Universidad del Desarrollo

Control 1

Por: Jan Frese

Ramo: Taller de inteligencia de negocios
Profesor: Mauricio Herrera

Para la realización de este control, es importante mencionar que se trabajó con la base de datos “theatre.csv” indicada por el profesor. Para esto, se decidió borrar todas las variables registradas por el programa, junto con sus gráficos, para luego poder trabajar sin inconvenientes con la base de datos y con las variables que se van a definir a lo largo del control. Tras esto, se decidió eliminar la variable “X” de la base, ya que esta hacía referencia solamente al número de las columnas y eso era irrelevante para los análisis que se realizarán.

```
rm(list=ls()) # Para limpiar todas las variables
graphics.off() # Borra todos los gráficos
setwd("C:/Users/pc/Desktop/bases de datos csv y excel/")
data <- read.csv("theatre.csv")
data$X <- NULL
```

Inciso a:

Para este inciso se pedía construir un modelo de regresión lineal para explicar la variable “Theatre” en función del resto de las variables indicadas en la base de datos.

```
modelo <- lm(Theatre ~ ., data=data)
summary(modelo)
```

Tras construir el modelo, se realizó un análisis global de este modelo mediante el comando *summary()*. El análisis hace referencia a que las variables significativas para explicar el modelo son “Sex”, “Income” y “Culture”. Mientras que la variable “Age” es relevante para explicar el modelo, pero no es tan significativa como las variables mencionadas anteriormente. Es importante mencionar que la variable “Theatre_ly” no es significativa para explicar el modelo y se podría eliminar de este sin inconvenientes.
Además, la fórmula obtenida de esta regresión es:

Theatre = -127,22271 + 0.39757*Age + 22.22059*Sex + 1.34817*Income + 0.53664*Culture + 0.17191*Theatre_ly + E

En donde “E” corresponde al error del modelo.
Por otra parte, el “R^2” y el “R^2 ajustado” de este modelo es 0.2966 y 0.2915 respectivamente. Como ambos son menores a 0.3, se puede afirmar que el modelo abarca menos del 30% de la varianza de los datos y, por lo tanto, el modelo no representa un buen ajuste, en función de las variables que se analizan.

Inciso b:

Para este inciso se decidió trabajar con matrices, en donde se ocuparon la cantidad de observaciones, la matriz diseño (la cual correspondía a las variables “Age”, “Sex”, “Income”, “Culture” y “Theatre_ly”), la matriz de la variable respuesta (“Theatre”), sus coeficientes ajustados y las matrices de proyección “H” e “I”. Además, se determinó que la cantidad de parámetros estimados eran 6.
Estas matrices serán nuestra base para calcular los valores pedidos que entrega el comando *summary()* de manera manual.
Para esto se procedió a determinar la suma cuadrática de error (SCE) mediante las matrices de proyección y la matriz de la variable respuesta. Tras esto se calculó la varianza estimada de la variable “Theatre” en un formato vector, para así calcular la varianza estimada de los coeficientes estimados. Luego se calculó el “valor t” pedido mediante el cociente de los betas estimados y la desviación estimada de los coeficientes estimados. Como este cálculo contempla los valores t de todas las variables, se escogió el “valor t” de la variable “Age” que solicitaba el inciso. Continuamente se determinó el “valor p” del estadístico t calculado anteriormente mediante la cantidad de observaciones, los parámetros estimados y el valor de t.

```
n <- dim(data)[1]
x <- as.matrix(cbind(rep(1, n), data$Age, data$Sex, data$Income,
                     data$Culture, data$Theatre_ly)) # Matriz de diseño
y <- as.matrix(data$Theatre)
beta.gorro <- solve(t(x)%*%x)%*%t(x)%*%y
H <- x%*%solve(t(x)%*%x)%*%t(x)
I <- diag(1,n)
p <- 6 # parametros estimados
SCE <- t(y)%*%(I-H)%*%y
var.gorro.y <- c(SCE/(n-p))
sd.g.b.g <- sqrt(var.gorro.y * diag(solve(t(x)%*%x)))
t <- beta.gorro/sd.g.b.g
t[2,] #valor t para "Age"

pvalor <- 2*(1-pt(abs(t), n-p))
pvalor[2,] #pvalor para "Age"
```

