Modelos Lineares Generalizados

Adriana Eva Fernandes

Descrição do Conjunto de Dados

Para este conjunto de dados foi realizado um estudo que consiste em encontrar uma estimativa para o **volume** de uma árvore (e, portanto, o rendimento de madeira), dada a sua **altura** e **diâmetro**. Os dados fornecem o volume (pés cúbicos), altura (pés) e diâmetro (polegadas) (a 54 polegadas acima do solo), o tamanho amostral é de 31 cerejeiras pretas na Floresta Nacional de Allegheny, Pensilvânia.

Descrição do conjunto de Dados

Os dados estão dispostos na Tabela 1. A dimensão da base de dados é definida por 31 linhas e 3 colunas. Além disso, as variáveis são quantitativas contínuas.

Tabela 1: Tabela dos dados

ID	Diâmetro	Altura	Volume
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
.	.	.	.
.	.	.	.
.	.	.	.
28	17.9	80	58.3
29	18	80	51.5
30	18	80	51
31	20.6	87	77

Objetivo da análise realizada

O presente estudo se objetiva em estimar a variável resposta (ou variável de interesse) **volume** de uma árvore. por meio das covariáveis **altura** e **diâmetro**. Nesse sentido, será aplicado um método de Modelos Lineares Generalizados para realizar essa estimação.

Justificativa da análise a ser realizada

Apesar de poderosa, a Análise de Regressão Linear exige fortes suposições para sua utilização, como normalidade, independência e homoscedasticidade dos erros. Como os nossos dados são assimétricos positivos, esperamos que, ao menos uma das suposições do modelo de Regressão Linear seja violada, assim, a partir das limitações deste modelo, aplicamos a metodologia de Modelos Lineares Generalizados.

Análise Descritiva

A Tabela 2 contém as medidas resumo das variáveis **Diâmetro**, **Altura** e **Volume**.

Tabela 2: Tabela dos dados

	Diâmetro	**Altura**	**Volume**
Média	13.24	76.0	30.17
Desvio Padrão	3.13	6.37	16.43
Mínimo	8.30	63.0	10.20
1° Quartil	11.05	72.0	19.40
Mediana	12.90	76.0	24.20
3° Quartil	15.25	80.0	37.30
Máximo	20.60	87.0	77.0
CV	0.23	0.08	0.54

Análise Descritiva

Podemos perceber pela Figura 1 que a mediana e a média do **Diâmetro** e do **Volume** são diferentes, essa informação é sustentada pela Tabela 2. Assim, temos um indicativo de que os dados são assimétricos.

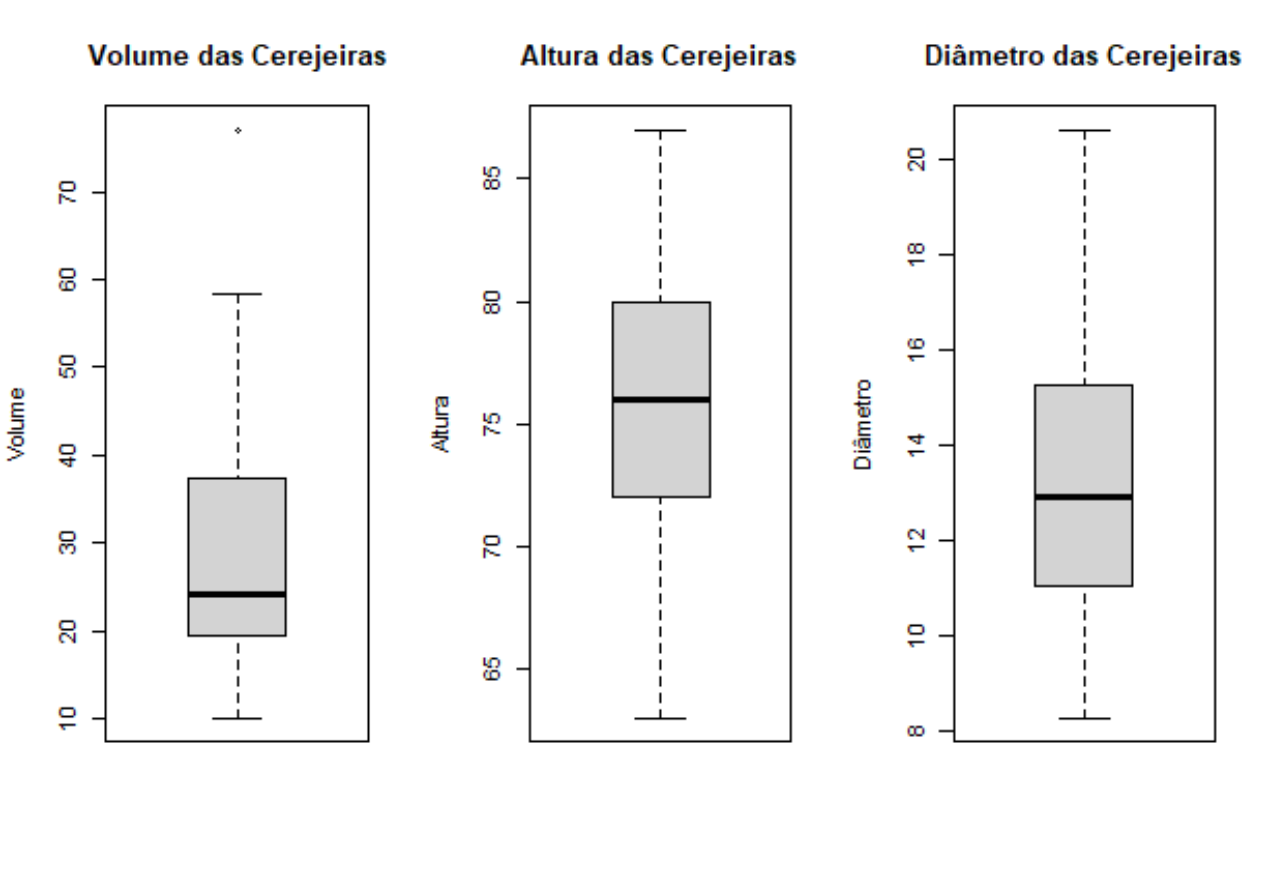

Figura 1: Boxplot do Volume, Altura e Diâmetro das cerejeiras

Análise Descritiva

Dado que temos um indicativo de que os dados são assimétricos, vamos realizar histogramas para investigar mais indicíos de assimetria dos dados. A partir da Figura 2 temos um indicativo de que os dados do **Diâmetro** são assimétricos.

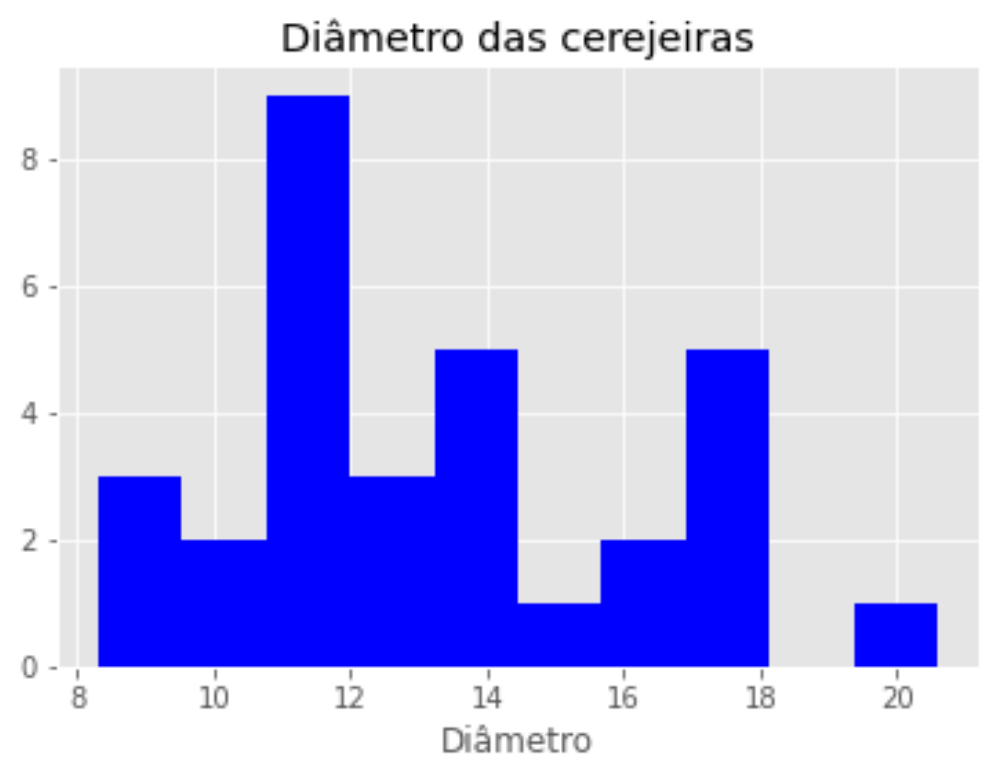

Figura 2: Histograma do Diâmetro das cerejeiras

Análise Descritiva

Por meio da Figura 3 temos um indicativo de que os dados da **Altura** são assimétricos.

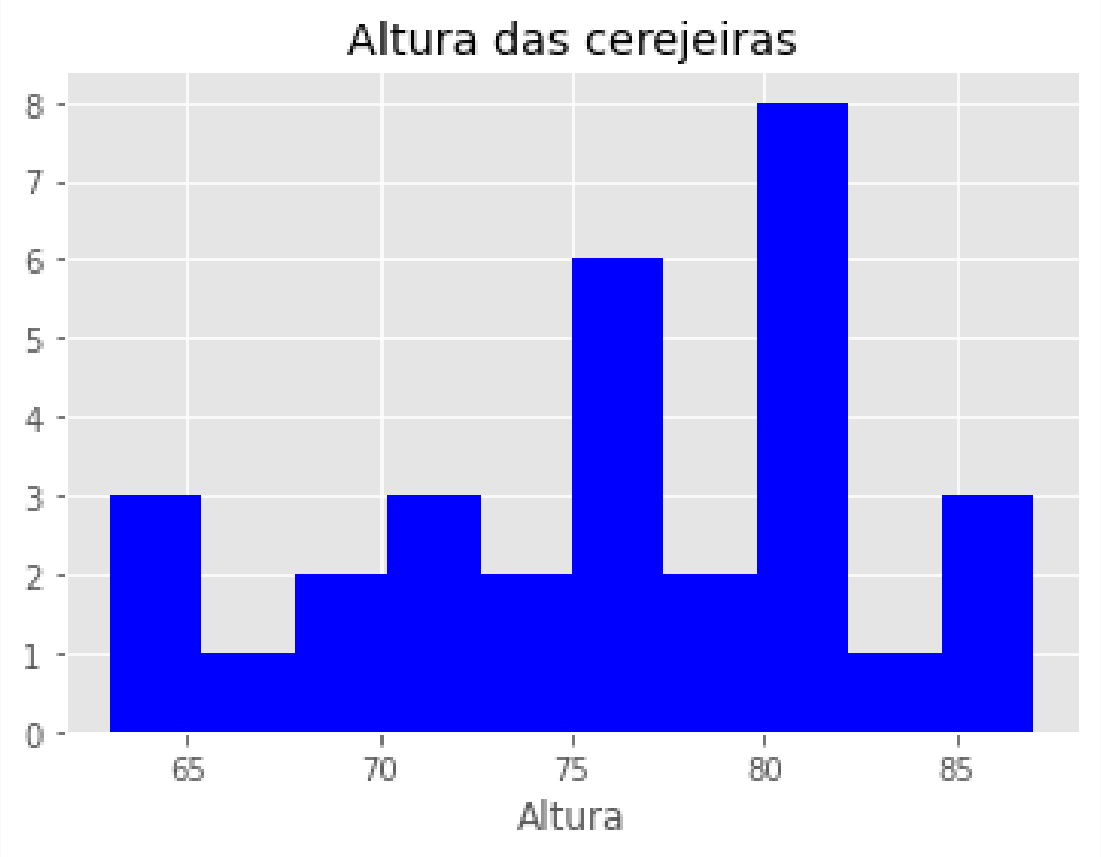

Figura 3: Histograma do Altura das cerejeiras

Análise Descritiva

Com base na Figura 4 temos um indicativo de que os dados de **Volume** são assimétricos.

