Lista 03

Klysman Rezende e Matheus Cougias

17/08/2020

Leitura dos dados

Inicialmente é feita a leitura dos dados do arquivo txt, mas como pode-se perceber, algumas informações presentes no arquivo não nos interessa, além de que a leitura do mesmo não gera uma “tabela” de dados, e sim uma grande tabela com somente uma coluna e uma grande quantidade de linhas. Dessa maneira, também é realizada a seleção de quais dados serão necessários na análise. Realizei três testes para definir qual a melhor maneira de realizar a leitura dos dados, decidindo utilizar o operador “delim”, que quebrou o arquivo em TABs e quebras de linha.

Realizei a leitura do arquivo através do “delim”, que gerava basicamente um vetor gigante com todas as informações do arquivo .txt. Em seguida, realizei o corte das informações que não eram relevantes para a análise, ou seja, o texto explicado inicial do arquivo (as 5 primeiras linhas do mesmo). Assim, fiz a seleção de quais eram os índices (nomes de coluna) que seriam retirados e também dos valores. Montei uma matriz relacionando os valores de acordo com o índice que ele representa.

Como nem todas as variáveis devem ser analizadas, realizei também o corte de algumas colunas desnecessárias.

```
#require(tidyverse)
#require(readtext)
#dados1 <- readLines("boston_corrected.txt")
#dados2 <- readtext("boston_corrected.txt")
arquivo <- read.delim("boston_corrected.txt")
arquivo <- arquivo$http...lib.stat.cmu.edu.datasets.boston_corrected.txt[5:10651]
indices <- arquivo[1:21]

arquivo <- arquivo[22:10647]
valores <- as.numeric(arquivo)
```

```
## Warning: NAs introduzidos por coerção
```

```
dados <- matrix(valores, ncol = 21, byrow = TRUE)

colnames(dados) <- indices

dados <- data.frame(dados)

dados <- dados[c(7, 9 : 21)]
```

