

# LABORATÓRIO 5: Regressão Linear

Fernado Bispo, Jeff Caponero

# Sumário

<b>Introdução</b>	<b>2</b>
<b>Parte 1: Regressão Linear Simples - Diagnóstico do modelo</b>	<b>3</b>
Metodologia . . . . .	3
Resultados . . . . .	4
Ajuste do Modelo . . . . .	4
Significância do Modelo . . . . .	5
Análise de Resíduos . . . . .	6
Testes de Diagnósticos do Modelo . . . . .	7
Transformações dos Dados . . . . .	7
Conclusão . . . . .	10
<b>Parte 2: Regressão Linear Múltipla - Estimação pontual</b>	<b>11</b>

# 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 análise, 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 conjunto de dados *trees*, disponível no pacote *datasets*, contém informações de 31 cerejeiras (*Black cherry*) da Floresta Nacional de Allegheny, relativas a três características numéricas contínuas:

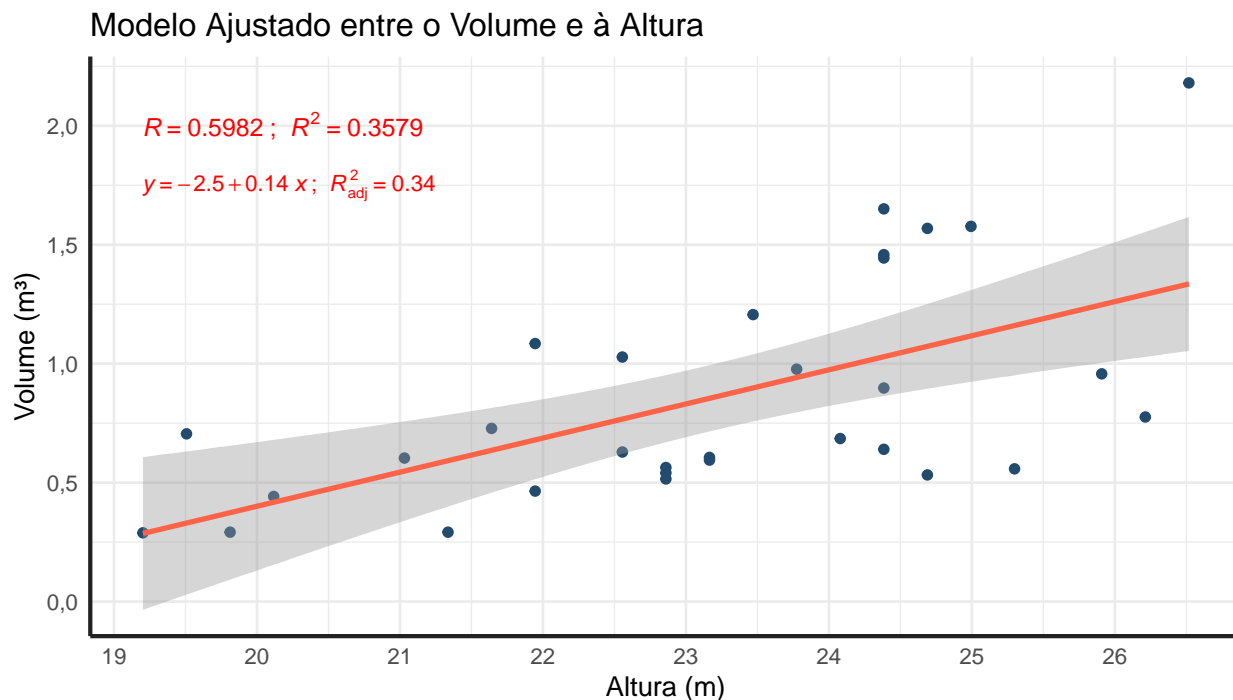
- Volume de madeira útil (em metros cúbicos ( $m^3$ ));
- Altura (em metros (m));
- Circunferência (em metros(m)) a 1,37 de altura.

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) Avaliação gráfica dos resíduos Jackknife para diagnóstico do modelo ajustado;
- (c) Transformações da característica 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**) que após o teste de hipótese para avaliar a significância da correlação, demonstrou possuir correlação não nula. 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  $\rho \neq 0$ .

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.

