Ministério da Educação
Universidade Federal da Bahia
Campus Ondina
Departamento de Estatística - UFBA

LABORATÓRIO 8

LABORATORY 8

Gabriel Penha*, Moisés Augusto†, Marla Lorrani‡

RESUMO

Em trabalhos anteriores, foi feito um estudo para a compreensão de aspectos que influenciariam no preço dos imóveis de um determinado país. A relação entre o preço dos imóveis e o seu tempo de construção não apresentaram uma comportamento linear, mas aparentavam ter uma relação quadrática. Baseado nisso, é plausível o ajuste de uma regressão polinomial. Desta maneira, diversos modelos de regressão polinomial foram comparados através de métodos de seleção de modelos como seleções *Backward*, *Forward* e *Stepwise*, avaliando-os com critérios como *AIC* e *BIC*. Para este trabalho, foram feitas alterações na base de dados anterior, como a retirada de algumas observações.

Palavras-chave: Preço do imóvel. Idade do imóvel. Modelo de Regressão Polinomial. Seleção de Modelos.

1 INTRODUÇÃO

No Relatório do Laboratório 3, disponível previamente para o leitor, as questões levantadas por Morais et al. (2015), foram discutidas. Uma base de dados que considerava informações como localização, tempo de construção e distância de imóveis até o metrô foi considerada.

Naquela ocasião, verificou-se que havia uma diferença relevante no preço das residências a depender de sua distância para com o metrô. Verificou-se ainda que a variável idade do imóvel - isto é, seu tempo de construção -, não apresentava uma relação linear muito forte para com o preço.

Dito isto, neste relatório 8, retomou-se o conjunto de dados do laboratório 3. Com as seguintes diferenças:

- A idade do imóvel foi analisada mais minunciosamente;
- Ajustaram-se alguns modelos lineares; sendo que um destes foi selecionado após métodos de seleção de modelos.

Na seguinte seção, ter-se-á uma análise exploratória dos dados, incluindo uma análise visual. Em seguida, a seção de resultados tratará dos modelos ajustados, das escolhas feitas para a seleção do melhor deles e, finalmente, análise de pressupostos de alguns destes modelos. A última seção conterá algumas considerações finais.

* Instituto de Matemática e Estatística, Departamento de Estatística, Bacharelando em Estatística; penha.gabriel@ufba.br.
† Instituto de Matemática e Estatística, Departamento de Estatística, Bacharelando em Estatística; moises.augusto@ufba.br.
‡ Instituto de Matemática e Estatística, Departamento de Estatística, Bacharelando em Estatística; marla.lorrani@ufba.br.

2 ANÁLISE EXPLORATÓRIA

O conjunto de dados utilizado possuía 360 observações, não haviam observações faltantes ou repetidas. O sumário das descritivas dos dados pode ser visualizado na Tabela 1

Tabela 1 – Estatísticas descritivas dos dados

Descritiva	Distância - Metrô	Nº Lojas	Idade	Preço do Imóvel ($)
Mínimo	21,38	0,00	0,00	11,20
1° Quartil	264,56	2,00	9,93	29,50
Mediana	450,10	5,00	16,25	39,40
Média	916,93	4,24	17,92	38,58
3° Quartil	1243,73	6,00	27,70	46,12
Máximo	5932,72	10,00	43,80	78,30

Pelas descritivas, a distribuição do número de lojas e do preço do imóvel parecem ser simétricas, bem como o tempo de construção. No entanto, a distribuição da distância do metrô parece ter uma assimetria considerável à direita.

Na Figura 1, é possível visualizar as distribuições de cada uma dessas variáveis.

