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

LABORATÓRIO 5
LABORATORY 5

Gabriel Penha*, Moisés Augusto†

RESUMO

Na base de dados *trees* do pacote *datasets* presentes no *software* estatístico R, encontram-se informações sobre 31 cerejeiras da Floresta Nacional de Allengheny, relativas a três variáveis: volume de madeira útil, em pés cúbicos (1 Pé = 0,3048 Metros), altura em pés e circunferência do tronco a 4,5 pés de altura (1,37 metros). Este trabalho consistiu em ajustar modelos de regressão linear simples em que a altura das árvores era a variável explicativa para seu volume (ou para transformações deste). Ao final do trabalho, 5 modelos foram ajustados, de modo que os modelos escolhidos pela equipe de análise foram o modelo 5, que aplicava a transformação Box-Cox e, no caso de uma necessidade maior por interpretabilidade, o modelo 3, que explicava o logaritmo natural do volume pela altura.

Palavras-chave: Cerejeiras. Volume. Altura. Modelo linear simples. Transformações.

1 INTRODUÇÃO

Na base de dados *trees* do pacote *datasets* presentes no *software* estatístico R, encontram-se informações sobre 31 cerejeiras da Floresta Nacional de Allengheny, relativas a três variáveis: volume de madeira útil, em pés cúbicos (1 Pé = 0,3048 Metros), altura em pés e circunferência do tronco a 4,5 pés de altura (1,37 metros).

Considerando essas informações, analisou-se a relação entre a variável resposta $Y = Volume$ e a variável explicativa $X = Altura$. Nas análises feitas, desconsideraram-se as informações tangentes à circunferência.

O objetivo deste trabalho, além de analisar os dados mencionados, foi de ajustar diferentes modelos de regressão linear simples considerando algumas transformações na variável resposta, com intuito de entender a relação entre o volume da árvore e sua altura.

Nas seguintes seções, ter-se-ão a análise exploratória dos dados, os modelos ajustados, suas interpretações e algumas considerações finais.

2 ANÁLISE EXPLORATÓRIA

O conjunto de dados continha 31 linhas, com nenhuma informação faltante ou duplicada. A seguir, na Tabela 1 é possível visualizar as estatísticas descritivas do volume e da altura das cerejeiras.

Tabela 1 – Estatísticas descritivas dos dados

Descritiva	Altura	Volume
Mínimo	63,00	10,20
1° Quartil	72,00	19,40
Mediana	76,00	24,20
Média	76,00	30,17
3° Quartil	80,00	37,30
Máximo	87,00	77,00

* 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.

Como é possível visualizar na Tabela 1, a distribuição da variável altura possivelmente é simétrica; média e mediana coincidem e estão igualmente distantes do 1° e 3° quartis. Através da análise visual, essa hipótese foi fortalecida.

Diferentemente da altura, a distribuição do volume não necessariamente se mostrou simétrica na análise visual; no entanto, como são apenas 31 observações, optou-se por não tomar conclusões a respeito disso.

2.1 Análise visual

Após a análise das distribuições das variáveis e de suas estatísticas descritivas, uma análise bivariada considerando volume e altura foi feita. Na Figura 1, é possível observar um diagrama de dispersão envolvendo ambas as variáveis.

