UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA

CURSO DE GRADUAÇÃO EM ESTATÍSTICA

CAIO BRUNO LOPES DE CARVALHO

ROMULO BARROS DE FREITAS

MODELOS DE REGRESSÃO COM RESPOSTA NORMAL INVERSA

FORTALEZA

2025

CAIO BRUNO LOPES DE CARVALHO
ROMULO BARROS DE FREITAS

MODELOS DE REGRESSÃO COM RESPOSTA NORMAL INVERSA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Estatística.

Orientador: Prof. Dr. Gualberto Segundo Agamez Montalvo.

FORTALEZA
2025

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Ao longo das últimas décadas, os Modelos Lineares Generalizados (MLGs) tornaram-se ferramentas essenciais para a análise de dados em diferentes áreas do conhecimento. Esses modelos estendem os modelos lineares clássicos, permitindo que a variável resposta assuma distribuições pertencentes à família exponencial, possibilitando a modelagem de situações em que a suposição de normalidade dos erros não é válida. Assim, os MLGs se mostram adequados para lidar com variáveis de natureza diversa, como contagens, proporções e medidas assimétricas.

De acordo com (CORDEIRO; DEMÉTRIO, 2008), a formulação de um MLG envolve três componentes principais: o componente aleatório, que especifica a distribuição da variável resposta; o componente sistemático, que estabelece a relação linear entre as covariáveis e o preditor linear; e a função de ligação, responsável por associar a média da variável resposta ao preditor linear. Essa estrutura confere flexibilidade aos MLGs, permitindo que sejam aplicados em contextos variados, desde experimentos agrícolas até estudos biomédicos e ambientais.

Essa generalização torna os GLMs particularmente úteis para situações em que a suposição de normalidade da variável resposta não é válida, como no caso de contagens, proporções ou dados assimétricos (MCCULLAGH; NELDER, 1989).

Dentre as distribuições que pertencem à família exponencial, destaca-se a distribuição Normal Inversa, apropriada para modelar dados contínuos e positivos, caracterizados por alta assimetria à direita e variância crescente com a média (DUNN; SMYTH, 2018). Modelos de regressão com resposta Normal Inversa são recomendados, por exemplo, para analisar tempos de falha, tempos de reação e pesos, sendo uma alternativa viável quando modelos baseados na distribuição normal ou gama não apresentam bom desempenho.

Neste trabalho, propõe-se o estudo teórico e aplicado de modelos de regressão com resposta Normal Inversa, abordando desde a dedução de suas propriedades até a implementação computacional. Adicionalmente, serão realizadas análises de diagnóstico para avaliação da adequação dos modelos ajustados, bem como aplicações práticas utilizando dados do pacote R para ilustrar a utilidade e as particularidades deste tipo de modelagem.

1.1 Família Exponencial de Distribuições

Os MLGs constituem uma classe flexível de modelos estatísticos que estende os modelos lineares clássicos, permitindo que a variável resposta siga distribuições que pertencem

à família exponencial, como a binomial, Poisson, normal e gama. Uma distribuição de probabilidade é dita pertencer à família exponencial quando a sua função de densidade (ou função de massa de probabilidade, no caso discreto) pode ser expressa na forma:

$$f(y) = \exp\left(\phi(y\theta - b(\theta)) + c(y;\phi)\right)$$

em que θ representa o parâmetro da distribuição, ϕ é o parâmetro de dispersão. Distribuições que pertencem a essa família apresentam propriedades analíticas particularmente convenientes. Em especial, a esperança e a variância da variável aleatória Y podem ser obtidas diretamente a partir da função cumulante $b(\theta)$, sendo dadas por:

$$E(Y) = b'(\theta) \qquad \text{e} \qquad Var(Y) = \phi\, b''(\theta),$$

em que $b'(\theta)$ e $b''(\theta)$ correspondem, respectivamente, à primeira e à segunda derivada de $b(\theta)$ em relação a θ.

1.2 Modelos Lineares Generalizados

Muitas das distribuições conhecidas podem ser reunidas em uma família denominada família exponencial de distribuições (CORDEIRO; DEMÉTRIO, 2008).

Conforme (DEMÉTRIO, 2001) o modelo linear generalizado envolve os três componentes:

Componente Aleatório: Representado por um conjunto de variáveis aleatórias independente $Y_1, \ldots Y_n$ provenientes de uma mesma distribuição que faz parte da família exponencial.

Componente Sistemático: A forma linear que explicativas entram na estrutura do modelo.

$$\eta_i = \sum_{i=1}^{n} x_{ij}\beta_j = \mathbf{x}_i^{\top}\beta \quad \text{ou} \quad \eta = \mathbf{X}\beta.$$

Função de ligação: uma função que liga o componente aleatório ao componente sistemático, ou seja, relaciona a média ao preditor linear, isto é

$$\eta_i = g(\mu_i)$$

2 METODOLOGIA

2.1 Função Escore e Informação de Fisher

2.1.1 Função Escore e Informação de Fisher para β

Considere a partição $\theta = (\beta^\top, \phi)^\top$ e denote o logaritmo da verossimilhança por $L(\theta)$, ou seja

$$L(\theta) = \prod_{i=1}^{n} f_Y(y_i;\theta_i,\phi) = \prod_{i=1}^{n} \exp\{\phi\,[y_i\theta_i - b(\theta_i)] + c(yi;\phi)\}$$
$$= \exp\left\{\phi \sum_{i=1}^{n}(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(yi;\phi)\right\}.$$

Consequentemente, o logaritmo da função de verossimilhana será denotado por

$$l(\theta) = \phi \sum_{i=1}^{n}(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(yi;\phi). \tag{2.1}$$

Para obter a função escore para o parâmetro β, deriva-se 2.1 com relação a cada coeficiente, ou seja

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j}\left\{\phi \sum_{i=1}^{n}(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(yi;\phi)\right\} \\
&= \phi \sum_{i=1}^{n}\left\{y_i\frac{\partial\theta_i}{\partial\beta_j} - \frac{\partial b(\theta_i)}{\partial\beta_j}\right\} \\
&= \phi \sum_{i=1}^{n}\left\{y_i\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j} - \frac{\partial b(\theta_i)}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j}\right\} \\
&= \phi \sum_{i=1}^{n}\left\{y_i V(\mu_i)^{-1}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)x_{ij} - \mu_i V(\mu_i)^{-1}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)x_{ij}\right\} \\
&= \phi \sum_{i=1}^{n}\left\{(y_i-\mu_i) V(\mu_i)^{-1}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)x_{ij}\right\} \\
&= \phi \sum_{i=1}^{n}\left\{(y_i-\mu_i)x_{ij}\sqrt{\frac{w_i}{V(\mu_i)}}\right\},
\end{aligned}$$

onde $w_i = \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \frac{1}{V(\mu_i)}$ e $V(\mu_i)^{-1}$ é a função de variância de μ_i. Consequentemente, é possível escrever a função escore na forma matricial

$$\mathbf{U}_\beta(\theta) = \frac{\partial L(\theta)}{\partial\theta} = \phi\mathbf{X}^\top\mathbf{W}^{\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}(\mathbf{y}-\mu), \tag{2.2}$$

em que $\mathbf{X}$ é uma matriz de dimensão $n \times p$, de posto completo cujas linhas serão denotadas por $\mathbf{x}_i^\top$, $i = 1,\ldots,n$, $\mathbf{W} = \text{diag}\{w_1,\ldots w_n\}$ é a matriz de pesos, $\mathbf{V} = \text{diag}\{V_1,\ldots,V_n\}$, $\mathbf{y} = (y_1,\ldots,y_n)^\top$ e $\mu = (\mu_1,\ldots,\mu_n)^\top$.

