## Análisis de regresión en R

El lenguaje R permite la elaboración de análisis de regresión, además de formulación de diferentes tipos de pruebas para observar el cumplimiento o no de los supuestos que conlleva este tipo de análisis, a continuación se presentará la forma por medio de la cual se realizan en modelos simple y múltiple.

## Análisis de regresión simple

Dentro de los análisis de regresión que se formulan en el programa R es muy utilizada la función *lm()*, esta hace referencia a *linear model* (por sus siglas en inglés), para el desarrollo de la presente sección se utilizará la librería faraway, dentro de la cual se llamará el conjunto de datos gala, este archivo presenta la información de la diversidad de pájaros que hay en la las Islas Galapagos; entre las variables que se pueden encontrar dentro de este archivo están:

* + - *Species*: Número de especies de pájaros que se pueden encontrar en la isla.
    - *Endemics*: Cantidad de especies endémicas (las especies endémicas son aquellas que solo pueden ubicarse en una región específica)
    - *Area*: Se refiere al área en kilómetros cuadrados de la isla
    - *Elevation*: La mayor altura de la isla en metros
    - *Nearest:* Incorpora la distancia que hay la isla más cercana en Kilómetros
    - *Scruz*: Distancia que se tiene hasta la isla de Santa Cruz
    - *Adjacent*: Área de la isla que adyacente en kilómetro cuadrados

Con respecto a estas variables se realizará en primera instancia un análisis de regresión simple en el cual se utilizará como variable dependiente la cantidad de especies que hay en la isla, como variable explicativa se utilizará la cantidad de especies endémicas que hay en la isla, a continuación se presentará el programa para poder obtener las salidas en R:

library(faraway)

data(gala)

names(gala)

[1] "Species" "Endemics" "Area" "Elevation" "Nearest" "Scruz" "Adjacent"

modelo<-lm(gala$Species~gala$Endemics)

modelo

Call:

lm(formula = gala$Species ~ gala$Endemics)

Coefficients:

(Intercept) gala$Endemics

-21.048 4.072

summary(modelo)

Call:

lm(formula = gala$Species ~ gala$Endemics)

Residuals:

Min 1Q Median 3Q Max

-71.791 -15.894 3.507 12.088 78.200

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -21.0480 7.1138 -2.959 0.00622 \*\*

gala$Endemics 4.0721 0.1899 21.443 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.95 on 28 degrees of freedom

Multiple R-squared: 0.9426, Adjusted R-squared: 0.9406

F-statistic: 459.8 on 1 and 28 DF, p-value: < 2.2e-16

Para la obtener los gráficos correspondientes para análisis de residuales es necesario pedir a R que agregue los residuos dentro del archivo que se maneja; para realizar este procedimiento se deben utilizar los siguientes comandos:

gala$fitted <- fitted(modelo)

gala$residuals <- residuals(modelo)

gala$rstudent <- rstudent(modelo)

gala$cooks <- cooks.distance(modelo)

De esta forma se integran dentro del archivo gala los valores correspondientes a; residuales (residuals) distancias de cook (cooks.distance), r estudentizados (rstudent), y los valores estimados (fitted), para el para el gráfico se usarán los r estudentizados dado que se encuentran en su forma estandarizada; a continuación se presenta un programa por medio del cual se puede ver cuales son identificar cuales son los datos que presentan valores extremos, para la utilización de este comando es necesario llamar la librería car.

library(car)

scatterplot(gala$rstudent~gala$Endemics, ylab="rstudent", xlab="Endemics", labels=rownames(gala), smooth=FALSE, boxplot=TRUE, reg.line= TRUE)

abline(h=2.1, v=0, col="gray", lty=14)

abline(h=-1.6, v=0, col="gray", lty=14)

Con éste programa se obtiene un gráfico como el siguiente:



Otra de las formas por medio de las cuales se pueden diseñar gráficos para el control de residuos es a partir del siguiente comando:

influencePlot(Modelo, labels=row.names(gala))



