Explorando a Relação entre Variáveis Ambientais e a Produção de Clorofila-a em Cianobactérias: Uma Abordagem com Modelos Lineares Generalizados

Thiago Tavares Lopes

10 dezembro 2024

Sumário

Resumo

Este trabalho investiga a relação entre variáveis ambientais e a produção de clorofila-a em cianobactérias, utilizando Modelos Lineares Generalizados (MLGs). O estudo utiliza um conjunto de dados obtido do catálogo público dos Estados Unidos, contendo informações genéticas, ambientais e geográficas de diversas espécies de cianobactérias. A variável resposta, clorofila-a, foi modelada considerando características ambientais, como nitrogênio total, fósforo total, oxigênio dissolvido, pH da água e dióxido de carbono dissolvido.

Os modelos foram ajustados com distribuição Gamma e função de ligação logarítmica devido à natureza assimétrica e positiva da variável resposta. A seleção do modelo baseou-se na técnica stepwise para minimizar o critério de informação de Akaike (AIC). O modelo final identificou que o nitrogênio total, fósforo total, dióxido de carbono e pH da água possuem influência significativa na produção de clorofila-a. Oxigênio dissolvido também mostrou tendência relevante, mas com significância marginal.

A análise de resíduos confirmou a adequação do modelo, indicando ausência de observações influentes e distribuição aleatória dos resíduos, sugerindo homocedasticidade. O desempenho geral do modelo, medido pelo R^2 de Nagelkerke, foi moderado, explicando cerca de 41,96

1 Introdução

Foi proposto um modelo linear generalizado para avaliar a produção de clorofila a em cianobactérias em diferentes condições climáticas. O *dataset* utilizado foi obtido do catálogo de dados público do governo dos Estados Unidos, disponível em *Data.Gov*. Esse *datset* possui informações detalhadas sobre diferentes espécies de cianobactérias, sendo estas informações: Informações genéticas, condições climáticas do local de coleta das amostras e localização geográfica da coleta. Para a construção do modelo, foram consideradas exclusivamente as informações de clorofila a quantificada e condições ambientais (físicas e químicas) do local da coleta das amostras. O trabalho aqui desenvolvido foi fundamentado no artigo disponível no seguinte link.

Sobre os dados utilizados, temos as seguintes variáveis e suas respectivas descrições (tabela 1):

Tabela 1: Variáveis utilizadas

Variável	Descrição
chlorophyll_a	Quantidade de Clorofila a (μg/L)
total_nitrogen	Quantidade Total de nitrogênio (μg/L)
total_phosphorus	Quantidade Total de Fósforo (μg/L)
temp_water_celsius	Temperatura da água (°C)
dissolved_oxigen	Oxigênio dissolvido (mg/L)
pH_water	pH da água
carbon_dioxide_water	Dióxido de Carbono (mg/L)
total_nitrogen_water	Quantidade de Nitrato, Nitrito, Amônia e Nitrogênio Orgânico (mg/L)
nitrite_water	Quantidade de Nitrito (mg/L)
nitrate_water	Quantidade de Nitrato (mg/L)
phosphorus_water	Quantidade de Fósforo (mg/L)
sulfate_water	Quantidade de Sulfato (mg/L)
total_nitrogen_water	Quantidade de Nitrato, Nitrito, Amônia e Nitrogênio Orgânico (mg/L) - água filtrada
ammonia(NH3 + NH4+)_water	Quantidade de NH3+ e NH4+ (mg/L) como NH4

1.1 Análise Descritiva

Nesta seção são apresentados os resultados referentes a análise descritiva dos dados. As tabelas 2 e 3 mostram os valores de mínimo, máximo, média dos dados em estudo. Destaca-se a variável **chlorophyll_a** que possui a característica de ser contínua e positiva, e justifica o uso da distribuição Gamma para modelar a mesma.