A matriz de informação de Fisher para o parâmetro β é obtida derivando-se novamente $l(\theta)$ em relação aos coeficientes (PAULA, 2024), ou seja

$$\begin{aligned}\frac{\partial^2 l(\theta)}{\partial\beta_j\partial\beta_l} &= \phi\sum_{i=1}^{n}(y_i-\mu_i)\frac{d^2\theta_i}{d\mu_i^2}\left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{ij}x_{il}\\ &\quad+\phi\sum_{i=1}^{n}(y_i-\mu_i)\frac{d^2\theta_i}{d\mu_i^2}\frac{d^2\mu_i}{d\eta_i^2}x_{ij}x_{il}\\ &\quad-\phi\sum_{i=1}^{n}\frac{d\theta_i}{d\mu_i}\left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{ij}x_{il},\end{aligned}$$

cujos valores esperados ficam dados por

$$\begin{aligned}E\left\{\frac{\partial^2 l(\theta)}{\partial\beta_j\partial\beta_\ell}\right\} &= -\phi\sum_{i=1}^{n}\frac{d\theta_i}{d\mu_i}\left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{ij}x_{i\ell}\\ &= -\phi\sum_{i=1}^{n}\frac{(d\mu_i/d\eta_i)^2}{V_i}x_{ij}x_{i\ell}\\ &= -\phi\sum_{i=1}^{n}w_i x_{ij}x_{i\ell}.\end{aligned}$$

Consequentemente, a submatriz de informação de Fisher para β fica expressa na seguinte forma matricial

$$\mathbf{K}_{\beta\beta}(\theta) = E\left\{-\frac{\partial^2 l(\theta)}{\partial\beta\partial\beta^\top}\right\} = \phi\mathbf{X}^\top\mathbf{W}\mathbf{X}. \tag{2.3}$$

Em particular, quando se utiliza a ligação canônica $\theta_i = \eta_i$, essas quantidades tomam forma simplificadas, ou seja

$$\mathbf{U}_\beta = \phi\mathbf{X}^\top(\mathbf{y}-\mu) \quad \text{e} \quad \mathbf{K}_{\beta\beta} = \phi\mathbf{X}^\top\mathbf{V}\mathbf{X}.$$

2.1.2 Função Escore e Informação de Fisher para ϕ

A função escore para o parâmetro ϕ fica dada por

$$\mathbf{U}_\phi(\theta) = \frac{\partial l(\theta)}{\partial\phi} = \sum_{i=1}^{n}\{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n}c'(y_i;\phi), \tag{2.4}$$

em que

$$c'(y_i;\phi) = \frac{d\,c(y_i;\phi)}{d\phi}.$$

Para a obtenção da Informação de Fisher é necessário calcular a segunda derivada a segunda derivada do logaritmo da função de verossimilhança em relação ao parâmetro de

disperção ϕ, ou seja $\frac{\partial^2 l(\theta)}{\partial \phi^2} = \sum_{i=1}^{n} c''(y_i;\phi)$. Assim, a informação de Fisher para ϕ fica dada por

$$\mathbf{K}_{\phi\phi}(\theta) = -\sum_{i=1}^{n} E\left\{c''(Y_i;\phi)\right\}. \tag{2.5}$$

2.2 Estimação de β

O estimador de máxima verossimilhança (EMV) do vetor β, denotado por $\hat{\beta}$, é obtido ao resolver $\mathbf{U}_\beta = 0$. Em geral, esse sistema de equações têm que ser resolvido mediante mediante métodos numéricos. O método de Newton-Raphson é um dos métodos iterativos mais utilizados para obter os EMV de β, dado que têm como objetivo estimar as raízes de uma função de $f(x)$ e sua implementação é relativamente fácil.

Em geral, esse método utiliza a aproximação de Taylor de primeira ordem da função $f(x)$ ao redor do ponto x_0 dada por

$$f(x) = f(x_0) + (x - x_0) f'(x_0).$$

Para resolver o processo a equação $f(x) = 0$ o processo iterativo é dado por

$$x^{(m+1)} = x^m - \frac{f(x^m)}{f'(x^m)},$$

em que $x^{(m+1)}$ e x^m são os valores de x no passo $(m+1)$ e no passo m, respectivamente, $f(x^{(m)})$ a função de $f(x)$ avaliada em $x^{(m)}$ e $f'(x^m)$ a derivada da função $f(x)$ avaliada em x^m.

A expansão de Taylor de primeira ordem da função escore para β ao redor de $\beta^{(0)}$ é dada por

$$\mathbf{U}_\beta = \mathbf{U}_\beta^{(0)} + \mathbf{U}'^{(0)}_\beta (\beta - \beta^{(0)}),$$

em que $\mathbf{U}_\beta^{(0)}$ e $\mathbf{U}'^{(0)}_\beta$ representam $\mathbf{U}_\beta$ e $\mathbf{U}'_\beta$ avaliados em $\beta^{(0)}$, respectivamente. Portanto, o processo iterativo de Newton-Raphson para a obtenção do EMV de β é dada por

$$\beta^{(m+1)} = \beta^{(m)} + \left[(-\mathbf{U}'_\beta)^{-1}\right]^{(m)} \mathbf{U}_\beta^{(m)}.$$

É importante destacar que a matriz $-\mathbf{U}'_\beta$ pode não ser positiva definida, então, neste caso, pode ser substituída por $\mathbf{K}_{\beta\beta}$. Portanto, o processo iterativo de escore de Fisher, ou simplesmente Fisher, para a obtenção da estimativa de máxima verossimilhança de β é dado por

$$\beta^{(m+1)} = \beta^{(m)} + \left[\mathbf{K}_{\beta\beta}^{-1}\right]^{(m)} \mathbf{U}_\beta^{(m)}.$$

O processo iterativo de Fisher pode ser considerado como um processo iterativo de mínimos quadrados ponderados dado que $\beta^{(m+1)}$ pode ser reescrito como

$$\beta^{(m+1)} = \left(\mathbf{X}^\top \mathbf{W}^{(m)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}^{(m)} \mathbf{Z}^{(m)}, \tag{2.6}$$

em que

$$\mathbf{Z} = \eta + \mathbf{W}^{-\frac{1}{2}} \mathbf{V}^{-\frac{1}{2}} (\mathbf{y} - \mu).$$

Observações:

- ⊗ As funções de variância e de ligação entram no processo iterativo através de $\mathbf{W}$ e de $\mathbf{Z}$;
- ⊗ No caso dos modelos lineares clássicos temos que $\mathbf{Z} = \mathbf{y}$ e $\mathbf{W} = \mathbf{I}$, portanto, não precisamos de métodos iterativos;
- ⊗ $\beta^{(m+1)}$ não depende de ϕ;
- ⊗ O processo iterativo finaliza quando o critério de parada pré-estabelecido é satisfeito.

2.2.1 Análise de Diagnóstico

2.2.1.1 Deviance

A qualidade do ajuste de um MLG é avaliada através da função desvio (PAULA, 2024). Nessa perspectiva, sem perda de generalidade podemos supor que o logaritmo da verossimilhança seja dado por

$$l(\mu; \mathbf{y}) = \sum_{i=1}^{n} l(\mu_i; y_i),$$

em que $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = \mathbf{x}_i^\top \beta$. A estimativa de máxima verossimilhança de μ_i para o modelo saturado, ou seja, quando $n = p$ será denotada por $\tilde{\mu}_i = y_i$, enquanto que para o modelo ajustado $(p < n)$ será denotado por $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$, em que $\hat{\eta}_i = \mathbf{x}_i^\top \hat{\beta}$. Consequentemente, o logaritmo da função de verossimilhança usando a estimativa de máxima verossimilhança de μ_i é dado por

$$l(\hat{\mu}; \mathbf{y}) = \sum_{i=1}^{n} l(\hat{\mu}_i; y_i).$$

A função desvio fica dada por

$$D^*(\mathbf{y}; \hat{\mu}) = \phi D(\mathbf{y}; \hat{\mu}) = 2\phi [\underbrace{l(\mathbf{y}; \mathbf{y})}_{\text{modelo saturado}} - \underbrace{l(\hat{\mu}; \mathbf{y})}_{\text{modelo ajustado}}],$$

que é duas vezes a diferença do logaritmo da função de verossimilhança do modelo saturado e do logaritmo da função de verossimilhança do modelo ajustado, escalonado pelo parâmetro de dispersão ϕ.

Além disso, considerando que

$$\hat{\theta}_i = \theta(\hat{\mu}_i) \quad \text{e} \quad \tilde{\theta}_i = \theta(\tilde{\mu}_i),$$

são as estimativas de máxima verossimilhança de θ para modelos sob ajuste ($p < n$) e saturado ($p = n$), respectivamente. A função desvio para os MLGs é dada por

$$D^*(\mathbf{y};\hat{\mu}) = 2\phi \sum_{i=1}^{n} \left\{ y_i(\tilde{\theta}_i - \hat{\theta}_i) + b(\hat{\theta}_i) - b(\tilde{\theta}_i) \right\}. \tag{2.7}$$

Sob a hipótese de que o modelo ajustado está corretamente especificado, espera-se que o desvio $D^*(\mathbf{y};\mu)$ tenha uma distribuição aproximada qui-quadrado com $n-p$ graus de liberdade para ϕ grande, onde p é o número de parâmetros estimados. Valores excessivamente altos para esse desvio podem sugerir inadequação do modelo ou problemas de ajuste.

