Sección de Matemáticas
Universidad de La Laguna

Eduardo Hernández Córdoba

Selección del mejor conjunto de Regresión.

The Best Linear Regression Model.

Trabajo Fin de Grado
Grado en Matemáticas
La Laguna, julio de 2019

Dirigido por
María Mercedes Suárez Rancel
Coromoto León Hernández

María Mercedes Suárez Rancel
Departamento de Matemáticas, Estadística e Investigación Operativa
Universidad de La Laguna
38200 La Laguna, Tenerife

Coromoto León Hernández
Departamento de Ingeniería Informática y de Sistemas.
Universidad de La Laguna
38200 La Laguna, Tenerife

Agradecimientos

A mis tutoras María Mercedes Suárez Rancel y Coromoto León Hernández por su dedicación y constancia en la elaboración de este TFG.

A mis padres y mi hermana por ser mis pilares fundamentales de inspiración y gran apoyo en todos los proyectos que me planteo.

A mis amigos de toda la vida por regalarme momentos únicos.

A mis amistades creadas en los años de estudio del Grado de Matemáticas que han sido clave para superar esta etapa.

En general a toda persona ya sea de la comunidad de la Universidad de La Laguna o del ámbito social, que me ha aportado algo contribuyendo así a conseguir realizar este trabajo y obtener el titulo de graduado en matemáticas.

Eduardo Hernández Córdoba
La Laguna, 7 de julio de 2019

Resumen · Abstract

Resumen

Este Trabajo de Fin de Grado realiza un estudio de la regresión lineal múltiple implementada en el Software libre Rcommander y Rstudio con la finalidad de encontrar el mejor conjunto de regresión. Primeramente se estudia la regresión lineal múltiple, la estimación de los parámetros, y se estudian diferentes criterios que se usan para la comparación de modelos. Además, se estudia los algoritmos de los métodos de selección de variables y se da una ampliación de los criterios de información de Akaike (AIC) y criterio de información Bayesiano (BIC). A continuación, se desarrolla el análisis de sensibilidad de la regresión lineal múltiple, donde se estudia la validación del modelo e hipótesis asociadas y el efecto de las observaciones anómalas. Por ultimo, se desarrolla la implementación en Rcommander y Rstudio.

Palabras clave: *Regresión Lineal Múltiple – Análisis de sensibilidad – Rcommander – Rstudio – AIC y BIC – Modelos*

Abstract

This work performs a multiple linear regression study implemented in the free Software Rcommander and Rstudio in order to find the best linear regression model. Firstly multiple linear regression, the estimation of the parameters and the different criteria used to compare models are studied. In addition, the algorithms of the variables selection methods are studied and an extension of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) is made. Afterwards, the sensitivity analysis of the multiple linear regression is developed, where the validation of the model and the associated hypotheses and the effect of the anomalous observations are studied. Finally, the implementation is carried out in Rcommander and Rstudio.

Keywords: *Multiple Linear Regression – Sensitivity Analysis – Rcommander – Rstudio – AIC y BIC – Models*

Contenido

Introducción

La regresión lineal múltiple es una herramienta utilizada en la sociedad para predecir y describir ciertos sucesos, acontecimientos, fenómenos, etc. Como por ejemplo; predecir la altura de un cadáver en función las medidas de su tibia y su fémur. El objetivo de la regresión es el de relacionar un variable dependiente con un conjunto de variables independientes, todas ellas continuas.

Dichas relaciones pueden verse como algo lejano a la realidad, pero en muchas ocasiones la modelización de esta función puede acercarse tanto a la realidad que la diferencia entre lo que sucede en la realidad y el modelo es despreciable a la hora de predecir. Cabe destacar que, si se logra encontrar relaciones que consiguen predecir fenómenos que suceden en la realidad, estamos ante algo muy práctico y que nos permite simplificar y abordar problemas mayores.

El modelo de regresión lineal múltiple lleva consigo una serie de hipótesis que debe satisfacer a la hora de dar un modelo fiable. A lo largo de este trabajo, desarrollaremos un estudio en profundidad de dichas hipótesis y abordaremos situaciones que pueden ocasionar problemas a la hora de modelar nuestra ecuación de regresión lineal múltiple.

El hecho de encontrar una relación entre la variable dependiente y el conjunto de las variables independiente cumpliendo una serie de hipótesis y restricciones, no siempre es fácil de satisfacer. Seleccionar el mejor modelo de regresión, bajo unas hipótesis, y que además con el mejor conjunto de variables el cual aproxime mejor los datos a la realidad, no es algo sencillo. Para ello existen diversas técnicas y criterios que te permiten determinar cual es el mejor modelo de regresión.

La condición de linealidad de linealidad del modelo es algo positivo, ya que te permite realizar transformaciones en las variables y esto a su vez permite englobar modelos que a priori no son lineales, transformarlos y convertirlos en lineales. Además, en lo modelos lineales, los cálculos son mucho mas sencillos que otros modelos de regresión.

En la actualidad para llevar a cabo las pruebas que corroboran las hipótesis se utilizan software específicos en estadística y análisis multivariante. En este caso utilizaremos un software de libre distribución basado en el lenguaje R.

1

Capítulo 1: Análisis estadístico de la regresión lineal.

En este primer capítulo estudiaremos en profundidad el modelo de la regresión lineal múltiple, la cual nos permitirá predecir una variable dependiente en función de un conjunto de variables independientes. Veremos como calcular los estimadores de la ecuación y estudiaremos criterios que nos permiten comparar modelos.

1.1. Modelo teórico de la regresión lineal.

La regresión trata de encontrar una relación entre un conjunto de variables independientes $(X_0, X_1, X_2, ..., X_{k-1},)$ y una variable dependiente (Y). Existen diversas técnicas de regresión en función del tipo de variable, en este caso nosotros estudiaremos las lineales, las cuales se desarrollan por variables continuas. La regresión lineal es la mas importante entre las regresiones, ya que nos permite hacer transformaciones en la variables y además sus cálculos son muy sencillos. El ejemplo más básico es la regresión lineal simple, la cual relaciona la variable dependiente con una sola variable independiente. El modelo que describe la regresión lineal simple es el siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

por ejemplo:

$$PESO = \beta_0 + \beta_1 ALTURA + \varepsilon$$

Este tipo de relación es muy básica a la hora de explicar sucesos que ocurren en la vida, ya que la simplicidad de su ecuación no permite aproximarnos a la realidad de los sucesos. Es por eso que los modelos que mejor se aproximan a la realidad (o "modelo real") son los modelos de regresión lineal múltiples, donde estos relacionan la variable dependiente con mas de una variable independiente.

Por tanto en el modelo se introducirían mas variables que pueden describir un mismo acontecimiento.

La ecuación de la regresión lineal múltiple se define de la siguiente manera:

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{k-1} X_{k-1} + \varepsilon \tag{1.1}$$

Donde:

- Y ≡ Variable dependiente.
- β_i ≡ Coeficientes de regresión. ∀ i=0,...,k-1
- X_i ≡ Variables independientes. ∀ i=0,...,k-1
- ε ≡ Error experimental.

El modelo de regresión lineal múltiple debe de satisfacer una serie de hipótesis. Estas hacen que encotnrar el modelo lineal de regresión sea complejo,ya que si no se satisfacen no se puede aproximar el modelo estimado, al modelo real. Las hipótesis asociadas al modelo son:

1. E[ε]=0. Esta hipótesis no revoca en una nueva restricción, ya que si no se verifica bastaría con añadir un término constante al modelo.
2. La varianza del error es siempre constante y no depende de X (homocedástica).
3. Los errores son independientes entre si, es decir, (cov(ε_i,ε_j)=$\delta_{i,j}\sigma^2$) (Autocorrelación).
4. El error tiene una distribución normal (Normalidad).
5. Multicolinealidad.

Estas hipótesis pueden expresarse también con respecto a la distribución de la variable dependiente de la siguiente manera:

1. E[ε]=0
2. La varianza de Y es constante.
3. Las observaciones Y son independientes entre sí.
4. La distribución de Y es normal.
5. Multicolinealidad.

1.2. El modelo muestral en la regresión lineal.

Al tomarse una muestra de n observaciones de n valores de las variables independientes $X_0, X_1, X_2, ..., X_{k-1},$, y observando los valores que pertenecen a $Y_1, Y_2, ..., Y_n$, entonces se tiene que:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{k-1} X_{ik-1} + \varepsilon_i \qquad \forall i = 1, ..., n$$

donde X_{ij} es el í-ésimo valor de la variable X_j. Estas ecuaciones se pueden expresar de forma mas compacta como:

$$Y = X\beta + \varepsilon$$

teniendo que:

- Y matriz de orden (n x 1)
- X matriz de orden (n x k)
- β es el vector de coeficientes desconocidos de orden (k x 1), y
- ε el vector aleatorio de errores.

El modelo matricial seria:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{10} & x_{11} & \vdots & x_{1k-1} \\ x_{20} & x_{21} & \vdots & x_{2k-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \vdots & x_{nk-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Con $x_{10} = x_{20} = ... = x_{n0} = 1$

1.3. Estimación de los parámetros β_i y de σ^2.

Antes de empezar a hacer uso del modelo, debemos realizar previamente la estimación de los coeficientes de regresión y de σ^2.

· Estimación de β y σ^2 por el método de mínimos cuadrados.

Supongamos que el modelo (1.1) verifica las hipótesis salvo la normalidad. El método que se utiliza en este caso para la estimación de los parámetros es el *" criterio de mínimos cuadrados"*. El criterio está basado en minimizar las perturbaciones del modelo, es decir:

$$min\ \varepsilon^t\ \varepsilon = min \sum_{i=1}^{n} \varepsilon_i^2 = min\ (Y - X\beta)^t (Y - X\beta) \tag{1.2}$$

La estimación de los valores de β que minimizan $\varepsilon^t\varepsilon$ viene dado de resolver la siguiente expresión:

$$\frac{\partial}{\partial \beta} \varepsilon^t \varepsilon = 0$$

se sigue que: $-2X^tY + 2X^tX\hat{\beta} = 0 \Rightarrow X^tY = X^tX\hat{\beta}$. Por tanto se tiene que, como $rg((X^tX)^{-1}) = rg(X) = k$, entonces:

$$\hat{\beta} = (X^tX)^{-1}X^tY$$

El cual se demuestra que es un mínimo de (1.2).

El inconveniente de minimizar la suma de cuadrados de $\varepsilon^t\varepsilon$, es que no proporciona un estimador de σ^2, por lo que se debe proponer el siguiente estimador:

$$S^2 = \frac{(Y - X\hat{\beta})^t(Y - X\hat{\beta})}{n-k} = \frac{Y^t(I - X(X^tX)^{-1}X^t)Y}{n-k}$$

Previamente a estudiar las propiedades de los estimadores MCO en el modelo de regresión lineal múltiple, definimos un conjunto de supuestos muy simples donde los estimadores, bajo estos supuestos, poseen muy buenas propiedades. Definimos el conjunto de supuestos:

1. La relación entre la variable dependiente, las variables independientes y el error es lineal.
2. La matriz X, no contiene errores de medición.
3. La matriz X, tiene rango igual a k, $rg(X) = k$ (Recordar que la dimensión de X es de $n \times k$). Para este supuesto además se debe de tener que:
   - El número de observaciones, n, debe ser igual o mayor que el número de variables.
   - Las variables predictoras deben ser independientes. Así evitaremos el problema de multicolinealidad.
4. Los parámetros β_i son constantes.
5. La media de los residuos son 0, es decir, $E(\varepsilon) = 0$.
6. Los residuos tienen una varianza constante (supuesto de homocedasticidad), $var(\varepsilon_i) = \sigma^2 \ \forall i = 1, 2, ..., n$,
7. Los residuos no están correlacionados entre si (supuesto de no autocorrelación), $E(\varepsilon_i\varepsilon_j) = 0 \ i \neq j$.

Una vez visto los supuestos, veamos ahora las propiedades que tienen los estimadores bajo esto supuestos:

- El estimador $\hat{\beta}$, es linealmente insesgado. Por 1) tenemos que $\hat{\beta}$ se puede expresar en función de ε, es decir:

  $$\hat{\beta} = [X^tX]^{-1}X^tY = [X^tX]^{-1}X^t[X\beta + \varepsilon] = \beta + [X^tX]^{-1}X^t\varepsilon$$

  Tomando las esperanza y por el supuesto 6 tenemos:

  $$E[\hat{\beta}] = \beta + [X^tX]^{-1}X^tE[\varepsilon] = \beta$$

  Por tanto se tiene que $\hat{\beta}$ es un estimador insesgado.