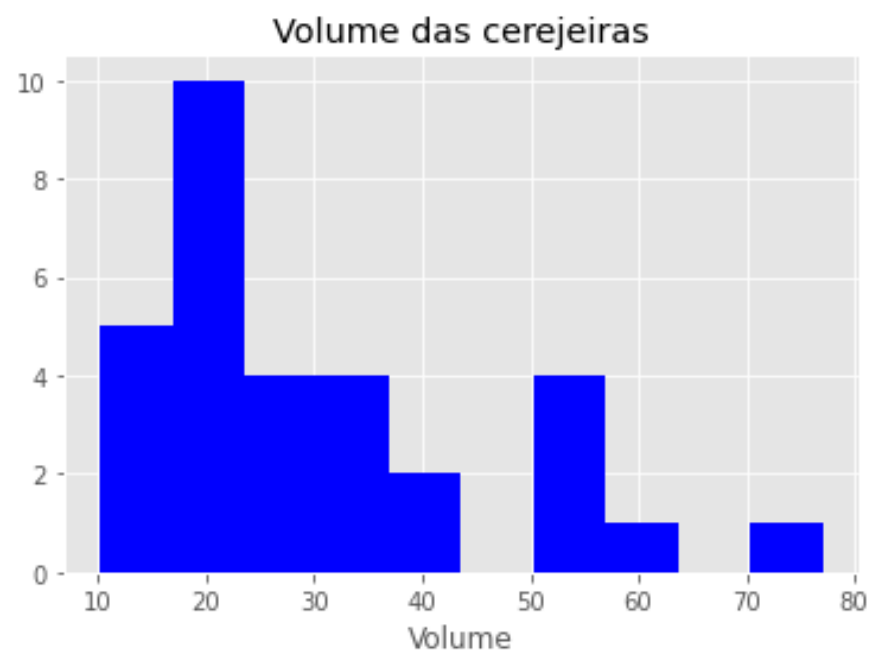

Figura 4: Histograma do Volume das cerejeiras

Portanto, com as análises realizadas para a Tabela Resumo e Box-Plot alinhadas com os Histogramas, temos evidências de que os dados são assimétricos, contínuos e positivos.

Análise Descritiva

Com base na Figura 5 temos um indicativo de que os dados de **Volume**, **Diâmetro** e **Altura** não são normalmente distribuídos.

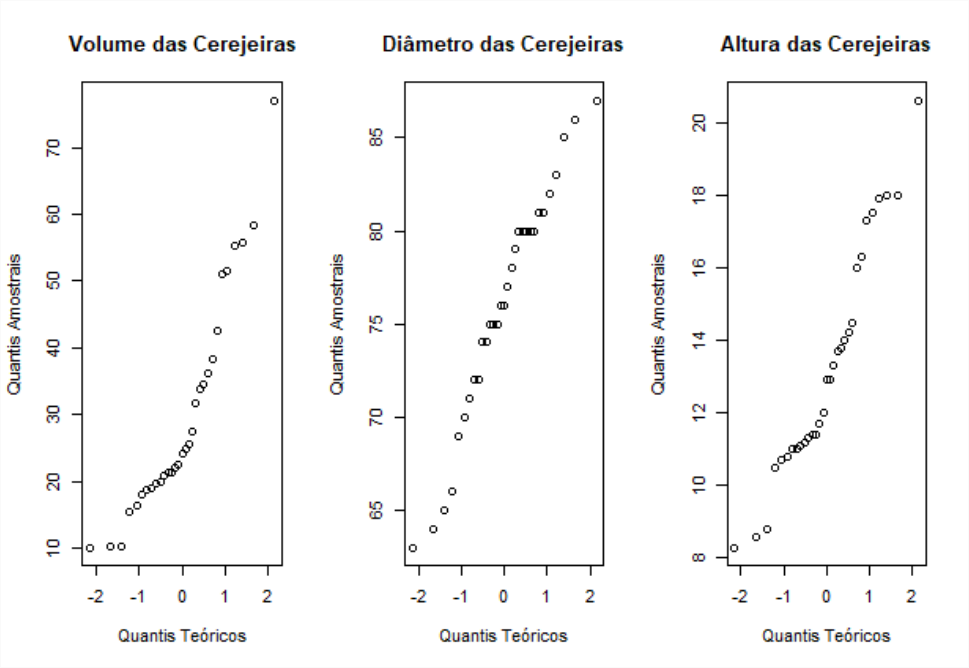

Figura 5: Verificando nromalidade dos dados

Portanto, como os dados não são simétricos e não são normalmente distribuídos, temos evidências de que não é adequado o ajuste do modelo Normal para os dados.

Análise Descritiva

Com base na Figura 6, vemos que a nossa variável resposta Volume e nossa variável explicativa Altura, possuem uma relação linear intermediária com correlação de Person de 0.59, talvez seja necessário fazer uma transformação na variável resposta para melhorar essa relação linear.

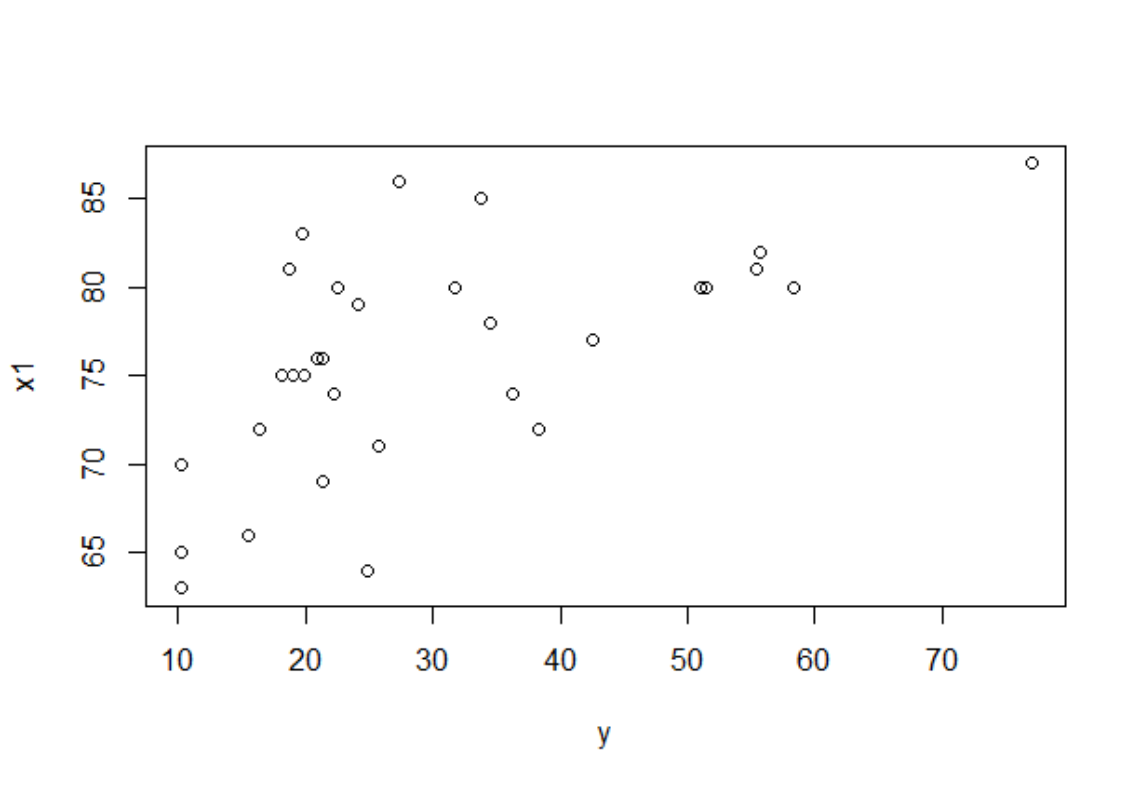

Figura 6: Diagrama de dispersão do Volume das cerejeiras pela Altura

Análise Descritiva

Com base na Figura 7, vemos que a nossa variável resposta Volume e nossa variável explicativa Diâmetro, possuem uma boa relação linear com correlação de Person de 0.967.

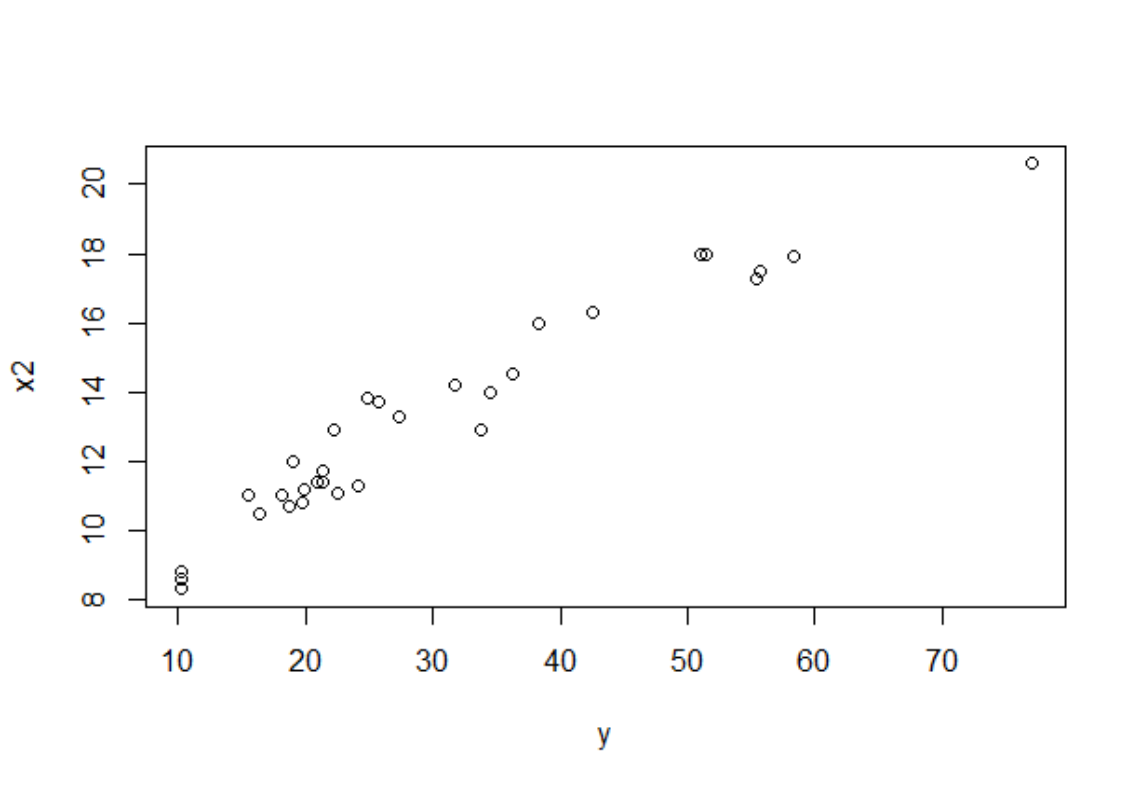

Figura 7: Diagrama de dispersão do Volume das cerejeiras pelo Diâmetro

No sentido de melhorar

Com base na Figura 8, fazendo uma transformação logaritimica na Variável Volume com a variável Diâmetro, temos que a nossa correlação aumenta para 0.969, pouco aumento, mas ainda há um incremento.

