R Forward MTCARS

Yosef Guevara Salamanca

26/11/2020

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(MASS)
library(nortest)
```

```
data("mtcars");
attach(mtcars);
```

Paso 1. Modelo nulo

Realizaremos el análisis del modelo nulo para extraer la suma de los errores cuadrados SCE para ser utilizada como medida comparativa para encontrar una covariable que explique mejor a "mpg" que el modelo nulo

```
#SLR modelo nulo

mt.fit0 <- lm(mpg ~ 1)
anova(mt.fit0)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 31   1126  36.324
```

```
SCE.0=anova(mt.fit0)[1,2]
cat("Suma de los cuadrados del error Modelo nulo: ", SCE.0)
```

```
## Suma de los cuadrados del error Modelo nulo:  1126.047
```

```r
y_hat <- mean(mpg);
cat("y_hat:", y_hat )
```

```
## y_hat: 20.09062
```

```r
# Calculo de la media

media <- y_hat
cat("La media es: ", media, "\n")
```

Calculo conceptual del SSRnulo

```
## La media es:  20.09062
```

```r
# Caculo del las Sumas de cuadrados de la regresion

SCR.nulo = sum ( (mpg-media)^2 )
cat("Las suma de cuadrados de la regresion es: ", SCR.nulo, "\n")
```

```
## Las suma de cuadrados de la regresion es:  1126.047
```

"Partial f Test" Se realiza el cálculo del Fcritico que nos servirá como medida comparativa para identifica un mejor modelo, para ellos se definen los siguientes parametros

- Alpha: un alpha input de alpha <- 0.01
- df1: Los grados de libertad dados por la suma de cuadrados de la regresion; (1)
- df2: Los grados de liberdad dados por la suma de cuadrados de error (n-2)

```r
# F critico modelo nulo para el 10%


alpha <- 0.05

f.critico1<-qf( 1- alpha , 1 , 50 )
cat("F_Critico modelo nulo: ", f.critico1 )
```

```
## F_Critico modelo nulo:  4.03431
```

TRES REGRESIONES LINEALES SIMPLES Se realiza el calculo de los modelos SLR para cada covariable

```r
# Regresiones lineales necesarias

mt.fit11<-lm(mpg~ wt)
mt.fit12<-lm(mpg~ disp)
mt.fit13<-lm(mpg~ hp)

print("mpg ~ wt ")
```

```
## [1] "mpg ~ wt "
```

```
anova(mt.fit11)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## wt         1 847.73  847.73  91.375 1.294e-10 ***
## Residuals 30 278.32    9.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print("mpg~ disp")
```

```
## [1] "mpg~ disp"
```

```
anova(mt.fit12)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value   Pr(>F)
## disp       1 808.89  808.89  76.513 9.38e-10 ***
## Residuals 30 317.16   10.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print("mpg~ hp")
```

```
## [1] "mpg~ hp"
```

```
anova(mt.fit13)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## hp         1 678.37  678.37   45.46 1.788e-07 ***
## Residuals 30 447.67   14.92
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

fparcial Se extrer la suma cuadrados de los residuales, el cuadrado medio del error de todos los SLR y del modelo nulo.

```
# SCE Modelo nulo

SCE.0<-anova(mt.fit0)[1,2]
```

```r
# SCE SLR
SCE.11<-anova(mt.fit11)[2,2]
SCE.12<-anova(mt.fit12)[2,2]
SCE.13<-anova(mt.fit13)[2,2]

# CME SLR

CME.11<-anova(mt.fit11)[2,3]
CME.12<-anova(mt.fit12)[2,3]
CME.13<-anova(mt.fit13)[2,3]

cbind(SCE.11,SCE.12,SCE.13,CME.11,CME.12,CME.13)
```

```
##        SCE.11   SCE.12   SCE.13   CME.11   CME.12   CME.13
## [1,] 278.3219 317.1587 447.6743 9.277398 10.57196 14.92248
```

Cálculos de fparciales Se realiza el calculo de los Fparciales

```r
fparcial.11=(SCE.0-SCE.11)/CME.11
fparcial.12=(SCE.0-SCE.12)/CME.12
fparcial.13=(SCE.0-SCE.13)/CME.13
cbind(f.critico1,fparcial.11, fparcial.12, fparcial.13)
```

```
##      f.critico1 fparcial.11 fparcial.12 fparcial.13
## [1,]    4.03431    91.37533    76.51266     45.4598
```

Gracias al anterior paso evidenciamos que el fparcial.11 (wt) es es mas grande para todas las covaraibles que pueden explicar a mpg, por ende el modelo dado por lm(mpg ~ wt) explica mejor a mpg que el modelo nulo.

PASO 2. Evaluacion de modelos con 2 covariables Para encontrar si es posible explicar de una mejor forma la variable de salida mpg mediante un modelo con 2 covariables calcularemos un segundo f.critico, bajo la siguiente premisa.

