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$rm)
b$zn = ifelse(b$zn > 0, 1, 0)

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

Ajustando valores

```
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
## -10.8140  -2.1758  -0.4094   1.8289  11.3871
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       52.535978   3.916696  13.413  < 2e-16 ***
## crim              -0.138100   0.023065  -5.987 4.20e-09 ***
## zn                 0.277078   0.518403   0.534    0.593
## indus             -0.048217   0.045578  -1.058    0.291
## chasbounds river   0.698504   0.692180   1.009    0.313
```

```
## nox             -15.016129   2.943598  -5.101 4.88e-07 ***
## rm                2.774363   0.344769   8.047 6.79e-15 ***
## age              -0.004892   0.010365  -0.472    0.637
## dis              -4.884288   0.680164  -7.181 2.67e-12 ***
## rad               0.490316   0.102869   4.766 2.49e-06 ***
## tax              -0.014161   0.002941  -4.815 1.98e-06 ***
## ptratio          -0.741094   0.097881  -7.571 1.93e-13 ***
## lstat            -6.936314   0.530948 -13.064  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.498 on 477 degrees of freedom
## Multiple R-squared:  0.807,  Adjusted R-squared:  0.8022
## F-statistic: 166.2 on 12 and 477 DF,  p-value: < 2.2e-16
```

Testando multicolinearidade

VIF > 5 indica alta chance de multicolinearidade.

