LABORATÓRIO 5: Regressão Linear

Fernado Bispo, Jeff Caponero

Sumário

Introdução

O laboratório desta semana está subdividido em duas partes com análises de dois conjuntos de dados distintos que visa a continuidade da aplicação das técnicas de Regressão Linear Simples com a aplicabilidade das técnicas de análise de resíduos e transformação de variáveis inclusive. Para melhor desenvolvimento do processo de analise, este relatório foi dividido em duas partes contendo as análises de cada um dos conjuntos de dados e contando com suas respectivas apresentações sobre o contexto a ser analisado.

Parte 1: Regressão Linear Simples - Diagnóstico do modelo

Metodologia

O primeiro conjunto de dados a ser analisado é denominado *trees*, disponível no pacote *datasets*, contém informações de 31 cerejeiras (*Black cherry*) da Floresta Nacional de Allegheny, relativas a três características numéricas contínuas:

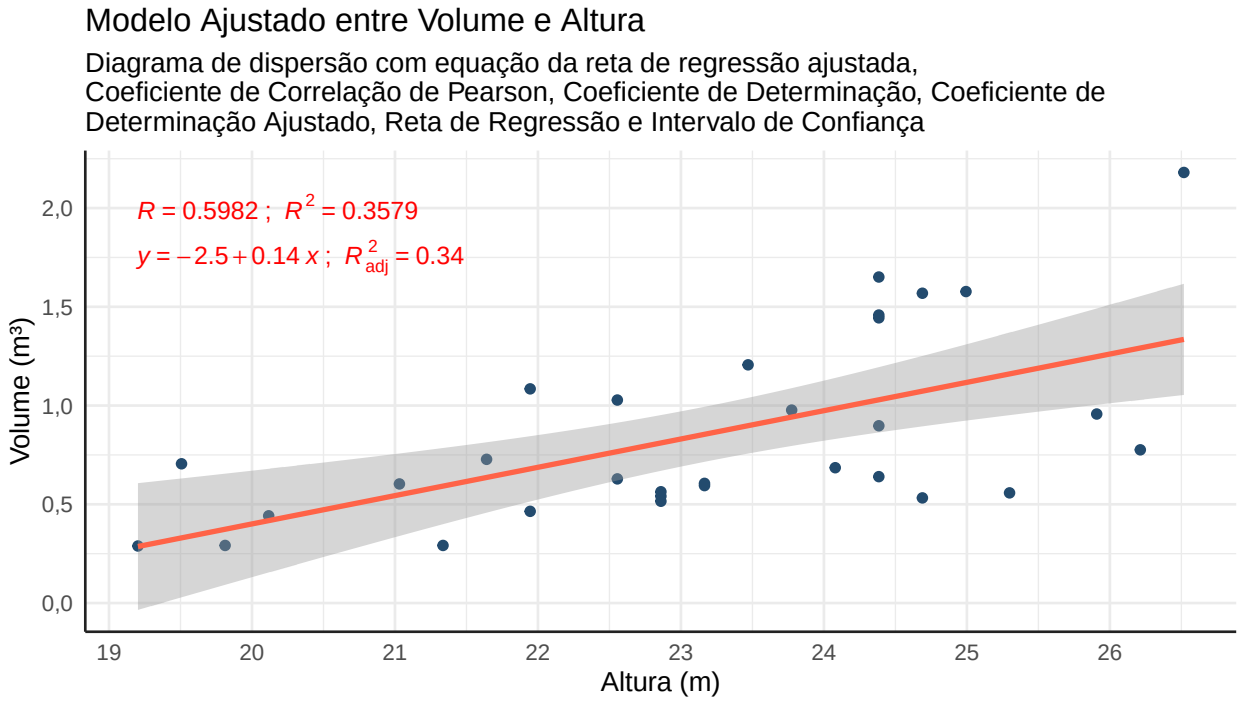
- Volume de madeira útil (em metros cúbicos (m^3));
- Altura (em metros (m));
- Circunferência (em metros(m)) a 1,37 de altura.