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

	t	p-valor	LI	LS
<b>Altura</b>	4,02051	0,00038	0,30952	0,78598

*Nota:* Teste realizado com 5% de significância

## 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
<b>Regressão</b>	1	2,326	2,326	16,1645	0,0004
<b>Resíduos</b>	29	4,174	0,144		

*Legenda:*

<sup>1</sup> GL: Graus de Liberdade

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

	LI	LS
<b>Beta 0</b>	-4,162	-0,772
<b>Beta 1</b>	0,070	0,216

*Legenda:*

<sup>1</sup> LI: Limite Inferior (alpha = 2,5%)

<sup>2</sup> LS: Limite Superior (alpha = 97,5%)

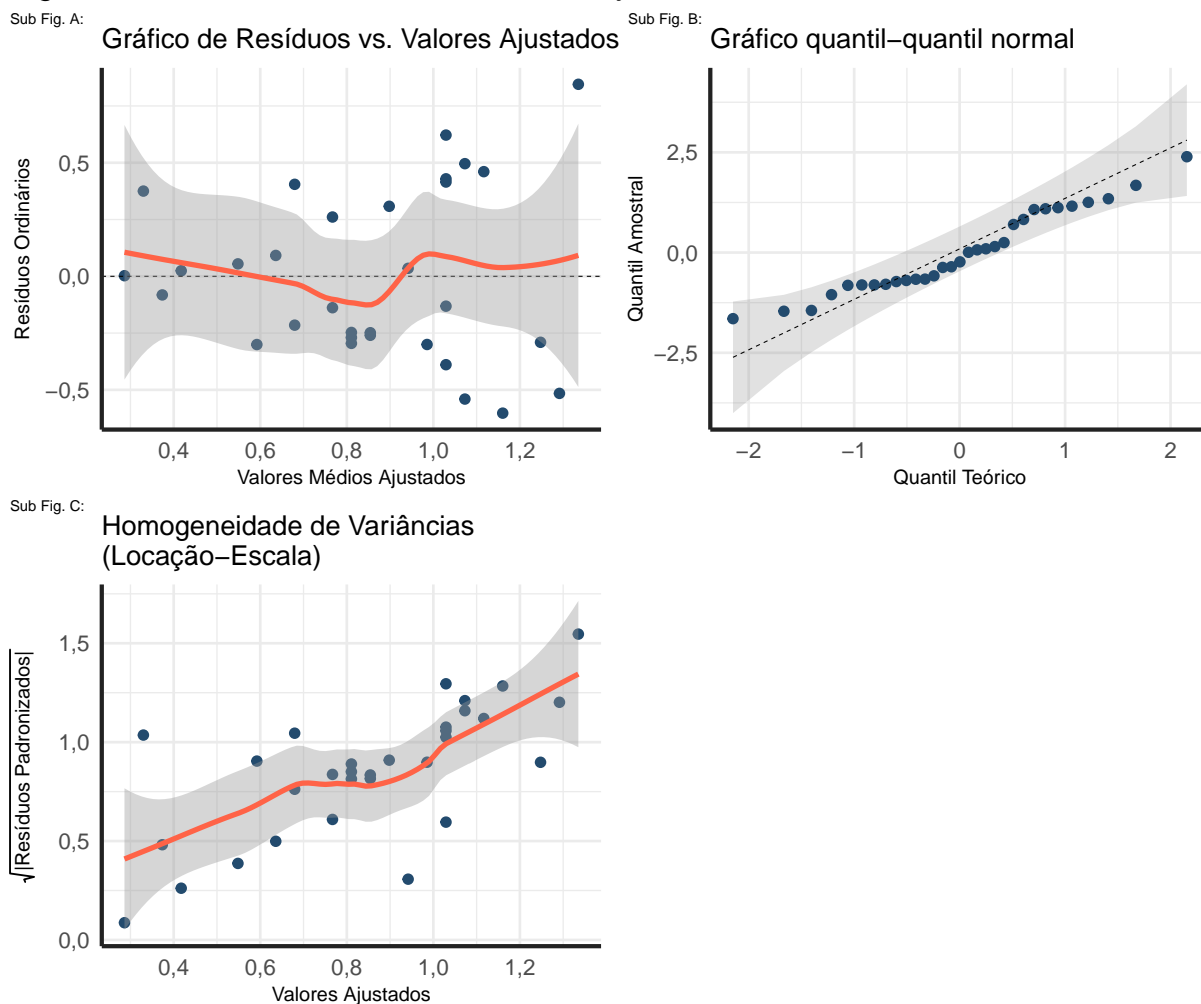
\* Nível de Significância de 5%.

Com base na Tabela 1, avaliando o p-valor é possível afirmar que o modelo é significativo rejeitando assim  $H_0$  que tem como pressuposto  $\hat{\beta}_1 = 0$ .