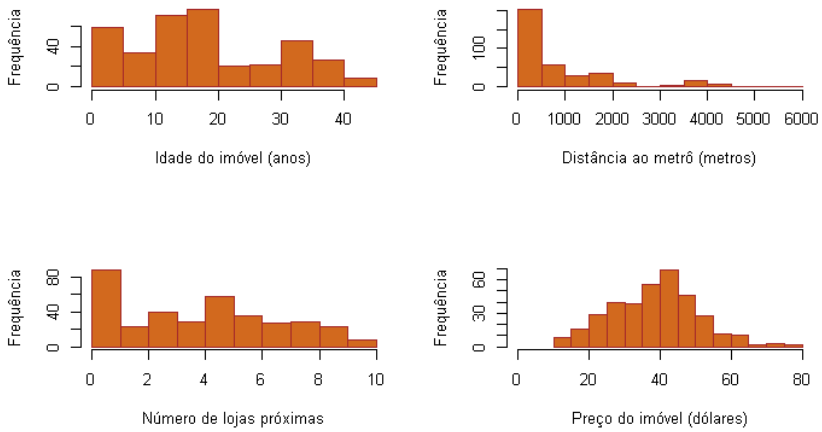

Figura 1 – Histogramas: Distância até o metrô; Lojas; Idade; Preço do Imóvel

A seguir, é possível visualizar um diagrama de dispersão entre a variável resposta preço do imóvel e o tempo de construção da casa.

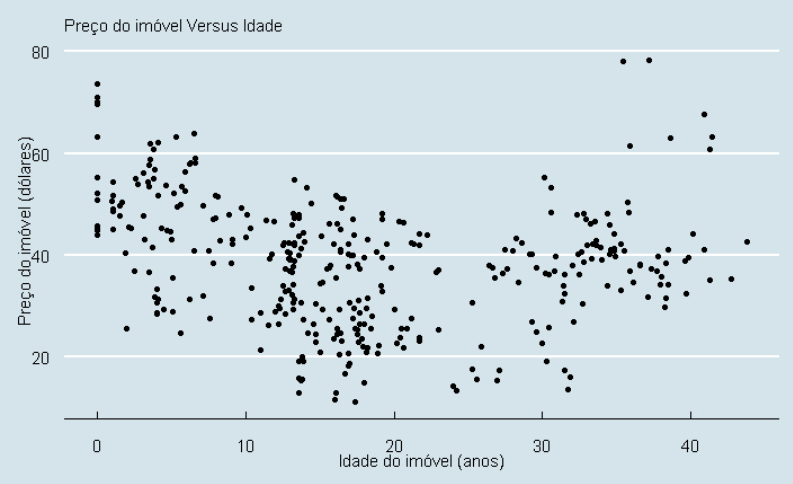

Figura 2 – Diagrama de Dispersão: Preço do Imóvel Vs. Idade

Como foi dito inicialmente, a relação entre as características do preço e a idade do imóvel não parecem apresentar um comportamento linear, mas um quadrático. Dito isso, alguns ajustes se atentando a uma relação polinomial entre as duas variáveis serão considerados. Para uma análise descritiva bi-variada mais detalhada levando em conta as outras co-variáveis, um *link* para o relatório do laboratório 3 estará disponível nos Anexos deste documento.

3 RESULTADOS

Considere que *tempo* indicará a idade do imóvel, *metro* a distância em metros do imóvel ao metrô e que *lojas* será o número de lojas de conveniência próximas ao imóvel.

Vários ajustes de modelos de regressão foram considerados; estes serão explicitados adiante;

O Modelo 1 (polinomial usual de grau 2) estimado ficou da seguinte forma:

$$\hat{Y} = 54{,}35 - 2{,}05tempo + 0{,}04tempo^2 \tag{1}$$

Em que $\hat{Y}$ indica o valor estimado para o preço do imóvel. Este modelo implica que, fixado o $tempo^2$, o acréscimo de um ano no tempo de construção da casa diminuirá o preço desta em \$2,05. Além disso, com o tempo fixado, a cada acréscimo em $tempo^2$, aumenta em \$0,04. Obviamente não se trata de uma interpretação que possa ser realizada tão simplesmente; mas de modo geral, ela pode significar que imóveis muito velhos podem começar a valorizar - possivelmente por questões históricas ou afins -. Além disso, para imóveis que não possuem nem um ano de idade, o preço médio do imóvel seria de \$54,35. Neste modelo, tanto o tempo quanto o tempo^2 foram consideradas significativas ao nível de 95%.

