Regressão Múltipla

Load data

```
boston = readxl::read_excel("boston.xlsx")
b = boston[,-1]
```

Manipulacao inicial dos dados

Adicionando labels

```
b$chas=as.factor(b$chas)
levels(b$chas)=c("otherwise", "bounds river")
```

Sumario dos dados

```
summary(b)
```

```
##       crim                zn             indus                  chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   otherwise   :471
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   bounds river: 35
##  Median : 0.25651   Median :  0.00   Median : 9.69
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74
##       nox               rm             age              dis
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       rad              tax           ptratio          lstat
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   : 1.73
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.: 6.95
##  Median : 5.000   Median :330.0   Median :19.05   Median :11.36
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :12.65
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.95
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :37.97
##       medv
##  Min.   : 5.00
##  1st Qu.:17.02
##  Median :21.20
##  Mean   :22.53
##  3rd Qu.:25.00
##  Max.   :50.00
```

Transformando log

```
b$crim = log(b$crim)
```

Regressao incial

Fazendo Regressão com todas as variaveis

```
reg.mlt=lm(data=b, medv ~ crim + zn + indus + chas + nox + rm + age + dis +
             rad + tax + ptratio + lstat)

summary(reg.mlt)
```

```
## 
## Call:
## lm(formula = medv ~ crim + zn + indus + chas + nox + rm + age + 
##     dis + rad + tax + ptratio + lstat, data = b)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -14.5196  -2.7591  -0.6185   1.8580  26.8435 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       42.038331   5.128612   8.197 2.15e-15 ***
## crim               0.241790   0.279864   0.864 0.388033    
## zn                 0.045271   0.014355   3.154 0.001710 ** 
## indus              0.016041   0.063182   0.254 0.799692    
## chasbounds river   3.031613   0.879877   3.445 0.000619 ***
## nox              -18.950940   4.038672  -4.692 3.50e-06 ***
## rm                 3.676786   0.425626   8.639  < 2e-16 ***
## age                0.002473   0.013590   0.182 0.855696    
## dis               -1.400774   0.202931  -6.903 1.58e-11 ***
## rad                0.184917   0.077317   2.392 0.017145 *  
## tax               -0.012508   0.003850  -3.249 0.001238 ** 
## ptratio           -0.912405   0.134699  -6.774 3.59e-11 ***
## lstat             -0.591869   0.051245 -11.550  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.86 on 493 degrees of freedom
## Multiple R-squared:  0.7274, Adjusted R-squared:  0.7208 
## F-statistic: 109.6 on 12 and 493 DF,  p-value: < 2.2e-16
```

Testando multicolinearidade

VIF > 5 indica alta chance de multicolinearidade.

