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 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 através do link:

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

Variáveis a serem analisadas

As amostras contêm 13 atributos de casas em diferentes locais nos subúrbios de Boston no final dos anos 1970. Os objetivos são os valores médios das casas em um local (em k$). As variáveis são descritas a seguir:

1. CRIM: Índice de criminalidade per capita por bairro.
2. ZN: Proporção de terreno residencial zoneado para lotes acima de 25.000 sq.ft.
3. INDUS: Proporção de hectares de negócios não varejistas por bairro.
4. CHAS: Variável fictícia que representa imóveis próximos a margem do rio Charles (1 se o trecho margeia o rio; 0 caso contrário).
5. NOX: Concentração de óxidos nítricos (partes por 10 milhões).
6. RM: Número médio de cômodos por habitação.
7. AGE: Proporção de unidades ocupadas pelo proprietário 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 $10.000.
11. PTRATIO: Proporção aluno-professor por bairro.
12. B: O resultado da equação $B = 1000(Bk - 0,63)^2$ onde Bk é a proporção de negros por bairro.
13. LSTAT: % da população de "classe baixa".

Sobre a interpretabilidade de algumas variáveis

Existem variáveis que em virtude da sua natureza e da falta de um dicionário, inviabilizam a interpretação. Ressalto também que o conjunto de dados tem um problema ético, os autores deste conjunto de dados incluíram uma variável, "**B**", que pode parecer assumir que a auto-segregação racial influencia os preços das casas e portanto será excluída das análises, sendo estas:

- ZN: Proporção de terreno residencial zoneada para lotes acima de 25.000 sq.ft.
- RAD: Índice de acessibilidade às rodovias radiais.
- B: O resultado da equação $B = 1000(Bk - 0,63)^2$ onde Bk é a proporção de negros por bairro.

Resultados

Análise Descritiva dos Dados

Iniciando a análise das variáveis contidas no conjunto de dados foi construída a Tabela 1 com a sumarização das variáveis em análise, a fim de identificar o comportamento das variáveis por meio das métricas de posição e dispersão.