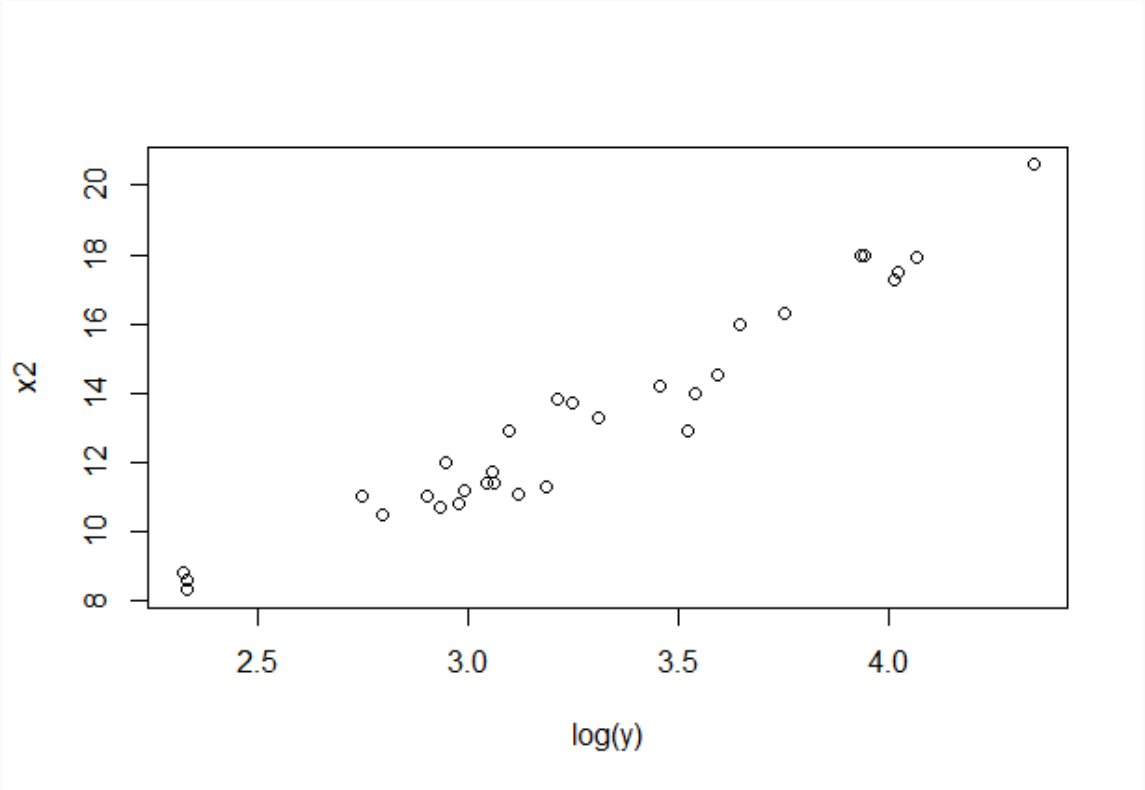

Figura 8: Diagrama de dispersão do Volume das cerejeiras pelo Diâmetro

Análise Descritiva

Com base na Figura 9, fazendo uma transformação logaritimica na Variável Volume com a variável Altura, temos que a nossa correlação aumenta para 0.64, aumento de 0.05.

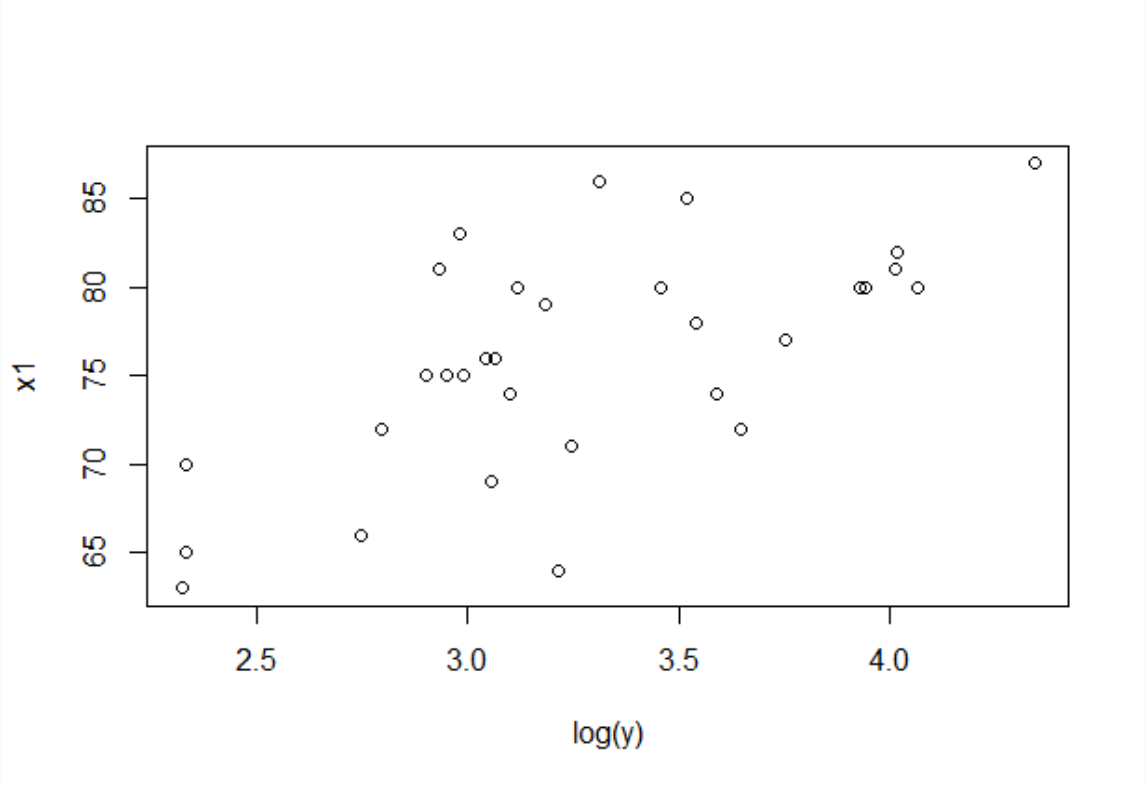

Figura 9: Diagrama de dispersão do Volume das cerejeiras pelo Altura

Análise

Para utilização do modelo linear generalizado ajustado com resposta gama ou normal inversa, vamos ajustar um modelo de regressão linear com resposta normal e fazer a análise diagnóstico para sabermos se podemos simplificar o nosso problema em um problema de regressão linear. Para isso, as seguintes suposições devem ser atendidas:

- Independência
- Homocedásticidade dos resíduos
- normalidade dos resíduos

Análise Diagnóstico

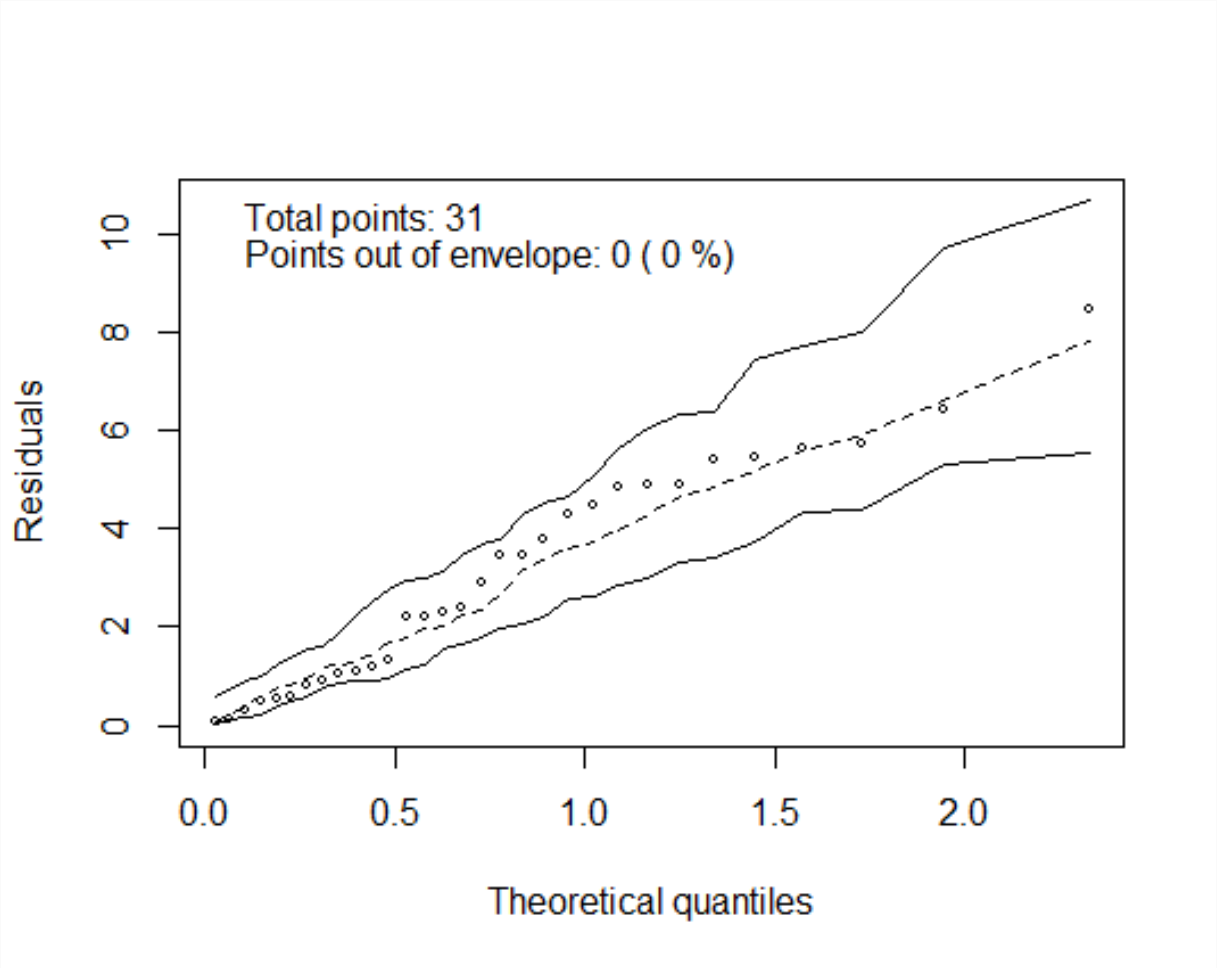

Figura 10: Gráfico de Envelope do modelo Normal

Análise Descritiva

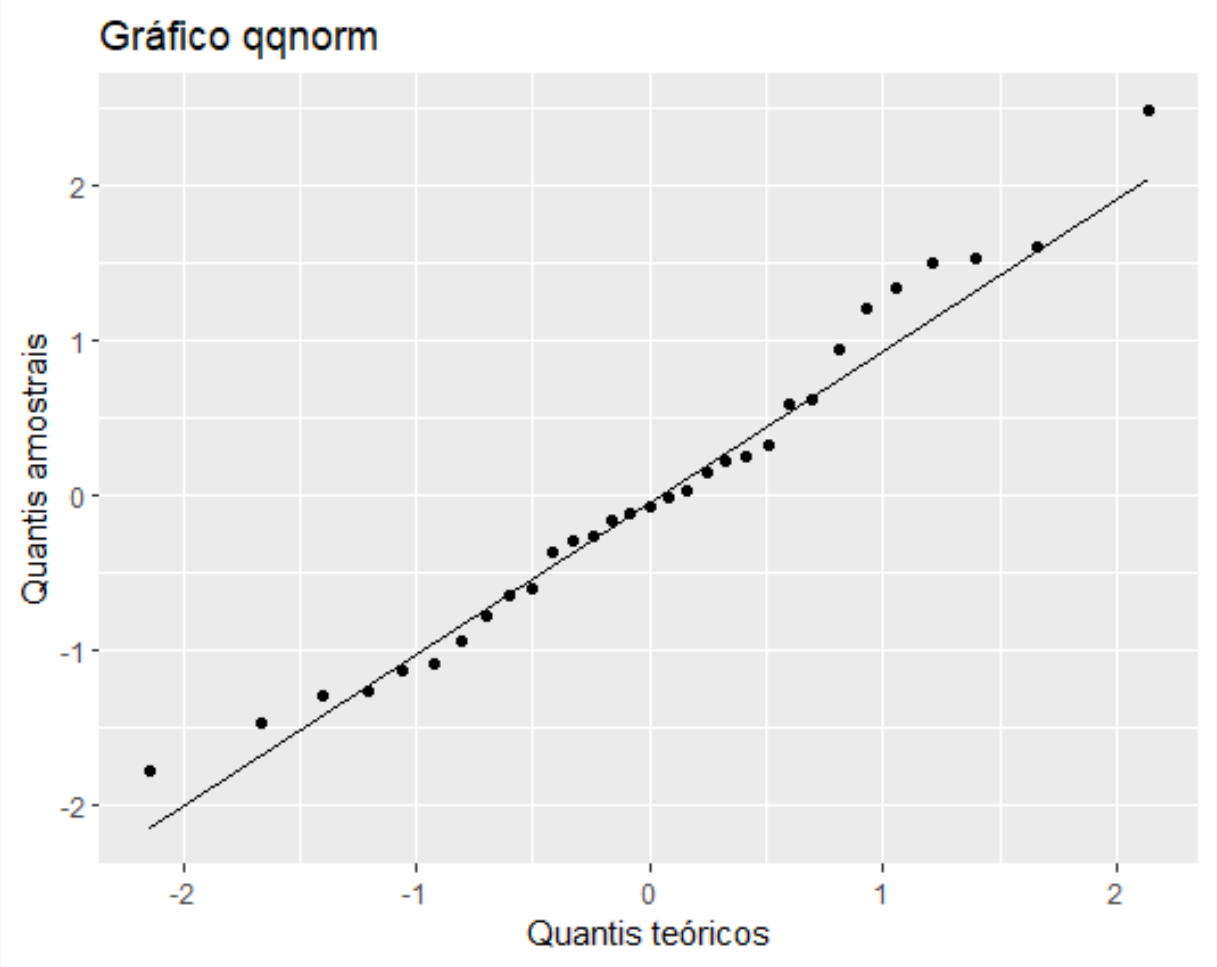

Figura 11: Gráfico QQnorm do modelo normal

Análise Diagnóstico do modelo normal

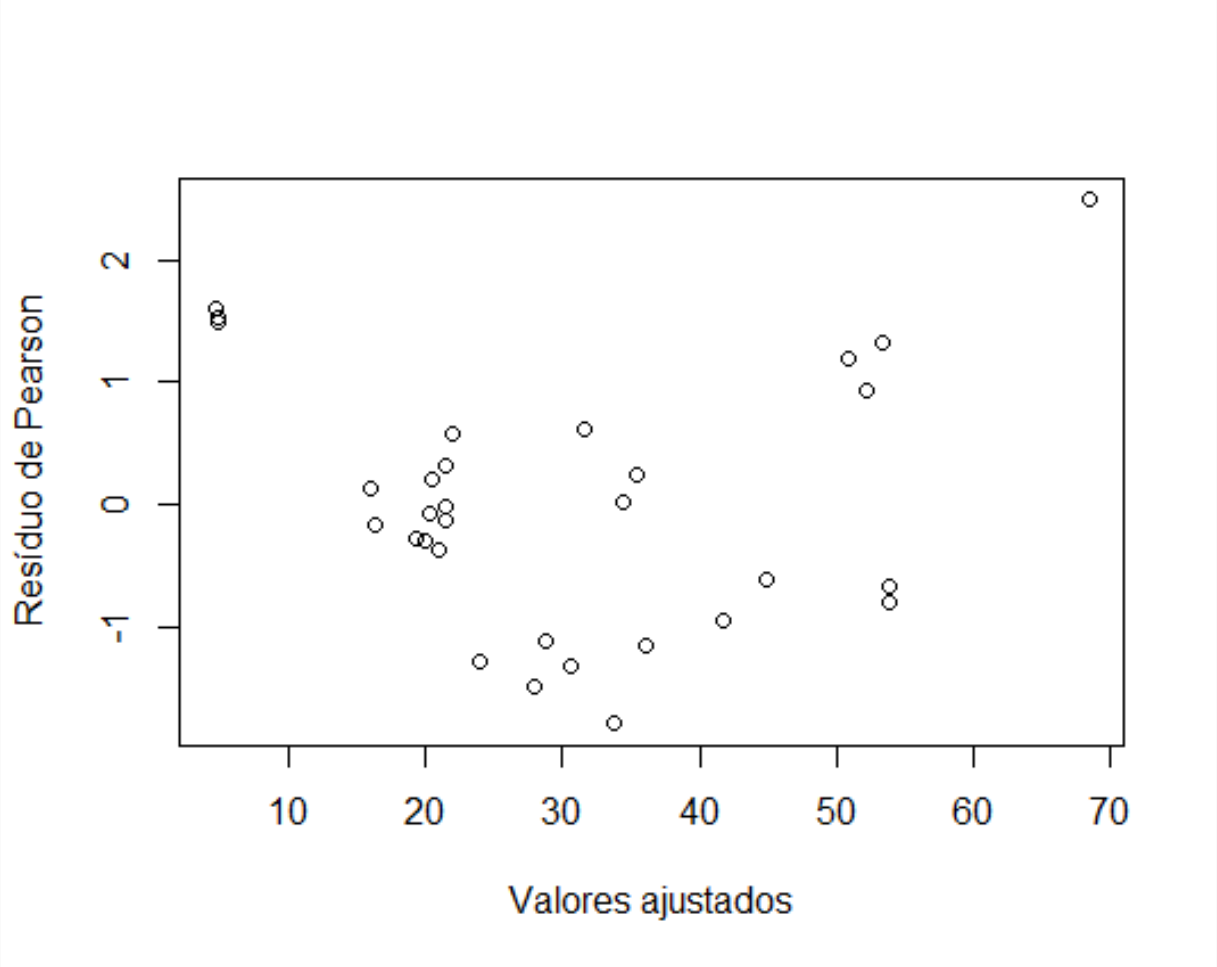

Figura 12: Gráfico Resíduos vs Valores Ajustados do modelo normal

Análise diagnóstico do modelo normal - conclusão

Para a suposição de independência, como não temos o tempo de coleta dos dados, vamos garantir que essa suposição é atendida.
Já para verificar a suposição de normalidade dos resíduos, verificamos através dos gráficos 11 e 10 que a normalidade está sendo satisfeita, já que, no gráfico do QQplot, os pontos parecem seguir uma reta, e no gráfico do Envelope, não há nenhum ponto está fora dos limites gerados pelo gráfico.
A suposição de homocedasticidade dos resíduos, pode ser verificada através do gráfico de Resíduos vs Valores ajustados que foi feito em 12. Porém, notamos uma certa tendência nos pontos, chegando até a ter um formato de funil, onde podemos perceber que os resíduos são heterocedásticos e concluímos que o modelo de regressão linear com resposta normal não é adequada para o nosso problema.

Escolha da distribuição e função de ligação

Agora, sabendo que nossos dados são assimétricos, contínuos e positivos, vamos ajustar um modelo com resposta gama ou normal inversa. Para isso, ajustamos essas distribuições com as funções de ligação: raiz quadrada, logaritmica, identidade e canônica. Para escolha da melhor distribuição e função de ligação que provavelmente irá se ajustar bem aos nossos dados, vamos utilizar o método do menor critério de informação de Akaike (AIC).

Tabela 3: AIC's das distribuições Gamma e Gaussiana Inversa sem interação

	Raiz Quadrada	**Log**	**Identidade**	**Canônica**
Gamma	143.25	151.01	170.47	200.87
Gaussiana Inversa	142.89	149.16	167	201.95

Escolha da distribuição e função de ligação

Observando a tabela anterior, podemos notar que o menor valor de AIC se deve a distribuição gaussiana inversa com função de ligação raiz quadrada, porém, podemos perceber que a gaussiana inversa com função de ligação logaritmica, não está com muita diferença, então para fins de interpretabilidade dos parâmetros, iremos escolher a distribuição gaussiana inversa com função de ligação logaritimica.

Conclusão da função de ligação adequada

Segue da tabela, que os valores dos AIC´s são muito próximos, optamos pela função de ligação log. A função log é utilizada com mais frequência pois, a função de ligação canônica pode gerar valores ajustados negativos, o que não é adequado, uma vez que estamos ajustados valores da variável resposta positivos.