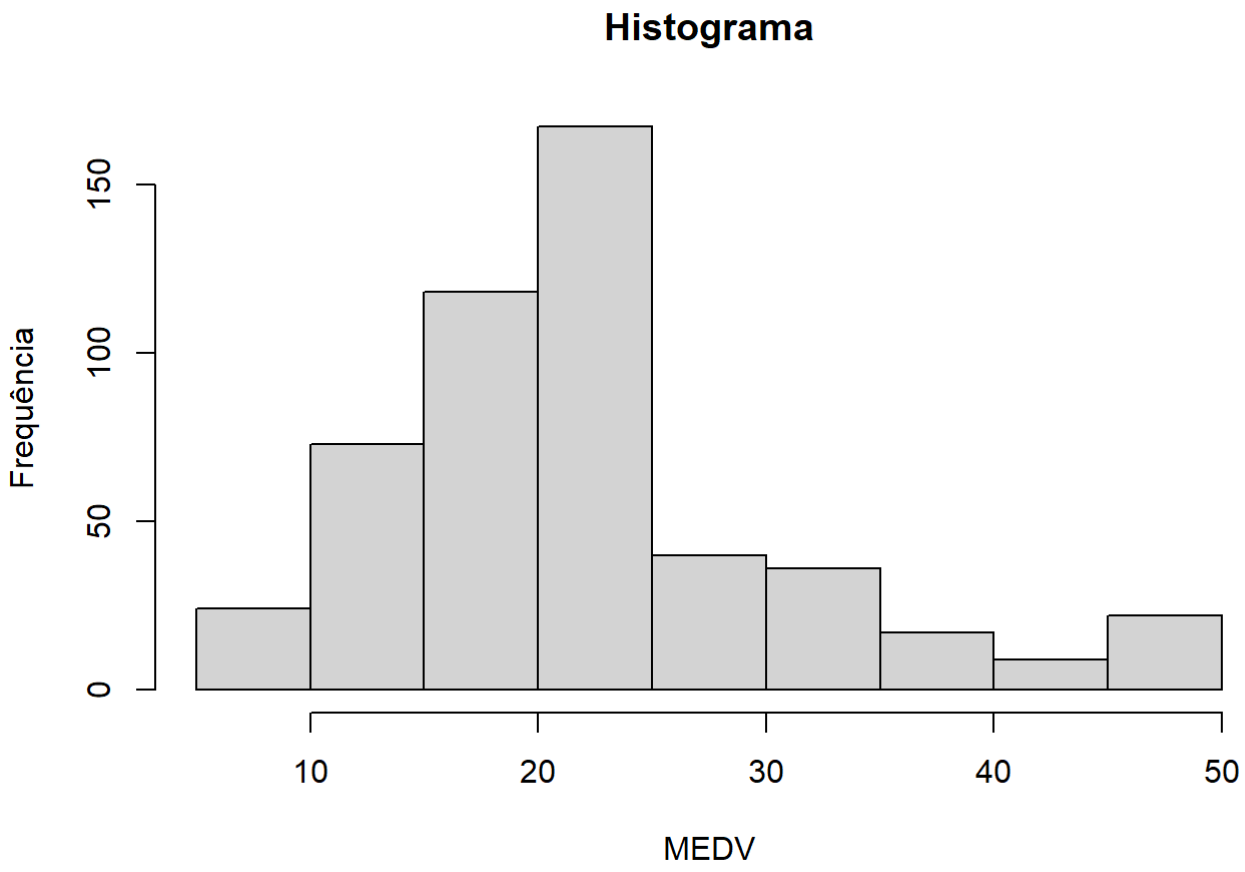
Análise descritiva da variável resposta

Com a análise do sumário da variável resposta MEDV, temos que a média apresentada para seu valor é de 22.53, com o menor valor sendo de 5.00 e o valor máximo 50.00. Através do histograma e o boxplot, percebe-se que a maior parte dos valores apresentados estão na faixa entre 15 a 25. Ainda no boxplot, existe uma quantidade considerável de pontos fora do intervalo de 0.25 a 0.75, prncipalmente acima do valor 40.00, que não necessariamente são outliers.

```
summary(dados$MEDV)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   17.02   21.20   22.53   25.00   50.00
```

```
hist <- hist(dados$MEDV, xlab = "MEDV", ylab = "Frequência", main = "Histograma")
```




```
boxplot <- boxplot(dados$MEDV, xlab = "MEDV", ylab = "Valor", main = "Boxplot")
```