Tabela 1: Medidas Resumo dos dados

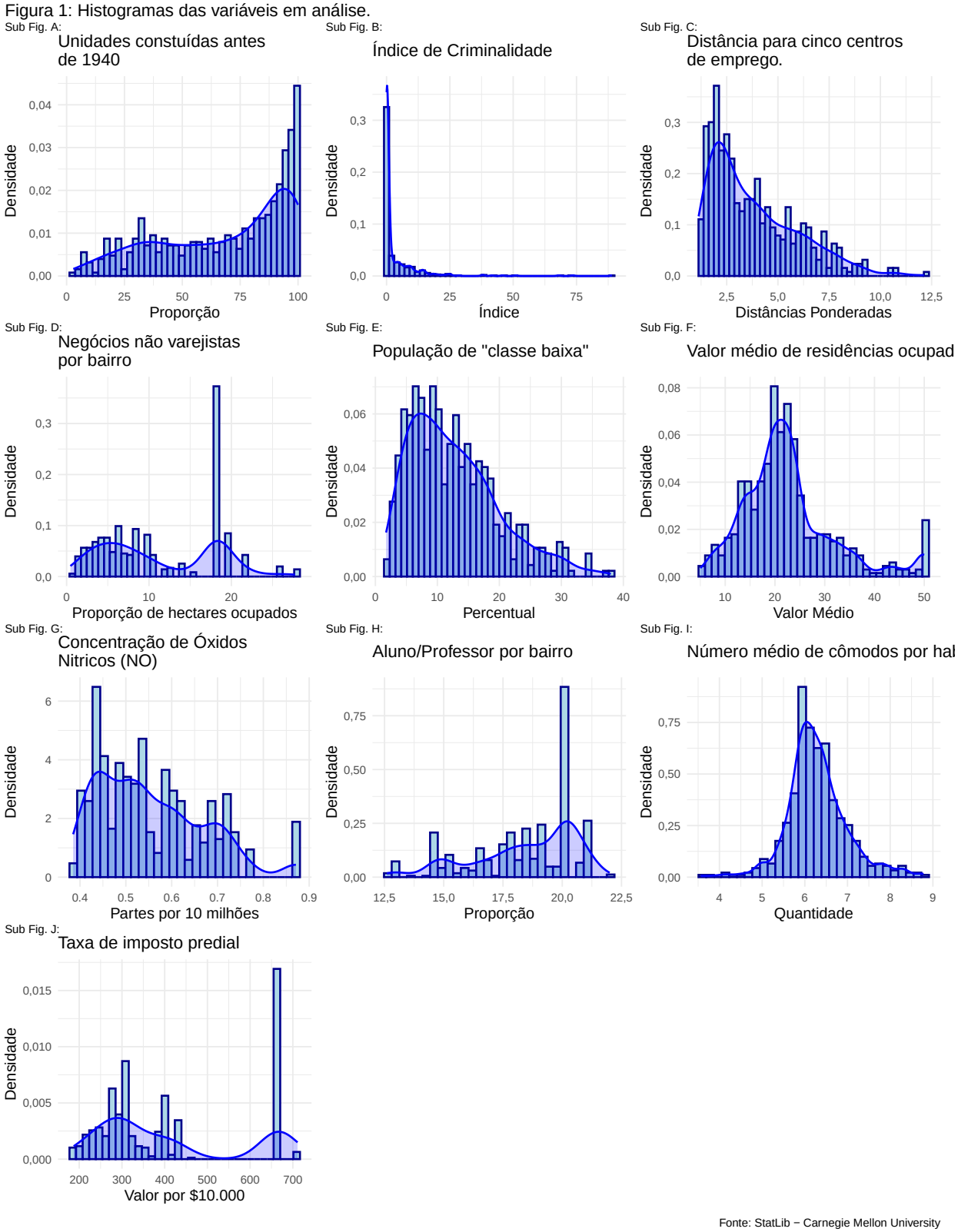
	Mín	Q1	Med	Média	Q3	Máx	Desv.padrão	CV	Assimetria	Curtose
AGE	2,90	45,00	77,50	68,57	94,10	100,00	28,15	0,41	-0,60	-0,98
CRIM	0,01	0,08	0,26	3,61	3,68	88,98	8,60	2,38	5,19	36,60
DIS	1,13	2,10	3,21	3,80	5,21	12,13	2,11	0,55	1,01	0,46
INDUS	0,46	5,19	9,69	11,14	18,10	27,74	6,86	0,62	0,29	-1,24
LSTAT	1,73	6,93	11,36	12,65	16,96	37,97	7,14	0,56	0,90	0,46
MEDV	5,00	17,00	21,20	22,53	25,00	50,00	9,20	0,41	1,10	1,45
NOX	0,38	0,45	0,54	0,55	0,62	0,87	0,12	0,21	0,72	-0,09
PTRATIO	12,60	17,40	19,05	18,46	20,20	22,00	2,16	0,12	-0,80	-0,30
RM	3,56	5,88	6,21	6,28	6,62	8,78	0,70	0,11	0,40	1,84
TAX	187,00	279,00	330,00	408,24	666,00	711,00	168,54	0,41	0,67	-1,15

Fonte: StatLib - Carnegie Mellon University
Legenda:
AGE: Proporção de unidades próprias construídas antes de 1940.
CRIM: Índice de criminalidade per capita por bairro.
DIS: Distâncias ponderadas para cinco centros de emprego de Boston.
INDUS: Proporção de hectares de negócios não varejistas por bairro.
LSTAT: Percentual da população de "classe baixa".
MEDV: Valor médio de residências ocupadas pelo proprietário.
NOX: Concentração de óxidos nítricos (partes por 10 milhões).
PTRATIO: Proporção aluno-professor por bairro.
RM: Número médio de cômodos por habitação.
TAX: Valor total do imposto predial por $10.000.

Analisando a Tabela 1 é possível identificar a presença de pontos atípicos para a variável **crim**, sendo esta a variável que mais se destaca pelos indicadores obtidos, gerando assim uma grande variabilidade, confirmada pelo Coeficiente de Variação - CV, sendo esta medida que avalia a dispersão dos dados em relação a média, além de verificar um elevado Coeficiente de Assimetria Positiva, indicando que a maioria dos dados são menores que a média e um comportamento Leptocúrtico, com base no Coeficiente de Curtose, ou seja, os dados possuem um comportamento mais alongado.