G.L SST = 50 - 1 G.L SSR = 2 parametros G.L SSE = 50 - 1 - 2

```r
f.critico2<- qf(1- alpha , 2 ,47 )
f.critico2
```

```
## [1] 3.195056
```

MODELOS 21 Y 22

con disp Se rocede a evaluar wt y disp como covariables

```r
mt.fit21<-lm(mpg ~ wt + disp)
anova(mt.fit21)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## wt         1 847.73  847.73 99.6586 6.861e-11 ***
## disp       1  31.64   31.64  3.7195   0.06362 .
## Residuals 29 246.68    8.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SCE.21<-anova(mt.fit21)[3,2]
CME.21<-anova(mt.fit21)[3,3]
```

con hp Se rocede a evaluar wt y hp como covariables

```
mt.fit22<-lm(mpg~ wt + hp)
SCE.22<-anova(mt.fit22)[3,2]
CME.22<-anova(mt.fit22)[3,3]
anova(mt.fit22)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## wt         1 847.73  847.73 126.041 4.488e-12 ***
## hp         1  83.27   83.27  12.381  0.001451 **
## Residuals 29 195.05    6.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cbind(SCE.21,CME.21,SCE.22,CME.22)
```

```
##        SCE.21   CME.21   SCE.22   CME.22
## [1,] 246.6825 8.506293 195.0478 6.725785
```

fparciales 21 y 22 Para verificar cual de los 2 modelos con 2 covaraiables es mejor calculamos los fparciales vs el fcritico de modelo con la cavarible "wt"

```
fparcial.21<-(SCE.12-SCE.21)/CME.21
fparcial.22<-(SCE.12-SCE.22)/CME.22

cbind(f.critico2, fparcial.21, fparcial.22)
```

```
##      f.critico2 fparcial.21 fparcial.22
## [1,]   3.195056    8.285182    18.15564
```

Elmodelo que cuyo fparcial es mayor al f.critico2 es el modelo 22.Este modelo es mejor que el modelo que solo tiene a wt

Conclusión: Tanto el modelo con 21 como el modelo 22 son mejores que el modelo 11 pero de estos 2 el mejor que contiene a las covariables wt y hp

```
mt.fit22<-lm(mpg ~ wt + hp)
summary(mt.fit22)
```

Posible modelo con las covariabes "wt" y "hp"

```
##
## Call:
## lm(formula = mpg ~ wt + hp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.941 -1.600 -0.182  1.050  5.854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
## wt          -3.87783    0.63273  -6.129 1.12e-06 ***
## hp          -0.03177    0.00903  -3.519  0.00145 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.593 on 29 degrees of freedom
## Multiple R-squared:  0.8268, Adjusted R-squared:  0.8148
## F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

PASO 3 DE FORWARD ... (GUARDAR SCE.22) Sólo queda chequear si el modelo con las 3 covariables es mejor que el modelo que solo selecciona las covaribales wt y hp. Para ello calculamos un f.critico con los siguientes parametros

G.L SST = 50 - 1 G.L SSR = 3 parametros G.L SSE = G.L SST - G.L SSR = 46

```
f.critico3<- qf(1- alpha , 3 , 46 )
f.critico3
```

```
## [1] 2.806845
```

```
mt.fit3<-lm(mpg ~ wt + hp + disp)
anova(mt.fit3)
```

Se ajusta modelo con las 3 covariables

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## wt         1 847.73  847.73 121.7305 1.052e-11 ***
## hp         1  83.27   83.27  11.9579  0.001758 **
## disp       1   0.06    0.06   0.0082  0.928507
```

```
## Residuals 28 194.99    6.96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SCE.3<-anova(mt.fit3)[4,2]
CME.3<-anova(mt.fit3)[4,3]
```

Se obtiene SCE.3 y CME.3 del modelo con 3 covaraibles

```
fparcial.3<-(SCE.22-SCE.3)/CME.3
cbind(f.critico3,fparcial.3)
```

Calculo del fparcial.3

```
##      f.critico3  fparcial.3
## [1,]   2.806845 0.008196498
```

Como el fcritico 3 es mayor que el fparcial.3 el modelo propuesto con 3 covariables no es mejor que el modelo con 2 variables dado por las variables wt y hp

Por ende podemos decir que el modelo propuesto que mejor explica a mpg mediante la metodologia forward es:

mpg = B0 + B1 * wt + B2 * hp

Validacion de supuestos

Para la validacion de los de supuestos, se utiliza el modelo propuesto modelo y sus residuales, para lo cual se ha creado la siguiente funcion.

```
ValidarSupuestos <- function (respuesta ,modelo,confianza){

  print("En conclusión: ")

  # Test de normalidad de los residuales

  shapiro <- shapiro.test(modelo$residuals)
  shapiro <- shapiro$p.value

  lillie <- lillie.test(modelo$residuals)
  lillie <- lillie$p.value

  ifelse((shapiro > confianza) & (lillie > confianza), print("Normalidad de los residuales, no se recha

  # Test de homocedasticidad (varianza constante de los residuales)
```

```
  Breusch <- bptest(modelo)
  Breusch <- Breusch$p.value

  Goldfeld <- gqtest(modelo)
  Goldfeld <- Goldfeld$p.value

  ifelse((Breusch > confianza) & (Goldfeld > confianza), print("Existe homocedasticidad, no se rechaza H

  # Test de independencia de errores

  indepenencia <- dwtest(mpg ~ modelo$residuals)
  indepenencia <- indepenencia$p.value

  ifelse((indepenencia > confianza), print("Hay independencia de errores, No se rechaza Ho"), print("No

  # Construccion de la Tabla de respuestas

  tabla <- rbind(shapiro, lillie, Breusch, Goldfeld, indepenencia)
  rownames(tabla) <- c("Shapiro", "Lillie", "Breusch", "Goldfeld", "independencia")
  colnames(tabla) <- c("p.value")

  print(tabla)
}
```

```
modelo_propuesto <- lm(mpg ~ wt + hp)


ValidarSupuestos(mpg,modelo_propuesto,alpha)
```

```
## [1] "En conclusión: "
## [1] "No existe normalidad de los residuales, se rechaza H0"
## [1] "No existe homocedasticidad, se rechaza H0"
## [1] "No hay independencia de errores, se rechaza Ho"
##                    p.value
## Shapiro       0.0342747606
## Lillie        0.3674906935
## Breusch       0.6438038145
## Goldfeld      0.0004864276
## independencia 0.0022188827
```