2.2.1.2 Resíduos e Observações Influentes

A definição de resíduos com propriedades conhecidas ou próximas das de resíduos normalizados é fundamental. Uma primeira proposta é considerar o resíduo ordinário da regressão linear ponderada de $\hat{z}$ contra $\mathbf{X}$, definido por

$$\mathbf{r}^* = \mathbf{W}^{\frac{1}{2}}(\hat{\mathbf{z}} - \hat{\eta}) = \hat{\mathbf{V}}^{-\frac{1}{2}}(\mathbf{y} - \hat{\mu}). \tag{2.8}$$

Assumindo que $\text{Var}(\mathbf{z}) \approx \mathbf{W}^{-1}\phi^{-1}$, tem-se aproximadamente:

$$\text{Var}(\mathbf{r}^*) \approx \phi^{-1}(\mathbf{I}_n - \mathbf{H}). \tag{2.9}$$

Com isso, pode-se definir o resíduo studentizado como:

$$t_{S_i} = \frac{\sqrt{\phi}(y_i - \hat{\mu}_i)}{\sqrt{\hat{V}_i(1-h_{ii})}}, \tag{2.10}$$

onde h_{ii} é o elemento da diagonal principal da matriz $\mathbf{H}$. Embora $\hat{\eta}$ e $\mathbf{z}$ não sejam fixos nem sigam distribuição normal, essa definição de resíduo fornece uma base adequada para o diagnóstico do ajuste.

Outra alternativa amplamente utilizada envolve os resíduos componentes do desvio. O resíduo de desvio padronizado é definido por:

$$t_{D_i} = \frac{d(y_i;\hat{\mu}_i)}{\sqrt{1-h_{ii}}}, \tag{2.11}$$

em que $d(y_i;\hat{\mu}_i)$ representa a raiz do componente do desvio da i-ésima observação, e seu sinal coincide com o de $y_i - \hat{\mu}_i$.

A partir dos resíduos, pode-se calcular a distância de Cook, que verifica a influência de cada observação sobre os parâmetros estimados do modelo. Essa medida é definida como:

$$LD_i = 2\left\{L(\hat{\beta}) - L(\hat{\beta}_{(i)})\right\}. \tag{2.12}$$

Como geralmente não se obtém uma forma analítica para $\hat{\beta}_{(i)}$, utiliza-se a aproximação de Pregibon (1981):

$$\hat{\beta}^{(1)}_{(i)} = \hat{\beta} - \frac{\hat{r}_i\sqrt{\hat{\phi}}}{(1-h_{ii})}(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_i. \tag{2.13}$$

Substituindo essa expressão na definição de LD_i, chega-se à aproximação:

$$LD_i \approx \frac{h_{ii}}{1-h_{ii}}t^2_{S_i} \tag{2.14}$$

Essa expressão permite identificar pontos potencialmente influentes de forma prática, e sua simplicidade possibilita a implementação direta no R, utilizando a expressão $D = h \times (t^2 S)/(1-h)$. Embora essa aproximação possa subestimar o verdadeiro valor de LD_i, é suficiente para destacar observações que merecem maior atenção no diagnóstico.

Além das medidas numéricas, os gráficos desempenham um papel fundamental na avaliação da qualidade do ajuste de um modelo linear generalizado. As principais representações gráficas têm como objetivo identificar observações discrepantes, avaliar a variabilidade, verificar a adequação da função de ligação e identificar pontos influentes.

Para detectar observações aberrantes, recomenda-se utilizar gráficos dos resíduos, como o resíduo de desvio t_{D_i} ou o resíduo studentizado t_{S_i}, contra a ordem das observações.

A avaliação da variabilidade pode ser feita com gráficos de t_{D_i} ou t_{S_i} em função dos valores ajustados $\hat{\mu}_i$, permitindo identificar possíveis heterocedasticidades.

Outra ferramenta gráfica importante envolve a verificação da adequação da função de ligação utilizada no modelo. Para isso, traça-se um gráfico do escore ajustado $\hat{z}_i$ contra o preditor linear $\hat{\eta}_i$. Uma tendência linear nesse gráfico indica que a função de ligação escolhida é adequada.

Gráficos para identificação de pontos influentes também são essenciais. A distância de Cook contra ordem das observações, facilitando a identificação visual de observações com alto impacto nas estimativas do modelo.

3 MODELO COM RESPOSTA NORMAL INVERSA

A distribuição Normal Inversa (NI), também conhecida como distribuição Inversa Gaussiana foi originalmente proposta por (SCHRÖDINGER, 1915), e é apropriada para modelar variáveis contínuas com assimetria positiva. Propriedades estatísticas mais abrangentes dessa distribuição foram posteriormente exploradas por (TWEEDIE, 1957). A NI tem sido amplamente aplicada em contextos das ciências médicas, da saúde, engenharias e ciências sociais (AMIN *et al.*, 2016; AMIN *et al.*, 2022). Quando a variável resposta está associada a um conjunto de covariáveis, adota-se o Modelo de Regressão Normal Inversa (MRNI), uma classe particular dos Modelos Lineares Generalizados (MLG). Nesse modelo, a média da variável resposta é uma função contínua, assimétrica positivamente, e não assume a hipótese de observações independentes e identicamente distribuídas (AKRAM *et al.*, 2020).

Sendo Y uma variável aleatória com distribuição normal inversa (ou inversa gaussiana) de média μ e parâmetro de dispersão ϕ^{-1}. Denota-se $Y \sim \text{NI}(\mu, \phi)$, cuja função de densidade de probabilidade é expressa na seguinte forma (PAULA, 2024)

$$f(y;\mu,\phi) = \sqrt{\frac{\phi}{2\pi y^3}} \exp\left\{-\frac{\phi(y-\mu)^2}{2\mu^2 y}\right\}, \quad \text{em que} \quad y>0, \quad \mu>0 \quad \text{e} \quad \phi>0. \tag{3.1}$$

A densidade 3.1 pode ser expressa em termos da família exponencial de distribuições, ou seja

$$\begin{aligned}
f(y;\mu,\phi) &= \exp\left\{\frac{1}{2}\log\left(\frac{\phi}{2\pi y^3}\right) - \frac{\phi(y-\mu)^2}{2\mu^2 y}\right\} \\
&= \exp\left\{\frac{1}{2}\log\left(\frac{\phi}{2\pi y^3}\right) - \phi\left[\frac{y^2-2\mu y+\mu^2}{2\mu^2 y}\right]\right\} \\
&= \exp\left\{\frac{1}{2}\log\left(\frac{\phi}{2\pi y^3}\right) - \frac{\phi y^2}{2\mu^2 y} + \frac{2\phi\mu y}{2\mu^2 y} - \frac{\phi\mu^2}{2\mu^2 y}\right\} \\
&= \exp\left\{\frac{1}{2}\log\left(\frac{\phi}{2\pi y^3}\right) - \frac{\phi y}{2\mu^2} + \frac{\phi}{\mu} - \frac{\phi}{2y}\right\} \\
&= \exp\left\{\phi\left[y\left(-\frac{1}{2\mu^2}\right) + \frac{1}{\mu}\right] - \frac{\phi}{2y} + \frac{1}{2}\log\left(\frac{\phi}{2\pi y^3}\right)\right\}.
\end{aligned}$$

Temos então que, $\theta = -\dfrac{1}{2\mu^2}$; $\quad b(\mu) = -\dfrac{1}{\mu} \quad$ e $\quad c(y;\phi) = \dfrac{1}{2}\log\left(\dfrac{\phi}{2\pi y^3}\right) - \dfrac{\phi}{2y}$. Fazendo μ em termos de θ, temos que

$$\theta = -\frac{1}{2\mu^2} \Leftrightarrow \mu^2 = -\frac{1}{2\theta} \Leftrightarrow \mu = \sqrt{-\frac{1}{2\theta}} \Leftrightarrow \mu = \frac{1}{\sqrt{-2\theta}} \Leftrightarrow \boxed{\mu = (-2\theta)^{-1/2}}.$$

