LABORATÓRIO 5: REGRESSÃO LINEAR SIMPLES - ANÁLISE DE RESÍDUOS E TRANSFORMAÇÕES

Fernando Bispo

Sumário

Introdução

O laboratório desta semana visa a continuidade da aplicação das técnicas de Regressão Linear Simples - RLS com a aplicabilidade das técnicas de análise de resíduos e transformação de variáveis para a condição de quebra dos pressupostos do modelo.

Metodologia

O 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 que tiveram suas unidades de medidas convertidas do padrão americano para o Sistema Internacional - SI:

- 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:

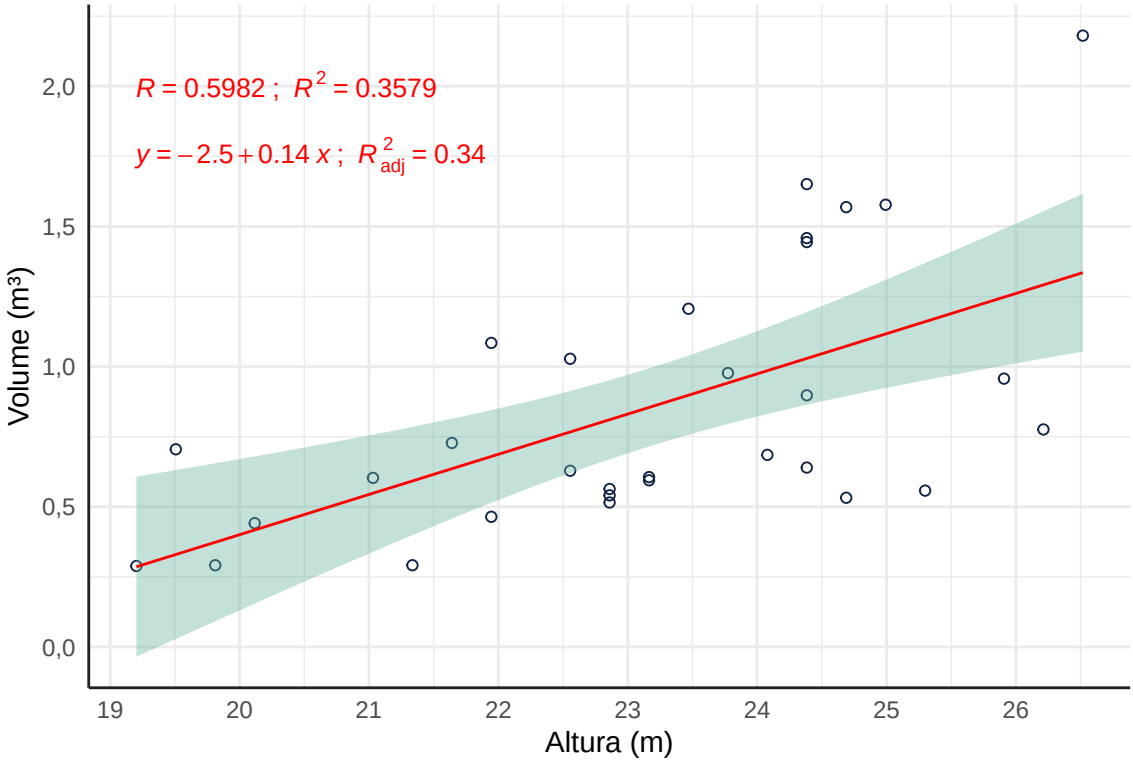
(a) Ajuste de Modelo Regressão Linear Simples - MRLS para volume como função da altura da árvore;
(b) Avalação gráfica dos resíduos Jacknife para diagnóstico do modelo ajustado;
(c) Transformações da característrica utilizada como variável resposta do modelo;
(d) Avaliação da transformação mais apropriada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

Resultados

Item a: Ajuste do MRLS

Figura 1: Modelo Ajustado entre Volume e Altura

Diagrama de dispersão com equação da reta de regressão ajustada



OBS.: Para facilitar a interpretação estão inseridas as seguintes estimativas: Coeficiente de Correlação Linear de Pearson, Coeficiente de Determinação, Coeficiente de Determinação Ajustado, Reta de Regressão e Intervalo de Confiança.

