Regressão Múltipla

Load data

```
boston = readxl::read_excel("boston.xlsx")
b = boston#[,-1]
#b = b[-c(365),]
```

Manipulacao inicial dos dados

Adicionando labels

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

Sumario dos dados

```
summary(b)
```

```
##        id             crim                zn             indus
##  Min.   :  1.0   Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46
##  1st Qu.:127.2   1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19
##  Median :253.5   Median : 0.25651   Median :  0.00   Median : 9.69
##  Mean   :253.5   Mean   : 3.61352   Mean   : 11.36   Mean   :11.14
##  3rd Qu.:379.8   3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10
##  Max.   :506.0   Max.   :88.97620   Max.   :100.00   Max.   :27.74
##            chas          nox               rm              age
##  otherwise   :471   Min.   :0.3850   Min.   :3.561   Min.   :  2.90
##  bounds river: 35   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02
##                     Median :0.5380   Median :6.208   Median : 77.50
##                     Mean   :0.5547   Mean   :6.285   Mean   : 68.57
##                     3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08
##                     Max.   :0.8710   Max.   :8.780   Max.   :100.00
##       dis              rad              tax           ptratio
##  Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60
##  1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40
##  Median : 3.207   Median : 5.000   Median :330.0   Median :19.05
##  Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46
##  3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20
##  Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

Filtrando dados

```
b = filter(b,medv <50)

#b$dis = ifelse(b$dis >= 3, 3, b$dis)
b$rad = ifelse(b$rad >= 9, 9, b$rad)
b$tax = ifelse(b$tax >= 500, 500, b$tax)
#b$nox = ifelse(b$nox >= 0.8, 0.75, b$nox)
#b$rm = ifelse(b$rm >= 7.5, 7.5, b$rm)

#b = filter(b, id != 366)
```

Ajustando valors

```
b = filter(b,medv <50)
```

Transformando log

```
#b$crim = log(b$crim)
b$lstat = log(b$lstat)
b$dis = log(b$dis)
#b$medv = log(b$medv)
```

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 
## -10.764  -2.142  -0.445   1.824  11.331 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       52.449044   3.898524  13.454  < 2e-16 ***
## crim              -0.140924   0.023190  -6.077 2.51e-09 ***
## zn                 0.011309   0.009855   1.148    0.252    
## indus             -0.044448   0.045653  -0.974    0.331    
## chasbounds river   0.688448   0.690888   0.996    0.320    
## nox              -14.966018   2.939639  -5.091 5.13e-07 ***
## rm                2.761520   0.344175   8.024 8.04e-15 ***
## age              -0.003908   0.010400  -0.376    0.707    
## dis              -4.975750   0.671793  -7.407 5.92e-13 ***
## rad               0.516139   0.104161   4.955 1.01e-06 ***
## tax              -0.015118   0.003071  -4.923 1.18e-06 ***
## ptratio          -0.730224   0.094220  -7.750 5.58e-14 ***
## lstat            -6.895288   0.531185 -12.981  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.494 on 477 degrees of freedom
## Multiple R-squared:  0.8075, Adjusted R-squared:  0.8026 
## F-statistic: 166.7 on 12 and 477 DF,  p-value: < 2.2e-16
```

Testando multicolinearidade

VIF > 5 indica alta chance de multicolinearidade.