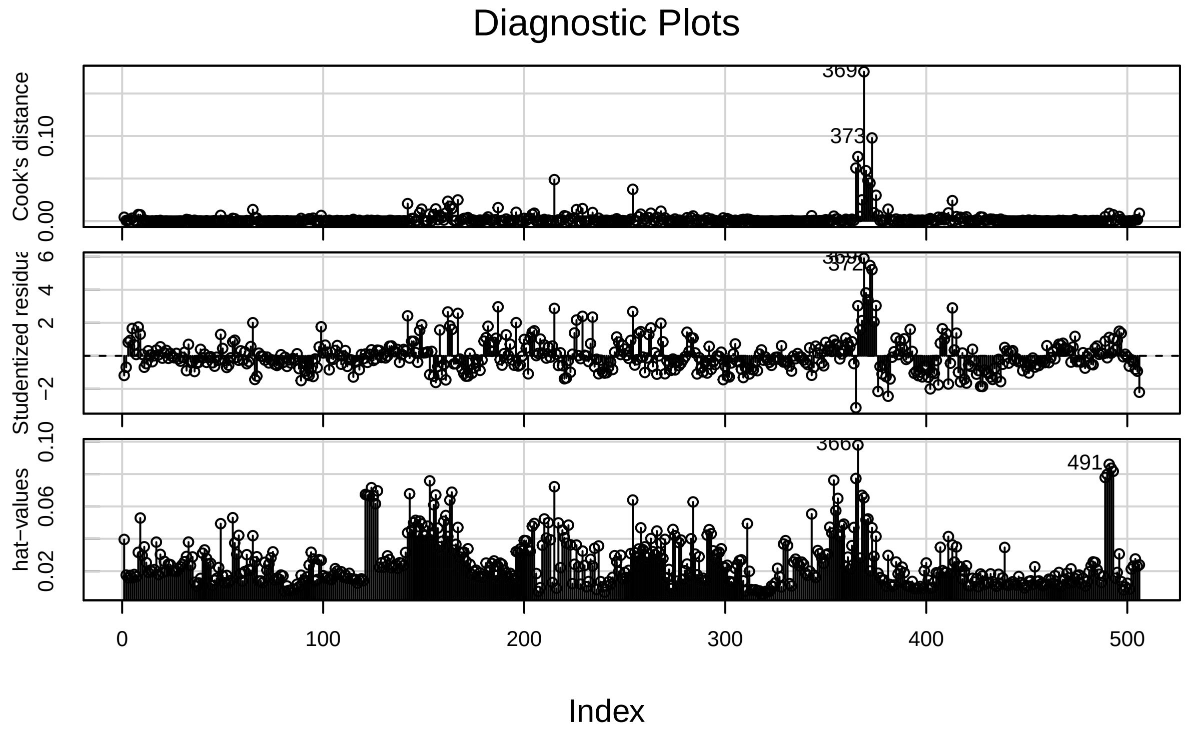
```
round(vif(reg.mlt),1)
```

```
##             crim               zn            indus chasbounds river 
##              7.8              2.4              4.0              1.1 
##              nox           rm          age          dis
##              4.7          1.9          3.1          3.9
##              rad          tax      ptratio        lstat
##              9.7          9.0          1.8          2.9
```

Detecção de anomalias

Cooks Distances -> Pontos Influentes Studentized residuals -> Outliers em Y hat-values -> Outliers em X

```
influenceIndexPlot(reg.mlt , vars=c("Cook","Studentized","hat"))
```



Regressao com seleção de variáveis

```
reg.mlt2=step(reg.mlt)
```

```
## Start:  AIC=1612.79
## medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad + 
##     tax + ptratio + lstat
## 
##           Df Sum of Sq   RSS    AIC
## - age      1      0.78 11644 1610.8
## - indus    1      1.52 11645 1610.9
## - crim     1     17.63 11661 1611.6
## <none>                 11643 1612.8
## - rad      1    135.09 11778 1616.6
## - zn       1    234.91 11878 1620.9
## - tax      1    249.27 11893 1621.5
## - chas     1    280.37 11924 1622.8
## - nox      1    520.01 12163 1632.9
## - ptratio  1   1083.61 12727 1655.8
## - dis      1   1125.29 12769 1657.5
## - rm       1   1762.42 13406 1682.1
## - lstat    1   3150.54 14794 1732.0
##
## Step:  AIC=1610.82
## medv ~ crim + zn + indus + chas + nox + rm + dis + rad + tax +
##     ptratio + lstat
##
##           Df Sum of Sq   RSS    AIC
## - indus    1       1.5 11646 1608.9
## - crim     1      18.7 11663 1609.6
## <none>                 11644 1610.8
## - rad      1     134.8 11779 1614.7
## - zn       1     234.4 11878 1618.9
## - tax      1     248.7 11893 1619.5
## - chas     1     282.9 11927 1621.0
## - nox      1     537.4 12182 1631.7
## - ptratio  1    1087.5 12732 1654.0
## - dis      1    1247.3 12891 1660.3
## - rm       1    1850.8 13495 1683.5
## - lstat    1    3461.5 15106 1740.5
##
## Step:  AIC=1608.89
## medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
##     lstat
##
##           Df Sum of Sq   RSS    AIC
## - crim     1      19.9 11666 1607.8
## <none>                 11646 1608.9
## - rad      1     138.9 11784 1612.9
## - zn       1     232.9 11878 1616.9
## - tax      1     287.4 11933 1619.2
## - chas     1     290.3 11936 1619.3
## - nox      1     555.4 12201 1630.5
## - ptratio  1    1097.4 12743 1652.5
## - dis      1    1324.2 12970 1661.4
## - rm       1    1857.1 13503 1681.8
## - lstat    1    3465.5 15111 1738.7
##
## Step:  AIC=1607.75
## medv ~ zn + chas + nox + rm + dis + rad + tax + ptratio + lstat
##
##           Df Sum of Sq   RSS    AIC
## <none>                 11666 1607.8
## - zn       1     213.6 11879 1614.9
## - rad      1     279.4 11945 1617.7
## - tax      1     282.9 11948 1617.9
## - chas     1     290.6 11956 1618.2
## - nox      1     548.0 12214 1629.0
## - ptratio  1    1133.4 12799 1652.7
## - dis      1    1361.4 13027 1661.6
## - rm       1    1867.5 13533 1680.9
## - lstat    1    3523.5 15189 1739.3
```

