UNICAMP - Universidade Estadual de Campinas

IMECC - Instituto de Matemática, Estatística e Computação Científica

Projeto 1 de ME613 - Análise de Regressão

Alex Dos Santos Lima - RA: 230560

Professor Aluísio de Souza Pinheiro

20 de Agosto a 28 de Setembro de 2021

1. Introdução ao conjunto de dados

O objetivo deste relatório é realizar uma análise de Regressão Linear Simples do pacote de dados disponível do Shoftware *R* cuja o nome da biblioteca é *Stat2Data* sendo o pacote com o nome *HousesNY*, com 5 variável e 53 observações do conjunto de dados. As variáveis que será explorada ao longo do relatório é *Price*, *Beds*, *Baths*, *Size*, *Lot* e traduzindo para o português, *Dinheiro*, *Cama*, *Banho*, *Tamanho*, *Quantidade*.

Sendo assim, uma breve descrição do conjunto de dados onde temos, CasasNY são preços de casas na zona rural de NY, Argumentos do formato das variáveis. O quadro de dados se encontra com 53 observações nas 5 variáveis a seguir.

- *Price*, Preço estimado (em $ 1.000),
- *Beds*, Números de quartos,
- *Beths*, Números de banheiros,
- *Size*, Área da casa (em 1.000 pés quadros),
- *Lot*, Tamanho do lote (em ocres).

Detalhes, os dados extraídos de Zillow.com para uma amostra de casas próximas ao código de área 13617 (Canton, NY, uma pequena cidade no interior do estado de NY). Casas em lotes maiores que cinco acres (geralmente fazendas) foram excluídas.

2. Análise Descritiva dos dados

Esta seção terá como objetivo a ser destacados algumas correlações entre as variáveis pressentes no banco de dados, por meio de uma breve análise descritiva, que podem auxiliar no caminho de um possível modelo em Regressão Linear Simples.

Sendo assim, para ilustrar análise descritiva e melhor o entendimento, o gráfico da frequência de cada uma das observações do conjunto de dados realizado na figura 1 e é realizado as possíveis tendencia das maiores e menores frequência relativa em cada observação.

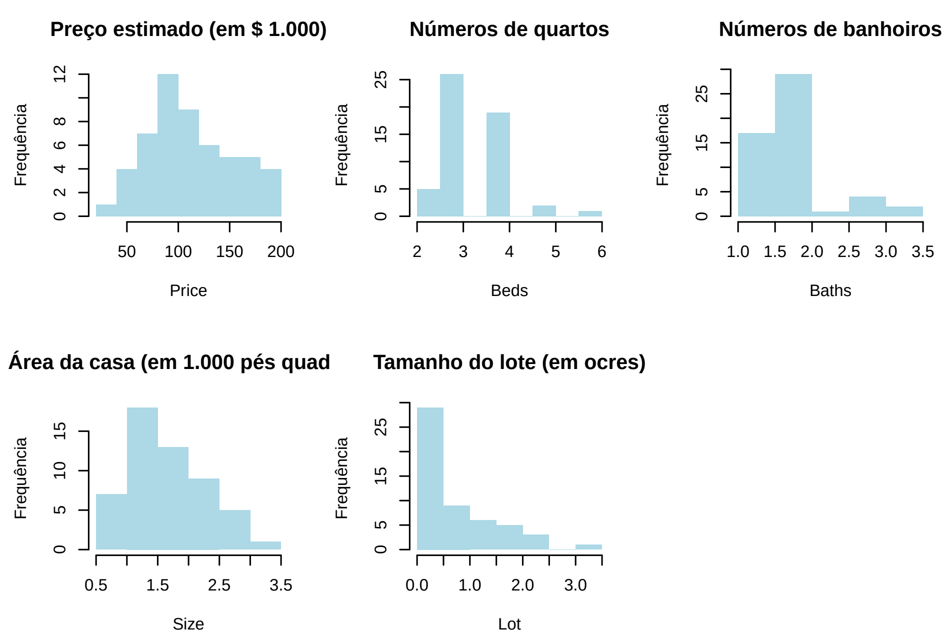


Figura 1: Gráficos das frequências de cada observação

Logo, o que se pode perceber é que o *número de quarto* tem uma frêquencia de 25 quando é alugado em até 2 ou 3 pessoas, já o *tamanho do lote* sua frêquencia é apróximadamente 30 pelo o pico da altura da barra.