A Figura 1 traz os Histogramas das variáveis a fim de tentar identificar o comportamento das variáveis e uma possível associação a algum modelo de distribuição

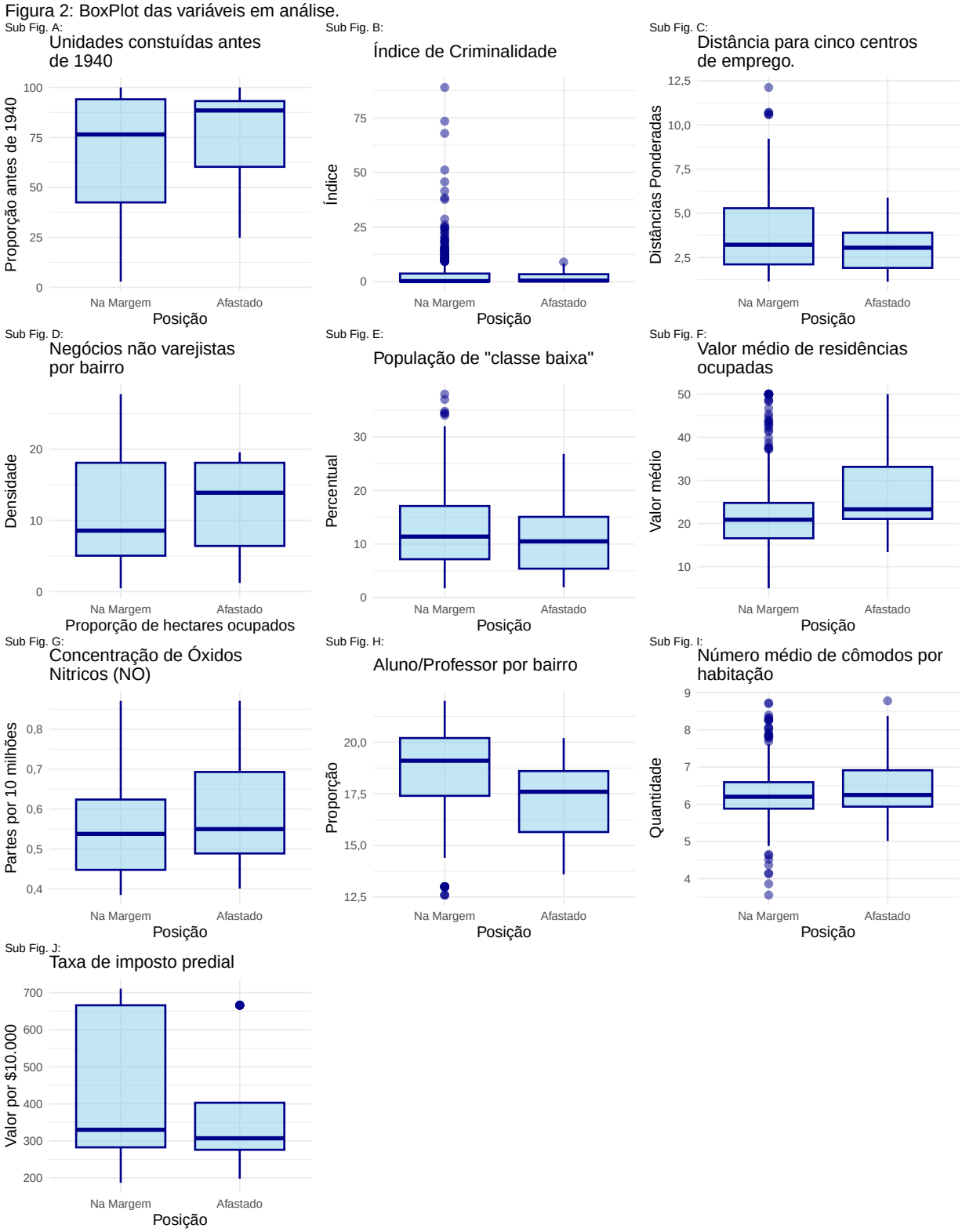
Figura 1: Histogramas das variáveis em análise.



Fonte: StatLib − Carnegie Mellon University

A Figura 2 traz os Gráficos de Caixa (*BoxPlot*) das variáveis em análise com o intuito de identificar a presença de pontos influentes, bem como a variabilidade das variáveis por meio da análise visual.