Interação das variáveis do modelo escolhido

Como temos indicativo de que, o melhor modelo será a Gaussiana inversa com função de ligação logaritimica, e temos poucas variáveis iremos ver se as variáveis diâmetro e altura possuem interação.
Para isso, iremos usar o teste da razão de verossilhanças é um teste usado para comparar a adequação de ajuste de dois modelos aninhados.
A estatística de teste é a diferença entre os log-verossimilhanças dos dois modelos, multiplicada por -2.
Onde as hippóteses são: H_0: o modelo sem interação, ou seja, o modelo mais simples, é melhor para explicar os dados contra H_1: o modelo com interação, ou seja o modelo mais completo é melhor para explicar os dados.
Fazendo o teste em R, obtemos um p-valor de 0,005421, ou seja, usando um nivel $\alpha = 0{,}05$, rejeitamos H_0. Temos evidências de que o modelo com interação é melhor para explicar os nossos dados.

Análise diagnóstico

Para verificar se o modelo escolhido se ajusta bem aos dados, iremos fazer uma análise diagnóstico para verificar se as seguintes suposições serão atendidas:

- Independência dos resíduos
- Normalidade dos resíduos
- Homocedasticidade dos resíduos

Se as suposições forem verificadas então temos que nosso modelo escolhido, vai ser o modelo adequado para os dados.