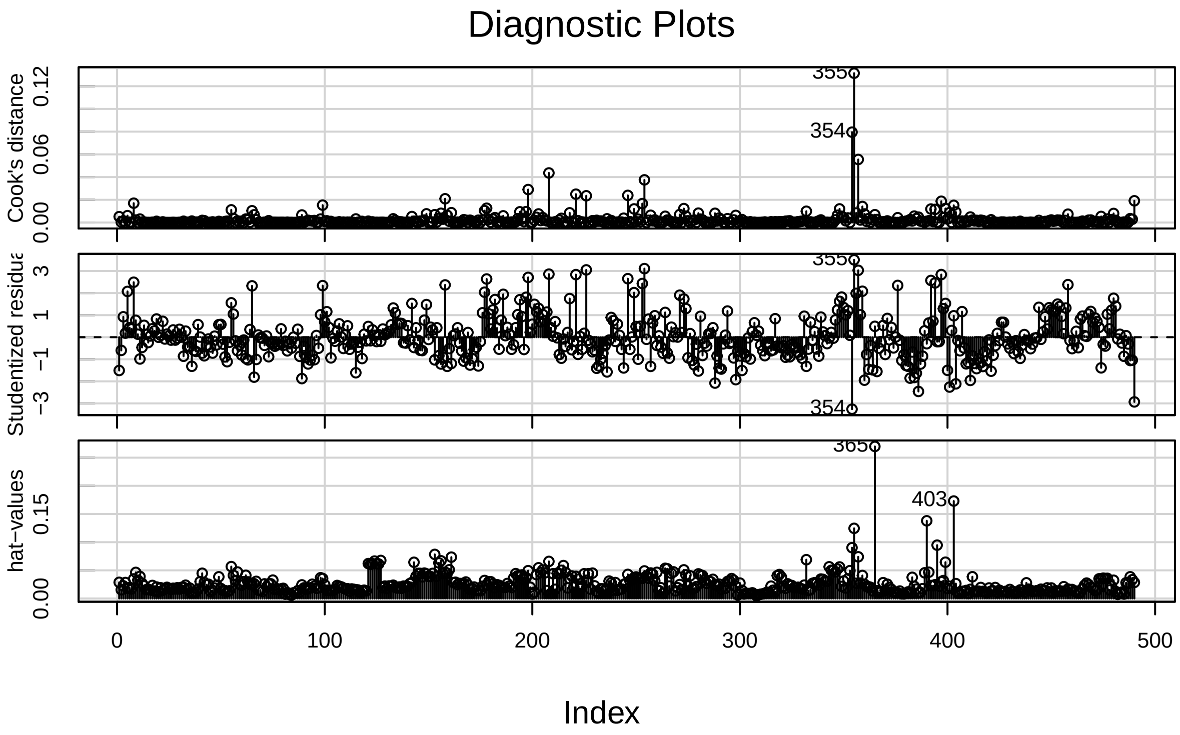
```
round(vif(reg.mlt),1)
```

```
##             crim               zn            indus chasbounds river 
##              1.6              2.0              3.9              1.1 
##              nox               rm              age              dis 
##              4.7              2.0              3.4              5.2 
##              rad              tax          ptratio            lstat 
##              2.6              3.9              1.6              3.7 
```

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=1238.98
## medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
##     tax + ptratio + lstat
##
##           Df Sum of Sq    RSS    AIC
## - age      1      1.72 5826.4 1237.1
## - indus    1     11.57 5836.3 1238.0
## - chas     1     12.13 5836.8 1238.0
## - zn       1     16.08 5840.8 1238.3
## <none>                 5824.7 1239.0
## - tax      1    295.90 6120.6 1261.3
## - rad      1    299.83 6124.6 1261.6
## - nox      1    316.51 6141.2 1262.9
## - crim     1    450.95 6275.7 1273.5
## - dis      1    669.89 6494.6 1290.3
## - ptratio  1    733.47 6558.2 1295.1
## - rm       1    786.13 6610.8 1299.0
## - lstat    1   2057.64 7882.4 1385.2
##
## Step:  AIC=1237.12
## medv ~ crim + zn + indus + chas + nox + rm + dis + rad + tax +
##     ptratio + lstat
##
##           Df Sum of Sq    RSS    AIC
## - indus    1     11.07 5837.5 1236.0
## - chas     1     11.84 5838.3 1236.1
## - zn       1     17.39 5843.8 1236.6
## <none>                5826.4 1237.1
## - tax      1    294.37 6120.8 1259.3
## - rad      1    301.96 6128.4 1259.9
## - nox      1    336.84 6163.3 1262.7
## - crim     1    449.24 6275.7 1271.5
## - ptratio  1    740.55 6567.0 1293.8
## - dis      1    740.68 6567.1 1293.8
## - rm       1    828.74 6655.2 1300.3
## - lstat    1   2603.44 8429.9 1416.1
##
## Step:  AIC=1236.05
## medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
##     lstat
##
##           Df Sum of Sq    RSS    AIC
## - chas     1     10.31 5847.8 1234.9
## - zn       1     20.63 5858.1 1235.8
## <none>                5837.5 1236.0
## - rad      1    330.18 6167.7 1261.0
## - nox      1    383.80 6221.3 1265.2
## - tax      1    388.43 6225.9 1265.6
## - crim     1    442.40 6279.9 1269.8
## - dis      1    744.80 6582.3 1292.9
## - ptratio  1    795.62 6633.1 1296.7
## - rm       1    848.48 6686.0 1300.5
## - lstat    1   2669.94 8507.5 1418.6
##
## Step:  AIC=1234.92
## medv ~ crim + zn + nox + rm + dis + rad + tax + ptratio + lstat
##
##           Df Sum of Sq    RSS    AIC
## - zn       1     20.14 5868.0 1234.6
## <none>                5847.8 1234.9
## - rad      1    344.49 6192.3 1261.0
## - nox      1    375.18 6223.0 1263.4
## - tax      1    413.69 6261.5 1266.4
## - crim     1    452.99 6300.8 1269.5
## - dis      1    744.58 6592.4 1291.6
## - ptratio  1    815.92 6663.7 1296.9
## - rm       1    855.49 6703.3 1299.8
## - lstat    1   2663.57 8511.4 1416.8
##
## Step:  AIC=1234.6
## medv ~ crim + nox + rm + dis + rad + tax + ptratio + lstat
##
##           Df Sum of Sq    RSS    AIC
## <none>                5868.0 1234.6
## - rad      1    326.38 6194.3 1259.1
## - nox      1    392.34 6260.3 1264.3
## - tax      1    395.85 6263.8 1264.6
## - crim     1    433.55 6301.5 1267.5
## - dis      1    757.11 6625.1 1292.1
## - rm       1    870.67 6738.6 1300.4
## - ptratio  1   1001.80 6869.8 1309.8
## - lstat    1   2770.00 8638.0 1422.1
```

Novo sumario da regressao

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