Tabela 2: Análise descritiva

chlorophyll_a	total_nitrogen	total_phosphorus	temp_water_celsius	dissolved_oxigen	pH_water	carbon_dioxide_water
Min. : 0,70	Min. : 15	Min. : 1,60	Min. : 8,8	Min. : 5,30	Min. :6,90	Min. : 0,70
1st Qu.: 2,02	1st Qu.: 58	1st Qu.: 4,00	1st Qu.:21,3	1st Qu.: 6,95	1st Qu.:7,40	1st Qu.: 2,60
Median : 4,85	Median : 129	Median : 7,15	Median :23,8	Median : 7,80	Median :7,60	Median : 3,40
Mean : 6,75	Mean : 236	Mean :22,40	Mean :23,4	Mean : 7,72	Mean :7,58	Mean : 4,15
3rd Qu.:11,55	3rd Qu.: 326	3rd Qu.:22,00	3rd Qu.:26,3	3rd Qu.: 8,50	3rd Qu.:7,80	3rd Qu.: 4,75
Max. :20,20	Max. :1300	Max. :97,00	Max. :30,6	Max. :10,90	Max. :8,30	Max. :16,00

Tabela 3: Análise descritiva

total_nitrogen_water...8	nitrite_water	nitrate_water	phosphorus_water	sulfate_water	total_nitrogen_water...13	ammonia(NH3 + NH4+)_water
Min. : 0,180	Min. :0,00100	Min. : 0,040	Min. :0,0160	Min. : 2,88	Min. : 0,15	Min. :0,0130
1st Qu.: 0,647	1st Qu.:0,00725	1st Qu.: 0,359	1st Qu.:0,0400	1st Qu.: 7,53	1st Qu.: 0,58	1st Qu.:0,0138
Median : 1,450	Median :0,01250	Median : 0,907	Median :0,0715	Median : 15,80	Median : 1,29	Median :0,0370
Mean : 2,572	Mean :0,03250	Mean : 1,789	Mean :0,2240	Mean : 40,04	Mean : 2,36	Mean :0,0757
3rd Qu.: 3,475	3rd Qu.:0,02250	3rd Qu.: 2,345	3rd Qu.:0,2200	3rd Qu.: 74,30	3rd Qu.: 3,25	3rd Qu.:0,0785
Max. :13,000	Max. :0,26600	Max. :10,800	Max. :0,9700	Max. :128,00	Max. :13,00	Max. :0,6880

A imagem 1, apresenta as correlações entre as variáveis.

