Banco de dados: Boston House Prices

Fernado Bispo, Jeff Caponero

Sumário

Sobre o banco de dados

Contexto

Os dados de preços de 506 casas em Boston publicados em HHarrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.
Os dados podem ser acessados em:
https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data.

Objetivo

O objetivo deste trabalho será determinar, a partir de técnicas de regressão linear, o preço de casas em Boston com base nos dados fornecidos pelo banco de dados análisado.

Informações do conteúdo do banco de dados

1) CRIM: índice de criminalidade per capita por bairro.

2) ZN: proporção de terreno residencial zoneada para lotes acima de 25.000 sq.ft.

3) INDUS: proporção de hectares de negócios não varejistas por bairro.

4) CHAS: Margem do rio Charles (1 se o trecho margeia o rio; 0 caso contrário).

5) NOX: concentração de óxidos nítricos (partes por 10 milhões) [partes/10M].

6) RM: número médio de cômodos por habitação.

7) AGE: proporção de unidades próprias 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: valor total do imposto predial por $10.000 [$/10k].

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".

Variável de saída:

1) MEDV: Valor médio de residências ocupadas pelo proprietário em US$1.000 [k$].

Fonte

StatLib - Carnegie Mellon University

Análise Descritiva

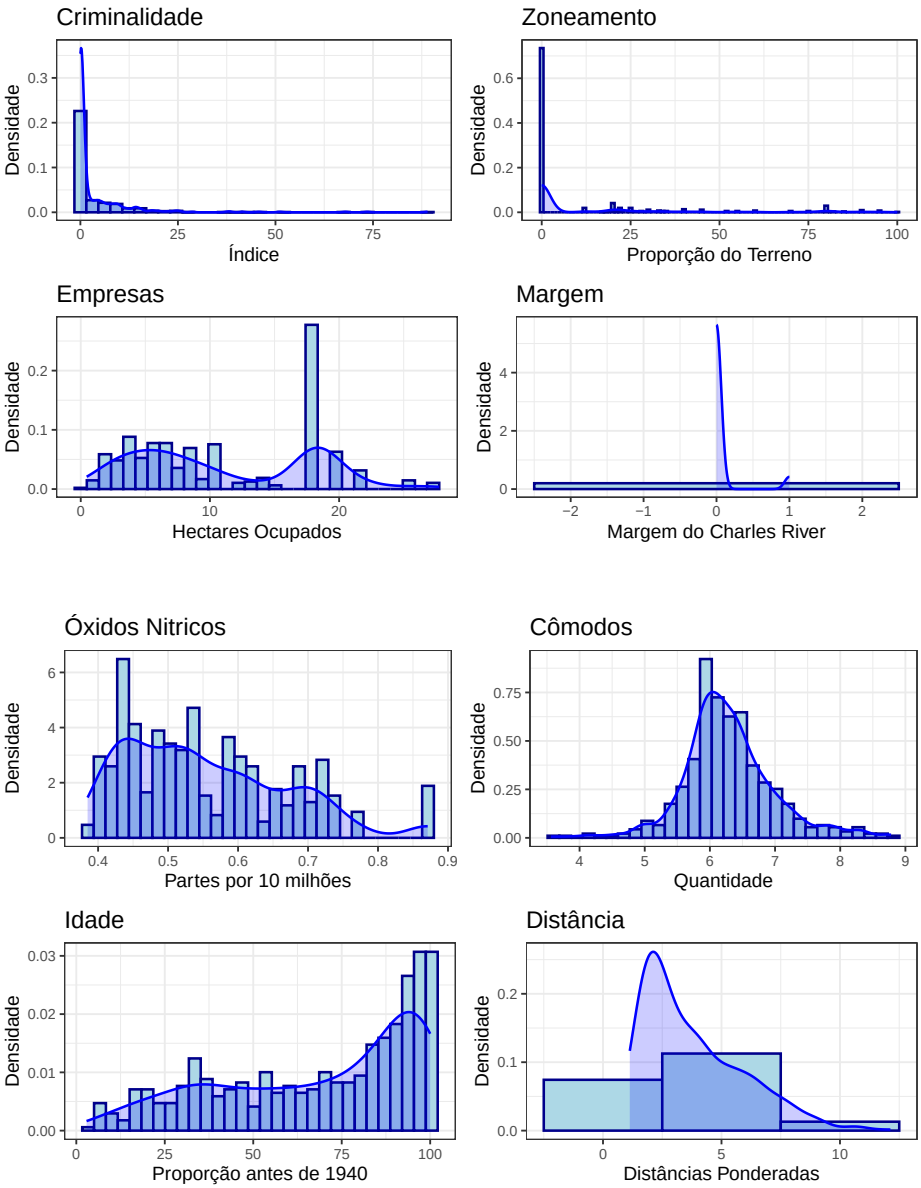
De modo a conhecer melhor o banco de dados analisado é importante realizar uma análise descritiva das variáveis que o compõem. Na Tabela 1 pode se ver as medidas de resumo de posição e de tendência central destas variáveis.