Análise diagnóstico do modelo escolhido

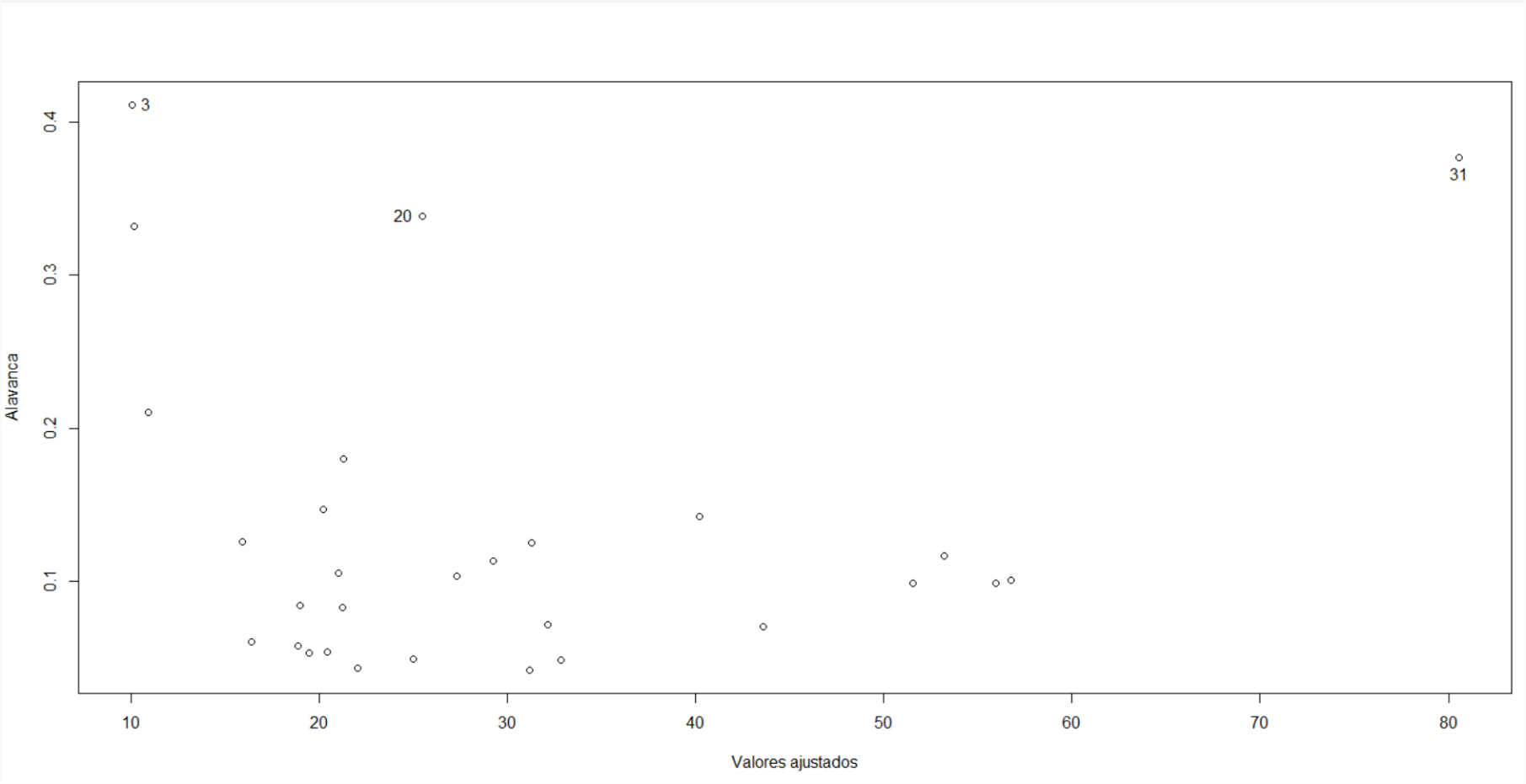

Figura 13: Pontos alavanca do modelo ajustado final

Análise diagnóstico do modelo escolhido

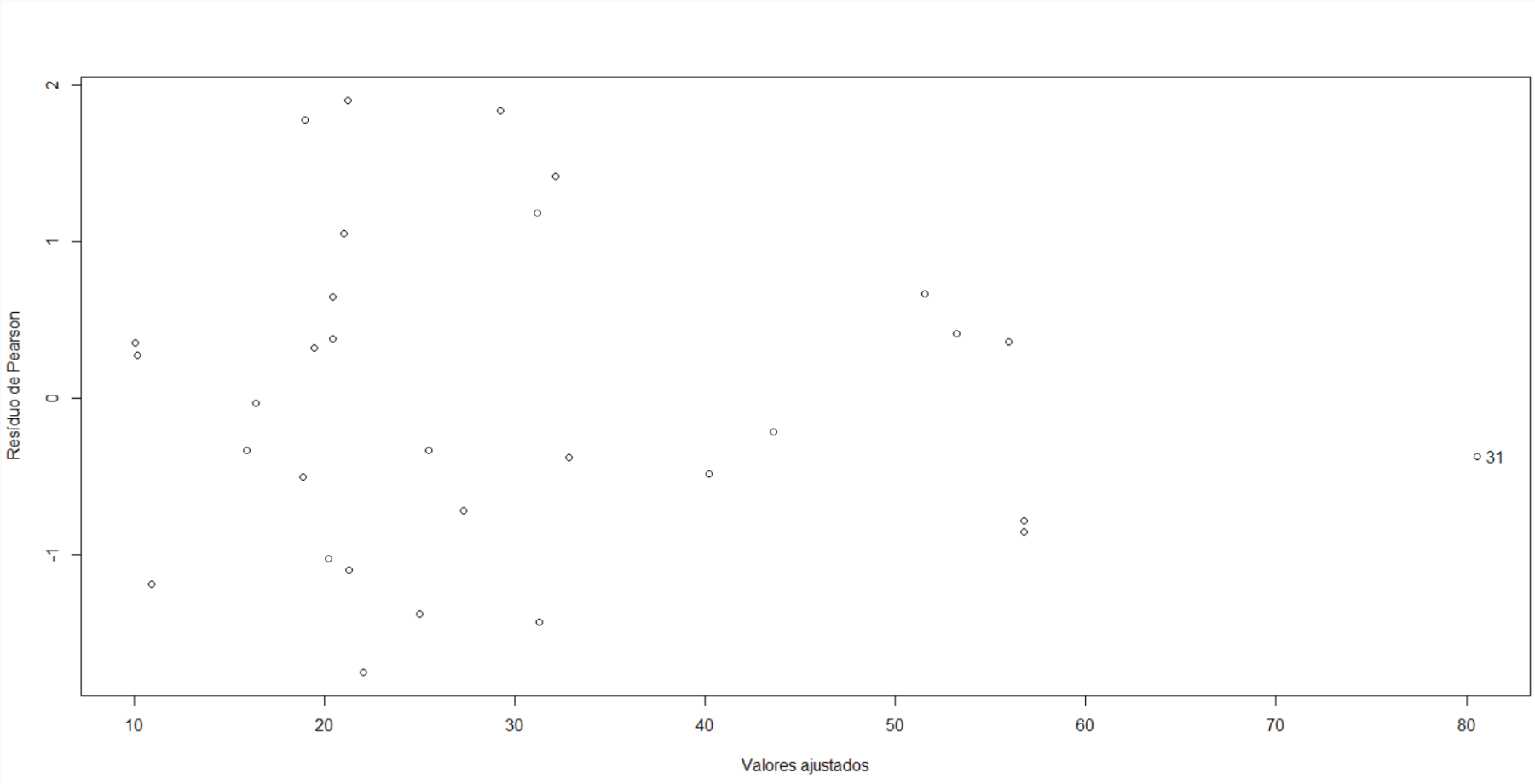

Figura 14: Gráfico Resíduos vs Valores Ajustados do modelo final

Análise diagnóstico do modelo escolhido

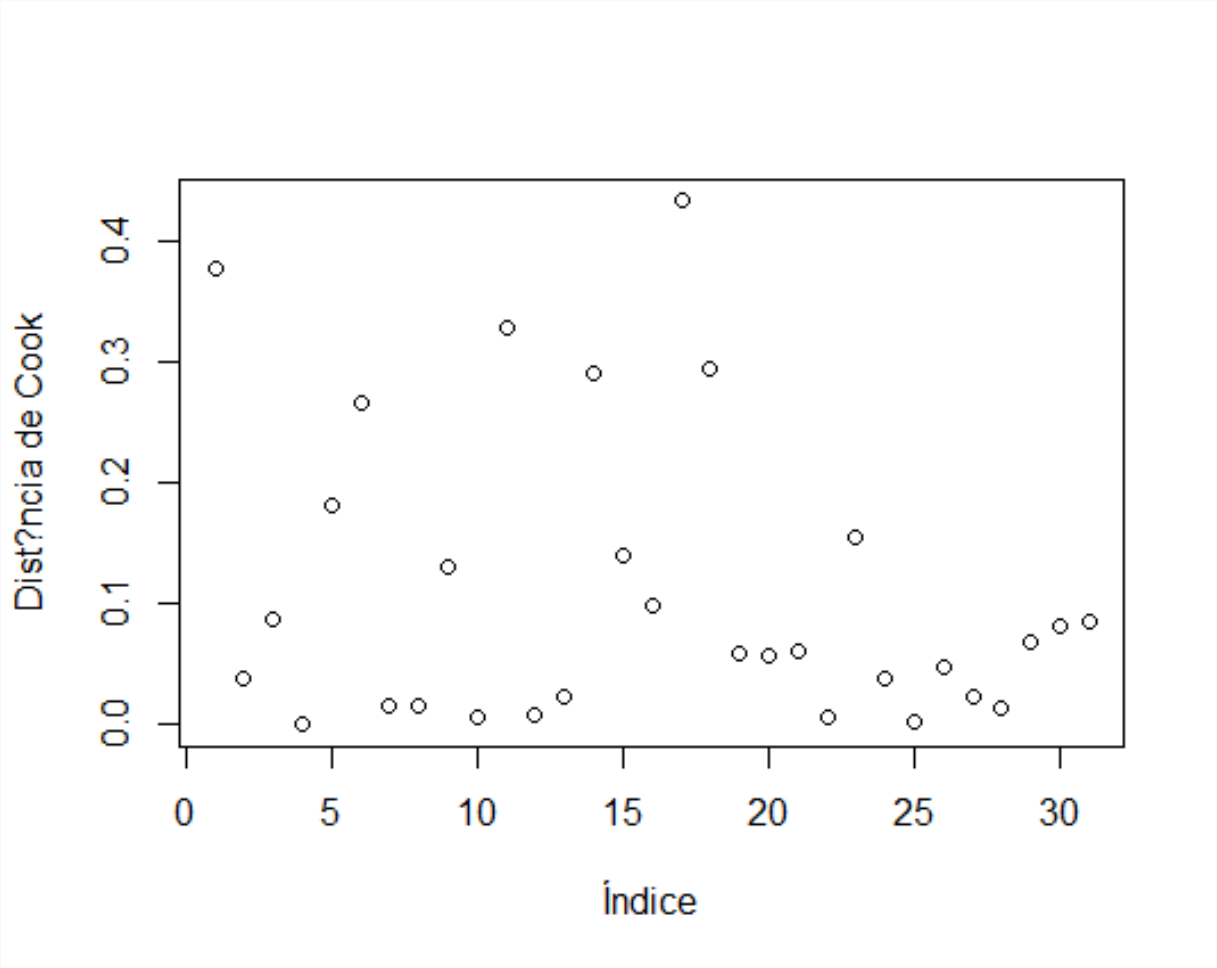

Figura 15: Pontos influentes

Análise diagnóstico do modelo escolhido

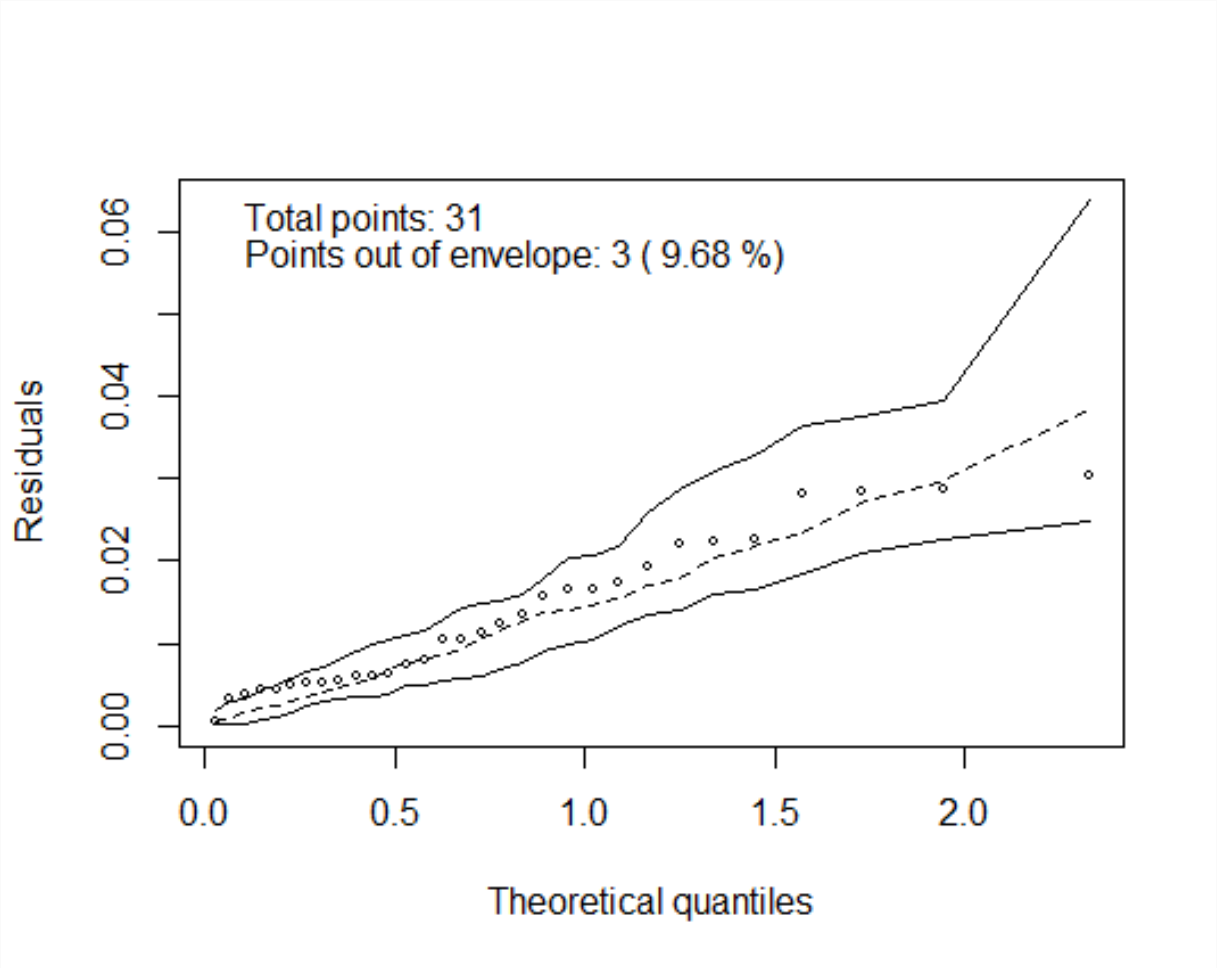

Figura 16: Gráfico Envelope

Análise diagnóstico do modelo escolhido

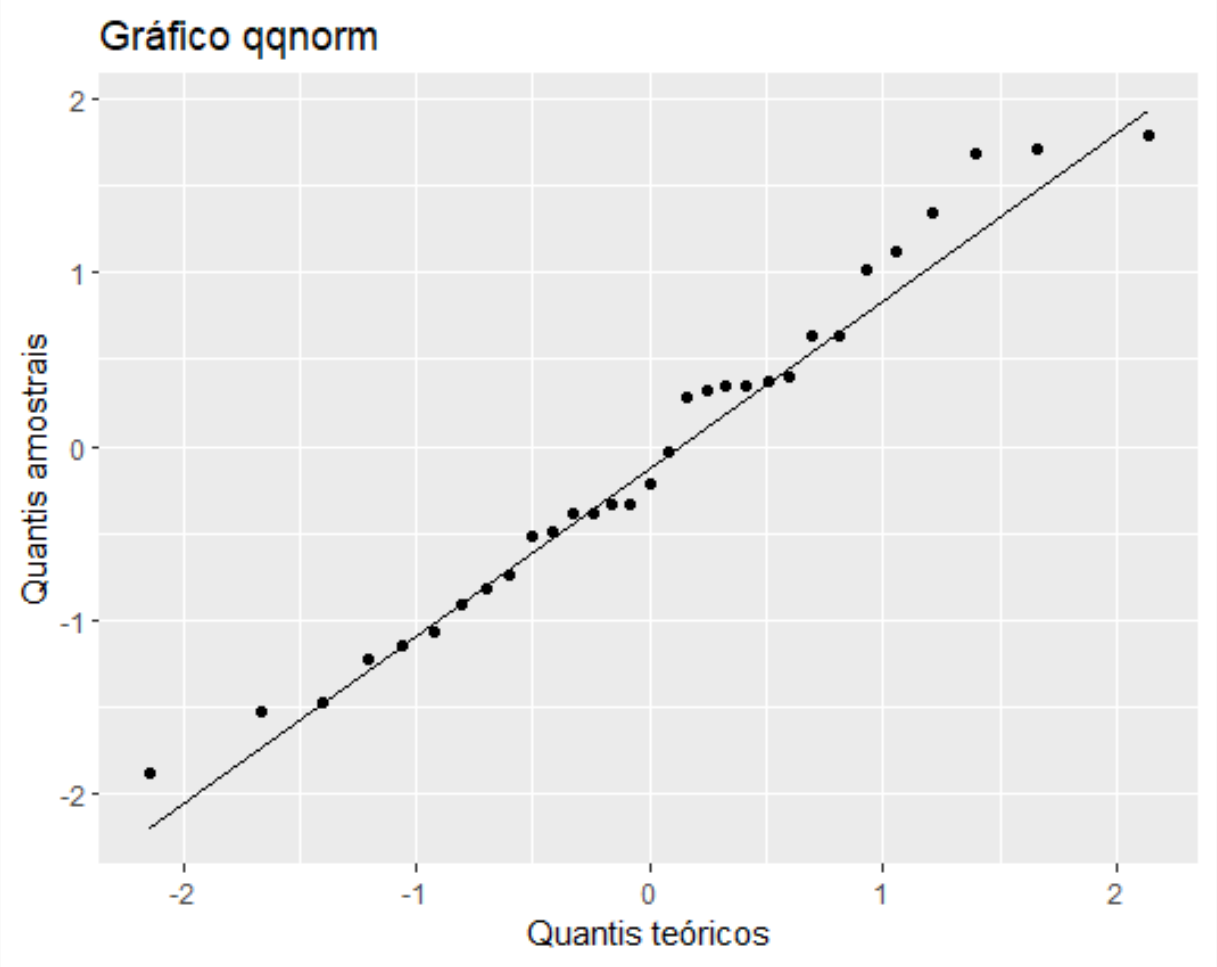

Figura 17: Gráfico QQNorm

Teste de normalidade dos resíduos

Utilizando um $\alpha = 0.05$, temos que H_0: os resíduos são normais contra H_1: os resíduos não são normais.