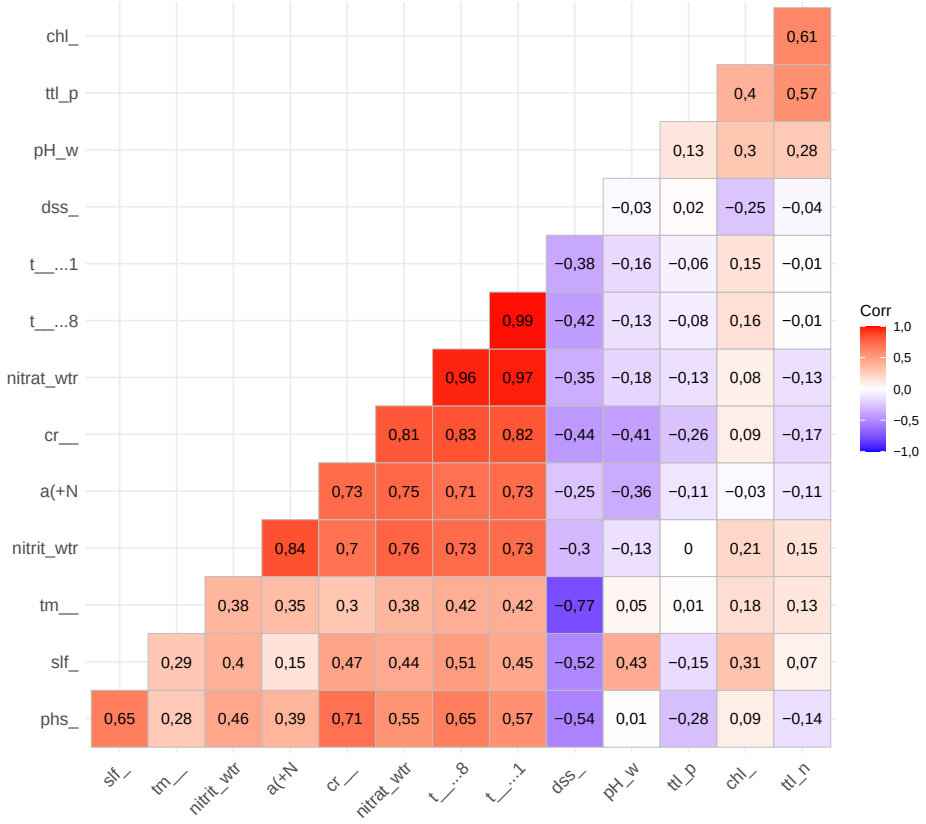


Figura 1: Matriz de Correlação

1.2 Cianobactérias

Devido a atividade fotossintetizante das cianobactérias estima-se que as primeiras tiveram origem entre 2,6 a 3,5 bilhões de anos atrás (LAU e colab., 2015) e são uma das principais responsáveis pela atmosfera oxigenada como conhecemos hoje, participando do "Grande Evento de Oxigenação" (HUISMAN e colab., 2018; PLANAVSKY e colab., 2014; RASMUSSEN e colab., 2008). O estromatólitos são uma evidência da atividade de microrganismos que ocorreu a , aproximadamente 3,700 milhões de anos atrás (NUTMAN e colab., 2016). As cianobactérias são classificadas como microrganismo procariontes autotróficos com sistemas adaptativos particulares, como a capacidade de fixar nitrogênio do ar atmosférico devido a presença da enzima nitrogenase localizada nos heterócitos (PETERS e colab., 2015). São capazes de realizar a fotossíntese na presença ou ausência de oxigênio. Existem espécies que se desenvolvem na ausência de luz ou em condições anaeróbicas utilizando sulfetos como doadores de elétrons para a fotossíntese, além disso são bactérias gram – negativas e dispõem da estrutura chamada bainha mucilaginosa e tricoma, podem ou não apresentar o acineto e o heterócitos que são estruturas especializadas na sobrevivência da espécie em ambientes não favoráveis (ABED e colab., 2009; COHEN e colab., 1986; HUISMAN e colab., 2018; LAU e colab., 2015; STAL e MOEZELAAR,

1997). A figura 2, apresenta algumas das cianobactérias encontradas na região amazônica. As mesmas podem apresentar estrutura filamentosa, colonial, etc.

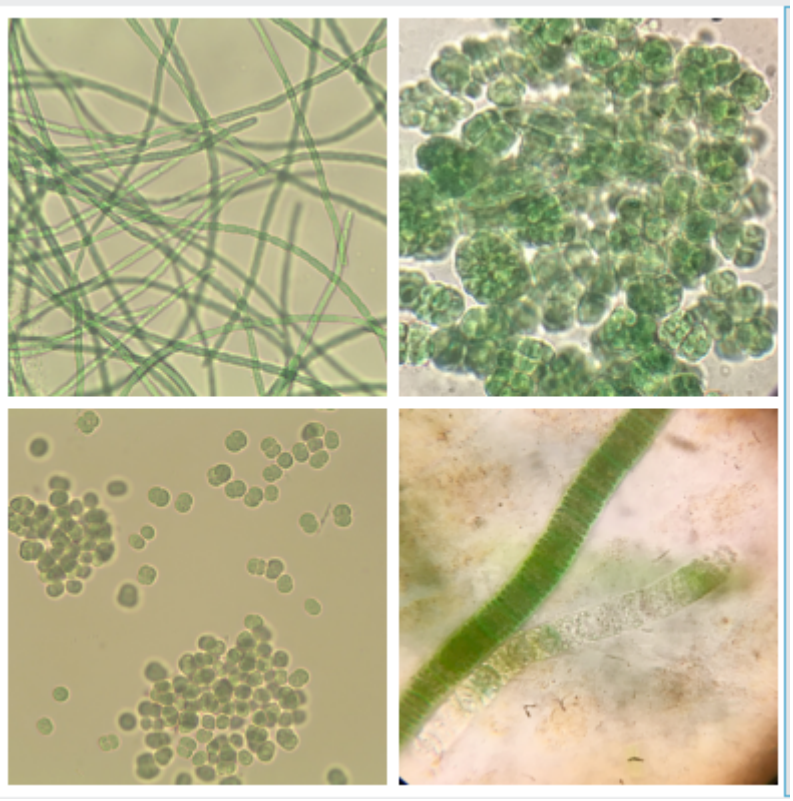


Figura 2: Cianobactérias-Fonte: Coleção Amazônica de Cianobactérias e Microalgas

A clorofila a 3 é um pigmento verde ou azul que capta a luz natural ou sintética e é essencial para a fotossíntese, é encontrado em todos os grupos de vegetais e outros organismos autótrofos, utilizada como indicadora da biomassa em ambientes aquáticos.

