Banco de dados: Boston House Prices

Fernado Bispo, Jeff Caponero

Sumário

Introdução

A busca pela moradia própria é o desejo da grande maioria das pessoas, contudo a conquista desse bem nos grandes centros não é tarefa fácil. Levando isso em consideração a procura por imóveis na região metropolitana torna-se uma opção viável economicamente, mesmo havendo penalizações no que diz respeito a distância e congestionamentos.

O objetivo deste relatório é trazer a luz as análises e conclusões acerca da utilização das técnicas de regressão linear a fim de determinar o preço das casas em Boston, baseado nos dados fornecidos pelo conjunto de dados obtido. Neste primeiro momento, em que se utilizará a regressão linear simples, se buscará determinar uma função que descreva a relação entre o Valor Médio dos imóveis e o Percentual da população de "classe baixa".

Composto por 506 observações e 14 variáveis, o conjunto de dados, publicado no *Jornal of Environmental Economics & Management*, vol.5, 81-102, 1978.t, traz inúmeras características que servirão de parâmetros para resolução do seguinte questionamento: O valor médio dos imóveis é influenciado pelas diversas características externas observadas?

Metodologia

Sobre o conjunto de dados

Os dados utilizados apresentam os preços de 506 casas em Boston publicados por *'D. Harrison e D.L.Rubinfeld'* [1] e usados por *'Belsley, Kuh & Welsch'* para avaliar a demanda por ar limpo no valor de casas do município.

Os dados podem ser acessados na plataforma para aprendizado de ciência de dados Kaggle através do link:

https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data.

Variáveis a serem analisadas

A amostra contêm 14 atributos de casas em diferentes locais nos subúrbios de Boston no final dos anos 1970, sendo duas delas classificadas como categóricas e 12 como numéricas. O objetivo é o valor médio das casas em um local (em milhares de dólares - k$). As variáveis presentes no banco de dados são descritas a seguir bem como a forma que estas variáveis serão representadas ao longo deste relatório a fim de facilitar o entendimento:

1. CRIM: Índice de criminalidade per capita por bairro. Taxa de criminalidade por cidade. Uma vez que o CRIM mede a ameaça ao bem-estar que as famílias percebem em vários bairros da área metropolitana de Boston (assumindo que as taxas de criminalidade são geralmente proporcionais às percepções de perigo das pessoas), ele deve ter um efeito negativo nos valores das moradias. Será representada como **Índice Criminalidade**.

2. ZN: Proporção da área residencial de uma cidade dividida em lotes com mais de 25.000 pés quadrados. Uma vez que tal zoneamento restringe a construção de pequenas casas em lotes, esperamos que o ZS esteja positivamente relacionado aos valores das moradias. Um coeficiente positivo também pode surgir porque o zoneamento representa a exclusividade, a classe social e as comodidades externas de uma comunidade. Será representada como **Prop. Terreno Zoneado**.

3. INDUS: Proporção de hectares de negócios não varejistas por bairro. O INDUS serve como um *proxy* para as externalidades associadas ao ruído da indústria, tráfego intenso e efeitos visuais desagradáveis e, portanto, deve afetar negativamente os valores das habitações. Será representado por **Área Industrial**.

4. CHAS: Variável fictícia categórica que representa imóveis próximos a margem do rio Charles (1 se o trecho margeia o rio; 0 caso contrário). Esperamos que o CHAS esteja positivamente relacionado aos valores das moradias, uma vez que podem denotar imóveis de mais alto padrão. Será representada como **Margem**.

5. NOX: Concentração de óxidos nítricos em pphm (partes por 100 milhões). Como o aumento dos valores de NOX representam uma piora da qualidade do ar, esperamos que esta variável esteja negativamente correlacionada com o valor dos imóveis. Será representada como **Índice Oxido Nítrico**.

6. RM: Número médio de quartos em unidades proprietárias. RM representa espaço e, em certo sentido, quantidade de habitação. Deve estar positivamente relacionado com o valor da habitação. Verificou-se que a forma RM^2 fornece um ajuste melhor do que as formas linear ou logarítnica. Será representada por **N° de Cômodos**.

7. AGE: Proporção de unidades próprias construídas antes de 1940. A idade da unidade geralmente está relacionada à qualidade da estrutura e portanto negativamente correlacionada ao valor do imóvel. Será representada por **Idade**

8. DIS: Distâncias ponderadas para cinco centros de emprego na região de Boston. De acordo com as teorias tradicionais de gradientes de renda da terra urbana, os valores das moradias devem ser maiores perto de locatários de emprego. DIS é inserido na forma logarítmica; o sinal esperado é negativo. Será representada como **Dist. Empregos**.

9. RAD: Variável categórica que representa o índice de acessibilidade às rodovias radiais. O índice de acesso rodoviário foi calculado com base na cidade. Boas variáveis de área de estrada são necessárias para que todas as variáveis de poluição não capturem as vantagens locacionais de estradas. O RAD captura outros tipos de vantagens locacionais além da proximidade do local de trabalho. é inserido na forma logarítmica; o sinal esperado é positivo. Será representada como **Acessibilidade Rodovias**.

10. TAX: Valor total do imposto sobre a propriedade (\$/\$10,000). Mede o custo dos serviços públicos na comunidade terrestre. As taxas de imposto nominais foram corrigidas pelos índices de avaliação locais para gerar o valor total da taxa de imposto para cada cidade. Diferenças intramunicipais na taxa de avaliação eram difíceis de obter e, portanto, não eram usadas. O coeficiente desta variável deve ser negativo. Será representada como **Imposto**.

11. PTRATIO: Proporção aluno-professor por distrito escolar da cidade. Mede os benefícios do setor público em cada cidade. A relação do rácio aluno-professor com a qualidade da escola não é totalmente clara, embora um rácio baixo deva significar que cada aluno recebe mais atenção individual. Esperamos o sinal em PTRATIO seja negativo. Será representada como **Prop. Prof.-Aluno**.

12. B: O resultado da equação $B = 1000(Bk - 0,63)^2$ onde Bk é a proporção de negros por bairro. Em níveis baixos a moderados de B, um aumento em B deve ter uma influência negativa no valor da habitação se os negros forem considerados vizinhos indesejáveis

pelos brancos. No entanto, a discriminação de mercado significa que os valores das moradias são mais altos em níveis muito altos de B. Espera-se, portanto, uma relação parabólica entre a proporção de negros em um bairro e os valores das moradias. Será representada por **Prop. Negros/bairro**.

13. LSTAT: Proporção da população de "classe baixa", ou seja, com status inferior = 1/2 (proporção de adultos sem nível de ensino médio e proporção de trabalhadores do sexo masculino classificados como trabalhadores). A especificação logarítmica implica que as distinções de status socioeconômico significam mais nas camadas superiores da sociedade do que nas classes inferiores. Será representada por **Pop. Classe Baixa**.

Variável de Saída (Resposta):

- Valor do Imóvel: Valor médio de residências ocupadas pelo proprietário em US$1.000 [Milhares de dólares - k$].

Fonte

StatLib - Carnegie Mellon University

PARTE 1 - Análise de Regressão Linear Simples

Resultados

Análise Descritiva

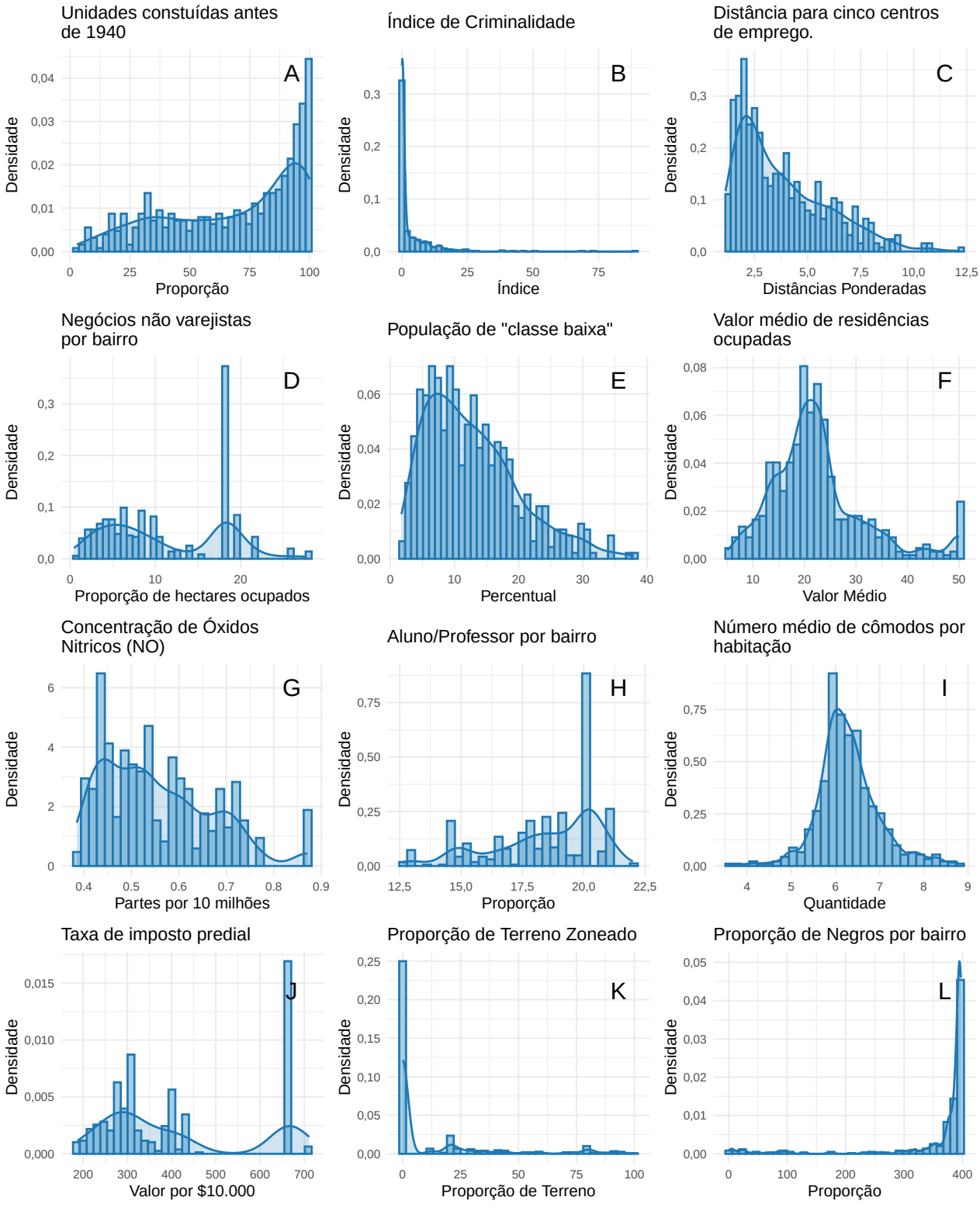
De modo a conhecer melhor o banco de dados analisado é importante realizar uma análise descritiva das variáveis que o compõem. Na Tabela 1 pode se ver as medidas de resumo de posição e de tendência central destas variáveis.