Además se pueden realizar los gráficos para ver la normalidad que presentan los datos, para esto se realiza la siguiente función:

qq.plot(RegModel.1, simulate=TRUE, labels=names(residuals(Modelo)))

Por medio del cual se puede obtener un gráfico como el siguiente:



O con un comando como el siguiente se puede obtener un gráfico para ver la normalidad que se puede encontrar entre el conjunto de datos; las líneas serían de la siguiente forma:

cr.plots(Modelo, ask=FALSE)

Con lo cual se puede generar un gráfico para normalidad, es importante mencionar que este gráfico no puede emitir ningún tipo de etiqueta con la cual se busquen los valores que influyan; el gráfico obtenido es el siguiente:



Para un análisis más sencillo en R se pueden obtener de forma gráfica una serie de resultados sobre el contexto de regresión, por lo que se presentará una explicación de estos y el gráfico que se puede obtener:

oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))

plot(modelo)

par(oldpar)

Con el programa que se acaba de estructurar, el R ofrece un conjunto de gráficos en los cuales se obtienen los residuales contra los predichos, gráfico de normalidad Q-Q, un gráfico con la raíz de los residuales contra los predichos y un gráfico en el cual se evalúan los hat-values contra los residuos estandarizados. A continuación se presenta la forma gráfica por medio de la cual se obtendrían estos resultados:



## Regresión Múltiple

En el siguiente análisis se tomará en cuenta el archivo correspondiente a gala pero para este caso se meterán además como variables explicativas, Area y Elevation, a continuación se presenta como se obtienen los resultados en un modelo múltiple:

library(faraway)

data(gala)

names(gala)

modelo<-lm(gala$Species~gala$Endemics + gala$Area + gala$Elevation)

modelo

Call:

lm(formula = gala$Species ~ gala$Endemics + gala$Area + gala$Elevation)

Coefficients:

(Intercept) gala$Endemics gala$Area gala$Elevation

-15.89124 4.33179 0.01267 -0.04144

summary(modelo)

Call:

lm(formula = gala$Species ~ gala$Endemics + gala$Area + gala$Elevation)

Residuals:

Min 1Q Median 3Q Max

-67.691 -10.530 2.387 10.529 72.723

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -15.891237 7.569210 -2.099 0.0456 \*

gala$Endemics 4.331791 0.304686 14.217 8.97e-14 \*\*\*

gala$Area 0.012669 0.008936 1.418 0.1681

gala$Elevation -0.041439 0.023653 -1.752 0.0916 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.29 on 26 degrees of freedom

Multiple R-squared: 0.9492, Adjusted R-squared: 0.9433

F-statistic: 161.8 on 3 and 26 DF, p-value: < 2.2e-16

influencePlot(modelo, labels=row.names(gala))



qq.plot(modelo, simulate=TRUE, labels=row.names(gala))



cr.plots(modelo, ask=FALSE)



oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))

plot(modelo)

par(oldpar)



Por medio de lo cual se puede decir que se han incorporado las técnicas básicas con las cuales se pueden realizar análisis de regresión, además de algunos de los resultados que se pueden obtener al igual que las herramientas gráficas que facilitan la interpretación dentro de este programa.

## Análisis de Diseños Experimentales

A continuación se hará una introducción básica a la forma por medio de la cual se realizan los pueden realizar diseños experimentales, así como la forma por medio de la cual se pueden obtener los resultados pertinentes para realizar algún tipo de análisis.