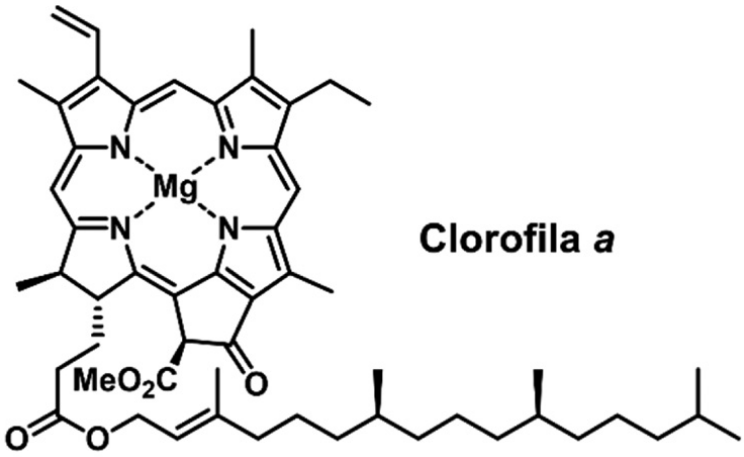


Figura 3: Clorofila a

1.3 Modelos Lineares Generalizados

Um modelo linear generalizado (MLG), é uma extensão dos modelos lineares generalizados em que a variável dependente não segue, obrigatoriamente, uma distribuição normal. Podemos definir a variável resposta como o componente aleatório do modelo, pertecente a mais variadas distribuições como a normal, binomial, exponencial, etc. O compomente sistemático, variável independente, do modelo se encaixa de forma linear ao mesmo. Por último, pode-se definir a função de ligação, que a estrutura responśavel por fazer a ligação entre o componente aleatório e o componente sistemático. Em sua definição, temos que $g(\cdot)$ é uma função que transforma a média da variável responsta $E(y) = \mu$ de modo a estabelecer uma relação lienar com os preditores da seguinte forma $g(\mu) = X\beta$.

2 Seleção do modelo

Nesta seção são apresentados os resultados da elaboração dos modelos para explicar a variabilidade de produção de clorofila a. Devido a natureza positiva assimétrica da variável resposta $Y(chlorophyll_a)$, foi utilizado a distribuição Gamma com função de ligação logarítmica. Seja $Y\ Gamma(\mu, \phi)$ em que: μ é a média e ϕ é o parâmetro de dispersão.

A tabela 4, apresenta os resultados do primeiro modelo construído. Nota-se, que apenas as variáveis *total_nitrogen*, *pH_water* e *total_nitrogen_water...8*, foram significativas para o modelo. Por conseguinte, foi utilizado a técnica **stepwise** para escolher um novo modelo baseado no menor AIC.

Tabela 4: Resultados para o modelo 1

	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	-7,17864	5,53834	-1,29617	0,20167	
total_nitrogen	0,00109	0,00054	2,02436	0,04903	*
total_phosphorus	0,00698	0,00441	1,58323	0,12053	
temp_water_celsius	-0,02700	0,05012	-0,53868	0,59282	
dissolved_oxigen	-0,25390	0,18544	-1,36920	0,17788	
pH_water	1,40239	0,61068	2,29644	0,02647	*
carbon_dioxide_water	0,24729	0,13331	1,85499	0,07031	.
total_nitrogen_water...8	0,28761	0,79233	0,36300	0,71834	
nitrite_water	-0,07889	5,14524	-0,01533	0,98784	
nitrate_water	0,18158	0,21271	0,85367	0,39791	
phosphorus_water	-1,14418	1,13485	-1,00823	0,31886	
sulfate_water	-0,00460	0,00567	-0,81218	0,42106	
total_nitrogen_water...13	-0,46093	0,84540	-0,54521	0,58836	
‘ammonia(NH3 + NH4+)_water‘	-1,31316	1,93832	-0,67747	0,50165	

