SME0820 - Modelos de Regressão e Aprendizado Supervisionado I - Exercício 2

Brenda da Silva Muniz 11811603 Francisco Rosa Dias de Miranda 4402962
Heitor Carvalho Pinheiro 11833351 Mônica Amaral Novelli 11810453

Novembro 2021

Exercício 4

Queremos mostrar que o estimador F_0 pode ser escrito na forma $F_0 = \frac{R^2(n-p)}{k(1-R^2)}$

Note que podemos escrever $SQ_{reg} = SQ_{total}R^2$. Por definição, temos que

$$F_0 = \frac{QM_{reg}}{QM_{res}} = \frac{\frac{SQ_{reg}}{k}}{\frac{SQ_{res}}{n-p}} = \frac{SQ_{reg}}{SQ_{res}}\frac{n-p}{k} = \frac{SQ_{total}\ R^2}{SQ_{total} - SQ_{reg}}\frac{n-p}{k} = \frac{SQ_{total}\ R^2}{SQ_{total} - SQ_{total}\ R^2}\frac{n-p}{k} = \frac{SQ_{total}\ R^2}{SQ_{total}(1-R^2)}\frac{n-p}{k} = \frac{R^2(n-p)}{k(1-R^2)}.$$

Portanto, os dois são equivalentes.

Exercício 5

a) Usando o exercício 4, teste a significância da regressão com $\alpha = 0.05$

Do exercício anterior temos que:

$$F_0 = \frac{R^2(n-p)}{k(1-R^2)}$$

Para $k = 2$, $n = 25$, $p = k + 1$ e $R^2 = 0.90$:

$$F_0 = \frac{0.9(25-3)}{2(1-0.9)} = 99$$

Podemos calcular $F_{(0.95,2,22)}$ com o seguinte comando:

```
#calculando o valor crítico de F(0.95,2,22)
qf(0.95,2,22)
```

```
## [1] 3.443357
```

Logo, $F_{(0.95,2,22)} \approx 3.443$.

Como, $F_0 > F_{(0.95,2,22)}$, rejeitamos H_0 em favor de H_1 e concluímos que o modelo testado é significante para $\alpha = 0.05$. Isto é, ele capta melhor a tendência dos dados se comparado ao modelo restrito, $y = \beta_0 + \epsilon$.

b) Qual o menor valor de R^2 para que o modelo seja significativo?

Como no item anterior, vamos considerar um nível de significância $\alpha = 0.05$.

Sabemos que o modelo será significante se, e somente se, $F_0 > F_{(0.95,2,22)}$. Isso implica que:

$$\frac{R^2(n-p)}{k(1-R^2)} > 3.443$$

Logo, segue que:

$$\frac{R^2(25-3)}{2(1-R^2)} > 3.443 R^2 > 0.2383$$

Portanto, considerando apenas dois algarismos significativos, podemos considerar que o valor mínimo de R^2 para que o modelo possa ser considerado significativo é $R^2 = 0.24$

Exercício 11

Conjunto de dados

O dataset contém dados de um experimento para determinar **tempo**, **temperatura**, **percentual de solvatação**, **rendimento de óleo** e **carvão total** sob o **rendimento (y)**

Significância: 97%

Leitura dos pacotes utilizados:

Leitura do pacote de dados:

```
library(readr)
dados <- read_csv("data-table-B5.csv")
```

```
#Renomeando as colunas
names(dados) <- c('y', 'Tempo', 'Temperatura', 'Perc_solvatação', 'Rendimento_Òleo', 'Carvão_Total','x6
head(dados)
```

```
## # A tibble: 6 x 8
##       y Tempo Temperatura Perc_solvatação Rendimento_Òleo Carvão_Total    x6
##   <dbl> <dbl>       <dbl>           <dbl>           <dbl>        <dbl> <dbl>
## 1 37.0    5.1         400            51.4            4.24        1485. 2227.
## 2 13.7   26.4         400            72.3           30.9          290.  435.
## 3 10.1   23.8         400            71.4           33.0          321.  481.
## 4  8.53  46.4         400            79.2           44.6          165.  247.
## 5 36.4    7           450            80.5           33.8         1097. 1646.
## 6 26.6   12.6         450            89.9           41.3          605.  908.
## # ... with 1 more variable: x7 <dbl>
```

Para facilitar nosso trabalho em termos computacionais, estaremos nomeando nossas variáveis e criando um data frame com as mesmas; sendo:

- Y : Rendimento total
- X_1 : Tempo
- X_2 : Temperatura
- X_3 : Perc_solvatação
- X_4 : Rendimento_Òleo
- X_5 : Carvão_Total

```
x1 <- dados$Tempo
x2 <- dados$Temperatura
x3 <- dados$Perc_solvatação
x4 <- dados$Rendimento_Òleo
x5 <- dados$Carvão_Total
y <- dados$y

tabela01 <- data.frame(cbind(x1, x2, x3, x4, x5, y))
```