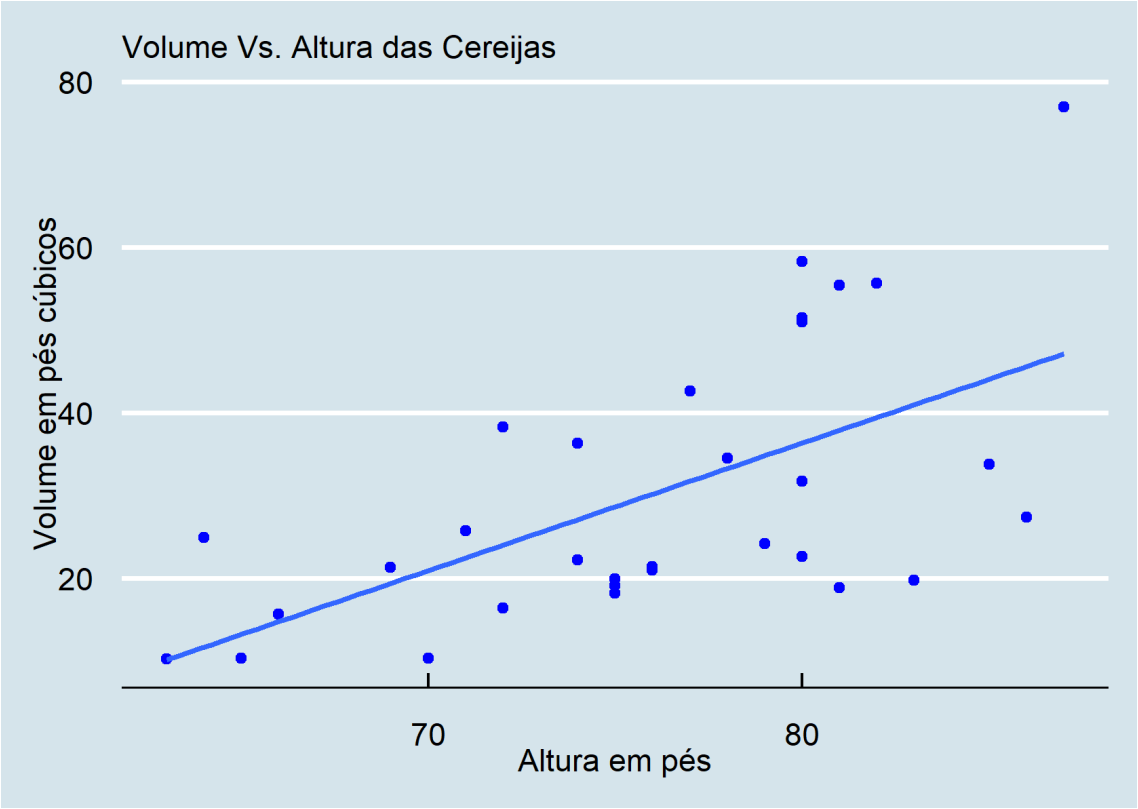

Figura 1 – Diagrama de Dispersão: Volume (pés cúbicos) Vs. Altura (pés) das cerejeiras

Como é possível visualizar pela Figura 1, parece existir, de fato, uma relação crescente entre o volume e a altura das cerejeiras. Tal afirmação foi verificada através de um teste de hipóteses. A correlação pontual foi de aproximadamente 60%, e o teste apontou favoravelmente pela relação entre as variáveis a um nível de 95%.

3 RESULTADOS

Vários modelos de regressão linear simples foram testados, o primeiro considerando a variável Y como estava disposta no banco de dados, e os outros com as seguintes transformações: $\sqrt{Y}$, $log(Y)$, Y^2 e a transformação Box-Cox.

Na Equação 1, é possível observar a forma do primeiro modelo, que será chamado de agora em diante de

modelo 1:

$$\hat{Y} = -87{,}12 + 1{,}54X \tag{1}$$

Sendo, $\hat{Y}$ o valor estimado para o volume, quando a altura for X.

Este modelo indica que a cada acréscimo na altura em pés, o volume da árvore cresceria em média 1,54 pés cúbicos. Vale salientar que o modelo ajustado não deve ser considerado (em termos de interpretação) para alturas menores que 57 pés (pois o valor do intercepto apontaria para volumes negativos, que não existem). A 95% de confiança, a variável altura foi considerada significante, assim como o intercepto. Isto indica que ambos, a altura e o intercepto, são estatisticamente diferentes de 0.

Vale dizer, finalmente, que o coeficiente de determinação ajustado foi igual a: $R^2 = 0{,}3358$, o que indica que a variabilidade explicada pelo modelo foi inferior a 34%.

Na Equação 2, é possível visualizar a forma do modelo 2, que envolve a transformação $\sqrt{Y}$:

$$\hat{Y} = -5{,}27 + 0{,}14X \tag{2}$$

Em que $\hat{Y}$ é o valor estimado para a raiz quadrada do volume quando a altura é X.

O modelo 2 indica que a cada acréscimo na altura em pés, o volume da árvore cresceria em média 0,02 pés cúbicos. Além disso, vale dizer que o modelo não é interpretável para árvores com alturas inferiores a 40 pés.

Assim como no modelo 1, a 95% tanto o intercepto quanto a altura foram considerados significantes.

O coeficiente de determinação ajustado obtido foi: $R^2 = 0{,}3698$, o que indica que pouco menos que 37% da variabilidade do volume é explicada pelo modelo.