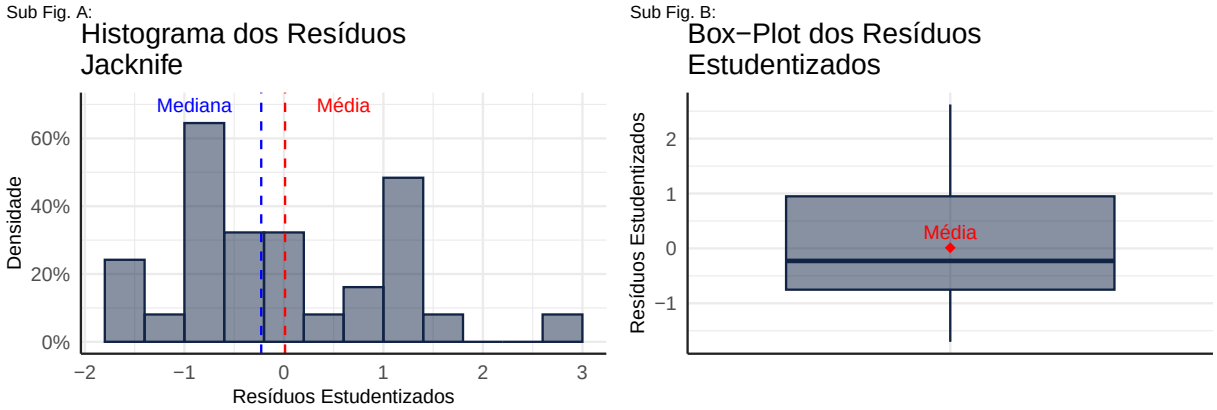
Com base na Figura 1 é possível identificar uma aparente relação positiva entre as variáveis **Volume** e **Altura**, fato indicado pela estimação do Coeficiente de Correlação Linear de Pearson ($\widehat{R} = 0,598$) .

A reta de regressão ajustada é dada pela seguinte equação:

$$\widehat{Y_i} = -2,467 + 0,143X_i; i = 1, \ldots, n.$$

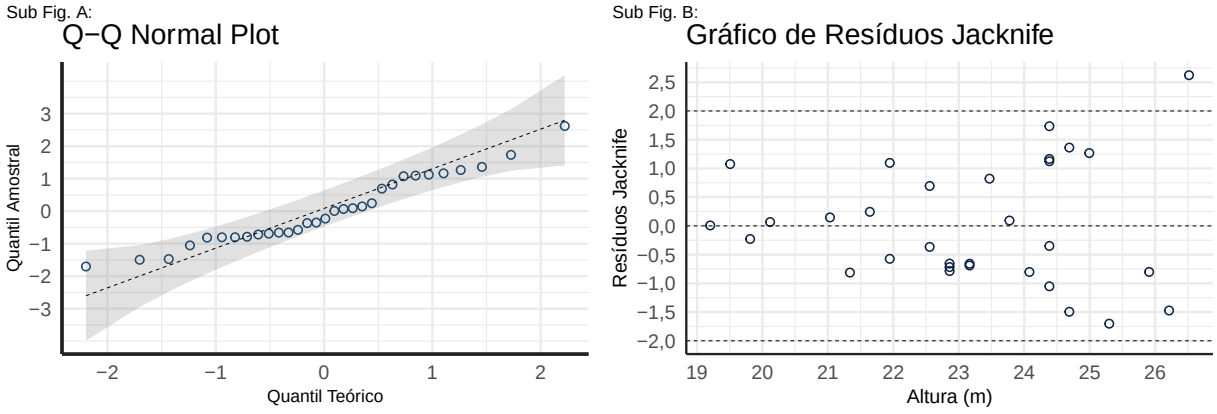
Item b: Diagnóstico do Modelo.

Figura 2: Avaliação do comportamento dos resíduos.



Com base na Figura 2 é possível constatar que os resíduos apresentam um comportamento assimétrico, fato esse identificado tanto no histograma (Sub. Fig. A) quanto no *box-plot* (Sub. Fig. B), percebe-se também uma aparente variabilidade acentuada nos resíduos, fato a ser confirmado nas analises a seguir.

