

Banco de dados: Boston House Prices

Fernado Bispo, Jeff Caponero

Sumário

| | |
|---|-----------|
| Introdução | 2 |
| Metodologia | 3 |
| Sobre o conjunto de dados | 3 |
| Variáveis a serem analisadas | 3 |
| Variável de Saída (Resposta): | 5 |
| Fonte | 5 |
| Resultados | 6 |
| Análise Descritiva | 6 |
| Análise de Dados Atípicos | 8 |
| Relação entre as variáveis | 10 |
| Análise de Resíduos | 13 |
| Testes de diagnóstico | 13 |
| Conclusão | 15 |
| Referências | 16 |

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 de preços de 506 casas em Boston, publicados em Harrison, D. and Rubinfeld, D.L. *'Hedonic prices and the demand for clean air'*, J. Environ. Economics & Management, vol.5, 81-102, 1978.

Usado em Belsley, Kuh & Welsch, *'Regression diagnostics: identifying influential data and sources of collinearity'*. New York: Wiley 1980. Os dados podem ser acessados na plataforma para aprendizado de ciência de dados [Kaggle](https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data) através do link:

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

Variáveis a serem analisadas

A amostra contém 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 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). Será representada como **Margem**.
5. NOX: Concentração de óxidos nítricos em pphm (partes por 100 milhões). Será representada como **Índice Óxido 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 logarítmica. Será representada por **Nº de Cômodos**.
7. AGE: Proporção de unidades próprias construídas antes de 1940. A idade da unidade geralmente está relacionada à qualidade da estrutura. 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 locais de estradas. O RAD captura outros tipos de vantagens locais 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 [k\$].

Fonte

StatLib - Carnegie Mellon University

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.

