Análise de Regressão Linear

Boston House Prices

Rafael Reis

08/08/2022

Universidade Federal de Juiz de Fora

Prof. Tiago M. Magalhães

1 Introdução

Utilizaremos para a seguinte análise o banco de dados referente aos preços médios de residências em Boston, Massachusetts, e principalmente como os preços e a demanda por casas podem ser afetadas pela poluição no ar.

Realizaremos diversas análises derivadas da análise de regressão linear para tentarmos encontrar o melhor modelo que possa nos justificar os preços encontrados para residências em Boston.

Nosso banco de dados provém de um total de 13 variáveis explicativas e uma variável resposta, todas descritas a seguir em ordem:

Variáveis Explicativas

1) CRIM: taxa de criminalidade per capita dos bairros
2) ZN: proporção de terrenos residenciais zoneados para lotes acima de 25.000 sq.ft.
3) INDUS: proporção de acres de negócios não varejistas por bairro
4) CHAS: variável dummy relacionada ao Rio Charles (1 em caso do rio passar próximo; 0 caso contrário)
5) NOX: concentração de óxidos nítricos (partes por 10 milhões) [parts/10M]
6) RM: número médio de quartos por habitação
7) AGE: proporção de unidades ocupadas pelos proprietários construídas antes de 1940
8) DIS: distâncias ponderadas para cinco centros de emprego de Boston
9) RAD: índice de acessibilidade às rodovias radiais
10) TAX: taxa de imposto de propriedade de valor total por US$ 10.000 [US$/10l]
11) PTRATIO: relação aluno-professor por cidade
12) B: O resultado da equação B=1000(Bk - 0,63)^2 onde Bk é a proporção de negros por cidade
13) LSTAT: % status mais baixo da população (proporção de adultos sem níveis altos de escolaridade)

Variável Resposta

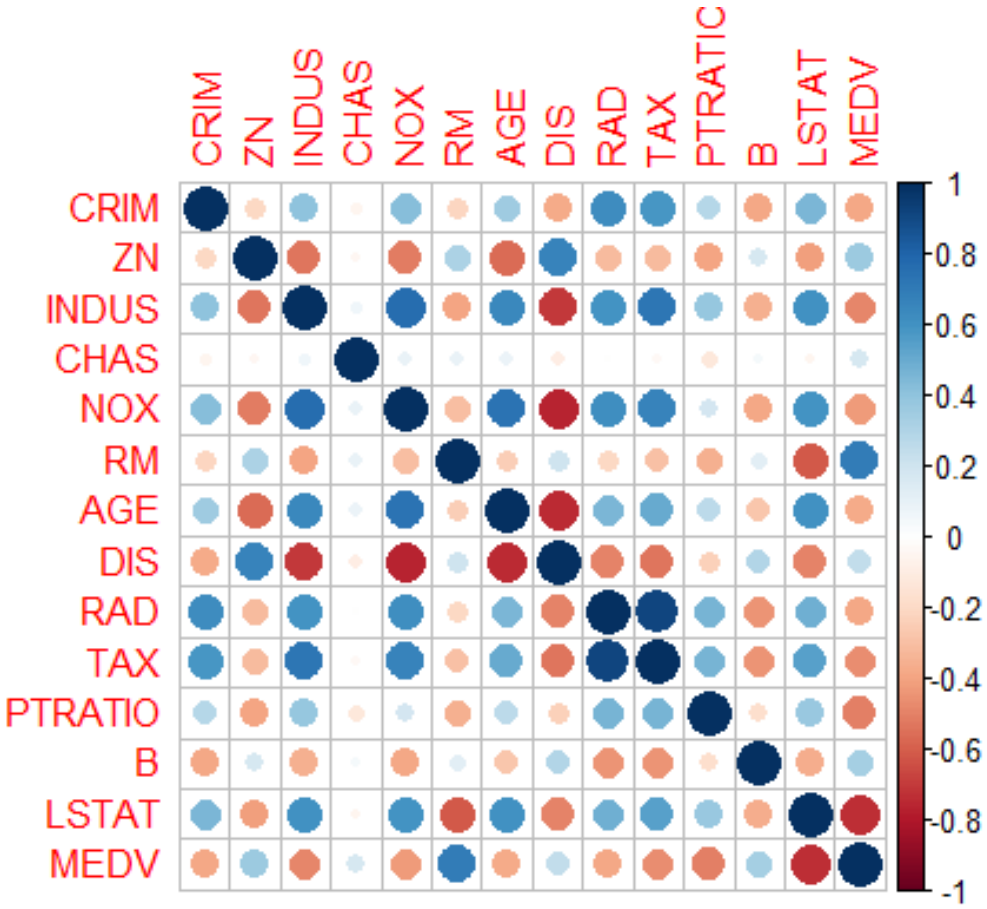
1) MEDV: Valor médio das casas ocupadas pelos proprietários em $ 1.000 [k$]

É notável a atenção da volatividade dos dados aqui estudados. O mercado residencial norte-americano vem enfrentando uma evidente crise nos últimos anos relativa ao número de casas disponíveis em relação a quantidade de pessoas com a renda suficiente para comprá-las. Este estudo não busca se aprofundar severamente neste assunto, mas é recomendada certa noção de que nossa variável resposta pode não ser tão fácil de ser analisada completamente enquanto ignoramos fatores socio-econômicos importantes.

