Fundação Getúlio Vargas (FGV)
MBA Business Analytics e Big Data

Ana Paula Gonçalves dos Santos
Marcelo K. Costa

Trabalho apresentado para disciplina Análise Preditiva, como requisito para avaliação (Banco de Dados Boston).

Professor Abraham Laredo Sicsu

Brasília (DF), 30 de setembro de 2019.

Sumário

1. Introdução

O presente trabalho é requisito de avaliação da disciplina Análise Preditiva e compreende a análise e tratamento dos dados do *data frame* "Boston", disponibilizado pelo professor, abrangendo também a definição de um modelo de regressão múltipla, seleção de variáveis, avaliação da capacidade preditiva do modelo e a validação do mesmo.

Será utilizada a Linguagem R para execução das análises apresentadas a seguir.

O código fonte do trabalho está disponível no endereço https://github.com/anapaulagsantos/analise_preditiva.

2. Análise e tratamento dos dados

O data frame Boston fornece informações sobre os Valores de Habitação nos Subúrbios de Boston, sendo composto por 506 linhas e 13 colunas.

As colunas compreendem as seguintes variáveis, cujas descrições foram traduzidas para o Português.

Variável	Descrição da variável	Tipo de Variável
Crim	taxa de criminalidade per capita por cidade.	Quantitativa
Zn	proporção de terrenos residenciais divididos em lotes com mais de 25.000 pés quadrados	Quantitativa
Indus	proporção de hectares de negócios não comerciais por cidade	Quantitativa
Chas	variável do rio Charles (= 1 se o trecho limita o rio; 0 caso contrário)	Qualitativa
Nox	concentração de óxidos de nitrogênio (partes por 10 milhões)	Quantitativa
Rm	número médio de cômodos por moradia	Quantitativa
Age	proporção etária de unidades construídas antes de 1940, ocupadas pelos proprietários	Quantitativa
Dis	média ponderada das distâncias para cinco centros de emprego em Boston	Quantitativa
Rad	índice de acessibilidade às rodovias radiais	Qualitativa
Tax	valor total da taxa de imposto sobre a propriedade por \ $ 10.000	Quantitativa
Ptratio	proporção de alunos-professor por cidade	Quantitativa
Lstat	status da população (%)	Qualitativa
Medv	valor médio das casas ocupadas pelos proprietários em \ $ 1000s.	Quantitativa

2.1 Análise univariada

A análise univariada compreende a análise de cada variável isoladamente. Para tanto, será avaliada, por variável, a existência de outliers e missing values, além das medidas descritivas e gráficos.

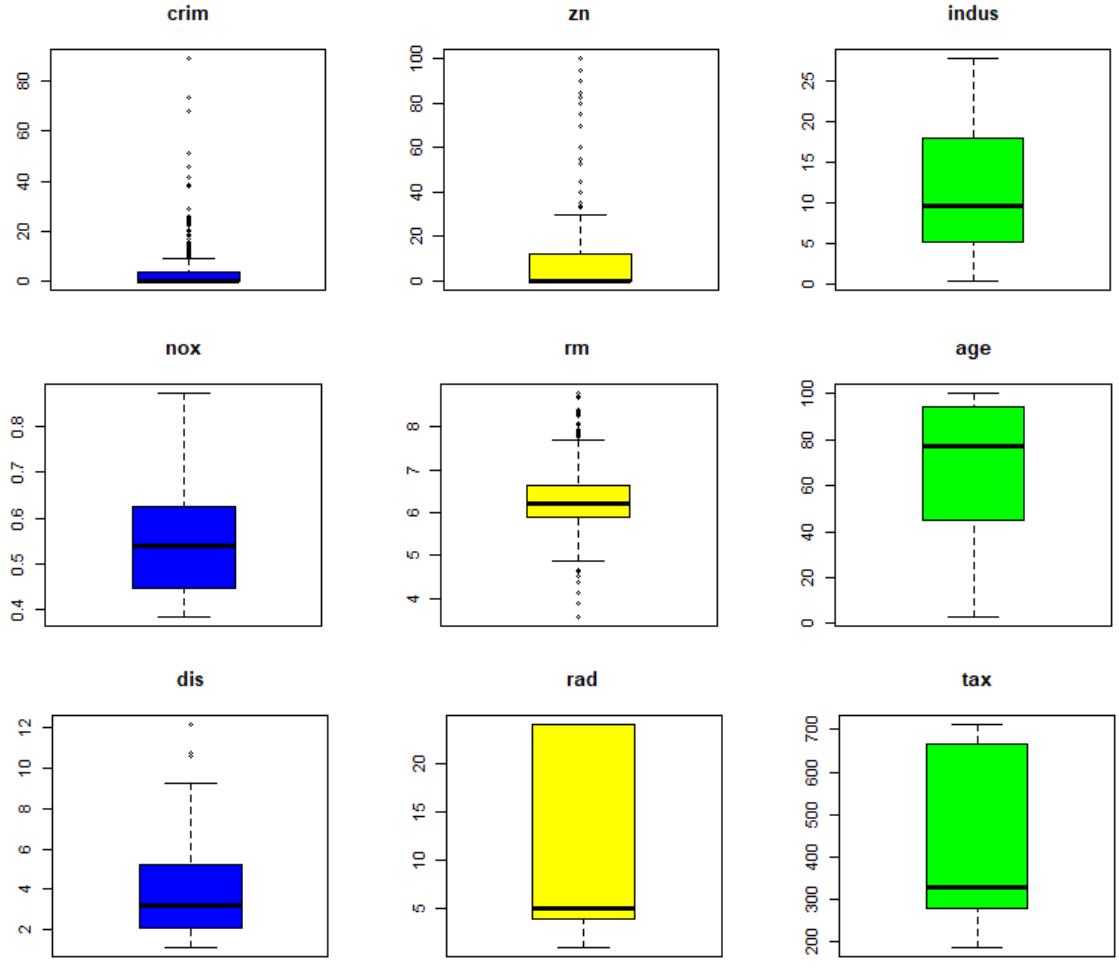
2.1.1 Missing values

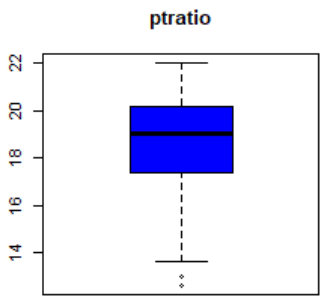
Missing values são valores ausentes nas observações de determinada variável.

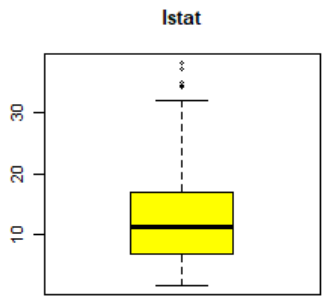
Não foram encontrados casos de missing value na base de dados fornecida.

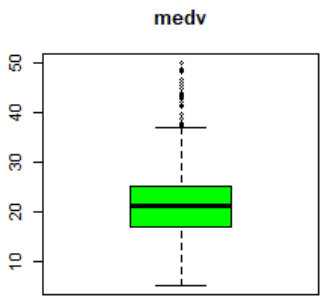
2.1.2 Outliers

Outliers são observações que se afastam da maioria dos dados da série avaliada. Para identificação de outliers foram gerados gráficos do tipo boxplot, exceto para a variável “chas”, (por se tratar de variável dummy):





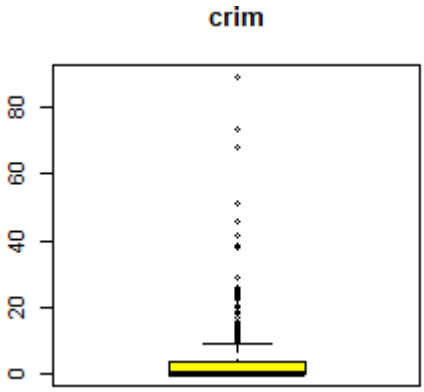


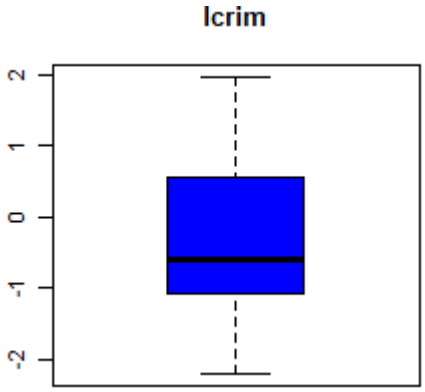


As variáveis que apresentaram outliers foram “crim”, “zn”, “rm”, “dis”, “ptratio”, “lstat” e “medv”.