- Aceptados los supuestos anteriores, definimos la varianza del estimador mímimo cuadrado:

$$var(\hat{\beta}) = E[\hat{\beta} - E[\hat{\beta}]][\hat{\beta} - E[\hat{\beta}]]^t = E[\hat{\beta} - \beta][\hat{\beta} - \beta]^t$$

$$por\ tanto\ se\ tiene\ que:$$

$$var(\hat{\beta}) = E[[X^tX]^{-1}X^t\varepsilon\varepsilon^tX[X^tX]^{-1}] = [X^tX]^{-1}X^tE[\varepsilon\varepsilon^t]X[X^tX]^{-1}$$

$$por\ 7)\ y\ 8)$$

$$var(\hat{\beta}) = \sigma^2[X^tX]^{-1}$$

  Entonces se tiene que $var(\hat{\beta}) = \sigma^2[X^tX]^{-1}$ es la matriz de covarianzas del vector $\hat{\beta}$.
- Bajo los supuestos 1) a 6), el estimador de mínimo cuadrado, $\hat{\beta}$ es consistente. Hagamos la prueba que verifica esta hipótesis.
  Expresamos $\hat{\beta}$ de la siguiente manera:

$$\hat{\beta} = \beta + \left(\frac{1}{n}X^tX\right)^{-1}\left(\frac{1}{n}X^t\varepsilon\right)$$

  Definimos el limite siguiente:

$$\lim_{n\to\infty}\frac{1}{n}X^tX = Q$$

  Como X es fija, de acuerdo con 2), entonces se tiene que $Q = \frac{1}{n}X^tX$. De acuerdo con 3) entonces existe inversa de Q, Q^{-1}. Por tanto, se tiene que:

$$\lim_{n\to\infty}\hat{\beta} = \beta + Q^{-1}\lim_{n\to\infty}\frac{1}{n}X^t\varepsilon$$

  se tiene que: $\frac{1}{n}X^t\varepsilon = \frac{1}{n}\sum_{i=1}^n x_i\varepsilon_i = \overline{x_i\varepsilon_i}$ donde x_i es el vector de la columna correspondiente a la observación i-ésima. Veamos ahora la esperanza y la varianza de $\overline{x_i\varepsilon_i}$.

$$E[\overline{x_i\varepsilon_i}] = \frac{1}{n}\sum_{i=1}^n E[x_i\varepsilon_i] = \frac{1}{n}\sum_{i=1}^n x_iE[\varepsilon_i] = \frac{1}{n}X^tE[\varepsilon] = 0$$

$$var(\overline{x_i\varepsilon_i}) = E[\overline{x_i\varepsilon_i(x_i\varepsilon_i)^t}] = \frac{1}{n^2}X^tE[\varepsilon\varepsilon^t]X = \frac{\sigma^2}{n}Q$$

$$Se\ observa\ que:\ \lim_{n\to\infty}var(\overline{x_i\varepsilon_i}) = \lim_{n\to\infty}\frac{\sigma^2}{n}Q = 0$$

  dado que la esperanza de $\overline{x_i\varepsilon_i}$ es 0 y la varianza converge a 0, $\overline{x_i\varepsilon_i}$ converge en media cuadrática a 0. Entonces $\lim_{n\to\infty}\overline{x_i\varepsilon_i} = 0$. Por tanto:

$$\lim_{n\to\infty} \hat{\beta} = \beta + Q^{-1} \lim_{n\to\infty} \frac{1}{n} X^t \varepsilon = \beta + Q^{-1} \lim_{n\to\infty} \overline{x_i \varepsilon_i} = \beta + Q^{-1} x0 = \beta$$

Se concluye entonces que, $\hat{\beta}$ es consistente.

· Estimación de β y σ^2 por el principio de máximo verosimilitud.

Suponiendo que el modelo cumple las hipótesis asociadas, en este caso, utilizaremos el método del principio máximo verosímil para estimar $\beta_1, \beta_2, ..., \beta_{k-1}$ y σ^2. La función de verosimilitud se define de la siguiente manera:

$$L(\varepsilon; \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\left(-\frac{\varepsilon^t \varepsilon}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\left(-\frac{(Y - X\beta)^t (Y - X\beta)}{2\sigma^2}\right)$$

Aplicando logaritmos en ambos miembros de la ecuación, se tiene que:

$$\ln L(\varepsilon; \beta, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{(Y - X\beta)^t (Y - X\beta)}{2\sigma^2}$$

Los estimadores máximo verosímiles de β y σ^2 se obtienen derivando la función $\ln L(\varepsilon; \beta, \sigma^2)$ con respecto de β y σ^2:

$$\frac{\partial}{\partial \beta}[\ln L(\varepsilon; \beta, \sigma^2)] = 0$$

$$\frac{\partial}{\partial \sigma^2}[\ln L(\varepsilon; \beta, \sigma^2)] = 0$$

es decir,

$$\frac{2}{2\sigma^2}(X^t Y - X\beta) = 0$$

$$-\frac{n}{2\sigma^2} + \frac{(Y - X\beta)^t (Y - X\beta)}{2\sigma^4} = 0$$

Por lo cual, se obtiene $\hat{\beta}$ y $\hat{\sigma}^2$ de:

$$X^t X \hat{\beta} = X^t X \qquad \hat{\sigma}^2 = \frac{(Y - X\hat{\beta})^t (Y - X\hat{\beta})}{n} \tag{1.3}$$

y como el $rg(X) = k$, se tiene que $rg(X^t X) = k$ por tanto:

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

se observa que el estimador de β por mínimo cuadrático y el de máximo verosímil coinciden, esto se da ya que es equivalente minimizar el exponente de la función de mínimo cuadrático a maximizar la función de máximo verosímil.

Los estimadores $\hat{\beta}$ y $\hat{\sigma^2}$ por ser máximo verosímiles cumplen las siguientes propiedades:

1. Consistentes.
2. Eficientes.
3. Se demuestra que $\hat{\beta}$ y $S^2 = \frac{n}{n-k}\hat{\sigma^2}$ son insesgados.
4. Suficientes.
5. $\hat{\beta}$ y S^2 son independientes

1.4. Contraste de la regresión.

Es importante conocer antes de modelar nuestro problema mediante una regresión lineal múltiple, si nuestras variables independientes tienen o no influencia en el modelo. Así se conoce si las variables son o no significativas.

Primero definimos las siguientes sumas de cuadrados:

$$SSY = \sum_{i=1}^{n}(y_i - \overline{y})^2; \quad SSX = \sum_{i=1}^{n}(x_i - \overline{x})^2$$
$$SXY = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}); \quad SSR = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

Además se define $SS_{reg} = SSY - SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$.

Para ello, es necesario realizar un test de hipótesis en el cual se contraste los siguiente:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = ... = \beta_n = 0 \\ H_1 : Existe\ al\ menos\ un\ \beta_i \neq 0 \end{cases}$$

Si se acepta la hipótesis nula H_0, el modelo no es explicativo, es decir, ninguna de las variables independientes influye en la variable dependiente Y.

Por el contrario, si se rechaza la hipótesis nula, el modelo es explicativo, es decir, al menos una de las variables independientes influye en la variable dependiente Y.

El estadístico que se utiliza para este contraste es el F-test, el cual se define de la siguiente manera:

$$F = \frac{SS_{reg}/k}{SSR/(n-k-1)} = \frac{(SSY - SSR)/k}{SSR/(n-k-1)} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2/k}{\sum_{i=1}^{n}(y_i - \hat{y})^2/(n-k-1)}$$

donde n es el tamaño de la muestra y k los grados de libertad.

Bajo la hipótesis nula el estadístico F sigue una distribución $F_{k,n-(k+1)}$.

1.5. Test de bondad de ajuste.

A la hora de discutir sobre la bondad de ajuste del modelo, es necesario hablar antes de la dispersión del error residual y como se relaciona con la varianza de Y.

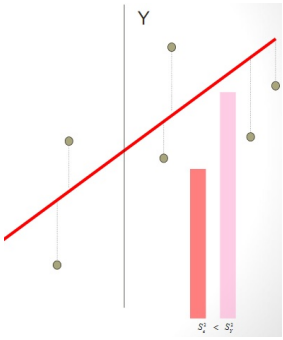

Figura 1.1. Dispersion error residual frente variable Y

Se observa gráficamente que la dispersión del error residual es mas pequeña que la dispersión original de Y. Por tanto cuanto menor sea la dispersión del error residual el ajuste de regresión es mucho mejor.

2

Capítulo 2: Selección de modelos.

En este capitulo se estudiaran diferentes estadísticos que se usan para la comparación de modelos, los algoritmos de los métodos de selección de variables y ampliaremos los conceptos AIC y BIC y determinaremos diferencias entres ambos criterios. Los criterios de información de Akaike y Bayesiano son utilizados de forma usual para la comparación y selección de modelos. Existe discrepancia entre diferentes autores en cual de los dos criterios es más efectivo a la hora de determinar el mejor conjunto de regresión. Es fundamental conocer que dichos criterios están basados en el método de máxima verosimilitud, el cual se rige por que las variables observadas siguen una distribución normal.

2.1. Eligiendo el mejor modelo de regresión.

El mejor modelo de regresión es aquel que proporciona los mejores valores predichos. Por ello debemos utilizar un estadístico que nos permita comparar modelos con diferente número y diferentes variables. Existen multitud de estadísticos de los cuales estudiaremos los siguientes:.

2.1.1. Comparación de diferentes modelos.

a) Coeficiente de determinación o de correlación múltiple R^2.
El "*coeficiente de determinación o de correlación múltiple*" se define como:

$$R^2 = \frac{SSR_{reg}}{SSY} = 1 - \frac{SSR}{SSY} = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2}$$

El coeficiente de determinación se utiliza para comparar distintas regresiones. Pero el coeficiente R^2 tiene consigo una serie de desventaja a la hora de compararse con otros modelos ya que aumenta su valor al introducir nuevas

variables en el modelo. Por tanto el valor se puede incrementar de manera artificial.

El coeficiente de determinación tiene una serie de propiedades:

1. $|R| \leq 1$. Cuando $R = 1$ existe una relación funcional exacta entre la respuesta y las variables explicativas.
2. 100 $(1 - R^2)$ representa el % de variabilidad no explicada por el modelo.

b) Coeficiente de determinación ajustado o corregido.
Este coeficiente se define como un cociente de varianza, en lugar de un cociente de sumas de cuadrados, el cual sirve para evitar que R_a^2 aumente siempre al añadir nuevas variables.

El "*coeficiente de determinación ajustado o de correlación múltiple ajustada*" se define como:

$$R_a^2 = 1 - \frac{Varianza\ Residual}{Varianza\ de\ Y} = 1 - \frac{(SSR)/(n-k)}{(SSY)/n} = 1 - (1 - R^2)\left(\frac{n}{n-k}\right)$$

Este coeficiente, según se define, es independiente del número de variables independientes y al número de casos que existan. De esta manera, podemos comparar dos modelos totalmente independientes entre sí.

d) Criterio de información Akaike (AIC).
El criterio de información Akaike (AIC) se define mediante la siguiente función:

$$AIC(k) = -2 \ln L[\hat{\Theta}(k)] + 2k$$

donde $L[\hat{\Theta}(k)]$ es la función de máxima verosimilitud de las observaciones, $\hat{\Theta}(k)$ la estimación máximo verosímil del vector de parámetros Θ y k el nº de parámetros independientes estimados dentro del modelo.

El primer termino de AIC puede ser interpretado como una medida de bondad de ajuste, mientras el segundo término es una penalización creciente conforme aumenta el número de parámetros. El mejor modelo que se ajusta, según AIC, es aquel con menor valor. Este criterio no pretende identificar el modelo verdadero, sino el mejor modelo entre los modelos candidatos.

e) Criterio de información Bayesiano (BIC). Schwarz (1978) sugirió un criterio de información alternativo a partir de un enfoque bayesiano, el criterio de información bayesiano (BIC). Con este criterio, se penaliza el numero de parámetros con ln n en lugar de 2k. Luego, el BIC se define de la siguiente manera:

$$BIC(k, n) = -2 \ln L[\hat{\Theta}(k)] + k \ln n$$

donde $L[\hat{\Theta}(k)]$ es la función de máxima verosimilitud de las observaciones, $\hat{\Theta}(k)$ la estimación máximo verosímil del vector de parámetros Θ y k el nº

de parámetros independientes estimados dentro del modelo,mientras n es el tamaño de la muestra.