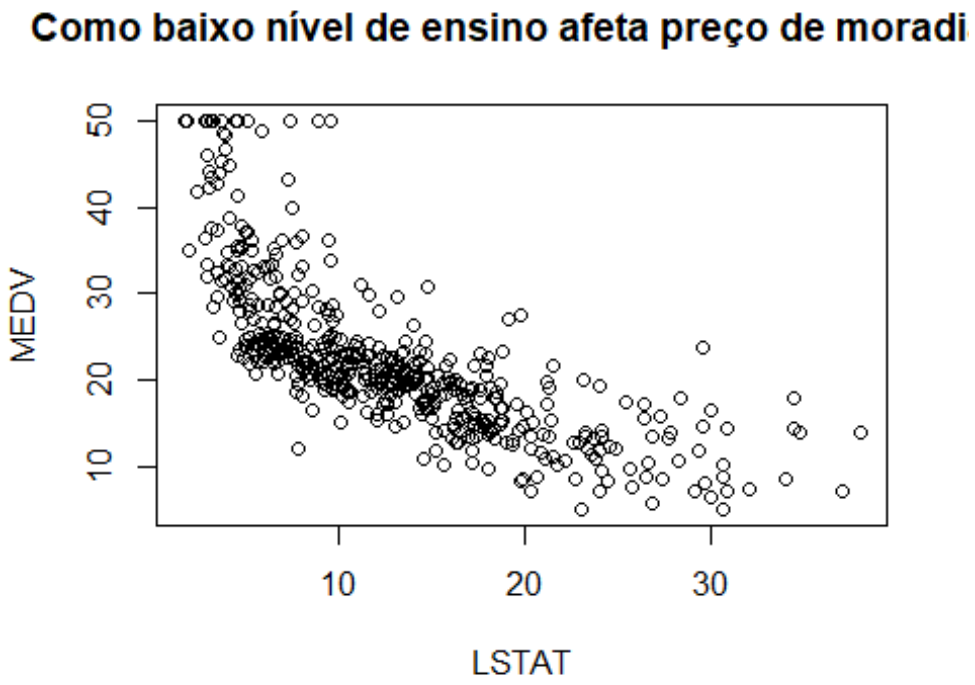
2 Analise Exploratoria

Aqui tentaremos encontrar relações e comparações entre as variáveis de nosso banco de dados de forma que possamos visualizar como elas interagem entre si.

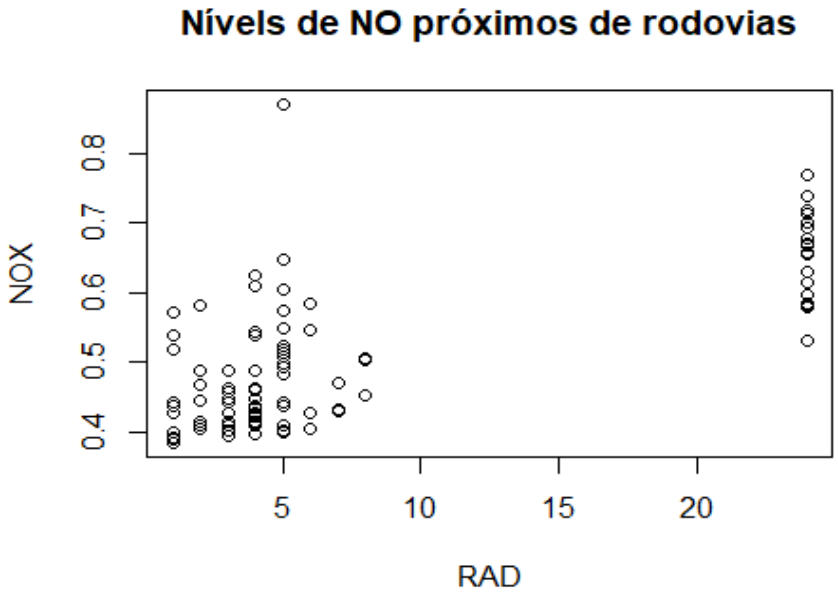
Gráfico de correlações



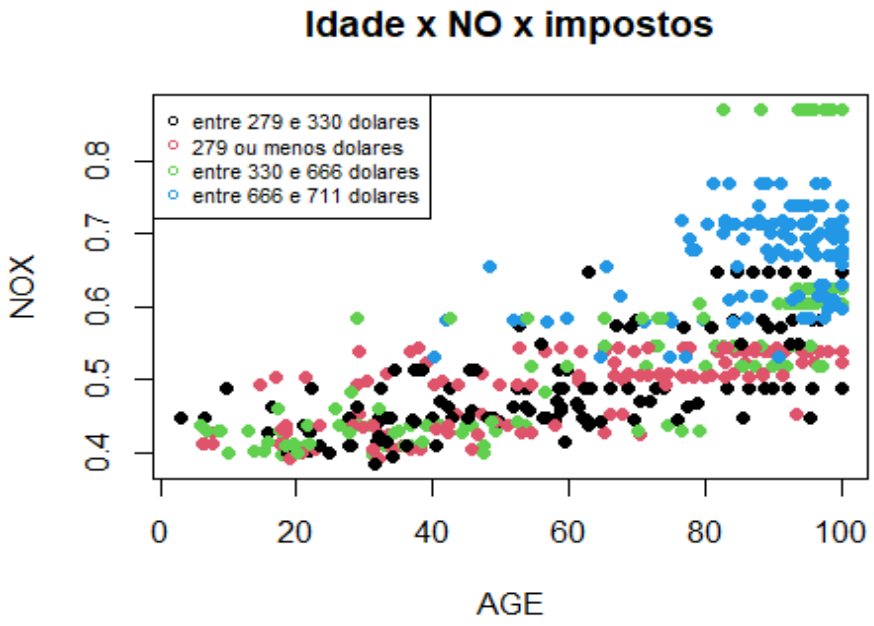
Aqui já podemos perceber algumas correlações que podemos considerar importantes para nossa análise, como por exemplo como a proporção de adultos sem ensino superior afeta negativamente o preço das residências da região, entre outros pontos interessantes.