Figura 3: Avaliação dos pressupostos do modelo ajustado.



A análise dos pressupostos é fundamental para avaliar se a adequação do modelo é satisfatória. Observando a Figura 3, Sub. Fig. A, figura que avalia a condição de normalidade, percebe-se um comportamento que foge a reta de referencia na região central da figura, o que dificulta bastante uma possível correção, contudo os pontos permanecem dentro da região de confiança, ainda assim, não se rejeita esse pressuposto.

A Sub. Fig. B traz o gráfico dos resíduos estudentizados vs a variável explicativa e através dele se constata que não há homogeneidade de variâncias, ou seja, a medida que a variável explicativa aumenta há um aumento na dispersão dos dados, caracterizando assim um comportamento heterocedástico, além da presença de um ponto atípico no extremo canto superior direito da figura. para contribuir com esta interpretação segue a Tabela 1 com os resultados dos testes não paramétricos.

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.

- **Independência**

  H_0: Existe correlação serial entre os resíduos.

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

- **Linearidade**

  H_0 : **Não** há falta de ajuste (a regressão de fato é linear)

  H_1 : Há falta de ajuste (a regressão de fato não é linear)

Para tanto serão utilizados os seguintes testes:

- *Kolmogorov−Smirnov*, para avaliar a Normalidade;
- *Breush-Pagan*, para avaliar a Homoscedasticidade;
- *Durbin-Watson*, para avaliar a Independência.

Tabela 1: Testes de Diagnósticos dos Resíduos

	Estatística de teste	p-valor
Kolmogorov−Smirnov	0,1373	0,5565
Breush-Pagan	12,2068	0,0005
Durbin-Watson	0,5009	0,0000

A Tabela 1 traz os resultados dos testes não paramétricos realizados para avaliação dos pressupostos do modelo de regressão ajustado.

Para a avaliação do pressuposto da normalidade, o teste de *Kolmogorov−Smirnov* não trouxe indícios para rejeitar a hipótese nula H_0 a um nível de 5%, com base no p-valor, contudo, conforme a prévia análise gráfica, há a confirmação da heterocedasticidade conforme o teste de *Breush-Pagan* (p-valor = 0,00048) bem como da dependência entre as características, confirmado pelo teste de *Durbin-Watson* (p-valor = 0,0000001) em ambos os testes rejeita-se a hipótese nula (H_0) a um nível de 5%, com base p-valor.

Como tentativa de contornar a quebra dos pressupostos se faz necessária a utilização de algumas técnicas sendo uma destas a técnica de transformação da variável resposta e uma nova avaliação.

Item c: 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.

Após as transformações das variáveis acima descritas, seguem os MRLS ajustados:

$$\widehat{T_1} = -0,888 + 0,077X_i; i = 1, ..., n.$$

$$\widehat{T_2} = -4,361 + 0,176X_i; i = 1, ..., n.$$

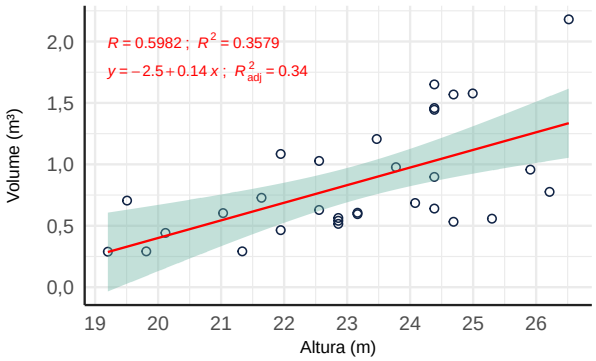
$$\widehat{T_3} = -5,910 + 0,296X_i; i = 1, ..., n.$$

A Figura 4 traz os diagramas de dispersão com a reta ajustada para o MRLS sem transformação (Sub. Fig. A) e para cada transformação realizada, a fim de possibilitar uma melhor identificação das diferenças entre cada modelo.