Agora, para $b(\theta)$, temos

$$b(\mu) = -\frac{1}{\mu} \Leftrightarrow b(\theta) = -\frac{1}{1/\sqrt{-2\theta}} \Leftrightarrow \boxed{b(\theta) = -\sqrt{-2\theta}}.$$

A Esperança de Y é dado por

$$\begin{aligned} E(Y) = b'(\theta) &= \frac{d}{d\theta}[-(-2\theta)^{1/2}] \\ &= \frac{1}{2}(-2\theta)^{-1/2}(-2) = -2\theta^{-1/2}, \end{aligned}$$

mas $-2\theta^{-1/2} = \mu$, portanto,

$$E(Y) = \mu.$$

A função de variância $V(\mu)$ pode ser encontrada por meio da segunda derivada de $b(\theta)$ em termos da média, ou seja

$$\begin{aligned} V(\theta) = b''(\theta) = \frac{d\mu_i}{d\theta_i} &= \frac{\partial(-2\theta_i)^{-1/2}}{\partial\theta_i} = -\frac{1}{2}(-2\theta_i)^{-3/2}(-2) \\ &= (-2\theta_i)^{-3/2} = [(-2\theta_i)^{-1/2}]^3, \end{aligned}$$

mas, $-2\theta^{-1/2} = \mu$, portanto,

$$\boxed{V(\mu) = \mu_i^3}.$$

Consequentemente, a variância de Y é definida por

$$\text{Var}(Y) = \frac{V(\mu)}{\phi} = \frac{\mu^3}{\phi}.$$

3.1 Função Geradora de Momentos

A função geradora de momentos de Y é dada por

$$\begin{aligned} M(t;\theta,\phi) &= \exp\{\phi[b(\phi^{-1}t - \theta) - b(\theta)]\} \\ &= \exp\left\{\phi - [(-2\phi^{-1}t - 2\theta)^{1/2} - (-2\theta)^{1/2}]\right\} \\ &= \exp\left\{\phi\left[\left(2\phi^{-1}t + 2\left(-\frac{\mu^{-2}}{2}\right)\right)^{1/2} - \left(-\frac{2}{2\mu^2}\right)^{1/2}\right]\right\} \\ &= \exp\left\{\phi\left[(2\phi^{-1}t - \mu^{-2})^{1/2} + \left(\frac{1}{\mu^2}\right)^{1/2}\right]\right\} \\ &= \exp\left\{\phi\left[\left(2\phi^{-1}t - \frac{1}{\mu^2}\right)^{1/2} + \left(\frac{1}{\mu}\right)\right]\right\}. \end{aligned}$$

3.2 Estimação dos Parâmetros

3.2.1 Função de Ligação Canônica

Inicialmente, iremos considerar a função de ligação identidade da norma inversa, ou seja

$$\eta_i = \mu_i^{-2}, \quad \text{ou seja} \quad \mu_i = (\eta_i)^{-1/2} = (\mathbf{x}_i^\top \beta)^{-1/2}.$$

Agora, vamos encontrar v_i que corresponde ao $i-$ésimo elemento da matriz $\mathbf{V}_{(n\times n)}$, onde $\mathbf{V}_{(n\times n)} = \text{diag}\{v_1, \ldots, v_n\}$.

$$\begin{aligned} v_i = b''(\theta) = \frac{d\mu_i}{d\theta_i} = \frac{\partial(-2\theta_i)^{-1/2}}{\partial\theta_i} &= -\frac{1}{2}(-2\theta_i)^{-3/2}(-2) \\ &= (-2\theta_i)^{-3/2} = [(-2\theta_i)^{-1/2}]^3 = \mu_i^3. \end{aligned}$$

A matriz $\mathbf{V}_{(n\times n)}$ será, portanto, $\mathbf{V}_{(n\times n)} = \text{diag}\{(\mathbf{x}_i^\top \beta)^{-3/2})\}, \quad i = 1, 2, \ldots, n.$

Agora, encontraremos a matriz de pesos associada à nossa matriz de parâmetros β, onde o $i-$ésimo elemento é dado por

$$\begin{aligned} w_i &= \left(\frac{d(\eta_i)^{-1/2}}{d\eta_i}\right)^2 \frac{1}{\mu_i^3} = \left(-\frac{1}{2}(\eta_i)^{-3/2}\right)^2 \frac{1}{(\eta_i^{-1/2})^3} = \frac{1}{4}\eta_i^{-3}\frac{1}{(\eta_i^{-1/2})^3} \\ &= \frac{1}{4}\eta_i^{-3}\frac{1}{\eta_i^{-3/2}} = \frac{\eta_i^{-3/2}}{4}. \end{aligned}$$

Em termos da média, temos que

$$wi = \frac{(\mu_i^{-2})^{-3/2}}{4} = \frac{\mu_i^3}{4}.$$

A matriz $\mathbf{W}_{(n\times n)}$ será, portanto, $\mathbf{W}_{(n\times n)} = \text{diag}\left\{\frac{(\mathbf{x}_i^\top \beta)^{-3/2}}{4}\right\}, \quad i = 1, \ldots, n.$

O i-ésimo elemento da matriz $\mathbf{Z}_{(n\times 1)}$ será portanto

$$z_i = \eta_i - \frac{2(y_i - \eta_i^{-1/2})}{\eta_i^{-3/2}}, \quad \text{onde} \quad \eta_i = \mathbf{x}_i^\top \beta, \quad i = 1, 2, \ldots n.$$

No `R`, o processo iterativo dado em (2.6) para estimação dos parâmetros é feita por meio da função `glm`. Para especificar o ajuste do modelo com resposta normal inversa e função de ligação logarítmica é necessário acrescentar o parâmetro: `family = inverse.gaussian( link = '1/mu^2')`.

3.2.2 Função de Ligação logarítmica

Ao ajustarmos um modelo com resposta normal inversa utilizando função de ligação logarítmica, ou seja, $\log(\mu_i) = \eta_i$, onde $\eta_i = \mathbf{x}_i^\top \beta$, temos que

$$\mathbf{V}_{(n\times n)} = \text{diag}\{\exp(3\eta_i)\} \quad i = 1,2,\ldots n.$$

Além disso, o i-ésimo elemento da matriz de pesos $\mathbf{W}$ é

$$w_i = \exp(2\eta_i)\frac{1}{\exp(3\eta_i)} = \exp(-\eta_i),$$

portanto,

$$\mathbf{W}_{(n\times n)} = \text{diag}\{\exp(-\eta_i)\} \quad i = 1,2,\ldots n.$$

O i-ésimo elemento da matriz $\mathbf{Z}_{(n\times 1)}$ será portanto

$$z_i = \eta_i + \exp(-\eta_i)(y_i - \exp(\eta_i)), \quad i = 1,2,\ldots n.$$

No `R`, o processo iterativo dado em (2.6) para estimação dos parâmetros é feita por meio da função `glm`. Para especificar o ajuste do modelo com resposta normal inversa e função de ligação logarítmica é necessário acrescentar o parâmetro `family = inverse.gaussian(link = 'log')`.