Através dos Intervalos de Confiança calculados é possível afirmar **com 95% de confiança que o verdadeiro valor de  $\beta_0$  está entre (-4,1624; -0,7717) e que o verdadeiro valor de  $\beta_1$  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 constatar uma pequena variabilidade inicial e um aumento desta à medida que os valores ajustados aumentam, caracterizando uma heterocedasticidade. A Sub.Fig C, que trata da Homogeneidade de Variâncias (Locação-Escala) mostra que há um problema na variabilidade dos dados, corrobora 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
<b>Shapiro-Wilk</b>	0,95078	0,1640
<b>Breush-Pagan</b>	7,49015	0,0062
<b>Durbin-Watson</b>	0,50090	0,0000

## 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:

- $T1 = \sqrt{Y}$ ;
- $T2 = \log(Y)$ ;
- $T3 = Y^2$ .

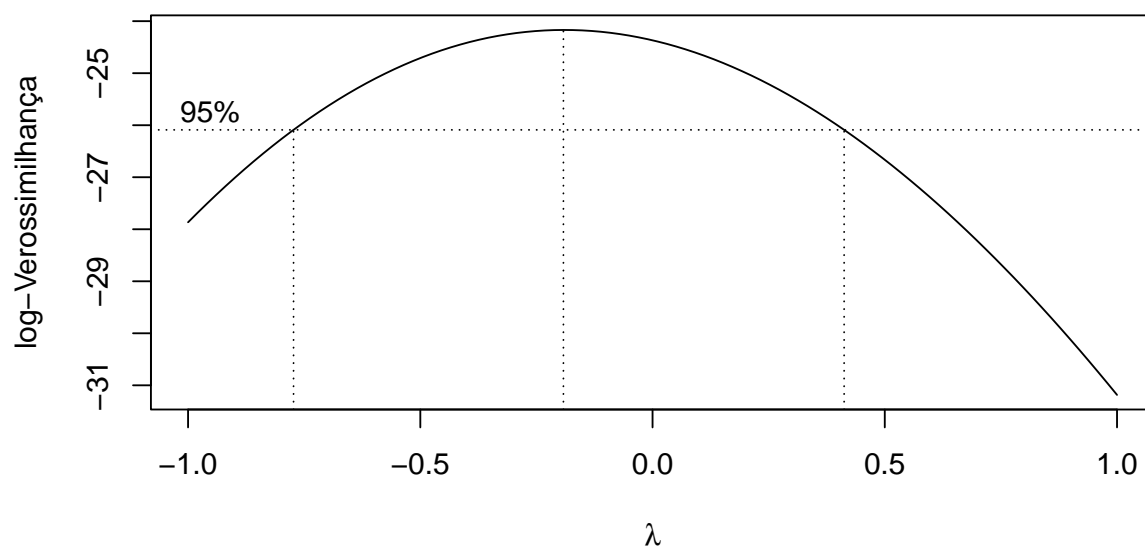
Sendo Y a variável resposta do modelo representada pelo Volume.

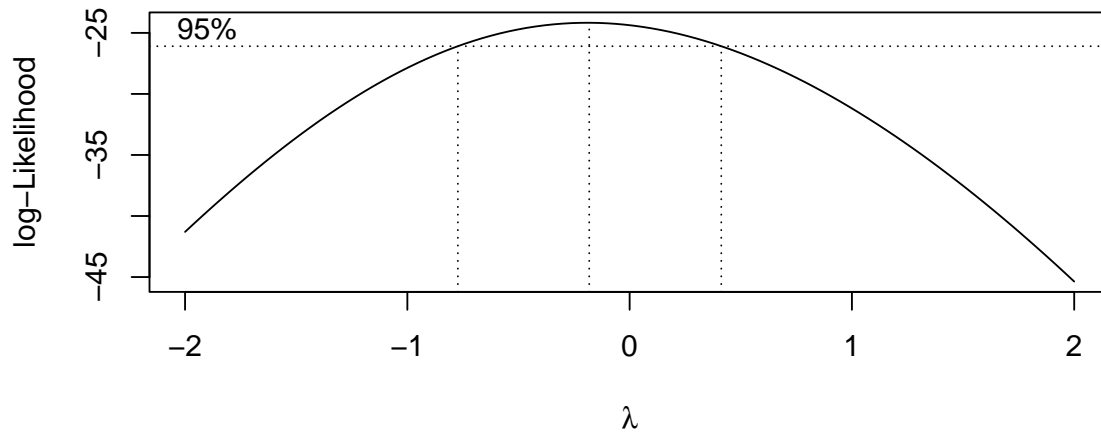


Tabela 5: Medidas Resumo da característica Volume com e sem transformações.