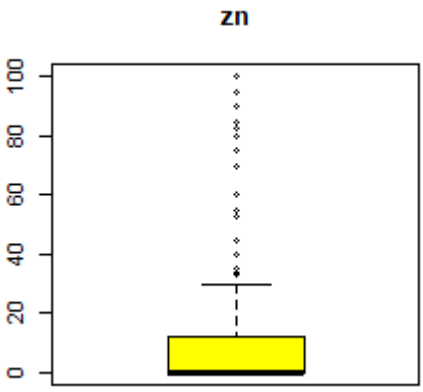
Para tratamento dos outliers foram utilizadas as técnicas descritas abaixo, sugeridas por Tabachnick and Fidell (2007) e Howell (2007):

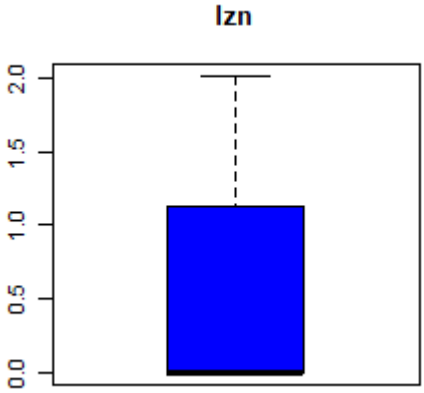
i. **Crim:** optou-se pela transformação dos dados por meio da função logaritmo (log10), sendo criada a variável lcrim.





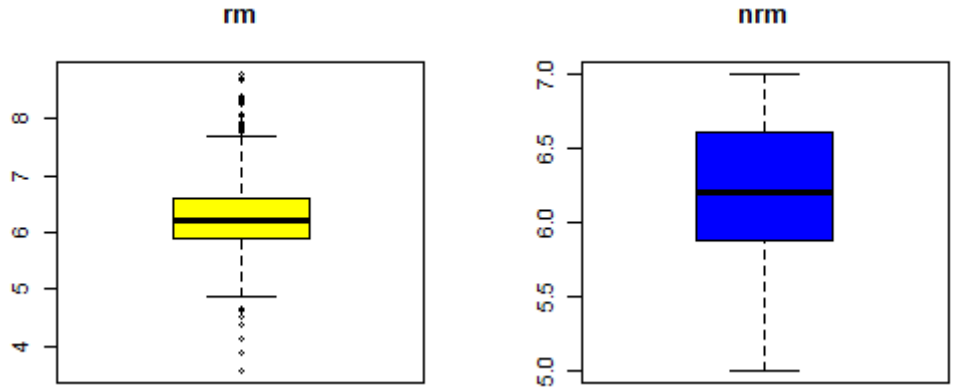
ii. **Zn:** optou-se pela transformação dos dados por meio da função logaritmo (log10 + 1 em virtude do valor mínimo de observação da variável ser igual a zero), sendo criada a variável lzn.



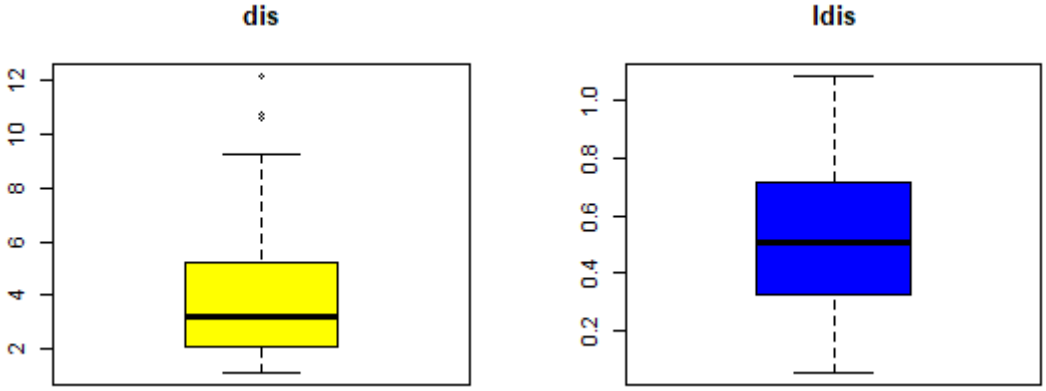


iii. **Rm:** a transformação dos dados por meio da função logaritmo não foi suficiente para o tratamento dos outliers dessa variável. Dessa forma, como se trata da descrição da média

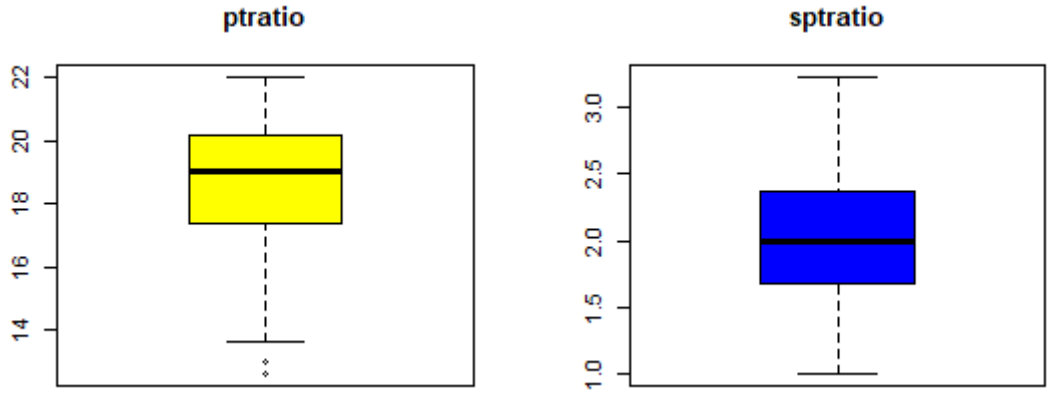
de cômodos por residência, optou-se por, a partir da variável original (rm), agrupar os imóveis com até 5 cômodos no mesmo grupo e os que possuem 7 ou mais em outro grupo, sendo criada a variável nrm.



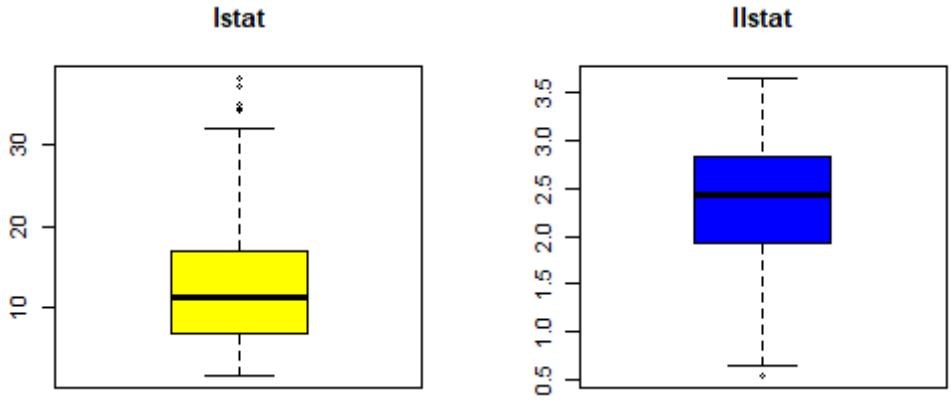
iv. **Dis:** optou-se pela transformação dos dados por meio da função logaritmo (log10), sendo criada a variável ldis.