## Análisis de factores

Entre las técnicas utilizadas para el desarrollo de los diseños experimentales se pueden encontrar los análisis de factores, los cuales buscan ver si el incorporar o no un tratamiento influye significativa en una aumento de producción, a continuación se presenta un conjunto de datos con los cuales se aplicará dicha técnica, los factores a utilizar dentro del análisis son los siguientes:

K=Potasio

N=Nitrógeno

D= Compuesto Orgánico

A continuación se presenta el conjunto de datos con el cual se pretende realizar el análisis factorial:



El archivo que se presenta anteriormente se maneja en SPSS, por lo cual se levantará ese archivo hacia R para poder aplicar los análisis respectivos. A continuación se presenta el programa con el cual se pueden obtener los resultados para realizar un diseño factorial para analizar un experimento:

library(foreign)

nkd <- read.spss("Analisis factores.sav", to.data.frame=TRUE)

factorial <- lm(PRODUCCI ~ D +K +N +D \*K +D \*N +K \*N +D \*K \*N , data=nkd)

summary(factorial)

Call:

lm(formula = PRODUCCI ~ D + K + N + D \* K + D \* N + K \* N + D \*

K \* N, data = nkd)

Residuals:

Min 1Q Median 3Q Max

-28.750 -11.062 -1.125 6.938 33.250

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 106.250 9.164 11.594 2.54e-11 \*\*\*

D 214.500 12.960 16.551 1.25e-14 \*\*\*

K 173.250 12.960 13.368 1.30e-12 \*\*\*

N 0.250 12.960 0.019 0.984769

D:K -75.750 18.328 -4.133 0.000376 \*\*\*

D:N 28.000 18.328 1.528 0.139651

K:N 21.000 18.328 1.146 0.263173

D:K:N -15.750 25.919 -0.608 0.549124

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.33 on 24 degrees of freedom

Multiple R-squared: 0.9827, Adjusted R-squared: 0.9777

F-statistic: 195.1 on 7 and 24 DF, p-value: < 2.2e-16

anova(factorial)

Analysis of Variance Table

Response: PRODUCCI

Df Sum Sq Mean Sq F value Pr(>F)

D 1 278818 278818 830.0464 < 2.2e-16 \*\*\*

K 1 161170 161170 479.8066 < 2.2e-16 \*\*\*

N 1 3465 3465 10.3162 0.003732 \*\*

D:K 1 13986 13986 41.6375 1.132e-06 \*\*\*

D:N 1 810 810 2.4115 0.133537

K:N 1 345 345 1.0257 0.321278

D:K:N 1 124 124 0.3692 0.549124

Residuals 24 8062 336

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Con base en los resultados que se encontraron para los análisis factoriales se puede realizar los análisis que se consideren relevantes, dado que ya se conoce cual es la probabilidad asociada a alguna diferencia en el modelo anteriormente citado.

## Análisis de covarianza

El análisis de Covarianza o ANCOVA (por sus siglas en inglés Analisys of covariance), es una técnica estadística utilizada en procedimientos de regresión lineal múltiple, con la cual se pretende comparar los resultados que se obtienen a diferentes niveles de una variable categórica, en si es una fusión entre un ANOVA (por sus siglas en inglés Analisys of Variance) y una regresión lineal múltiple.

A continuación se presentará un código con el cual se crea un archivo de datos y luego es utilizado dicho conjunto para la realización de los análisis respectivos:

x<-c(79,65,57,51,57,66,44,41,44,36,62,73,58,43,50,44,48,51,40,38)

y<-c(1.96,1.77,1.62,1.76,1.88,1.50,1.60,1.49,1.77,1.27,1.22,1.39,1.28,1.28,1.45,1.22,1.31,1.57,1.21,1.06)

trat<-c("a","a","a","a","a","a","a","a","a","a","b","b","b","b","b","b","b","b","b","b")

covarianza<-data.frame(pesoini=x, ganancia=y, tratamientos=trat)

modelo<-lm(ganancia~pesoini+tratamientos,data=covarianza)

modelo

Call:

lm(formula = ganancia ~ pesoini + tratamientos, data = covarianza)

Coefficients:

(Intercept) pesoini tratamientos[T.b]

1.230843 0.007984 -0.336652

anova(modelo)

Analysis of Variance Table

Response: ganancia

Df Sum Sq Mean Sq F value Pr(>F)

pesoini 1 0.27408 0.27408 11.747 0.0032119 \*\*

tratamientos 1 0.55537 0.55537 23.803 0.0001413 \*\*\*

Residuals 17 0.39665 0.02333

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

De igual forma que en los análisis anteriores con estas líneas de comandos se obtiene la información necesaria para poder realizar los análisis pertinentes en cuanto a la interpretación de la información con la cual se cuenta y de que tan probable es que cada una de las variables causen diferencias en el análisis o cada uno de los tratamientos proporcione resultados diferentes.

