas las necesidades de cualquier usuario de técnicas estadísticas, básicas o avanzadas. Frente a estas opciones, en los últimos años *R* ha surgido con fuerza como alternativa de software libre en muy distintos ambientes docentes y de investigación.

R es un lenguaje de programación especialmente indicado para el análisis estadístico. A diferencia de la mayoría de los programas que solemos utilizar en nuestros ordenadores, que tienen interfaces tipo ventana, R es manejado a través de una consola en la que se introduce código propio de su lenguaje para obtener los resultados deseados [4].

El código de *R* está disponible como software libre bajo las condiciones de la licencia GNU-GPL, y puede ser instalado tanto en sistemas operativos tipo Windows como en Linux o MacOS X.

La página principal desde la que se puede acceder tanto a los archivos necesarios para su instalación como al resto de recursos del proyecto R es

`http://www.r-project.org.`

Despues de instalar *R* lo primero que nos aparece es una ventana, también llamada *consola*, donde podemos manejar R mediante la introducción de código. Sin embargo, esta no es la manera más eficiente de trabajar en R, para ello debemos acceder a un documento en blanco del editor, llamado *script*. La utilidad de un script o guión de trabajo radica en que podemos modificar nuestras líneas de código con comodidad y guardarlas para el futuro.

En este caso el código que se utilizó fue guardado en un script del editor, en el cual se fueron introduciendo todos los pasos de nuestro trabajo para hacer el análisis estadístico requerido [4].

`prop<-read.table(file="clipboard", head=T)`: Con esta indicación se pueden leer los datos desde Exel, solo seleccionando el conjunto deseado.

`data.frame(prop)`: Esta línea crea un macro(conjunto) de datos y los muestra en pantalla.

La sintaxis básica de la función `lm()`, que nos hace un análisis básico de regresión lineal es la siguiente:

`lm(formula, data, subset)`; en este caso

```
g <- lm(COS ~ ., data = prop)
```

`summary(g)`: Esta indicación nos muestra un resumen de los resultados obtenidos con la función `lm()`.

`plot(fitted(g),residuals(g),xlab="Fitted",ylab="Residuals")`: Esta línea nos muestra la gráfica de residuales.

`abline(h=0)`: Crea una línea horizontal.

`qqnorm(residuals(g),ylab="Residuals")`: Esta línea nos muestra la gráfica normal de residuales.

`qqline(residuals(g))`: Esta línea une el primer y tercer cuantil.

Ahora para detectar los problemas de multicolinealidad se utilizaron los siguientes códigos:

`x<-model.matrix(g)[,-1]`: Crea la matriz `x` que corresponde a la matriz de las variables de respuesta.

`print(x)`: Imprime en pantalla la matriz `x`.

`e<-eigen(t(x) %*% x)`: Nos hace un análisis del eigensistema de la matriz `x'x`.

`e$val`: Muestra los eigenvalores correspondientes a `x'x`.

`sqrt(e$val[1]/e$val)`: Calcula todos los números de condición asociados a `x'x`.

Posteriormente se procede a hacer el cálculo de los *VIF´s* y para ello primero se debe cargar el paquete "faraway".

`vif(x)`: Muestra los valores VIF.

Para llevar a cabo el análisis de componentes principales, se hace uso de la matriz de correlaciones.

`prop.prcomp<-princomp(prop[,-1],cor=T)`: Calculo los componentes principales basados en la matriz de correlaciones.

`print(summary(prop.prcomp,loadings=T))`: Muestra un resumen de los resultados obtenidos, así como los eigenvalores para elegir el número de componentes a utilizar.

`S=cor(prop[,-1])`: Crea la matriz de correlación correspondiente a las variables de respuesta.

`eigen(S)`: Muestra los eigenvalores de la matriz de correlación, mismos que serán las componentes principales.

`screeplot(prop.prcomp,type="lines")`: Grafica los eigenvalores asociados a la matriz de correlación para elegir el número de componentes.

Después se debe hacer el nuevo análisis con las componentes principales elegidas.

`comp <-data.frame(prop.prcomp$scores,prop)`: Se crea un nuevo conjunto de datos, a saber las componentes principales.