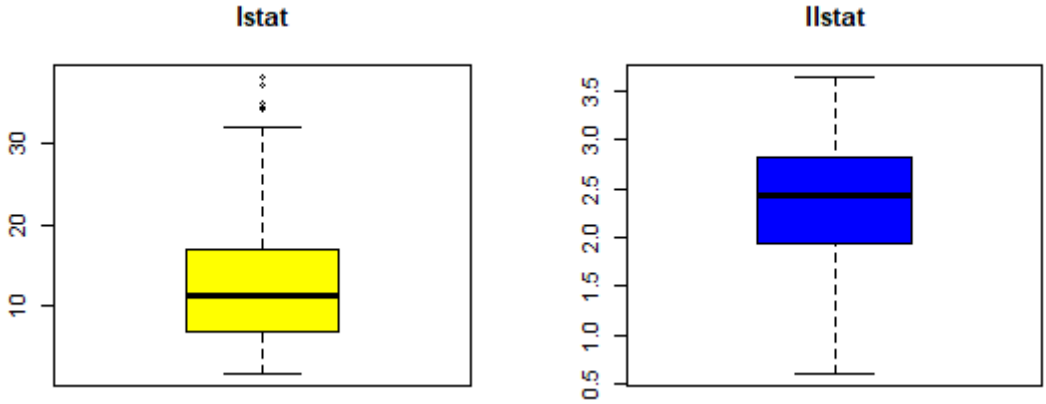
v. **Ptratio:** optou-se pela transformação dos dados por meio da função sqrt, sendo criada a variável sptratio.



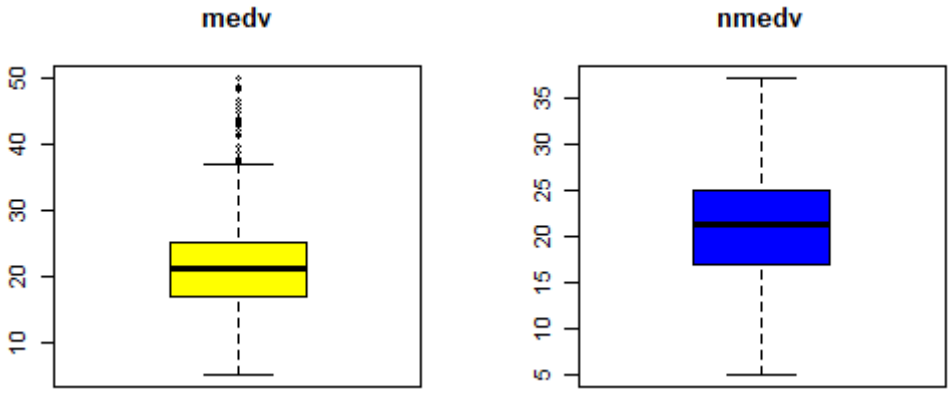
vi. **Lstat:** optou-se pela transformação dos dados por meio da função logaritmo, sendo criada a variável llstat.



Porém, como continuou apresentando 1 (um) outlier, foi utilizada a função ifelse, de forma a agrupar os valores abaixo do valor mínimo (consideramos 0,6 como mínimo para facilitar a transformação):



vii. **Medv**: a transformação dos dados por meio da função logaritmo não foi suficiente para o tratamento dos outliers dessa variável. Dessa forma, como se trata da descrição do valor médio das casas ocupadas, optou-se por agrupar os imóveis com valor médio superior a $37 no mesmo grupo, sendo criada a variável nmedv.



2.1.3 Medidas descritivas

Após os tratamentos listados no item 2.1.2 e seleção das variáveis substituídas, o data frame "nboston" apresenta 506 linhas e 13 variáveis (colunas). Segue a sumarização dos dados por variável, extraída do software R.

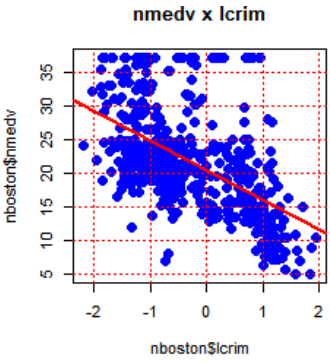
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
lcrim	-2,19930	-1,08590	-0,59090	0,33890	0,56550	1,94930
lzn	0,00000	0,00000	0,00000	0,41430	1,13030	2,00430
indus	0,46000	5,19000	9,69000	11,14000	18,10000	27,74000
chas	0,00000	0,00000	0,00000	0,06917	0,00000	1,00000
nox	0,38500	0,44900	0,53800	0,55470	0,62400	0,87100
nrm	5,00000	5,88600	6,20800	6,22600	6,62300	7,00000
age	2,90000	45,02000	77,50000	68,57000	94,08000	100,00000
ldis	0,52920	0,32226	0,50616	0,51596	0,71502	1,08374
rad	1,00000	4,00000	5,00000	9,54900	24,00000	24,00000
tax	187,00000	279,00000	330,00000	408,20000	666,00000	711,00000
sptrat	1,00000	1,67300	1,98700	2,07400	2,36600	3,22500
llstat	0,60000	1,93900	2,43000	2,37100	2,83100	3,63700
nmedv	5,00000	17,02000	21,20000	21,88000	25,00000	37,00000

2.2 Análise bivariada (resposta vs. variável previsora)

Nesta seção, considerando que a variável "nmedv" é a variável dependente (y), será verificada a sua relação com as demais variáveis por meio de modelo de regressão linear.

2.2.1 Verificar relação entre as variáveis

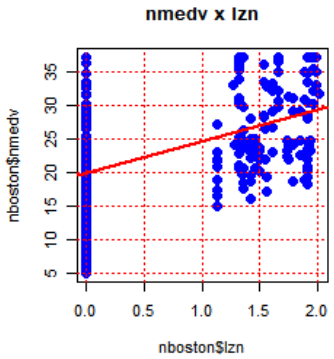
a) **Nmedv x lcrim:** considerando o p-value é possível afirmar que existe relação entre as variáveis. O valor de "Multiple R-squared" nos fornece a informação de que 29,6% da variabilidade no valor médio das casas (nmedv) é explicada pela taxa de criminalidade (lcrim).




```
Residuals:
    Min      1Q  Median      3Q     Max
-16.634  -4.515  -1.819   3.273  20.872
Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)  20.3849     0.3020   67.49 <0.0000000000000002 ***
lcrim        -4.4101     0.3028  -14.56 <0.0000000000000002 ***
Residual standard error: 6.39 on 504 degrees of freedom
Multiple R-squared:  0.2962,    Adjusted R-squared:  0.2948
F-statistic: 212.1 on 1 and 504 DF,  p-value: < 0.00000000000000022
```

Nmedv = 20,3849 -4,4101 lcrim

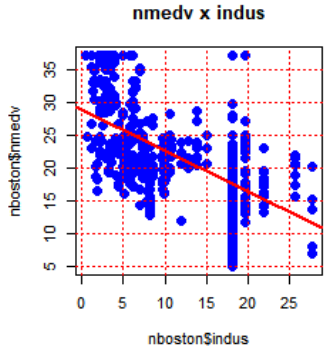
b) **Nmedv x lzn:** considerando o p-value é possível afirmar que existe relação entre as variáveis. O valor de "Multiple R-squared" nos fornece a informação de que 18,6% da variabilidade no valor médio das casas (nmedv) é explicada pela proporção de terrenos residenciais divididos em lotes com mais de 25.000 pés quadrados (lzn).




```
Residuals:
     Min       1Q   Median       3Q      Max
-14.9402  -5.0079  -0.4077   3.6744  17.0598

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)  19.9402     0.3542   56.29 <0.0000000000000002 ***
lzn           4.6818     0.4340   10.79 <0.0000000000000002 ***

Residual standard error: 6.865 on 504 degrees of freedom
Multiple R-squared:  0.1876,    Adjusted R-squared:  0.186
F-statistic: 116.4 on 1 and 504 DF,  p-value: < 0.00000000000000022
```

Nmedv = 19,9402 + 4,6818 lzn

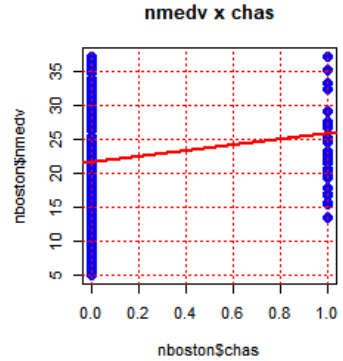
c) **Nmedv x indus:** considerando o p-value é possível afirmar que existe relação entre as variáveis. O valor de "Multiple R-squared" nos fornece a informação de que 31,5% da variabilidade no valor médio das casas (nmedv) é explicada pela proporção de hectares de negócios não comerciais por cidade (indus).




```
Residuals:
    Min      1Q  Median      3Q     Max
-12.547  -4.247  -0.855   3.811  20.373

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)  28.8085     0.5348   53.86 <0.0000000000000002 ***
indus        -0.6222     0.0409  -15.21 <0.0000000000000002 ***

Residual standard error: 6.306 on 504 degrees of freedom
Multiple R-squared:  0.3146,    Adjusted R-squared:  0.3133
F-statistic: 231.4 on 1 and 504 DF,  p-value: < 0.00000000000000022
```