2.1.2. Métodos Paso a Paso.

Los métodos paso a paso que se estudian a continuación son técnicas sistemáticas para examinar solo algunos subconjuntos de cada tamaño. En los diferentes métodos se parte de un modelo inicial y se añadirán o eliminarán variables hasta alcanzar un modelo satisfactorio.

Estudiaremos tres algoritmos básicos para la selección de modelos:

- **Selección hacia delante (Forward Selection):** Las variables independientes se van añadiendo en cada paso en un modelo vacío inicialmente.
- **Selección hacia detrás (Backward Selection):** Las variables independientes se van eliminando en cada paso en un modelo completo inicialmente.
- **Modo paso a paso (Stepwise Selection):** donde las variables independientes se añaden, se eliminan, o se intercambian.

2.2. Algoritmos métodos de selección hacia delante y hacia detrás.

Como estudiamos en el capítulo 1, exiten diferentes métodos de selección de variables. Estos métodos se utilizan en la práctica para determinar los mejores conjuntos de varriables que determinen el mejor modelo posible.

A continuación mostraremos los algoritmos de los métodos de selección.
Forward Selection:

- Paso 0: Se parte del modelo vacío de regresión. $Y = \beta_0 + \varepsilon$.
- Paso 1: Entra la variable mas significativa al modelo según el criterio que estemos utilizando, es decir, el modelo queda de la forma: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ y se realizan los siguientes contrastes:

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases} \quad y \quad \begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

- Paso 2: Entra la segunda variable más significativa. El modelo queda de la forma: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. Y se realizan los siguientes constrastes nuevamente:

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases} \quad y \quad \begin{cases} H_0 : \beta_1 = \beta_2 = 0 \\ H_1 : Otro\ caso. \end{cases}$$

- Se va realizando este proceso tantas veces hasta que encontremos una variable que no es significativa.

Backward Selection:

- Paso 0: Este modelo parte con todas las variables a estudiar, es decir, $Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$.
- Paso 1: Se identifica la variable menos significativa en el modelo. Se realiza un contraste de hipótesis, si al variable es no significativa, se elimina. El modelo quedaría de la siguiente manera: $Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{k-1} X_{k-1} + \varepsilon$-
- Se va realizando este proceso tantas veces hasta que todas las variables del modelo sean significativas.

Stepwise selection: Este procedimiento es una combinación de los dos anteriores.

- Paso 0: El modelo empieza vacio, $Y = \beta_0 + \varepsilon$.
- Paso 1: Introduce la primera variable en el modelo y comprueba si es significativa. Si lo es entra en el modelo. $Y = \beta_0 + \beta_1 X_1 + \varepsilon$.
- Paso 2: Se elimina o añaden mas variables según los test de hipotesis.
- En cada paso se va comprobando si el modelo es significativo y las variables que no lo sean se eliminan.
- El proceso termina cuando no haya mejoras significativas a la hora de añadir o eliminar alguna variable.

A la hora de aplicar estos algoritmos se debe de tener en cuenta la condición para suprimir o incluir un término. Se pueden utilizar diferentes criterios. Por ejemplo, el criterio de significación de cada coeficiente. También se pueden utilizar criterios globales, es decir, una medida global de cada modelo de modo que tenga en cuenta el ajuste y el exceso de parámetros. Dos criterios globales pueden ser el AIC y el BIC, los cuales con valores AIC y BIC pequeños son los mejores modelos.

2.3. Criterio de información de Akaike (AIC).

El motivo fundamental de la selección de modelos mediante el criterio de información de Akaike es estimar la perdida de información al aproximar el modelo real con una función de densidad g, mediante una función de densidad f. Por tanto de aquí que se obtenga la relación con la información de Kullback-Leibler.

El criterio de información Akaike (AIC) es un estimador muestral de la esperanza de la log-verosimilitud. El AIC se define mediante la siguiente función:

$$AIC(k) = -2\ln \mathcal{L}[\hat{\Theta}(k)] + 2k$$

donde $\mathcal{L}[\hat{\Theta}(k)]$ es la función de máxima verosimilitud de las observaciones, $\hat{\Theta}(k)$ la estimación máximo verosímil del vector de parámetros Θ y k el nº de parámetros independientes estimados dentro del modelo.

El primer termino de AIC puede ser interpretado como una medida de bondad de ajuste, mientras el segundo término es una penalización creciente conforme aumenta el número de parámetros, de acuerdo al Principio de Parsimonia, es decir, cuanto mas aumenta el numero de parámetros aumenta el valor de AIC. El mejor modelo que se ajusta, según AIC, es aquel con menor valor. Este criterio no pretende identificar el modelo verdadero, sino el mejor modelo entre los modelos candidatos. El modelo al que se ajusta puede cambiar en función al tamaño muestral, al aumentarlo los parámetros se pueden estimar de una manera mas fiable. Las ventajas que posee el AIC son su simplicidad y facilidad a la hora de implementarse. AIC es ua medida global de la bondad de ajuste del modelo.

Cuando obtenemos muestras pequeñas el AIC deja de ser tan fiable a la hora de determinar un modelo, por ello se define nuevo concepto denominado AIC corregido y se denota por AIC_c. Este criterio determina mejores estimaciones que AIC cuando las muestras son pequeñas. Para muestras grandes el comportamiento de ambas son similares.

El AIC_c queda definido de la siguiente manera:

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$$

2.4. Criterio de información Bayesiano (BIC).

El criterio de información Bayesiano(BIC) es propuesto por Schwarz (1978), argumentando que el AIC no es justificable asintóticamente y por ello presentó una modifición de AIC, donde en BIC se penaliza el número de parámetros con ln n, en vez de 2. Luego, el BIC se define de la siguiente manera:

$$BIC(k,n) = -2\ln \mathcal{L}[\hat{\Theta}(k)] + k\ln n$$

donde $\mathcal{L}[\hat{\Theta}(k)]$ es la función de máxima verosimilitud de las observaciones, $\hat{\Theta}(k)$ la estimación máximo verosímil del vector de parámetros Θ y k el nº de parámetros independientes estimados dentro del modelo,mientras n es el tamaño de la muestra.

Este criterio posee muchas características semejantes al AIC como por ejemplo: que el objetivo fundamental es el de estimar la perdida de información de aproximar un modelo real con un modelo estimado, es decir, cuanto menor

sea el valor BIC mejor modelo obtendremos, al aumentar el tamaño muestral la estimación es mucho mas fiable, es una medida global de la bondad del ajuste,...

En cambio, este modelo hace una mayor restricción en los parámetros ya que se penaliza con el valor de $ln(n)k$. Por tanto, si se introducen mas parámetros en el modelo, mejorara el ajuste y disminuirá la desviación que es medida por $-2\ln\mathcal{L}[\hat{\Theta}(k)]$. Esto supone una compensación entre la desviación y la penalización.

2.5. Diferencias entre el criterio AIC y el BIC.

Los criterios de información AIC y BIC son de los mas usado a la hora de determinar que modelo es el que mejor se ajusta al modelo real, pero existe la discrepancia de cual de los dos criterios encuentra el mejor modelo de regresión. Para ello en esta sección del capitulo veremos algunas diferencias notorias entre ambos criterios.

La primera diferencia que se observa, la cual ya hemos citado, es la del termino de penalización donde el BIC penaliza el número de parámetros con el ln del número de observaciones tomadas. Por el motivo anterior, al tomar una restricción mas fuerte en la penalización usando el criterio BIC, este aborda modelos con menor dimensiona que AIC. Cuando el tamaño muestral es grande, existe diferencias entre AIC y BIC.

Anteriormente se comentó que ambos criterios tienen la misma finalidad, encontrar el modelo que minimice la perdida de información en ajustar el modelo estimado a uno real. Pero el BIC trata de acercarse mas al modelo real, asumiendo así que el modelo real esta incluido entre los posible modelos. En cambio el criterio AIC trata de obtener modelos que proporcione mejores predicciones entre todos los modelos dados, no asume que entre los posibles modelos se encuentre el modelo real. EL criterio BIC se demuestra que es un modelo consistente, es decir, el error de medida o sesgo se aproxima a cero cuando el tamaño de la muestra tiende a infinito. Y que AIC es un modelo eficiente, es decir, si comparamos dos modelos estimados insesgados por AIC, se dice que 1 es mas eficiente que 2 si y solo si la varianza de 1 es menor que 2. Por el contrario, cuando el tamaño muestral es pequeño, AIC tiende a seleccionar modelos con mas parámetros de los necesarios.

En consecuencia de las diferencias entre ambos criterios, es difícil determinar cual de los dos es mejor que el otro. Todo depende de las condiciones en las que se encuentre la estimación del modelo. Pero la diferencia básica entre ambos criterios es que BIC penaliza mas los modelos con un mayor numero de parámetros que AIC, de tal manera que se obtienen modelos de orden inferior a los de AIC. Además de que los modelos que selecciona AIC se ajusta mejor para tamaños muestrales grandes.

3

Capítulo 3: Análisis de sensibilidad en la regresión lineal.

A lo largo del tiempo se ha descubierto que la regresión lineal múltiple es una herramienta que permite realizar predicciones de ciertos acontecimientos que suceden en la realidad. Bien es cierto que a priori, se puede tratar de un estudio simple, pero nada mas lejos de la realidad ya que dicho modelo requiere del cumplimiento de ciertas condiciones e hipótesis para poder hacerse uso de él. En este capitulo se llevará a cabo un seguimiento de lo importante que debe satisfacer un estudio de la regresión y como se puede diagnosticar, resolver e implementar.

3.1. Validación del modelo e hipótesis asociadas

El modelo de regresión lineal múltiple, como ya vimos en el capitulo 1, lleva consigo unas hipótesis asociadas que debe de satisfacer. Es importante saber como diagnosticar las hipótesis y en caso que no se satisfagan saber que solución abordar.

3.1.1. Distribución normal de la variable dependiente

En el capitulo 1 estudiamos que satisfacer las hipótesis asociadas al error eran equivalentes a satisfacerlas para la variable dependiente. Entonces, la variable dependiente elegida en el modelo debe de seguir una distribución normal.

Existen diferentes maneras de diagnosticar si la variable dependiente sigue o no una distribución normal, en este capitulo estudiaremos las siguientes:

- **Gráficos:** Gráficamente no se puede diagnosticar si la variable dependiente sigue una distribución normal o no con rigurosidad, ya que no deja de ser una prueba visual. Sin embargo, nos permite tener una aproximación. Para

ello los gráficos mas usados son es el histograma con la campana de Gauss superpuesta, la función de densidad de la variable y los gráficos P-P ó Q-Q.

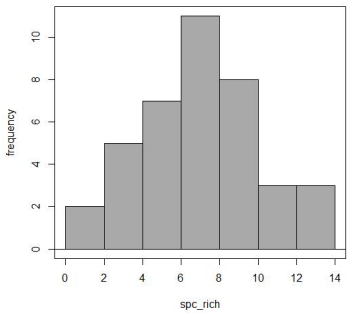

Figura 3.1. Histograma de una posible variable que sigue una distribución normal.

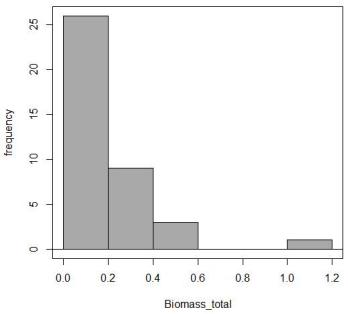

Figura 3.2. Histograma de una posible variable que no sigue una distribución normal.

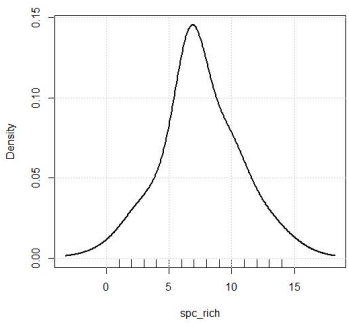

Figura 3.3. Función de densidad de una posible variable que sigue una distribución normal.

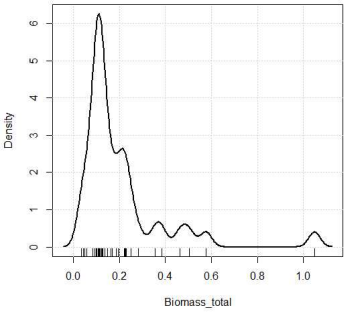

Figura 3.4. Función de densidad de una posible variable que no sigue una distribución normal.