Tabela 1: Medidas Resumo dos dados

	Mín	Q1	Med	Média	Q3	Máx	Desv.padrão	CV	Assimetria	Curtose
Área Industrial	0,46	5,19	9,69	11,14	18,10	27,74	6,86	0,62	0,29	-1,24
Dist. Empregos	1,13	2,10	3,21	3,80	5,21	12,13	2,11	0,55	1,01	0,46
Idade do Imóvel	2,90	45,00	77,50	68,57	94,10	100,00	28,15	0,41	-0,60	-0,98
Imposto Propriedade	187,00	279,00	330,00	408,24	666,00	711,00	168,54	0,41	0,67	-1,15
Índice Criminalidade	0,01	0,08	0,26	3,61	3,68	88,98	8,60	2,38	5,19	36,60
Índice Oxido Nítrico	0,38	0,45	0,54	0,55	0,62	0,87	0,12	0,21	0,72	-0,09
N° Cômodos	3,56	5,88	6,21	6,28	6,62	8,78	0,70	0,11	0,40	1,84
Pop. Classe Baixa	1,73	6,93	11,36	12,65	16,96	37,97	7,14	0,56	0,90	0,46
Prop. Negros/bairro	0,32	375,33	391,44	356,67	396,23	396,90	91,29	0,26	-2,87	7,10
Prop. Prof.-Aluno	12,60	17,40	19,05	18,46	20,20	22,00	2,16	0,12	-0,80	-0,30
Prop. Terreno Zoneado	0,00	0,00	0,00	11,36	12,50	100,00	23,32	2,05	2,21	3,95
Valor do Imóvel	5,00	17,00	21,20	22,53	25,00	50,00	9,20	0,41	1,10	1,45

Para facilitar a compreenção das medidas apresentadas na Tabela 1, a Figura 1 mostra graficamente estas distribuições.

Figura 1: Histogramas das variáveis em análise.



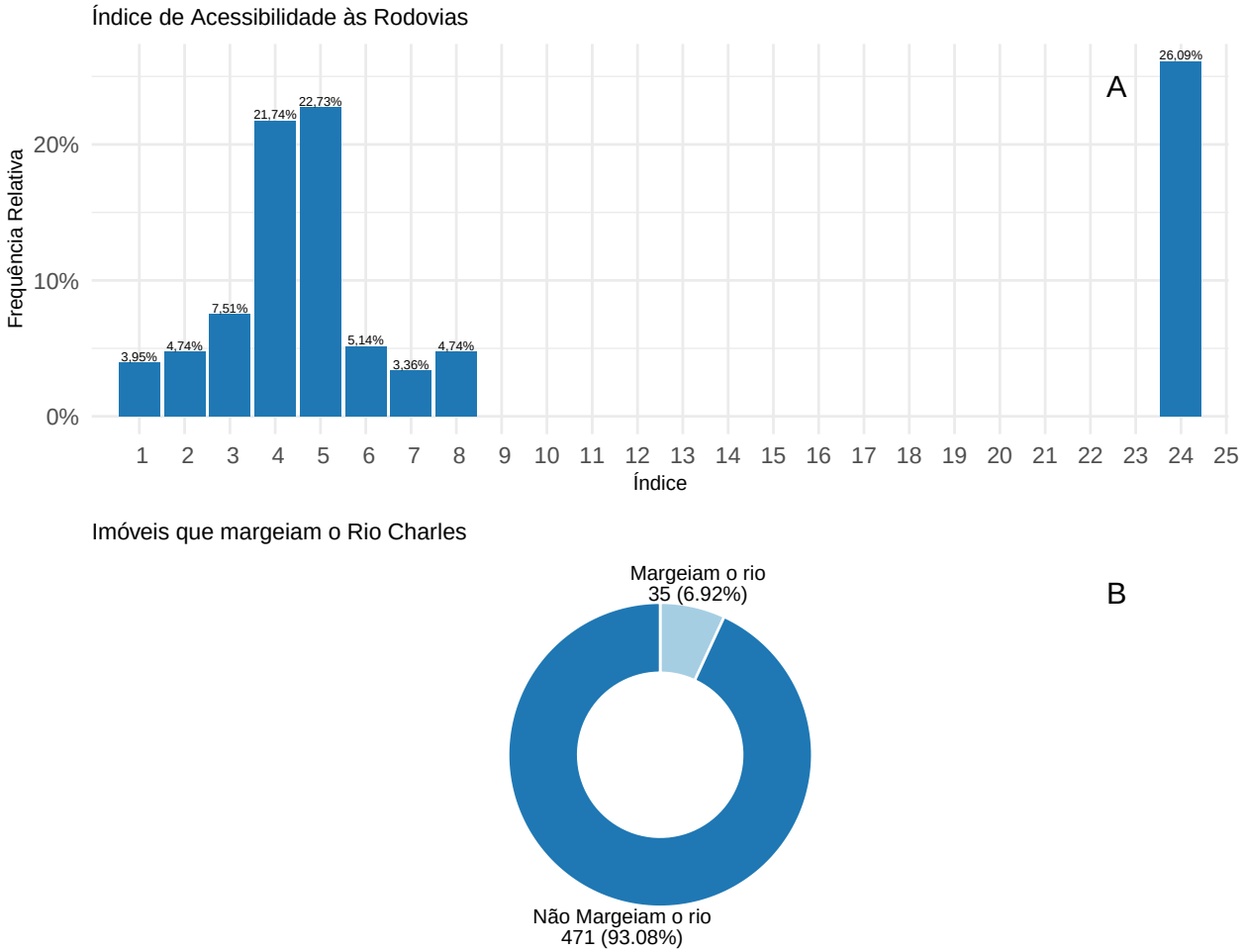
Fonte: StatLib − Carnegie Mellon University

Desta análise inicial, destancam-se algumas características:

- O Índice de criminalidade per capita por bairro é bastante baixo na maioria dos bairros (Figura 1B);
- A Proporção de terreno residencial zoneada para lotes acima de 25.000 sq.ft. contém uma alta concentração de valores zeros (Figura 1K);
- Verifica-se uma concentração de empresas com cerca de 18 hectares em diversos bairros (Figura 1D);
- Há uma concentração de imóveis com alto valor total do imposto predial (Figura 1J);
- A maior parte dos bairros tinha alta proporção de negros (Figura 1L).

Tendo em vista o fato de as variáveis categóricas não estarem representadas na Figura 1, foi construída a figura 2 com gráficos que possibilitam a avaliação do comportamento dessas variáveis de maneira mais adequada.

Figura 2: Distribuição de Frequência das Variáveis Categóricas.



Fonte: StatLib − Carnegie Mellon University

Avaliando a Figura 2 é possível constatar na Figura 2A que representa a frequência do Índice de Acessibilidade às Rodovias a existência de um comportamento tri modal e que há

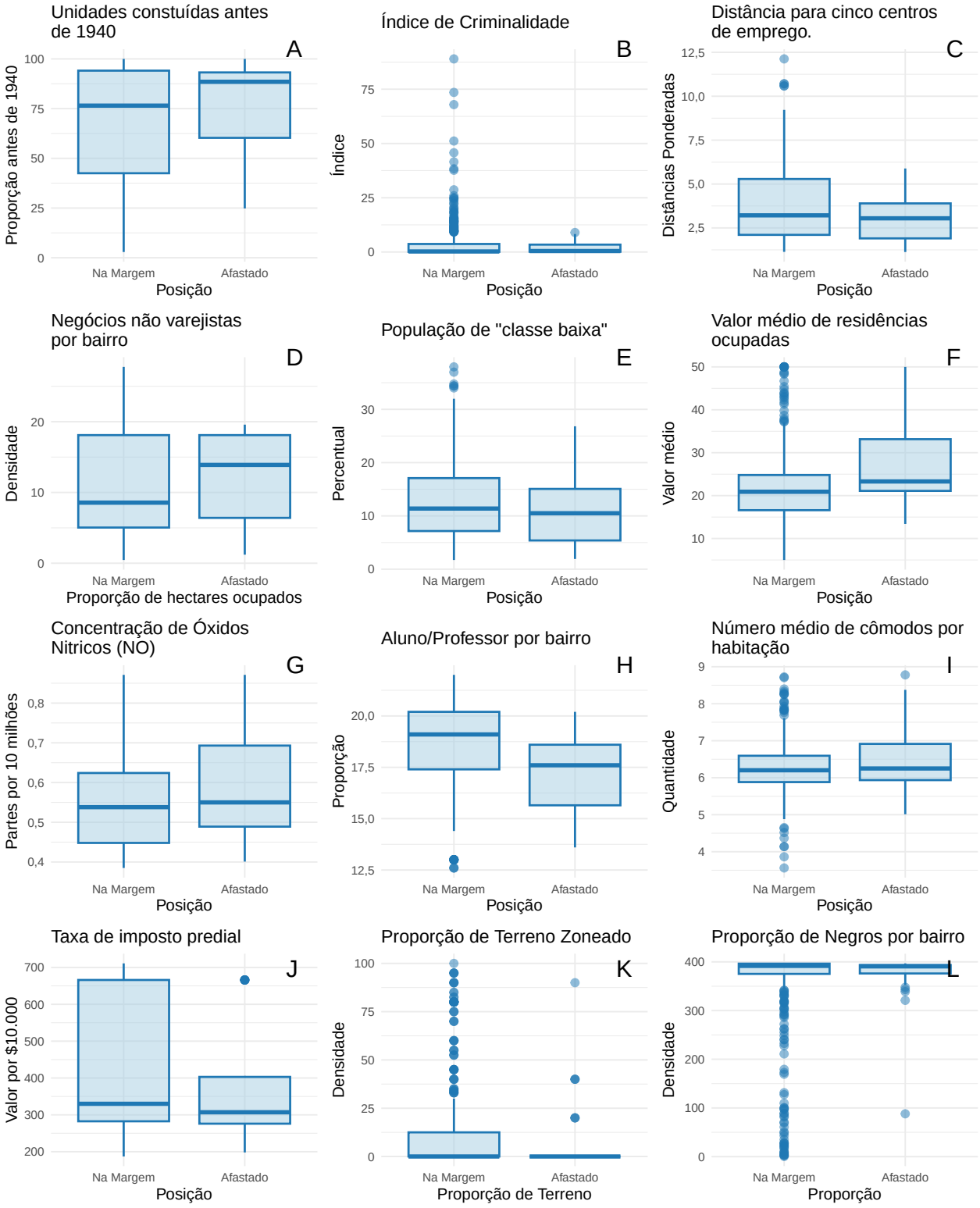
uma lacuna entre os índices, sendo este *gap* entre o índice 8 e o índice 24 indicando assim que há uma concentração de imóveis próximos a acessos a rodovias e que há uma parcela, cerca de 26%, que estão distantes desses acessos, podendo pressupor que esses imóveis são desvalorizados em relação aos mais próximos.