Nmedv = 28,8085 – 0,6222 indus

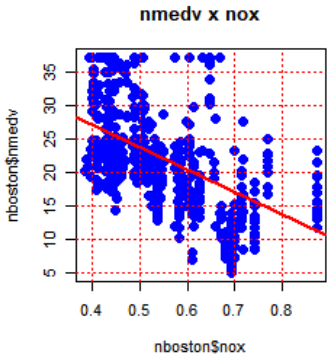
d) **Nmedv x chas:** considerando o p-value é possível afirmar que existe relação entre as variáveis. O valor de "Multiple R-squared" nos fornece a informação de que o trecho ser limitado pelo Rio Charles não é relevante na variabilidade do valor médio das casas (nmedv), pois a variável responde por 2% dessa variabilidade.




```
Residuals:
     Min       1Q   Median       3Q      Max
-16.5769  -4.9769  -0.7769   3.3731  15.4231

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)  21.5769     0.3472  62.147 < 0.0000000000000002 ***
chas          4.3774     1.3201   3.316             0.000979 ***

Residual standard error: 7.535 on 504 degrees of freedom
Multiple R-squared:  0.02135,   Adjusted R-squared:  0.01941
F-statistic:    11 on 1 and 504 DF,  p-value: 0.0009791
```

Nmedv = 21,5769 + 4,3774 chas

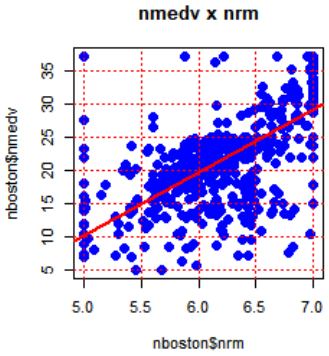
e) **Nmedv x nox:** considerando o p-value é possível afirmar que existe relação entre as variáveis. O valor de "Multiple R-squared" nos fornece a informação de que 25,6% da variabilidade no valor médio das casas (nmedv) é explicada pela concentração de óxidos de nitrogênio (nox).




```
Residuals:
    Min      1Q  Median      3Q     Max
-13.074  -4.435  -1.470   3.603  18.887

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)   40.319      1.429   28.21 <0.0000000000000002 ***
nox          -33.242      2.522  -13.18 <0.0000000000000002 ***

Residual standard error: 6.569 on 504 degrees of freedom
Multiple R-squared:  0.2563,    Adjusted R-squared:  0.2548
F-statistic: 173.7 on 1 and 504 DF,  p-value: < 0.00000000000000022
```

Nmedv = 40,319 – 22,242 nox

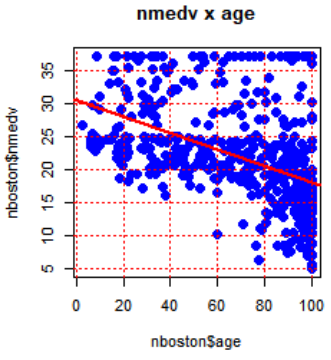
f) **Nmedv x nrm:** considerando o p-value é possível afirmar que existe relação entre as variáveis. O valor de "Multiple R-squared" nos fornece a informação de que o número de cômodos por moradia é relevante pois 42,6% da variabilidade no valor médio das casas (nmedv) é explicada por essa variável (nrm).




```
Residuals:
     Min       1Q   Median       3Q      Max
-19.6828  -2.4149   0.3451   3.5293  26.8017

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept) -37.4574     3.0778  -12.17 <0.0000000000000002 ***
nrm           9.5312     0.4927   19.35 <0.0000000000000002 ***

Residual standard error: 5.77 on 504 degrees of freedom
Multiple R-squared:  0.4262,    Adjusted R-squared:  0.425
F-statistic: 374.3 on 1 and 504 DF,  p-value: < 0.00000000000000022
```

Nmedv = –37,4574 + 9,5312 nrm

g) **Nmedv x age:** considerando o p-value é possível afirmar que existe relação entre as variáveis. O valor de "Multiple R-squared" nos fornece a informação de que 21% da variabilidade no valor médio das casas (nmedv) é explicada pela proporção etária de unidades construídas antes de 1940, ocupadas pelos proprietários (age).

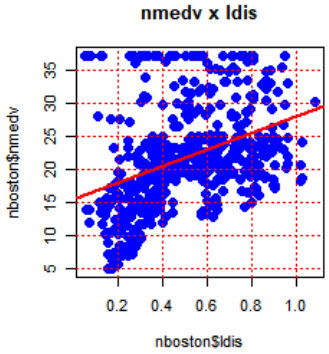



```
Residuals:
    Min      1Q  Median      3Q     Max
-14.437  -4.501  -1.298   3.039  19.014

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)  30.3763     0.7932   38.30 <0.0000000000000002 ***
age          -0.1239     0.0107  -11.58 <0.0000000000000002 ***

Residual standard error: 6.769 on 504 degrees of freedom
Multiple R-squared:  0.2101,    Adjusted R-squared:  0.2085
F-statistic: 134.1 on 1 and 504 DF,  p-value: < 0.00000000000000022
```

Nmedv = 30,3763 – 0,1239 age

h) **Nmedv x ldis:** considerando o p-value é possível afirmar que existe relação entre as variáveis. O valor de "Multiple R-

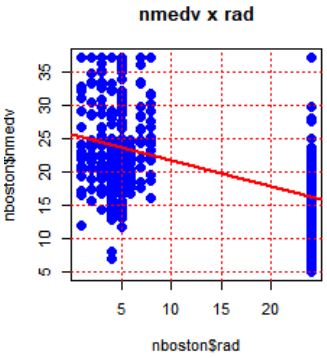
squared" nos fornece a informação de que a distância para os centros de emprego em Boston não é relevante na variabilidade do valor médio das casas (nmedv), pois a variável responde por 1,5% dessa variabilidade.




```
Residuals:
    Min      1Q  Median      3Q     Max
-12.602  -4.843  -1.519   2.952  20.896

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)  15.4433     0.7566  20.410 <0.0000000000000002 ***
ldis         12.4745     1.3355   9.341 <0.0000000000000002 ***
Residual standard error: 7.032 on 504 degrees of freedom
Multiple R-squared:  0.1476,    Adjusted R-squared:  0.1459
F-statistic: 87.25 on 1 and 504 DF,  p-value: < 0.00000000000000022
```

Nmedv = 15,4433 + 12,4745 ldis

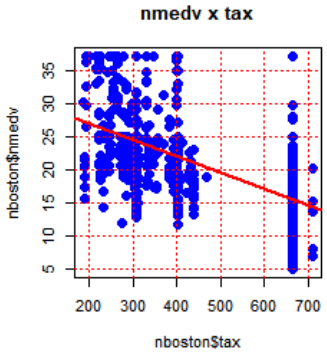
i) **Nmedv x rad:** considerando o p-value é possível afirmar que existe relação entre as variáveis. O valor de "Multiple R-squared" nos fornece a informação de que 20,5% da variabilidade no valor médio das casas (nmedv) é explicada pelo índice de acessibilidade às rodovias radiais (rad).




```
Residuals:
    Min      1Q  Median      3Q     Max
-17.074  -4.544  -1.266   3.932  20.834

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) 25.65547    0.44841   57.22 <0.0000000000000002 ***
rad         -0.39540    0.03471  -11.39 <0.0000000000000002 ***
Residual standard error: 6.792 on 504 degrees of freedom
Multiple R-squared:  0.2047,    Adjusted R-squared:  0.2031
F-statistic: 129.7 on 1 and 504 DF,  p-value: < 0.00000000000000022
```

Nmedv = 25,65547 -0,3954 rad

j) **Nmedv x tax:** considerando o p-value é possível afirmar que existe relação entre as variáveis. O valor de "Multiple R-squared" nos fornece a informação de que 29,5% da variabilidade no valor médio das casas (nmedv) é explicada pelo valor total da taxa de imposto sobre a propriedade (tax).