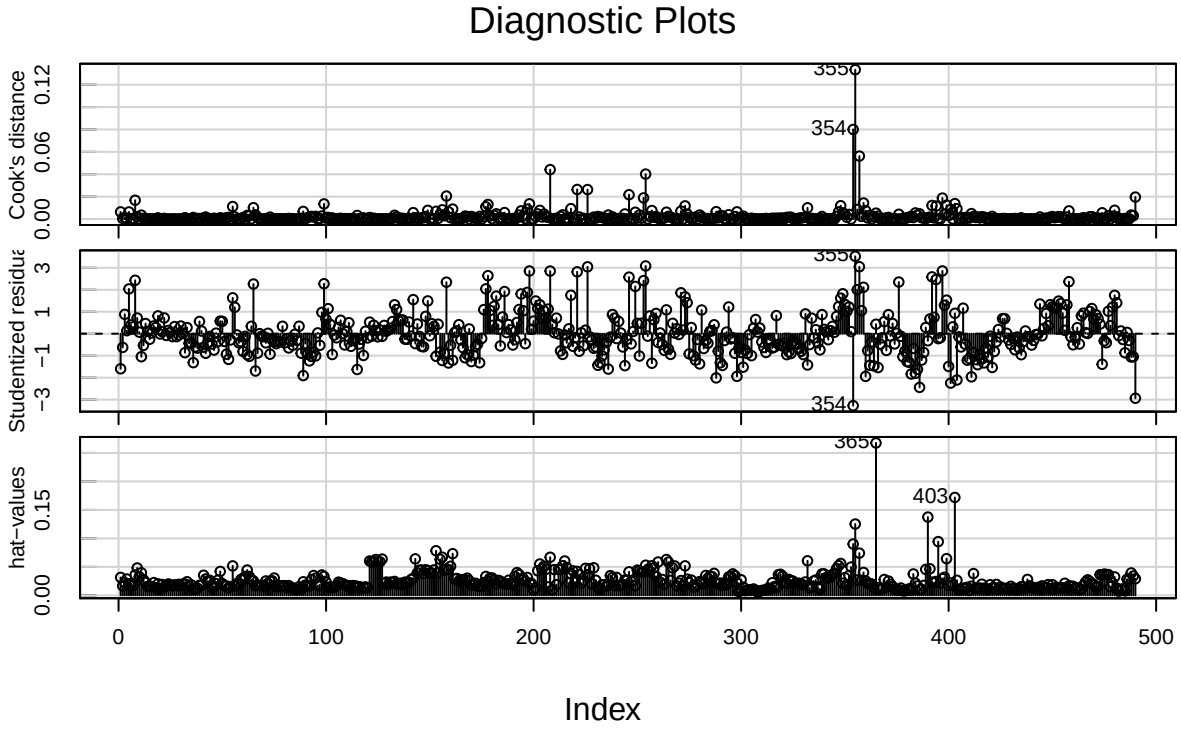
```
round(vif(reg.mlt),1)
```

```
##             crim               zn            indus chasbounds            river
##              1.6              2.1              3.9              1.1
##              nox               rm              age              dis
##              4.7              2.0              3.4              5.3
##              rad              tax          ptratio            lstat
##              2.5              3.5              1.7              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=1240.03
## medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
##     tax + ptratio + lstat
##
##           Df Sum of Sq    RSS    AIC
## - age      1      2.73 5840.0 1238.3
## - zn       1      3.50 5840.8 1238.3
## - chas     1     12.46 5849.8 1239.1
## - indus    1     13.70 5851.0 1239.2
## <none>                 5837.3 1240.0
## - rad      1    278.02 6115.3 1260.8
## - tax      1    283.74 6121.0 1261.3
## - nox      1    318.46 6155.8 1264.1
## - crim     1    438.71 6276.0 1273.5
## - dis      1    631.06 6468.4 1288.3
## - ptratio  1    701.53 6538.8 1293.6
## - rm       1    792.44 6629.7 1300.4
## - lstat    1   2088.57 7925.9 1387.9
##
## Step:  AIC=1238.26
## medv ~ crim + zn + indus + chas + nox + rm + dis + rad + tax +
##     ptratio + lstat
##
##           Df Sum of Sq    RSS    AIC
## - zn       1      3.81 5843.8 1236.6
```

```
## - chas     1     12.11 5852.1 1237.3
## - indus    1     13.07 5853.1 1237.4
## <none>                 5840.0 1238.3
## - rad      1    278.91 6118.9 1259.1
## - tax      1    281.08 6121.1 1259.3
## - nox      1    341.70 6181.7 1264.1
## - crim     1    436.03 6276.1 1271.5
## - dis      1    695.08 6535.1 1291.4
## - ptratio  1    711.05 6551.1 1292.6
## - rm       1    828.92 6669.0 1301.3
## - lstat    1   2688.93 8529.0 1421.8
##
## Step:  AIC=1236.58
## medv ~ crim + indus + chas + nox + rm + dis + rad + tax + ptratio +
##     lstat
##
##           Df Sum of Sq    RSS    AIC
## - chas     1     11.57 5855.4 1235.5
## - indus    1     14.31 5858.1 1235.8
## <none>                 5843.8 1236.6
## - tax      1    278.16 6122.0 1257.4
## - rad      1    285.77 6129.6 1258.0
## - nox      1    346.31 6190.1 1262.8
## - crim     1    432.36 6276.2 1269.6
## - dis      1    745.77 6589.6 1293.4
## - rm       1    840.21 6684.0 1300.4
## - ptratio  1    886.12 6730.0 1303.8
## - lstat    1   2689.46 8533.3 1420.1
##
## Step:  AIC=1235.55
## medv ~ crim + indus + nox + rm + dis + rad + tax + ptratio +
##     lstat
##
##           Df Sum of Sq    RSS    AIC
## - indus    1     12.55 5868.0 1234.6
## <none>                 5855.4 1235.5
## - rad      1    302.01 6157.4 1258.2
## - tax      1    305.29 6160.7 1258.5
## - nox      1    338.43 6193.8 1261.1
## - crim     1    442.89 6298.3 1269.3
## - dis      1    742.69 6598.1 1292.1
## - rm       1    848.83 6704.2 1299.9
## - ptratio  1    913.32 6768.7 1304.6
## - lstat    1   2684.21 8539.6 1418.5
##
## 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