A Figura 2B que retrata os Imóveis que margeiam o Rio Charles é possível identificar que mais de 93% não margeiam o rio, apenas uma pequena parcela margeia, podendo pressupor uma maior valorização dos imóveis que margeiam o rio.

Análise de Dados Atípicos

Com base na variável que indica se o imóvel margeia ou não o Charles River, pode-se realizar a análise de dispersão dos dados por meio de gráficos do tipo BoxPlot, como se vê na Figura 1.

Figura 3: BoxPlot das variáveis em análise.



Fonte: StatLib − Carnegie Mellon University

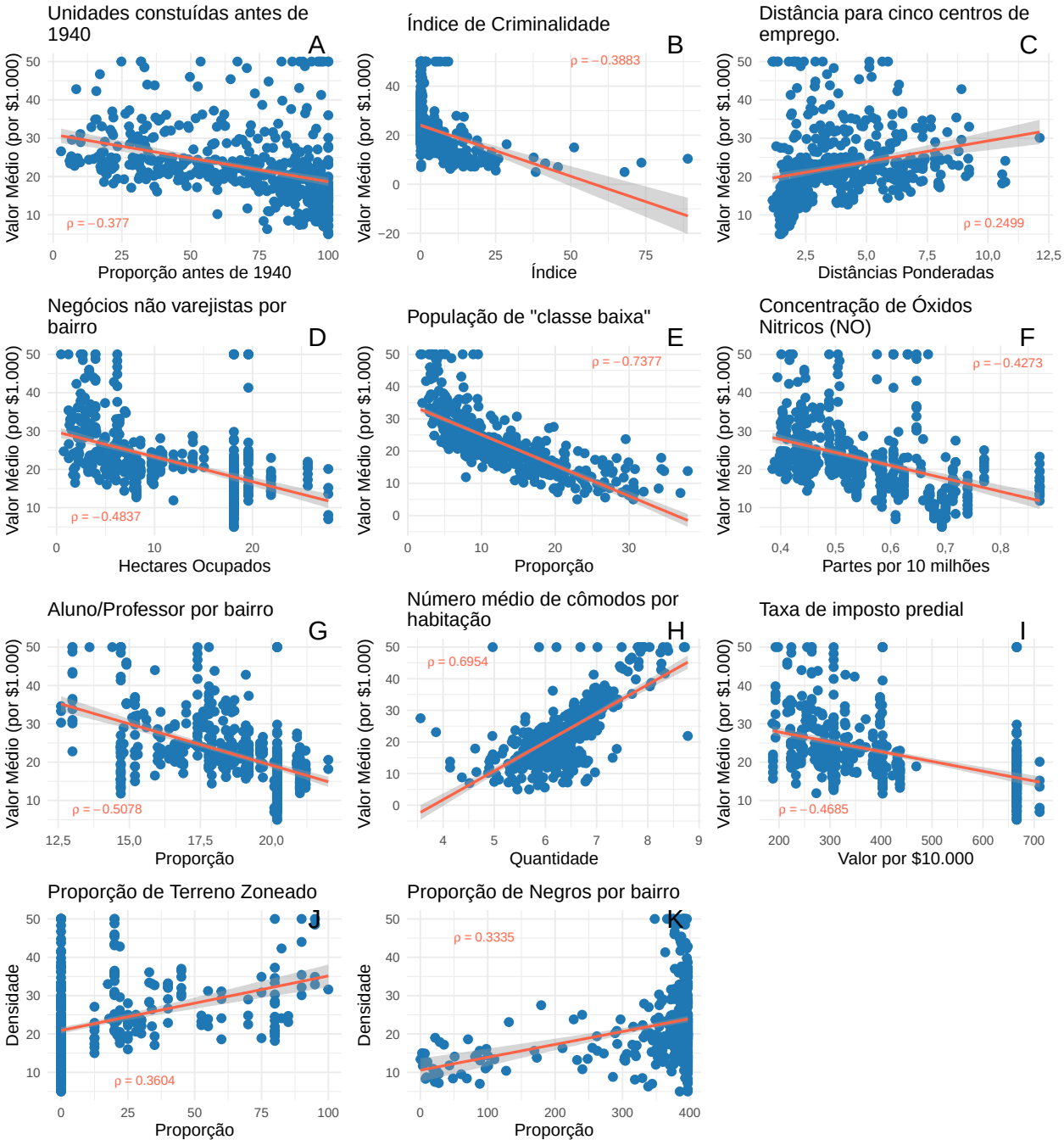
Pode-se verificar pela Figura 3 que cerca de 2/3 das variáveis apresentam valores atípicos

(*outliers*), entretanto o tratamento destes *outliers* em todos os casos parece, salvo melhor juízo, ser o mesmo. Observa-se que há coerência entre eles, isto é, são realizações possíveis e não devem ser desprezadas como se fossem erros ou dados irrelevantes. Isto se deve a tremenda variedade em tipos, propósitos e status dos imóveis avaliados. Esses dados por sua vez, representam um maior desafio ao modelamento a que esse trabalho se propõe.

Relação entre as variáveis

Antes da proposição do modelo de regressão mais bem elaborado é conveniente uma avaliação gráfica da dispersão dos valores das variáveis em relação à variável resposta **Valor Médio do Imóvel**. A Figura 4 apresenta essas dispersões de pontos e já apresenta uma linha de tendência para os valores observados.

Figura 4: Reta de regressão ajustada entre o Valor médio dos imóveis e demais medições



Para avaliar a significância das correlações entre as variáveis com relação ao **Valor Médio do Imóvel** segue a Tabela 2 com os resultados do Teste de Hipóteses com nível de significância de 5% que tem como hipóteses:

$$H_0 : \hat{\rho} = 0$$

$$H_1 : \hat{\rho} \neq 0.$$

Tabela 2: Teste de Hipótese para Correlação

	t	p-valor	LI	LS
Índice Criminalidade	-9.46	<0,0001	-0.46	-0.312
Prop. Terreno Zoneado	8.675	<0,0001	0.282	0.434
Área Industrial	-12.408	<0,0001	-0.548	-0.414
Índice Oxido Nítrico	-10.611	<0,0001	-0.496	-0.353
N° Cômodos	21.722	<0,0001	0.647	0.738
Idade do Imóvel	-9.137	<0,0001	-0.449	-0.3
Dist. Empregos	5.795	<0,0001	0.166	0.33
Imposto Propriedade	-11.906	<0,0001	-0.534	-0.398
Prop. Prof.-Aluno	-13.233	<0,0001	-0.57	-0.44
Prop. Negros/bairro	7.941	<0,0001	0.254	0.409
Pop. Classe Baixa	-24.528	<0,0001	-0.775	-0.695

Nota: Teste realizado com 5% de significância

Conforme expresso na Tabela 2, levando em consideração o **p-valor** a Hipótese Nula foi rejeitada, e com 95% de confiança se pode afirmar que **é significativa a relação linear entre todas as variáveis em estudo.**

Na avaliação da Figura 4, observa-se que nenhuma das variáveis tem uma forte correlação com o valor médio dos imóveis. A Tabela 3, apresenta os valores calculados de $\hat{\beta}_0$ e $\hat{\beta}_1$ que estimam os valores do modelo $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ com seus respectivos erros padrão (σ_0 e σ_1), além de calcular o p-valor desta regressão linear como forma de identificar a rejeição ou não do modelo proposto. Nesta mesma linha, o valor estimado do Coeficiente de Determinação (R^2) também foi calculado.

Tabela 3: Sumarização dos Modelos Ajustados de Regressão Linear Simples - RLS.

	Sumarização para Beta 0			Sumarização para Beta 1			
	Estimativa	Erro Padrão	p-valor	Estimativa	Erro Padrão	p-valor	R^2
Índice Criminalidade	24.033	0.409	<0,0001	-0.415	0.044	<0,0001	0.151
Área Industrial	29.755	0.683	<0,0001	-0.648	0.052	<0,0001	0.234
Índice Oxido Nítrico	41.346	1.811	<0,0001	-33.916	3.196	<0,0001	0.183
N° Cômodos	-34.671	2.65	<0,0001	9.102	0.419	<0,0001	0.484
Idade do Imóvel	30.979	0.999	<0,0001	-0.123	0.013	<0,0001	0.142
Dist. Empregos	18.39	0.817	<0,0001	1.092	0.188	<0,0001	0.062
Imposto Propriedade	32.971	0.948	<0,0001	-0.026	0.002	<0,0001	0.22
Prop. Prof.-Aluno	62.345	3.029	<0,0001	-2.157	0.163	<0,0001	0.258
Pop. Classe Baixa	34.554	0.563	<0,0001	-0.95	0.039	<0,0001	0.544
Prop. Terreno Zoneado	20.918	0.425	<0,0001	0.142	0.016	<0,0001	0.13
Prop. Negros/bairro	10.551	1.557	<0,0001	0.034	0.004	<0,0001	0.111

Interpretação dos Parâmetros dos Modelos Ajustados

Baseado na análise dos gráficos de dispersão e considerando os modelos ajustados expressos na Tabela 3 é possível constatar que tanto $\hat{\beta}_0$ quanto $\hat{\beta}_1$ são significantes para todos os modelos ajustados, com base no p-valor, levando em consideração a ordem de precedência das variáveis na referida tabela, seguem as interpretações com base nos parâmetros estimados:

- Modelo que avalia o Índice de Criminalidade:
  - Para cada **valorização** de 0,415 no Índice de Criminalidade o Valor Médio dos Imóveis decresce em $24 033,00 ou seja, os imóveis se desvalorizam em regiões cujo Índice de Criminalidade é elevado.
- Modelo que avalia a Área Industrial:
  - O Valor Médio dos Imóveis decresce em $29 755,00 para cada **valorização** proporcional de 0,648 hectares de Negócios não Varejistas.
- Modelo que avalia o Índice Oxido Nítrico:
  - Há uma **desvalorização** de cerca de $41 346,00 no Valor Médio dos Imóveis para cada aumento de 33 916 pphm (partes por 100 milhões) no Índice Oxido Nítrico, ou seja, há uma desvalorização no valor dos imóveis que estão situados em regiões cujo ar é mais poluída.
- Modelo que avalia o N° Cômodos:
  - Há um **valorização** de cerca de $34 671,00 no Valor Médio dos Imóveis para cada aumento de aproximadamente 9 cômodos, ou seja, os imóveis são mais valorizados a medida que possuem mais cômodos.
- Modelo que avalia a Idade do Imóvel:

  - Há uma **desvalorização** de cerca de $30 979,00 para cada aumento proporcional de 0,123 unidades próprias construídas antes de 1940, ou seja, os imóveis são desvalorizados à medida que são mais antigos.