Teste	**p-valor**
Shapiro-Wilk	0.58
Kolmogorov-Smirnov	0.8966
Anderson-Darling	0.71

Análise de diagnóstico

Pela figura 15, podemos notar que não há pontos influentes já que o último ponto não está muito longe do penúltimo ponto visto.
Já no gráfico de envelope 16 e no QQNorm 17, podemos notar que os resíduos serão normais uma vez que somente 3 pontos ficaram para fora dos limites do envelope e no QQnorma maioria dos pontos consegue captar a linha reta formada e os testes de normalidade não rejeitaram a hipóetese de normalidade.
Agora, podemos notar na figura 13 , que há ponto alavanca referente a observação 31, e esse ponto volta a ser mostrado na fugura 14, onde, com esse ponto, podemos notar uma tendência de funil dos resíduos nos indicando heterocedasticidade, porém, antes de concluir que o modelo escolhido não é adequado, vamos retirar esse ponto conflitante e fazer a análise outra vez

Analise diagnóstico do modelo final mas sem a obervação 31

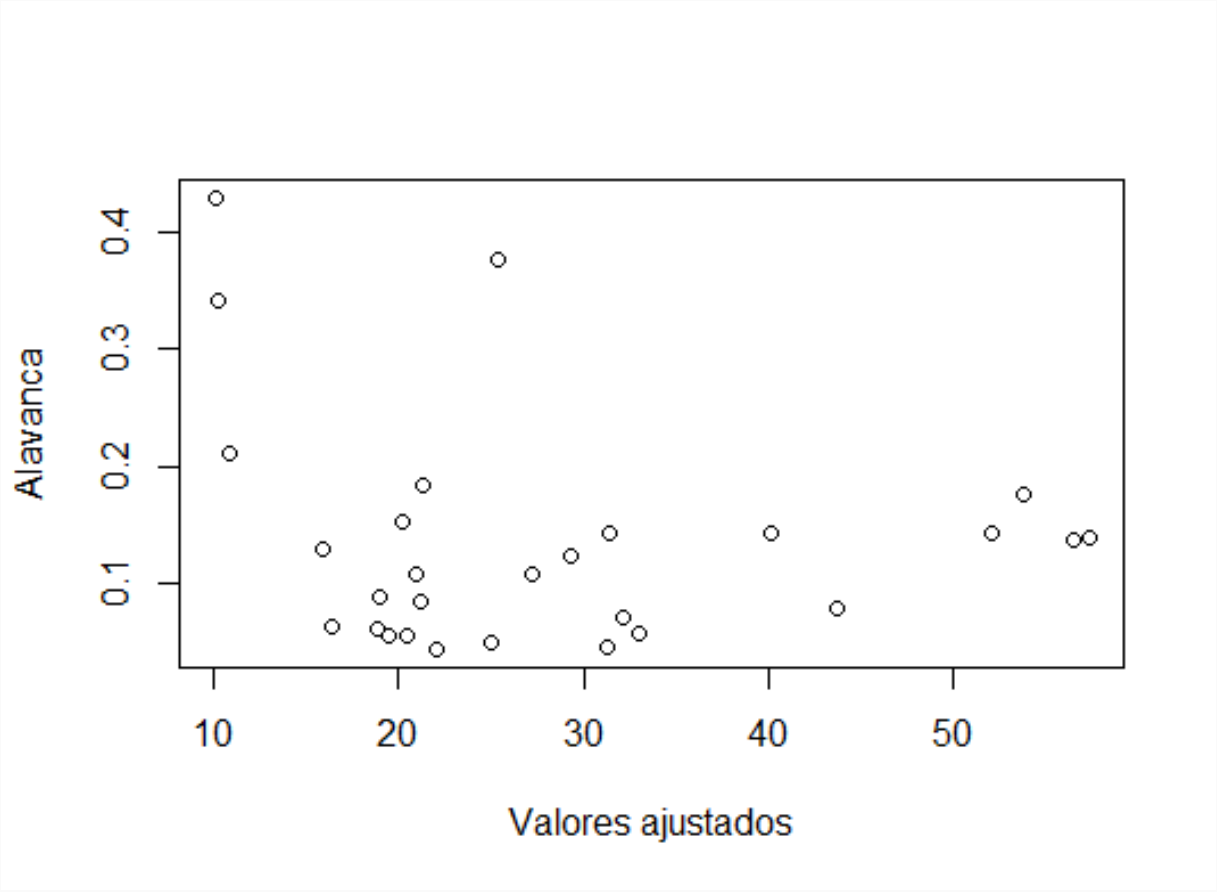

Figura 18: Pontos alavanca do modelo final sem a observação 31

Analise diagnóstico do modelo final mas sem a obervação 31

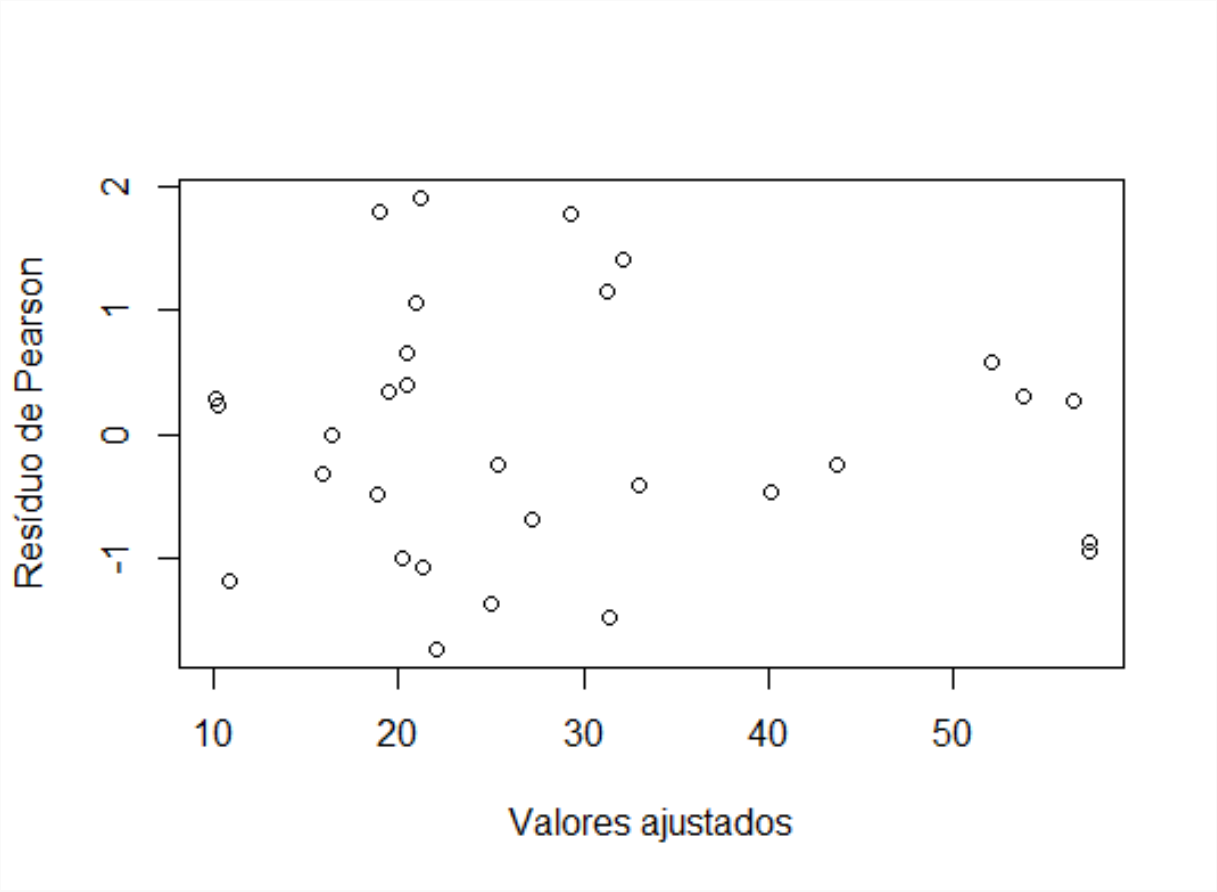

Figura 19: Gráfico de resíduos vcs Valores ajustados

Analise diagnóstico do modelo final mas sem a obervação 31

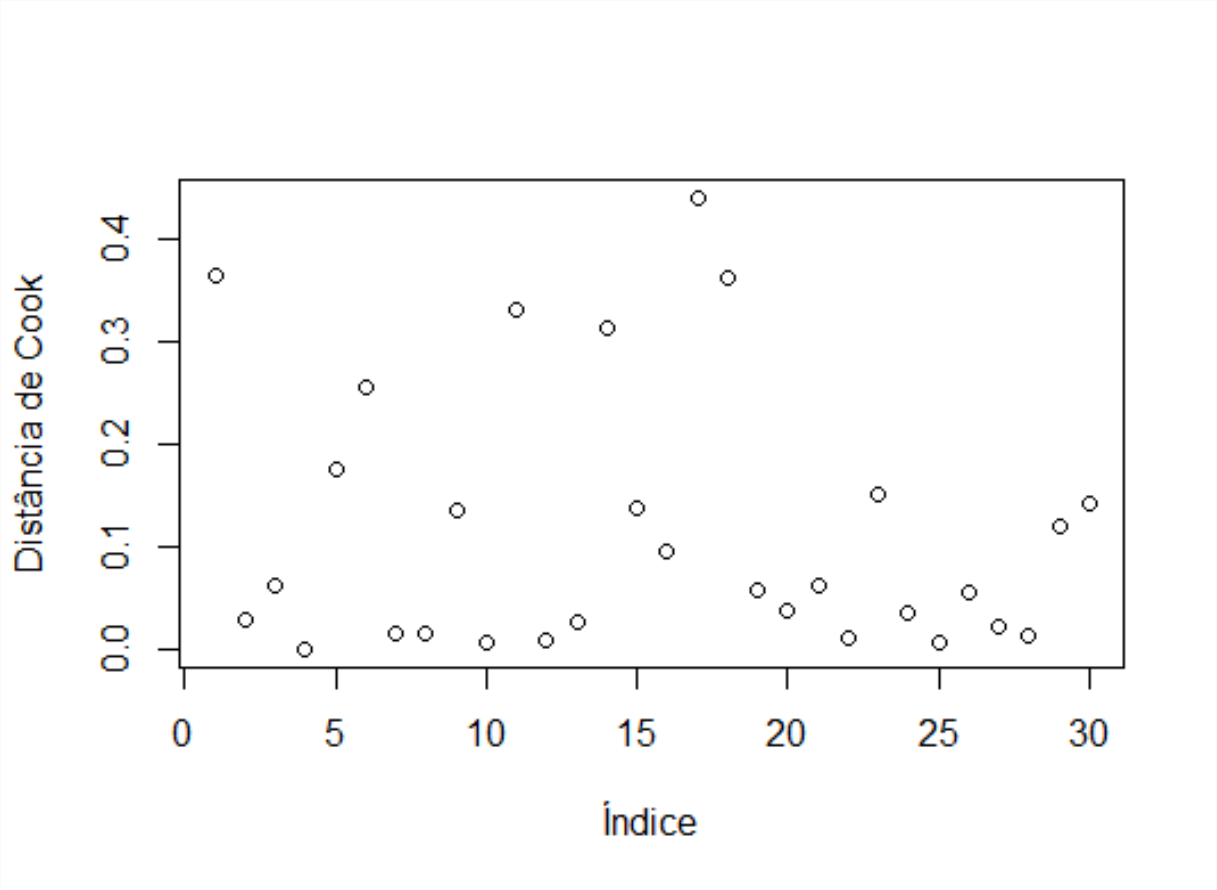

Figura 20: Pontos influentes do modelo final sem a observação 31

Analise diagnóstico do modelo final mas sem a obervação 31

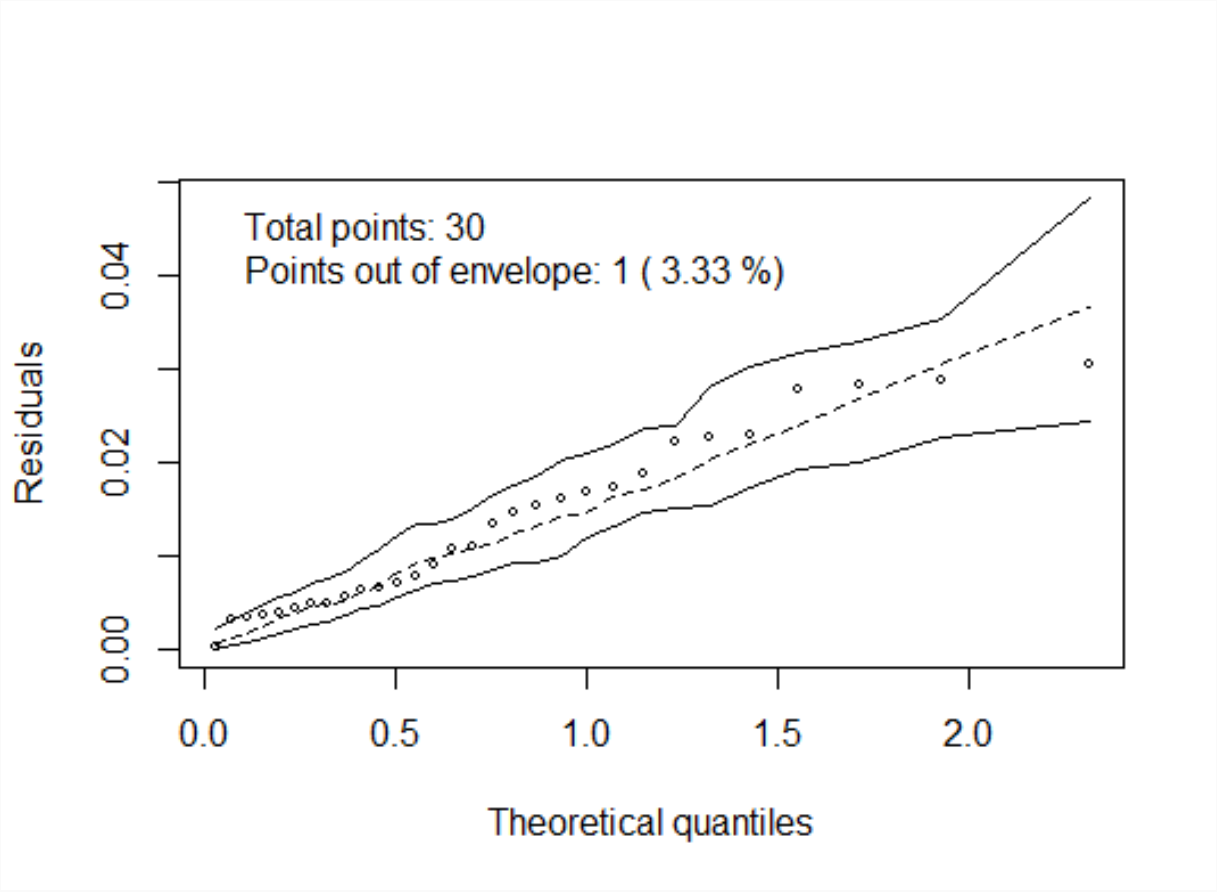

Figura 21: gráfico envelope do modelo final sem a observação 31

Teste de normalidade dos resíduos

Utilizando um $\alpha = 0.05$, temos que H_0: os resíduos são normais contra H_1: os resíduos não são normais.

Teste	**p-valor**
Shapiro-Wilk	0.5965
Kolmogorov-Smirnov	0.9824
Anderson-Darling	0.7808

Análise diagnóstico

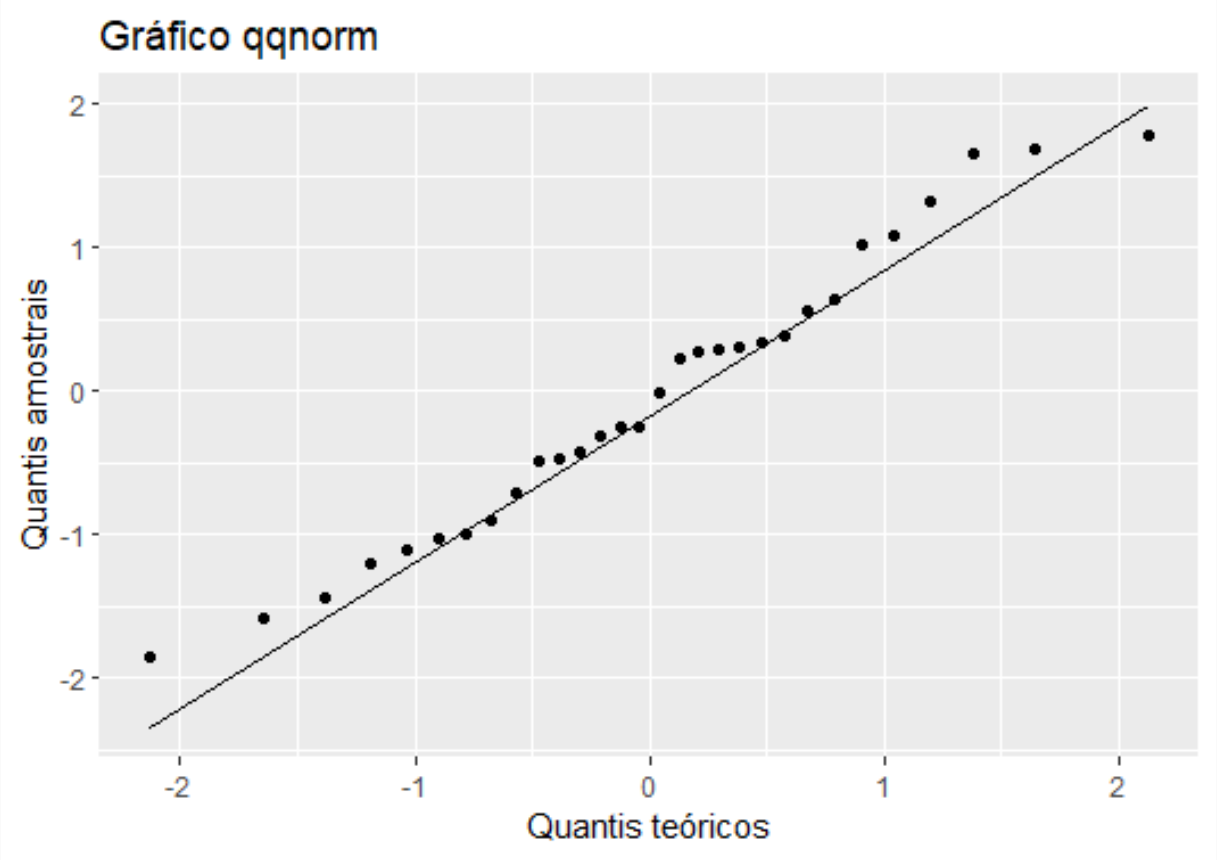

Figura 22: QQnorm do modelo final sem a observação 31

Análise diagnóstico

Podemos notar pela tabela **??** e os gráficos de QQNorm em 22 e o gráfico do envelope na figura 21 que os resíduos, ainda que sem a observação 31, continuaram sendo normais, jpa que os pontos do QQNorm conseguiram acompanhar o comportamento da linha reta, e tambe, diminuiram os pontos fora do envelope, que antes eram 3 pontos, agora é somente um ponto que fica fora dos limites do envelope. A normalidade também é testada nos testes de Shapiro-Wilk, Kolmogorov-Smirnov e Anderson ddarling, onde resultaram p-valor maior que 0.05, nos confirmando que o modelo ajustado final sem a observação 31 tem resíduos normais.
A heterocedasticidade que era verificada com a observação 31 agora não é mais, os pontos do gráfico 19 não seguem nenhuma tendência ou seja, verificamos que a homocedasticidade também é satisfeita.
A independência ainda é tida como verificada nesse caso.

Comparação entre o modelo com e sem a observação 31

P-valor sem a obs 31	P-valor com a obs 31
0.026932	0.00579
0.000891	6.49e-05