Devido à complexidade das fórmulas envolvidas para realizarmos uma regressão linear múltima, utilizaremos uma abordagem matricial. Desse modo, poderemos escrever

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, ..., 5.$$

como:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + ... + \beta_k x_{k1} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + ... + \beta_k x_{k2} + \varepsilon_2 \\ y_3 = \beta_0 + \beta_1 x_{13} + \beta_2 x_{23} + ... + \beta_k x_{k3} + \varepsilon_3 \\ y_4 = \beta_0 + \beta_1 x_{14} + \beta_2 x_{24} + ... + \beta_k x_{k4} + \varepsilon_4 \\ y_5 = \beta_0 + \beta_1 x_{15} + \beta_2 x_{25} + ... + \beta_k x_{k5} + \varepsilon_5 \end{cases}$$

Assim, alocamos essas equações em dois vetores coluna (5x1), fazendo:

```
n <- length(dados$y)
X <- matrix(c(rep(1,n), x1, x2, x3, x4, x5), ncol = 6, nrow = n, byrow = FALSE)
Y <- matrix(y, ncol = 1, nrow = n)
k <- ncol(X) -1
p <- k + 1
```

**

```
#Definindo os betas do modelo de regressão múltipla

betas <- solve(t(X)%*%X)%*%t(X)%*%Y

#Definindo a matrix C_jj
```

```
C_jj = solve(t(X)%*%X)

#Definindo uma matrix para os betas

betas <- matrix(data = betas, nrow = length(betas), ncol = 1, byrow = FALSE)
rownames(betas) <- c("beta0", "beta1","beta2","beta3","beta4","beta5")
betas
```

```
##              [,1]
## beta0 30.59319595
## beta1 -0.38612042
## beta2 -0.04556211
## beta3  0.31074404
## beta4 -0.37167815
## beta5  0.01409776
```

Abaixo, estaremos conferindo com os valores dos betas calculados pelo método lm. É importante ressaltar que não chamaremos o data frame do conjunto de dados completo - visto que esse conta com as variáveis x_6 e x_7 que não serão consideradas nessa análise de regressão -, e sim a tabela com as variáveis de x_1 até x_5.

```
#Modelo do R
model <- lm(formula = y ~., data = tabela01)
```

Segue abaixo a equação para o modelo determinado:

$$y = 30.59320 - 0.38612x_1 - 0.04556x_2 + 0.31074x_3 - 0.37168x_4 + 0.01410x_5$$

Teste de significancia da Regressão

A estimação de σ^2 é necessária para o seguimento do teste. Sendo assim, fazemos:

$$SQ_{res} = \boldsymbol{Y}^T\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y}$$

```
SQRes <- (t(Y)%*%Y)-(t(betas)%*%t(X)%*%Y)
SQRes
```

```
##          [,1]
## [1,] 3222.907
```

Logo, $\widehat{\sigma}^2 = \dfrac{SQres}{n-p}$.

```
p <- ncol(X)

#estimando o sigma^2

sigma2 <- SQRes/(n-p)
sigma2
```

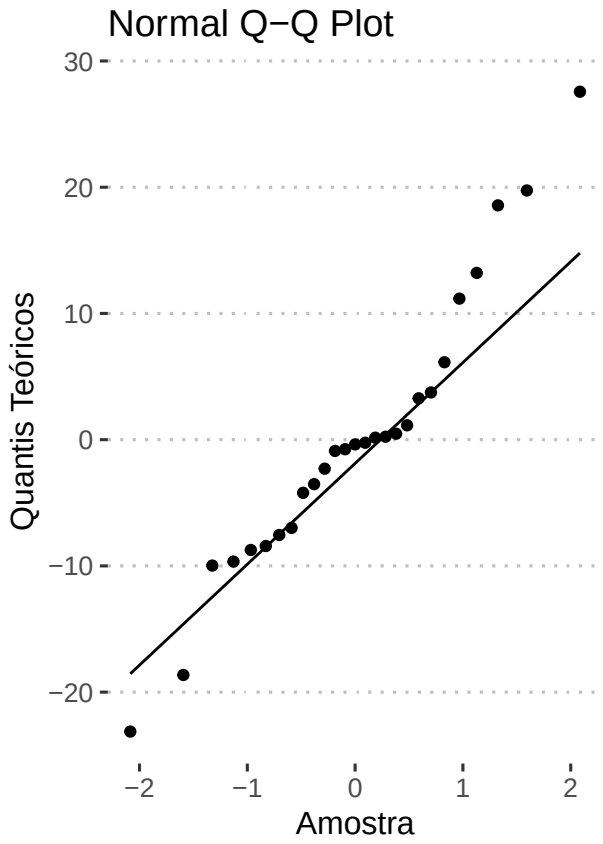
```
##          [,1]
## [1,] 153.4718
```