- Modelo que avalia a Dist. Empregos:
  - Há um **valorização** de cerca de $18 390,00 no Valor Médio dos Imóveis à medida que a distância para os centros de emprego na região de Boston aumenta.

- Modelo que avalia o Imposto de Propriedade:
  - Há uma **desvalorização** de cerca de $32 971,00 no Valor Médio dos Imóveis à medida que há um aumento de aproximadamente 0,026 no valor proporcional total do Imposto de Propriedade, ou seja, quanto maior o imposto pago na região, maior a desvalorização do imóvel.

- Modelo que avalia a Prop. Prof.-Aluno:
  - Há uma **desvalorização** de cerca de $62 345,00 no Valor Médio dos Imóveis à medida que há um aumento de aproximadamente 2,157 na proporção professor aluno. Como a própria descrição da variável descreve como confusa essa relação, de fato se mostra, com base no modelo ajustado da variável, pois demostra que os imóveis se desvalorizam a medida que os benefícios do setor público aumentam!

- Modelo que avalia a Pop. Classe Baixa:
  - Há uma **desvalorização** de cerca de $34 554,00 no Valor Médio dos Imóveis à medida que há um aumento de aproximadamente 0,950 na proporção da Pop. de Classe Baixa, ou seja, os imóveis se desvalorizam a medida que a classe social dos habitantes da região cai.

- Modelo que avalia a Prop. Terreno Zoneado:
  - Há um **valorização** de cerca de $20 918,00 no Valor Médio dos Imóveis à medida que há um aumento de aproximadamente 0,142 na Prop. Terreno Zoneado, ou seja, os imóveis são valorizados em regiões com maior proporção de zoneamento.

- Modelo que avalia a Prop. Negros/bairro:
  - Há um **valorização** de cerca de $10 551,00 no Valor Médio dos Imóveis à medida que há um aumento de aproximadamente 0,034 na proporção de Prop. Negros/bairro, ou seja, os imóveis são mais valorizados em regiões cuja proporção de negros é maior.

Tendo em vista que nesse primeiro momento a proposta é a implementação de técnicas de Regressão Linear Simples - RLS, a escolha de uma variável explicativa que aparenta melhor possibilidade de explicação do Preço Médio dos Imóveis se faz necessária. Após análise dos gráficos de dispersão, Coeficiente de Correlação e avaliação dos Coeficientes de Determinação a variável escolhida foi **Pop. Classe Baixa**, logo as análises a seguir serão direcionadas a avaliar o modelo com esta variável.

Significância do Modelo

Tendo em vista a necessidade de se avaliar a significância dos parâmetros, o teste de hipótese para tal situação será realizado, contendo as seguintes hipóteses:

$$H_0 : \hat{\beta}_0 = 0$$
$$H_1 : \hat{\beta}_0 \neq 0.$$

As Tabelas 4 e 5 trazem os principais resultados da tabela ANOVA e do Intervalo de Confiança para os parâmetros, possibilitando assim inferir sobre o modelo ajustado.

Tabela 4: Resultados da ANOVA.

	GL	Soma de Quadrados	Quadrado Médio	Valor F-Snedecor	p-valor
Regressão	1	23.243,91	23.243,91	601.62	<0,0001
Resíduos	504	19.472,38	38,64		

Tabela 5: Intervalo de Confiança.

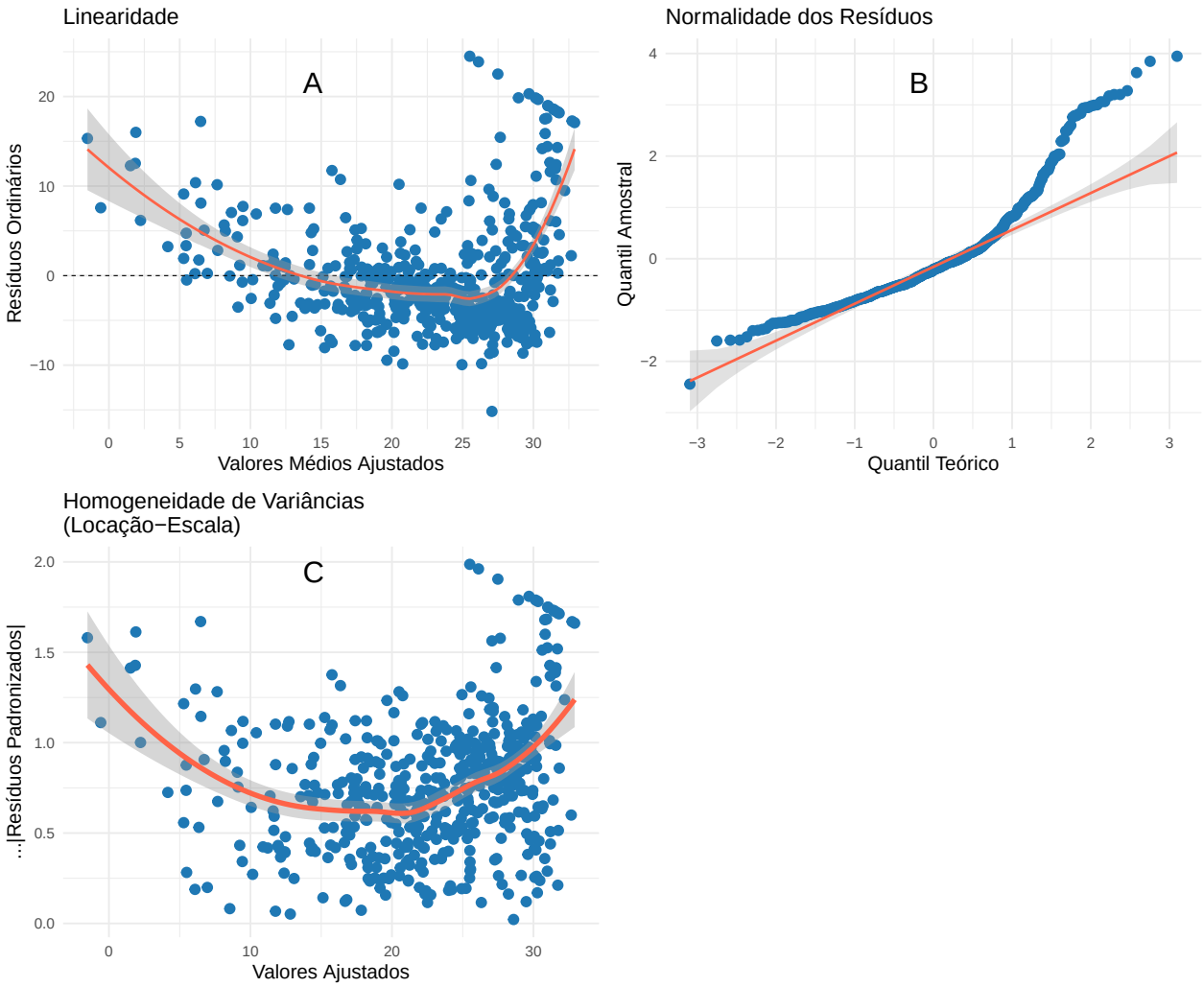
Parâmetro	LI	LS
Beta 0	33.448	35.659
Beta 1	-1.026	-0.874

A Tabela 4, que traz os resultados da tabela ANOVA, para o modelo que avalia o **Valor Médio dos Imóveis** em relação a **Pop. Classe Baixa**, corrobora com a significância do $\hat{\beta}_1$, pois sendo o p-valor menor que o nível de significância ($\alpha = 5\%$) possibilita rejeitar H_0, indicando ser significante para o modelo. O Intervalo de Confiança para os parâmetros estimados, Tabela 5, mostra que **com 95% de confiança é possível afirmar que o verdadeiro valor de** $\hat{\beta}_0$ está entre (33,4485; 35,6592) e que o verdadeiro valor de $\hat{\beta}_1$ está entre (-1,0261; -0,8740).

Análise de Resíduos

Sendo de fundamental importância para a verificação da bondade do modelo, a análise de resíduos possibilita avaliar se refletem o comportamento do modelo, para tanto se construiu a Figura 5 para iniciar as análises descritas.

Figura 5: Análise de resíduos do modelo ajustado



Analisando a Figura 5A, que traz o gráfico que avalia a **linearidade** do modelo se constata uma não aleatoriedade, possibilitando identificar um certo padrão, aparentando um afunilamento dos dados, indicando assim uma variância não constante (pela mudança da amplitude dos dados), em que mostra uma menor variabilidade dos resíduos no início e segue aumentando a medida que crescem os valores ajustados.

O gráfico que avalia a Normalidade dos Resíduos, Figura 5B, também não mostra um comportamento adequado para considerar o modelo como bom, pois é possível identificar que as caldas fogem e muito da reta de referência e do intervalo de confiança, inclusive, demonstrando haver pontos significativamente influentes que afetam o comportamento dos resíduos, logo, **se conclui que este não é um bom modelo para explicar o valor médio dos imóveis**.

O gráfico que avalia a Homogeneidade de Variâncias, Figura 5C, mostra que a Variância não é constante.

Ainda assim, para fins de implementação das técnicas apresndidas até o momento, serão

realizados os **Testes de Diagnóstico** para avaliação dos resultados, com a expectativa dos mesmos correoborarem com as interpretações obtidas através da análise gráfica.