[a] Nota:*** p<0.001; ** p<0.01; * p<0.05; . p<0.1

Foi ajustado um novo modelo com $AIC = 317$ e Residual Deviance (28,252) que foi menor que a Deviance Nula (46,207). Os resultados são apresentado na tabela 5, observa-se que as variáveis significativas para o modelo são, *total_nitrogen*, *total_phosporus*, *dissolved_oxigen*, *pH_whater*, *carbon_dioxide_water*. Ademais, teste de aderência baseado na razão entre a deviance residual $(28, 252)$ e o valor crítico de qui-quadrado $qchisq(0.95, 52)$ (69,832 para 52 GL) sugere que o modelo se ajusta adequadamente aos dados. Também foi testado a função de ligação utilizando o teste RESET, no qual foi obtido um $p - valor = 0.0996$ indicando que a função de ligação (log), foi adequada para o modelo.

Por conseguinte, a clorofila-a é positivamente influenciada por nitrogênio total, dióxido de carbono e pH da água, sendo essas variáveis significativas para o modelo. Por último, Fósforo total e oxigênio dissolvido têm tendências que podem ser relevantes em análises futuras, com mais dados ou em níveis de significância menos conservadores.

Tabela 5: Resultdos para o modelo 2

	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	-4,23976	2,71284	-1,5629	0,12415	
total_nitrogen	0,00089	0,00039	2,2833	0,02653	*
total_phosphorus	0,00693	0,00384	1,8039	0,07704	.
dissolved_oxigen	-0,13273	0,07892	-1,6817	0,09862	.
pH_water	0,83629	0,31895	2,6220	0,01144	*
carbon_dioxide_water	0,07590	0,03738	2,0305	0,04744	*

[a] Nota:*** p<0.001; ** p<0.01; * p<0.05; . p<0.1

Por conseguinte, as expressões 1 e 2, representam a estrutura do modelo e o modelo final com os β estimados, respectivamente

$$\begin{aligned}\eta = \beta_0 + \beta_1 \cdot \text{total\_nitrogen} + \beta_2 \cdot \text{total\_phosphorus} + \beta_3 \cdot \text{dissolved\_oxigen} \\ + \beta_4 \cdot \text{pH\_water} + \beta_5 \cdot \text{carbon\_dioxide\_water}\end{aligned} \tag{1}$$

$$\begin{aligned}\eta = -4.239760 + 0.000889 \cdot \text{total\_nitrogen} + 0.006927 \cdot \text{total\_phosphorus} \\ -0.132727 \cdot \text{dissolved\_oxigen} + 0.836286 \cdot \text{pH\_water} + 0.075904 \cdot \text{carbon\_dioxide\_water}\end{aligned} \tag{2}$$

Por último, a tabela 6 apresenta os coeficientes estimados para o modelo log-linear aplicado a μ, juntamente com a interpretação de cada estimativa. Esses coeficientes indicam como cada variável preditora está associada à resposta em termos do logaritmo natural da média da variável resposta.

Tabela 6: Resultados da regressão log-linear com coeficientes estimados e suas interpretações

Variável	**Coef. (Estimativa)**	**Interpretação**
(Intercept)	-4,239760	Valor médio de $\log(\mu)$ quando todos os preditores são iguais a zero.
total_nitrogen	0,000889	Para cada aumento unitário em total_nitrogen, μ cresce em $e^{0,000889}$.
total_phosphorus	0,006927	Para cada aumento unitário em total_phosphorus, μ cresce em $e^{0,006927}$.
dissolved_oxigen	-0,132727	Para cada aumento unitário em dissolved_oxigen, μ diminui em $e^{-0,132727}$.
pH_water	0,836286	Para cada aumento unitário em pH_water, μ cresce em $e^{0,836286}$.
carbon_dioxide_water	0,075904	Para cada aumento unitário em carbon_dioxide_water, μ cresce em $e^{0,075904}$.

2.1 Análise de Resíduos

Neste capítulo, é apresentada a análise de resíduos do segundo modelo desenvolvido. O R^2 generalizado foi calculado utilizando a métrica de Nagelkerke, cujo valor obtido foi 0,41957. Esse resultado indica que o modelo explica aproximadamente 41,96% da variância observada nos dados, sugerindo um ajuste moderado. Este desempenho é coerente com a natureza do conjunto de dados, composto por informações reais coletadas em ambiente natural, que geralmente apresentam alta variabilidade intrínseca.