```
##
## Call:
## lm(formula = medv ~ crim + nox + rm + dis + rad + tax + ptratio +
##     lstat, data = b)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2445  -2.2169  -0.3962   1.8000  11.3857
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  53.285115   3.851023  13.837  < 2e-16 ***
## crim         -0.135371   0.022708  -5.961 4.84e-09 ***
## nox         -15.828758   2.791156  -5.671 2.45e-08 ***
## rm            2.780929   0.329181   8.448 3.54e-16 ***
## dis          -4.449171   0.564771  -7.878 2.24e-14 ***
## rad           0.521068   0.100740   5.172 3.40e-07 ***
## tax          -0.015417   0.002707  -5.696 2.13e-08 ***
## ptratio      -0.792582   0.087463  -9.062  < 2e-16 ***
## lstat        -7.099162   0.471127 -15.068  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.493 on 481 degrees of freedom
## Multiple R-squared:  0.806,  Adjusted R-squared:  0.8028
## F-statistic: 249.8 on 8 and 481 DF,  p-value: < 2.2e-16
```

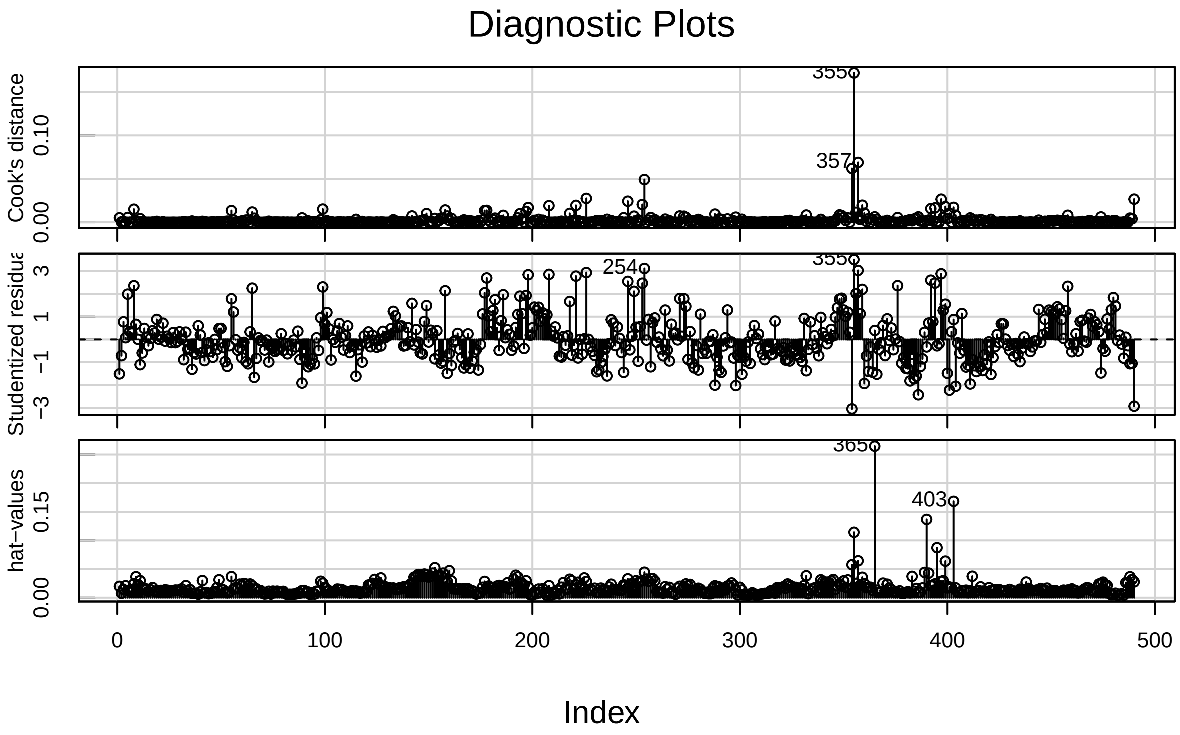
Nova deteccao de multicolinearidade

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