Para esta atividade **serão considerados apenas as informações referentes ao volume e altura das árvores**. Com base nestes dados se desenvolverá:

(a) Ajuste um modelo linear simples para volume como função da altura da árvore;
(b) Avalação gráfica dos resíduos Jacknife para diagnóstico do modelo ajustado;
(c) Transformações da característrica utilizada como variável resposta do modelo;
(d) Avaliação da transformação mais apropriada dentro da família proposta por Box e Cox;
(e) Indicação da melhor transformação analisada.

Resultados

Ajuste do Modelo



Com base na Figura 1 é possível sugerir uma relação positiva entre as variáveis **Volume** e **Altura**, fato confirmado pelo Coeficiente de Correlação de Pearson ($R = 0{,}598$) que após o teste de hipótese para avaliar a significância da correlação estimada, demonstrou **possuir correlação não nula**. A reta de regressão ajustada segue a seguinte equação:

$$\hat{Y}_i = -2,5 + 0,14X_i$$

A Tabela 1 traz os resultados do teste de hipóteses para correlação e o Intervalo de Confiança para o verdadeiro valor da correlação, podendo concluir, com base no p-valor menor que o nível de significância ($\alpha = 5\%$), que a hipótese nula (H_0) foi rejeitada, assumindo-se a hipótese alternativa (H_1) que afirma que $\rho \neq 0$.

Tabela 1: Teste de Hipótese para Correlação entre Volume e Altura

	t	p-valor	LI	LS
Altura	4,02051	0,00038	0,30952	0,78598

Nota: Teste realizado com 5% de significância

O Coeficiente de Determinação (R^2) apresenta um valor baixo, podendo afirmar que apenas

aproximadamente 36% da variabilidade dos dados está sendo explicada pelo modelo de regressão calculado.

Significância do Modelo

Após o ajuste do modelo existe a necessidade de se avaliar a significância do mesmo, o teste de hipótese para tal situação será realizado, contendo as seguintes hipóteses:

$$H_0 : \hat{\beta}_1 = 0$$
$$H_1 : \hat{\beta}_1 \neq 0.$$

As Tabelas 4 e 5 trazem os principais resultados da tabela ANOVA e do Intervalo de Confiança para os parâmetros, possibilitando assim inferir sobre o modelo ajustado.

Tabela 2: Análise de Variância (ANOVA)

	GL	Soma de Quadrados	Quadrado Médio	Estatística F-Snedecor	p-valor
Regressão	1	2,326	2,326	16,1645	0,0004
Resíduos	29	4,174	0,144		

Legenda:
[1] GL: Graus de Liberdade

Tabela 3: Intervalos de Confiança para os parâmetros estimados no MRLS.