3.2.3 Função de Ligação Raiz Quadrada

Ao ajustarmos um modelo com resposta normal inversa utilizando função de ligação raiz quadrada, ou seja, $\mu_i^{1/2} = \eta_i$, onde $\eta_i = \mathbf{x}_i^\top \beta$, temos que

$$\mathbf{V}_{(n\times n)} = \text{diag}\{\eta_i^6\} \quad i = 1,2,\ldots n.$$

Além disso, o i-ésimo elemento da matriz de pesos $\mathbf{W}$ é

$$w_i = 4\eta_i^2\frac{1}{\eta_i^4} = 4\eta_i^{-4},$$

portanto,

$$\mathbf{W}_{(n\times n)} = \text{diag}\{4\eta_i^{-4}\} \quad i = 1,2,\ldots n.$$

O i-ésimo elemento da matriz $\mathbf{Z}_{(n\times 1)}$ será portanto

$$z_i = \eta_i + \frac{(y_i - \eta_i^2)}{2\eta_i}, \quad i = 1,2,\ldots n.$$

3.3 Função Escore e Informação de Fisher para β

O vetor escore, apresentado em 2.2 (de dimensão $p \times 1$) e a matriz de informação de Fisher, apresentada em 2.3 (de dimensão $p \times p$) para β dependem da função de ligação escolhida para o ajuste do modelo de regressão. Nessa perspectiva, utilizando função de ligação canônica, temos que o vetor escore e a matriz de informação de Fisher são dados respectivamente, por

$$\mathbf{U}_\beta = \phi \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} 1/2 & 0 & \cdots & 0 \\ 0 & 1/2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/2 \end{bmatrix} \begin{bmatrix} (y_1 - \exp(\eta_1)) \\ (y_2 - \exp(\eta_2)) \\ \vdots \\ (y_n - \exp(\eta_n)) \end{bmatrix},$$

$$\mathbf{K}_{\beta\beta} = \phi \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \eta_1^{-3/2}/4 & 0 & \cdots & 0 \\ 0 & \eta_2^{-3/2}/4 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \eta_n^{-3/2}/4 \end{bmatrix} \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

Para o modelo ajustado com função de ligação logarítmica, temos que o vetor escore e a matriz de informação de Fisher são dados respectivamente, por

$$\mathbf{U}_\beta = \phi \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \exp(-2\eta_1) & 0 & \cdots & 0 \\ 0 & \exp(-2\eta_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \exp(-2\eta_n) \end{bmatrix} \begin{bmatrix} (y_1 - \exp(\eta_1)) \\ (y_2 - \exp(\eta_2)) \\ \vdots \\ (y_n - \exp(\eta_n)) \end{bmatrix},$$

$$\mathbf{K}_{\beta\beta} = \phi \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \exp(-\eta_1) & 0 & \cdots & 0 \\ 0 & \exp(-\eta_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \exp(-\eta_n) \end{bmatrix} \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

Para o modelo ajustado com função de ligação raiz quadrada, temos que o vetor escore e a matriz de informação de Fisher são dados respectivamente, por

$$\mathbf{U}_\beta = \phi \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} 2\eta_1^{-5} & 0 & \cdots & 0 \\ 0 & 2\eta_2^{-5} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2\eta_n^{-5} \end{bmatrix} \begin{bmatrix} (y_1 - \eta_1^2) \\ (y_2 - \eta_2^2) \\ \vdots \\ (y_n - \eta_n^2) \end{bmatrix},$$

$$\mathbf{K}_{\beta\beta} = \phi \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} 4\eta_1^{-4} & 0 & \cdots & 0 \\ 0 & 4\eta_2^{-4} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 4\eta_n^{-4} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

3.4 Função Escore e Informação de Fisher para ϕ

Utilizando 2.4, temos que a função escore para ϕ no modelo com resposta normal inversa é dada por

$$\begin{aligned} \mathbf{U}_\phi &= \sum_{i=1}^{n} \left\{ y_i \left(-\frac{1}{2\mu_i^2} + \frac{1}{\mu_i} \right) \right\} + \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \phi} \left[\frac{1}{2}\log\phi - \frac{1}{2}\log(2\pi y_i^3) - \frac{\phi}{2y_i} \right] \right\} \\ &= \sum_{i=1}^{n} \left\{ y_i \left(-\frac{1}{2\mu_i^2} + \frac{1}{\mu_i} \right) \right\} + \sum_{i=1}^{n} \left\{ \frac{1}{2\phi} - \frac{1}{2y_i} \right\}. \end{aligned}$$

A segunda derivada de $c(y_i;\phi)$ é dada por

$$\frac{\partial}{\partial \phi} \left[\frac{1}{2\phi} - \frac{1}{2y_i} \right] = -\frac{1}{2\phi^2},$$

consequentemente, utilizando 2.5, temos que a informação de fisher para ϕ é dada por

$$\mathbf{K}_{\phi\phi} = -\sum_{i=1}^{n} E\left[-\frac{1}{2\phi^2} \right] = \sum_{i=1}^{n} \frac{1}{2\phi^2} = \frac{n}{2\phi^2}.$$

3.5 Análise de Diagnóstico no Modelo com Resposta Normal Inversa

A verificação da adequação de um modelo é uma etapa essencial em qualquer análise estatística. No caso de modelos de regressão com resposta normal inversa, alguns procedimentos específicos podem ser adotados para avaliar a qualidade do ajuste e identificar possíveis observações influentes.

3.5.1 Deviance no Modelo com Resposta Normal Inversa

Uma medida comumente utilizada para avaliar a qualidade do ajuste desses modelos é o desvio, que mede a discrepância entre os valores observados e os estimados. Considerando n observações independentes, o desvio do modelo pode ser expresso como:

$$
\begin{aligned}
D^*(\mathbf{y};\mu) &= 2\phi \sum_{i=1}^{n} \left[y_i \left(-\frac{1}{2y_i^2} + \frac{1}{2\hat{\mu}_i^2} \right) - \frac{1}{\hat{\mu}_i} + \frac{1}{y_i} \right] \\
&= 2\phi \sum_{i=1}^{n} \left[\left(\frac{y_i}{2\hat{\mu}_i^2} - \frac{y_i}{2y_i^2} \right) - \frac{1}{\hat{\mu}_i} + \frac{1}{y_i} \right] \\
&= \phi \sum_{i=1}^{n} \left[\frac{y_i}{\hat{\mu}_i^2} - \frac{1}{y_i} - \frac{2}{\hat{\mu}_i^2} + \frac{2}{y_i} \right] \\
&= \phi \sum_{i=1}^{n} \left[\frac{1}{y_i} + \frac{y_i}{\hat{\mu}_i^2} - \frac{2}{\hat{\mu}_i} \right] \\
&= \phi \sum_{i=1}^{n} \left[\frac{\hat{\mu}_i^2}{y_i \hat{\mu}_i^2} + \frac{y_i^2}{\hat{\mu}_i^2 y_i} - \frac{2\hat{\mu}_i y_i}{\hat{\mu}_i^2 y_i} \right] \\
&= \phi \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{y_i \hat{\mu}_i^2}.
\end{aligned}
$$

3.5.2 Resíduos e Observações Influentes no Modelo com Resposta Normal Inversa

Para investigar a presença de observações discrepantes, podem ser utilizados os resíduos componentes do desvio padronizados da equação (2.11). No caso da normal inversa, o resíduo padronizado para a *i*-ésima observação é dado por:

$$
t_{D_i} = \sqrt{\frac{2\hat{\phi}}{1 - h_{ii}}} \cdot \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i \sqrt{y_i}}, \tag{3.2}
$$

Em que $\hat{\phi}$ é o estimador do parâmetro de dispersão. A dedução da expressão (3.2) parte da definição geral:

$$
t_{D_i} = \frac{d(y_i, \hat{\mu}_i)}{\sqrt{1 - h_{ii}}},
$$

$$
d(y_i;\hat{\mu}_i) = \pm \sqrt{2\hat{\phi} \left(y_i(\tilde{\theta}_i - \hat{\theta}_i) + \left(b(\hat{\theta}_i) - b(\tilde{\theta}_i) \right) \right)}, \tag{3.3}
$$

sendo que o sinal de $d(y_i;\hat{\mu}_i)$ é o mesmo de $y_i - \hat{\mu}_i$.

Além dos resíduos, é importante avaliar a influência de cada observação sobre o ajuste do modelo. Uma medida prática é a distância de Cook adaptada para modelos com resposta normal inversa, calculada a partir da seguinte expressão:

$$
LD_i = \frac{\hat{\phi} h_{ii}}{(1 - h_{ii})^2} \cdot \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^3}. \tag{3.4}
$$

Essa expressão pode ser deduzida a partir da fórmula da distância de Cook para MLGs, dada por

$$LD_i = \frac{h_{ii}}{1-h_{ii}} \cdot t_{S_i}^2,$$

em que t_{S_i} é o resíduo estudantilizado. No caso da normal inversa, ele assume a forma

$$t_{S_i} = \sqrt{\hat{\phi}} \cdot \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\mu}_i^3(1-h_{ii})}}.$$

Substituindo esse valor na expressão de LD_i, temos:

$$\begin{aligned} LD_i &= \frac{h_{ii}}{1-h_{ii}} \cdot \left(\sqrt{\hat{\phi}} \cdot \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\mu}_i^3(1-h_{ii})}} \right)^2 \\ &= \frac{h_{ii}}{1-h_{ii}} \cdot \frac{\hat{\phi}(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^3(1-h_{ii})} \\ &= \frac{\hat{\phi} h_{ii}(y_i - \hat{\mu}_i)^2}{(1-h_{ii})^2 \hat{\mu}_i^3}, \end{aligned}$$

chegando à fórmula apresentada anteriormente.