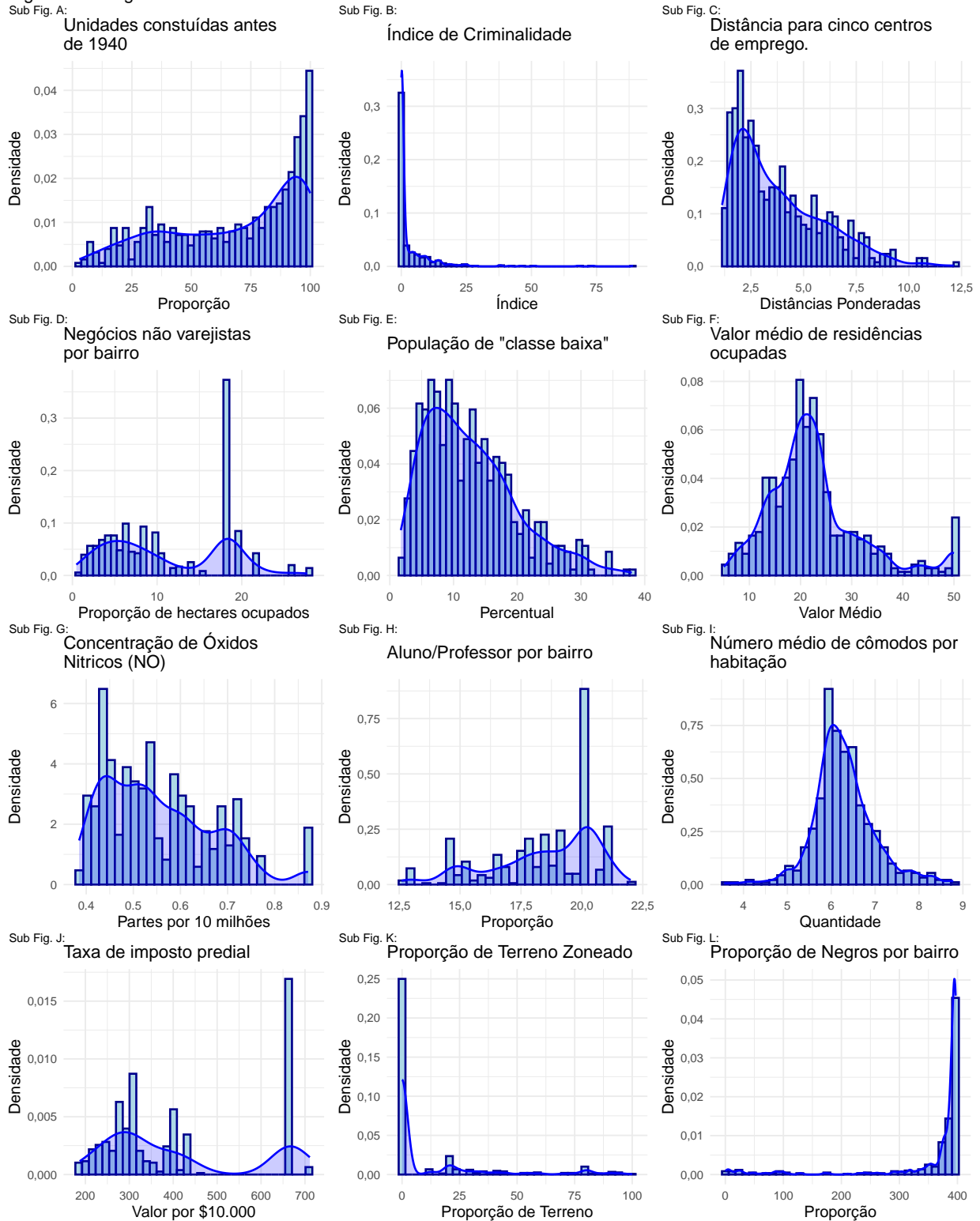
Tabela 1: Medidas Resumo dos dados

| | 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 |
| Nº 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 |

Fonte: StatLib - Carnegie Mellon University

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

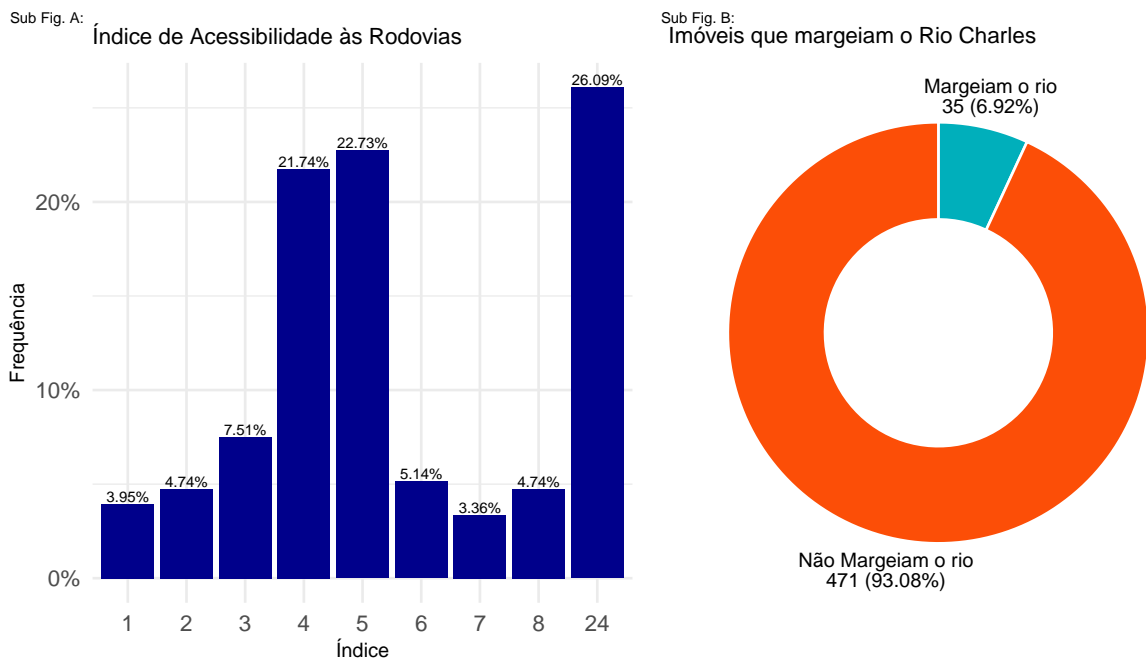
Figura 1: Histogramas das variáveis em análise.



Desta análise inicial, destancam-se algumas características:

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

Figura 2: Distribuição de Frequência das Variáveis Categóricas.