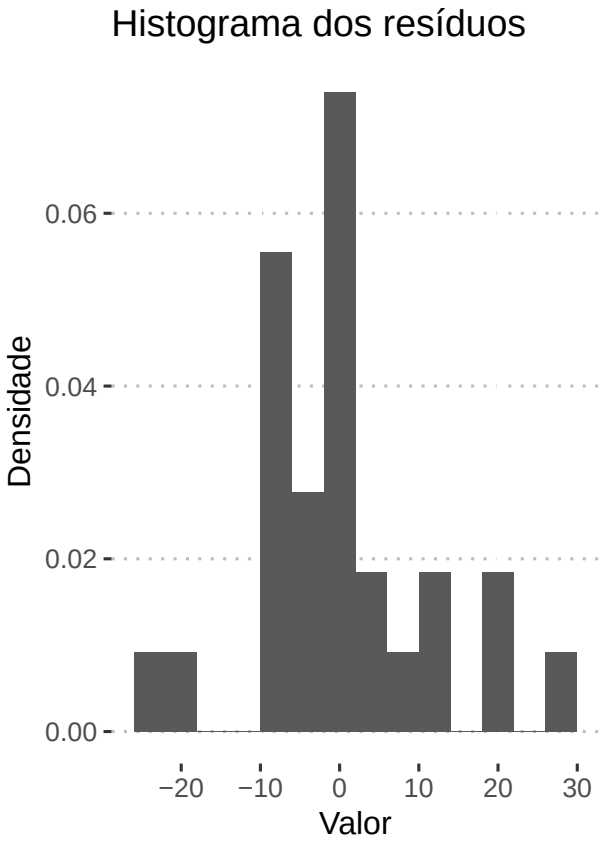
ANOVA

Agora que ajustamos um modelo inicial, é necessário que verifiquemos se ele é adequado em explicar a variabilidade de nossa amostra. Vamos assumir que $\xi \sim N_n(0, \sigma^2 I)$. Precisamos agora verificar nossa suposição graficamente:

```
# Obtendo uma estimativa para Y a partir do modelo ajustado
Y_est <- X%*%betas

# Cálculo dos resíduos
res <- Y - Y_est
```





Interpretando os gráficos acima, podemos observar no histograma que a distribuição - embora majoritariamente centralizada - ocorre de forma irregular. No gráfico de dispersão podemos constatar certa normalidade dos valores, que possuem uma aproximação evidente dos quantis teóricos.

Além disso, para determinar matematicamente se existe uma relação linear entre a variável resposta $\boldsymbol{Y}$ e qualquer as outras covariáveis $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$, é possível utilizar o teste **ANOVA**. Nele, queremos testar:

H_0: Nenhuma das variáveis contribui significativamente ao modelo, versus:

H_a: Pelo menos uma das covariáveis contribui significativamente ao modelo.

Tabela *ANOVA*:

Table 1: Tabela ANOVA

F.V.	G.L	S.Q.	Q.M.	F
Regressão	k	SQ_{reg}	$QMreg = \frac{SQreg}{k}$	$F = \frac{QMreg}{QMres}$
Resíduo	$n-p$	SQ_{res}	$QMres = \frac{SQres}{n-p}$	
Total	$n-1$	SQ_{total}	QM_{Total}	

Table 2: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	3909	3909	25.47	5.363e-05
x2	1	0.067	0.067	0.0004366	0.9835
x3	1	271.3	271.3	1.768	0.1979
x4	1	49.21	49.21	0.3206	0.5772
x5	1	417.2	417.2	2.718	0.1141
Residuals	21	3223	153.5	NA	NA

```
# Soma dos quadrados dos residuos
(SQRes <- t(Y-Y_est)%*%(Y-Y_est))
```

```
##          [,1]
## [1,] 3222.907
```

$$SQ_{Reg} = SQ_{Total} - SQ_{Res} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (Y - \hat{Y})^T \cdot (Y - \hat{Y}) =$$

$$\beta^T \cdot X^T \cdot Y - \frac{1}{n}(u^T \cdot Y)^2$$

```
# Soma dos quadrados totais
u <- c(rep(1,n))
(SQTotal <- t(Y)%*%Y - ((t(u)%*%Y)^2)/n)
```

```
##          [,1]
## [1,] 7870.112
```

```
#Soma dos quadrados da regressao
(SQReg <- SQTotal - SQRes)
```

```
##          [,1]
## [1,] 4647.205
```

```
# Calculando a anova

k <- 5
p <- k+1
```

```r
gl_sqreg <- k
QMReg <- SQReg/gl_sqreg

gl_sqres <- n-p
QMRes <- SQRes/gl_sqres

gl_sqtotal <- n-1
```

```r
#calculando a estatistica F
(F_0 <- QMReg/QMRes)
```

```
##          [,1]
## [1,] 6.056104
```

```r
(QMTotal <- QMRes + QMReg)
```

```
##          [,1]
## [1,] 1082.913
```