Valores elevados de LD_i indicam observações potencialmente influentes, cujo impacto sobre os parâmetros do modelo merece atenção. É recomendada a construção de gráficos dos resíduos t_{D_i}, dos valores de h_{ii} e dos índices LD_i para uma avaliação visual mais completa.

4 APLICAÇÃO E RESULTADOS

Modelos com distribuição Normal Inversa são especificados no `R` com `glm(family = inverse.gaussian)`. As funções de ligação `"inverse"`, `"identity"` , `"log"` e `"sqrt"` são permitidas na função `glm`. A distribuição Normal Inversa também permite a função de ligação `"1/mu^2"` (que é a função de ligação canônica prática para essa distribuição).

4.1 Análise de Biomassa Foliar em Tílias-de-Folhas-Pequenas (*Tilia cordata*)

Na primeira aplicação, foi adotado o exemplo apresentado por (DUNN; SMYTH, 2018, p. 439), o qual analisa o tamanho de folhas de limeira. Na exemplo apresentado no livro, o autor compara o ajuste do modelo com distribuição normal com o modelo com distribuição gama, ambos com função de ligação logarítmica. No ajuste apresentado, o modelo gama é mais satisfatório em comparação ao modelo com resposta normal inversa É possível acessar a base de dados no `R` por meio do pacote `GLMsData`. Todos os códigos utilizados na análise estão em (FREITAS, 2025). A base de dados é composta por 385 observações e 4 variáveis, sendo elas:

- ⊗ **Foliage**: biomassa foliar, em kg (matéria seca em estufa);
- ⊗ **DBH**: diâmetro à altura do peito da árvore, em cm (DAP em português - nomenclatura oficial do IBGE e EMBRAPA);
- ⊗ **Age**: idade da árvore, em anos;
- ⊗ **Origin**: origem da árvore, as categorias são: *Coppice* (Rebrote), *Natural* (Natural) e *Planted* (Plantado).

Na tabela 1 é mostrado as seis primeiras medições de *Foliage*, *DBH* e *Age* para cada categoria de origem de Tília-de-folhas-pequenas.

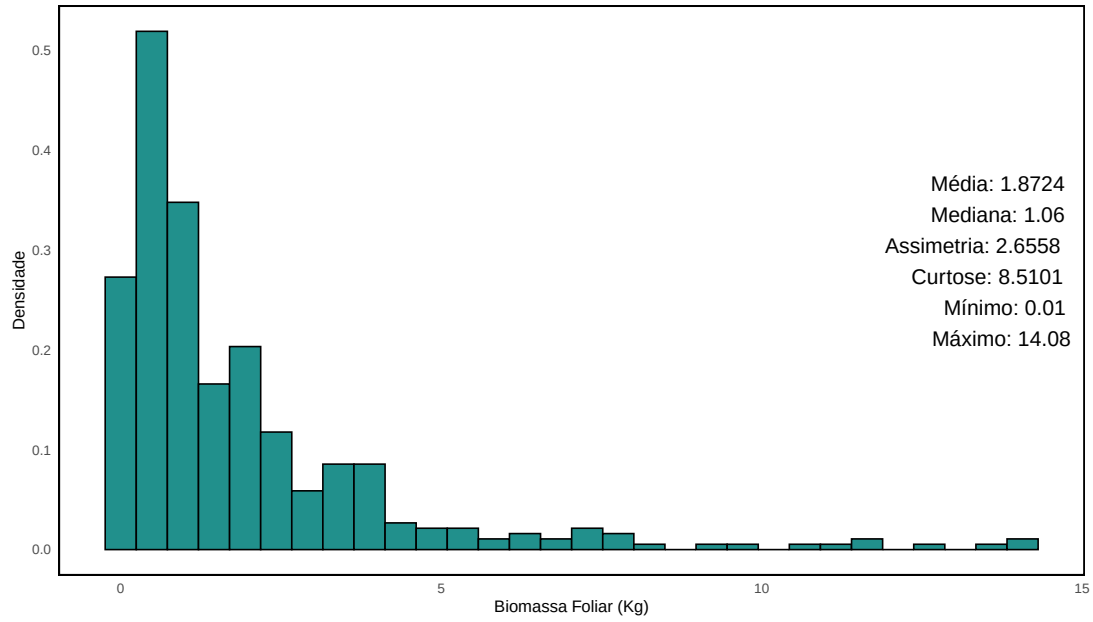
Tabela 1 – Medições de tílias-de-folhas-pequenas na Rússia, agrupadas pela origem da árvore.

Natural			**Coppice**			**Planted**		
Foliage (kg)	DBH (cm)	Age (anos)	Foliage (kg)	DBH (cm)	Age (anos)	Foliage (kg)	DBH (cm)	Age (anos)
0,10	4,00	38	0,27	7,20	24	0,92	16,40	38
0,20	6,00	38	0,03	3,10	11	3,69	18,40	38
0,40	8,00	46	0,04	3,30	12	0,82	12,80	37
0,60	9,60	44	0,03	3,10	11	1,09	14,10	42
0,60	11,30	60	0,01	3,30	12	0,08	6,40	35
0,80	13,70	56	0,07	3,30	12	0,59	12,00	32
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Fonte: (DUNN; SMYTH, 2018).

A Figura 1 apresenta o histograma de densidade da biomassa foliar (kg), o qual revela uma distribuição assimétrica positiva acentuada (coeficiente de assimetria = 2,6558) e leptocúrtica (curtose = 8,5101), caracterizada por uma concentração de dados próximos ao valor mínimo (0,0100 kg) e uma cauda direita alongada, com valores máximos atingindo 14,0800 kg. A discrepância entre a média (1,8724 kg) e a mediana (1,0600 kg) confirma a presença de assimetria, indicando que a média é influenciada por valores extremamente altos. A elevada curtose sugere uma distribuição mais pontiaguda que a normal, com excesso de observações próximas à média e nos extremos. A densidade máxima (0,5) ocorre na primeira classe de biomassa (próxima a 0 kg), com frequência decrescente à medida que os valores aumentam, evidenciando que apenas uma pequena proporção de árvores apresenta biomassa superior a 5 kg. Todas essas características presentes na distribuição da biomassa foliar direciona a análise por meio de distribuições assimétricas à direita como é o caso da distribuição Gama e da distribuição Norma Inversa, por exemplo.

Figura 1 – Distribuição de Biomassa Foliar em Tílias-de-Folhas-Pequenas (*Tilia cordata*).

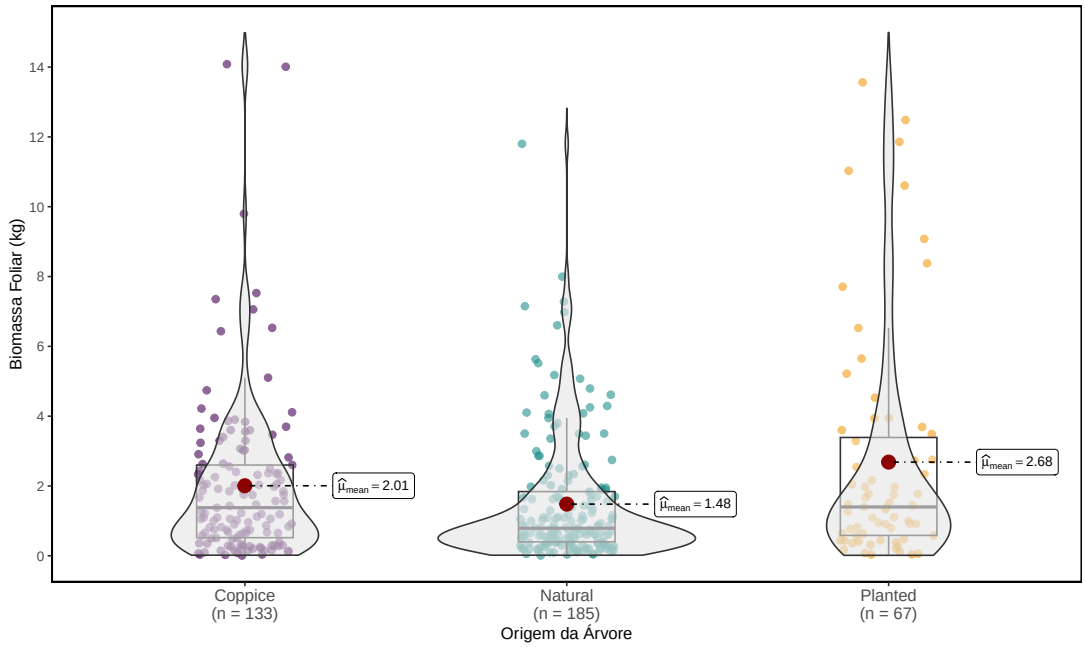


Fonte: Elaborado pelos autores a partir de (DUNN; SMYTH, 2018).