Sendo assim, ao analisar a tabela da Estatística descritiva dos cálculos de cada observação com *mediana, média, desvio padrão, variância, amplitude, valores máximos e mínimos.* Percebemos que a variável *Price* que está acima de todas as outras variáveis e comparação com estatísticas.

Table 1: Estatístia Descritiva

Nomes	Price	Beds	Baths	Size	Lot
Média	113.63	3.40	1.86	1.68	0.80
Médiana	107.00	3.00	2.00	2.00	0.00
Desvio Padrão	41.43	0.79	0.65	0.60	0.76
Variância	1716.00	1.00	0.00	0.00	1.00
Amplitude	159.00	4.00	2.00	2.00	4.00
Máximo	197.50	6.00	3.50	3.10	3.50
Mínimo	38.50	2.00	1.00	0.71	0.00

Quando é feito a realização entre todas as variávies do conjunto de dados na figura 2, temos que o valor *Price*, ou seja, *Preço estimado (em $ 1.000)*, está fazendo uma comparação com todas as outras observações nos gráficos. Afim de verificar o quanto de gasto estão ralacionado com as outras variáveis.

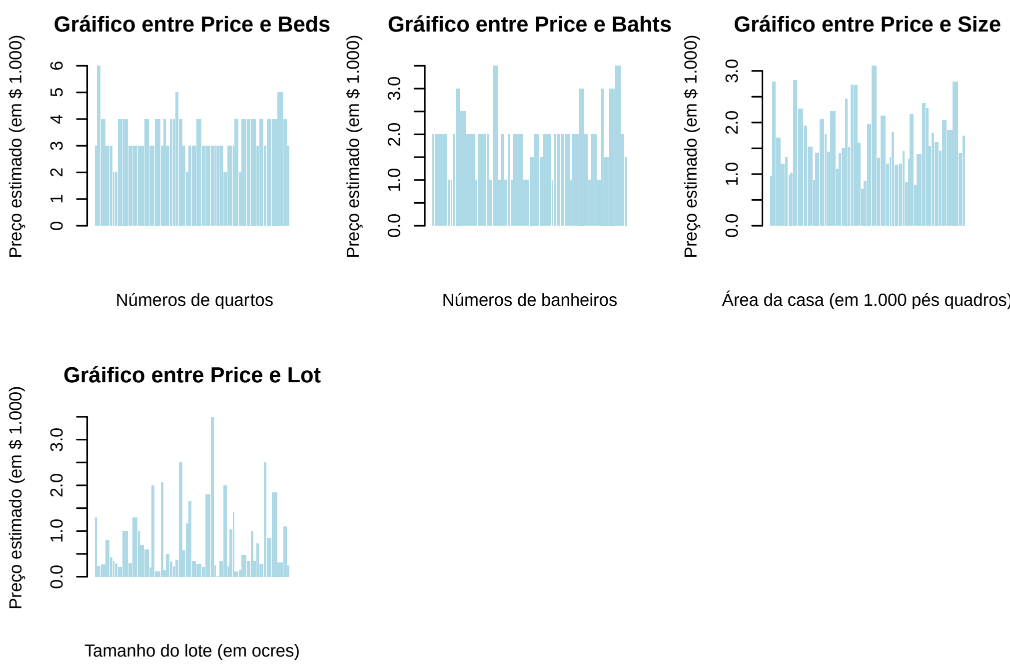


Figura 2: Graficos de barra contendo todas as observações

Logo, nos gráficos de barras onde se encontra todos na figura 2, temos que a observação do *Beds* e *Bath* que é respectivamente o *Números de quartos* e *Números de banheiros* fica quase constate nos valores em *4* e *2.0*, o que podemos análisar que houve um gasto maior com os quartos e os números de banho.