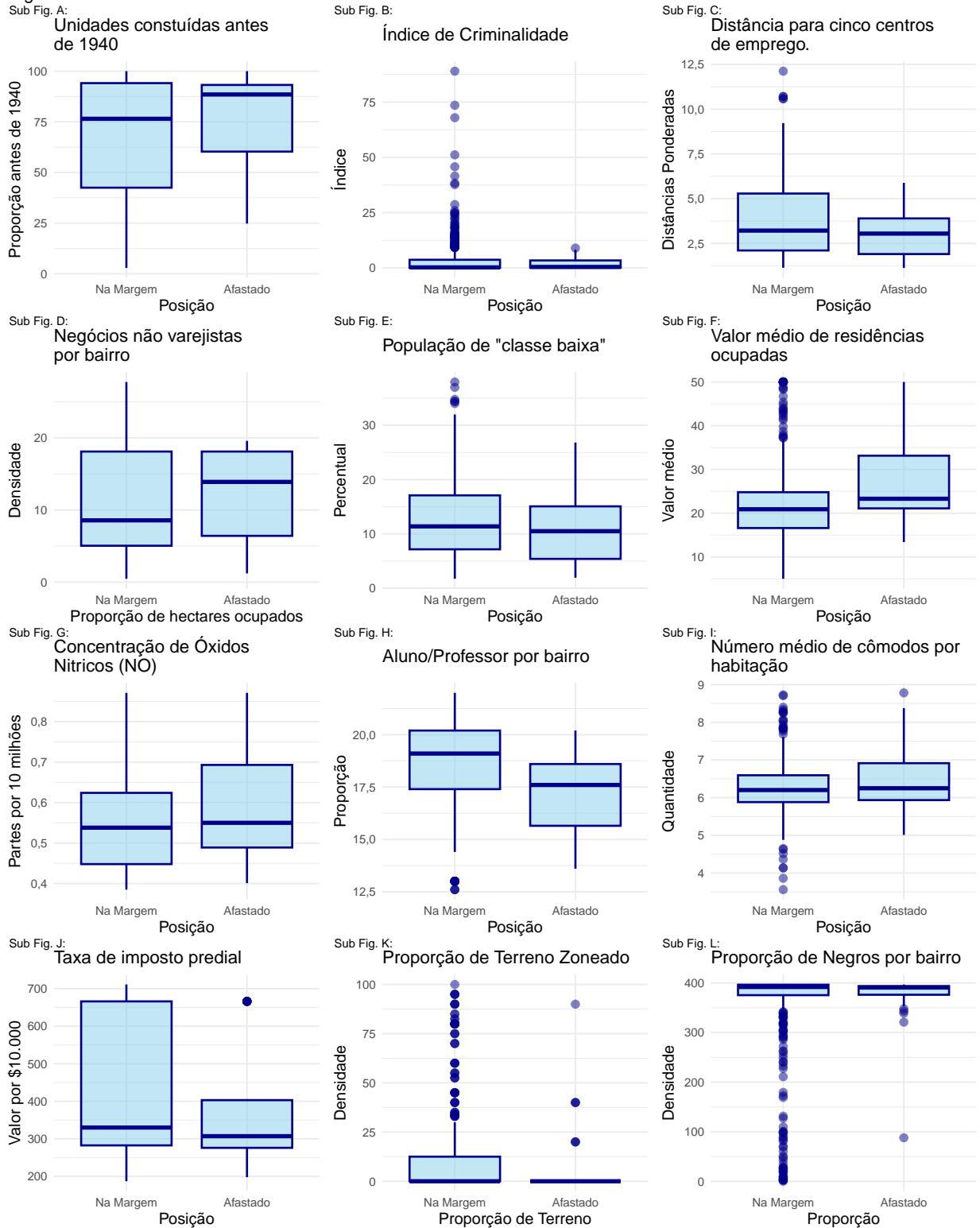


Fonte: StatLib – Carnegie Mellon University

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

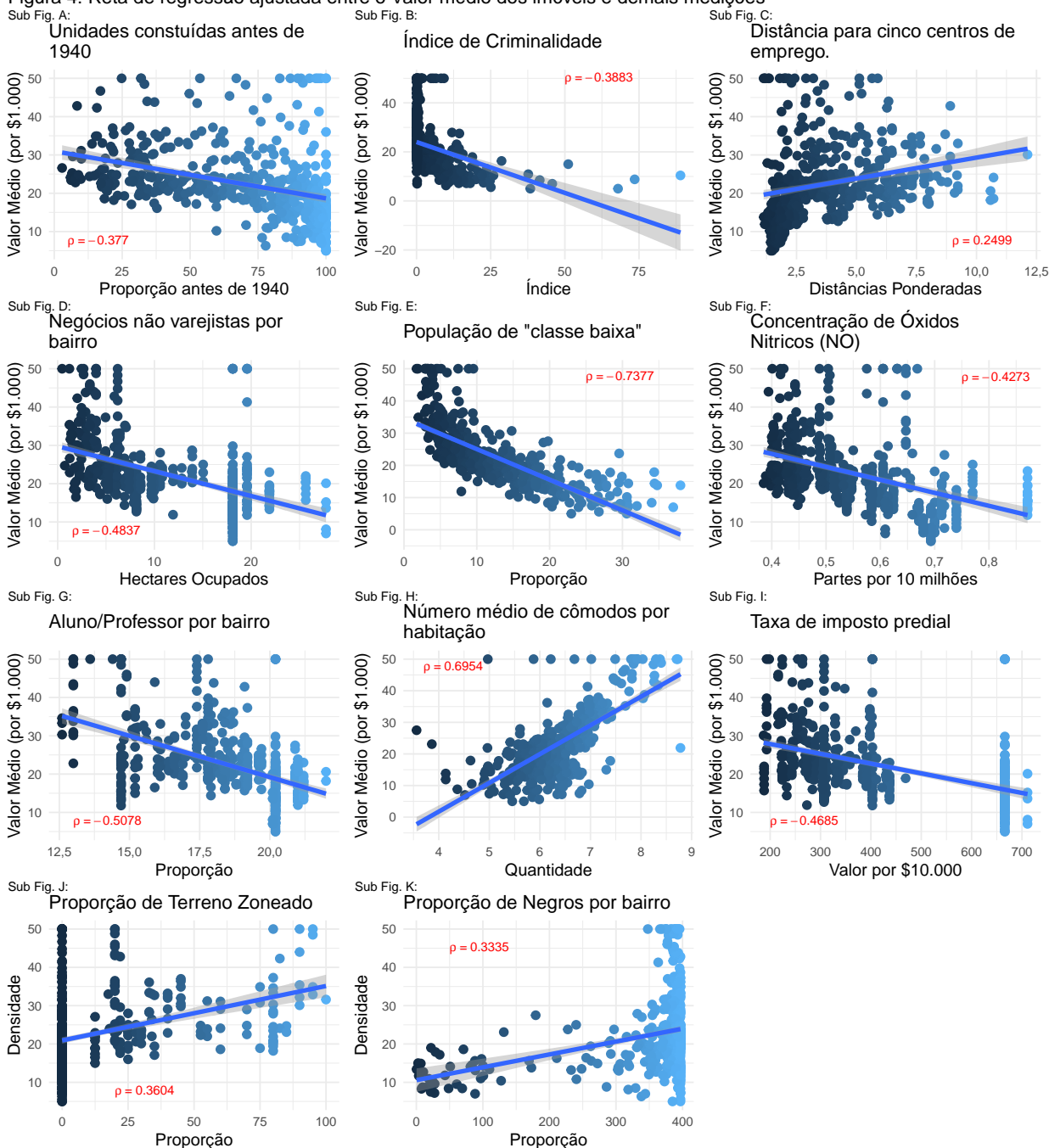
Pode-se verificar pela Figura 2 que praticamente todas as variáveis apresentam *outliers*,

entretanto o tratamento destes *outlayers* em todos os casos parece, salvo melhor juízo ser o mesmo. Observa-se que há coerência entre eles, isto é, são realizadas possíveis e não devem ser desprezados 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: Retas de regressão ajustada entre o Valor médio dos imóveis e demais medições



Fonte: StatLib – Carnegie Mellon University
Valor do Coeficiente de Correlação de Pearson em vermelho

Na avaliação da Figura 3, observa-se que nenhuma das variáveis tem uma correlação forte com o valor médio dos imóveis. A Tabela 2, apresenta os valores calculados de $\hat{\beta}_0$ e $\hat{\beta}_1$ que estimam os valores do modelo $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ como 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 Correlação ($\hat{\rho}$) também foi calculado.

Após o ajuste dos modelos exibidos na Figura 4, foi elaborada a Tabela 2 com os respectivos valores das regressões calculadas entre valor médio dos imóveis e demais medições.

Tabela 2: Valores dos modelos de regressão linear simples.