En las gráficas anteriores se observa que la variable puede asemejarse a la curva de la distribución normal.

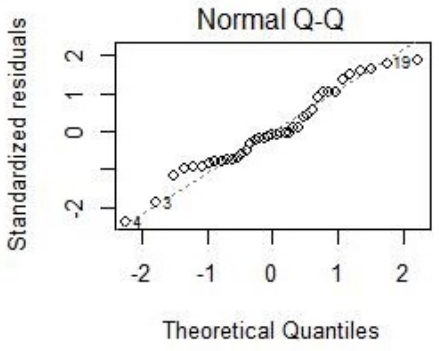

Figura 3.5. Gráfica Q-Q plot

Lo que se espera que ocurra en el gráfico Q-Q plot es que los residuos estandarizados estén lo más cerca posible a la línea punteada que aparece en el gráfico.

- **Teorema Central del Límite:** Téngase una variable aleatoria cuyas observaciones se miden en promedio, como por ejemplo: uso medio del teléfono. Entonces si se toman las medias en una muestra aleatoria de más de 30 elementos, se tiene que la distribución de la medias de la muestra es normal. Por tanto, podemos garantizar que al ser la distribución de las medias normal, la distribución de la población de la variable aleatoria es normal.
- **Test de normalidad Kolmogorov-Smirnov:** El test de Kolmogorov-Smirnov se utiliza para contrastar si un conjunto de datos se ajustan o no a una distribución normal. Este test diagnostica la normalidad de la variable a nivel poblacional. El contraste de hipótesis que se platea en este test es el siguiente:

$$\begin{cases} H_0 : Y \approx NORMAL \\ H_1 : Y \neq NORMAL \end{cases}$$

El estadistico de prueba que se emplea en el test Kolmogorov-Smirnov es la máxima diferencia:

$$D = max|F_n(x) - F_o(x)|$$

siendo $F_n(x)$ la función de distribución muestral y $F_o(x)$ la función teórica o correspondiente a la población normal.

En el supuesto caso que la variable dependiente de nuestro modelo no siguiese una distribución normal, entonces se tendría que salir de la estadística paramétrica e irnos a una estadística no paramétrica.

3.1.2. Homocedasticidad y linealidad

El modelo de regresión lineal múltiple debe asumir que la varianza de los errores es constante a lo largo de las observaciones. La homocedasticidad es una cualidad del modelo necesaria para que en un modelo los coeficientes estimados sean eficientes, lineales e insesgado. En el caso de no cumplirse se habla de heterocedasticidad.

Una forma de diagnosticar la homocedasticidad del modelo es mediante el gráfico de residuos frente a predichos. En dicho gráfico lo ideal es observar que los residuos no muestren ningún patrón, ya que si observamos por ejemplo que el tamaño de los residuos aumenta o disminuye a medida que lo hace los predichos entonces no se cumple la condición de homocedasticidad.

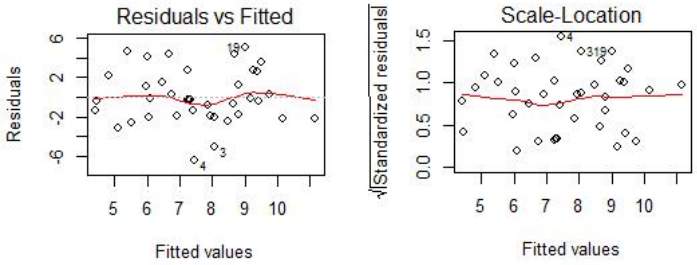

Figura 3.6. Gráfica residuos vs. predichos

En la gráfica se observa como en este caso existen mínimas perturbaciones en el modelo, por tanto se tiene una banda "casi" rígida y constante por lo que el modelo se puede describir como homocedástico. A continuación veremos un caso en el cual no se tiene homocedasticidad.

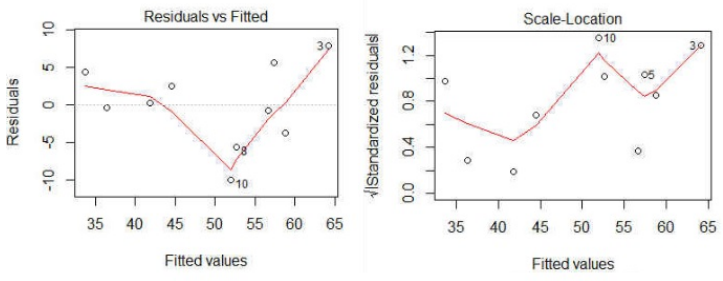

Figura 3.7. Gráfica residuos vs. predichos

En el supuesto caso que se obtuviese heterocedasticidad, como se da en el gráfico anterior, entonces se debe de realizar transformaciones en el modelo para obtener homocedasticidad. Una de las transformaciones mas conocidas es la transformación de Box & Cox. Esta transformación se introduce en 1964 con el objetivo de satisfacer las condiciones de normalidad y homocedasticidad. Esta trata de aplicar una transformación de la variable Y, la cual ajusta el modelo.

La transformación de Box & Cox es la siguiente:

$$W = \begin{cases} \frac{(Y^{\lambda}-1)}{\lambda} & si\ \lambda \neq 0 \\ \\ \ln Y & si\ \lambda = 0 \end{cases}$$

3.1.3. Autocorrelación, Independencia de errores

El modelo de regresión lineal múltiple debe de satisfacer que los errores del modelo sean independientes, es decir, que los valores no dependan unos de otros.

La no verificación de esta hipótesis supone un gran problema para el modelo, ya que:

- Los estimadores dejan de ser eficientes, es decir, de varianza mínima.
- Los contraste de significación dejan de ser validos tendiendo a detectar modelos inexistentes.
- Los predictores son ineficientes.
- La falta de independencia se suele dar en situaciones donde las observaciones son recogidas secuencialmente en el tiempo. Ésto ocurre en el estudio de muchas variables económicas, sociales y demográficas.

La existencia o no de autocorrelación en el modelo se puede detectar de las siguiente manera:

- **Gráfico de los residuos de la estimación en el tiempo.** Es importante conocer el comportamiento de los errores en el tiempo ya que si existen patrones de comportamiento sistemático en la distribución en el tiempo de los residuos, entonces podemos determinar autocorrelación en el modelo.

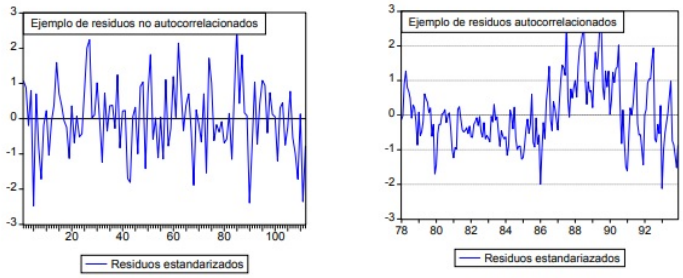

Figura 3.8. gráfica de los residuos de la estimación en el tiempo

En el gráfico anterior vemos un ejemplo de autocorrelación, y otro donde no.

- **Test de Durbin-Watson.** El test de Durbin-Watson sirve como diagnostico de la autocorrelación del modelo. Dicho test realiza el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : \rho = 0 \quad [ErroresIndependientes] \\ H_1 : \rho \neq 0 \quad [ErroresDependientes] \end{cases}$$

El estadístico que se utiliza en el test, siendo e_t el residual asociado a la observación en el tiempo t, donde la varianza y la media de e_t son constante, independientemente de t. Se tiene que $e_t = \rho e_{t-1} + z_t$, donde $z_t \sim N(0, \sigma^2)$, y es independiente de $e_{t-1}, e_{t-2}, ...$ y de z_{t-1}, z_{t-2}. Además para el contraste de hipótesis anterior se debe asumir que $e_t \sim N(0, \sigma^2)$. Entonces una ves calculados los e_t, podemos definir el estadístico de Durbin-Watson:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

donde n es el número de observaciones.
El valor de d siempre está entre 0 y 4. En dicho test se tiene de calcular una cota inferior ("lower-tailed") y una cota superior (üpper-tailed") en función del valor de probabilidad $\alpha = 0.05, 0.025, 0.01$, denominadas d_L *y* d_U respectivamente, tal que son los valores críticos que determinarán, según se encuentre el valor del estadístico d, si existe autocorrelación o no.
Si $d < d_L$ el test es significativo, rechazamos H_0 por tanto existe dependencia de los errores a un nivel 2α, si $d > d_U$ se concluye que d no es significativo, por lo que no se puede rechazar H_0 y existe independencia de errores a un nivel 2α, y si $d_L < d < d_U$ no se puede determinar nada. Es importante destacar de que los valores de los estadísticos son todos a nivel muestral.
Cuando existe autocorrelación, el método de mínimos cuadrados deja de ser

eficiente, por lo cual el modelo no se puede estimar bajo las condiciones de mínimos cuadrados. Un método alternativo para estimar los coeficientes puede ser el de mínimos cuadrados generalizados.

3.1.4. Multicolinealidad

La multicolinealidad es un problema del tipo muestral, debido a que está asociado con la matriz X. Por tanto, la no existencia de multicolinealidad es una hipótesis que debe satisfacer el modelo, pues si no estamos ante una situación grave a la hora de calcular los estimadores mínimos cuadráticos, ya que si existe multicolinealidad la matriz X'X es no singular. Este hecho es difícil que se presente en la práctica, lo que si es usual es encontrarse entre las variables una alta correlación, por lo que los estimadores que se calculan son pocos precisos. Es decir, las varianzas de los estimadores son elevadas.

Para detectar la multicolinealidad se han desarrollados numerosas reglas que determinan en cuanto afecta la multicolinealidad a la estimación y contraste del modelo.

La regla que nosotros veremos es factores de inflación de varianza (VIF), la cual se define de la siguiente manera:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R_j^2}$$

donde $\hat{\beta}_j$ es el estimador j del modelo y R_j^2 es el coeficiente de determinación obtenido al efectuar la regresión de la variable X_j sobre el resto de las variables del modelo.

El FIV determina en que grado aumenta la varianza del estimador como consecuencia de la no ortogonalidad de las variables. Muchos autores coinciden en que existe un problema grave de multicolinealidad cuando el FIV de algún estimador es mayor de 10.

El problema de multicolinealidad, una vez detectado, se puede solucionar de diferentes maneras, entre ellas:

- **Eliminar variables del modelo.** Es decir, las variables que mas estén afectadas por la multicolinealidad se pueden eliminar y así evitar este problema. Para ello se puede acudir a los procesos de selección de variables visto en el capitulo 1.
- **Aumentar el tamaño de la muestra**. Esto podría llevar consigo la reducción de la varianza.
- **Utilizar información extramuestral**, ya sea estableciendo restricciones sobre los parámetros del modelo, o bien aprovechando estimadores procedentes de otros estudios.

3.2. Efecto de algunas observaciones en el modelo

El objetivo nuestro es encontrar un modelo de regresión que se ajuste lo máximo posible a los datos reales, pero esto puede verse afectado por ciertas observaciones que influyen en el modelo, que se recogen en la muestra. Es necesario saber identificar estas posibles observaciones, en este caso estudiaremos tres tipo: outliers, alto potencia influyentes.

3.2.1. Outliers, Alto Potencial, e Influyentes

- **Outliers:** Tenemos una observación outliers cuando la magnitud de su residual, en valor absoluto, es superior al resto. Este tipo de datos suelen tener una alteración en el comportamiento con respecto al resto de datos. Por lo tanto es importante su detección para estudiar su posible eliminación si es un error en la toma de los datos o también puede indicar que el modelo no es el correcto.
  Gráficamente los outliers pueden presentarse ocultos, o podemos encontrarnos "falsos" outliers. Esto es mas conocido como las técnicas de:
  "Masking": dado un punto "x_i" outliers de forma que este tira de la recta de regresión hacia él obteniendo así un error residual pequeño.
  "Swamping": de la forma que la recta se aproxima al punto "x_i", entonces existirá un punto "x_j"que obtiene un valor del error residual alto.
  En consecuencia, existen medidas basadas en los residuales que son mas resistentes a los fenómenos de masking y swamping, que detectan los outliers. En este caso veremos las siguientes medidas basada en los residuales:

$$(1)\ r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-p_{ii}}} \qquad ; \qquad (2)\ r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-p_{ii}}}$$

  donde (1) se denomina el i-ésimo residual estudentizado interno y (2) el i-ésimo residual estudentizado externo.