Testes de Diagnóstico

Serão realizados os Testes de Significância, como forma secundária de avaliação, sendo descritos os testes aplicados para fins didáticos, tendo em vista a conclusão obtida com as análises gráficas, sendo estes:

- Normalidade:
  - Teste de Kolmogorov-Smirnov
  - Teste de Shapiro-Wilks
- Homocedasticidade:
  - Teste de Goldfeld-Quandt
  - Teste de Breush-Pagan
  - Teste de Park
- Linearidae:
  - Teste F para linearidade
- Independência:
  - Teste para avaliação da independência dos resíduos

Sendo estes uma forma secundária de avaliação:

Teste de Kolmogorov-Smirnov

Avalia o grau de concordância entre a distribuição de um conjunto de valores observados e determinada distribuição teórica. Consiste em comparar a distribuição de frequência acumulada da distribuição teórica com aquela observada. Realizado o teste obteve-se um p-valor de aproximadamente 0, o que inviabiliza rejeitar a hipótese de que haja normalidade entre os dados, com um grau de confiabilidade minimamente razoável.

Teste de Shapiro-Wilks

O teste de Shapiro-Wilks é um procedimento alternativo ao teste de Kolmogorov-Smirnov para avaliar normalidade. Realizado o teste obteve-se um p-valor de aproximadamente 0, o que, semelhantemente, inviabiliza rejeitar a hipótese de que haja normalidade entre os dados, com um grau de confiabilidade minimamente razoável.

Teste de Goldfeld-Quandt

Esse teste envolve o ajuste de dois modelos de regressão, separando-se as observações das duas extremidades da distribuição da variável dependente. Realizado o teste obteve-se um p-valor de aproximadamente 0.05815, o que demanda rejeitar a hipótese de que haja homocedasticidade entre os dados, com um grau de confiabilidade de 95%. Entretanto, como o p-valor obtido é próximo do necessário para a rejeição da hipotese nula, cabe um novo teste para a confirmação do resultado obtido.

Teste de Breush-Pagan

Esse teste é baseado no ajuste de um modelo de regressão em que a variável dependente é definida pelos resíduos do modelo de interesse. Se grande parte da variabilidade dos resíduos não é explicada pelo modelo, então rejeita-se a hipótese de homocedasticidade. Realizado o teste obteve-se um p-valor de aproximadamente 0, desta foram deve-se rejeitar a hipótese de que haja homocedasticidade entre os dados, com um grau de confiabilidade de 95%.

Teste de Park

Esse teste é baseado no ajuste de um modelo de regressão em que a variável dependente é definida pelos quadrados dos resíduos do modelo de interesse. Nesse caso, se β_1 diferir significativamente de zero, rejeita-se a hipótese de homocedasticidade. O valor de β_1 obtido no teste foi de -1.962 com p-valor de aproximadamente 0. Por esse teste não se deve rejeitar a hipótese de homocedasticidade, com confiabilidade de 95%.

Teste F para linearidade

O teste da falta de ajuste permite testar formalmente a adequação do ajuste do modelo de regressão. Neste ponto assume-se que os pressupostos de normalidade, variância constante e independência são satisfeitos, como demosntrado pelos testes realizados. A ideia central para testar a linearidade é decompor SQRes em duas partes: erro puro e falta de ajuste que vão contribuir para a definição da estatística de teste F. Realizado o teste obteve-se um valore de p-valor igual a 0,289, o que demanda a rejeição da hipótese que há uma relação linear entre as variáveis.

Teste para avaliação da independência dos resíduos

Tendo em vista, o resultado obtido no teste anterior esse teste pode esclarecer ainda mais o ajuste do modelo. O teste para avaliação da independência dos resíduos é utilizado para detectar a presença de autocorrelação provenientes de análise de regressão. Realizando o teste obteve-se um valor de p-valor aproximadadente igual a 0, indicando que se deve rejeitar a hipotese que não existe correlação serial entre os dados, com uma confiança de 95%.

PARTE 2 - Análise de Regressão Linear Múltipla

Introdução

Nesta parte o conjunto de dados será avaliado de forma múltipla incluindo ao modelo anterior uma variável categórica ao modelo, a fim de selecionar um melhor modelo com o auxílio de uma variável categórica.

Objetivo

Ajustar um modelo de regressão linear múltiplo utilizando técnicas de retas coincidentes e paralélas a fim de selecionar a melhor explicação para os valores médios dos imóveis de Boston.

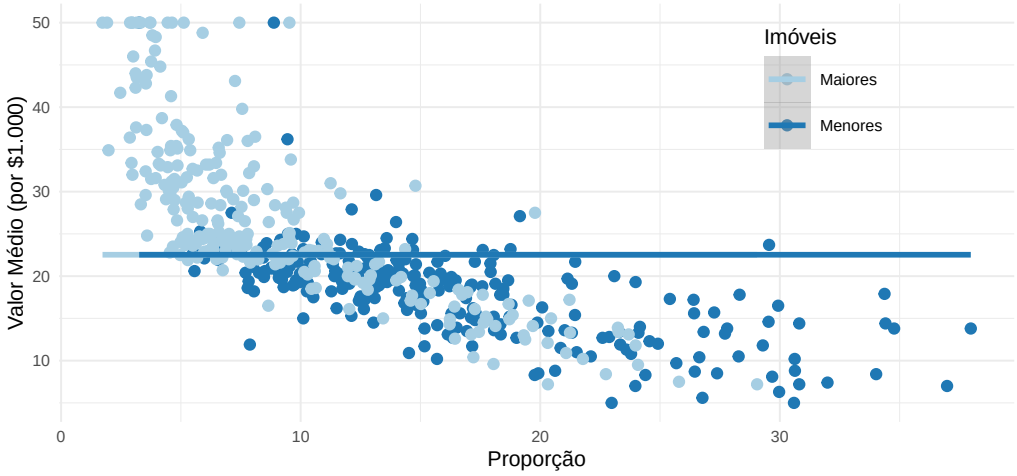
Criação de uma variável categórica

Embora o banco de dados estudado apresente duas variáveis categóricas: a CHAS, que indica a se o imóvel se encontra ou não as margens do rio Charles e RAD, que representa o índice de acessibilidade às rodovias radiais, verificou-se que a correlação destas variáveis com o preço médio dos imóveis é muito pequena, como se pode ver na Figura 4. Assim pecebeu-se que o tamanho dos imóveis baseado no seu número de cômodos poderia melhor explicar o preço médio destes imóveis. Desta forma, com base na tabela 1, verificou-se que a mediana do número médio de cômodos dos imóveis poderia ser usada para criar dois grupos: "Imóveis Pequenos" e "Imóveis Maiores".

Distribuição do valor médio dos imóveis

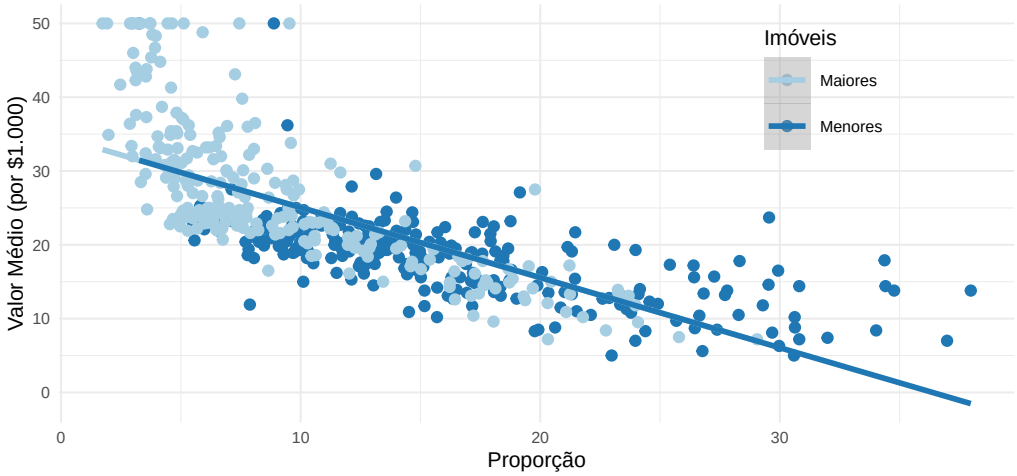
Com base na distribuição população de baixa renda e do tamanho dos imóveis segundo a classificação proposta foi avaliado o valor médio dos imóveis conforme as figuras a seguir.

Figura 6: Distribuição entre o valor médio dos imóveis e a população de baixa renda em função do tamanho do imóvel. Modelo Nulo



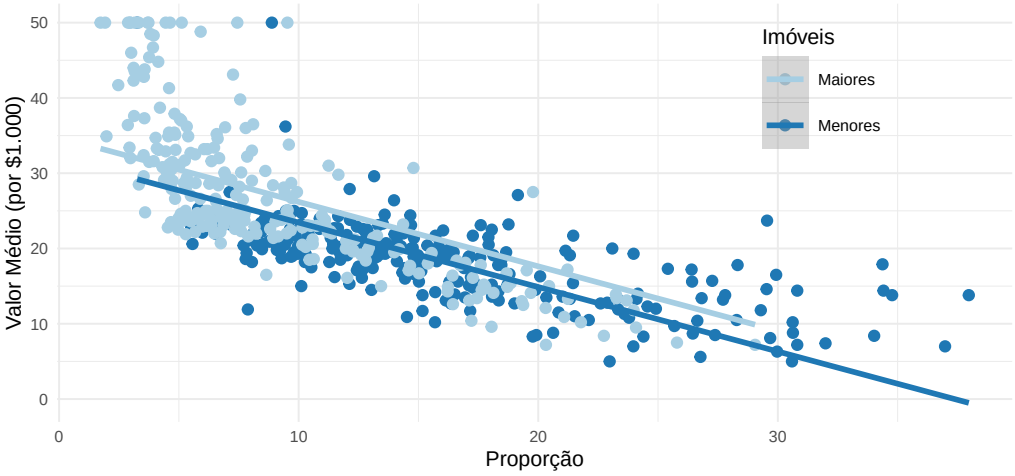
Utilizando-se o Modelo Nulo, verifica-se pela Figura 6 que há uma diferença expressiva entre os dois sub-conjuntos de dados (classificados segundo o tamanho do imóvel), indicando que há um efeito expressivo da variável categórica no modelo proposto.