O segundo modelo, ajustou um polinômio usual de grau 3; ficou da seguinte forma:

$$\hat{Y} = 54{,}83 - 2{,}22tempo + 0{,}06tempo^2 - 0{,}0002tempo^3 \tag{2}$$

Com as variáveis definidas como anteriormente. Neste modelo, para imóveis de idade nula o preço médio do imóvel seria de \$54,83. Este modelo indica que a cada acréscimo na idade do imóvel, fixadas as variáveis $tempo^2$ e $tempo^3$. As interpretações de $tempo^2$ e $tempo^3$ são análogas. Neste modelo, apenas o $tempo^3$ não foi significativo a 95%.

Os Modelos 3 e 4 foram ajustados com as mesmas variáveis do 1 e 2, com a diferença de que eram ortogonais.

$$\hat{Y} = 38{,}58 - 42{,}72tempo + 115{,}85tempo^2 \tag{3}$$

Para este modelo, casas com idade nula têm um preço médio de $38,58. Para $tempo^2$ fixado, há decréscimo de $42,72. Com o *tempo* fixado, há um aumento de $115,85. Neste modelo, tanto o *tempo* quanto o $tempo^2$ foram consideradas significativas.

$$\hat{Y} = 38{,}58 - 42{,}72tempo + 115{,}85tempo^2 - 5{,}14tempo^3 \tag{4}$$

Neste modelo, casas com idade nula tem um preço de $38,58. Para $tempo^2$ fixado, há decréscimo de $42,72 neste preço e com o *tempo* fixado, há um aumento de $115,85. Finalmente, para $tempo^3$ fixado, há um decréscimo de $5,14 no preço do imóvel.

3.1 Seleção de Modelos

Primeiramente, consideraram-se os Modelos 1 e 2, isto, é, em que a regressão polinomial não era ortogonal e, um terceiro modelo, que considerava somente a variável tempo (isto é, era um modelo de regressão linear simples).

Utilizando o princípio do resíduo condicional, analisaram-se estes três ajustes e, chegou-se a conclusão, que o modelo 1, que considerava uma regressão polinomial com acréscimo somente do tempo ao quadrado tinha resultados um pouco melhores que os outros dois.

Assim, verificou-se qual seria a previsão deste modelo para alguns valores, e um deles, por exemplo, foi com tempo igual a 45 anos. O resulto encontrado para tal valor de idade foi de $56,10. Vale dizer, no entanto, que 45 anos é um valor que não está presente nas observações fornecidas pela base de dados. Deste modo, não é possível afirmar que a relação do tempo pelo preço do imóvel iria se manter da mesma maneira e, portanto, este tipo de previsão (fora do intervalo da co-variável na base de dados) não é recomendada.

Finalmente, um novo modelo, que será chamado de Modelo 5, considerando, além do tempo e do tempo ao quadrado, a distância em metros ao metrô e o número de lojas próximas do imóvel foi ajustado.

Através dele, utilizaram-se alguns métodos de seleção de variáveis.

Consideraram-se três abordagens:

- Backward;
- Forward;
- Stepwise;

E as métricas utilizadas foram as estatísticas AIC e BIC, além dos coeficientes de determinação ajustado e de C_P de Mallows.

Para a abordagem *Backward*, os menores AIC e BIC foram obtidos através do modelo completo, isto é: o que considerou o tempo e o tempo quadrático, além da distância ao metrô e o número de lojas. O AIC deste modelo foi de 1462,10, enquanto o BIC dele foi 1481,60. Seu coeficiente de determinação ajustado foi próximo de 0,62.

Na abordagem *Forward* o Modelo 5 também foi o selecionado. Respectivamente, o AIC e BIC foram de 1462,14, muito próximo da abordagem *Backward* e 1481,57.

Finalmente, para o *Stepwise*, o AIC e o BIC do Modelo 5 foram de respectivamente 1462,10 e 1481,60, exatamente os valores da abordagem *Backward*. Ressalta-se que o modelo selecionado através do método de seleção *Stepwise* também foi o Modelo 5.

Dito isso, uma comparação de modelos utilizando o coeficiente de C_P de Mallows pode ser visualizada na Tabela 2.

Tal comparação é feita considerando um modelo que ajusta o tempo, o número de lojas próximas e a distância ao metrô (que foi chamado de Modelo 6).