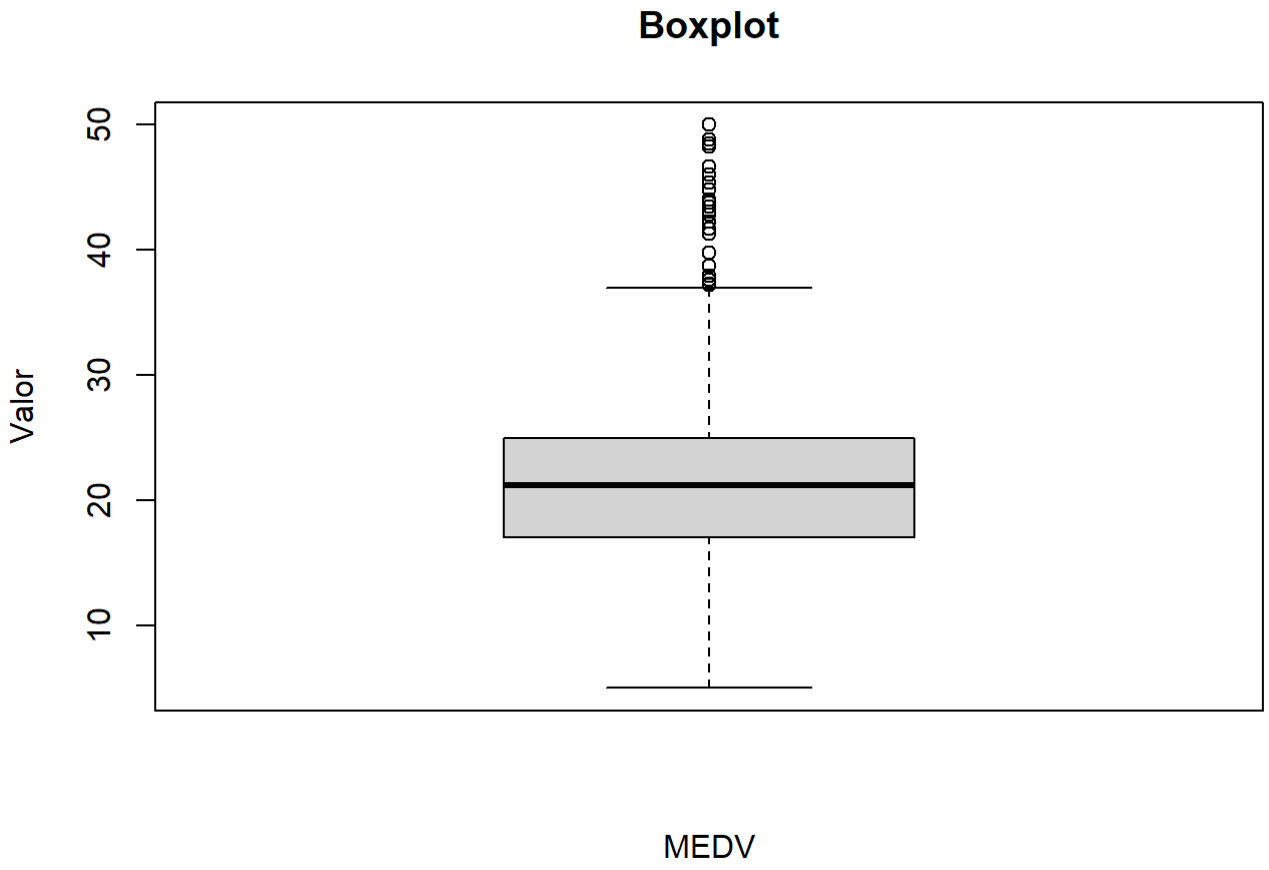


Gráfico de correlação linear

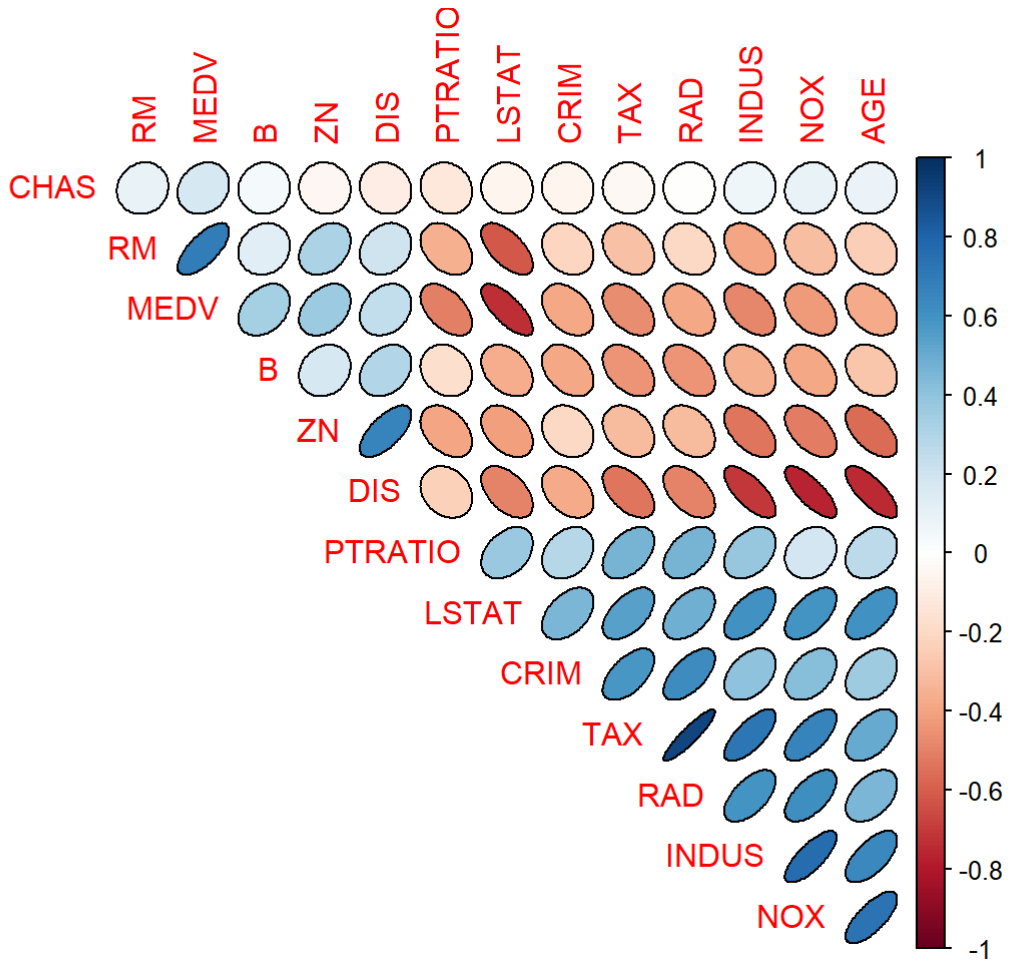
Através da leitura do gráfico de correlação linear entre a variável MEDV (valor mediano das residências), pode-se perceber que existe uma forte correlação negativa com a variável LSTAT (porcentagem de população de baixa renda), que é provado logicamente com o fato das variáveis representarem aspectos contrários da população. Por outro lado, a variável RM (número médio de quartos por habitação) possui uma forte correlação positiva com a variável resposta. Nas demais variáveis a correlação não se mostra tão presente, apesar de que em sua maioria ela seja levemente negativa.

```
 require(corrplot)
```

```
## Loading required package: corrplot
```

```
## corrplot 0.84 loaded
```

```
 corMat <- cor(dados)
 corrplot(corMat, method = "ellipse", type = "upper", order = "AOE", diag=FALSE, addgrid.col=
NA, outline=TRUE)
```