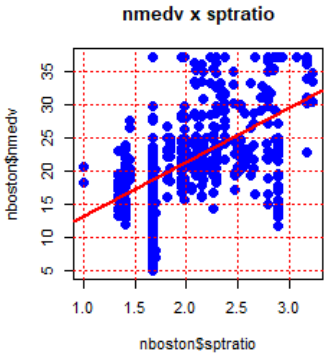



```
Residuals:
    Min      1Q  Median      3Q     Max
-13.297  -4.494  -1.432   3.674  21.443

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) 31.893580   0.745561   42.78 <0.0000000000000002 ***
tax         -0.024530   0.001688  -14.53 <0.0000000000000002 ***
Residual standard error: 6.394 on 504 degrees of freedom
Multiple R-squared:  0.2952,    Adjusted R-squared:  0.2938
F-statistic: 211.1 on 1 and 504 DF,  p-value: < 0.00000000000000022
```

Nmedv = 31,89358 - 0,024530 tax

k) **Nmedv x sptratio:** considerando o p-value é possível afirmar que existe relação entre as variáveis. O valor de "Multiple R-squared" nos fornece a informação de que 28,9% da

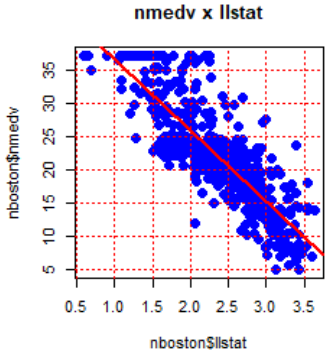
variabilidade no valor médio das casas (nmedv) é explicada pela proporção de alunos-professor por cidade (sptratio).




```
Residuals:
     Min       1Q   Median       3Q      Max
-16.7482  -4.3440  -0.1691   3.7294  18.4356

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)   4.7309     1.2411   3.812             0.000155 ***
sptratio      8.2671     0.5822  14.200 < 0.0000000000000002 ***

Residual standard error: 6.437 on 504 degrees of freedom
Multiple R-squared:  0.2858,    Adjusted R-squared:  0.2843
F-statistic: 201.6 on 1 and 504 DF,  p-value: < 0.00000000000000022
```

Nmedv = 4,7309 + 8,2671 sptratio

l) **Nmedv x llstat:** considerando o p-value é possível afirmar que existe relação entre as variáveis. O valor de "Multiple R-squared" nos fornece o status da população é muito relevante pois 70,8% da variabilidade no valor médio das casas (nmedv) é explicada por essa variável (nrm).

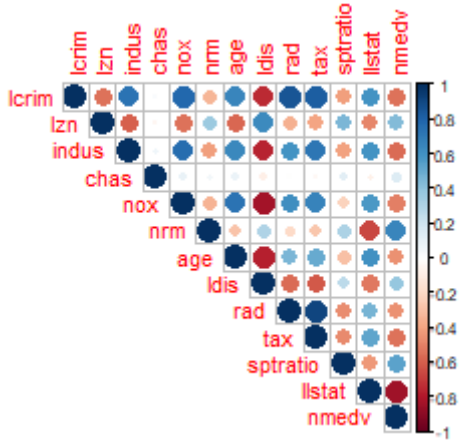



```
Residuals:
     Min       1Q   Median       3Q      Max
-13.2497  -2.8792  -0.5596   2.2083  13.8771

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  47.1567     0.7458   63.23 <0.0000000000000002 ***
llstat      -10.6606     0.3049  -34.96 <0.0000000000000002 ***

Residual standard error: 4.116 on 504 degrees of freedom
Multiple R-squared:  0.708,     Adjusted R-squared:  0.7074
F-statistic:  1222 on 1 and 504 DF,  p-value: < 0.00000000000000022
```

Nmedv = 47,1567 -10,6606 llstat

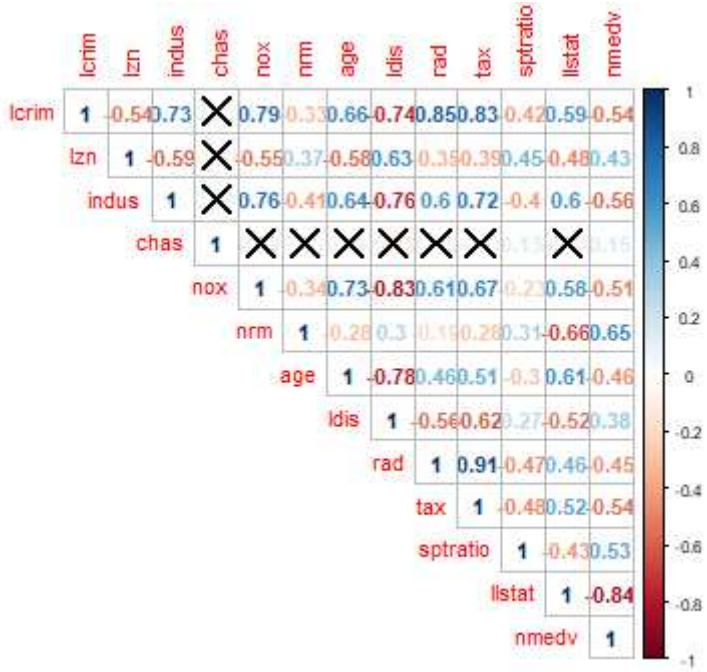
2.3 Correlações entre as variáveis previsoras

Para visualização das correlações foi utilizada a biblioteca "corrplot" do R. Para interpretação da matriz abaixo, considere que quanto maior o círculo maior a correlação entre as variáveis. Além disso, quanto mais azul escuro, mais próxima a correlação fica de 1, que significa que além de forte a correlação é positiva. Equivalentemente quanto mais próximo de vermelho escuro, mais próxima a correlação fica de -1, que significa que além de forte a correlação é negativa.



2.3.1 Identificar possíveis colinearidades

A matriz abaixo nos fornece os valores de "r" e "P" para a correlação das variáveis. Observa-se que as variáveis tax, rad e lcrim possuem correlação forte. Possivelmente serão reavaliadas quanto a sua permanência no modelo quando da execução do modelo de regressão múltipla.



3. Modelo de Regressão com todas as variáveis

Antes de rodar o modelo de regressão linear com todas as variáveis do data frame nboston, abrangendo as variáveis que sofreram transformação, optou-se por dividir esse *data frame* para viabilizar o treino e a validação do modelo.

Dessa forma foram criados os seguintes *dataset*:

a) **treino.nboston:** 70% dos dados de nboston, resultando em 361 observações e 13 variáveis, com a seguinte sumarização:

```
Call:
lm(formula = nmedv ~ ., data = treino.nboston)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5452 -2.1254 -0.1715  2.1293 13.3715

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)  34.632620   4.491072   7.711    0.000000000000132
lcrim        -0.923744   0.535443  -1.725              0.085380
lzn           0.067702   0.402741   0.168              0.866600
indus        -0.068812   0.054328  -1.267              0.206142
chas          1.692760   0.734644   2.304              0.021800
nox         -13.421890   3.421489  -3.923              0.000105
nrm           2.089801   0.479539   4.358    0.000017304820821
age           0.021486   0.011777   1.824              0.068947
ldis         -9.262113   1.729201  -5.356    0.000000154741474
```

```
rad           0.177656   0.064318   2.762             0.006047
tax          -0.009699   0.003085  -3.143             0.001813
sptratio      2.983850   0.521614   5.720    0.000000022916756
llstat       -7.853462   0.532947 -14.736 < 0.0000000000000002

(Intercept) ***
lcrim       .
lzn
indus
chas        *
nox         ***
nrm         ***
age         .
ldis        ***
rad         **
tax         **
sptratio    ***
llstat      ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.429 on 348 degrees of freedom
Multiple R-squared:  0.8058,       Adjusted R-squared:  0.7991
F-statistic: 120.3 on 12 and 348 DF,  p-value: < 0.0000000000000002
2
```

b) **teste.nboston**: 30% dos dados de nboston, resultando em 145 observações e 13 variáveis, com a seguinte sumarização:

```
Call:
lm(formula = nmedv ~ ., data = teste.nboston)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5067  -2.2414  -0.0247   2.0010   8.7957

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
(Intercept)  43.506353   8.003930   5.436 0.00000025503993322
lcrim        -0.090550   0.871246  -0.104            0.917381
lzn          -0.341695   0.613567  -0.557            0.578539
indus         0.045639   0.085578   0.533            0.594721
chas          1.771800   1.272824   1.392            0.166257
nox         -14.227821   5.814866  -2.447            0.015728
nrm           1.735634   0.919785   1.887            0.061356
age          -0.019938   0.020103  -0.992            0.323120
ldis        -10.437639   2.896662  -3.603            0.000444
rad           0.236170   0.110987   2.128            0.035204
tax          -0.017178   0.005566  -3.086            0.002469
sptratio      2.554348   0.753760   3.389            0.000926
llstat       -7.974849   0.911503  -8.749 0.00000000000000878

(Intercept) ***
lcrim
lzn
indus
chas
nox         *
nrm         .
age
ldis        ***
rad         *
tax         **
sptratio    ***
llstat      ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 3.545 on 132 degrees of freedom
Multiple R-squared:  0.7965,       Adjusted R-squared:  0.778
F-statistic: 43.04 on 12 and 132 DF,  p-value: < 0.0000000000000002
2
```

As análises a seguir serão feitas apenas para o data frame de treino (**treino.nboston**).