- **Alto potencial:** Un punto de alto potencia es aquel con un gran valor de p_{ii}, comparándolo con el resto. En el espacio X los puntos alto potencial pueden ser reconocidos como outliers. Una posibilidad común de que una observación sea alto potencial es que el rango de X este mal considerado, por lo que se tendría que aumentar el tamaño muestral.
  Existen varias medidas para localizar los datos alto potencial de las observaciones recogidas, entre ellas vamos a ver los valores de $p_{ii} = x_i^t(X^tX)^{-1}x_i$ $\forall = 1, 2, ..., n$, que son los elementos de la diagonal de P, matriz vista en el capitulo 1. Los datos alto potencial son aquellos que tienen un valor mucho mas alto de los p_{ii} con respecto al resto de observaciones.

- **Influyentes:** Son las observaciones que de una manera considerada influyen en la ecuación de regresión en comparación a las otras observaciones acorde con alguna medida de influencia.
  La influencia de las observaciones no siempre influyen de la misma manera en la ecuación de regresión, ya que estos pueden influir en $\hat{\beta}$, en la varianza de $\hat{\beta}$, en los valores predichos, en la bondad de ajuste,...
  Para detectar los valores influyentes, entre otras medidas, se utiliza la distancia de Cook. Cook propone que la influencia i-ésima debe ser medida por:

$$C_i = \frac{(\hat{Y} - \hat{Y}_{(i)})^t(\hat{Y} - \hat{Y}_{(i)})}{k\hat{\sigma}^2}$$

  donde :
  - $\hat{Y} = Xb \equiv$ vector valores predichos.
  - $\hat{Y}_{(i)} = Xb_{(i)} \equiv$ vector de los valores predichos cuando el i-ésimo punto es eliminado.
  - $b_{(i)} \equiv$ es el correspondiente estimador de mínimos cuadrados del vector de parámetros β cuando el i-ésimo punto es eliminado.

  como $\hat{Y} - \hat{Y}_{(i)} = X\{b - b_{(i)}\}$, entonces C_i se pede escribir de la siguiente manera:

$$C_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^t(X^tX)(\hat{\beta} - \hat{\beta}_{(i)})}{k\hat{\sigma}^2}$$

  Una tercera manera de presentar la distancia de Cook es combinando el i-ésimo residual estudentizado interno y los valores p_{ii}:

$$C_i = \left[\frac{e_i}{\hat{\sigma}\sqrt{1-p_{ii}}}\right]^2 \frac{p_{ii}}{1-p_{ii}}\frac{1}{k} = r_i^2\frac{p_{ii}}{1-p_{ii}}\frac{1}{k}$$

  donde :
  - $e_i \equiv$ es el residual i-ésimo.
  - $p_{ii} \equiv$ es el elemento i-ésimo de la diagonal de la matriz P.

  Por tanto las observaciones con un alto valor de C_i con respecto a las demás observaciones, se consideran observaciones influyentes en la ecuación de regresión.

3.2.2. Gráficos.

A continuación observaremos algunos gráficos donde podemos observar este tipo de observaciones anómalas:

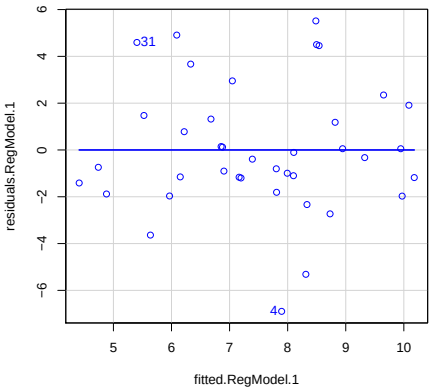

Figura 3.9. Gráfica de residuos vs. predichos.

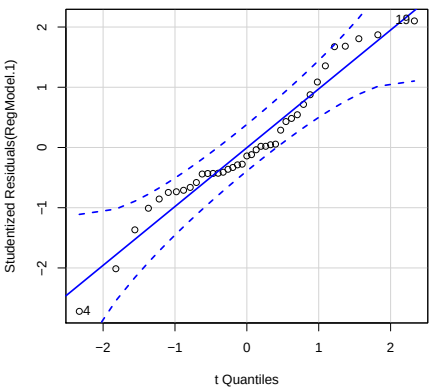

Figura 3.10. Gráfica residuos estudentizado

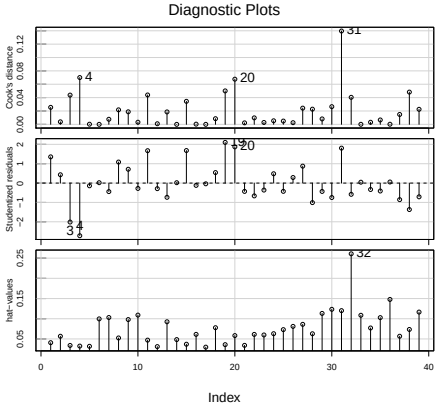

Figura 3.11. Gráfica de los valores, de las medidas utilizadas, de las observaciones.

4

Capítulo 4: Regresión lineal múltiple en R Studio y R commander.

Hoy en día para el desarrollo de todo lo estudiado en capítulos anteriores, existen infinidad de paquetes informáticos que nos permite aplicar esos conocimientos. Entre los más comunes hemos elegido R, R studio y R commander, ya que es de código abierto y software libre. En este capitulo haremos un recorrido completo de como realizar un estudio de regresión lineal mútliple (RLM) en R.

En este desarrollo vamos a intentar elegir el mejor subconjunto de regresión que se ajuste a lo publicado en el paper por JR. Arévalo "Manejo de pasturas en pastoreo de cabras en Canarias"[1].

4.1. Objetivo.

El objetivo de este estudio es el de predecir la riqueza (spc_rich) del suelo por una RLM en función de las variables independiente: MO: % materia orgánica, Ca, Na, Mg, K, P,C.I.C: complejo de intercambio cationes, pH, C.E: intercambio de cationes y Sat: % saturación del suelo.

4.2. Background.

En el Departamento de Ecología de la ULL se ha realizado un estudio de las comunidades vegetales en zonas con presencia de cabras de la península de Teno para estudiar el comportamiento de parcelas de suelo recogido de 0,5x0,5 m en 10 áreas aleatoriamente seleccionadas en 1992, 1993,1994 y 1999. En el paper publicado por el departamento se realizan dos partes:

1. Predecir la riqueza y la biomasa del suelo por una RLM en función de las variables independiente: MO: % materia orgánica, Ca, Na, Mg, K, P,C.I.C:

complejo de intercambio cationes, pH, C.E: intercambio de cationes y Sat: % saturación del suelo.

2. Realizar un análisis de componentes principales para las variables MO: % materia orgánica, Ca, Na, Mg, K, P,C.I.C: complejo de intercambio cationes, pH, C.E: intercambio de cationes y Sat: % saturación del suelo

En este caso solamente nos centraremos en el objetivo indicado anteriormente.

4.3. Naturaleza de los datos.

Los datos vienen dados por un paper publicado por JR Arévalo (ULL), del estudio de las comunidades vegetales en zonas de presencias de cabra en la península de teno. Las parcelas de suelo recogido de 0,5x0,5 m en 10 áreas aleatoriamente seleccionadas en 1992, 1993,1994 y 1999. Las variables a utilizar son todas cuantitativas y continuas.

4.4. Modelo.

El modelo que se empleará en este caso, es el que hemos estudiado en los capítulos 1 y 2. Es decir, regresión lineal múltiple.

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{k-1} X_{k-1} + \varepsilon$$

Con las siguientes hipótesis asociadas:

- Normalidad en la variable Y.
- Independencia de los errores.
- Homocedasticidad y linealidad.
- Multicolinealidad.

4.5. Estudio Piloto.

En este caso en concreto no procede realizar un estudio piloto ya que los datos los hemos heredado del profesor JR Arévalo (ULL) a través de la profesora María Mercedes Suárez Rodriguez. Si hubiésemos participado en el equipo de diseño, se debiera hacer el estudio piloto el cual nos permite ver ciertas conclusiones del estudio final.

4.6. Tamaño y diseño muestral.

Dado que no hemos participado en el análisis de la toma de la muestra a determinar y los datos nos vienen dados, no conocemos la manera que se ha llevado a cabo para la obtención de la muestra. El tamaño muestral del que disponemos es de 39 observaciones. La población son todas las parcelas en la península de Teno con presencia de cabras.

4.7. Análisis de los datos.

En este apartado del capitulo realizaremos el análisis de los datos obtenidos por el profesor JR Arévalo. Este análisis lo realizaremos en R Commander y R Studio como ya se comento anteriormente.

R Commander, también conocido como Rcmdr, es una librería que pertenece y se instala directamente de R, siendo una interfaz gráfica la cual proporciona una forma visual donde los usuarios se familiariza con los comandos de R, y así poder desarrollar su conocimiento y experiencia en el uso de la línea de comandos.. Dicha interfaz permite realizar estudios estadístico como paquetes como el SPSS, basándose en el lenguaje de programación de R. En nuestro caso, mostraremos los comandos utilizados por R para realizar cualquier prueba o estudio en Rcmdr.

Para la realización del estudio que queremos abordar debemos de introducir o genera los datos en Rcmdr o Rstudio. Existen muchas formas de generar o introducir los datos en Rcmdr y Rstudio pero nosotros los importaremos desde SPSS ya que es el formato que tiene. Entonces en Rcmdr debemos seguir la siguiente ruta:

Datos → Importar Datos → Desde datos SPSS.

y la sentencia de R que debemos usar es:

- *TENO <- readSPSS(" C:/Users/Eduardo Hernandez/Desktop/TFG Eduardo/Datos/teno.sav", rownames=FALSE, stringsAsFactors=TRUE, tolower=FALSE)*

4.7.1. Estadística descriptiva.

La estadística descriptiva en Rcmdr la realizamos en:

Estadístico → Resúmenes → Resúmenes numéricos.

donde podemos seleccionar los estadístico que queremos estudiar de la variable selecionada, como por ejemplo: la media, la desviación típica, error típico de la media, rango intercuartilico, coeficientes de variación, frecuencia por intervalos, asimetría, apuntamiento, cuantiles,...

Las sentencias utilizadas son:

- *numSummary(TENO[,"spc_rich", drop=FALSE], statistics=c("mean", "sd", "se(mean)", ÏQR", "quantiles", "cv", "skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1), type="2").*

- *binnedCounts(TENO[,"spc_rich", drop=FALSE]).*

  Observamos el resultado de las sentencias:

  - ```
    mean       sd  se(mean) IQR        cv  skewness  kurtosis 0% 25% 50%
    7.512821 3.136137 0.5021839   4 0.4174381 0.0627248 -0.355313  1   6   7
    75% 100%  n
    10   14 39
    ```
  - ```
    Binned distribution of spc_rich
    Count Percent
    (0, 2]       2    5.13
    (2, 4]       5   12.82
    (4, 6]       7   17.95
    (6, 8]      11   28.21
    (8, 10]      8   20.51
    (10, 12]     3    7.69
    (12, 14]     3    7.69
    Total       39  100.00
    ```

De esta manera obtenemos los estadístico que queremos a la hora de mostrar una estadística descriptiva de la variable que se desea. En nuestro caso hemos estudiado la que sera nuestra variable dependiente del modelo, spc_rich (riqueza del suelo).

Se observa que se tiene una media de 7.51, con desviación típica 3.13.

Ademas debemos de realizar un histograma para ver gráficamente como se agrupan las observaciones de la variable dependiente.

En Rcmdr debemos ir a:

Gráficas → Histogramas

La sentencia R a utilizar es:

- *with(TENO, Hist(spc_rich, scale="frequency", breaks="Sturges", col="darkgray"))*

destacar de que en la sentencia se observa que la escala usada es la frecuencia y la separación de los puntos es la de Sturges, que a diferencia de SPSS que hay que exigírselo, en Rcmdr viene predeterminada.