Figura 7: Distribuição entre o valor médio dos imóveis e a população de baixa renda em função do tamanho do imóvel. Modelo Retas Coincidentes



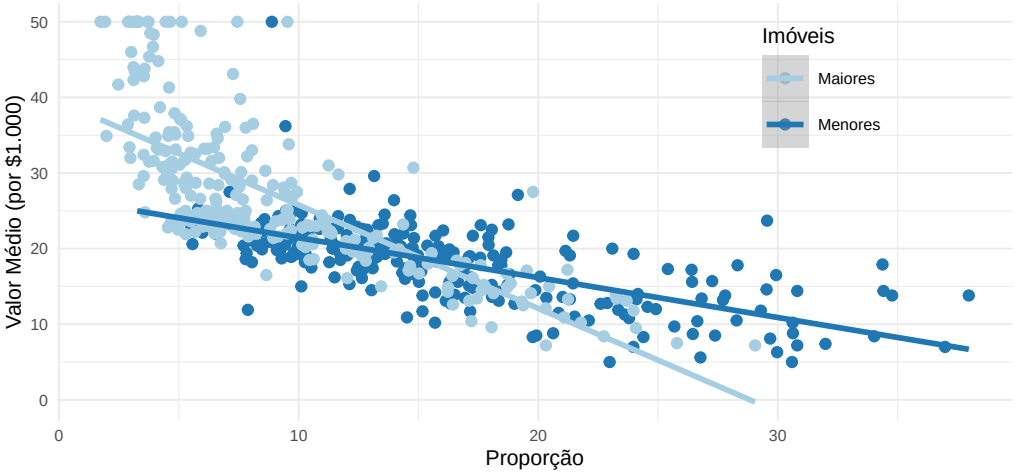
No caso da utilização do modelo com retas coincidentes, não se pode pela análise visual definir o efeito dos subconjuntos de dados no modelo proposto, como se vê na Figura 7, uma vez que esse modelo não leva em conta a variável categórica e a reta obtida é baseada nos valores de todo o conjunto.

Figura 8: Distribuição entre o valor médio dos imóveis e a população de baixa renda em função do tamanho do imóvel. Modelo Retas Paralelas



Na Figura 8, há uma clara distinção entre os subconjuntos baseados no tamanho do imóvel, porém este modelo de retas paralelas apresenta retas ainda muito próximas, denotando que o efeito aditivo da classificação é pouco relevante, mesmo que pela análise visual os subconjuntos apresentem características bastante disitintas.

Figura 9: Distribuição entre o valor médio dos imóveis e a população de baixa renda em função do tamanho do imóvel. Modelo Retas Concorrentes



Na Figura 9, observa-se claramente o efeito da interação entre as variáveis. A sensível diferença entre as inclinações das retas propostas pelo modelo deixam claro os efeitos multiplicativos da interação entre as variáveis. Pode-se verifiar assim que, pela análise gráfica, o modelo de retas concorrentes foi o mais adequado para descrever o comportamento

das variáveis selecionadas. Esta observação será verificada sob a ótica da análise dos testes F para os modelos encaixados.

Análise dos testes F para os modelos encaixados

Após o ajuste dos modelos encaixados existe a necessidade de se avaliar a significância dos mesmos, o teste de hipótese ANOVA para tal situação será realizado. A Tabelas 6 traz os principais resultados da tabela ANOVA para estes modelos, possibilitando assim inferir sobre o modelo mais bem ajustado.

Tabela 6: Análise de Variância (ANOVA) dos modelos encaixados.

Modelo	Est. F	p-valor
Nulo	1	0.5
Retas Coincidentes	601.613	<0.001
Retas Paralélas	321.795	<0.001
Retas Concorrentes	295.821	<0.001

```
[1] 0.6365505
```

Com base nos resultados apresentados na Tabela 6, verifica-se que não é possível definir o melhor modelo apenas pelo teste F aplicado. Nota-se qua a exceção do modelo nulo, todos os demais são, com base neste teste, apropriados para descrever a variável preço médio dos imóveis de Boston com base nas variáveis população de baixa renda e tamanho dos imóveis. Dentre estes, valor da estatística F foi mínimo para o modelo de retas concorrentes que conforme a análise gráfica foi o mais adequado e será selecionado para avaliação diagnóstica do ajuste obtido. Neste modelo o valor do Coeficiente de Determinação de Pearson obtido foi de $R^2 = 0.639$ e do valor do Coeficiente de Determinação de Pearson Ajustado obtido foi de $R^2_{aju} = 0.637$.

Diagnóstico do ajuste

Testes de Diagnósticos do Modelo

Para avaliar se o modelo atende aos pressupostos, além da análise gráfica podem ser realizados testes de diagnósticos, que são testes de hipóteses para avaliação dos pressupostos que são:

- Normalidade;

  H_0 : Os resíduos possuem normalidade.

  H_1 : Os resíduos **não** possuem normalidade.

- Homoscedasticidade (Homogeneidade de Variância);

  H_0 : Os resíduos possuem variância constante.

  H_1 : Os resíduos **não** possuem variância constante.

- Linearidade;
- Independência.

  H_0: Existe correlação serial entre os resíduos.

  H_1: **Não** existe correlação serial entre os resíduos.

Para tanto serão utilizados os seguintes testes:

- Shapiro-Wilk, para avaliar a Normalidade;
- Breush-Pagan, para avaliar a Homoscedasticidade;
- Durbin-Watson, para avaliar a Independência.

Tabela 7: Testes de Diagnósticos dos Resíduos

	Estatística de teste	p-valor
Shapiro-Wilk	0.9152	<0,0001
Breush-Pagan	88.6279	<0,0001
Durbin-Watson	0.9602	<0,0001

A Tabela 7 traz os testes de diagnósticos realizados para avaliar o modelo de regressão ajustado. Verifica-se que a hipótese de nula da homocedasticidade deve ser rejeitada com um nível de significância de 5%, uma vez que o teste de Breush-Pagam obteve um p-valor menor que 0.05. A normalidade da distribuição também foi rejeitada como indica o p-valor do teste de Shapiro-Wilk. Nota-se ainda que há dependência entre as características confirmado pelo p-valor do teste de Durbin-Watson. Este codependência era esperada uma vez que os tamanhos dos imóveis tem relação com o nível de renda da população. O que pode ser conferdo na matriz de correlação a seguir.

Correlação entre as variáveis do modelo

A correlação entre as variáveis do modelo pode ser medida pelo coeficiente de correlação entre elas.

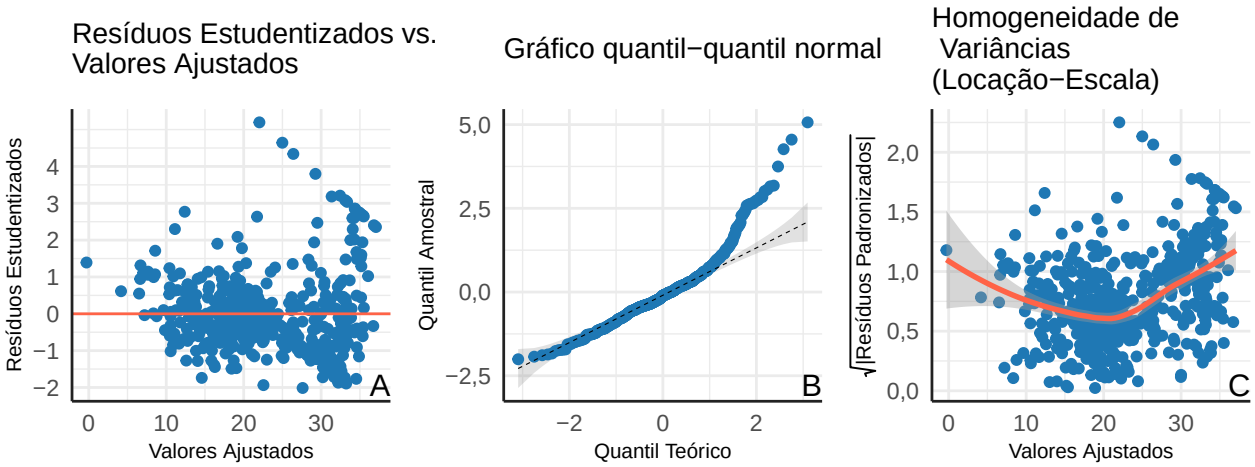
Tabela 8: Matriz de correlação das variáveis do modelo

	População de Baixa Renda	Valor Médio	Tamanho do Imóvel
População de Baixa Renda	1,000	-0,738	-0,486
Valor Médio	-0,738	1,000	0,473
Tamanho do Imóvel	-0,486	0,473	1,000

Nota-se que as variaveis tamanhos dos imóveis tem relação com o nível de renda da população em um nível similar ao com a variável resposta. Esta situação compromete a qualidade do modelo utilizado.

Análise de Resíduos

Figura 10: Análise de resíduos do modelo ajustado

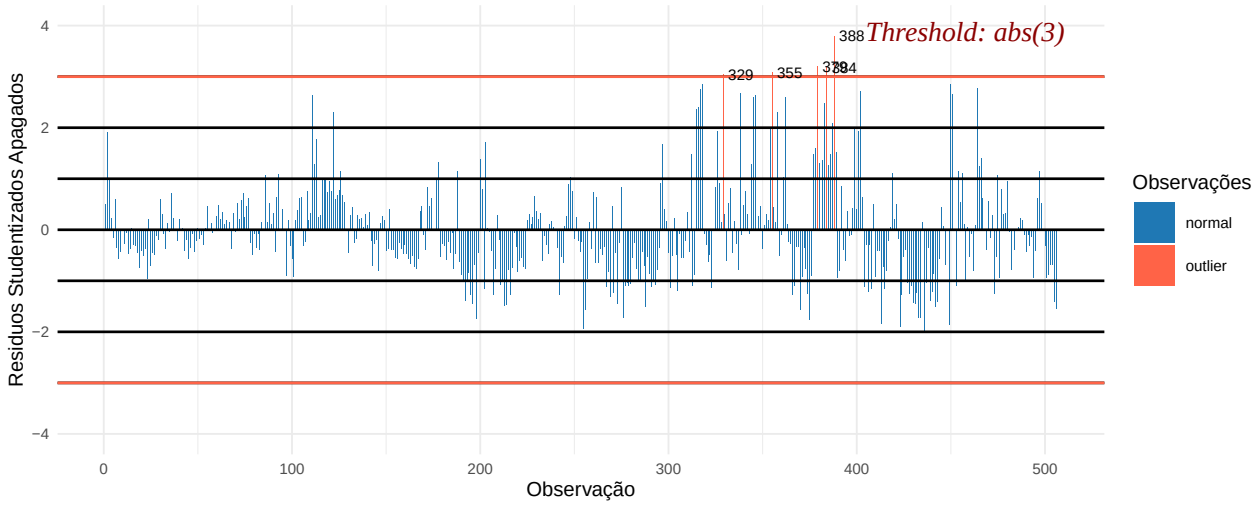


A Figura 10A apresenta um comportamento assimétrico dos resíduos, podendo ser constatado uma pequena variabilidade inicial e um aumento desta à medida que os valores ajustados aumentam, caracterizando uma maior heterocedasticidade. A Figura 10C, que trata da Homogeneidade de Variâncias (Locação-Escala) ressalta que há um problema na variabilidade dos dados, ampliando a interpretação feita na análise da Figura 10A, de que há uma mudança na variabilidade dos dados, caracterizando assim uma heterocedasticidade dos dados. A Figura 10B traz o gráfico para avaliação da normalidade dos dados e mostra que os dados estão bastante afastados da reta de referência, especialmente nas caudas da distribuição onde fogem inclusive da região pertencente ao Intervalo de Confiança - IC, podendo-se assumir que não há normalidade.