`g6 <- lm(COS ~ Comp.1+Comp.2+Comp.3+Comp.4+Comp.5+Comp.6, data=comp)`: Hace un nuevo análisis de regresión lineal.

`summary(g6)`: Muestra los resultados del nuevo modelo.

Cabe señalar que para la metodología PLS se utilizo como alternativa el paquete estadístico MINITAB 14 ya que con el paquete R no se obtenían los resultados requeridos para hacer un análisis completo. Así mismo se utilizó MINITAB para obtener cálculos como el $PRESS$ y el R_{pred}.

Bibliografía

[1] Castillo M., Linares G., Valera M. A., García N. E., Acevedo O. A. *Modelación de la materia orgánica en suelos volcánicos de la región de Teziutlán, Puebla, México.* En: Revista Latinoamericana de Recursos Naturales, 5 (2): 148-154, 2009.

[2] Castillo M., Linares G., Valera M. A., Torres E., García N. E., Acevedo O. A., Jiménez R. *Estimación de los regímenes de humedad y temperatura de suelos volcánicos de la región de Teziutlán, Puebla, México y su relación con los contenidos de carbono orgánico del suelo.* Extenso en el Congreso Internacional de Ciencias Ambientales. Chetumal, Quintana Roo, México, 2010.

[3] Chin, W.W. *How to write up and report PLS analyses.* In Esposito, V., et al. (eds.), Handbook of partial Least Squares (pp. 655-688). New York: Springer Verlag, 2010.

[4] Faraway Julian J. *Linear Models with R.* Chapman and Hall/CRC, 2005.

[5] Garson G. D. *Partial Least Squares.* Asheboro, NC: Staristical Associates Publishers, 2012.

[6] Linares Gladys, Mederos María V. *Sistema de Calibración Multivariada.* Departamento de Matemática Aplicada, Universidad de La Habana. En: Revista Investigación Operacional, Vol. 21, No. 1, pp. 46-51, 2000.

[7] Linares Gladys, Saldaña Adrian, Rivera Marcela, Cervantes Hortensia J. *Comparación de Métodos de Regresión en la Predicción de Dióxido de Carbono.* Revista de Ciencias Matemáticas. Vol.25, 2010-2012.

[8] Linares Gladys, Sandoval María de Lourdes, Rivera Marcela, Reyes Hortensia J. *Comparación de Métodos de Regresión en la Producción De la Materia Orgánica en Suelos.* Aportaciones y Aplicaciones de la Probabilidad y la Estadística. Volumen 2. BUAP, Dirección General de Fomento Editorial., pp. 61-75, 2008.

[9] López Pineda Gabriela, Linares Fleites Gladys, Reyes Cervantes Hortensia J. *Regresión de Componentes Principales en Estudios de Carbono Orgánico en Suelos de Teziutlán, Puebla.* Memorias en extenso en: 5 Semana Internacional de la Estadística y la Probalilidad. FCFM-BUAP, 2012.

[10] Montgomery D.C., Peck E.A., Vining G.G. *Introducción al Análisis de Regresión Lineal.* México: Compañía Editorial Continental, 2004.

[11] Rao C. Radhakrishna, Toutenburg Helge. *Linear Models: Least Squares and Alternatives.* Second Edition. Springer-Verlag, New York, 1999.

[12] Rencher, Alvin C., *Linear models in statistics.* John Wiley and Sons, Inc., Hoboken, New Jersey. 2000.

[13] Rosipal Roman, Krämer Nikole. *Overview and Recent Advances in Partial Least Squares.* In SLSFS 2005 LNCS3940, (eds. Saunders et al.) pp. 34-51 Springer-Verlag Berlin Heidelberg, 2006.

[14] Vega-Vilca José Carlos, Guzmán Josué. *Regresión PLS y PCA como solución al problema de multicolinealidad en regresión múltiple.* En: Revista Matemática: Teoría y Aplicaciones 18(1): pp. 9-20 CIMPA-UCR ISSN: 1409-2433, 2011.

[15] Weisberg S. *Applied Linear Regression.* Segunda edición, Wiley, Nueva York, 1985.

[16] Yeniay Özgür, Göktas Atilla. *A Comparison Of Partial Least Squares Regression With Other Prediction Methods.* Hacettepe Journal of Mathematics and Statistics, Volume 31, pp. 99-111, 2002.