A Figura 2 apresenta a análise da distribuição da biomassa foliar (kg) em Tilia cordata por categoria de origem (*Coppice*: n = 133; *Natural*: n = 185; *Planted*: n = 67) revelou padrão assimétrico à direita em todos os grupos, caracterizado por concentração de valores na região inferior e cauda alongada na extremidade superior, com diferenças marcantes entre

médias e medianas). Essas características na distribuição dos dados, evidenciada pelo gráfico de violino, justifica teoricamente a adoção da distribuição normal inversa para modelagem, por sua adequação a dados positivos com assimetria positiva e presença de *outliers* superiores, características típicas de variáveis biométricas de crescimento vegetal. A diferença entre médias por categoria sugere ainda que a origem da árvore constitui fator relevante na predição da biomassa foliar.

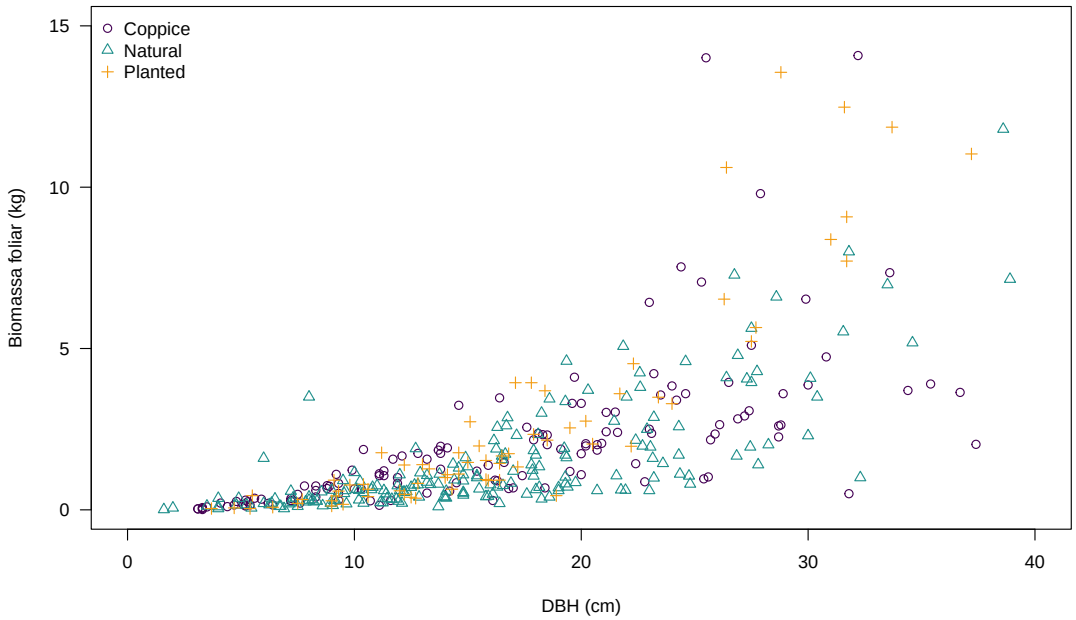
Figura 2 – Distribuição de Biomassa Foliar em Tílias-de-Folhas-Pequenas (*Tilia cordata*) por Origem da Árvore.



Fonte: Elaborado pelos autores a partir de (DUNN; SMYTH, 2018).

A Figura 3 apresenta a relação entre o diâmetro à altura do peito (DBH, em cm) e a biomassa foliar (kg) nas três categorias de origem da árvore (Coppice, Natural e Planted), revelando padrões distintos entre os grupos. Observa-se uma tendência geral positiva entre as variáveis, indicando que árvores com maior DBH tendem a apresentar maior biomassa foliar, conforme esperado em relações alométricas vegetais. A dispersão dos pontos indica variabilidade intra-grupos, sendo mais acentuada na categoria Planted, possivelmente devido a diferenças microambientais ou genéticas. A ausência de sobreposição completa entre os grupos sugere que a origem é um fator relevante na relação DBH-biomassa.

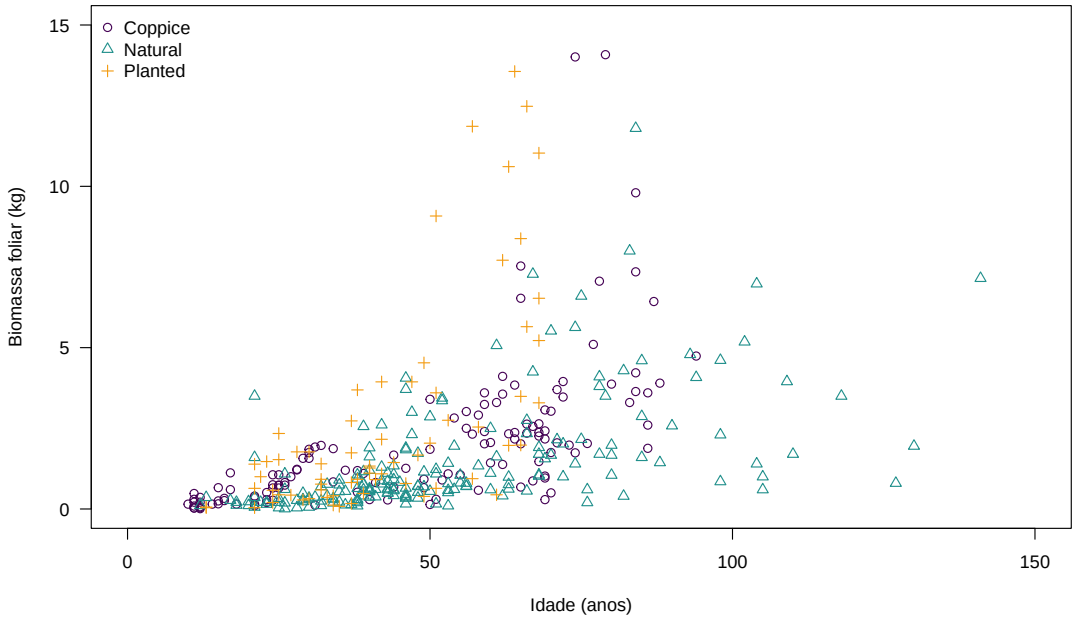
Figura 3 – Relação entre Diâmetro à Altura do Peito (DBH) e Biomassa Foliar em Tílias-de-Folhas-Pequenas (*Tilia cordata*).



Fonte: Elaborado pelos autores a partir de (DUNN; SMYTH, 2018).

A Figura 4 apresenta a relação entre a idade (em anos) e a biomassa foliar (em kg) das árvores. Verifica-se uma relação não linear, com padrões distintos conforme as diferentes fases de desenvolvimento. Nota-se um crescimento inicial da biomassa foliar proporcional à idade, que persiste até aproximadamente 30-40 anos, seguido por um período de estabilização ou mesmo um possível declínio nos indivíduos mais longevos. A significativa dispersão vertical dos pontos observada em árvores de mesma idade revela uma considerável variabilidade na produção de biomassa, o que sugere a influência de outros fatores não contemplados no gráfico, como condições edáficas, competição intraespecífica ou variabilidade genética.

Figura 4 – Relação entre Idade da Árvore e Biomassa Foliar em Tílias (*Tilia cordata*).



Fonte: Elaborado pelos autores a partir de (DUNN; SMYTH, 2018).