Gráficos de Diagnóstico

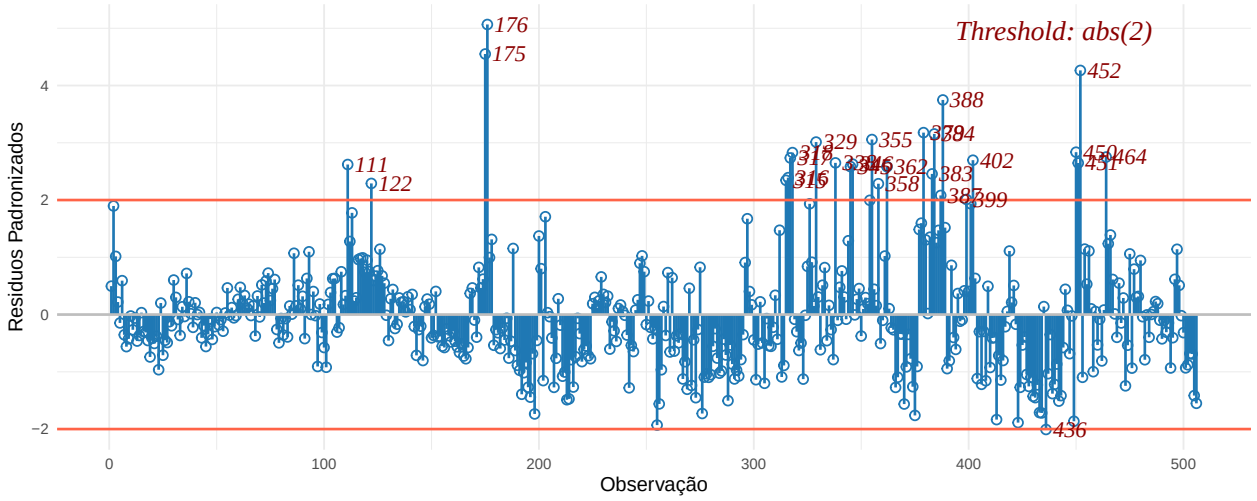
A análise dos gráficos de diagnóstico permite avaliar as observações realizadas e conhecer a influência de cada uma delas para o modelo de regressão proposto. Assim, com base no modelo, é possível fazer as seguintes análises:

Figura 11: Valores Ajustados e Resíduos Studentizados



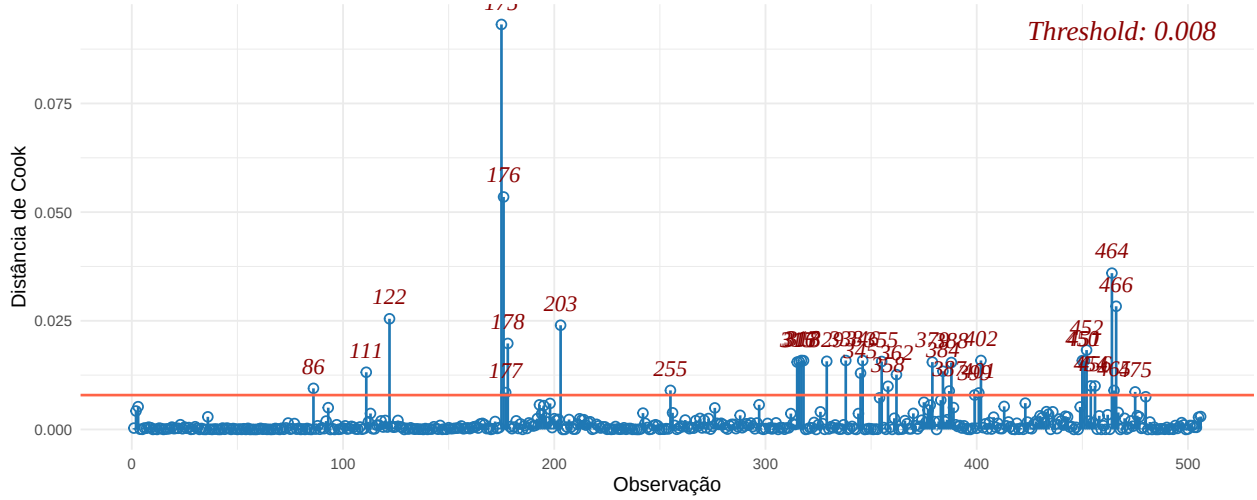
A Figura 11 demonstra que os resíduos em sua grande maioria dentro dos limites esperados, com exceção de poucas observações que ultrapassam o limite superior Não parece ser o caso de nenhuma intervenção por conta deste valor.

Figura 12: Valores Ajustados e Resíduos Padronizados.



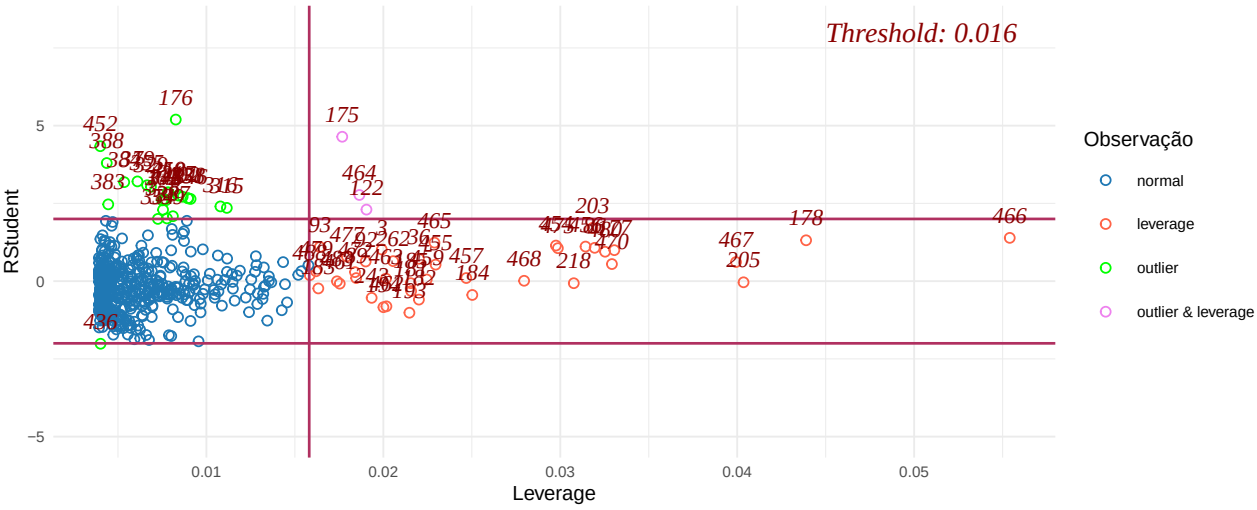
Na análise da Figura 12, onde os resíduos foram padronizados, verifica-se queo número de observações além do limite de aceitação aumentou consideravelmente, inclusive o limite inferior tanbém foi ultrapassado.

Figura 13: Distância de Cook.



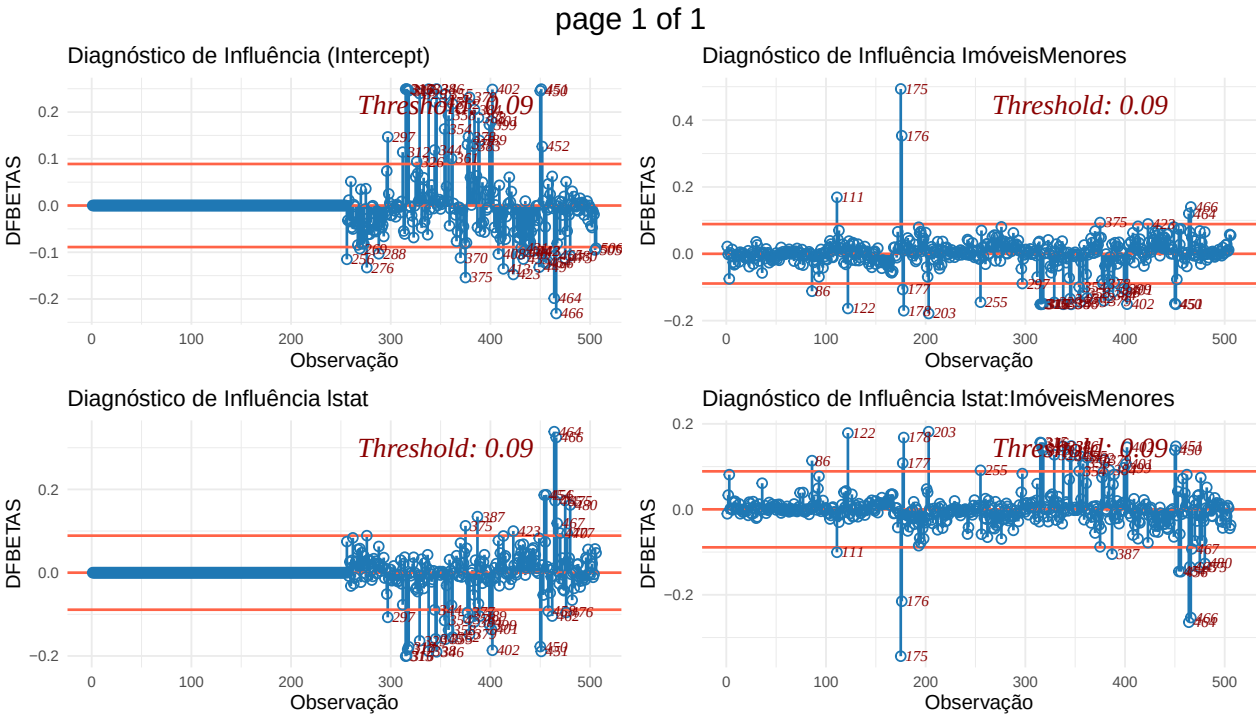
A analise da distância de Cook apresentada na Figura 13 demonstra mais uma vez que muitas observações destoam do conjunto, alguns com distância muito expressiva.