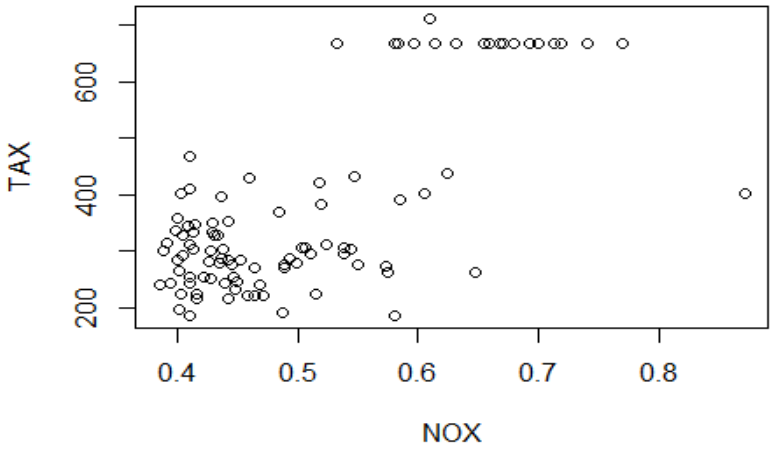
Percebemos aqui a correlação negativa vista no gráfico anterior.



Um pouco mais difícil de se visualizar devido à natureza da variável RAD, mas é possível perceber uma relação positiva entre a acessibilidade à rodovias e os níveis de óxidos nítricos no ar, estes derivados principalmente de automóveis como carros e caminhões.



Podemos analisar diversos fatores deste gráfico. Primeiro podemos perceber como os níveis de óxidos nítricos aumentam quando analisamos áreas com grandes proporções de residências antigas. Isso muito provavelmente se dá pelo fato de que estas áreas agora se instalaram como áreas residenciais onde um grande número de automóveis transitam diariamente, assim elevando os níveis de NO. Podemos também notar como as taxas de impostos para residências em áreas com altos níveis de NO são mais elevadas.



Aqui vemos em outro ângulo a reçação do nível de NO no ar e como isso afeta o valor dos impostos pagos na região.

3 Regressão Linear e Seleção de Variáveis

Agora veremos como podemos realizar uma regressão linear para explicarmos como encontramos de forma eficiente nossa variável resposta MEDV.

Como é de se imaginar, testar todas as possibilidades de modelos com nosso número considerável de variáveis seria impossível, então para facilitar nosso trabalho utilizamos o método Stepwise via AIC. Esse método avaliará os p-valores de cada variável em comparação a um $alpha$ específico (em nosso caso utilizamos nível de significância de 5%)

Os códigos para nossa análise podem ser encontrados no apêndice 3.1 ao final do relatório.

```
##
## Call:lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT, data = dado)
##
## 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
```

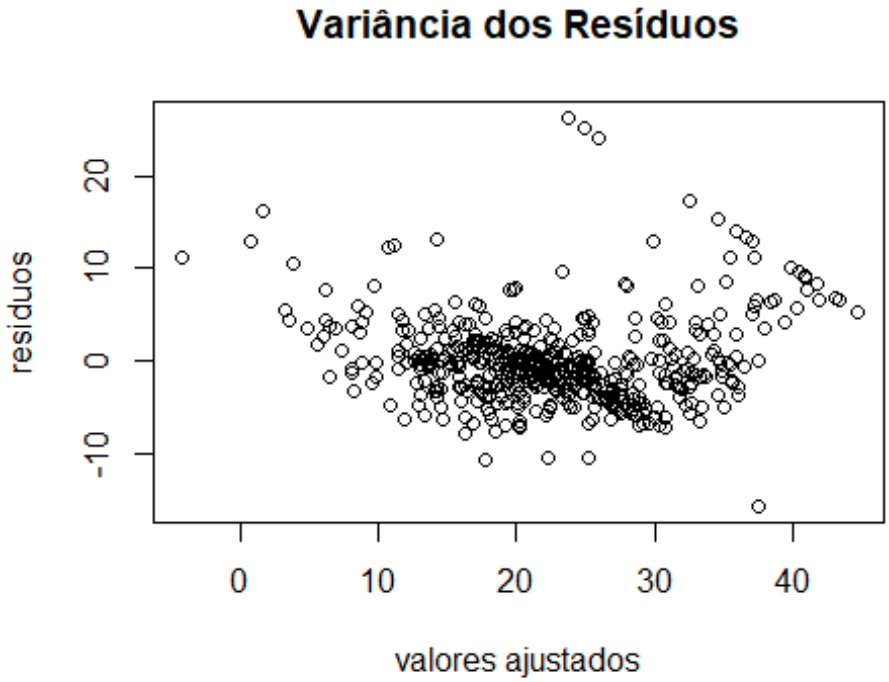
As variáveis selecionadas foram CRIM, ZN, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, B e LSTAT.