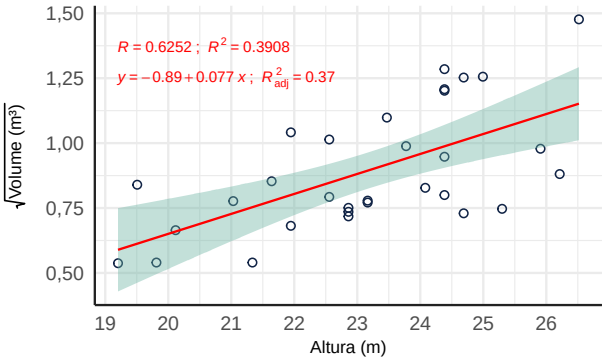
Figura 4: Modelo ajustado e suas transformações

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

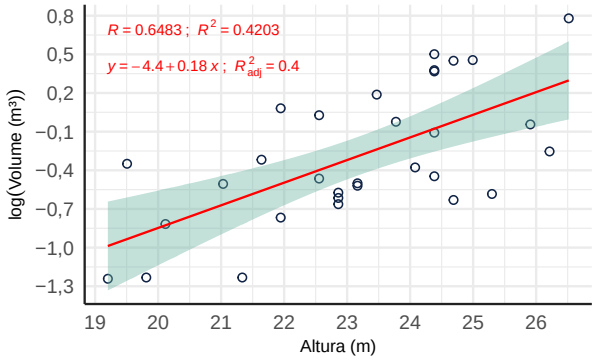
Sub Fig. A:



Sub Fig. B:



Sub Fig. C:



Sub Fig. D:

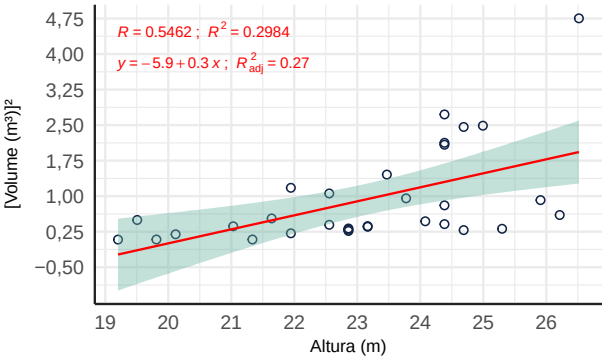
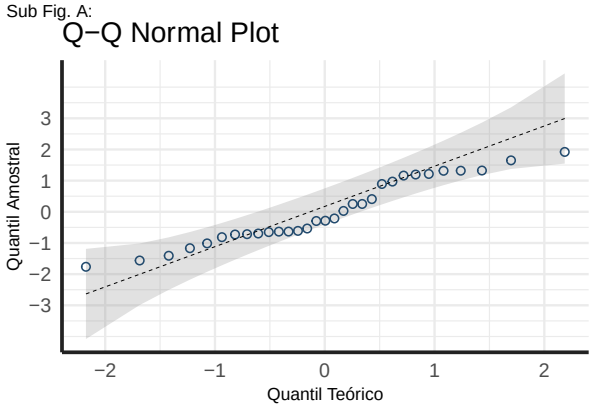
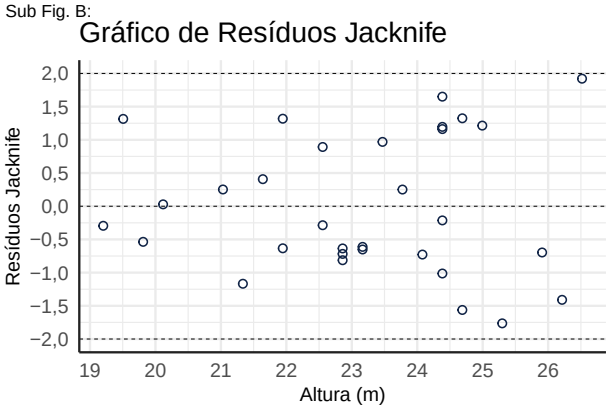


Figura 5: Avaliação dos pressupostos do modelo T1 ajustado

Sub Fig. A:

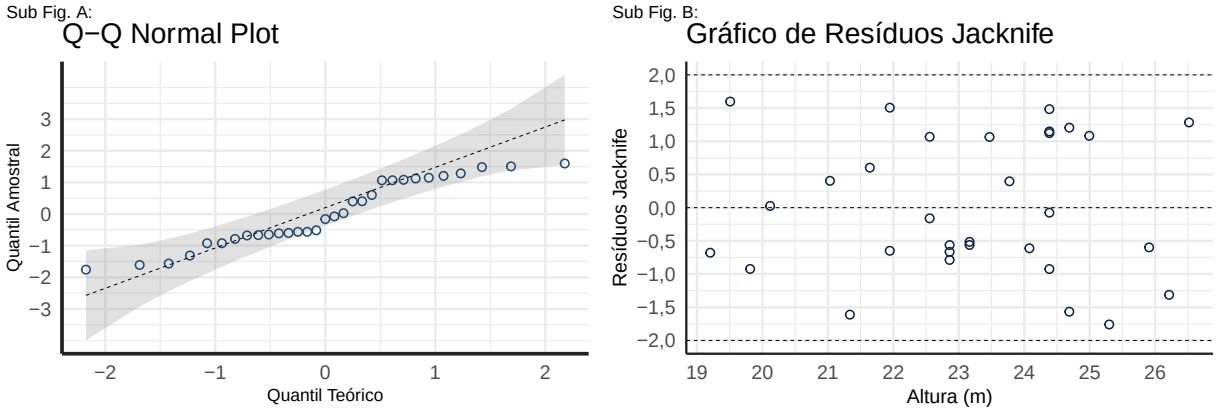


Sub Fig. B:



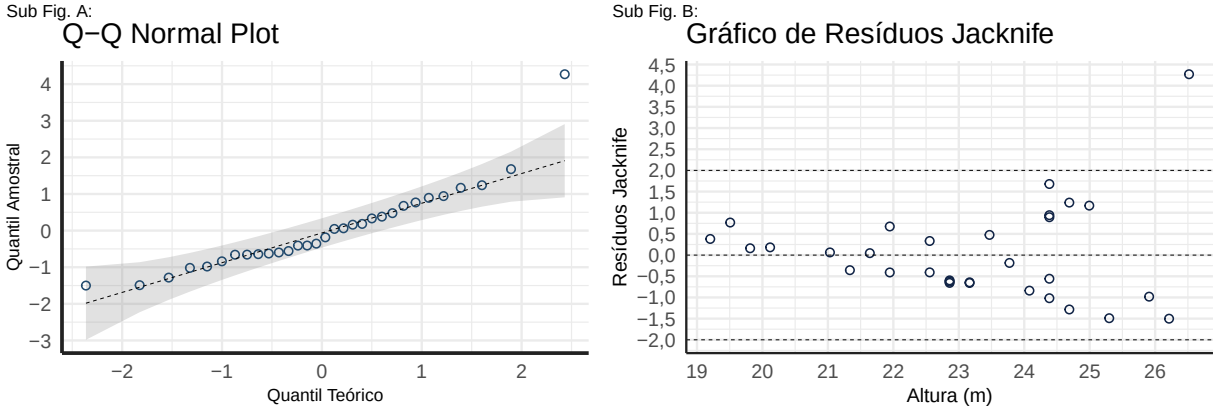
É possível constatar uma melhora no pressuposto de homoscedasticidade para os resíduos do modelo T1, mas também houve uma maior acentuação do desvio na região central do gráfico de normalidade, contudo os valores permanecem dentro da região de confiança.

Figura 6: Avaliação dos pressupostos do modelo T2 ajustado



Assim como no modelo anterior, o modelo T2 apresenta uma melhora no pressuposto de homoscedasticidade para os resíduo, mas também houve uma maior acentuação do desvio na região central do gráfico de normalidade, contudo os valores permanecem dentro da região de confiança.

Figura 7: Avaliação dos pressupostos do modelo T3 ajustado



Para o modelo T3 se constata a evidenciação de um ponto atípico (*outlier*), sendo este influente o suficiente para interferir na variabilidade total deste modelo, portanto este modelo transformado foi descartado pela quebra do pressuposto da homogeneidade de variâncias. Dentre os modelos sob a transformação restantes ($\sqrt{Y}$ e $log(Y)$), o modelo sob a transformação $log(Y)$ (T2) apresenta uma maior homogeneidade de variância, sendo o mais adequado dentre os modelos transformados.