Na Equação 3, é possível visualizar o modelo 3, que envolve a transformação $log(Y)$:

$$\hat{Y} = -0{,}8 + 0{,}05X \tag{3}$$

Em que $\hat{Y}$ é o valor estimado para o logaritmo natural do volume quando a altura é X.

O modelo indica que a cada acréscimo da altura em pés, o volume da árvore crescerá em média 1,05 pés cúbicos. Além disso, o modelo não é interpretável para alturas inferiores a 16 pés.

Vale dizer que assim como anteriormente, ao nível de 95%, a altura foi considerada significante. No entanto, o intercepto não era estatisticamente diferente de 0.

O coeficiente de determinação ajustado obtido foi: $R^2 = 0{,}4003$, o que indica que cerca de 40% da variabilidade do volume é explicada pelo modelo.

Na Equação 4, é possível visualizar o modelo 4 que envolve a transformação Y^2:

$$\hat{Y} = -7371{,}17 + 112{,}41X \tag{4}$$

Em que $\hat{Y}$ é o valor estimado para o volume ao quadrado quando a altura é X.

O modelo 4 indica que a cada acréscimo na altura em pés, o volume da árvore cresceria em média 10,60 pés cúbicos. Além disso, vale dizer que o modelo não é interpretável para árvores com alturas inferiores a 66 pés.

Tanto o intercepto como a altura foram significantes a 95% de confiança e o coeficiente de determinação ajustado obtido foi: $R^2 = 0{,}2742$, o que indica que menos de 30% da variabilidade do volume é explicada pelo modelo.

Finalmente, o último modelo ajustado foi o que aplicou a transformação Box-Cox. O parâmetro λ

utilizado foi: $\lambda = -0{,}1818....$ O modelo 5 foi ajustado da seguinte maneira:

$$\hat{Y} = -0{,}2 + 0{,}03X \tag{5}$$

Em que $\hat{Y}$ é o valor estimado para $\frac{Y^{\lambda}-1}{\lambda}$, em que Y é volume quando a altura é X.

O modelo indica que a cada acréscimo da altura em pés, o volume da árvore crescerá em média 1,03 pés cúbicos.

Vale dizer que assim como anteriormente, ao nível de 95%, a altura foi considerada significante. No entanto, o intercepto não era estatisticamente diferente de 0.

O coeficiente de determinação ajustado obtido foi: $R^2 = 0{,}409$, o que indica que cerca de 40% da variabilidade do volume é explicada pelo modelo.

3.1 Análise de resíduos

Na Figura 2, é possível visualizar, em ordem, os gráficos quantil-quantil para os resíduos *Jacknife*. Espera-se que os pontos estejam dentro da banda de confiança (ao nível de 95%). O gráfico foi plotado tendo a distribuição t-Student como base

