Prevendo os valores médios de casas nos bairros de Boston

Alunos:

Daniela Georg

Matheus Amaral

Raquel Fontes

**Objetivo do trabalho**

Criar um modelo preditivo para média de preços das casas de bairros em Boston.

**Resumo**

O modelo gerado consegue prever 80% da variabilidade dos preços com base nas variáveis crim, nox, rm, dis, rad, tax, ptratio, lstat.

Observação:

Para facilitar a leitura do documento, sinalizamos em amarelo pontos de atenção e delimitamos com caixas verdes as análises e conclusões realizadas a partir dos dados e modelos preditivos.

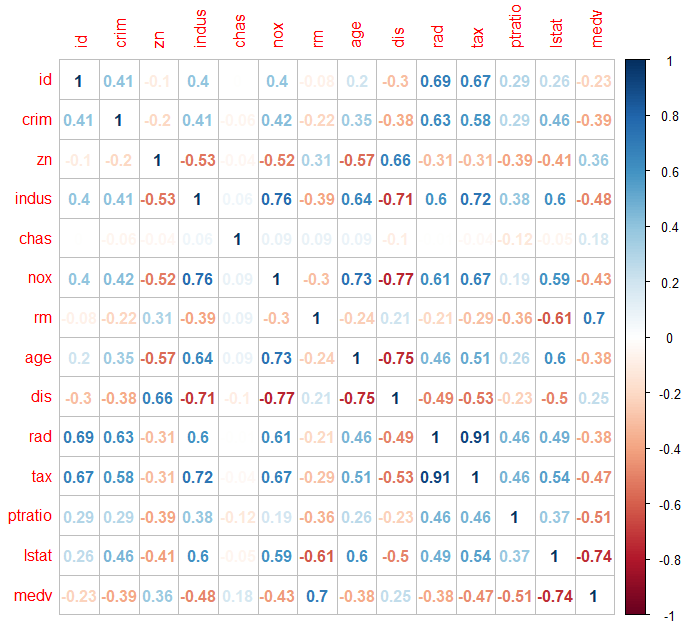
## Exploração e manipulação inicial dos dados

A exploração e análise dos dados se deu da seguinte forma

1. Buscar missing values. **Não havia nenhum.**
2. Analisar a distribuição das variáveis e entender como elas se comportam. Algumas variáveis precisaram de tratamento e transformação:

|  |  |  |
| --- | --- | --- |
| Var. | Análise | Tratamento |
| medv | Variável resposta.  Possui distribuição normal com outliers acima de 40 e um acúmulo de observações no valor 50. | Retirar as observações acumuladas no valor 50.  Just: em análise bivariada com as outras variáveis, essas observações se comportavam de forma anômala e com baixíssima correlação com qualquer variável disponível, atrapalhando o modelo. |
| dis | Distribuição de distância para centros comerciais.  A relação com a variável resposta é curvilínea. | Aplicar o log nas observações.  Just: assim diminuímos a assimetria dos dados e deixamos a distribuição mais adequada para o modelo de regressão linear. |
| rad | Índice de acesso a rodovias  Há um salto nos dados de 8 para 24, criando um hiato considerável. | Reduzir a discrepância transformando os valores 24 para 9.  Just: dessa forma, os valores continuam maiores que os outros da base e influenciarão o modelo, mas não criam um buraco tão grande e ficam mais adequados para o modelo de regressão linear. |
| Tax | Taxa de impostos  Há um salto nos dados de 450 para 666, criando um hiato considerável | Reduzir a discrepância transformando os valores 666 em 500.  Just: dessa forma, os valores continuarão maiores que os outros da base e influenciarão o modelo, mas sem criar um buraco tão grande que atrapalhe o modelo de regressão linear. |
| Chas | Indica se é às margens do rio Charles.  Variável qualitativa que foi considerada numérica na importação dos dados. | Transformar o tipo de dado para numérico.  Just: assim, o modelo pode interpretar corretamente a variável qualitativa. |
| Zn | Proporção de terras com mais de 25000m²  A distribuição se concentra em 0 e se dispersa nos outros valores | Classificar em sim e não (0 e 1).  Just: a concentração dos valores no 0 e a dispersão ao longo do restante indica que é mais importante saber se há ou não esse tipo de terreno do que saber a proporção exata. |
| lstat | Percentual da população de classes mais baixas  A relação com a variável resposta é curvilínea. | Aplicar o log nas observações.  Just: assim diminuímos a assimetria dos dados e deixamos a distribuição mais adequada para o modelo de regressão linear. |

## Entendendo a correlação entre as variáveis



Com o mapa de calor, podemos ver as principais variáveis que possuem correlação com o preço médio das casas. Além disso, conseguimos compreender se as relações são positivas ou negativas. Usaremos essas infos para avaliar o modelo.

## Regressão inicial

### Resultado da regressão com todas as variáveis

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 \*\*\*  
 ---  
 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.

crim zn indus chas nox rm  
1.6 2.1 3.9 1.1 4.7 2.0   
age dis rad tax ptratio lstat  
3.4 5.3 2.5 3.5 1.7 3.7

### Detecção de anomalias

### 

### Análise do modelo inicial

1. **Análise dos sinais dos coeficientes**

Os coeficientes das variáveis preditoras estão de acordo com o esperado a partir do mapa de calor. A única que não se comporta da forma esperada é a variável dis.

1. **Análise do p-value**

Apesar de resultar em um modelo com 80% de capacidade preditiva dos valores finais, as variáveis zn, indus, chas e age não passam no teste de hipóteses, como podemos verificar com o p-value.