Por nossa análise nos conseguimos encontrar o seguinte modelo com essas variáveis:

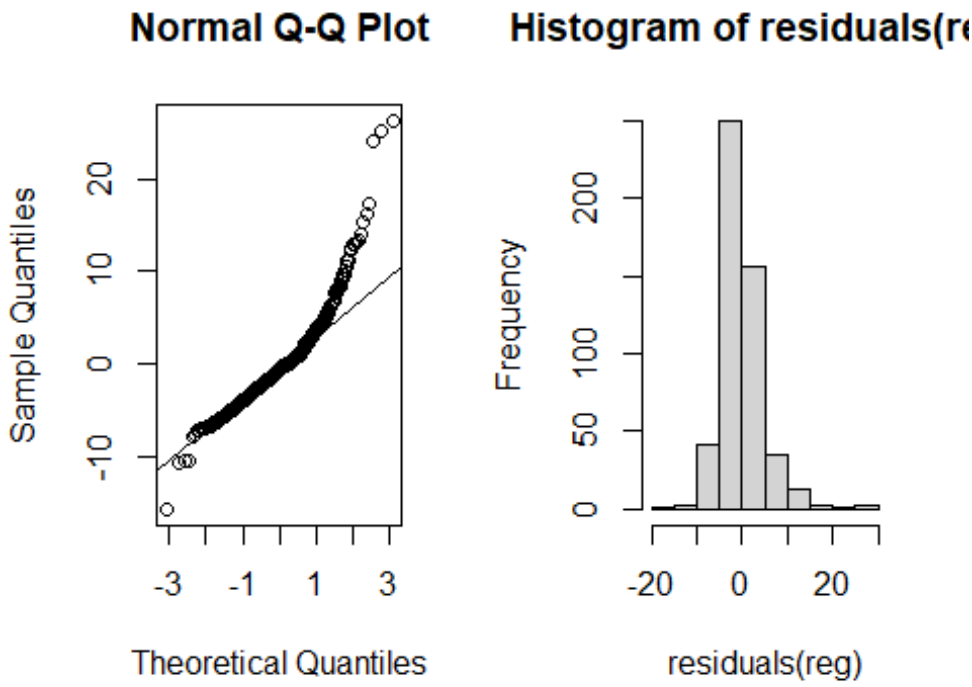
$$Y = 36.34 - 0.108x_2 + 0.045x_3 + 2.72x_4 - 17.38x_5 + 3.8x_6 - 1.49x_7 + 0.3x_8 - 0.12x_9 - 0.95x_{10} + 0.009x_{11} - 0.522x_{12}$$

4 Residuos

Para analisarmos se nosso modelo proposto é de fato o ideal para explicarmos os preços médios de residências em Boston, podemos analisar os resíduos que nosso modelo causa.



Para um bom modelo buscamos que a variância de nossos resíduos se situem entre 3 e -3. Como podemos ver em nosso gráfico, isso está longe de ser verdade.

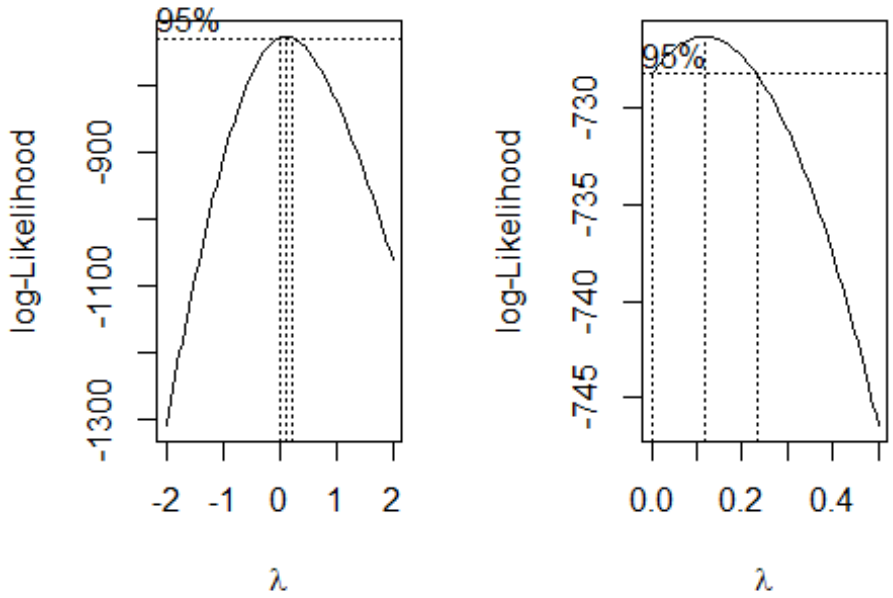


Analisamos então a normalidade de nosso modelo e também nos encontramos com resultados muito abaixo do esperado ou desejado.

Para buscar resultados mais satisfatórios para nosso modelo, devemos então tentar realizar uma transformação em nossa variável resposta.

5 Transformação

Realizamos nossa transformação via Procedimento de Box COX, e os códigos podem ser vistos no apêndice 5.1.



Note que nossos gráficos mostram os valores da log-verossimilhança para um intervalo de valores do parâmetro de transformação λ. O máximo da verossimilhança foi atingido com aproximadamente λ = 0.1162, com intervalo de confiança distante de 1. Com isso, há forte evidência da necessidade de transformação na variável resposta MEDV, dado por: MEDV* = (MEDV * 0.1162 − 1) / 0.1162, Sendo assim a nossa nova variável transformada deve ser inserida no banco de dados, para que o novo modelo de regressão linear simples seja ajustado.