Regressão linear simples

Aplicando a regressão linear simples com o pacote exploreR, temos a confirmação dos resultados encontrados no tópico anterior. Todas as variáveis possuem um p-valor considerável, mostrando que existe uma relação entre elas e a variável resposta. Outro resultado afirmado é dado pelo R^2, confirmando que as variáveis RM e LSTAT são as que melhor explicam os resultados gerados em MEDV.

```
require(exploreR)
```

```
## Loading required package: exploreR
```

```
require(knitr)
```

```
## Loading required package: knitr
```

```
reg_simples <- masslm(dados, "MEDV")
print(reg_simples)
```

```
##         IV Coefficient   P.value  R.squared
## 1     CRIM    -0.41520 1.174e-19 0.15078047
## 2       ZN     0.14210 5.714e-17 0.12992084
## 3    INDUS    -0.64850 4.900e-31 0.23399003
## 4     CHAS     6.34600 7.391e-05 0.03071613
## 5      NOX   -33.92000 7.065e-24 0.18260304
## 6       RM     9.10200 2.487e-74 0.48352546
## 7      AGE    -0.12320 1.570e-18 0.14209474
## 8      DIS     1.09200 1.207e-08 0.06246437
## 9      RAD    -0.40310 5.466e-19 0.14563858
## 10     TAX    -0.02557 5.638e-29 0.21952592
## 11 PTRATIO    -2.15700 1.610e-34 0.25784732
## 12       B     0.03359 1.318e-14 0.11119612
## 13   LSTAT    -0.95000 5.081e-88 0.54414630
```