Tabela 2 – Comparação dos coeficientes C_P de Mallows

Variáveis do modelo comparado	C_P de Mallows
tempo	483,42
tempo, tempo^2	270,55
tempo, tempo^2, metrô	10,43
tempo, tempo^2, loja	77,01
tempo, tempo^2, metrô, loja	-24,53

Note que, na segunda linha da tabela, o modelo comparado é justamente o Modelo 1. Note que os melhores resultados de coeficientes de C_P de Mallows são os menores e, portanto, o ajuste da última linha (que é o Modelo 5), foi o que apresentou o melhor resultado.

Dito isso, os pressupostos do Modelo 5 foram verificados.

3.2 Análise de resíduos - Modelo 5

Na Figura 3, é possível visualizar graficamente pressupostos como normalidade, homocedasticidade, alguns pontos influentes, entre outros.

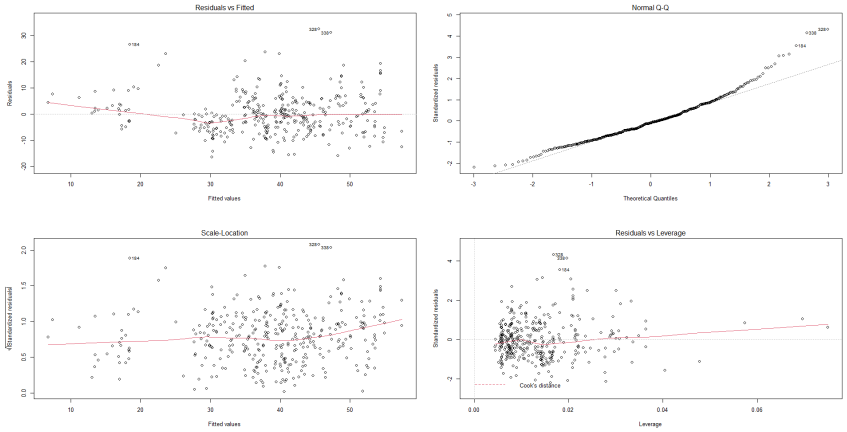

Figura 3 – Análise de Resíduos

É possível perceber uma leve tendência de variação quadrática nos resíduos; além de um escape do que era esperado nas caudas do gráfico quantil quantil. Todos os 4 gráficos apontaram para alguns pontos que precisam ser observados com maior cautela.

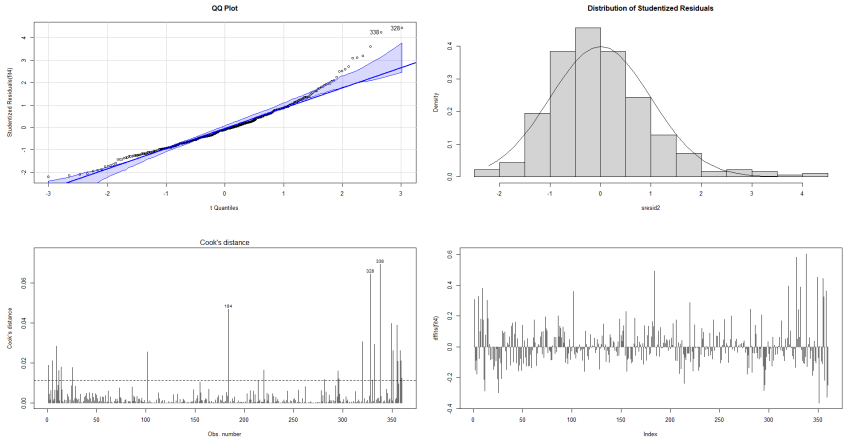

Figura 4 – Análise de Resíduos 2

Os gráficos da Figura 4 reforçam a possível não normalidade dos resíduos, bem como a possível influência dos pontos de $id = 184, 328 e 338$. Em particular, o gráfico da distância de Cook aponta que talvez existam mais informações influentes do que as mencionadas até aqui.

Testes de hipótese foram realizados para verificar o observado nas Figuras 3 e 4. Os testes apontaram tanto para a normalidade quanto para a homocedasticidade do modelo, ao nível de 95%, apesar do que foi visto nos gráficos.

Na Figura 5, é possível observar os $DFBetas$ do tempo e do tempo^2.