```
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
```

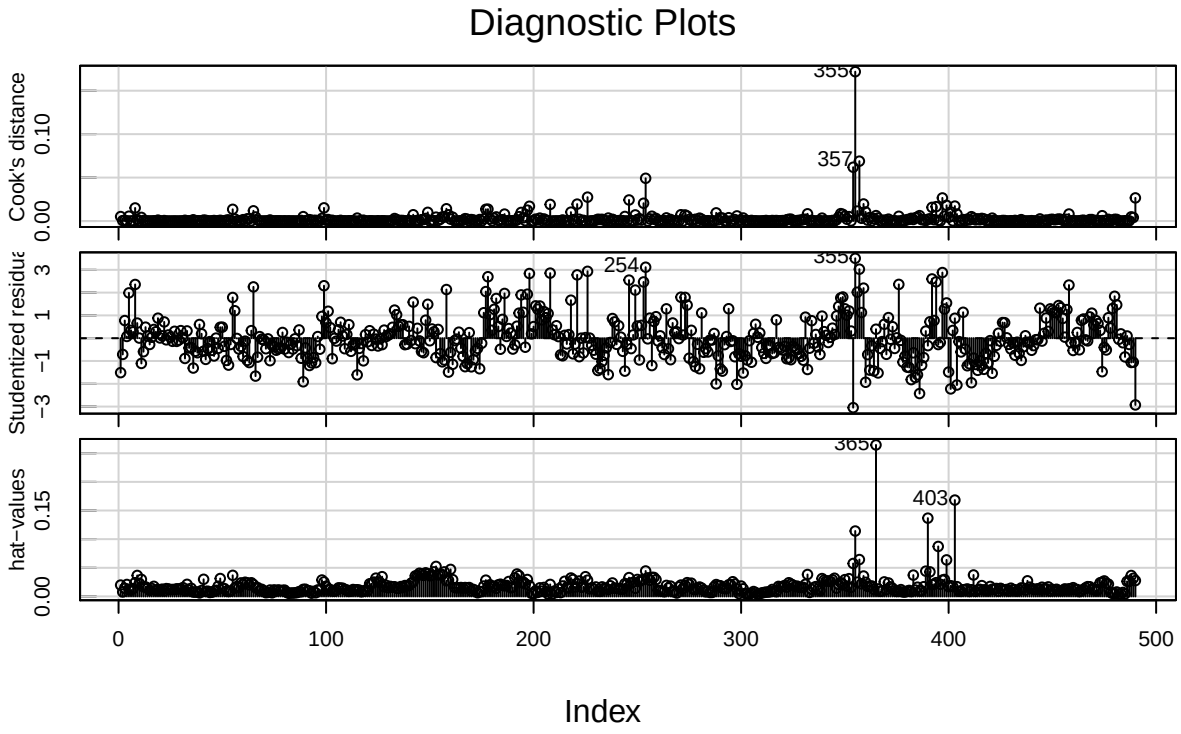
Nova deteccao de multicolinearidade

```
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

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

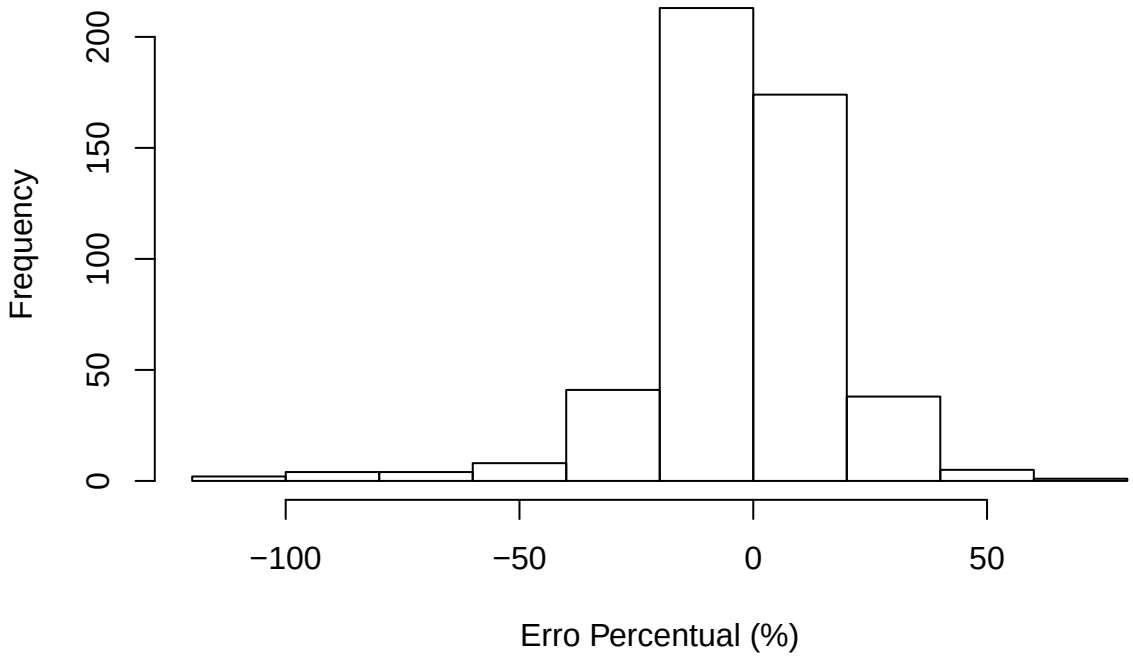


Criar previsoes

```
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 = '')
```