A função de regressão linear do treino é dada por:

Nmedv = 34,632620 – 0,923744 lcrim + 0,067702 lzn – 0,068812 indus + 1,692760 chas – 13,421890 nox + 2,089801 nrm + 0,021486 age – 9,262113 ldis + 0,177656 rad – 0,009699 tax + 2,983850 sptratio – 7,853462 llstat

3.1 Analisar sinais dos coeficientes

A análise dos sinais dos coeficientes evidencia que o valor o valor médio das casas ocupadas pelos proprietários no subúrbio de Boston (nmedv) tende a aumentar:

a) Quanto maior for:lzn, chas, nrm, age, rad e sptratio;
b) Quanto menor for: lcrim, indus, nox, ldis, tax e llstat.

Ou seja, taxa de criminalidade, proporção de hectares de negócios não comerciais pela cidade, concentração de óxido de nitrogênio, distância para os centros de emprego em Boston, taxa de imposto e status da população impactam negativamente o valor médio das casas, sendo o maior impacto pelo óxido de nitrogênio.

Já a proporção de terrenos residenciais divididos em lotes com mais de 25.000 pés quadrados, limitação pelo Rio Charles, número médio de cômodos por moradia, proporção etária de unidades construídas antes de 1940, ocupadas pelos proprietários, índice de acessibilidade às rodovias radiais e proporção de alunos-professor por cidade impactam positivamente o valor médio das casas.

3.2 Avaliar R2

O coeficiente de determinação (R2) indica que 80,58% da variabilidade do valor médio das casas ocupadas pelos proprietários no subúrbio de Boston (nmedv) pode ser explicada pelo modelo de regressão.

3.3 Teste de hipótese

O teste de hipótese a seguir tem como objetivo verificar se as variáveis previsoras são significantes para o modelo.

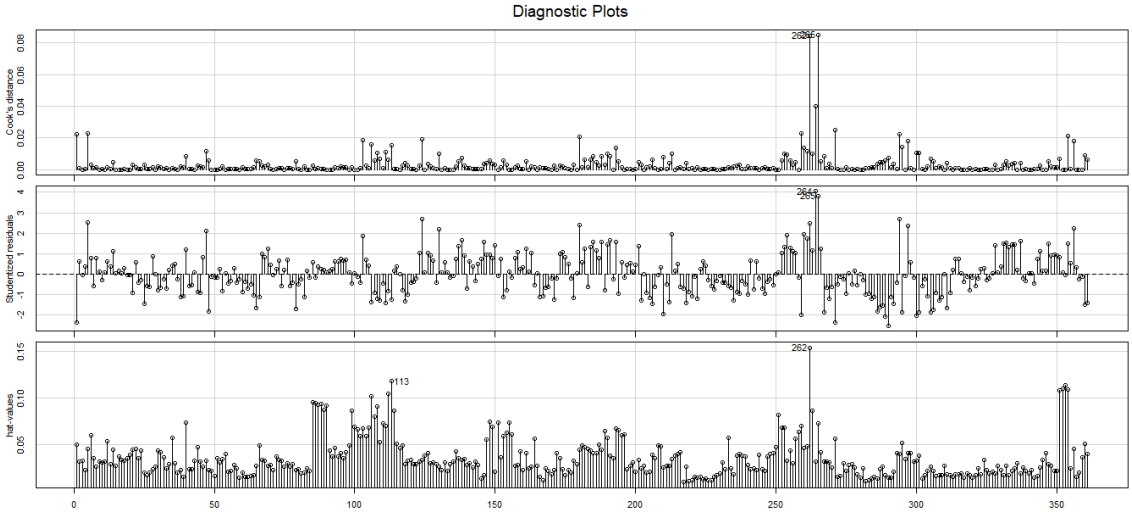
Ho: $\beta 1 = \beta 2 = \ldots = \beta p = 0$, ou seja, nenhum dos βj são significantes para o modelo;

Ha: βj ≠ 0 para pelo menos um j ≥ 1, ou seja, βj é significante para o modelo.

Observe que o p-value apresenta um valor baixo (p-value: < 0,00000000000000022), o que indica que o modelo tem bom desempenho para mensurar o valor médio das casas ocupadas pelos proprietários no subúrbio de Boston (nmedv). Ou seja, a regressão é significante (assumindo válidas as suposições estatísticas sobre o erro).

3.4 Diagnóstico (resíduos (TRES), alavancagem (hat), influentes (Cook))

Observe abaixo o diagnóstico do modelo de regressão:



O teste "*D de Cook*" mostra os pontos influentes para o modelo, cuja omissão ou inclusão podem afetar significativamente os resultados da regressão.

Por sua vez, o teste "*Studentized residuals*" serve para indicar os *outliers* em "y", ou seja, na variável *target* (dependente).

Já o teste "*hat values*" indica os *outliers* nas variáveis preditoras (independentes).

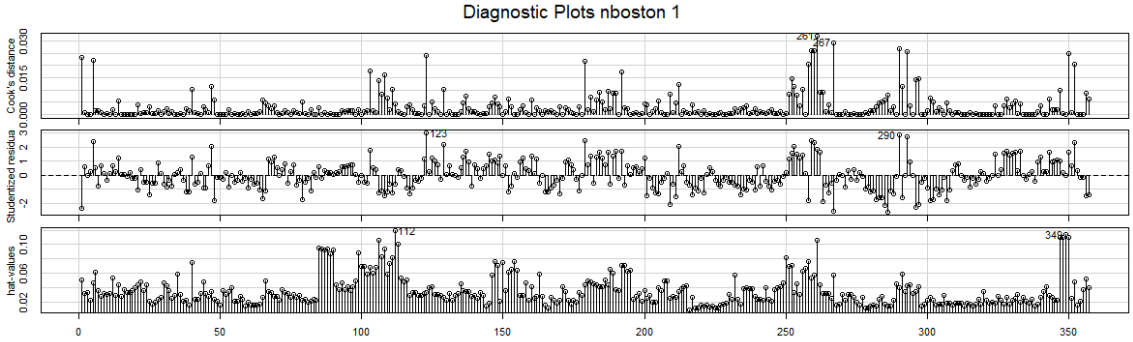
O gráfico indica que:

a) Os pontos 262 e 265 são influentes para o modelo;
b) Os pontos 264 e 265 são *outliers* em "y"; e
c) Os pontos 113 e 262 são *outliers* em "x".

Como se trata de apenas 4 pontos, optou-se pela exclusão dos mesmos, sendo criado novo data frame de nome treino_nboston1.

A sumarização do modelo de regressão de nboston1 apresentou R2 de 82,02%, ou seja, a exclusão das linhas acima melhorou em 1,44% a explicação da variabilidade de nmedv pelo modelo de regressão.

O diagnóstico no modelo de regressão de nboston1 ficou assim:



3.5 Verificar existência de multicolinearidade em nboston2

Para cálculo da multicolinearidade utilizou-se a função VIF da biblioteca RMS, a qual apresentou como saídas:

```
   lcrim      lzn    indus     chas      nox      nrm      age
     7.7      2.6      4.3      1.1      4.9      2.1      3.5
    ldis      rad      tax sptratio   llstat
     5.2      9.8      8.4      2.0      3.7
```

VIF superiores a 10 indicam a existência de multicolinearidade. Dessa forma, não há evidências de multicolinearidade na regressão.

4. Selecionar variáveis

Esta seção se destina aos testes para seleção das variáveis que irão compor o modelo de regressão final.