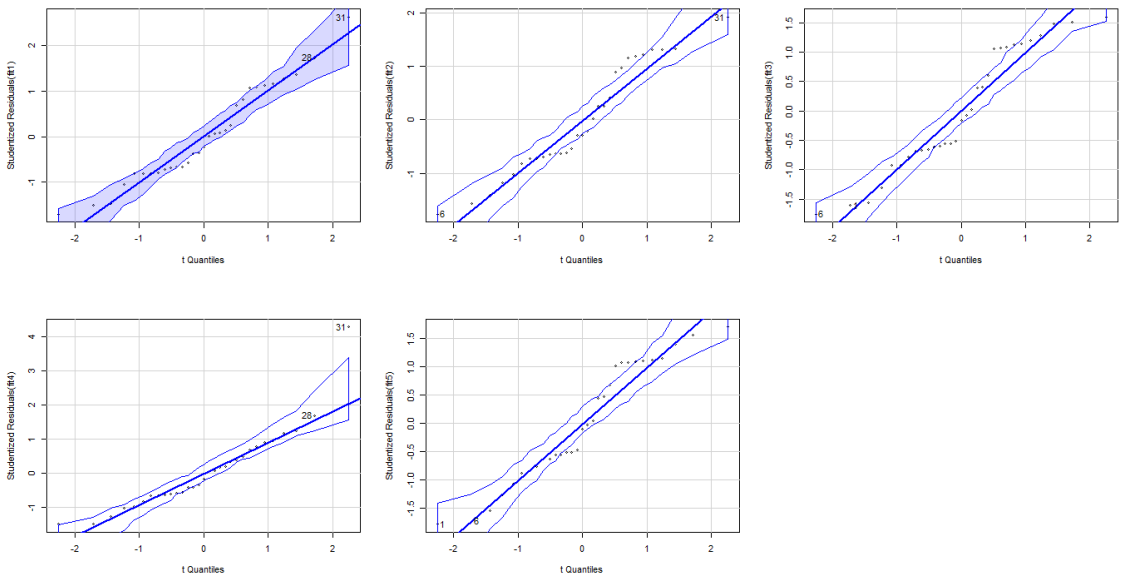

Figura 2 – QQPlots - t-Student - Resíduos *Jacknife*

Observando os gráficos quantil-quantil, aponta-se que provavelmente os resíduos *Jacknife* dos modelos 2 e 4 não seguem a distribuição t-Student. Possivelmente os resíduos dos modelos 3 e 5 também não o façam. Diferentemente dos resíduos do modelo 1, que parecem seguir esta distribuição.

Um teste de hipóteses para checar a normalidade nos resíduos comuns (não os *Jacknife*) foi realizado para cada um dos modelos. Como se suspeitava, apenas o modelo 1 tinha os resíduos seguindo o pressuposto de normalidade ao nível de 95%. Vale dizer, no entanto, que ao nível de 90% os modelos 3 e 5 também apontavam pela normalidade.

Na Figura 3, é possível visualizar o gráfico de resíduos *Jacknife* Vs. Valores ajustados para o modelo 1.

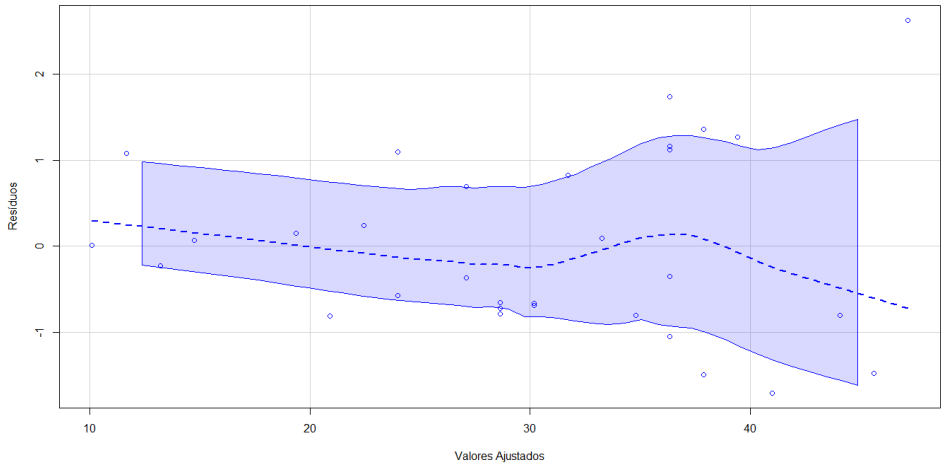

Figura 3 – Resíduos *Jacknife* Vs. Valores Ajustados - Modelo 1

Através do gráfico, é possível perceber indícios de heterocedasticidade, corroborados pelo teste de hipóteses de Goldfeld-Quandt, que apontou para rejeição da homocedasticidade.

Na Figura 4, é possível visualizar o mesmo gráfico, no entanto, agora para os resíduos do modelo 3.