Regressão linear múltipla

```
require(car)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
modelo <- lm(MEDV ~ ., data = dados)


## Caso o VIF seja maior que 5 (ou maior que 10) há forte evidencia de multicolinearidade

vif(modelo)
```

```
##     CRIM       ZN    INDUS     CHAS      NOX       RM      AGE      DIS
## 1.792192 2.298758 3.991596 1.073995 4.393720 1.933744 3.100826 3.955945
##      RAD      TAX  PTRATIO        B    LSTAT
## 7.484496 9.008554 1.799084 1.348521 2.941491
```

```
1 - 1/vif(modelo)
```

```
##       CRIM         ZN      INDUS       CHAS        NOX         RM        AGE
## 0.44202393 0.56498252 0.74947367 0.06889725 0.77240242 0.48286858 0.67750523
##        DIS        RAD        TAX    PTRATIO          B      LSTAT
## 0.74721589 0.86639048 0.88899439 0.44416160 0.25844689 0.66003636
```

```
modeloVIF <- lm(TAX ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD + B + LSTAT, data
= dados)

summary(modeloVIF)
```

```
##
## Call:
## lm(formula = TAX ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE +
##     DIS + RAD + B + LSTAT, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -219.696  -20.338   -3.531   14.345  261.000
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 211.05428   48.06919   4.391 1.38e-05 ***
## CRIM         -0.11805    0.39421  -0.299  0.76471
## ZN            0.73322    0.15302   4.792 2.19e-06 ***
## INDUS         7.36396    0.65228  11.290  < 2e-16 ***
## CHAS        -29.55484   10.20596  -2.896  0.00395 **
## NOX          50.79561   43.20960   1.176  0.24034
## RM           -9.64759    4.93123  -1.956  0.05098 .
## AGE           0.12665    0.15791   0.802  0.42291
## DIS           1.41527    2.38180   0.594  0.55265
## RAD          14.23444    0.44825  31.755  < 2e-16 ***
## B            -0.01732    0.03220  -0.538  0.59084
## LSTAT        -0.16134    0.60784  -0.265  0.79078
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.94 on 494 degrees of freedom
## Multiple R-squared:  0.8884, Adjusted R-squared:  0.8859
## F-statistic: 357.4 on 11 and 494 DF,  p-value: < 2.2e-16
```