Novo sumario da regressao

```r
summary(reg.mlt2)
```

```
##
## Call:
## lm(formula = medv ~ zn + chas + nox + rm + dis + rad + tax +
##     ptratio + lstat, data = b)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7539  -2.7900  -0.6344   1.9798  26.9675
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         40.826438   4.962506   8.227 1.70e-15 ***
## zn                   0.041556   0.013788   3.014 0.002711 **
## chasbounds river     3.064999   0.871925   3.515 0.000480 ***
## nox                -17.390316   3.602542  -4.827 1.85e-06 ***
## rm                   3.691673   0.414283   8.911  < 2e-16 ***
## dis                 -1.436472   0.188803  -7.608 1.41e-13 ***
## rad                  0.212885   0.061760   3.447 0.000615 ***
## tax                 -0.011960   0.003449  -3.468 0.000569 ***
## ptratio             -0.916459   0.132017  -6.942 1.22e-11 ***
## lstat               -0.578306   0.047247 -12.240  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.85 on 496 degrees of freedom
## Multiple R-squared:  0.7269, Adjusted R-squared:  0.722
## F-statistic: 146.7 on 9 and 496 DF,  p-value: < 2.2e-16
```

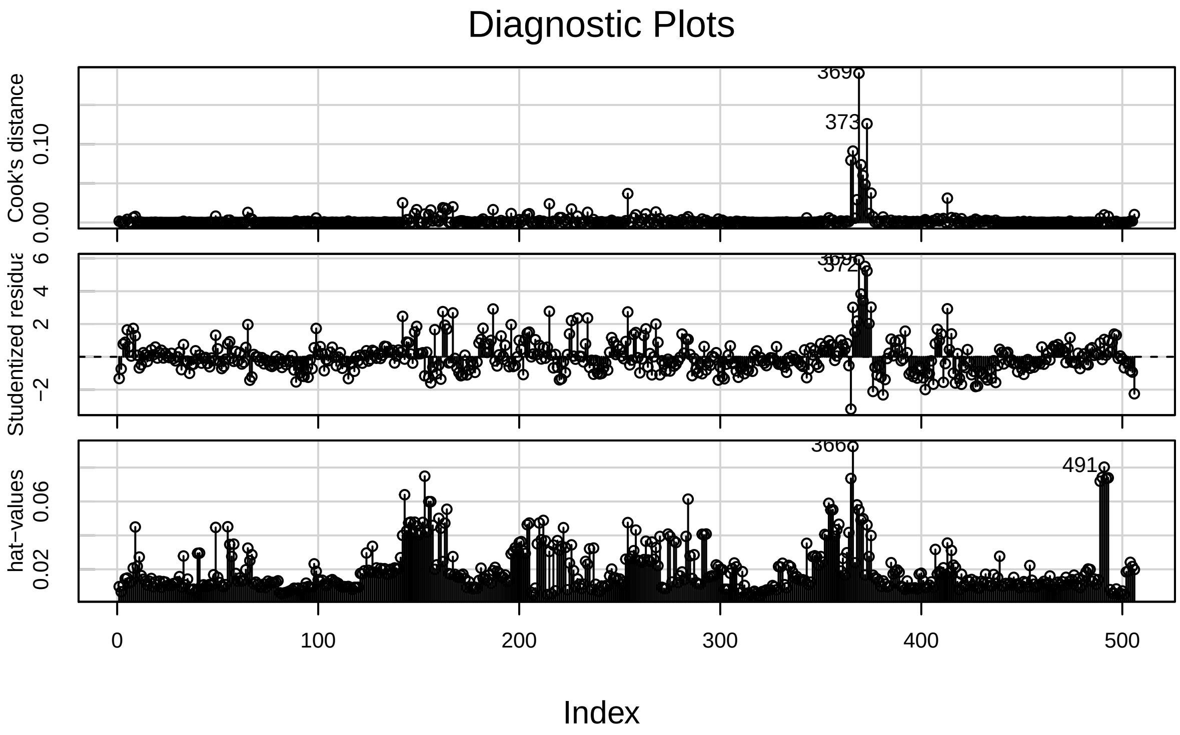
Nova deteccao de multicolinearidade

```r
round(vif(reg.mlt2),1)
```

```
##               zn chasbounds river              nox               rm
##              2.2              1.1              3.7              1.8
##              dis              rad              tax          ptratio
##              3.4              6.2              7.3              1.8
##            lstat
##              2.4
```