	LI	LS
Beta 0	-4,162	-0,772
Beta 1	0,070	0,216

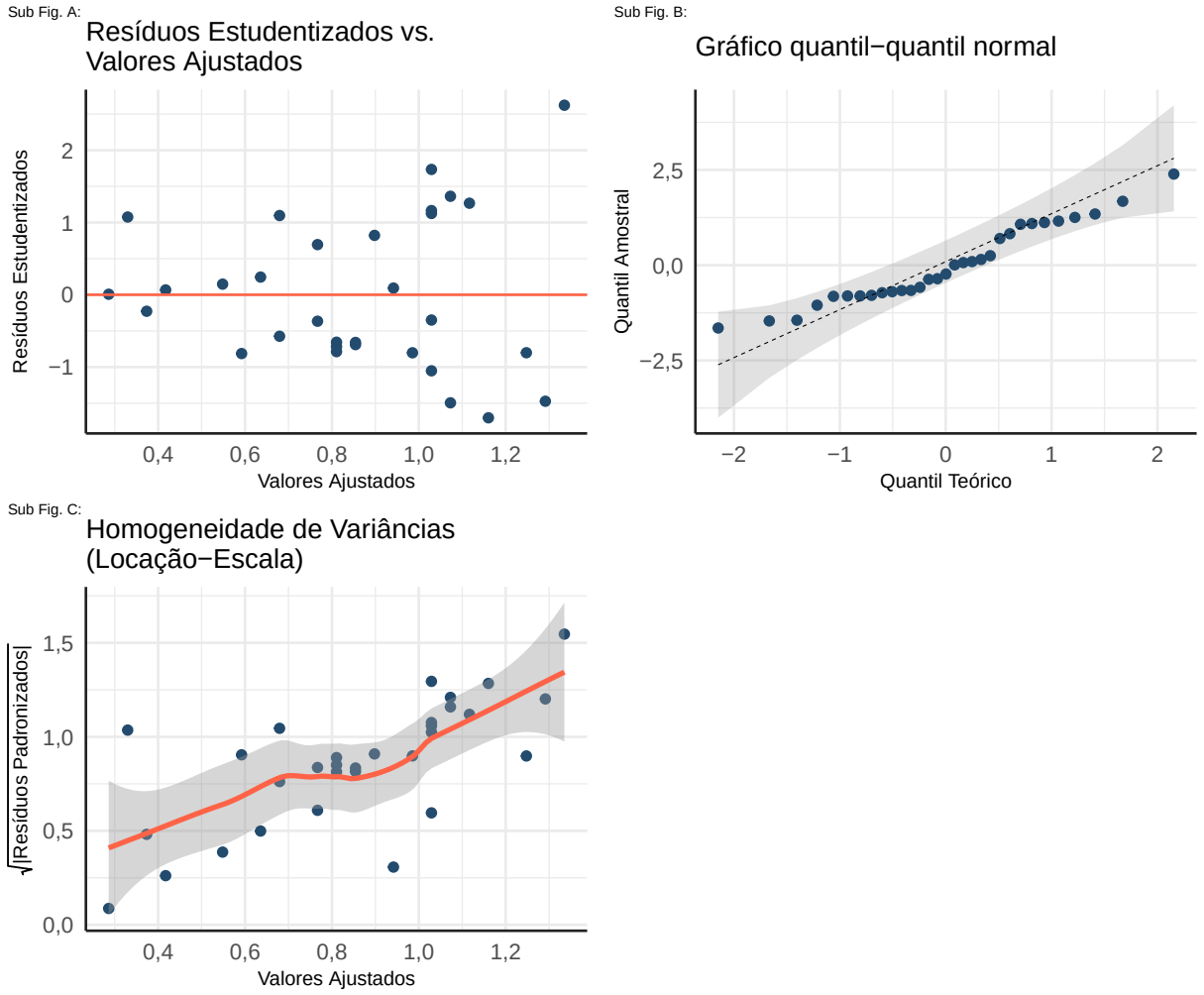
Legenda:
[1] LI: Limite Inferior (2,5%)
[2] LS: Limite Superior (97,5%)
[*] Nível de Significância de 5%.

Com base na Tabela 2, avaliando o p-valor é possível afirmar que o modelo é significante rejeitando assim H_0 que tem como pressuposto $\hat{\beta}_1 = 0$.

Através dos Intervalos de Confiança calculados (Tabela 3) é possível afirmar **com 95% de confiança que o verdadeiro valor de β_0 está entre (-4,1624; -0,7717) e que o verdadeiro valor de β_1 está entre (0,0704; 0,2163)**.

Análise de Resíduos

Figura 2: Análise de resíduos do modelo ajustado



A Figura 2 Sub.Fig A apresenta um comportamento assimétrico dos resíduos, podendo ser constatado uma pequena variabilidade inicial e um aumento desta à medida que os valores ajustados aumentam, caracterizando uma heterocedasticidade. A Sub.Fig C, que trada da Homogeneidade de Variâncias (Locação-Escala) ressalta que há um problema na variabilidade dos dados, corroborando com a interpretação feita na análise da Sub.Fig. A, de que há uma mudança na variabilidade dos dados, caracterizando Heterocedasticidade dos dados. A Sub.Fig. B que traz o gráfico para avaliação da normalidade dos dados, mostra que apesar dos dados não estarem precisamente sobre a reta de referência, os mesmos estão contidos na região pertencente ao Intervalo de Confiança - IC, podendo assumir que há normalidade, contudo tal avaliação será confirmada após os Testes de Diagnóstico.