El resultado de la sentencia es:

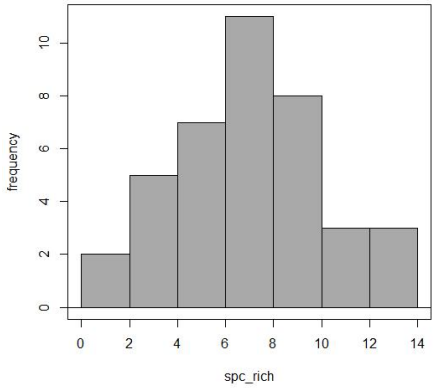

Figura 4.1. Histograma Spc_rich

4.7.2. Correlación entre las variables.

De manera orientativa e informativa realizamos la matriz de correlación pero 1 a 1, ya que esta no muestra la correlación múltiple de las variables. En esta matriz veremos la correlación entre todas las variables, una frente a otra.

Para ello debemos de utilizar la ruta siguiente:

Estadísticos → Resúmenes → Matriz de correlaciones.

La sentencia de R es:

- *cor(TENO[,c(" C.E"," C.I.C"," CA","K","Mg","MO","Na","P","PH","SAT","spc_rich")], use="complete")*

y el resultado de la sentencia es:

```
C.E        C.I.C          CA           K           Mg
C.E    1.000000000  0.534865759  0.35861513 -0.036646052  0.515242238
C.I.C  0.534865759  1.000000000  0.84622227 -0.097149931  0.930303555
CA     0.358615125  0.846222270  1.00000000  0.072844513  0.653104496
```

```
K         -0.036646052 -0.097149931  0.07284451  1.000000000 -0.272022455
Mg         0.515242238  0.930303555  0.65310450 -0.272022455  1.000000000
MO         0.005452799 -0.045073004 -0.15360189  0.035435377 -0.116767949
Na         0.410531629  0.442150077  0.40713810 -0.237526663  0.377365726
P         -0.317981184 -0.342687414 -0.15580486 -0.024925384 -0.454240546
PH        -0.049523292  0.179649303  0.49004931  0.157509172  0.178129885
SAT       -0.112258215  0.318019852  0.33354480  0.005770965  0.309291416
spc_rich -0.371487666 -0.008803347  0.06315516  0.076236105  0.001267852
MO           Na           P           PH           SAT
C.E        0.005452799  0.41053163 -0.31798118 -0.04952329 -0.112258215
C.I.C     -0.045073004  0.44215008 -0.34268741  0.17964930  0.318019852
CA        -0.153601886  0.40713810 -0.15580486  0.49004931  0.333544801
K          0.035435377 -0.23752666 -0.02492538  0.15750917  0.005770965
Mg        -0.116767949  0.37736573 -0.45424055  0.17812988  0.309291416
MO         1.000000000 -0.28264667  0.06315284 -0.58446177  0.142855457
Na        -0.282646669  1.00000000 -0.13252447  0.15798371 -0.103466213
P          0.063152843 -0.13252447  1.00000000 -0.23944759  0.095851353
PH        -0.584461770  0.15798371 -0.23944759  1.00000000  0.158648089
SAT        0.142855457 -0.10346621  0.09585135  0.15864809  1.000000000
spc_rich   0.008186121 -0.09089026 -0.19437832  0.18963533  0.216399843
spc_rich
C.E       -0.371487666
C.I.C     -0.008803347
CA         0.063155160
K          0.076236105
Mg         0.001267852
MO         0.008186121
Na        -0.090890258
P         -0.194378316
PH         0.189635326
SAT        0.216399843
spc_rich   1.000000000
```

4.7.3. F-test o Anova de RLM.

Para ver si el modelo de regresión lineal múltiple tiene sentido o no, debemos realizar el F-Test o anova de regresión, el cual realiza un contraste de hipótesis entre:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = ... = \beta_n = 0 \\ H_1 : Existe\ al\ menos\ un\ \beta_i \neq 0 \end{cases}$$

En Rcmdr una vez que generamos el modelo directamente muestra mucha información que a continuación detallaremos, y entre ella aparece el F-test. Para generar el modelo debemos de seguir la siguiente ruta en Rcmdr:

Estadísticos → Ajuste de modelos → Regresión lineal.

donde la sentencia R es:

- *RegModel.1 <- lm(spc_rich ~ C.E+C.I.C+CA+K+Mg+MO+Na+P+PH+SAT, data=TENO)*
- *summary(RegModel.1)*

y cuyo resultado es:

```
Call:
lm(formula = spc_rich ~ C.E + C.I.C + CA + K + Mg + MO + Na +
P + PH + SAT, data = TENO)

Residuals:
Min      1Q  Median      3Q     Max
-5.9852 -1.2547 -0.3902  1.1764  4.7972

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.05557   56.49337   2.001  0.05515 .
C.E         -11.94712    3.92323  -3.045  0.00502 **
C.I.C        -4.05158    2.17609  -1.862  0.07315 .
CA            4.90722    2.55909   1.918  0.06542 .
K             3.56083    2.01642   1.766  0.08831 .
Mg            4.23040    2.31370   1.828  0.07816 .
MO            0.07184    0.33217   0.216  0.83033
Na            3.71237    2.26433   1.640  0.11230
P            -0.14138    0.06167  -2.293  0.02959 *
PH          -14.85626    8.20600  -1.810  0.08098 .
SAT           0.01407    0.06059   0.232  0.81802
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.866 on 28 degrees of freedom
Multiple R-squared:  0.3847,    Adjusted R-squared:  0.1649
F-statistic: 1.751 on 10 and 28 DF,  p-value: 0.1181
```

En el resultado de la sentencia se obtienen los siguientes estadísticos: error residual estándar, R^2 múltiple, el R^2 ajustado, el F-test y su p-valor. Ademas de los coeficientes de regresión, estos se calculan en Rcommander por el método de mínimos cuadrados.

Como se observa el p-valor es 0.118 por tanto es mayor que alfa igual a 0.01 y 0.05 y por tanto no existen evidencias para rechazar H_0 por tanto el modelo es no significativo. Por tanto, el modelo propuesto utilizando las 10 variables inde-

pendiente no tiene sentido por lo que debemos hacer una selección de variables y así elegir el mejor subconjunto de regresión que se ajuste al problema inicial.

Como hemos estudiado en capítulos anteriores existen varios métodos de selección de variables. Nosotros utilizaremos selección hacia delante (forward selection) y selección hacia detrás (barckward selection). Ademas hemos visto también que existen dos criterios de información Akaike y Bayesiano, AIC y BIC respectivamente. Rcmdr utiliza uno u otro criterio, el que nosotros elijamos, para determinar la selección de las variables. Entonces utilizaremos ambos criterios por tanto debemos de obtener 4 subconjuntos de regresión.

La ruta que debemos seguir es:

Modelos → Selección de modelo paso a paso.

La sentencia de R a utilizar serán:

1. **Backward Selection BIC:** *stepwise(RegModel.1, direction='backward', criterion='BIC')*
2. **Forward Selection BIC:** *stepwise(RegModel.1, direction='forward', criterion='BIC')*
3. **Backward Selection AIC:** *stepwise(RegModel.1, direction='backward', criterion='AIC')*
4. **Forward Selection AIC:** *stepwise(RegModel.1, direction='forward', criterion='AIC')*

Los resultados de cada sentencia son los siguientes:

1.
```
Direction:  backward
Criterion:  BIC

Start:  AIC=109.5
spc_rich ~ C.E + C.I.C + CA + K + Mg + MO + Na + P + PH + SAT

Df Sum of Sq    RSS    AIC
- MO     1     0.384 230.35 105.90
- SAT    1     0.443 230.41 105.91
- Na     1    22.077 252.04 109.41
<none>               229.97 109.50
- K      1    25.612 255.58 109.95
- PH     1    26.919 256.89 110.15
- Mg     1    27.457 257.42 110.23
- C.I.C  1    28.471 258.44 110.39
- CA     1    30.200 260.17 110.65
- P      1    43.170 273.14 112.55
- C.E    1    76.163 306.13 116.99
```

```
Step:  AIC=105.9
spc_rich ~ C.E + C.I.C + CA + K + Mg + Na + P + PH + SAT

       Df Sum of Sq    RSS    AIC
- SAT    1     0.795 231.15 102.37
- Na     1    21.767 252.12 105.76
<none>               230.35 105.90
- K      1    25.354 255.71 106.31
- Mg     1    27.182 257.53 106.59
- C.I.C  1    28.350 258.70 106.76
- PH     1    28.707 259.06 106.82
- CA     1    30.402 260.75 107.07
- P      1    50.399 280.75 109.95
- C.E    1    75.780 306.13 113.33

Step:  AIC=102.37
spc_rich ~ C.E + C.I.C + CA + K + Mg + Na + P + PH

       Df Sum of Sq    RSS    AIC
<none>               231.15 102.37
- Na     1    22.987 254.13 102.41
- K      1    31.849 263.00 103.74
- Mg     1    35.079 266.23 104.22
- PH     1    35.147 266.29 104.23
- C.I.C  1    35.286 266.43 104.25
- CA     1    38.094 269.24 104.66
- P      1    49.605 280.75 106.29
- C.E    1    89.742 320.89 111.50

Call:
lm(formula = spc_rich ~ C.E + C.I.C + CA + K + Mg + Na + P +
PH, data = TENO)

Coefficients:
(Intercept)          C.E        C.I.C           CA            K
121.2593     -12.2755      -4.2630       5.1893       3.7443
Mg           Na            P           PH
4.4594       3.6969      -0.1433     -15.9250
```

2. ```
   Direction:  forward
   Criterion:  BIC

   Start:  AIC=91.8
   spc_rich ~ 1
   ```
   ```

```
Df Sum of Sq    RSS    AIC
+ C.E    1    51.578 322.17 89.676
<none>               373.74 91.804
+ SAT    1    17.502 356.24 93.597
+ P      1    14.121 359.62 93.965
+ PH     1    13.440 360.30 94.039
+ Na     1     3.088 370.66 95.144
+ K      1     2.172 371.57 95.240
+ CA     1     1.491 372.25 95.312
+ C.I.C  1     0.029 373.71 95.464
+ MO     1     0.025 373.72 95.465
+ Mg     1     0.001 373.74 95.467

Step:  AIC=89.68
spc_rich ~ C.E

Df Sum of Sq    RSS    AIC
+ P      1    40.605 281.56 88.085
<none>               322.17 89.676
+ Mg     1    18.889 303.28 90.983
+ C.I.C  1    18.877 303.29 90.984
+ CA     1    16.540 305.63 91.284
+ SAT    1    11.552 310.61 91.915
+ PH     1    10.986 311.18 91.986
+ Na     1     1.707 320.46 93.132
+ K      1     1.468 320.70 93.161
+ MO     1     0.039 322.13 93.335

Step:  AIC=88.09
spc_rich ~ C.E + P

Df Sum of Sq    RSS    AIC
<none>               281.56 88.085
+ SAT    1   14.5426 267.02 89.681
+ CA     1   14.2158 267.35 89.728
+ C.I.C  1    9.2615 272.30 90.445
+ Mg     1    4.9066 276.65 91.063
+ PH     1    2.7509 278.81 91.366
+ Na     1    1.6686 279.89 91.517
+ K      1    0.9335 280.63 91.619
+ MO     1    0.4033 281.16 91.693
```

```
Call:
lm(formula = spc_rich ~ C.E + P, data = TENO)

Coefficients:
(Intercept)          C.E            P
15.57031     -9.21101     -0.09817
```

3. ```
   Direction:  backward
   Criterion:  AIC

   Start:  AIC=91.2
   spc_rich ~ C.E + C.I.C + CA + K + Mg + MO + Na + P + PH + SAT

   Df Sum of Sq    RSS     AIC
   - MO     1     0.384 230.35  89.266
   - SAT    1     0.443 230.41  89.276
   <none>               229.97  91.200
   - Na     1    22.077 252.04  92.775
   - K      1    25.612 255.58  93.319
   - PH     1    26.919 256.89  93.518
   - Mg     1    27.457 257.42  93.599
   - C.I.C  1    28.471 258.44  93.753
   - CA     1    30.200 260.17  94.013
   - P      1    43.170 273.14  95.910
   - C.E    1    76.163 306.13 100.357

   Step:  AIC=89.27
   spc_rich ~ C.E + C.I.C + CA + K + Mg + Na + P + PH + SAT

   Df Sum of Sq    RSS    AIC
   - SAT    1     0.795 231.15 87.400
   <none>               230.35 89.266
   - Na     1    21.767 252.12 90.787
   - K      1    25.354 255.71 91.338
   - Mg     1    27.182 257.53 91.616
   - C.I.C  1    28.350 258.70 91.792
   - PH     1    28.707 259.06 91.846
   - CA     1    30.402 260.75 92.100
   - P      1    50.399 280.75 94.982
   - C.E    1    75.780 306.13 98.358