1. **Análise do VIF**

A análise de multicolinearidade traz VIF de dis>5, indicando que ela possui colinearidade com outras variáveis. Para melhorar o modelo, precisamos testar o modelo sem essa variável.

1. **Análise dos outliers do modelo**

Ao avaliarmos os outliers e pontos influentes, não detectamos nenhum valor muito discrepante dos outros.

## Regressão com seleção de variáveis

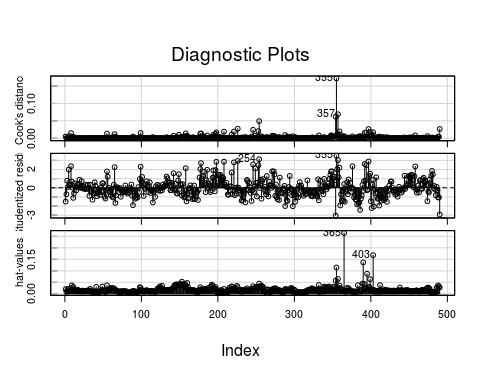
Para chegar no modelo mais adequado, comparamos os diferentes modelos a partir do Akaike Information Criterion (AIC), chegando no modelo sumarizado abaixo:

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 \*\*\*  
 ---  
 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 detecção de multicolinearidade

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



### Análise do modelo final

1. **Análise das variáveis selecionadas**

No modelo, as variáveis selecionadas foram: crim, nox, rm, dis, rad, tax, ptratio, lstat. Ficaram de fora: chas, indus, age. Essas 3 variáveis possuíam correlação menor que 0.5 com a variável resposta. Indus possuía alta colinearidade com as outras variáveis do modelo, o que também justifica sua retirada.

1. **Análise dos sinais dos coeficientes**

Os coeficientes das variáveis preditoras estão de acordo com o esperado a partir do mapa de calor. A única que não se comporta da forma esperada é a variável dis.

1. **Análise do p-value**

Todas as variáveis apresentam p-value abaixo de 0,01, indicando alta relevância estatística para sua presença no modelo.

1. **Análise do VIF**

Todos os VIFs estão abaixo de 5, indicando baixa colinearidade entre as variáveis.

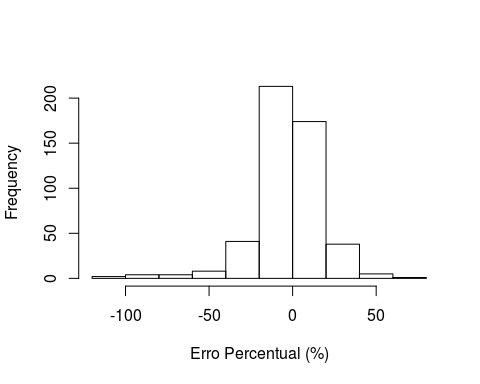
1. **Análise de outliers do modelo**

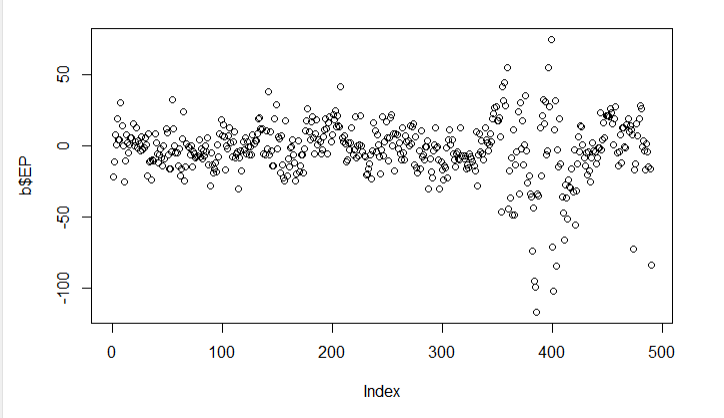
O modelo já não apresenta outliers significativos.

O modelo consegue prever 80% dos resultados, uma taxa bastante significativa. Além disso, possui relevância estatística e relevância econômica.

**Testando e validando o modelo**

### Erro percentual





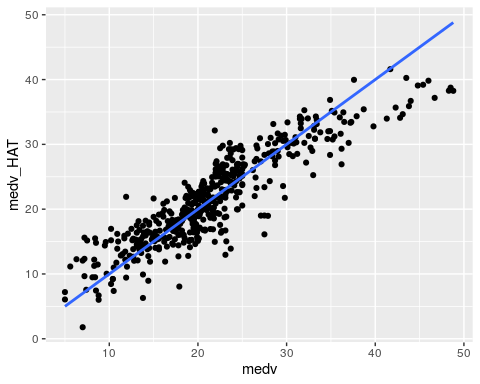
### MAPE

0.1407178

O percentual de erro do modelo se concentra em torno de 20%, com alguns pontos se dispersando além dessa faixa. Como sabemos que os valores reais da variável resposta possuem outliers, essa variação é esperada.

O valor do MAPE indica uma excelente capacidade preditiva do modelo.

### Valores reais VS Valores previstos



Preços previstos

Preços originais

O modelo se aproxima bastante dos valores reais. Vale ressaltar que os valores acima de 40 – onde o modelo mais se distancia – são outliers da variável Preço que preferimos não retirar para não enviesar o modelo.

### Teste anova

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

Com o ANOVA, conseguimos fazer um segundo teste de hipóteses para verificar se as variáveis explicam os preços. Como todas as variáveis apresentam p-value<0.05, podemos validar o modelo!