Testes de Diagnósticos do Modelo

Para avaliar se o modelo atende aos pressupostos, além da análise gráfica podem ser realizados testes de diagnósticos, que são testes de hipóteses para avaliação dos pressupostos que são:

- Normalidade;

  H_0 : Os resíduos possuem normalidade.

  H_1 : Os resíduos **não** possuem normalidade.

- Homoscedasticidade (Homogeneidade de Variância);

  H_0 : Os resíduos possuem variância constante.

  H_1 : Os resíduos **não** possuem variância constante.

- Linearidade;

- Independência.

  H_0: Existe correlação serial entre os resíduos.

  H_1: **Não** existe correlação serial entre os resíduos.

Para tanto serão utilizados os seguintes testes:

- Shapiro-Wilk, para avaliar a Normalidade;
- Breush-Pagan, para avaliar a Homoscedasticidade;
- Durbin-Watson, para avaliar a Independência.

Tabela 4: Testes de Diagnósticos dos Resíduos

	Estatística de teste	p-valor
Shapiro-Wilk	0,95078	0,1640
Breush-Pagan	7,49015	0,0062
Durbin-Watson	0,50090	0,0000

A Tabela 4 traz os testes de diagnósticos realizados para avaliar o modelo de regressão ajustado, conforme análise gráfica dos resíduos há a confirmação da heterocedasticidade de variância conforme o p-valor obtido pelo teste de Breush-Pagan (0,006) bem como a dependência entre as características confirmado pelo p-valor do teste de Durbin-Watson (0,000), em ambos os testes a hipótese nula (H_0) foi rejeitada com base p-valor, como tentativa de contornar a quebra dos pressupostos se faz necessária a transformação da variável resposta.

Transformações dos Dados

Tendo em vista que o modelo não atendeu aos pressupostos se faz necessário a utilização de técnicas para buscar uma melhora de performace do modelo antes da possibilidade de descarte e para tanto algumas transformações são sugeridas, sendo estas:

- $T_1 = \sqrt{Y}$;
- $T_2 = log(Y)$;
- $T_3 = Y^2$.

Sendo Y a variável resposta do modelo representada pelo Volume.

Tabela 5: Medidas Resumo da característica Volume com e sem transformações.

	Mín	Q1	Med	Média	Q3	Máx	Desv.padrão	CV
1. Volume	0,289	0,541	0,685	0,854	1,085	2,180	0,465	0,545
2. √Volume	0,537	0,735	0,828	0,894	1,041	1,477	0,239	0,267
3. log(Volume)	-1,242	-0,615	-0,378	-0,292	0,081	0,780	0,526	-1,805
4. Volume²	0,083	0,293	0,470	0,940	1,176	4,754	1,051	1,119

A Tabela 5 traz a sumarização da característica em analise em sua forma natural juntamente com as formas transformadas para tentar identificar o comportamento destas transformações, podendo constatar um aumento na variabilidade dos dados, com base no valor do Coeficiente de Variação, com exceção da variável sob a transformação Y^2. Seguem as equações das retas ajustadas após a transformação.

$$\sqrt{\hat{Y}_i} = -0,89 + 0,077X_i$$

$$\log(\hat{Y}_i) = -4,4 + 0,18X_i$$

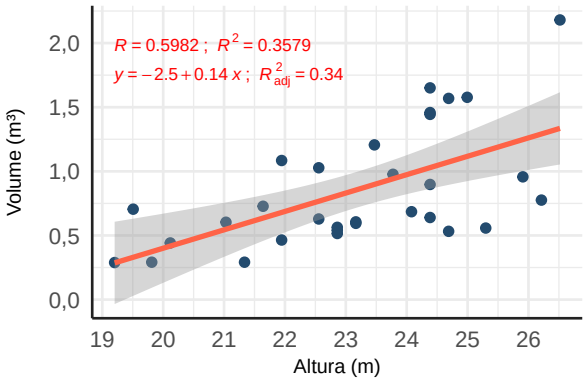
$$\hat{Y}_i^{\,2} = -5,9 + 0,3X_i$$

A Figura 3 traz os diagramas de dispersão com a reta ajustada para o modelo sem transformação e para cada transformação realizada, a fim de possibilitar a identificação das diferenças entre cada modelo.