Inciso c:

Para este inciso simplemente se ocupó el código entregado en el control, para luego analizar el gráfico determinado (**fig.0**).

plot(modelo,which = 2)

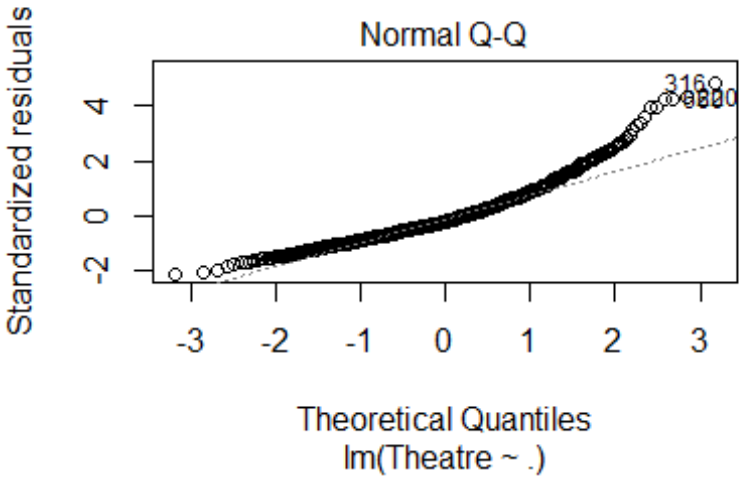

fig.0

Con relación al diagrama cuantil-cuantil del modelo (**fig.0**), se puede describir que, como varios datos se encuentran alineados según la recta del diagrama mostrado, podemos afirmar que los residuos estandarizados provienen de una distribución normal y, consecuentemente, también la regresión.

Inciso d:

Para este inciso se diseñó un histograma para los residuos del modelo (**fig.1**).

```
hist(residuals.lm(modelo), las=1, main = "Histograma de los residuos", xlab = "Residuos del modelo",
    ylab = "Frecuencia")
```

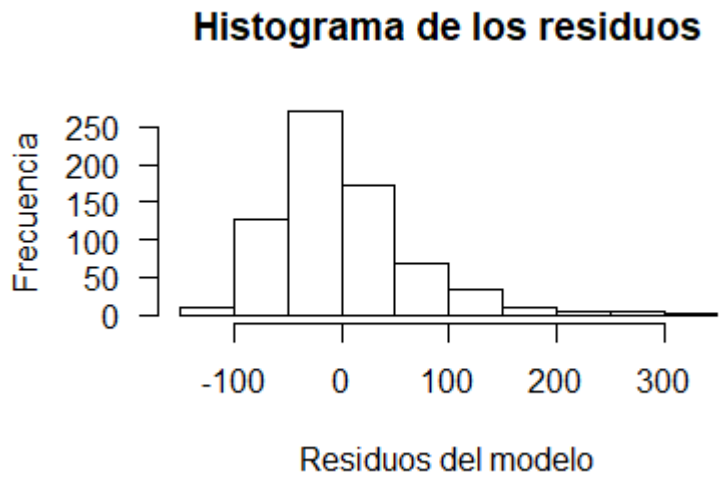

fig.1

Como en el histograma de los residuos del modelo representa un comportamiento cercano a una “campana” (centrado casi en el 0), podemos afirmar que los residuos de la regresión distribuyen normal.

Inciso e:

Para este inciso se calculó el logaritmo de todos los datos de la variable “Theatre” para luego realizar un nuevo modelo de regresión. Calculado esto, se procedió a eliminar la variable “Theatre”, para tener solamente la variable “log_Theatre” junto con las demás variables que explicarán en el modelo. “log_Theatre” corresponde a los valores obtenidos tras calcular el logaritmo de los datos de “Theatre”.

```
data$log_Theatre <- log(data$Theatre)
data$Theatre <- NULL
modelo1 <- lm(log_Theatre ~ ., data=data)
summary(modelo1)
```