   Step:  AIC=87.4
   spc_rich ~ C.E + C.I.C + CA + K + Mg + Na + P + PH
   ```
   ```

```
Df Sum of Sq    RSS    AIC
<none>               231.15 87.400
- Na     1    22.987 254.13 89.098
- K      1    31.849 263.00 90.434
- Mg     1    35.079 266.23 90.910
- PH     1    35.147 266.29 90.920
- C.I.C  1    35.286 266.43 90.941
- CA     1    38.094 269.24 91.350
- P      1    49.605 280.75 92.982
- C.E    1    89.742 320.89 98.194

Call:
lm(formula = spc_rich ~ C.E + C.I.C + CA + K + Mg + Na + P +
PH, data = TENO)

Coefficients:
(Intercept)          C.E        C.I.C           CA            K
121.2593     -12.2755      -4.2630       5.1893       3.7443
Mg           Na            P           PH
4.4594       3.6969      -0.1433     -15.9250
```

4. ```
   Direction:  forward
   Criterion:  AIC

   Start:  AIC=90.14
   spc_rich ~ 1

   Df Sum of Sq    RSS    AIC
   + C.E    1    51.578 322.17 86.349
   <none>               373.74 90.140
   + SAT    1    17.502 356.24 90.270
   + P      1    14.121 359.62 90.638
   + PH     1    13.440 360.30 90.712
   + Na     1     3.088 370.66 91.817
   + K      1     2.172 371.57 91.913
   + CA     1     1.491 372.25 91.984
   + C.I.C  1     0.029 373.71 92.137
   + MO     1     0.025 373.72 92.138
   + Mg     1     0.001 373.74 92.140

   Step:  AIC=86.35
   spc_rich ~ C.E
   ```
   ```

```
Df Sum of Sq    RSS    AIC
+ P      1    40.605 281.56 83.095
+ Mg     1    18.889 303.28 85.992
+ C.I.C  1    18.877 303.29 85.994
+ CA     1    16.540 305.63 86.293
<none>               322.17 86.349
+ SAT    1    11.552 310.61 86.925
+ PH     1    10.986 311.18 86.996
+ Na     1     1.707 320.46 88.142
+ K      1     1.468 320.70 88.171
+ MO     1     0.039 322.13 88.344

Step:  AIC=83.09
spc_rich ~ C.E + P

Df Sum of Sq    RSS    AIC
+ SAT    1   14.5426 267.02 83.026
+ CA     1   14.2158 267.35 83.074
<none>               281.56 83.095
+ C.I.C  1    9.2615 272.30 83.790
+ Mg     1    4.9066 276.65 84.409
+ PH     1    2.7509 278.81 84.712
+ Na     1    1.6686 279.89 84.863
+ K      1    0.9335 280.63 84.965
+ MO     1    0.4033 281.16 85.039

Step:  AIC=83.03
spc_rich ~ C.E + P + SAT

Df Sum of Sq    RSS    AIC
<none>               267.02 83.026
+ CA     1    5.8938 261.12 84.156
+ Na     1    2.3601 264.66 84.680
+ C.I.C  1    1.9481 265.07 84.741
+ PH     1    0.9884 266.03 84.882
+ K      1    0.9032 266.12 84.894
+ Mg     1    0.1679 266.85 85.002
+ MO     1    0.0099 267.01 85.025

Call:
lm(formula = spc_rich ~ C.E + P + SAT, data = TENO)

Coefficients:
```

```
(Intercept)          C.E            P          SAT
12.13535     -8.86519     -0.10193      0.06046
```

Al realizar los métodos de selección de variables utilizando los criterio BIC y AIC, el resultado obtenido es:

$$\textbf{(1) BIC} \Rightarrow \begin{cases} (1)\ BACWARD:\ spc\_rich \sim C.E + C.I.C + CA + K + Mg + Na + \\ \qquad\qquad + P + PH \\ (2)\ FORWARD:\ spc\_rich \sim C.E + P \end{cases}$$

$$\textbf{(2) AIC} \Rightarrow \begin{cases} (1)\ BACWARD:\ spc\_rich \sim C.E + C.I.C + CA + K + Mg + Na + \\ \qquad\qquad + P + PH \\ (2)\ FORWARD:\ spc\_rich \sim C.E + P + SAT \end{cases}$$

De los dos modelos obtenido para cada criterio, el de mínimo BIC es el modelo (1.2) y el de mínimo AIC es el modelo (2.2).

4.7.4. Bondad de ajuste.

Hemos de elegir un criterio para elegir el mejor subconjunto de regresión. Como el modelo inicial estaba constituido de una variable dependiente y de diez variables independientes elegiremos el criterio de información Bayesiano (BIC), ya que este penaliza más los modelos con un número mayor de parámetros estimados. Además de que nos encontramos ante un tamaño muestral pequeño.

Lo cual, por lo visto en el apartado anterior, según el BIC el mejor modelo de ajuste serie el siguiente:

$$spc\_rich \sim C.E + P$$

4.7.5. Modelo predicho.

En este caso debemos de repetir el proceso para obtener los coeficientes del modelo obtenido. Seguimos la ruta:

Estadísticos → Ajuste de modelos → Regresión lineal.

donde la sentencia R es:

- *RegModel.1 <- lm(spc_rich ~ C.E + P, data=TENO)*
- *summary(RegModel.1)*

y cuyo resultado es:

```
Call:
lm(formula = spc_rich ~ C.E + P, data = TENO)

Residuals:
Min     1Q Median     3Q    Max
-6.896 -1.304 -0.392  1.398  5.515

Coefficients:
Estimate Std. Error t value    Pr(>|t|)
(Intercept) 15.57031    2.40375   6.478 0.000000161 ***
C.E         -9.21101    2.91557  -3.159      0.0032 **
P           -0.09817    0.04308  -2.279      0.0287 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.797 on 36 degrees of freedom
Multiple R-squared:  0.2466,Adjusted R-squared:  0.2048
F-statistic: 5.893 on 2 and 36 DF,  p-value: 0.006109
```

El error residual estandarizado es: 2.797, el R^2: 0.2466, el R^2 ajustado: 0.2048, el F-test: 5.893 y su p-valor: 0.006.

Como se observa el p-valor es 0.006 por tanto es menor que alfa igual a 0.01 y 0.05 y por tanto se rechaza H_0 por tanto el modelo es significativo.

El modelo predicho que se ha obtenido es el siguiente:

spc_rich= 15.57 - 9.21 CE - 0.098 P

4.7.6. Diagnóstico de las hipótesis asociadas al modelo.

Para que el modelo tenga sentido y podamos utilizarlo, este debe de cumplir una serie de hipótesis asociadas al modelo. Las hipótesis asociadas al modelo de regresión lineal múltiple son:

1. La varianza de Y es constante. (Homocedasticidad)
2. Las observaciones Y son independientes entre si. (Autocorrelación)
3. La distribución de Y es normal. (Normalidad de la variable dependiente)
4. Multicolinealidad

Normalidad.

Una de las hipótesis mas rígidas del modelo es la normalidad del error o de la variable dependiente. Una de las pruebas mas utilizadas para diagnosticar

normalidad de una variable es la prueba de Kolmogorov-Smirnov, la cual realiza el siguiente test de hipótesis:

$$\begin{cases} H_0 : Y \approx NORMAL \\ H_1 : Y \neq NORMAL \end{cases}$$

La ruta que debemos seguir en Rcmdr es:

Estadísticos → Resúmenes → Test de normalidad.

La sentencia de R es:

- *normalityTest(˜ spc_rich, test="lillie.test", data=TENO)*

y el resultado de la sentencia es el siguiente:

```
Lilliefors (Kolmogorov-Smirnov) normality test

data:  spc_rich
D = 0.12905, p-value = 0.09994
```

Como podemos observar en el p-valor obtenido del test es 0.0999 por tanto mayor que 0.01 y 0.05, por lo que no existen evidencias para rechazar H_0. Por lo tanto se puede concluir con que la variable dependiente sigue una distribución normal.

Linealidad y homocedasticidad.

Para estudiar la homocdedasticidad se realizara el test de hipótesis de Box&Cox:

$$\begin{cases} H_0 : \lambda = 1 \\ H_1 : \lambda \neq 1 \end{cases}$$

En nuestro caso, para diagnosticar homocedasticidad, debemos obtener un p-valor mayor que 0.01 y 0.05.

La ruta que debemos seguir en Rcmdr es:

Modelos → Diagnósticos numéricos → Transformaciones de las respuesta.

La sentencia de R es:

- *summary(powerTransform(RegModel.1, family="bcPower"))*

y el resultado de la sentencia es el siguiente:

```
bcPower Transformation to Normality
Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1     0.896           1        0.384        1.408

Likelihood ratio test that transformation parameter is equal to 0
```

```
(log transformation)
LRT df      pval
LR test, lambda = (0) 14.44782  1 0.0001441

Likelihood ratio test that no transformation is needed
LRT  df    pval
LR test, lambda = (1) 0.1544995  1 0.69427
```

Como se observa en el LR test, lambda = (1), el p-valor es 0.69 por tanto no existen evidencias para rechazar H_0, por lo tanto nuestro modelo es homocedástico. En el caso que fuese heterocedástico tendríamos que realizar la transformación de Box&Cox.

Autocorrelación.

Debemos conocer también la independencia de los errores, autocorrelación. Para ello observaremos la gráfica de los residuos de la estimación en el tiempo, la cual nos permitirá tener una estimación de si los errores son independiente. Y realizaremos el test de Durbin-Watson que determina la independencia de errores.

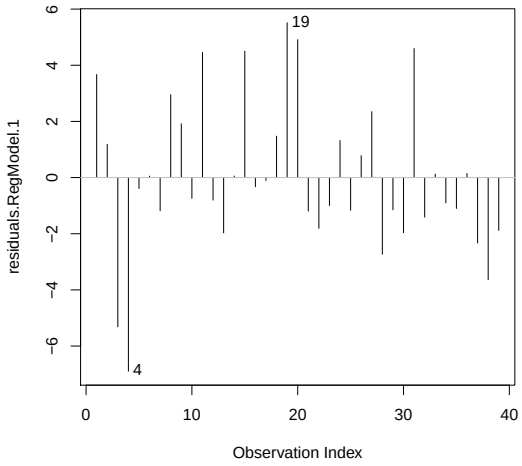

Figura 4.2. gráfica de los residuos de la estimación en el tiempo

En la gráfica no se observa ningún patrón por lo que podemos intuir que los errores son independientes.

El test de D-W realiza el siguiente test de hipótesis:

$$\begin{cases} H_0 : \rho = 0 \quad [ErroresIndependientes] \\ H_1 : \rho \neq 0 \quad [ErroresDependientes] \end{cases}$$

La ruta que debemos seguir en Rcmdr es:

Modelos → Diagnósticos numéricos → Test de Durbin-Watson para autocorrelación.

La sentencia de R es:

- *dwtest(spc_rich C.E + P, alternative="two.sided", data=TENO)*

y el resultado de la sentencia es el siguiente:

```
Durbin-Watson test

data:  spc_rich ~ C.E + P
DW = 1.5014, p-value = 0.07729
alternative hypothesis: true autocorrelation is not 0
```

Lo observado en la sentencia de Rcommander es que el p-valor es: 0.077,por tanto mayor que 0,01 y 0,05. Entonces no existen evidencias para rechazar H_0. De esta manera, nuestro modelo los errores son independiente.

Multicolinealidad.

La multicolinealidad de las variables se determina si los valores de factores de inflación de varianza (VIF) son altos (por encima de 10), esto indicara multicolinealidad en el modelo. Para resolver dicho problemas debemos realizar de nuevo una selección de variables.

La ruta que debemos seguir en Rcmdr es:

Modelos → Diagnósticos numéricos → Factores de inflación de varianza.

La sentencia de R es:

- *vif(RegModel.1)*

y el resultado de la sentencia es el siguiente:

```
C.E        P
1.112486 1.112486
```

Pero observamos que en este caso no existe multicolinealidad entre las variables, ya que los valores de la VIF son pequeños.

4.7.7. Diagnóstico de las observaciones anómalas.

Outliers, Alto potencia, Influyentes.

Como hemos visto en el capitulo 2, existen ciertas observaciones en la muestra que pueden interferir en los resultados de nuestro modelo. En Rcommander se realiza una gráfica donde se proporciona diagramas de índice de las distancias de Cook (" Cook") el cual nos indica datos influyentes en el modelo, los apalancamientos ("hat") que nos indica las observaciones con un alto potencial y los residuos de Student ("Studentized") en cual determina los outliers.