Novas anomalias

```r
influenceIndexPlot(reg.mlt2 , vars=c("Cook","Studentized","hat"))
```

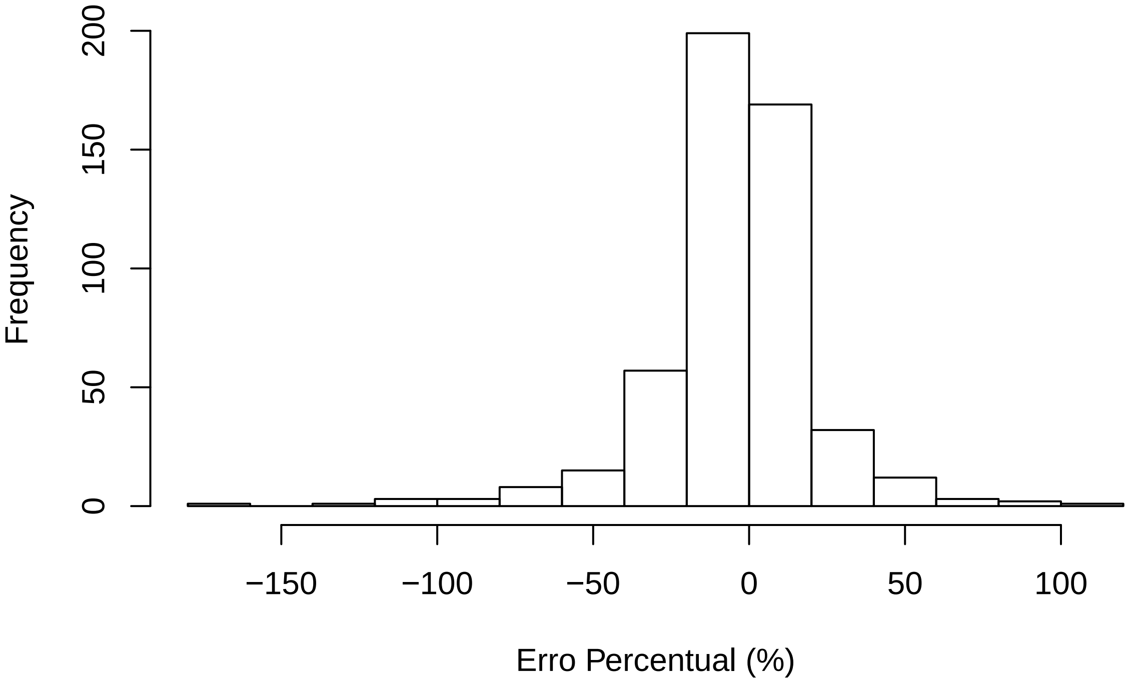


Criar previsoes

```r
b$medv_HAT=fitted.values(reg.mlt2) #Previsoes
b$RES=residuals(reg.mlt2) #Resuduais das previsoes
b$EP=b$RES/b$medv*100 #Erro percentual das previsoes
```