```
#Seleção automatica do modelo
 reg_multipla <- lm(MEDV ~ ., data = dados)
 reg_multipla <- step(reg_multipla)
```

```
## Start:  AIC=1589.64
## MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## - AGE      1      0.06 11079 1587.7
## - INDUS    1      2.52 11081 1587.8
## <none>                 11079 1589.6
## - CHAS     1    218.97 11298 1597.5
## - TAX      1    242.26 11321 1598.6
## - CRIM     1    243.22 11322 1598.6
## - ZN       1    257.49 11336 1599.3
## - B        1    270.63 11349 1599.8
## - RAD      1    479.15 11558 1609.1
## - NOX      1    487.16 11566 1609.4
## - PTRATIO  1   1194.23 12273 1639.4
## - DIS      1   1232.41 12311 1641.0
## - RM       1   1871.32 12950 1666.6
## - LSTAT    1   2410.84 13490 1687.3
##
## Step:  AIC=1587.65
## MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + DIS + RAD + TAX +
##     PTRATIO + B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## - INDUS    1      2.52 11081 1585.8
## <none>                 11079 1587.7
## - CHAS     1    219.91 11299 1595.6
## - TAX      1    242.24 11321 1596.6
## - CRIM     1    243.20 11322 1596.6
## - ZN       1    260.32 11339 1597.4
## - B        1    272.26 11351 1597.9
## - RAD      1    481.09 11560 1607.2
## - NOX      1    520.87 11600 1608.9
## - PTRATIO  1   1200.23 12279 1637.7
## - DIS      1   1352.26 12431 1643.9
## - RM       1   1959.55 13038 1668.0
## - LSTAT    1   2718.88 13798 1696.7
##
## Step:  AIC=1585.76
## MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO +
##     B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## <none>                 11081 1585.8
## - CHAS     1    227.21 11309 1594.0
## - CRIM     1    245.37 11327 1594.8
## - ZN       1    257.82 11339 1595.4
## - B        1    270.82 11352 1596.0
## - TAX      1    273.62 11355 1596.1
## - RAD      1    500.92 11582 1606.1
## - NOX      1    541.91 11623 1607.9
## - PTRATIO  1   1206.45 12288 1636.0
## - DIS      1   1448.94 12530 1645.9
## - RM       1   1963.66 13045 1666.3
## - LSTAT    1   2723.48 13805 1695.0
```