La ruta que debemos seguir en Rcmdr es:

Modelos → Gráficas → Gráfica de índice de influencia.

La sentencia de R es:

- *influenceIndexPlot(RegModel.1, id=list(method=" y", n=2), vars=c(" Cook", "Studentized", "hat"))*

y el resultado de la sentencia es la siguiente gráfica:

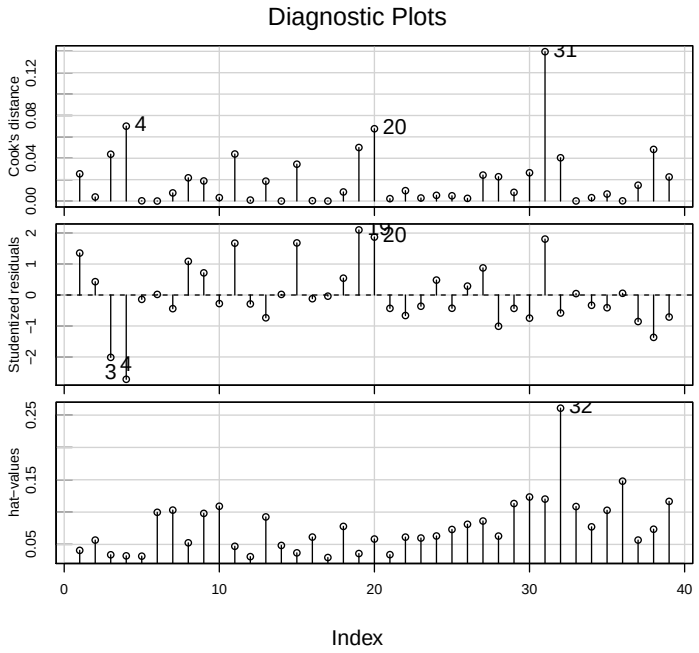

Figura 4.3. Gráfica de índice de influencia.

Como podemos observar existen varias observaciones que interfieren negativamente en los resultados obtenidos en el problema.

- Las observaciones consideradas **outliers** son la 3, la 4 y la 20.
- Las observaciones consideradas **alto potencia** es la 32.
- Las observaciones consideradas **influyentes** son la 4, la 20 y la 31.

Para subsanar la influencia de dichas observaciones anómalas existen diversas soluciones, entre ellas, rectificar errores en los datos, eliminar observaciones, usar estimadores alternativos, modificar el modelo, recoger mas datos, etc.

4.8. Conlclusión, Valoración personal y Acciones futuras..

La aplicación de los conocimientos aprendidos y estudiados en este trabajo, en el paquete informático Rcommander y Rstudio, ha sido, cuanto menos, laborioso. Hemos tenido que realizar una comparativa de los resultados obtenidos con diferentes paquetes como por ejemplo SPSS, ya que al tratarse de un Sofware Libre, debíamos asegurarnos de que los paquetes que utilizaban, verdaderamente realizaba lo que nosotros deseábamos. Verificando todos los paquetes que ibamos a utilizar, estudiando el codigo R intero, etc, llegamos a la conclusión que Rcommander era una herramienta que sí nos serviría para el desarrollo de nuestro trabajo. Decidimos realizarlo en Rcommander ya que se trata de un software de libre acceso, del que cualquiera puede usarlo. En mi opinión, Rcommander se trata de una herramienta muy útil, fácil de aprender y de usar, la cual te permite desarrollar infinidad de cálculos estadísticos.

Una vez pasamos a la fase de calcular todo lo necesario para nuestra implementación de la regresión lineal múltiple en el paquete informático Rcommander, donde hemos querido ajustar un modelo de regresión lineal múltiple relacionando la variable dependiente spc_rich en función de las variables : MO: % materia orgánica, Ca, Na, Mg, K, P,C.I.C: complejo de intercambio cationes, pH, C.E: intercambio de cationes y Sat: % saturación del suelo, hemos desarrollado los métodos de selección de variable bajo los criterios AIC y BIC, hemos diagnosticados las hipótesis asociadas al modelo y diagnosticado las observaciones anómalas, hemos concluido que a pesar de que todos los test para verificar cada hipótesis han dado correctamente, de que hemos ajustado un modelo significativo, etc, no considero que sea el mejor modelo que se ajustaría en la realidad, dado que el tamaño muestral del que se disponía era pequeño. Este, se podría considerar un estudio piloto del cual podemos sacar conclusiones positivas pero creo que con un aumento del tamaño muestral, podríamos considerar otros modelos ya que el criterio de información de Akaike funciona mejor para muestras grande. Es este uno de los motivos por lo que nos decantamos por el mejor

modelo ajustado por BIC. Además podríamos elaborar o encontrar un paquete en Rcommander o Rstudio, que calculase el criterio de información de Akaike corregido, visto en el capitulo 2. Este criterio ajusta modelos mejores que BIC y AIC para muestras pequeñas.

Por último decir que en una futura ampliación del estudio empezado por el profesor J.R. Arévalo, aparte de ampliar el tamaño muestral, podríamos aplicar transformaciones en las variables que permitan obtener resultados diferentes que puedan mejorar nuestra estimación. Es importante conocer cuantos modelos podemos ajustar, de que manera y con cuantas variables, ya que al utilizar un criterio podemos elegir el mejor conjunto de variables que ajusta nuestro modelo.

Bibliografía

[1] J.R. Arevalo, E. Chinea, and E. Barquin. Pasture management under goat grazing on canary islands. *Agriculture, Ecosystems and Environment*, 118(1):291 – 296, 2007.

[2] A.C. Atkinson. *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford science publications. Clarendon Press, 1987.

[3] Francisco Félix Caballero Díaz et al. *Selección de modelos mediante criterios de información en análisis factorial. Aspectos teóricos y computacionales*. Granada: Universidad de Granada, 2011.

[4] R.J. Carroll. *Transformation and Weighting in Regression*. CRC Press, 2017.

[5] S. Chatterjee and A.S. Hadi. *Sensitivity Analysis in Linear Regression*. Wiley Series in Probability and Statistics. Wiley, 1988.

[6] N.R. Draper and H. Smith. *Applied Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, 2014.

[7] Paul A. Murtaugh. In defense of p values. *Ecology*, 95(3):611–617.

[8] María Mercedes Suárez Rancel and Miguel Ángel González Sierra. *Análisis de regresión múltiple: teoría, métodos y aplicaciones*. 1999.

[9] Matthew R. E. Symonds and Adnan Moussalli. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using akaike's information criterion. *Behavioral Ecology and Sociobiology*, 65(1):13–21, Jan 2011.

The Best Linear Regression Model

Sección de Matemáticas
Universidad de La Laguna

Eduardo Hernández Córdoba
Facultad de Ciencias · Sección de Matemáticas
Universidad de La Laguna
alu0100884562@ull.edu.es

Abstract

This work performs a multiple linear regression study implemented in the free Software Rcommander and Rstudio in order to find the best linear regression model. Firstly multiple linear regression, the estimation of the parameters and the different criteria used to compare models are studied. In addition, the algorithms of the variables selection methods are studied and an extension of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) is made. Afterwards, the sensitivity analysis of the multiple linear regression is developed, where the validation of the model and the associated hypotheses and the effect of the anomalous observations are studied. Finally, the implementation is carried out in Rcommander and Rstudio.

1. Statistical analysis of linear regression.

The linear regression tries to find a linear relation between a set of independent variables $(X_0, X_1, X_2, ..., X_{k-1})$ and a dependent variable (Y). The most basic example is simple linear regression, which relates the dependent variable to an independent variable. The model that describes the simple linear regression is the following one:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon.$$

Since the previous model is very basic we study the multiple linear regression model. The model is defined as follows:

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{k-1} X_{k-1} + \varepsilon.$$

The multiple linear regression model must satisfy a serie of hypotheses. The hypotheses are the following: Normality, Homoscedasticity, Autocorrelation and Multicollinearity.
By taking a sample of n observations of n values of the independent variables $X_0, X_1, X_2, ..., X_{k-1}$, and observing the values belonging to $Y_1, Y_2, ..., Y_n$, then you have:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{k-1} X_{ik-1} + \varepsilon_i \qquad \forall i = 1, ..., n$$

where X_{ij} is the ith value of the variable X_j. These equations can be expressed more compactly as follows:

$$Y = X\beta + \varepsilon.$$

where Y order matrix (n x 1), X order matrix (n x k), β (k x 1), ε (k x 1).
In order to define the regression equation, it is necessary to calculate the parameters. We will use two methods: least squares and maximum likelihood.

· **Estimation of β and σ^2 by the method of least squares:** The estimator of β is: $\hat{\beta} = (X^t X)^{-1} X^t Y$. This method does not provide a σ^2 estimator, so the following is proposed: $S^2 = \frac{(Y-X\hat{\beta})^t (Y-X\hat{\beta})}{n-k} = \frac{Y^t (I - X(X^t X)^{-1} X^t) Y}{n-k}$.

· **Estimation of β and σ^2 by the principle of maximum likelihood:** The estimator of β is: $\hat{\beta} = (X^t X)^{-1} X^t Y$. an the σ^2 estimator is: $\hat{\sigma}^2 = \frac{(Y-X\hat{\beta})^t (Y-X\hat{\beta})}{n}$.
To know if the estimate of the model is significant, we will perform the following hypothesis test:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = ... = \beta_n = 0 \\ H_1 : Existe\ al\ menos\ un\ \beta_i \neq 0 \end{cases}$$

The statistic used for this contrast is the F-test, which is defined as follows:

$$F = \frac{SS_{reg}/k}{SSR/(n-k-1)} = \frac{(SSY - SSR)/k}{SSR/(n-k-1)} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 / k}{\sum_{i=1}^{n} (y_i - \hat{y})^2 / (n-k-1)}$$

2. Selection of Models.

There are statistics that allow us to compare models with different numbers and different variables, of which we will study the following:

- Coefficient of determination or multiple correlation:
  $R^2 = \frac{SSR_{reg}}{SSY} = 1 - \frac{SSR}{SSY} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$
- Coefficient of determination adjusted or corrected:
  $R_a^2 = 1 - \frac{Varianza\ Residual}{Varianza\ de\ Y} = 1 - \frac{(SSR)/(n-k)}{(SSY)/n} = 1 - (1 - R^2)\left(\frac{n}{n-k}\right)$
- Akaike information criterion (AIC):
  $AIC(k) = -2 \ln L[\hat{\Theta}(k)] + 2k$
- Bayesian information criterion (BIC):
  $BIC(k, n) = -2 \ln L[\hat{\Theta}(k)] + k \ln n$

We will study three basic algorithms for the selection of models:

- Forward Selection, Backward Selection and Stepwise Selection

3. Sensitivity analysis in linear regression.

The regression model requires the fulfillment of certain conditions and hypotheses to use it.

- Normal distribution of the dependent variable.
- Homocedasticity and linearity.
- Autocorrelation, Independence errors.
- Multicollinearity.

Our goal is to find a regression model that fits the real data as much as possible, but this can be affected by certain observations that influence the model, which are collected in the sample. It is necessary to know how to identify these possible observations, we will study three types: Outliers, High Power and Influential.

4. Multiple linear regression in Rcommander and Rstudio.

RegModel.1 <- lm(spc_rich ~ C.E + P, data=TENO)
summary(RegModel.1)

```
Call:
lm(formula = spc_rich ~ C.E + P, data = TENO)

Residuals:
Min      1Q Median      3Q     Max
-6.896 -1.304 -0.392  1.398  5.515

Coefficients:
Estimate Std. Error t value     Pr(>|t|)
(Intercept) 15.57031    2.40375   6.478 0.000000161 ***
C.E         -9.21101    2.91557  -3.159       0.0032 **
P           -0.09817    0.04308  -2.279       0.0287 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.797 on 36 degrees of freedom
Multiple R-squared:  0.2466,Adjusted R-squared:  0.2048
F-statistic: 5.893 on 2 and 36 DF,  p-value: 0.006109
```

References

[1] J.R. Arevalo, E. Chinea, and E. Barquin. Pasture management under goat grazing on canary islands. *Agriculture, Ecosystems and Environment*, 118(1):291 – 296, 2007.

[2] María Mercedes Suárez Rancel and Miguel Ángel González Sierra. *Análisis de regresión múltiple: teoría, métodos y aplicaciones*. 1999.