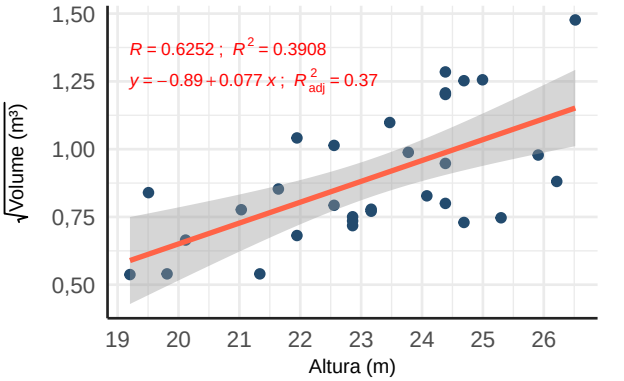
Figura 3: Modelo ajustado e suas transformações

Comparativo entre o modelo ajustado sem transformação com os modelos após a transformações da variável resposta

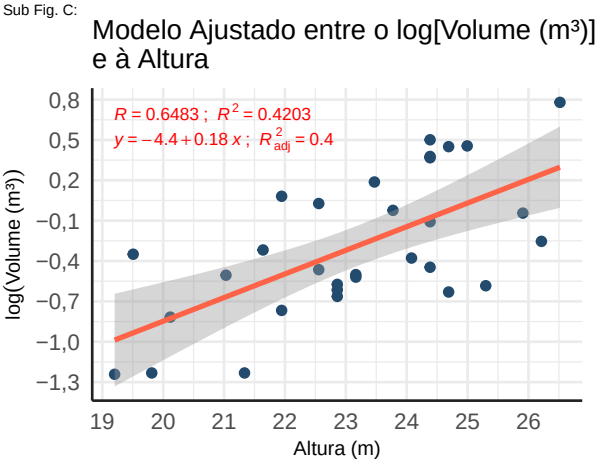
Sub Fig. A:



Sub Fig. B:



Sub Fig. C:



Sub Fig. D:

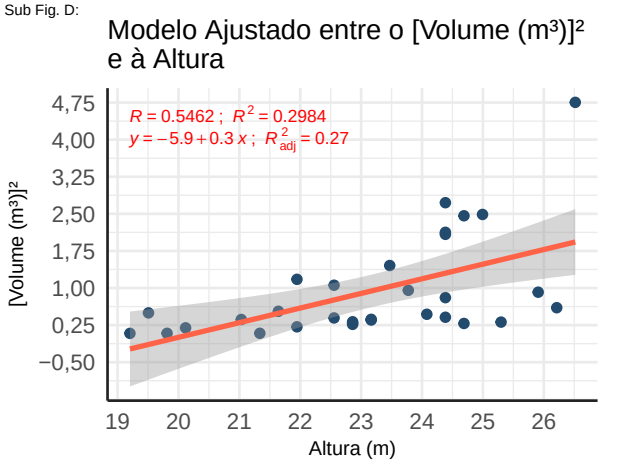
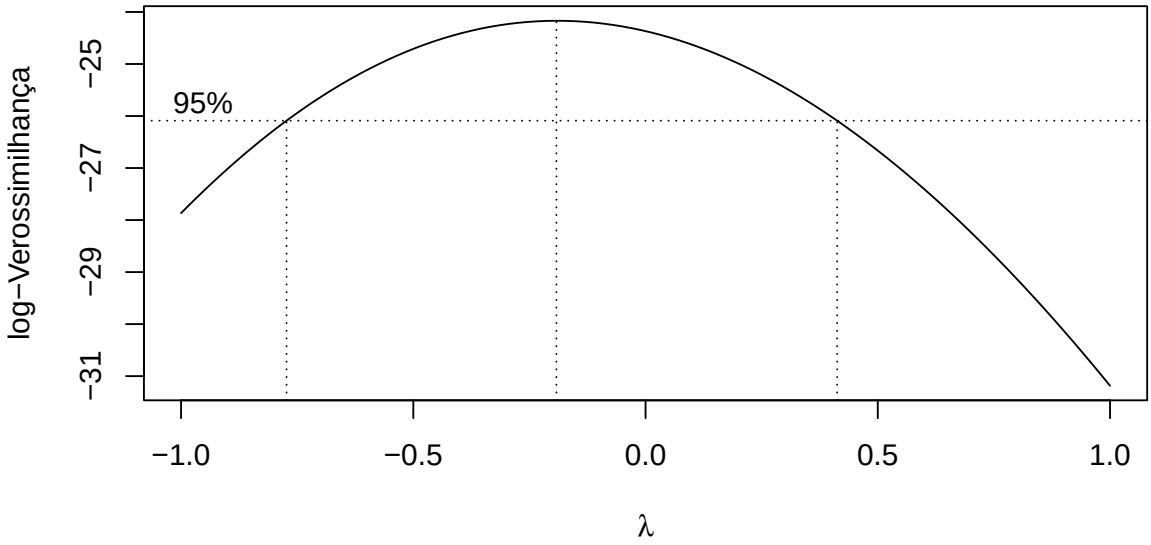