```
summary(reg_multipla)
```

```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5984  -2.7386  -0.5046   1.7273  26.2373
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
## CRIM         -0.108413   0.032779  -3.307 0.001010 **
## ZN            0.045845   0.013523   3.390 0.000754 ***
## CHAS          2.718716   0.854240   3.183 0.001551 **
## NOX         -17.376023   3.535243  -4.915 1.21e-06 ***
## RM            3.801579   0.406316   9.356  < 2e-16 ***
## DIS          -1.492711   0.185731  -8.037 6.84e-15 ***
## RAD           0.299608   0.063402   4.726 3.00e-06 ***
## TAX          -0.011778   0.003372  -3.493 0.000521 ***
## PTRATIO      -0.946525   0.129066  -7.334 9.24e-13 ***
## B             0.009291   0.002674   3.475 0.000557 ***
## LSTAT        -0.522553   0.047424 -11.019  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.736 on 494 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7348
## F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

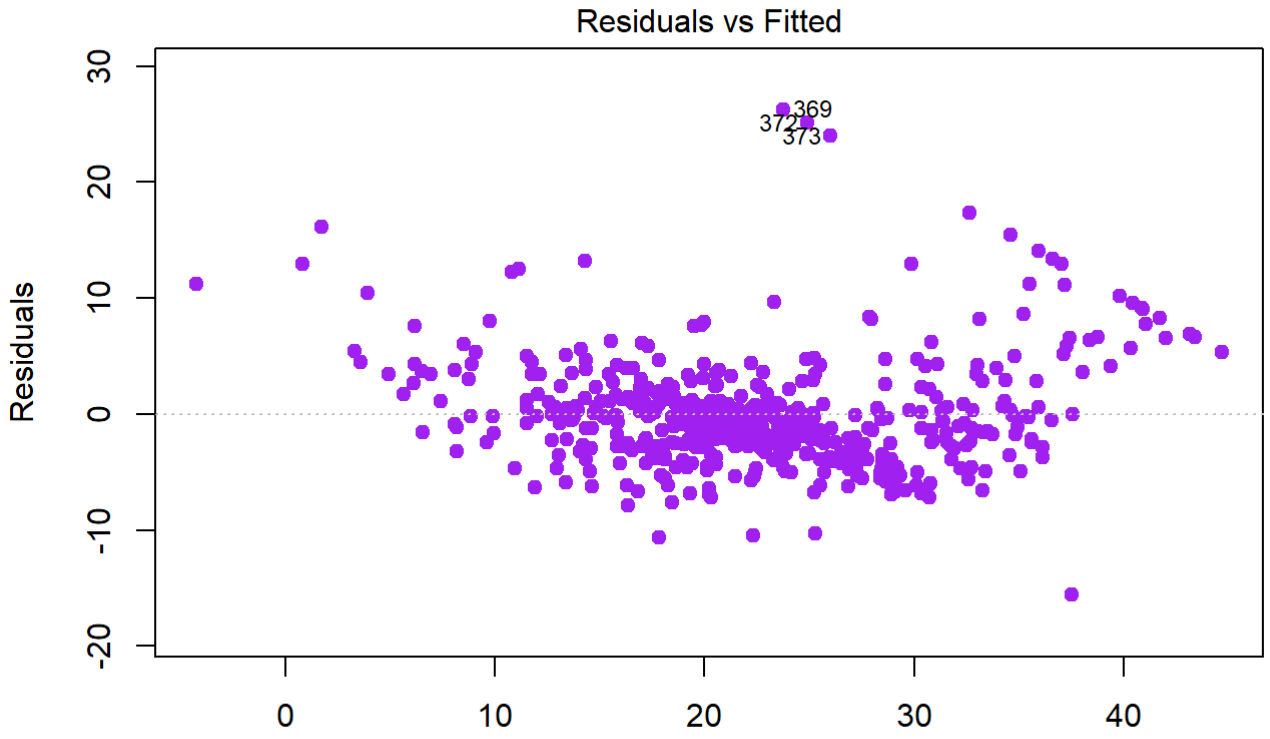
```
## Para remover o Intercepdo do modelo: lm(MEDV ~ -1 + Variaveis)
```

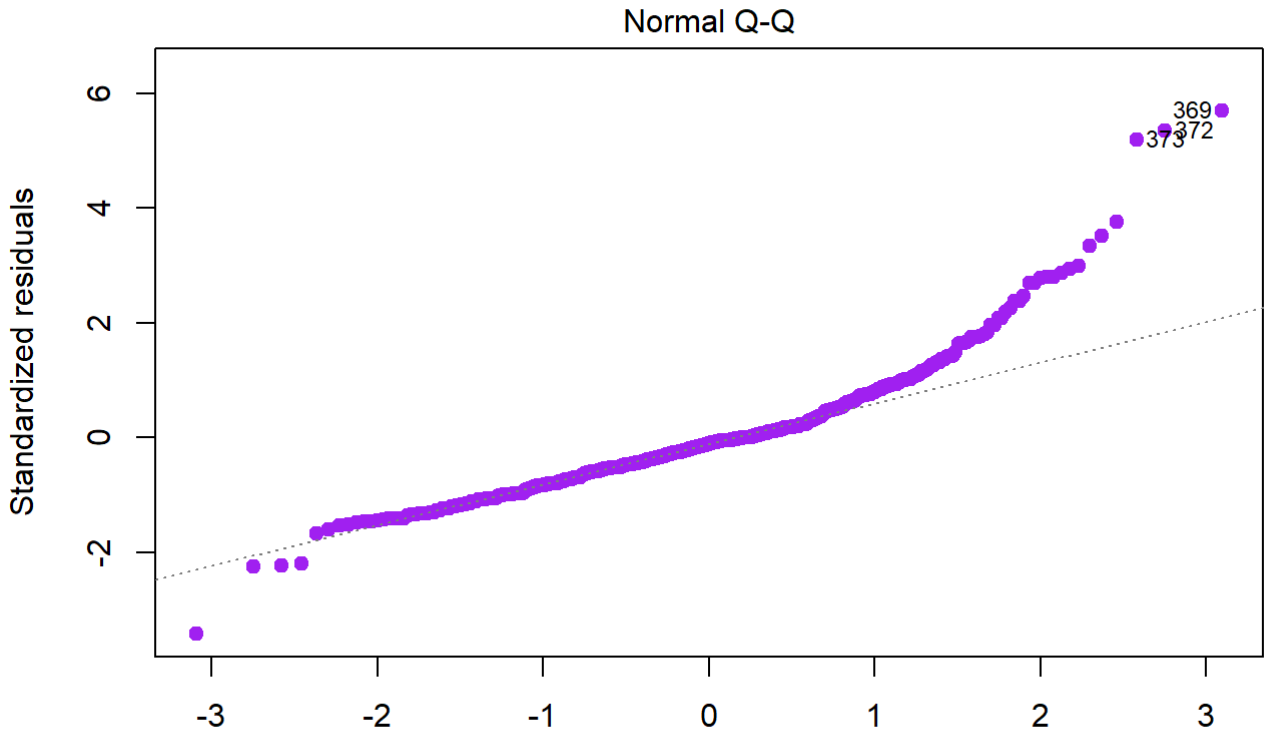
#Analises Para análises de regressão linear múltipla, o ideal é manter um modelo enxuto. Reduzir gradativamente a complexidade pela retirada de variáveis. Assim podemos aos poucos corrigir a Colinearidade (explica quando há pares de variáveis) e Multicolinearidade (relação linear com outras variáveis).

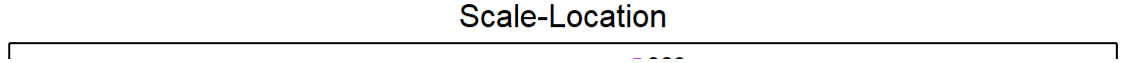
Em nosso gráfico Residuals vs Fitted, observamos uma certa homogeneidade em torno da média 0 com alguns pontos discrepantes porém não convem investigá-los nesse momento

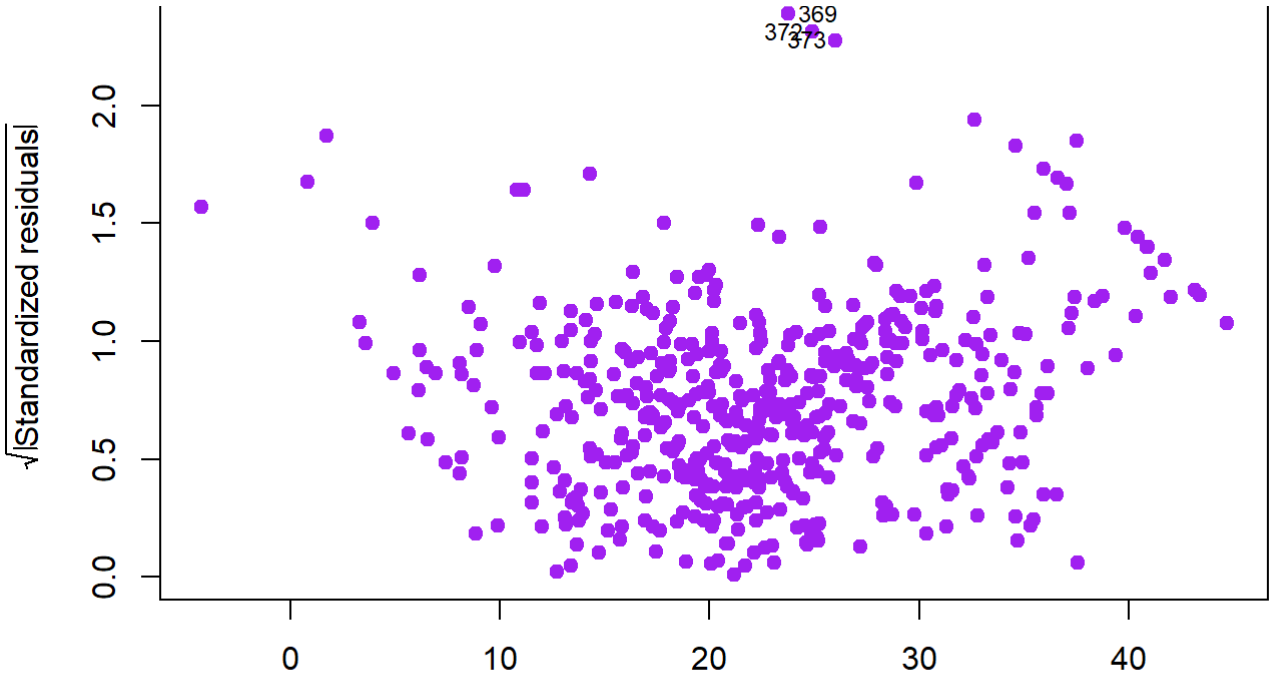
Pela analise do gráfico Normal Q-Q, não há um comportamento total pela distribuição normal, visto que o gráfico mostra o desalinhamento dos dados com a reta. Isso nos dias que os dados não são providemos de uma distribuição normal. Isso se confirma pela primeira analise descritiva dos dados (linha 45).