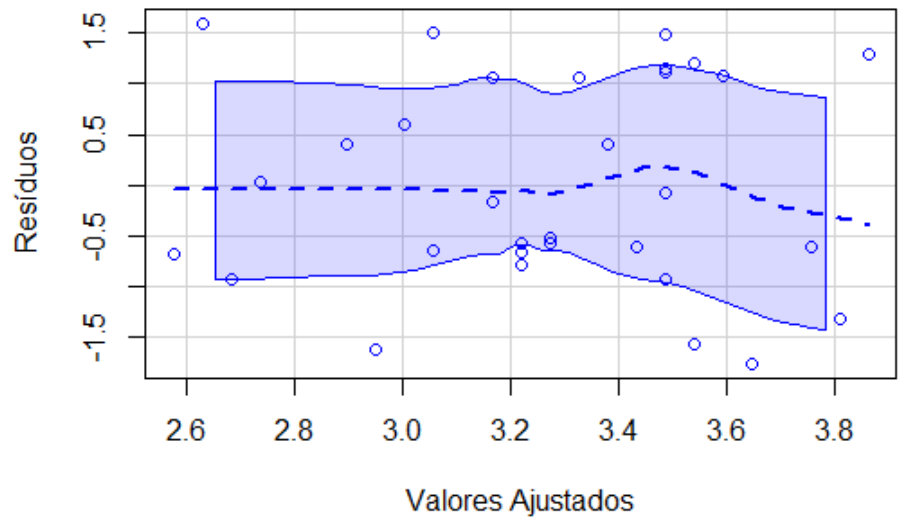

Figura 4 – Resíduos *Jacknife* Vs. Valores Ajustados - Modelo 3

A Figura 4 não parece apontar para uma heterocedasticidade. Apesar de leves desvios do que seria o gráfico "ideal". O teste de hipóteses de Goldfeld-Quandt também foi realizado para este modelo, apontando para a homocedasticidade.

Finalmente, a Figura 5 contém o gráfico de resíduos *Jacknife* por valores ajustados para o modelo 5

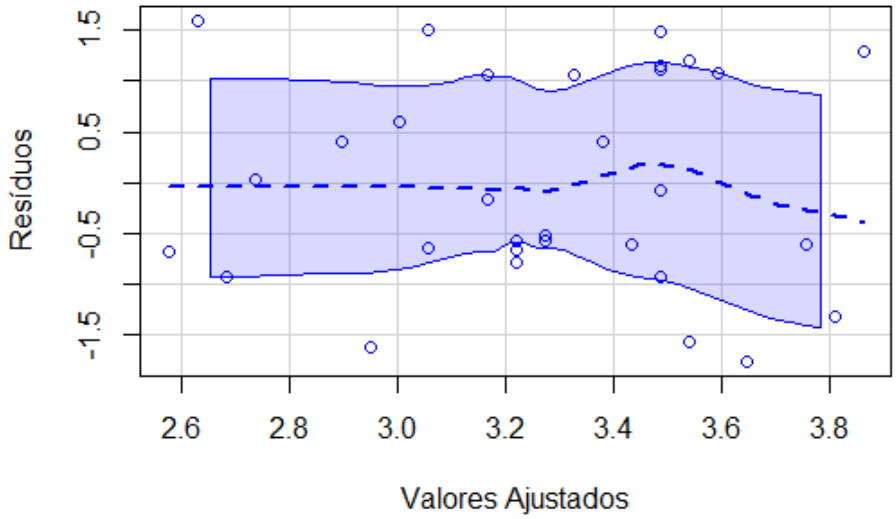

Figura 5 – Resíduos *Jacknife* Vs. Valores Ajustados - Modelo 5

Como é possível observar, o gráfico da Figura 5 também não parece apontar para uma heterocedasticidade e o teste de Goldfeld-Quandt corroborou essa interpretação.

Dentre todos os modelos considerados, a conclusão da equipe foi de que o modelo 5 era um pouco superior ao 3, que era superior ao 1. Os gráficos de resíduos vs. valores ajustados dos outros dois modelos foram omitidos pelo mau resultado deles em relação a normalidade. Vale dizer, no entanto, que estes também não obtiveram bons resultados em relação a homocedasticidade; de modo que não são transformações recomendadas para estes dados.

4 CONSIDERAÇÕES FINAIS

Considerando o mencionado, conclui-se que, possivelmente, os modelos ajustados podem ser melhorados. Uma alternativa seria o modelo de regressão linear múltipla, incluindo mais variáveis explicativas que não somente as alturas das árvores.

Apesar disso, a equipe ajustou os modelos com transformações na variável resposta. Entre eles, optou-se pelo modelo 5, que utilizava a transformação Box-Cox. No entanto, a diferença deste, para o modelo 3 é pequena, de modo que, caso a interpretabilidade fosse um fator importante, o modelo 3 seria o escolhido.

Uma análise mais rebuscada seria interessante. Para tal, recomenda-se fortemente um tamanho maior de amostra e, possivelmente, observações de mais informações relevantes sobre as árvores.

ANEXO A – CÓDIGOS UTILIZADOS PARA ANÁLISE (NO R)

É possível obter o projeto utilizado para análise de dados no R com o *link* a seguir:
https://drive.google.com/drive/folders/1UrVWEFnoXj5GUrwuaWl-CQf2QbelmIEU?usp=sharing

DECLARAÇÃO DE RESPONSABILIDADE

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