Figura 4: Transformação Box−Cox



Analisando o gráfico das **famílias de transformações Box-Cox** é possível identificar que $-0,5 < \lambda_{max} < 0$ ($\lambda_{max} \approx$ -0,192), partindo do princípio que o valor zero está incluso no intervalo de valores possíveis de λ que minimizam a variância residual, mesmo o zero não sendo o máximo valor assumido, ainda assim, visando a escolha de uma transformação que possibilite uma interpretação facilitada, a escolha da transformação $log(Y)$ torna-se a escolha mais assertiva. Para avaliar esta situação a Figura 3 foi construída.

Com base nos valores do Coeficiente de Determinação R^2 dos gráficos contidos da Figura 3, é possível constatar que o gráfico cuja transformação $log(Y)$ foi aplicada apresentou uma melhora no que diz respeito a possibilidade do modelo explicar a variabilidade dos dados, mesmo não sendo um valor considerado satisfatório, ainda assim é possível constatar a melhora de performance ao aplicar a transformação na característica utilizada como variável resposta.

Conclusão

Após as análises realizadas no modelo ajustado é possível afirmar que para cada aumento de um metro na atura da árvore, há uma redução média de $exp(-4,4)$ m^3 no volume ou aproximadamente 0,012 m^3.

Parte 2: Regressão Linear Múltipla - Estimação pontual

Metodologia

O segundo conjunto de dados a ser analisado, contém informações de 403 afro-americanos residentes no Estado da Virginia (EUA), entrevistados em um estudo referente à prevalência de obesidade, diabetes e outros fatores de risco cardiovasculares. As características apresentadas são:

- Número de identificação do sujeito;
- Colesterol total;
- Glicose estabilizada;
- Lipoproteína de alta densidade (colesterol bom) (hdl);
- Razão colesterol total e colesterol bom (chol/hdl);
- Hemoglobina glicada (glyhb);
- Município de residência (Buckingham ou Louisa);
- Idade em anos;
- Sexo;
- Altura (em polegadas);
- Peso (em libras);
- Pressão sanguínea sistólica (1ª medida) (bp.1s);
- Pressão sanguínea diastólica (1ª medida) (bp.1d);
- pressão sanguínea sistólica (2ª medida) (bp.2s);
- Pressão sanguínea diastólica (2ª medida) (bp.2d);
- Cintura (em polegadas);
- Quadril (em polegadas).

Com base nestes dados se desenvolverá:

(a) Verificação de possíveis inconsistências nos dados, dados ausentes e identificar a escala de cada característica;
(b) Conversão das medidas das características expressas em escalas não usuais ao padrão brasileiro;
(c) Tentar identificar uma variável que poderia ser considerada a variável resposta em uma análise de regressão e na medida do possível, explorar a relação dessa variável com as

demais;

(d) Proposição de novas variáveis, baseadas nas variáveis disponíveis;

(e) Ajuste de um Modelo de Regressão Linear Múltiplo - MRLM, interpretação e verificação da significância dos parâmetros do modelo ajustado (apenas para variáveis quantitativas).

Resultados