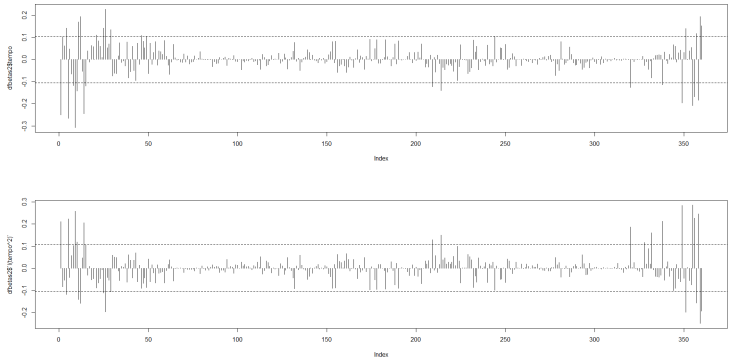

Figura 5 – DFBetas

Novamente, em ambas as variáveis é possível observar alguns pontos fora do intervalo esperado, o mesmo se repete para as variáveis lojas e distância até o metrô, que podem ser vistas na Figura 6.

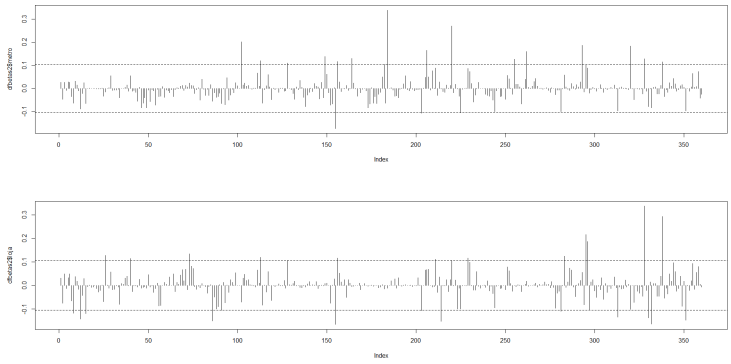

Figura 6 – DFBetas 2

Finalmente, analisaram-se os $COVRatio's$ do modelo. Uma visualização gráfica destes valores pode ser vista na Figura 7.

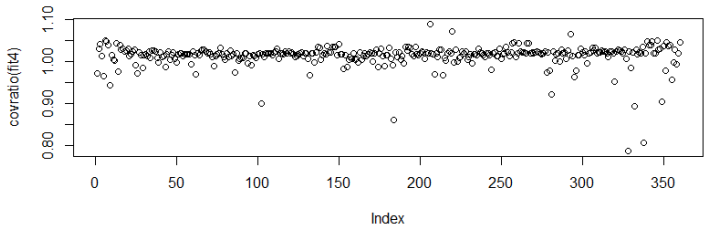

Figura 7 – Covratios

4 CONSIDERAÇÕES FINAIS

Considerando o mencionado, o Modelo 5 foi superior aos demais modelos ajustados através de diversos métodos de seleção de variáveis.

Apesar de um comportamento não muito esperado de alguns dos resíduos e, sobretudo de alguns pontos que foram considerados influentes, decidiu-se não realizar outros ajustes sem estas variáveis; tendo em vista que possivelmente são valores que podem acontecer na vida real.

Deste modo, algumas coisas poderiam ser consideradas para melhorar o ajuste; entre elas um maior tamanho de amostra e mais co-variáveis que possam explicar o preço dos imóveis, como, por exemplo, o tamanho do imóvel, o número de quartos e afins.

REFERÊNCIAS

MORAIS, A. C. de et al. Influência da proximidade à estação do metrô no valor da propriedade residencial urbana: um estudo no bairro de Samambaia em Brasília. **ANPET (esc) Aspectos Econômicos, Sociais, Políticos e Ambientais do Transporte, Ouro Preto, MG**, p. 2153–2161, 2015.

ANEXO A – CÓDIGOS UTILIZADOS PARA ANÁLISE (NO R)

É possível todos os laboratórios feitos até agora através do *link* a seguir:
https://github.com/Mkyou/labs-regressao

DECLARAÇÃO DE RESPONSABILIDADE

O(s) autor(es) é(são) o(s) único(s) responsável(eis) pelas informações contidas neste documento.