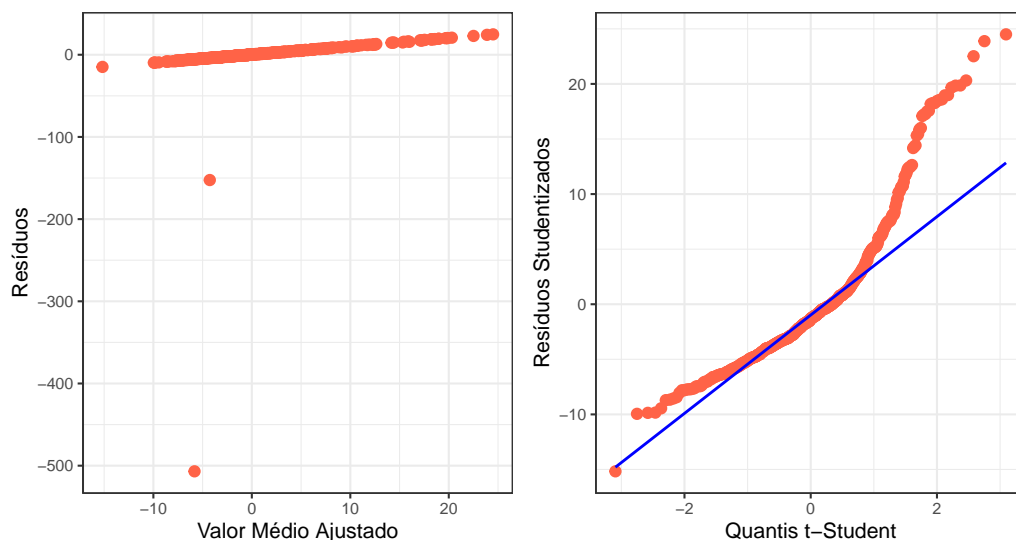
| | β_0 | σ_0 | β_1 | σ_1 | p-valor | $\hat{\rho}$ |
|-----------------------|-----------|------------|-----------|------------|---------|--------------|
| Índice Criminalidade | 24,033 | 0,409 | -0,415 | 0,044 | 0 | 0,151 |
| Área Industrial | 29,755 | 0,683 | -0,648 | 0,052 | 0 | 0,234 |
| Índice Oxido Nítrico | 41,346 | 1,811 | -33,916 | 3,196 | 0 | 0,183 |
| Nº Cômodos | -34,671 | 2,650 | 9,102 | 0,419 | 0 | 0,484 |
| Idade do Imóvel | 30,979 | 0,999 | -0,123 | 0,013 | 0 | 0,142 |
| Dist. Empregos | 18,390 | 0,817 | 1,092 | 0,188 | 0 | 0,062 |
| Imposto Propriedade | 32,971 | 0,948 | -0,026 | 0,002 | 0 | 0,220 |
| Prop. Prof.-Aluno | 62,345 | 3,029 | -2,157 | 0,163 | 0 | 0,258 |
| Pop. Classe Baixa | 34,554 | 0,563 | -0,950 | 0,039 | 0 | 0,544 |
| Prop. Terreno Zoneado | 20,918 | 0,425 | 0,142 | 0,016 | 0 | 0,130 |
| Prop. Negros/bairro | 10,551 | 1,557 | 0,034 | 0,004 | 0 | 0,111 |

Fonte: StatLib - Carnegie Mellon University

Nota-se da Tabela 2, que todas as tentativas de apresentar um modelo de regressão linear para os dados se mostraram infrutíferas. Os dados da forma como apresentados não comportam a simples regressão linear. Observe que para todas os pares de variáveis apresentadas o p-valor do modelo foi de aproximadamente 0,000 e os valores absolutos do coeficiente de correlação abaixo de 0,74. Desta forma, é necessário uma análise dos resíduos dos modelos para fundamentar a aplicação de alguma técnica de tratamento dos dados a fim de adequá-los a um modelo de regressão linear mais adequado.

Análise de Resíduos

Figura 5: Análise de resíduos do modelo de regressão da classe social com valor dos imóveis.



Testes de diagnóstico

Pode-se ainda utilizar um conjunto de testes de diagnóstico para confirmar este novo teste de significância. Como:

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

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 0.058, 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 hipótese 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 forma 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 demonstrado 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 valor 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 aproximadamente igual a 0, indicando que se deve rejeitar a hipótese que não existe correlação serial entre os dados, com uma confiança de 95%.

Conclusão

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.