3. Proposta do modelo

Devemos modelar a variável *Price* onde representa o *Preço estimado (em $ 1.000)* com essa resposta e as demais variáveis usadas como preditoras, propõe-se o desenvolvimento através de uma representação linear no seguinte formato em análise de Regressão Linear Simples:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_{2i} + ... + \beta_n X_{in} + \epsilon_i$$

Onde é representado o seguinte valor, Y_i será a variável *Price* onde é o *Dinheiro*, X_{i1} até X_{in} as demais n variáveis preditoras, β_0 até β_n os n coeficientes a serem estimados e ϵ_i o erro associado.

Dentre esse modelo, supõe-se que os erro ϵ iid $N(0, \sigma^2)$ e que a variância σ^2 é constante sob a reta.

Assim, ao usar a metodologia para avaliar a qualidade do modelo que será aplicado na Regressão Linear Simples em uma variável preditora, onde temos por definição, seja uma variável de resposta Y de interesse e X uma variável regressão. O Modelo de Regressão Linear Simples de Y em X é dado por:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, ..., n$$

em que:

$\epsilon_1, ..., \epsilon_n$ são v.a.'s não-correlacionadas de forma que $E(\epsilon_i) = 0$ e $Var(\epsilon_i) = \sigma^2$;

$X_1, ..., X_n$ são valores fixos e conhecidos;

$\beta_0, \beta_1, \sigma^2$ são parâmetros, com valores fixos e desconhecidos.

Explicação do módelo

Na análise descritiva dos dados afim de averiguar uma variável resposta para o modelo. O que se faz perceber é que a observação *Price* é o dinheiro estimado que as pessoas possivelmente irão usar no Hotel em NY, sendo assim, essa variável será resposta do modelo para Análise de Regressão Linear Simples. Como todas as outras estão em função desta variável. Vamos usar o fato de X estar relacionado somente com uma só variável e Y a resposta para o modelo.

Vamos realizar um modelo contendo todas as variáveis em Análise de Regressão Linear Simples, onde os valores são:

- β_0 é o *Price*, Preço estimado (em $ 1.000)
- β_1 é o *Beds*, Números de quartos,
- β_2 é o *Baths*, Números de banhoiros
- β_3 é o *Size*, Área da casa (em 1.000 pés quadros),
- β_4 é o *Lot*, Tamanho do lote (em ocres),

Table 2: Modelo com todas as variáveis

	Estimadores	S. dos Q. Erros	t-Valor	Pr(>|t|)
(β_0)	14.5899	23.2658	0.63	0.5336
β_1	2.7708	8.7303	0.32	0.7523
β_2	26.2384	7.8438	3.35	0.0016
β_3	22.1551	11.9308	1.86	0.0695
β_4	4.6211	6.1839	0.75	0.4585

Logo, temos o modelo de Regressão Linear Simples da seguinte forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i, i = 1, ..53$$

Onde temos:

Y_i é a respota estimada méida de uma possível observação em i

X_i é uma constante conhecida, no nivel da variável preditora em i

β's são parametros de estimação do conjunto de dados

ϵ_i são independente em $N(0, \sigma^2)$

$i = 1, ..., n$

Então temos o seguinte modelo:

$$Y_i = 14.5899 + 2.7708X_{1i} + 26.2384X_{2i} + 22.1551X_{3i} + 4.6211X_{4i} + \epsilon_i$$

Neste modelo onde tem todas as variáveis, percebemos que somente o $\beta_2 = 26.2384$ tem o $p - valor < \alpha$ do que o usual, agora se adotar um nível de significância para adequabilidade, do modelo sem fazer nenhuma transformação nas variáveis porque ainda precisamos avaliar o diagnóstico do modelo se está adequado para Análise de Regressão Linear Simples ou não. Pois, se adotarmos um nível de significância de $\alpha = 0.05$ só essa o β_2 terá menor que α definido isso, ainda precisamos verificar o diagnóstico do modelo. Com isso vamos verificar adequabilidade, do modelo, se ele tem homoscedasticidade, variância constante, fazendo gráficos dos resíduos vs variável resposta, resíduos vs variável preditora.