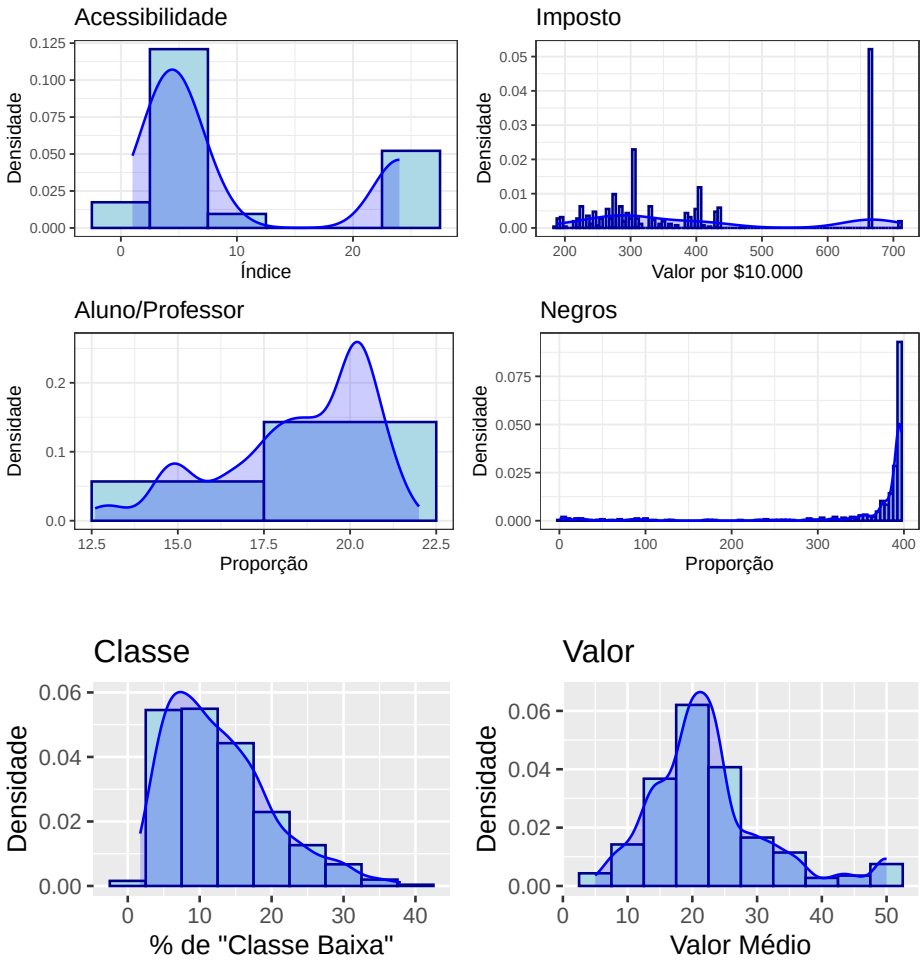
Tabela 1: Medidas Resumo dos dados

	Min	Q1	Med	Média	Q3	Max	D.Padrão	CV
AGE	2,90	45,00	77,50	68,57	94,10	100,00	28,15	0,41
B	0,32	375,33	391,44	356,67	396,23	396,90	91,29	0,26
CHAS	0,00	0,00	0,00	0,07	0,00	1,00	0,25	3,67
CRIM	0,01	0,08	0,26	3,61	3,68	88,98	8,60	2,38
DIS	1,13	2,10	3,21	3,80	5,21	12,13	2,11	0,55
INDUS	0,46	5,19	9,69	11,14	18,10	27,74	6,86	0,62
LSTAT	1,73	6,93	11,36	12,65	16,96	37,97	7,14	0,56
MEDV	5,00	17,00	21,20	22,53	25,00	50,00	9,20	0,41
NOX	0,38	0,45	0,54	0,55	0,62	0,87	0,12	0,21
PTRATIO	12,60	17,40	19,05	18,46	20,20	22,00	2,16	0,12
RAD	1,00	4,00	5,00	9,55	24,00	24,00	8,71	0,91
RM	3,56	5,88	6,21	6,28	6,62	8,78	0,70	0,11
TAX	187,00	279,00	330,00	408,24	666,00	711,00	168,54	0,41
ZN	0,00	0,00	0,00	11,36	12,50	100,00	23,32	2,05

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

Figura 1: Histogramas das variáveis em análise.