	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. $\sqrt{\text{Volume}}$	0,537	0,735	0,828	0,894	1,041	1,477	0,239	0,267
3. $\log(\text{Volume})$	-1,242	-0,615	-0,378	-0,292	0,081	0,780	0,526	-1,805
4. $\text{Volume}^2$	0,083	0,293	0,470	0,940	1,176	4,754	1,051	1,119

Figura 4: Transformação Box-Cox



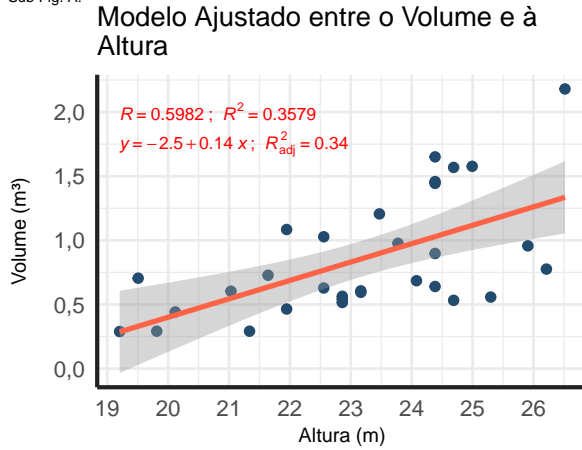


Analisando o gráfico das **famílias de transformações Box-Cox** é possível identificar que  $-0,5 < \lambda_{mx} < 0$  ( $\lambda_{\text{máx}} = -0,182$ ), partindo do princípio que o valor zero está incluso no intervalo de valores possíveis de  $\lambda$  que minimizam a variância residual, mesmo o 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.

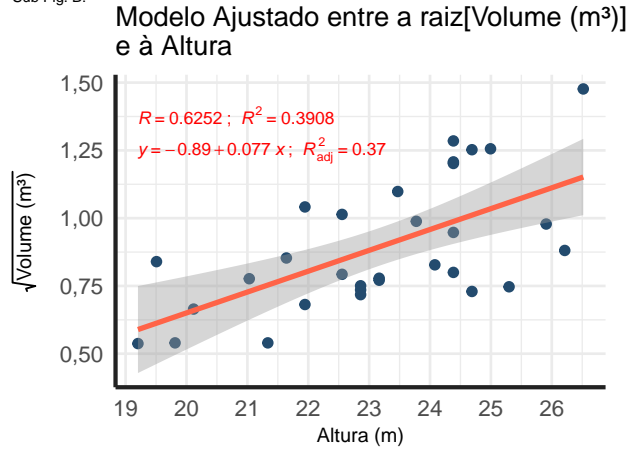
### Figura 3: Modelo ajustado e suas transformações

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

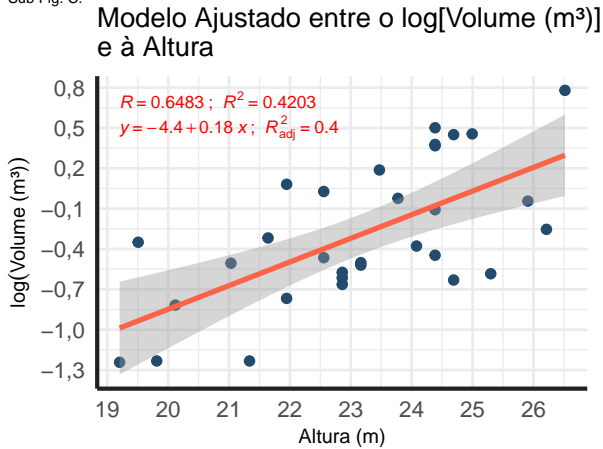
Sub Fig. A:



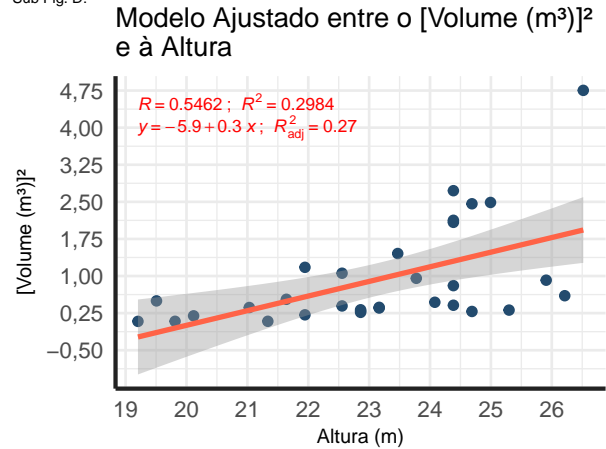
Sub Fig. B:



Sub Fig. C:



Sub Fig. D:



Com base nos valores do Coeficiente de Determinação  $R^2$  dos gráficos 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 1 metro na altura da árvore, há uma redução média de aproximadamente  $\exp(-4,4)$  m³ no volume ou aproximadamente 0,012 m³ no volume da árvore.

## Parte 2: Regressão Linear Múltipla - Estimação pontual