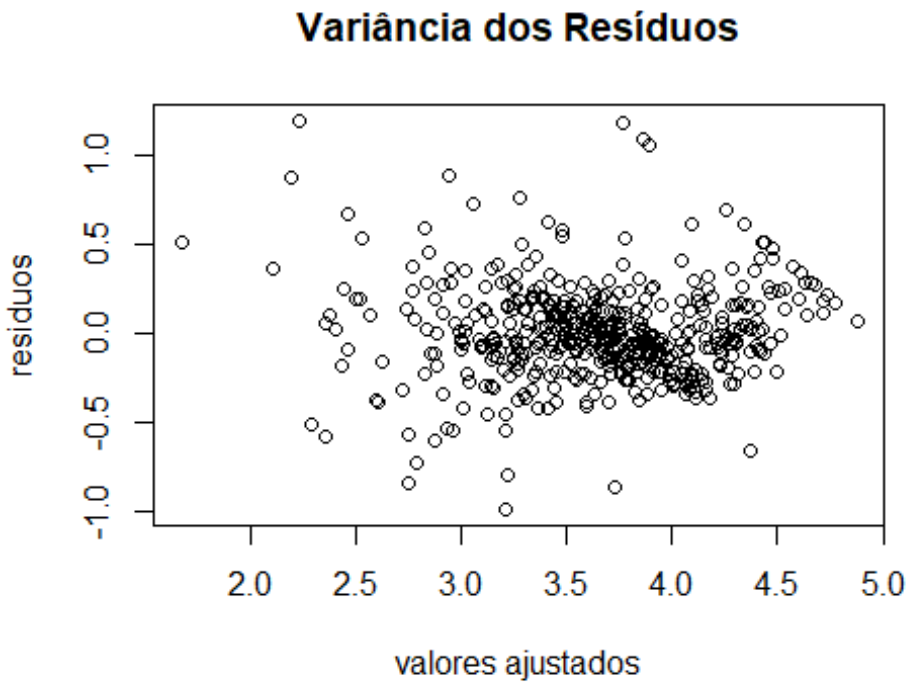
```
##
## Call:
## lm(formula = MEDVtrans ~ CRIM + ZN + CHAS + NOX + RM + DIS +
##     RAD + TAX + PTRATIO + B + LSTAT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98931 -0.14248 -0.03039  0.13603  1.20009
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.0767971  0.2877586  17.643  < 2e-16 ***
## CRIM        -0.0135454  0.0018614  -7.277 1.35e-12 ***
```

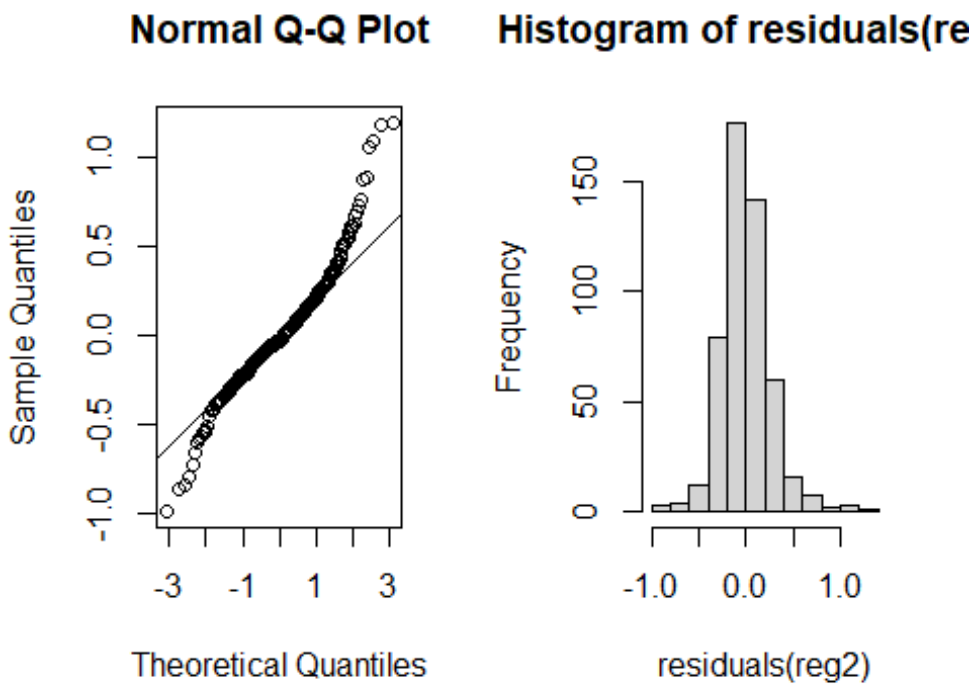
```
## ZN           0.0017170  0.0007679   2.236  0.02580 *
## CHAS         0.1517342  0.0485082   3.128  0.00186 **
## NOX         -1.0430584  0.2007495  -5.196 2.99e-07 ***
## RM           0.1423236  0.0230727   6.168 1.44e-09 ***
## DIS         -0.0761098  0.0105468  -7.216 2.02e-12 ***
## RAD          0.0191145  0.0036003   5.309 1.67e-07 ***
## TAX         -0.0007900  0.0001915  -4.126 4.34e-05 ***
## PTRATIO     -0.0542294  0.0073290  -7.399 5.93e-13 ***
## B            0.0005921  0.0001518   3.900  0.00011 ***
## LSTAT       -0.0397671  0.0026930 -14.767  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2689 on 494 degrees of freedom
## Multiple R-squared:  0.7884, Adjusted R-squared:  0.7837
## F-statistic: 167.3 on 11 and 494 DF,  p-value: < 2.2e-16
```

Nosso modelo transformado utiliza das mesmas variáveis mas agora é apresentado como:

$$Y = 5.076 - 0.014x_2 + 0.002x_3 + 0.15x_4 - 1.04x_5 + 0.14x_6 - 0.076x_7 + 0.019x_8 - 0.0008x_9 - 0.054x_{10} + 0.0006x_{11} - 0.04x_{12}$$