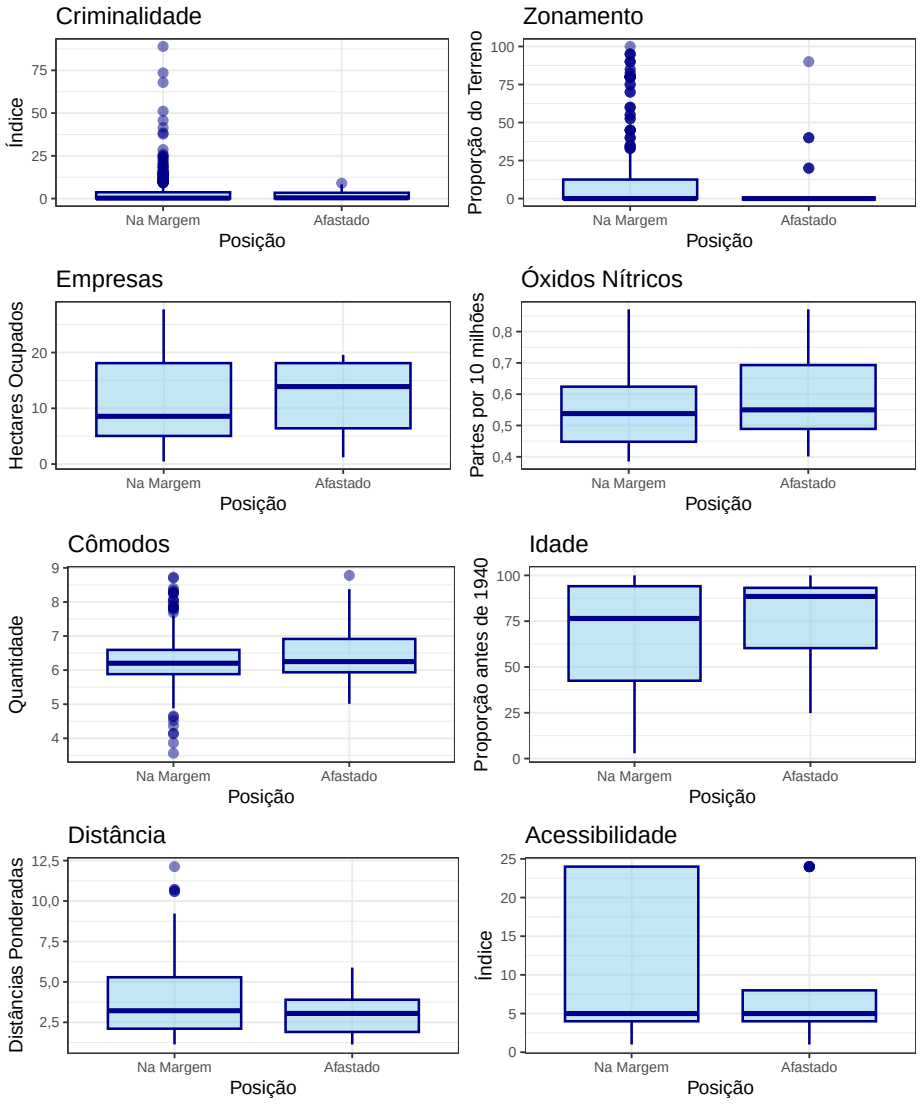
Desta análise inicial, destancam-se algumas características: 1) CRIM: índice de criminalidade per capita por bairro é bastante baixo na maioria dos bairros;
2) ZN: proporção de terreno residencial zoneada para lotes acima de 25.000 sq.ft. também qtem concentração de valores zeros;
3) INDUS: Verifica-se uma concentração de empresas com cerca de 18 hectares em diversos bairros; 4) TAX: há uma concentração de imóveis com alto valor total do imposto predial;
5) B: A maior parte dos bairros tinha alta proporção de negros. :::