Como nosso estimador $F \sim F(k,\ n-k-1)$, podemos obter os quantis com o auxílio do R, assumindo $\alpha = 0.03$ - a partir de instruções fornecidas em trabalhos anteriores:

```r
alpha <- 0.03
(RR <- qf(alpha, df1 = k, df2 = n - k -1, lower.tail = F) )
```

```
## [1] 3.098817
```

Rejeitamos H_0 se $F_0 > F_{crit}$, sendo F_{crit} o quantil teórico da distribuição F com k e $n-p$ graus de liberdade.

```r
if(RR < F_0){
  cat("Rejeita-se H0")
}
```

```
## Rejeita-se H0
```

Dessa forma, podemos concluir com 97% de confiança que pelo menos uma das variáveis contribui significativamente ao modelo.

Verificando a importancia do subconjunto com x3, x4 e x5.

Sob $H_0 : \beta_2 = 0$ VS $H_1 : \beta_2 \neq 0$:

Soma de quadrados extra devido a x_3,x_4 e x_5:

$$SQ_{reg}(\beta_3, \beta_4, \beta_5 | \beta_0, \beta_1, \beta_2) = SQ_{reg}(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) - SQ_{reg}(\beta_0, \beta_1, \beta_2)$$

(Não corrigida)

$$SQ_{reg}(\beta_3, \beta_4, \beta_5|\beta_0, \beta_1, \beta_2) =$$
$$SQ_{reg}(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5|\beta_0) - SQ_{reg}(\beta_1|\beta_0) - SQ_{reg}(\beta_2|\beta_1, \beta_0) -$$
$$SQ_{reg}(\beta_3|\beta_2, \beta_1, \beta_0) - SQ_{reg}(\beta_4|\beta_3, \beta_2, \beta_1, \beta_0) - SQ_{reg}(\beta_5|\beta_4, \beta_3, \beta_2, \beta_1, \beta_0)$$

(Corrigida)

Temos que:

$$SQ_{reg}(\beta_0, \beta_1, \beta_2) = SQ_{reg}(\beta) = \widehat{\beta}^T X^T \boldsymbol{y}$$

```
SQReg_r <- (t(betas)%*%t(X)%*%Y)
SQReg_r
```

```
##           [,1]
## [1,] 21160.17
```

e

$$SQ_{reg}(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5|\beta_0) = \widehat{\beta}^T X^T \boldsymbol{Y} - \frac{(I_n^T \boldsymbol{Y})^2}{n}$$

coincide com o valor obtido na tabela ANOVA, sendo este:

```
SQReg # SQreg encontrado na tabela ANOVA
```

```
##           [,1]
## [1,] 4647.205
```

**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \xi$$

Sob $H_0 : \beta_2 = 0$:

Considerando o modelo reduzido:

$$Y = \beta_1 X_1 + x_i$$

```
Sxx <- sum(x1^2) - n*(mean(x1))^2
Sxy <- sum(x1*y) - n*(mean(x1)*mean(y))
beta_est_modredu <- Sxy/Sxx
beta_est_modredu
```

```
## [1] -0.8650001
```

Agora, tendo o modelo reduzido como sendo:

$$SQ_{reg} = \beta_1 S_{XY}$$

podemos realizar:

```
SQReg_modredu <- beta_est_modredu*Sxy
SQReg_modredu # 3 G.L. (3 covariáveis no subconjunto)
```

```
## [1] 3909.467
```

```
gl_modredu <- 3
```

e portanto,

```
SQReg_teste <- SQReg-SQReg_modredu
SQReg_teste
```

```
##          [,1]
## [1,] 737.7382
```

Este valor representa o aumento na SQ_{reg} com o acréscimo do subconjunto no modelo que já possui X_1

```
QMReg_modredus <- (SQReg_teste/gl_modredu)
F_testeparcial <- (QMReg_modredus/QMRes)
F_testeparcial
```

```
##          [,1]
## [1,] 1.602332
```

```
alpha <- 0.03
RR <- qf(alpha, df1 = gl_modredu, df2 = n - gl_modredu -1, lower.tail = F)
RR
```

```
## [1] 3.556927
```

```
if(RR < F_testeparcial){
 cat("Rejeita-se H0")
}
```