Verificamos então os resíduos da nossa nova transformação





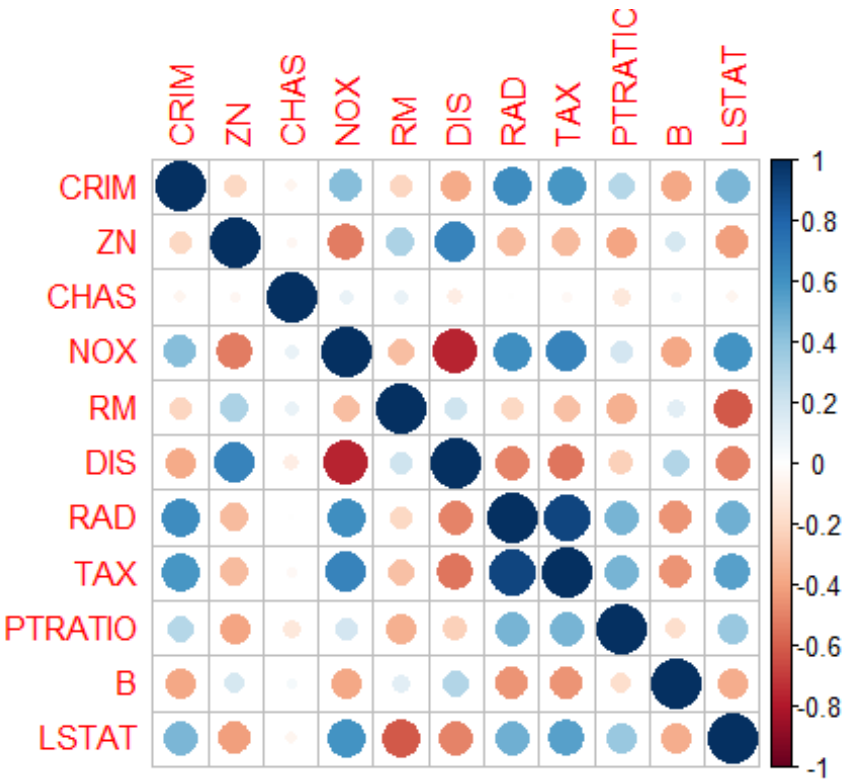
Infelizmente mesmo após extensivo trabalho relacionado às transformações, nosso modelo final ainda não é como desejávamos. A variância dos residuos de fato diminuiu e se instalou entre 3 e -3, apesar de que parece ainda não se apresentar de forma totalmente aleatória. Da mesma forma, a normalidade do modelo não foi alcançada em níveis ideais.

Diferentes versões do modelo apresentado foram testadas e incansáveis testes foram feitos em busca de algum que conseguísse nos entregar os resultados desejados, mas não importa o quão fundo procurávamos, os outros modelos possíveis eram ainda piores do que o nosso atual.

Mais análises seriam feitas a seguir em relação à multicolinearidade e uma conclusão sobre qual modelo devemos usar será feita ao seu fim.

6 Multicolinearidade

Para decidirmos de vez se nosso modelo é de fato o melhor para o banco de dados em questão, buscamos então realizar testes para verificar a existência de Multicolinearidade. Fizemos isso primeiro analisando a correlação das variáveis de nosso modelo, após disso nós calculamos o vif de cada uma e de nosso modelo. Os códigos para a análise em questão podem ser encontrados no apêndice 6.1 ao final do relatório.



Sabemos que são indícios de multicolinearidade caso alguma correlação entre as variáveis se aproxime de 1 ou -1, assim como se o VIF for maior que 5.

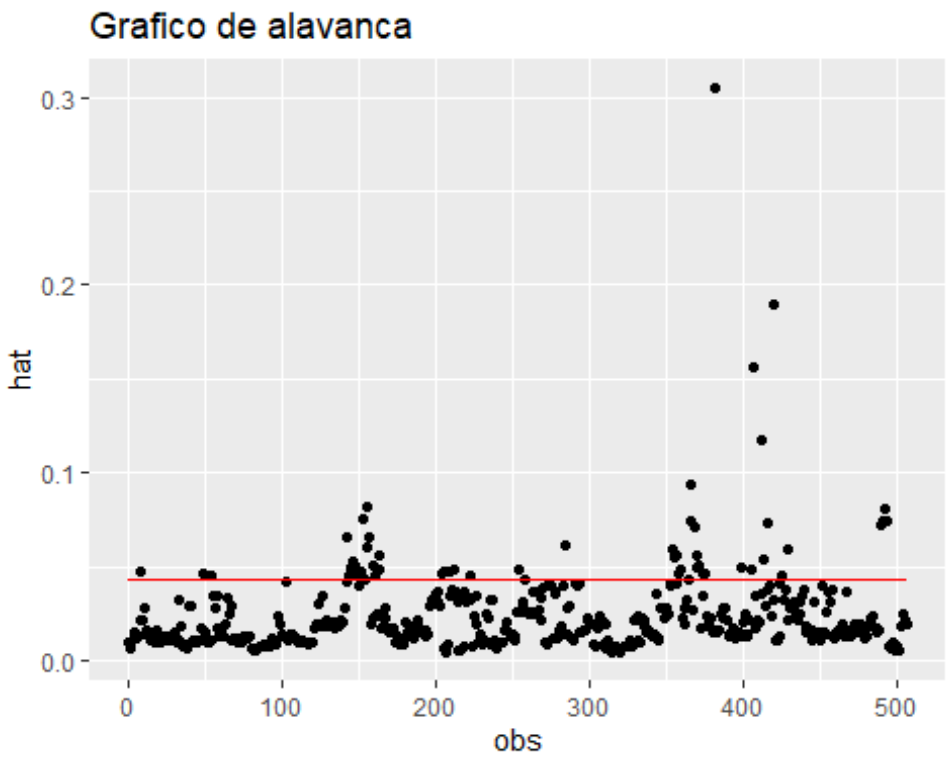
Por mais que tenhamos de fato visualizado correlações que nos chamem atenção, não foi possível realizar futuras transformações e modificações ao modelo em questão.