```
plot(reg_multipla, lty=0, pch=19, col="purple")
```

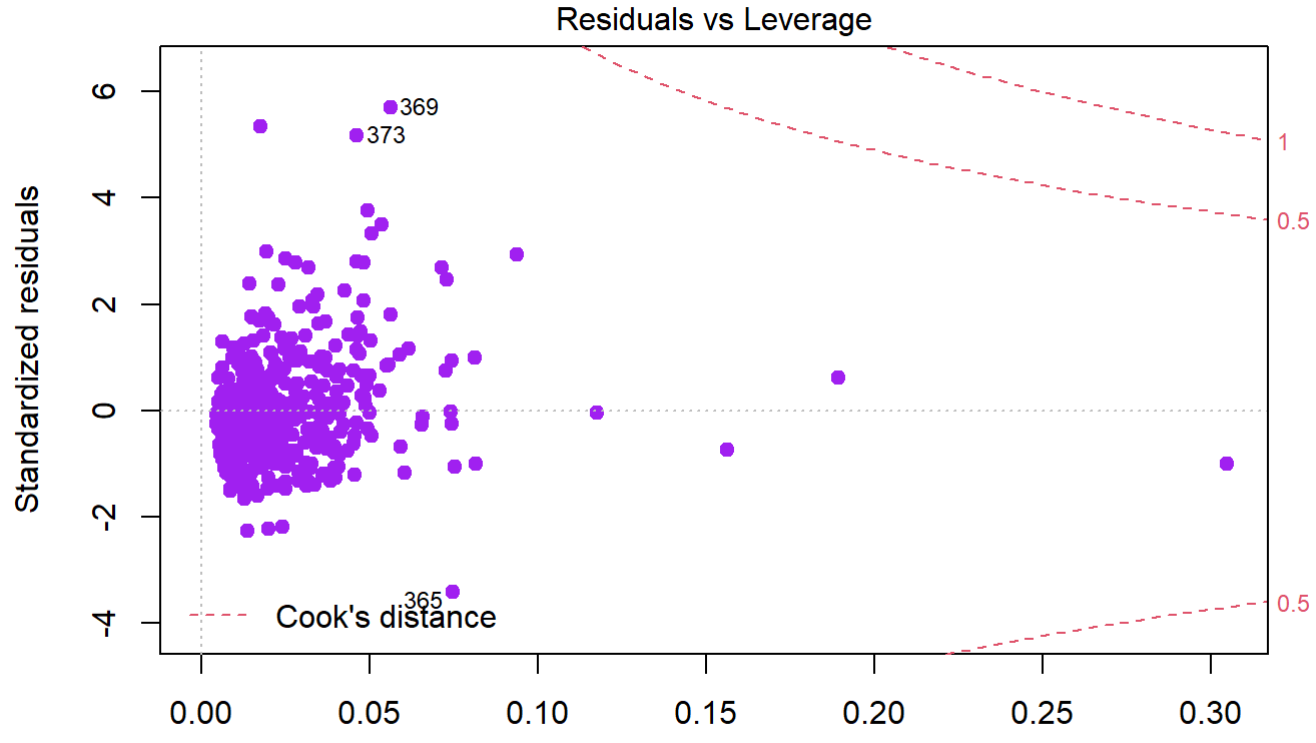

Residuals vs Fitted
Residuals
Fitted values
lm(MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO + B + LST .
369
372
373
30
20
10
0
-10
-20
0
10
20
30
40


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO + B + LST .
369
372
373
6
4
2
0
-2
-3
-2
-1
0
1
2
3


Scale-Location



lm(MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO + B + LST .



lm(MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO + B + LST .

install_tinytex()