4.1 Definir método(s) de seleção

Serão utilizados os seguintes métodos de seleção/exclusão de variáveis:

a) *Information Criterion de Akaike* (AIC);e
b) VIF – exclusão da variável de maior VIF

4.2 Modelos

4.2.1 Modelo AIC

Utilizando a função *step*, foram excluídas da seleção as variáveis **lzn** e **indus**, apresentando a seguinte sumarização:

```
Call:
lm(formula = nmedv ~ lcrim + chas + nox + nrm + age + ldis +
    rad + tax + sptratio + llstat, data = treino.nboston1)

Residuals:
   Min     1Q Median     3Q    Max
-8.305 -2.024 -0.148  1.991  9.756

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
```

```
(Intercept)  26.497643   4.384081   6.044    0.000000003889456 ***
lcrim        -0.771651   0.501339  -1.539             0.124675
chas          1.271427   0.709442   1.792             0.073982 .
nox         -12.132929   3.195129  -3.797             0.000173 ***
nrm           2.764910   0.475310   5.817    0.000000013640818 ***
age           0.018144   0.011338   1.600             0.110445
ldis         -6.885689   1.538521  -4.476    0.000010359275965 ***
rad           0.157218   0.058994   2.665             0.008061 **
tax          -0.011120   0.002635  -4.220    0.000031284939496 ***
sptratio      3.254420   0.437456   7.439    0.000000000000806 ***
llstat       -7.140258   0.551267 -12.952 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.243 on 346 degrees of freedom
Multiple R-squared:  0.8195,   Adjusted R-squared:  0.8143
F-statistic: 157.1 on 10 and 346 DF,  p-value: < 0.00000000000000022
```

A função de regressão do modelo é dada por:

Nmedv = 26,497643 - 0,771651 lcrim + 1,271427 chas - 12,132929 nox + 2,764910 nrm + 0,018144 age - 6,885689 ldis + 0,157218 rad - 0,011120 tax + 3,254420 sptratio -7,140258 llstat

a) Análise dos sinais dos coeficientes:

A análise dos sinais dos coeficientes evidencia que o valor o valor médio das casas ocupadas pelos proprietários no subúrbio de Boston (nmedv) tende a aumentar:

i) Quanto maior for: chas, nrm, age, rad e sptratio;
ii) Quanto menor for: lcrim, nox, ldis, tax e llstat.

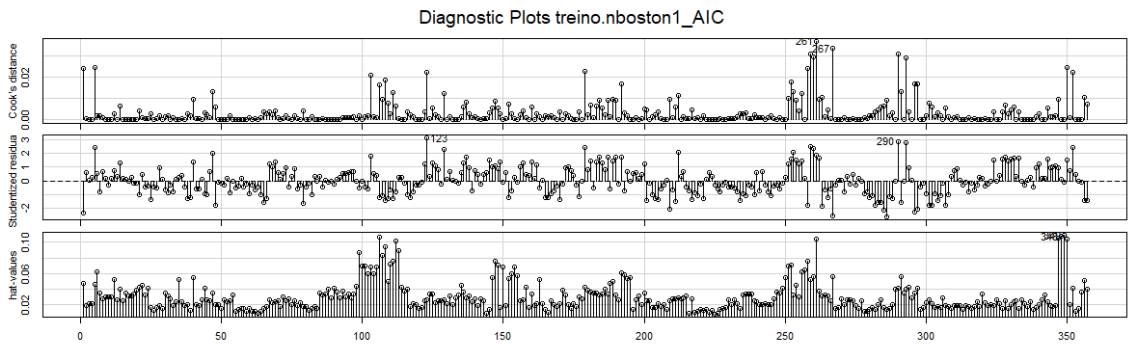
b) Avaliar R2:

O coeficiente de determinação (R2) indica que 81,95% da variabilidade do valor médio das casas ocupadas pelos proprietários no subúrbio de Boston (nmedv) pode ser explicada pelo modelo de regressão.

c) Análise do p-value:

Observe que o p-value apresenta um valor baixo (p-value: < 0,00000000000000022), o que indica que o modelo tem bom desempenho para mensurar o valor médio das casas ocupadas pelos proprietários no subúrbio de Boston (nmedv). Ou seja, a regressão é significante (assumindo válidas as suposições estatísticas sobre o erro).

d) Diagnóstico (resíduos (TRES), alavancagem (hat), influentes (Cook)):



Não foi possível identificar pontos que justifiquem exclusão, tendo em vista a similaridade dos mesmos.

e) **Multicolinearidade:** não há evidências de multicolinearidade no modelo AIC.

```
lcrim     chas      nox      nrm      age     ldis      rad
  7.4      1.1      4.7      2.1      3.5      4.5      9.0
  tax sptratio   llstat
  6.8      1.6      3.6
```

4.2.2 Modelo VIF – exclusão da variável de maior VIF

Excluindo a variável rad, que possui maior VIF, o modelo apresenta a seguinte sumarização:

```
Call:
lm(formula = nmedv ~ ., data = treino.nboston1_vif)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4841 -2.2754 -0.2243  1.9620  9.8800

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)  25.921380   4.563591   5.680        0.00000002863 ***
lcrim        -0.089922   0.422860  -0.213             0.831724
lzn          -0.071593   0.384105  -0.186             0.852250
indus        -0.086923   0.050446  -1.723             0.085772 .
chas          1.429847   0.720288   1.985             0.047924 *
nox         -11.578717   3.288000  -3.522             0.000487 ***
nrm           2.938632   0.476306   6.170        0.00000000192 ***
age           0.014655   0.011384   1.287             0.198808
ldis         -7.167501   1.676973  -4.274        0.00002486553 ***
tax          -0.005044   0.002118  -2.382             0.017778 *
sptratio      3.019628   0.491901   6.139        0.00000000229 ***
llstat       -7.011806   0.558074 -12.564 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.267 on 345 degrees of freedom
Multiple R-squared:  0.8174,  Adjusted R-squared:  0.8116
F-statistic: 140.4 on 11 and 345 DF,  p-value: < 0.00000000000000022
```

Nmedv = 25,921380 - 0,089922 lcrim - 0,071593 lzn - 0,086923 indus + 1,429847 chas - 11,578717 nox + 2,938632 nrm + 0,014655 age - 7,167501 ldis - 0,005044 tax + 3,019628 sptratio -7,011806 llstat

a) **Análise dos sinais dos coeficientes:**

A análise dos sinais dos coeficientes evidencia que o valor o valor médio das casas ocupadas pelos proprietários no subúrbio de Boston (nmedv) tende a aumentar:

i) Quanto maior for: chas, nrm, age e sptratio;
ii) Quanto menor for: lcrim, lzn, indus, nox, ldis, tax e llstat.

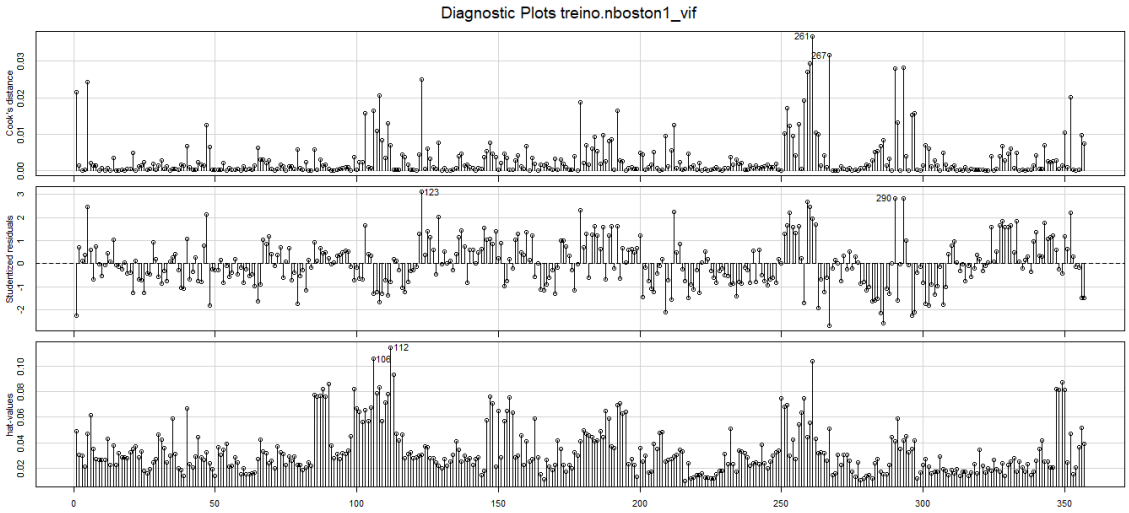
b) **Avaliar R2:**

O coeficiente de determinação (R2) indica que 81,74% da variabilidade do valor médio das casas ocupadas pelos proprietários no subúrbio de Boston (nmedv) pode ser explicada pelo modelo de regressão.