Não tendo acesso à uma das principais maneiras de se lidar com multicolinearidade, esta sendo a coleta de dados adicionais ou modificação da amostragem feita, nos limitamos ao banco de dados que já temos. Infelizmente, ao tentarmos outras permutações das variáveis, nosso modelo só ficava cada vez pior, apresentando diversos problemas no que as variáveis estavam fortemente ligadas umas às outras. Caso retirássemos alguma variável que estivesse causando a multicolinearidade, não só estaríamos perdendo grande significado na análise de nosso modelo em questão, como diversas outras variáveis estariam deixando de ser significativas, eventualmente causando nosso modelo a ficar ainda mais defasado.

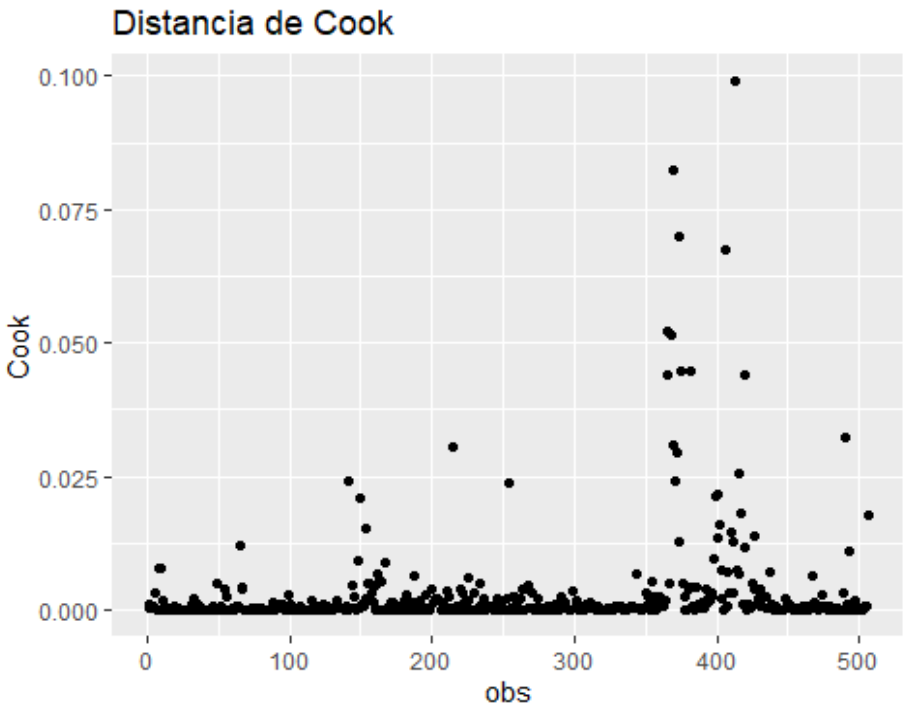
Em conclusão, após análise de resíduos e multicolinearidade, nos encontramos com nosso modelo transformado que foi visto anteriormente, sendo considerado o mais decente que conseguimos encontrar. É notável, ao menos, que nosso VIF encontrado foi pequeno para todas as variáveis e nosso VIF total foi abaixo de 5.

7 Análise de Diagnóstico

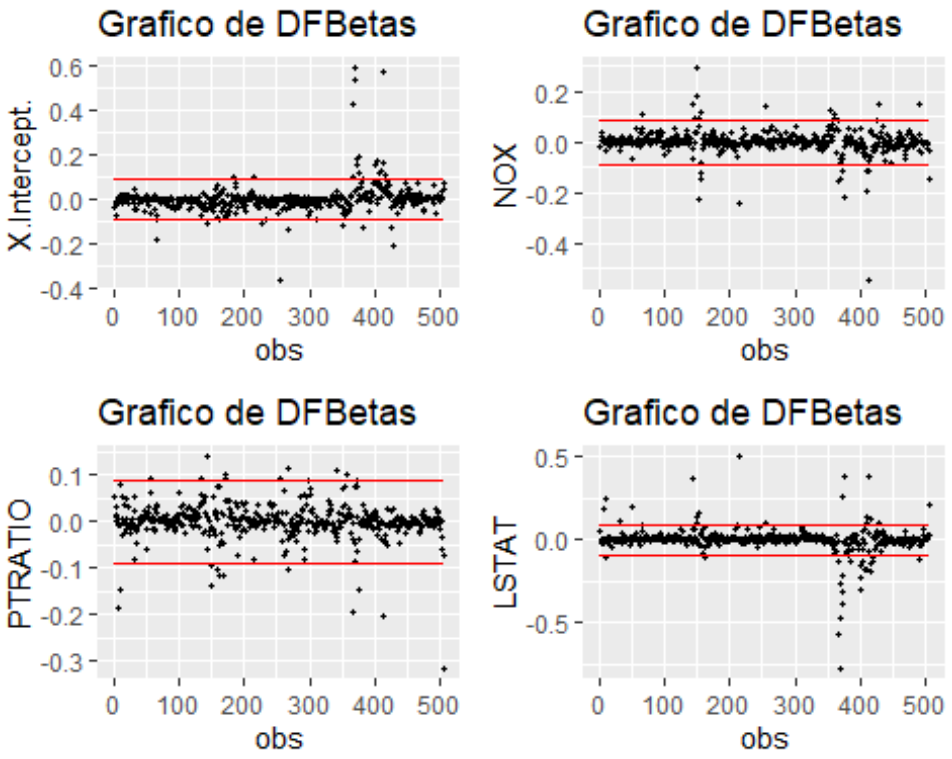
Vemos agora então como os pontos do nosso modelo, em busca de pontos de alavanca que possam influenciar as outras observações.