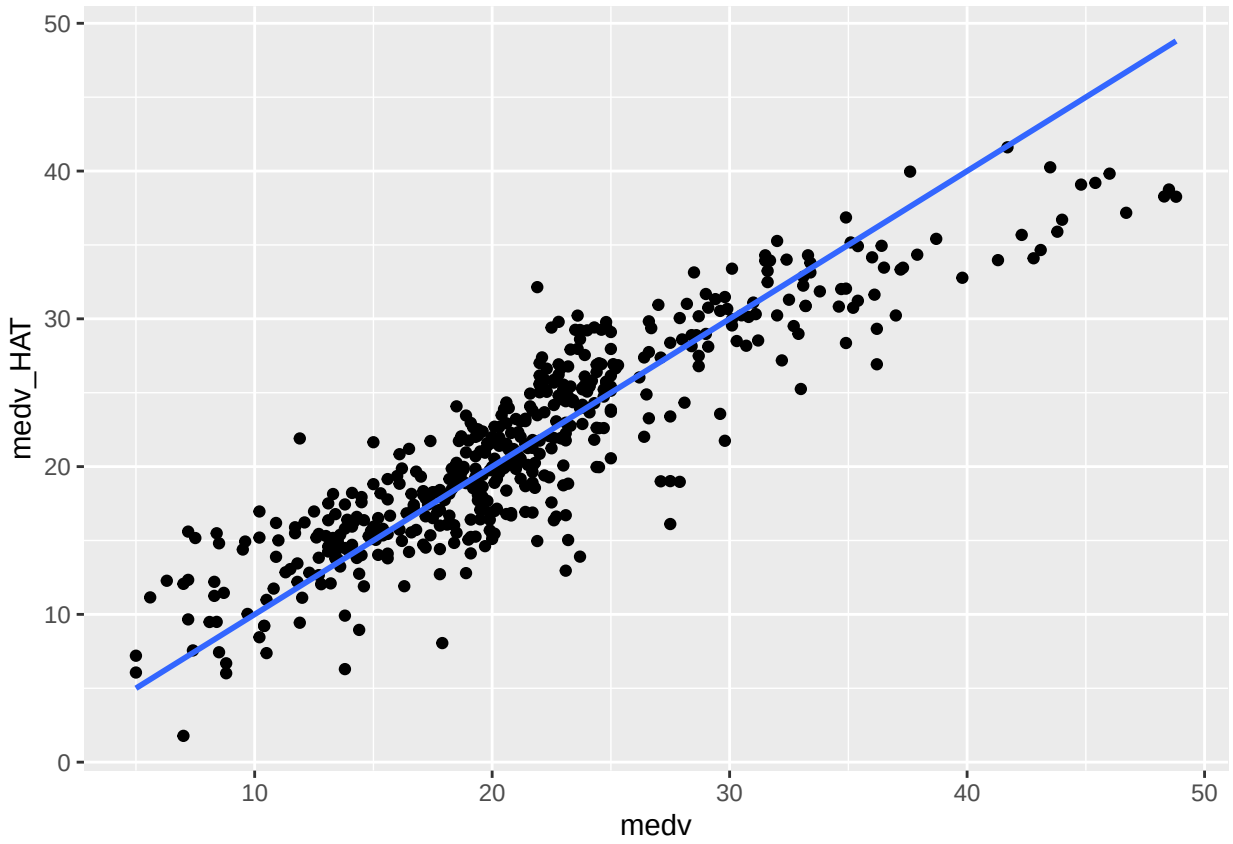
Previsao e real

```
#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

```
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

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

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

Fazer previsao

```
crim = 0.2651
zn = 0.0
indus = 9.69
chas = 0.0
nox = 0.5380
rm = 6.208
age = 77.50
dis = log(3.207)
rad = 5.0
tax = 330
ptratio = 19.05
lstat = log(11.36)

novo=data.frame(
  crim = crim,
  zn = zn,
  indus = indus,
  chas = chas,
  nox = nox,
  rm = rm,
  age = age,
  dis = dis,
  rad = rad,
  tax = tax,
  ptratio = ptratio,
  lstat = lstat
  )
lprice.hat=predict(reg.mlt2, novo)
lprice.hat
```

```
##        1
## 21.97985
```