Como forma de corroborar com a avaliação feita sobre a análise gráfica, foi construída a Tabela 2 com os testes de diagnósticos dos resíduos do modelo sob a transformação escolhida ($log(Y)$).

Tabela 2: Testes de Diagnósticos dos Resíduos após transformação log(Y)

	Estatística de teste	p-valor
Kolmogorov−Smirnov	0,1832	0,2204
Breush-Pagan	1,4156	0,2341
Durbin-Watson	0,5066	0,0000

Baseado nos testes de diagnóstico para a o MRLS após a transformação da variável resposta (Y), se conclui, com 5% de significância, que o modelo é homocedástico.

Item d: Box e Cox

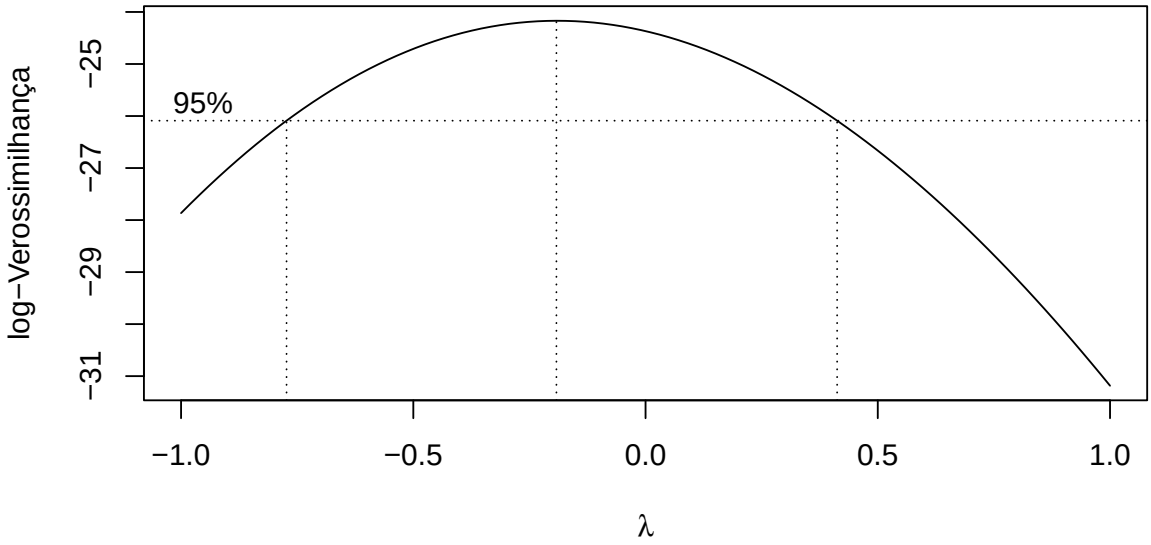
Tendo em vista não ser a tarefa mais simples a aplicação de diversas transformações e os devidos testes para avaliar o melhor possível modelo a ser utilizado, que minimize a variância residual, a opção mais adequada é a escolha do modelo baseado na família de transformações de Box-Cox, definida por:

$$Y(\lambda) = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda}, \text{ se } \lambda \neq 0 \\ log(Y), \text{ se } \lambda = 0 \end{cases}$$

sendo λ o parâmetro da transformação.

Para tanto a Figura 5 foi construída com base na função de λ para a escolha da transformação apropriada.

Figura 5: 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**, conforme conclusão anteriormente feita baseado na análise gráfica.

Item e: Conclusão

Após as análises realizadas sobre modelo ajustado foi possível constatar uma melhora no desempenho deste, quanto a variabilidade dos dados, após a transformação da variável resposta para a transformação $log(Y)$, fato constatado na Figura 3 por meio do valor do Coeficiente de Determinação R^2 bem como através da análise dos resíduos (Figura 4).

Apesar de não ter conseguido um modelo que preenchesse todos os pressupostos, ainda assim, na possibilidade desse modelo ter sido satisfatório, poder-se-ia interpretar a sua utilização da seguinte forma: Para cada aumento de um metro na altura da árvore, há um aumento médio de $exp(0,18)\ m^3$ no volume ou aproximadamente 1,197 m^3.