```
##    crim     nox      rm     dis     rad     tax ptratio   lstat
##     1.6     4.3     1.9     3.7     2.4     3.0     1.4     2.9
```

Novas anomalias

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

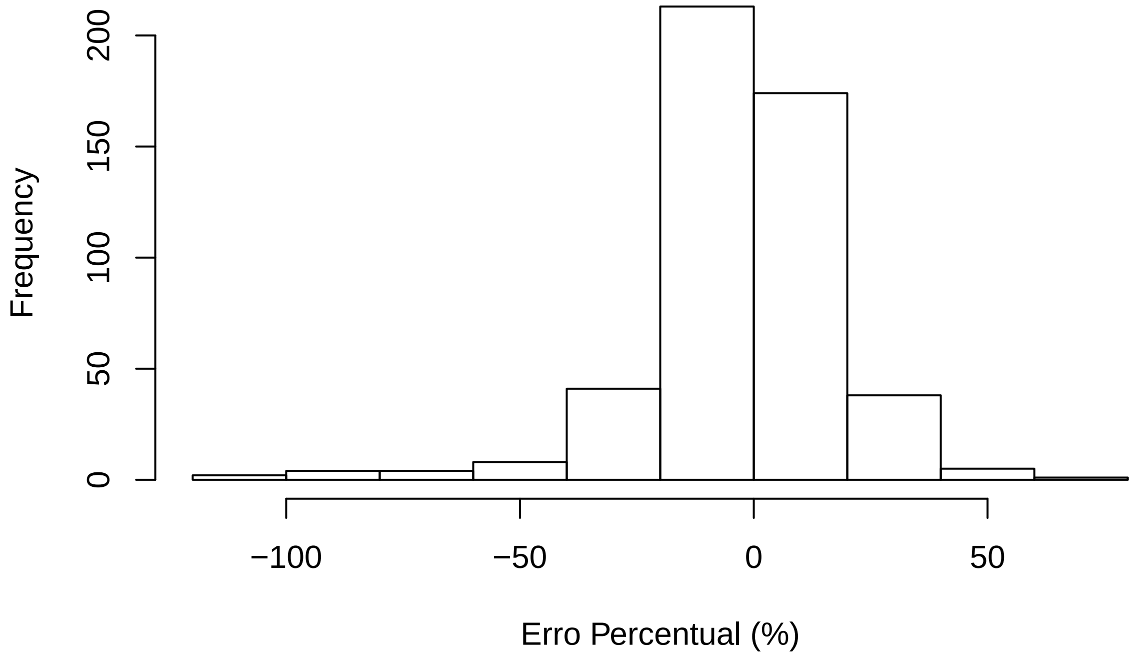


Criar previsoes