El *summary()* de este modelo hace referencia a que la única variable que no es significativa en este modelo es “Theatre_ly”. Sin embargo, la variable “Age” aumentó su nivel de significancia en comparación al modelo que tenía los valores de “Theatre”.
Además, la fórmula de este modelo según este comando es:

log_Theatre = 2.9541546 + 0.0038690*Age + 0.1794468*Sex + 0.0087906*Income + 0.0035360*Culture + 0.0013492*Theatre_ly + E

En donde “E” corresponde al error del modelo.
Finalmente, cabe destacar que el “R^2” y el “R^2 ajustado” de este modelo es 0.32 y 0.315 respectivamente. Dado esto, se puede afirmar que el modelo abarca aproximadamente el 32% de la varianza de los datos y, por lo tanto, el modelo no representa un buen ajuste, en función de las variables que se analizan.

Inciso f:

Para este inciso se ocuparon dos métodos para confirmar las variables relevantes del modelo. Primero que todo se hizo el *summary()* del modelo. Tras esto, como segundo método, se seleccionaron las variables que explicaban a “log_Theatre” en el modelo. Luego, validamos los parámetros. Continuamente, se “entrena” el modelo para el caso de una regresión lineal. Finalmente, se estima la importancia de las variables en el modelo y se grafican (**fig.2**). Gracias al *summary()*, se puede confirmar que los resultados obtenidos son correctos.

```
summary(modelo1)
library(caret)
Teatro1 <- dplyr::select(data,Age,Sex,Income,Culture,Theatre_ly)
# introducimos los parámetros, usamos crossvalidation
control <- trainControl(method="repeatedcv", number=10, repeats=3)
# entrenamos el modelo, en este caso es una regresión lineal.
model <- train(log_Theatre~., data=data, method="lm", preProcess="scale", trControl=control)
# estimamos la importancia de las variables en el modelo
importancia <- varImp(model, scale=FALSE)
# resumen de la importancia
print(importancia)
# gráfico de la importancia
plot(importancia)
```

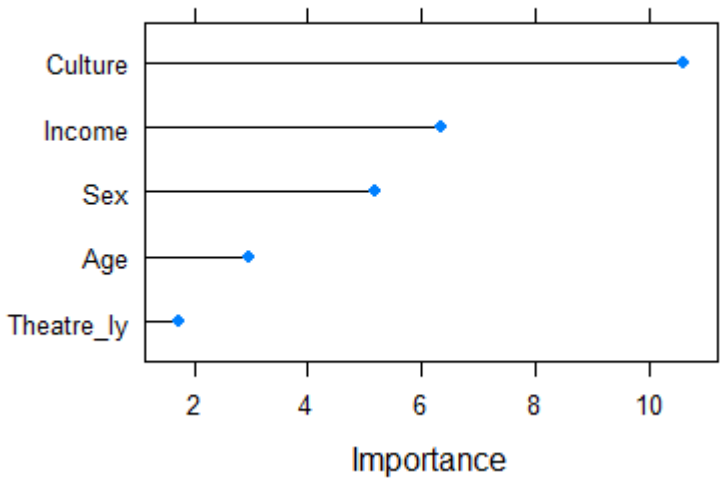

fig.2

Como se puede apreciar en el gráfico (**fig.2**), la variable “Culture” es la más relevante del modelo, mientras que la variable “Theatre_ly” es la menos relevante para explicar el modelo propuesto. Además, cabe mencionar que la variable “Culture” tiene mucha más importancia en comparación a la segunda variable más importante del modelo, la cual es “Income”. La variable “Income” y la variable “Sex” no poseen una amplia diferencia en cuanto a la importancia de las variables para el modelo.

Inciso g:

Para este inciso se establecieron los valores propuestos en el control, para luego crear el “dataframe” que servirá para estimar el gasto en visitas al teatro de un residente de la ciudad.

```
new_data=data.frame(Age = 38, Sex = 0, Income =59.3,
                    Culture = 187, Theatre_ly = 150)
predict(modelo1,new_data)
```

Como el valor obtenido de la predicción fue 4.486059, podemos afirmar que el gasto estimado en visitas al teatro de un residente será 4.49 euros mensuales aproximadamente.