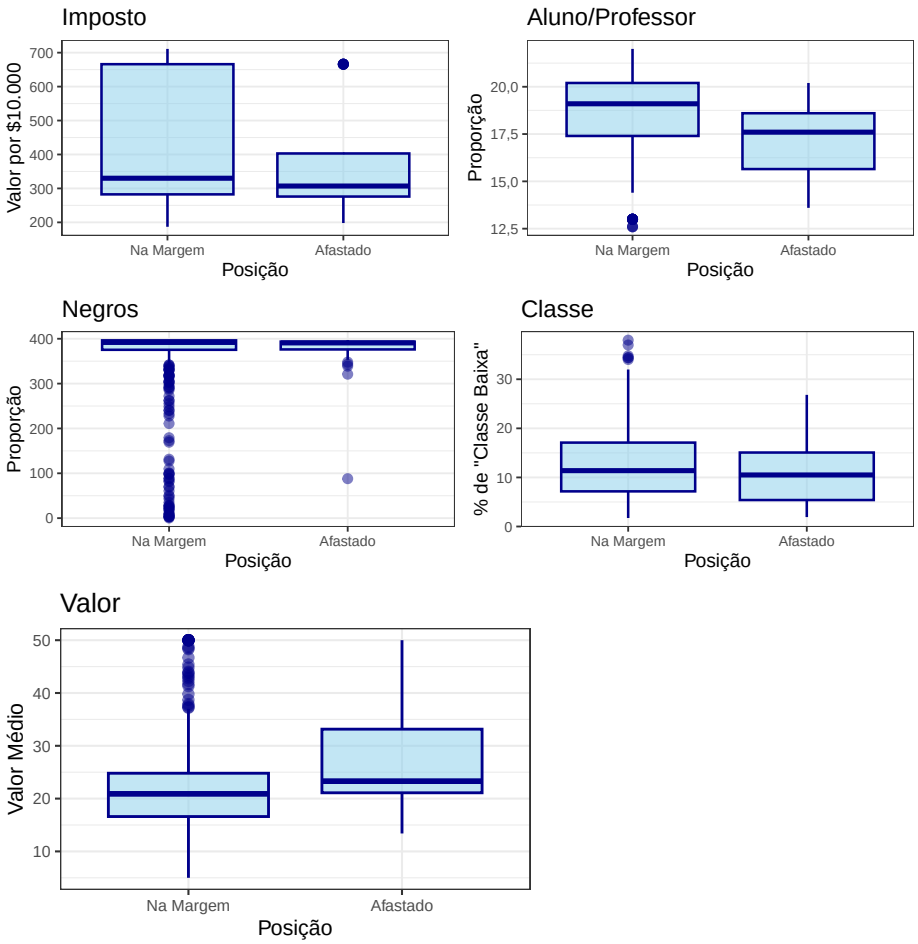
Análise de Dados Atípicos

Com base na variável CHAR, que indica se o imóvel margeia ou não o Charles River, pode-se realizar a análise de disperção dos dados por meio de gráficos do

tipo BoxPlot, como se vê na Figura 3.

Figura 2: BoxPlots entre a posição em relação ao Charles River e demais variáveis em análise.