Percebemos um número elevado de pontos influentes ao analisarmos nosso modelo. Podemos utilizar de diferentes métodos para percebelos:



Nenhum dos pontos ultrapassa 1, o que nos dá que pelo método de Cook não encontramos pontos influentes.



Buscamos aqui pontos influentes caso retiremos algumas variáveis específicas de nosso modelo, assim como o intercepto. Pelo número exorbitante de pontos influentes, nos encontramos incapazes de retirá-los do banco de dados para uma segunda análise.

8 Análise de variancia

Analisaremos agora a variância de nosso modelo como um todo, em busca de testar novamente o quão significativo ele é agora que foi transformado. Poderíamos ter feito análise para nosso primeiro modelo criado, mas como vimos que os dados ainda não estavam limpos foi desejado realizar todas as etapas até então para encontrarmos o modelo ideal para tal. Os códigos utilizados para a criação da tabela a seguir podem ser encontrados no apêndice 8.1 ao final do relatório.

	Soma dos Quadrados	**GL**	**Quadrado Médio**	**F_c**
Regressão	133.12	11	12.102	
Resíduos	35.73	494	0.072	
Total	168.85	505		167.32

Comparando nossa estatística encontrada (167.32) com o ponto crítico da f (1.808), percebemos que de fato nosso modelo é significativo.

Calculamos também o coeficiente de determinação (R^2) que nos deu 0.788. Isso nos diz que a proporção da variabilidade da varável resposta que é explicada pelo nosso modelo de regressão é, aproximadamente, de 78.8%.

A proporção encontrada não é completamente ideal, mas considerando nosso modelo e todos os problemas encontrados até então, podemos considerar que é suficiente.

9 Intervalo de Confiança

Por fim podemos ver que os intervalos de confiança baseados na estatística t são então:

```
##                 2.5 %   97.5 %
## (Intercept)  4.51142  5.64218
## CRIM        -0.01720 -0.00989
## ZN           0.00021  0.00323
## CHAS         0.05643  0.24704
## NOX         -1.43749 -0.64863
## RM           0.09699  0.18766
## DIS         -0.09683 -0.05539
## RAD          0.01204  0.02619
## TAX         -0.00117 -0.00041
## PTRATIO     -0.06863 -0.03983
## B            0.00029  0.00089
## LSTAT       -0.04506 -0.03448
```

10 Conclusão

Como foi previsto ao início de nosso trabalho, percebemos que análises relacionadas ao custo de residências em Boston podem não ser tão fáceis de serem feitas. Com a crise imobiliaria nos Estados Unidos ainda crescente, teríamos de realizar análises muito mais complexas envolvendo um número talvez ainda maior de variáveis e observações. É possível também que uma regressão linear por sí própria não seja o bastante para explicarmos a variável resposta desejada pelo banco de dados, mas este trabalho não se propoz a se a profundar nesta questão.

Foi visto que mesmo analisando cada faceta de nosso modelo ideal, ele ainda apresentava problemas intrinsecos em sua natureza.

O que conseguimos de resultados, por outro lado, foi que de fato o preço médio das residências na região estudada é influenciado por nossas variáveis escolhidas.

Podemos concluir que de fato é justo dizer que a busca por ar limpo é um fator importante quando analisamos o custo médio de uma moradia, assim como os causadores de poluição próximos.

11 Apêndices

3.1

```
nulo = lm(MEDV~1,data=dado)
completo = lm(MEDV~.,data=dado)

step(completo, data=dado, direction="backward",trace=FALSE)

step(nulo, scope=list(lower=nulo,
upper=completo),data=dado,direction="forward",trace=FALSE)

step(completo, data=dado, direction="both",trace=FALSE)

reg <- lm(MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO
+ B + LSTAT, data = dado)

summary(reg)
```

5.2

```
MEDVtrans <- ((MEDV^0.1162)-1)/0.1162
regtrans <- data.frame(cbind(dado,MEDVtrans))
head(regtrans)

reg2 <- lm(MEDVtrans ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX +
PTRATIO + B + LSTAT)

summary(reg2)
```

6.2

```
corre2 <- cor(dado[,c(1,2,4,5,6,8,9,10,11,12,13)])

corrplot(corre2, method = "circle")

vif(dado[,c(1,6,8,11,12,13)])

x <- summary(reg2)$r.squared

1/(1-x)
```

8.1

```
n<-nrow(dado)
p<-12
```

```
anova(reg2)

SQReg2<-133.12
QMReg2<-SQReg2/(p-1)
SQRes2<-35.73
QMRes2<-SQRes2/(n-p)
SQT2<-SQReg2 + SQRes2
QMT2<-SQT2/(n-1)
EstF2<-QMReg2/QMRes2


qf(.95,p-1,n-p)

summary(reg2)$r.squared
```

12 Referências

Dados adquiridos a partir de: https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data