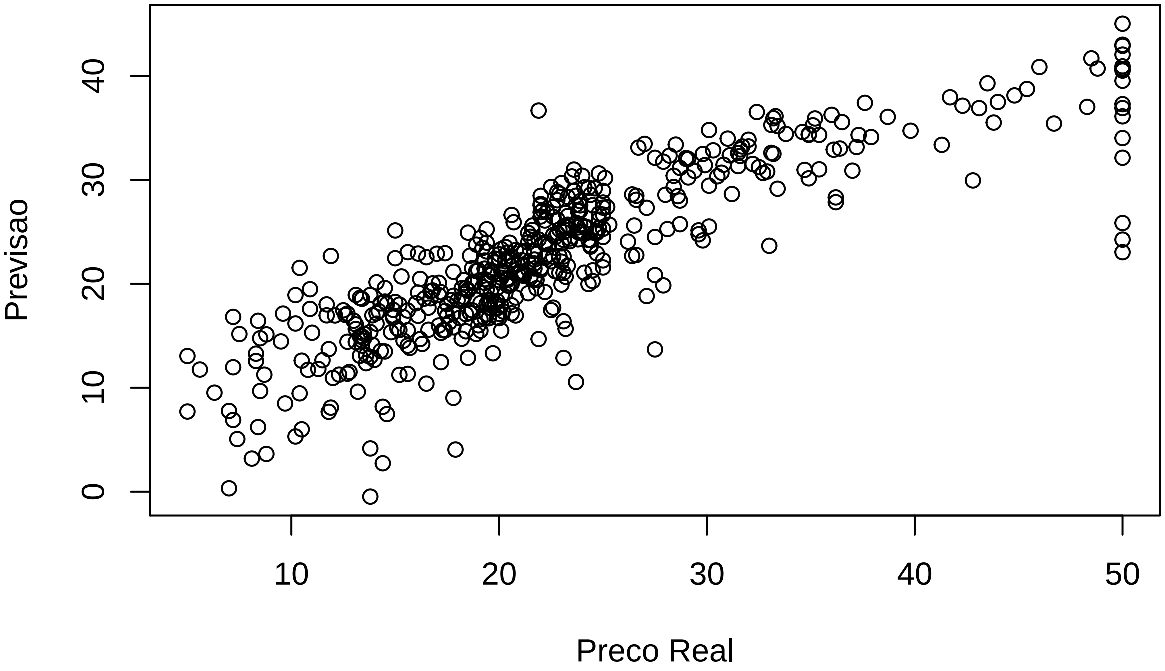
Erro percentual

```r
hist(b$EP, xlab = 'Erro Percentual (%)', main = '')
```



Previsao e real

```r
plot(x = b$medv, y = b$medv_HAT, xlab = 'Preco Real', ylab = 'Previsao')
```



Teste anova

```r
anova(reg.mlt2)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## zn          1  5549.7  5549.7 235.968 < 2.2e-16 ***
## chas        1  1555.5  1555.5  66.136 3.397e-15 ***
## nox         1  3793.4  3793.4 161.290 < 2.2e-16 ***
## rm          1 12955.2 12955.2 550.837 < 2.2e-16 ***
## dis         1   802.5   802.5  34.123 9.371e-09 ***
## rad         1   745.1   745.1  31.679 3.046e-08 ***
## tax         1   638.8   638.8  27.161 2.751e-07 ***
## ptratio     1  1487.2  1487.2  63.233 1.251e-14 ***
## lstat       1  3523.5  3523.5 149.816 < 2.2e-16 ***
## Residuals 496 11665.5    23.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```