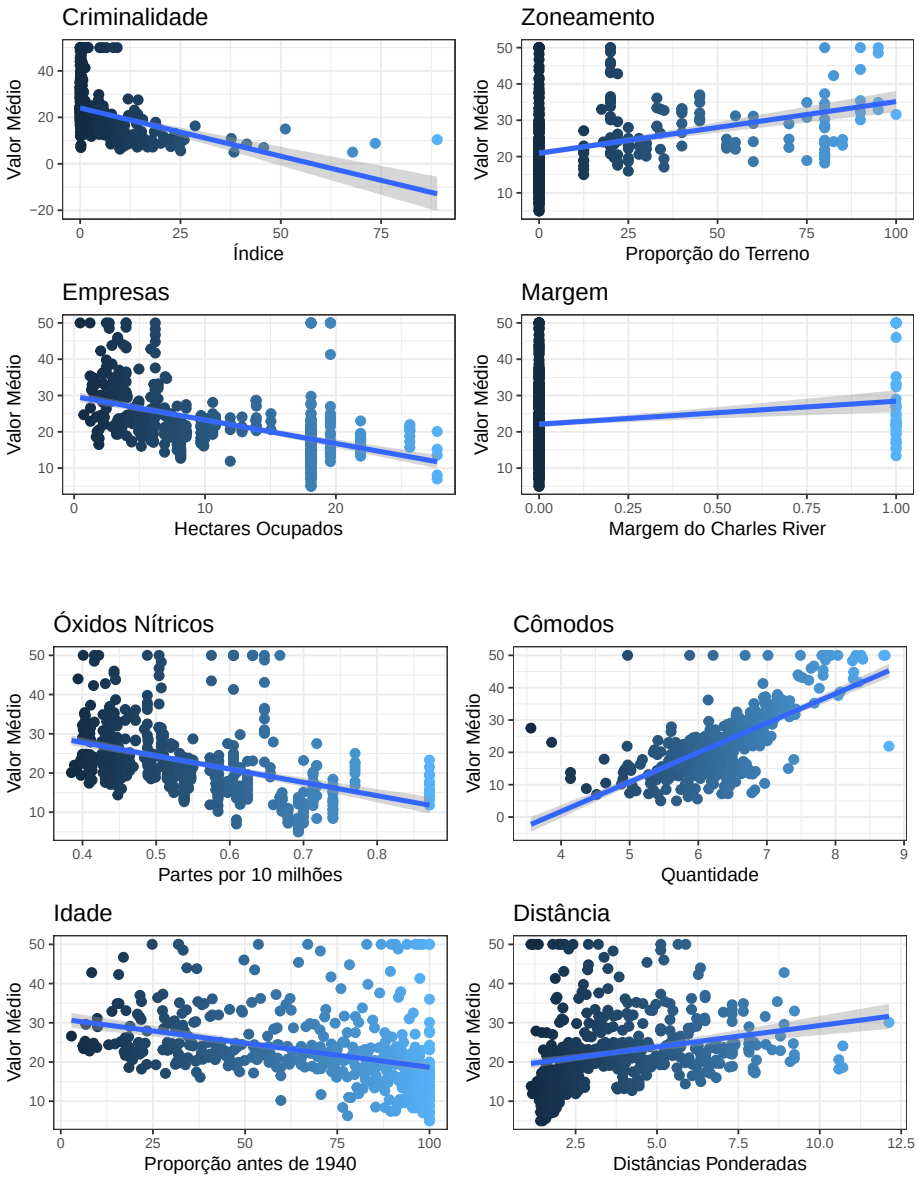


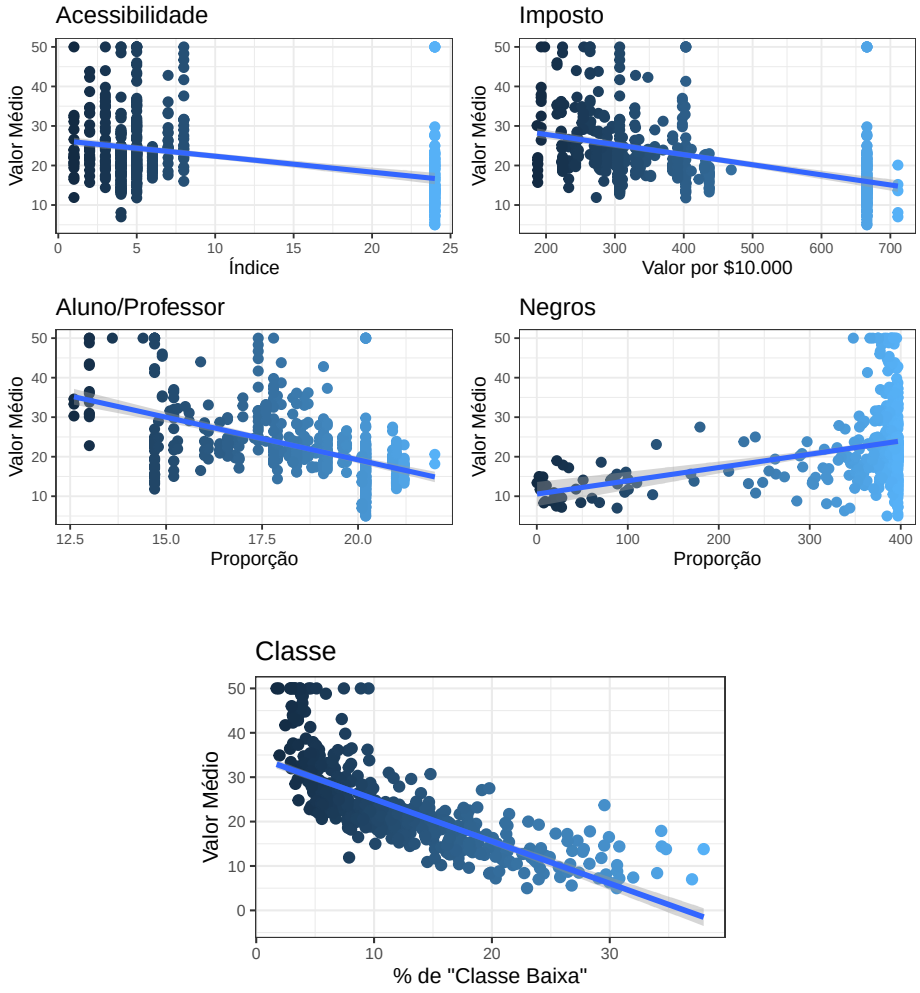
Pode-se verificar pela Figura 2 que praticamente todas as variáveis apresentam *outlayers*, entretanto o tratamento destes *outlayers* em todos os casos parece, salvo melhor juízo ser o mesmo. Observa-se que há cooerência entre eles, isto é, são realizades possíveis e não devem ser desprezados como se 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 3: Relação entre o Valor médio dos imóveis e demais medições