4. Adequação do modelo inferencial

Para o modelo se adequar os diagnósticos pelos gráficos onde são mais eficientes do que os testes. Será feito uma inferência em função de verificar todas as características empíricas esperadas dos resíduos, de acordo com a estrutura proposta dos erros. Depois, usando-se os instrumentos adequados (gráficos e testes). Se for necessário uma construção de uma Tabela ANOVA, assume-se o Modelo de Regressão Linear Simples e testando-se para situação em que $\beta_k = 0$. Também se testa a relação linear entre X e Y é estatisticamente razoável. Onde temos $Y_1, ..., Y_n$(ou condicionados a X) são normais com variâncias iguais e independentes.

Afim de fazer uma análise a respeito do modelo ajustado, devemos verificar a Tabela ANOVA.

Table 3: Tabela ANOVA

	Df	Soma Q.	Média Q.	Valor de F	Pr(>F)
β_1	1	15679.90	15679.90	14.37	0.0004
β_2	1	17109.98	17109.98	15.69	0.0002
β_3	1	3498.48	3498.48	3.21	0.0796
β_4	1	609.14	609.14	0.56	0.4585
Residuos	48	52357.90	1090.79		

Para obtermos um gral de confiança do modelo qua vamos usar temos que testar a situação assumindo-se um Modelo de Regressão Linear Simples na situação em que $\beta_k = 0$. Nesta caso será feito o intervalo de confiança para os parâmetros de β_k, usaremos a distribuição t-student, assumindo

$T = \frac{\beta_k}{\sqrt{Var(\hat{\beta_k})}}$ ~$ $t_{48,2.5\%} = 2.010635$

onde $IC(95\%, \beta_k) = (\beta_k \pm t_{48,2.5\%}\sqrt{\hat{Var(\hat{\beta_k})}})$

Concluimos pelo teste do intervalo de confiaça, que existe dentro dos intervalo o zero, então não podemos atestar com 95% de confiança que β_k é não nulo, ou seja, as variávies de fato precisa ser análisada mais para verificar o ajusto do modelo.

Afim de chegar em uma segundo teste, onde temos a heterocedasticidade do modelo, e para verifica-la, utilizaremos o teste de Breush-Pagan. Também é importante ressaltar que o modelo apresenta $R^2 = 0.4134$

Table 4: Intervalos de Confiança para os parâmetros

Parâmetro	2.5 %	97.5 %
(β_0)	-32.19	61.37
β_1	-14.78	20.32
β_2	10.47	42.01
β_3	-1.83	46.14
β_4	-7.81	17.05

Table 5: Testes de hipótese para os resíduos do modelo final

test	tema	p-valor
Shapiro-Wilk	Normalidade	0.1673
Breush-Pagan	Homocedasticidade	0.345
Durbin-Watson	Linearidade	0.5746

e o a R ajustdo $R^2_{ajut} = 0.3645$, foi realizado os testes de hipótese para verificar se os resíduos seguem os requisitos de normalidade, homocedasticidade e linearidade.

5. Analise dos Residuos e diagnósticos

Para verificar a normalidade, primeiro será observado o histograma dos resídos padronizados afim de si visualizar uma possível tendencia de uma normalidade.

Figura 3: Distribuição do histogrma dos resíduos

Pelo o histograma acima nota-se descritivamente uma distribuição próxima da normal centrada no zero. Para comfirmar a normalidade será usado o gráfico Q-Q Plot e o teste de Lilliefors.