## Cuadrado Latino

Un cuadrado latino es una matriz cuadrada (de n\*n elementos), la cual en cada una de sus casillas esta ocupada por uno de los n tratamientos que se aplicaron, de esta forma se cuenta con n filas, n columnas y n tratamientos, donde cada uno de estos tratamientos no se repite en ninguna fila o columna. A continuación se presenta lo que se conoce como un cuadrado Latino con 5 filas, 5 columnas y 5 tratamientos (A,B,C,D,E):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A | B | C | D | E |
| B | C | D | E | A |
| C | D | E | A | B |
| D | E | A | B | C |
| E | A | B | C | D |

Como se observa en cuadro anterior se tienen 5 filas y 5 columnas dentro de las cuales se aleatorizan los 5 tratamientos entre las filas y las columnas de manera que no se encuentre repetido en alguno de estos.

A continuación se presentará el programa con el cual se realiza este tipo de análisis:

aciertos<-c(1,2,1,2,3,2,1,1,3,2,3,2,2,2,1,3,2,1)

tiradores<-c("1","1","1","2","2","2","3","3","3","4","4","4","5","5","5","6","6","6")

blancos<-c("1","2","3","1","2","3","1","2","3","1","2","3","1","2","3","1","2","3")

cuadrado<-c("1","1","1","1","1","1","1","1","1","2","2","2","2","2","2","2","2","2")

trtamientos<-c("a","b","c","b","c","a","c","a","b","a","b","c","b","c","a","c","a","b")

cuadrado<-data.frame(y=aciertos, columnas=tiradores, hileras=blancos, replica=cuadrado, tratamientos=trtamientos)

cuadrado

y columnas hileras replica tratamientos

1 1 1 1 1 a

2 2 1 2 1 b

3 1 1 3 1 c

4 2 2 1 1 b

5 3 2 2 1 c

6 2 2 3 1 a

7 1 3 1 1 c

8 1 3 2 1 a

9 3 3 3 1 b

10 2 4 1 2 a

11 3 4 2 2 b

12 2 4 3 2 c

13 2 5 1 2 b

14 2 5 2 2 c

15 1 5 3 2 a

16 3 6 1 2 c

17 2 6 2 2 a

18 1 6 3 2 b

En la presentación del conjunto de datos que se utilizó para el análisis se encontró que hay 2 cuadrados, por lo cual también se unirá en la trabajo del archivo la replica con la cual se trabaja.

modelo<-lm(y~replica+columnas+hileras+tratamientos,data=cuadrado)

modelo

Call:

lm(formula = y ~ replica + columnas + hileras + tratamientos, data = cuadrado)

Coefficients:

(Intercept) replica[T.2] columnas[T.2] columnas[T.3]

0.8889 0.6667 1.0000 0.3333

columnas[T.4] columnas[T.5] columnas[T.6] hileras[T.2]

0.3333 -0.3333 NA 0.3333

hileras[T.3] tratamientos[T.b] tratamientos[T.c]

-0.1667 0.6667 0.5000

anova(modelo)

Analysis of Variance Table

Response: y

Df Sum Sq Mean Sq F value Pr(>F)

replica 1 0.2222 0.2222 0.3478 0.5716

columnas 4 2.2222 0.5556 0.8696 0.5220

hileras 2 0.7778 0.3889 0.6087 0.5675

tratamientos 2 1.4444 0.7222 1.1304 0.3695

Residuals 8 5.1111 0.6389