```
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

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



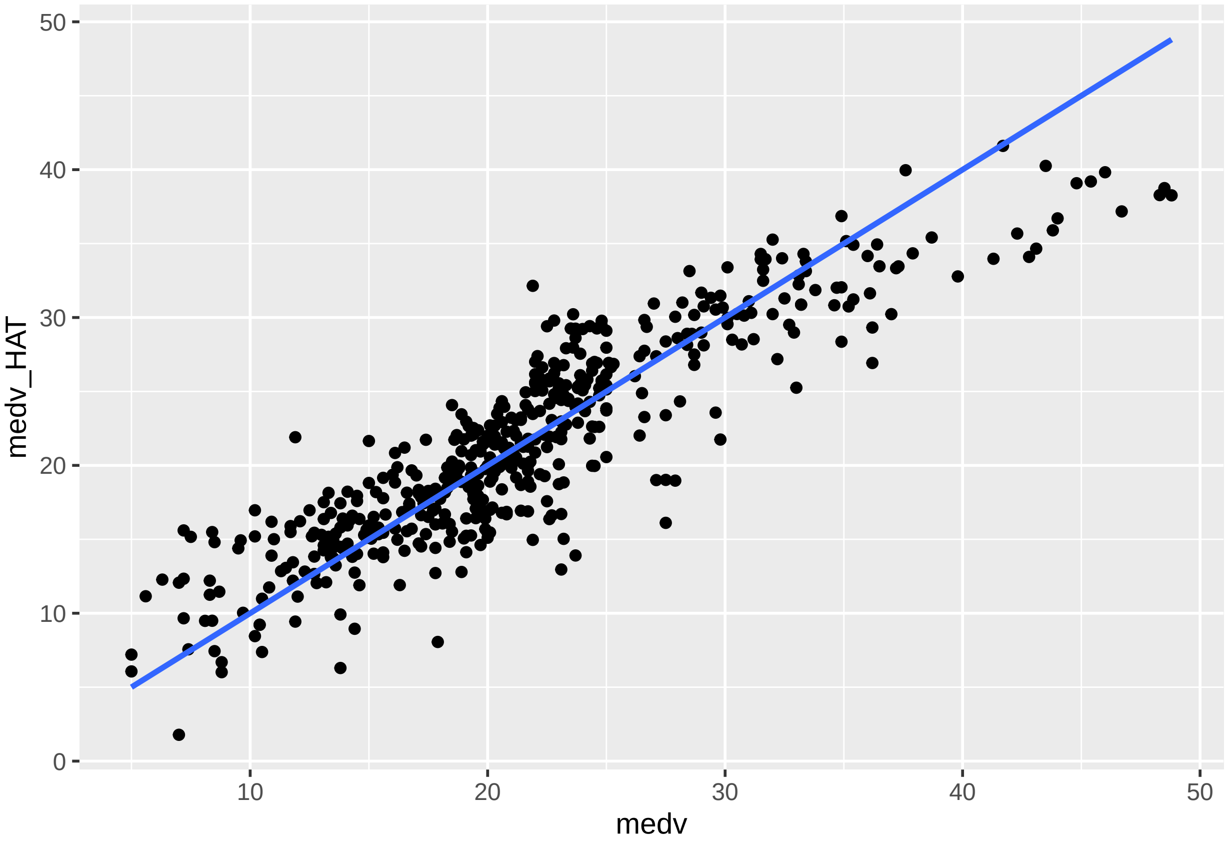
Previsao e real

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

p = ggplot(b) +
  geom_point(aes(x = medv, y = medv_HAT,
                 name = id #rownames(b)
                 )) +
  geom_smooth(method='lm',aes(x  = medv, y = medv) )
```

```
## Warning: Ignoring unknown aesthetics: name
```

```
#ggplotly(p)
p
```



Teste anova

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

```
## Analysis of Variance Table
##
## Response: medv
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## crim        1 6129.0  6129.0 502.3943 < 2.2e-16 ***
## nox         1 4128.7  4128.7 338.4309 < 2.2e-16 ***
## rm          1 8322.5  8322.5 682.2021 < 2.2e-16 ***
## dis         1  171.9   171.9  14.0905 0.0001955 ***
## rad         1   54.8    54.8   4.4911 0.0345843 *
## tax         1 1038.3  1038.3  85.1134 < 2.2e-16 ***
## ptratio     1 1767.8  1767.8 144.9093 < 2.2e-16 ***
## lstat       1 2770.0  2770.0 227.0586 < 2.2e-16 ***
## Residuals 481 5868.0    12.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Root mean sqared error

```r
mean((b$medv - b$medv_HAT) ** 2) **0.5
```

```
## [1] 3.460554
```