A figura 4, apresenta o envelope simulado com $\alpha = 5\%$ para os resíduos do modelo, No eixo horizontal, são representados os quantis teóricos de uma distribuição normal padrão, enquanto no eixo vertical, estão os quantis dos resíduos observados. Neste caso, os pontos seguem a linha diagonal central com poucas discrepâncias, indicando que os resíduos estão próximos de uma distribuição normal. Pequenas variações, especialmente nas extremidades, são comuns e podem ser aceitáveis.

```
## Gamma model
```

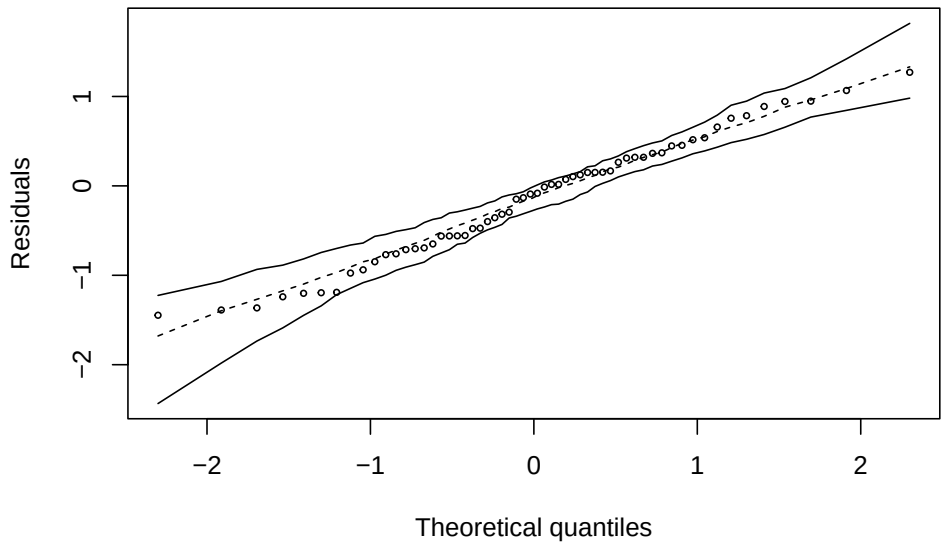


Figura 4: Envelope Simulado

```
## Total points: 58
## Points out of envelope: 0 ( 0 %)
```

A figura 5, apresenta o histograma dos resíduos.