Figura 2: BoxPlot das variáveis em análise.



Fonte: StatLib − Carnegie Mellon University

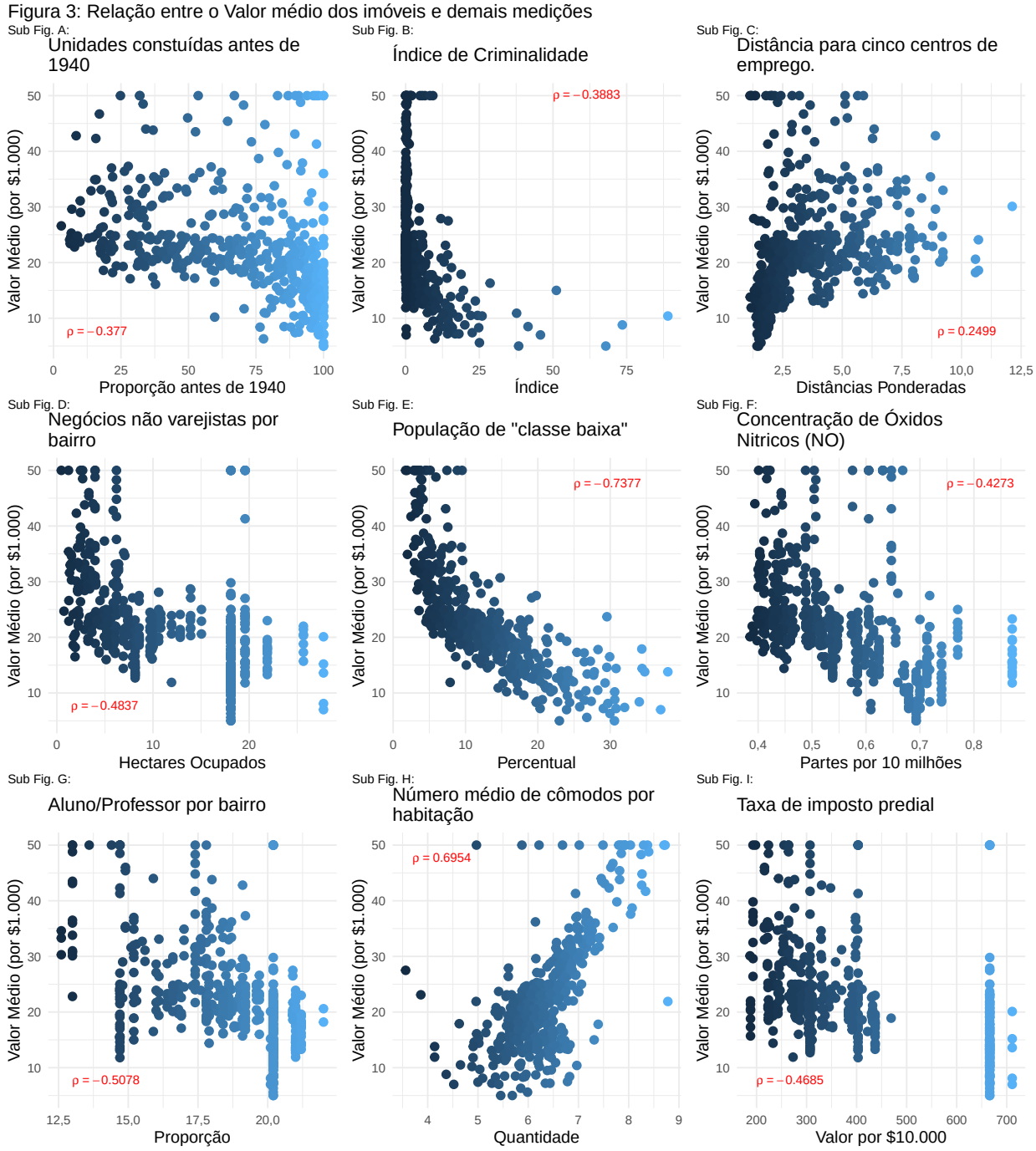
Corroborando com a identificação feita com base na Tabela 1, a variável que avalia o **Índice de Criminalidade** possui uma grande quantidade de valores atípicos principalmente para a região próxima a

margem do rio Charles.