Figura 14: Análise dos pontos de Alavanca e Resíduo Studentizado.



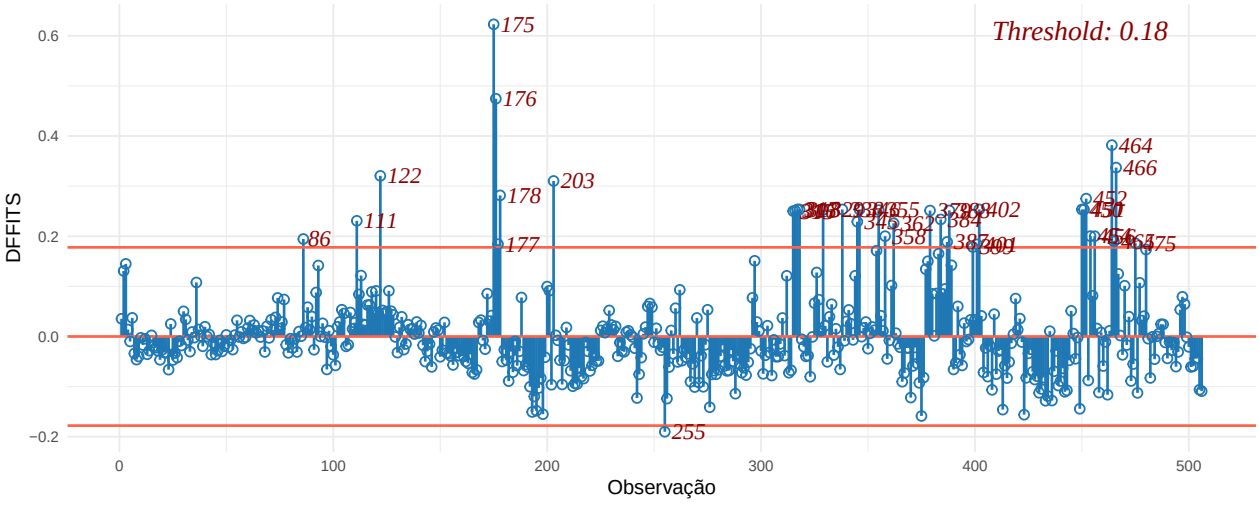
Pela Figura 14, observamos que diversas observações que podem ser consideradas como *Outliers* e como pontos de alavanca e algumas com as duas características sendo possívelmente as observações com maior influência negativa ao modelo.

Figura 15: DFBetas para as variáveis do modelo.



A Figura 15 apresenta os DFBetas para cada uma das variáveis utilizadas no modelo. Nota-se que as observações já identificadas como anômalas pelos gráficos anteriores se repetem com maior frequência na Figura 15.

Figura 16: DfFit para as variáveis do modelo.



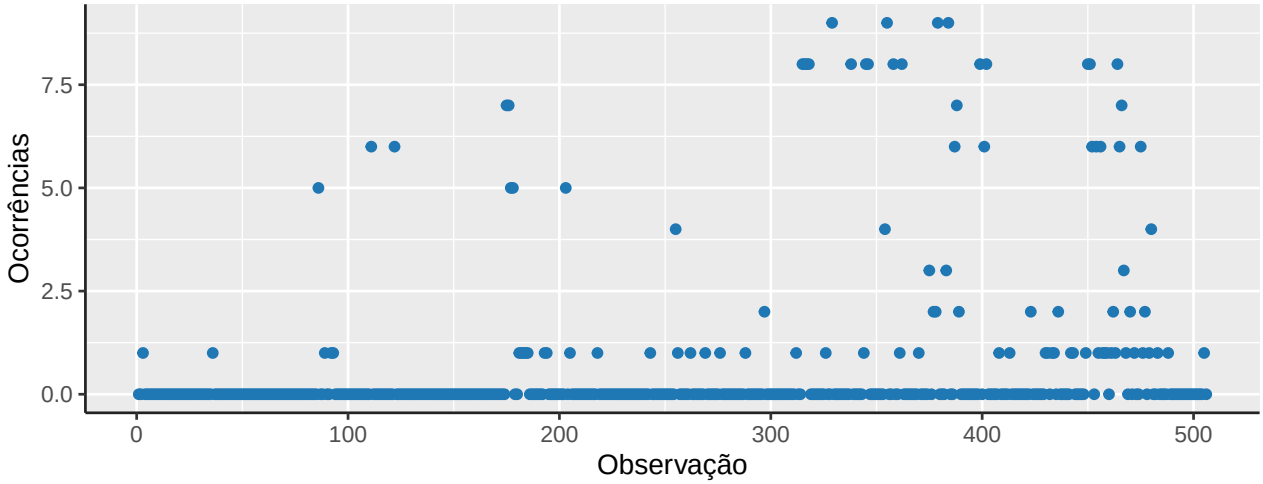
A Figura 16 acompanha os gráficos anteriores apresentando mais uma vez muitas observações discrepante.

Este comportamento com muitas observações discrepantes deve-se ao pressupostos da regressão linear múltipla terem sido violados como mostraram a Figura 10 e a Tabela 8.

Eliminação de observações anômalas

A fim de melhorar o ajuste do modelo, identificando ss observações que apresentaram comportamento anômalo nos diagnósticos dos valores ajustados e resíduos studentizados, valores ajustados e resíduos padronizados, distância de CooK, pontos de alavanca e *outliers*, análise de DfFit e todas as análises de BFBetas, chegamos as frequências de observações anômalas apresentadas na Figura 17.

Figura 17: Número de ocorrências para cada observação



Considerando apenas as observações com 5 ou mais ocorrências temos a Tabela a seguir.

Tabela 9: Observações com maior número de ocorrências.

Observação	Ocorrências
86	5
111	6
122	6
175	7
176	7
177	5
178	5
203	5
315	8
316	8
317	8
318	8
329	9
338	8
345	8
346	8
355	9
358	8
362	8
379	9
384	9
387	6
388	7
399	8
401	6
402	8
450	8
451	8
452	6
454	6
456	6
464	8
465	6
466	7
475	6

Observou-se que alguns dos pontos elencados na Tabela 9, correspondem a *outliers* também verificados na análise descritiva dos dados (Figura 3). A não eliminação dos pontos

atípicos naquele momento era desaconselhável, entretanto após as análises realisadas há maior segurança no procedimento. Essas observaçãoes correspondem a 6,9% do total de observações do conjunto, desta forma entende-se que a eliminação delas não provocará uma perda significativa de informação para o modelo.
Desta forma, obteve-se o seguinte modelo resultante:

Y_i = 26.452 -0.539 X_{1i}+ 10.142 X_{2i} -0.697 $X_{1,2i}$

Onde:
Y_i - Valor médio dos imóveis de Boston;
X_{1i} - População de Baixa Renda;
X_{2i} - Tamanho do Imóvel;
$X_{1,2i}$ - Interação entre as variáveis explicativas $(X_{1i} * X_{2i})$;

Neste novo modelo o coeficiente de determinação calculado foi de $R^2 = 0.713$, o que denota que 71.3% da variância dos dados é explicada pelo modelo. O valor deste novo coeficiente permite concluir que a eliminação das observações com maior impacto e das variáveis pouco relevantes ao modelo foi benéfica. Pode-se calcular o coeficiente de determinação ajustado igual a $R^2_{aju} =$. Estes valores indicam que a eliminação das observações anômalas contribuiu signicativamente para a melhora do modelo proposto.

Conclusão

Parte 1

Da análise descritiva das variáveis deste banco de dados não se observa, situações impeditivas da proposta de modelamento por Regressão Linear Simples dos dados como forma de predizer o Valor Médio dos Imóveis de Boston. Mesmo a análise de valores atípicos contribui com essa possibilidade uma vez que os valores candidatos a valores atípicos na verdade compõem o rol de dados relevantes e que há enorme variedade em tipos, propósitos e status dos imóveis avaliados. Esses dados por sua vez, representam um maior desafio ao modelamento a que esse trabalho se propõe.
No teste da hipótese de correlação, todas as variáveis apresentaram significativa relação linear com o valor médio do imóvel, mesmo em casos que o coeficiente de determinação (R^2) se apresentou muito baixo.
A implementação de técnicas de Regressão Linear Simples - RLS, para a variável explicativa que aparenta melhor possibilidade de explicação do Preço Médio dos Imóveis - Pop. Classe Baixa, não se mostrou muito eficiente como observado pela análise gráfica dos resíduos. Para melhor compreensão desta análise foram feitos testes de normalidade, homoscedasticidade e de independência dos resíduos, de onde se concluiu que se deve rejeitar as hipóteses de normalidade, homoscedasticidade e de independência serial dos dados, confirmando assim o que a análise gráfica demonstrou.

Parte 2

A criação da variável Tamanho do Imóvel, dividindo os imóveis em dois grupos equivalentes, quanto ao número médio de cômodos, foi bastante útil para os propósitos desta segunda parte do relatório.

A análise gráfica para a determinação do modelo retas concorrentes foi mais útil que a realização do teste F, uma vez que os resultados obtidos foram os mesmos para todos os modelos análisados, a exceção do modelo nulo.

O ajuste do modelo aos pressupostos da Regressão Linear Múltipla não foi satisfatório e o proseguimento dos calculos visou apenas observar o comportamento do modelo. Entretanto, a eliminação de observações anômalas fez com que a qualidade do modelo proposto fosse significativamente melhorada.

Referências

- Harrison, David & Rubinfeld, Daniel. (1978). Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management. 5. 81-102. 10.1016/0095-0696(78)90006-2.
- Belsley, David A. & Kuh, Edwin. & Welsch, Roy E. (1980). Regression diagnostics: identifying influential data and sources of collinearity. New York: Wiley.