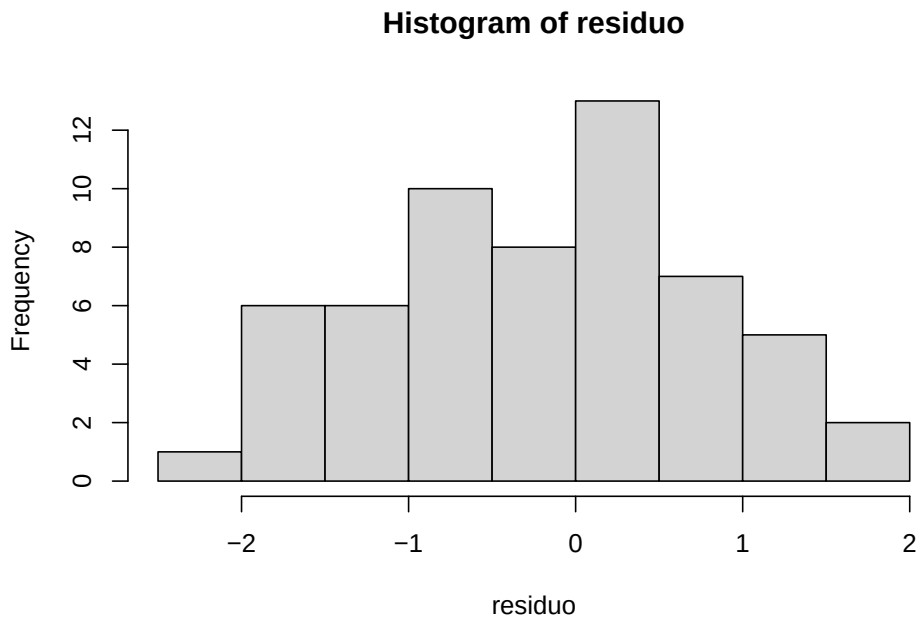
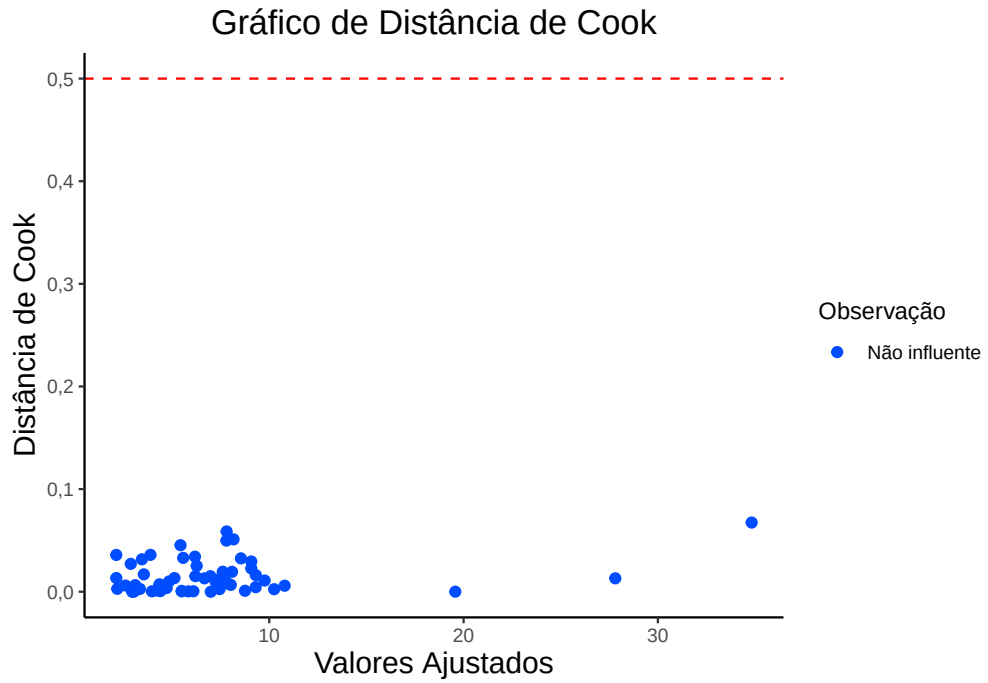