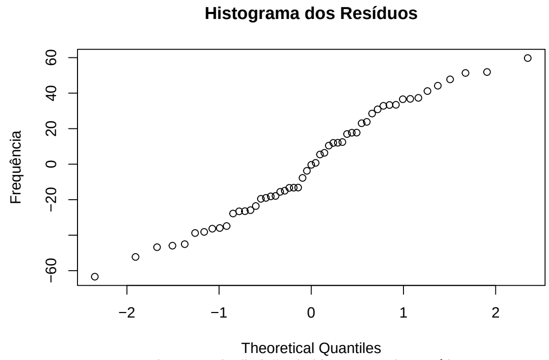
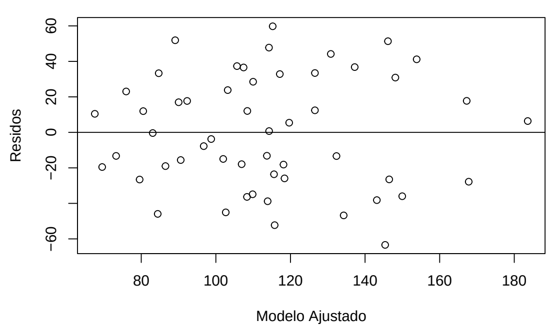


Figura 3: Distribuição do histograma dos resíduos

Graficamente, pode-se notar uma tendência de normalidade dos resíduos padronizados, porém, foi-se realizado todos os testes de hipótese nula não foi rejeitada nenhum com (p-valor $< \alpha$), ou seja, com base na amostra e com 95% de confiança, pode-se concluir que os resíduos padronizados seguem uma distribuição normal.



Com o fim de chegar a uma análise mais consistente, foi feito uma análise de resíduos (o resíduo é um termo adotado na estatística para diferenciar entre valor predito no modelo e valor real observado), esta análise é feita levando em consideração os resíduos padronizados, que consiste no resíduo dividido pela sua variância. Desta maneira, para que o modelo esteja estatisticamente correto, os resíduos padronizados devem seguir distribuição normal, conter variância constante e ser independentes, ou seja, não correlacionados.

6. Interpretação do modelo e resultados e conclusões

Consideremos o modelo de Regressão Linear Simples, então temos:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i + \beta_3 X_i + \beta_4 X_i, i = 1, ..., 53$$

O valor de β_0 é o intercepto do coeficiente ângular em 14.5899, ou seja, quando todos os $X_i = 0$, sendo a esperança da resposta média estimada do preço em \$ 1.000 no modelo de Regressão Linear Simples, uma suposição é que se todas os outros βi onde $i = 1, .., 4$ forem iguais a zero (0).

O coeficiente β_1 indica o acréscimo de 2.7708 na média final do *Preço estimado (em \$ 1.000)* da na resposta para a variável (*Price*) quando é *Beds* estiver no *números de quartos*, quando as outras variaveis estiverem fixas no modelo.

O coeficiente β_2 indica um acréscimo de 26.2384 na média final do *Preço estimado (em $ 1.000)* na resposta (*Price*) quando a *Baths* estiver no *números de banheiros* e quando as outras variáveis estiverem fixas no modelo.

O coeficiente β_3 indica um acréscimo de 22.1551 na média final do *Preço estimado (em $ 1.000)* na resposta (*Price*) quando a *Size* for a variável *área da casa (em 1.000 pés quadros)*, mantendo as outras variáveis fixas.

O coeficiente β_4 indica um acréscimo de 4.6211 na média final do *Preço estimado (em $ 1.000)* na resposta (*Price*) quando a *Lot* estiver no *tamanho do lote (em ocres)*, mantendo as outras variáveis fixas.

Bibilografica de referencia

https://www.rdocumentation.org/packages/Stat2Data/versions/2.0.0/topics/HousesNY, consultado em 24-09-2021