Análise Correlacional

A Figura 3 traz os gráficos de dispersão das variáveis, tendo como variável resposta o Valor Médio dos Imóveis por $1.000 e as demais variáveis como explicativas, a fim de se identificar a possível existência de correlação entre elas, sendo calculado e descrito o Coeficiente de Correlação de Pearson ($\hat{\rho}$) de cada gráfico, a fim de se ter numericamente a força da relação entre a variável resposta com as demais variáveis do conjunto de dados.

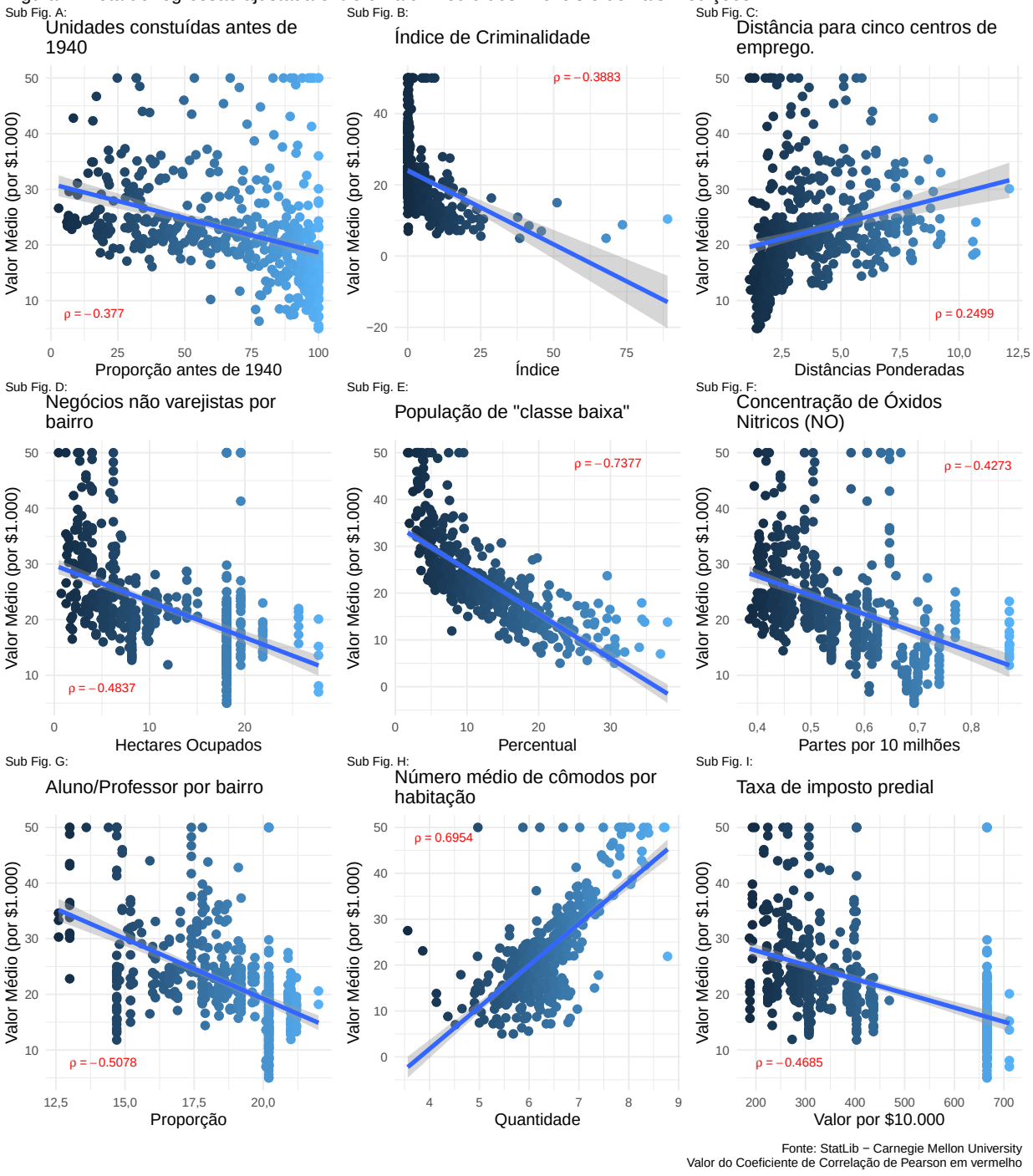
Figura 3: Relação entre o Valor médio dos imóveis e demais medições



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

Após análise da Figura 3 é constatado tanto pelo comportamento dos dados quanto pelo valor do coeficiente de correlação que as variáveis que apresentam maior correlação são as variáveis que representam o **Percentual da população de "classe baixa"** e o **Número médio de cômodos por habitação**, ainda assim segue a Figura 4 com a Reta de Regressão para todas as variáveis.