c) **Análise do p-value:**

Observe que o p-value apresenta um valor baixo (p-value: < 0,00000000000000022), o que indica que o modelo tem bom desempenho para mensurar o valor médio das casas ocupadas pelos proprietários no subúrbio de Boston (nmedv). Ou seja, a regressão é significante (assumindo válidas as suposições estatísticas sobre o erro).

d) **Diagnóstico (resíduos (TRES), alavancagem (hat), influentes (Cook)):**



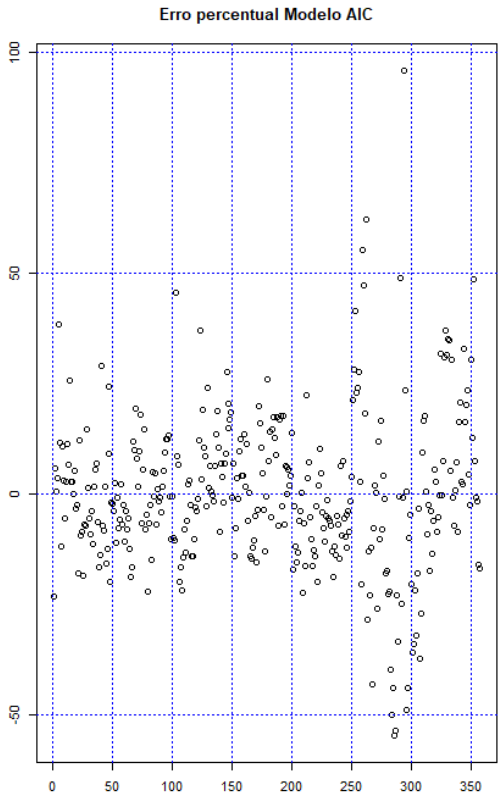
Não foi possível identificar pontos que justifiquem exclusão, tendo em vista a similaridade dos mesmos.

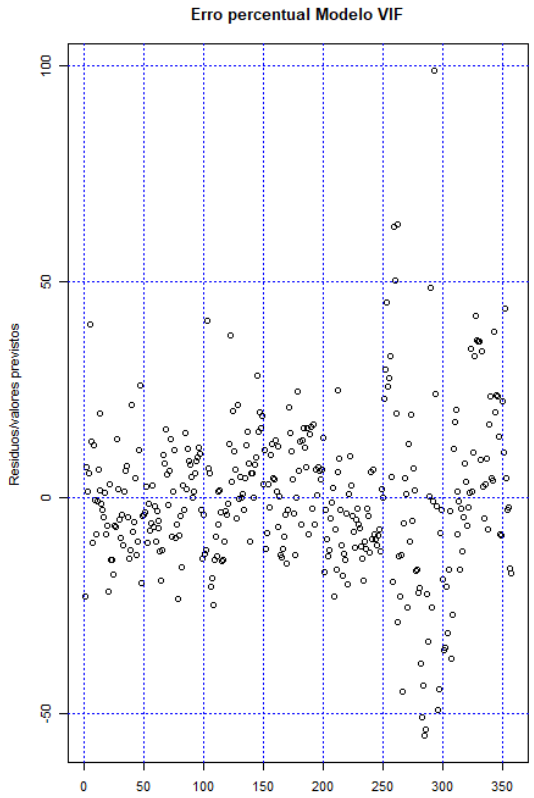
e) **Multicolinearidade:** não há evidências de multicolinearidade no modelo VIF.

lcrim	lzn	indus	chas	nox	nrm	age
5.2	2.6	4.0	1.1	4.9	2.1	3.5
ldis	tax	sptratio	llstat			
5.2	4.3	2.0	3.7			

5. Avaliar capacidade preditiva dos modelos

```
> summary (cp_modelo1)
   nmedv_hat           RES                EP          
 Min.   : 6.969   Min.   :-8.305   Min.   :-54.5998  
 1st Qu.:16.053   1st Qu.:-2.024   1st Qu.: -9.4626  
 Median :21.159   Median :-0.148   Median : -0.8183  
 Mean   :21.839   Mean   : 0.000   Mean   :  0.2214  
 3rd Qu.:26.807   3rd Qu.: 1.991   3rd Qu.:  9.0535  
 Max.   :39.399   Max.   : 9.756   Max.   : 95.6598  
> summary (cp_modelo2)
   nmedv_hat           RES                EP          
 Min.   : 7.137   Min.   :-8.4841   Min.   :-55.055  
 1st Qu.:16.058   1st Qu.:-2.2754   1st Qu.: -9.979  
 Median :21.089   Median :-0.2243   Median : -1.218  
 Mean   :21.839   Mean   : 0.0000   Mean   :  0.245  
 3rd Qu.:26.916   3rd Qu.: 1.9620   3rd Qu.:  8.736  
 Max.   :39.448   Max.   : 9.8800   Max.   : 98.913  
```





Mape do modelo 1 = 0.1341746

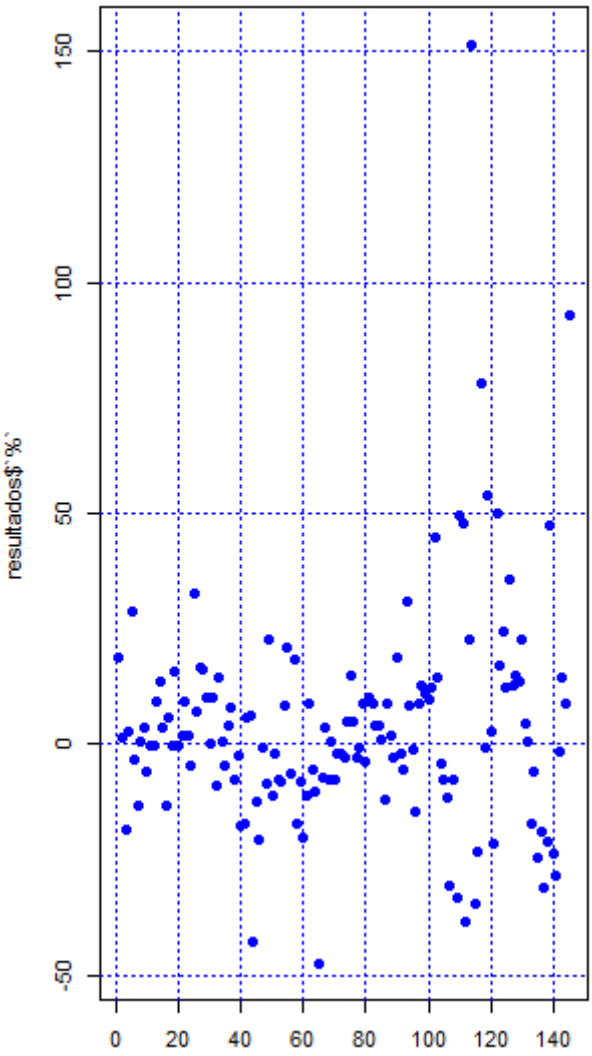
Mape do modelo 2 = 0.1358396

Considerando o coeficiente de determinação dos dois modelos de regressão e a capacidade preditiva dos mesmos, conclui-se que o modelo AIC é o mais adequado.

6. Validação do modelo AIC

```
> summary (resultados)
```

Previsto	Real	Diferença	%
Min. : 8.20	Min. : 5.00	Min. :-11.20000	Min. :-47.300
1st Qu.:16.20	1st Qu.:16.70	1st Qu.: -1.90000	1st Qu.: -7.700
Median :21.00	Median :20.90	Median : 0.10000	Median : 0.800
Mean :21.55	Mean :21.56	Mean : -0.01103	Mean : 3.704
3rd Qu.:25.50	3rd Qu.:24.60	3rd Qu.: 2.10000	3rd Qu.: 11.100
Max. :38.50	Max. :37.00	Max. : 11.00000	Max. :151.500


resultados$`%`
150
100
50
0
-50
0
20
40
60
80
100
120
140