0.000224	1.21e-05
0.039699	0.00776

Estimativas	Estivativas sem a obs 31	Estivativas com a obs 31
Intercepto	-1.9387155	-2.0965383
x1	0.0420835	0.0442921
x2	0.3151770	0.3296283
x1*x2	-0.0021200	-0.0023203

Comparação entre o modelo com e sem a observação 31

AIC sem a obs 31	136.73
AIC com a obs 31	143.42

Podemos notar pelas tabelas que os p-valores diminuiram porém ainda continuaram significativos e que não houve mudança bruta nas estimativas dos parâmetros, porém, temos apenas 31 observações, que é um banco de dados de mediano para pequeno, então essaas estimativas mudadas serão consideradas altas, temos também uma mudança considerada alta no AIC, então concluímos que o modelo sem o ponto influente deixa o modelo escolhido bem ajustado.

Verificando a qualidade do ajuste

A deviance pode ser usada para verificar a qualidade do ajuste do modelo aos dados. Geralmente, quanto menor a deviance, melhor é o ajuste do modelo aos dados. No entanto, é importante lembrar que a deviance é uma medida relativa e deve ser usada em comparações entre modelos.

Uma maneira comum de usar a deviance para avaliar a qualidade do ajuste é comparar a deviance do modelo ajustado com a deviance do modelo nulo :

Deviance do modelo nulo	**ajuste do modelo final**
0.2771229	0.007313106

Podemos perceber que a deviance do modelo final foi muito menor que a deviance do modelo sem nenhuma variável. Concluímos então que o modelo final escolhido, se ajustou bem aos dados.

Modelo final

Com a análise que se decorreu neste trabalho, concluimos que o modelo final escolhido é a normal inversa com a função de ligação log com interação. Segue a equação do modelo:

$$log(\mu_i) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2) \quad (1)$$

Substituindo, temos:

$$log(\mu_i) = exp(-2.09654 + 0.04429 x_1 + 0.32963 x_2 + -0.00232 x_1 x_2) \quad (2)$$

Interpretação dos parâmetros

- β_0: Como o volume da árvore não pode assumir valor zero, então não há uma interpretação prática para o parâmetro β_0.
- β_1: $e^{0.04429}$ é o valor pelo qual é multiplicada a média do volume das cerejeiras quando o diâmetro aumenta uma unidade e a altura permance constante.
- β_2: $e^{0.32963}$ é o valor pelo qual é multiplicada a média do volume das cerejeiras quando a altura aumenta uma unidade e o diâmetro permance constante.
- β_3: $e^{-0.00232}$ é o valor pelo qual é multiplicada a média do volume das cerejeiras quando o diâmetro e a altura aumentam uma unidade cada.

Referências

[1] Zar, J. H. (1999). Biostatistical Analysis, Fourth Edition. Prentice-Hall International, Upper Saddle River, New Jersey. Exercise 14.4.

[2] Statci. Proteínas plásmaticas de tartarugas em jejum `http://www.statsci.org/data/general/turtles.html`

[3] ZUANETTI, D. A.. Notas de aula da disciplina Planejamento e Análise de Experimentos 1

[4] PAULA, Gilberto A.. Modelos de regressão com apoio computacional. São Paulo: Instituto de Matemática e Estatística, 2013. 434 p.