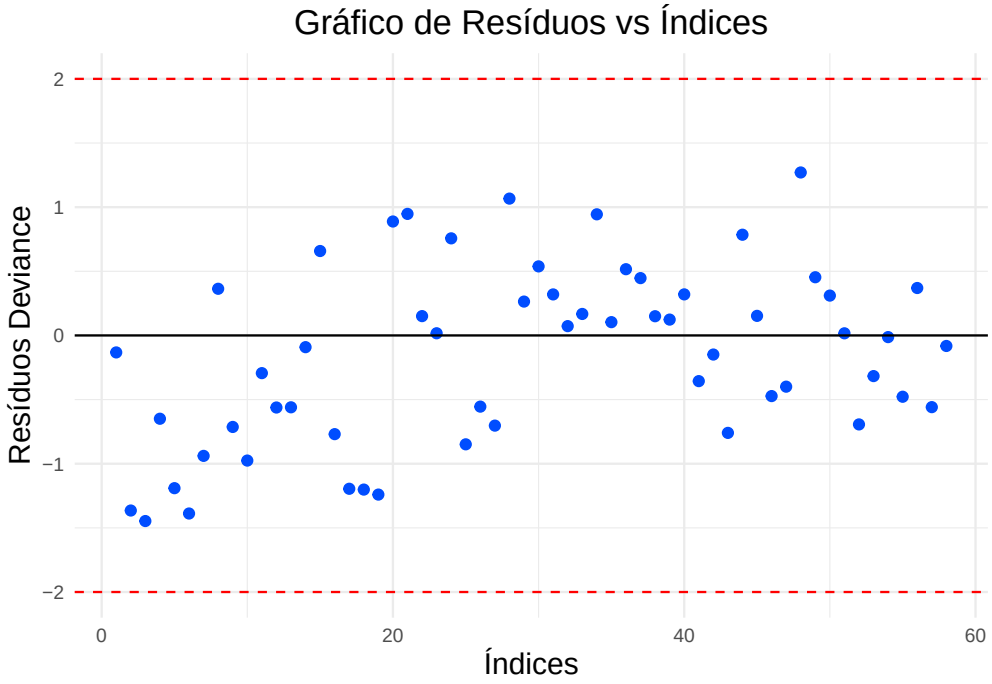
Figura 5: Histograma Resíduos

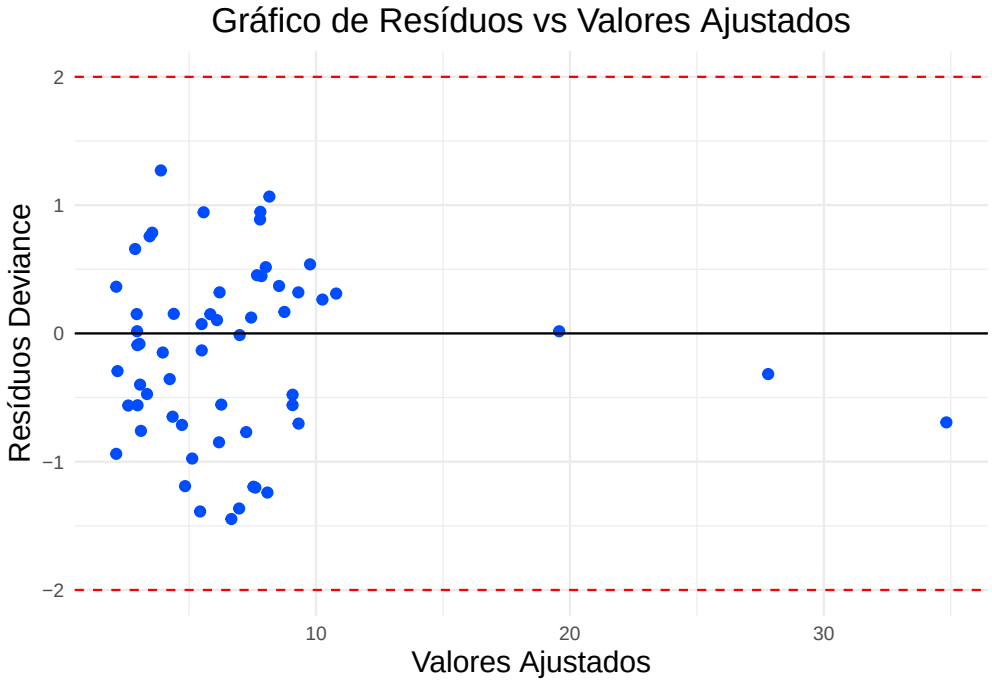
A figura **??**, apresenta a distância de cook e auxilia na identificação de observações influentes. No eixo horizontal estão os valores ajustados (valores preditos pelo modelo), enquanto no eixo vertical estão os valores da Distância de Cook. Neste gráfico, as observações representadas pelos pontos azuis estão abaixo do limite estabelecido, sugerindo que não há observações influentes significativas. Isso indica que os dados não possuem valores que exerçam grande impacto na estabilidade do modelo, o que é desejável em análises de regressão.



A figura **??** apresenta os gráficos para os resíduos. No gráfico, gráfico de Resíduos vs Índices, temos no eixo horizontal os índices das observações, enquanto no eixo vertical estão os resíduos deviance do modelo, que representam as diferenças entre os valores observados e os valores ajustados. Os resíduos devem estar distribuídos aleatoriamente ao redor da linha zero. Por último, os resíduos parecem dispersos de forma relativamente aleatória, sugerindo que o modelo não apresenta violações de homocedasticidade.

Por conseguinte, no gráfico de Resíduos vs Valores Ajustados, temos no eixo horizontal estão os valores ajustados pelo modelo, e no eixo vertical estão os resíduos deviance, que representam as diferenças entre os valores observados e os valores preditos pelo modelo.





A imagem **??** apresenta o gráfico de alavancagem, nota-se que apesar de existir alguns pontos influentes acima da linha vermelha, não são valores que influenciam na performace do modelo.