Figura 4: Reta 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

Analisando a Figura 4 é possível identificar que o **Valor médio dos imóveis** se concentra em regiões em que o **Percentual da população de "classe baixa"** (Sub.Fig E) é baixo, como esperado, demonstrando uma relação linear negativa com um Coeficiente de Correlação de Pearson ($\hat{\rho}$) de -0.7377 corroborando com

a existência da correlação negativa entre as variáveis confirmada através do teste de hipóteses.

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.

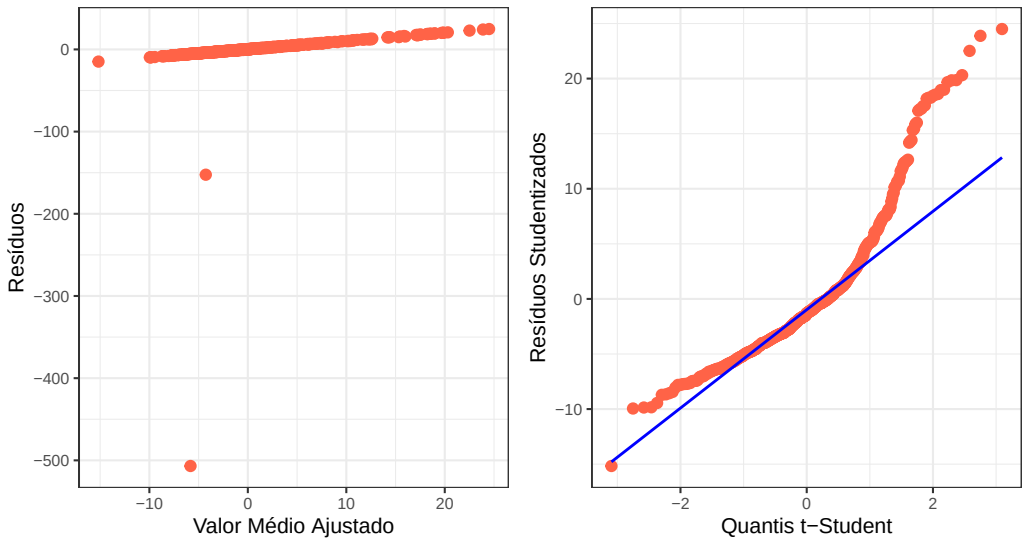
	CRIM	INDUS	NOX	RM	AGE	DIS	TAX	PTRATIO	LSTAT
β_0	24,033	29,755	41,346	-34,671	30,979	18,390	32,971	62,345	34,554
σ_0	0,409	0,683	1,811	2,650	0,999	0,817	0,948	3,029	0,563
β_1	-0,415	-0,648	-33,916	9,102	-0,123	1,092	-0,026	-2,157	-0,950
σ_1	0,044	0,052	3,196	0,419	0,013	0,188	0,002	0,163	0,039
p-valor	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
$\hat{\rho}$	0,151	0,234	0,183	0,484	0,142	0,062	0,220	0,258	0,544

Fonte: StatLib - Carnegie Mellon University

Legenda:
AGE: Proporção de unidades próprias construídas antes de 1940.
CRIM: Índice de criminalidade per capita por bairro.
DIS: Distâncias ponderadas para cinco centros de emprego de Boston.
INDUS: Proporção de hectares de negócios não varejistas por bairro.
LSTAT: Percentual da população de "classe baixa".
NOX: Concentração de óxidos nítricos (partes por 10 milhões).
PTRATIO: Proporção aluno-professor por bairro.
RM: Número médio de cômodos por habitação.
TAX: Valor total do imposto predial por $10.000.

Mesmo com os modelos calculados, se faz necessário a avaliação da bondade dos modelos.

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 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%.

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.