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 valore do modelo $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ como seus respectivos erros padrão (σ_0 e σ_1), alem 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.

Tabela 2: Valores dos modelos de regressão linear simples.

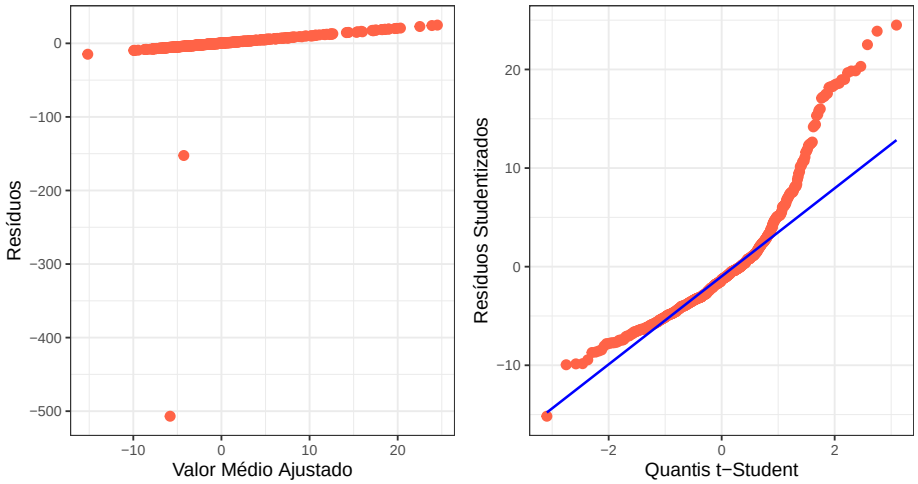
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
$\hat{\beta}_0$	24,033	20,918	29,755	22,094	41,346	-34,671	30,979
σ_0	0,409	0,425	0,683	0,418	1,811	2,650	0,999
$\hat{\beta}_1$	-0,415	0,142	-0,648	6,346	-33,916	9,102	-0,123
σ_1	0,044	0,016	0,052	1,588	3,196	0,419	0,013
p-valor	0,000	0,000	0,000	0,000	0,000	0,000	0,000
$\hat{\rho}$	-0,388	0,360	-0,484	0,175	-0,427	0,695	-0,377

	DIS	RAD	TAX	PTRATIO	B	LSTAT
$\hat{\beta}_0$	18,390	26,382	32,971	62,345	10,551	34,554
σ_0	0,817	0,562	0,948	3,029	1,557	0,563
$\hat{\beta}_1$	1,092	-0,403	-0,026	-2,157	0,034	-0,950
σ_1	0,188	0,043	0,002	0,163	0,004	0,039
p-valor	0,000	0,000	0,000	0,000	0,000	0,000
$\hat{\rho}$	0,250	-0,382	-0,469	-0,508	0,333	-0,738

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 linar. Observe que para todas os pares de varáveis 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 4: 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.