4.1.1 Estimação e Seleção dos Modelos

O processo de modelagem iniciou com o ajuste do Modelo 1, que incluiu todas as covariáveis disponíveis (Origin, DBH e Age) utilizando a função de ligação raiz quadrada, conforme a tabela 8. Este modelo completo apresentou o menor desvio (385,0035) e AIC (1066,3150), indicando um bom ajuste aos dados. Em seguida, para avaliar a contribuição individual do diâmetro à altura do peito (DBH), foi ajustado o Modelo 2, excluindo esta variável. Observou-se um aumento no desvio (414,5281) e no AIC (1092,7620), confirmando a importância do DBH na explicação da biomassa foliar. O Modelo 3, alternativo com apenas Idade e Origem, mostrou desempenho inferior (desvio de 956,9817 e AIC de 1414,8700), reforçando a relevância do DBH. Por fim, o Modelo 4 foi proposto para testar uma especificação distinta, mantendo apenas DBH e Idade, mas sem as variáveis de Origem. Embora seu AIC (1107,2420) tenha sido superior ao do Modelo 1, a comparação entre os modelos sugere que a inclusão simultânea de Origem, DBH e Idade (Modelo 1) fornece o melhor equilíbrio entre parcimônia e qualidade de ajuste, justificando sua seleção como modelo final.

Tabela 2 – Estimativas de máxima verossimilhança referentes ao modelos com resposta normal inversa ajustados aos dados sobre biomassa foliar em Tílias-de-Folhas-Pequenas (*Tilia cordata*).

Efeito	Modelo 1		Modelo 2		Modelo 3		Modelo 4	
	Estimativa	EP	Estimativa	EP	Estimativa	EP	Estimativa	EP
Intercepto	-0,0455	0,0141	-0,0667	0,0157	-4,7416	0,1357	0,0597	0,0210
OriginNatural	0,1455	0,0213	0,0634	0,0131	-2,1780	0,0759	-	-
OriginPlanted	-0,0301	0,0193	-0,0602	0,0194	-2,1426	0,1372	-	-
DBH	0,1049	0,0039	0,0901	0,0033	-	-	0,0834	0,0029
Age	-0,0061	0,0011	-	-	0,36523	0,0109	-0,0027	0,0009
ϕ	0,7817		1,0175		0,3321		1,1879	
Desvio	385,0035		414,5281		956,9817		432,6568	
AIC	1066,3150		1092,762		1414,87		1107,242	

Nota: A categoria de referência para a variável Origem da planta é *Coppice*. EP: Erro Padrão.

A Tabela 3 apresenta a análise de deviance comparando três modelos para biomassa foliar de Tilia cordata revelou que o modelo completo (Foliage $\sim$ Origin + DBH + Age; GL=380, Deviance = 385,0035) apresentou o melhor ajuste. A remoção da variável Age (Modelo 2: Foliage $\sim$ Origin + DBH) resultou em aumento significativo da deviance ($\Delta = -29,53; F = 37,77, p < 0,001$), assim como a exclusão de Origin (Modelo 3: Foliage $\sim$ DBH + Age; $\Delta = -47,65; F = 23,19, p < 0,001$). Ambos os testes demonstraram efeitos altamente significativos $(p < 0,001)$, indicando que tanto a idade quanto a origem das árvores contribuem substancialmente para explicar a variação na biomassa foliar, mesmo após controlar pelo diâmetro à altura do peito (DBH). O impacto mais pronunciado na deviance ao remover Origin $(-47,65 versus -29,53)$ sugere que esta variável possui maior peso explicativo no modelo. Esses resultados justificam a manutenção do modelo completo para análises subsequentes.

Tabela 3 – Análise de Deviance para Modelos de Biomassa Foliar

Modelo	Gl Resid.	Deviance	ΔDeviance	F
`Foliage ~ Origin + DBH + Age`	380	385,0035	–	–
`Foliage ~ Origin + DBH`	381	414,5281	-29.5246	$F = 37,7690 ***$
`Foliage ~ DBH + Age`	381	432,6568	-47.6533	$F = 23,1910 ***$

Nota: *** p < 0.001. ΔDeviance compara com o modelo completo. Valores negativos indicam pior ajuste ao remover variáveis.

O modelo final selecionado para a modelagem da biomassa foliar em em Tílias-de-Folhas-Pequenas possui as seguintes estimativas

$$\hat{\mu}_i = (-0,0455 + 0,1455 \cdot \text{OriginNatural}_i - 0,0301 \cdot \text{OriginPlanted}_i + 0,1049 \cdot \text{DBH}_i - 0.0061 \cdot \text{Age}_i)^2.$$

No processo de modelagem, foram consideradas outras funções de ligação, como

a logarítmica e a identidade. Na Tabela 4, estão as estimativas de ϕ, do desvio e do AIC para o modelo final ajustado para cada função de ligação. É importante destacar que, apesar de apresentar o menor desvio, a função de ligação identidade não é recomendada para esse tipo de modelagem, uma vez que o suporte da variável resposta pertence ao conjunto dos números reais positivos, e o uso dessa função de ligação pode violar essa restrição, pois os valores preditos pela ligação identidade podem ser menores do que zero.

Tabela 4 – Resumo do ajuste do modelo (`Foliage` $\sim$ `Origin + DBH + Age`) com resposta normal inversa para diferentes funções de ligação.

Função de Ligação	$\hat{\phi}$	Desvio	AIC
$\mu^{1/2}$	0,7817	385,0035	1066,3150
$\log \mu$	1,2571	532,0373	1190,8480
μ	0,7059	367,2297	1048,1180

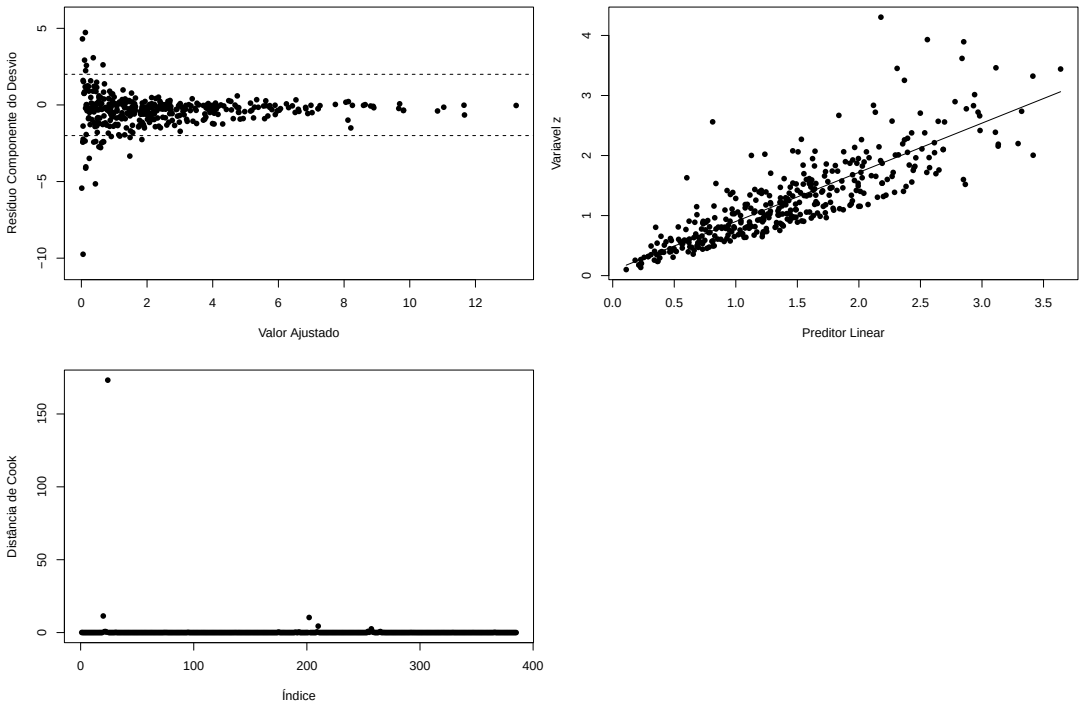
Tabela 5 – Coeficientes estimados do modelo com distribuição normal inversa

Variável	Estimativa	Erro Padrão	t	p-valor
Intercepto	-0,0455	0,0141	-3,2280	0,0014
OriginNatural	0,1455	0,0213	6,8260	<0,0001
OriginPlanted	-0,0301	0,0193	-1,5540	0,1209
DBH	0,1049	0,0039	26,6420	<0,0001
Age	-0,0061	0,0011	-5,6980	<0,0001

4.1.2 Análise de Diagnóstico do Modelo

Na Figura 7 são apresentados os gráficos de diagnóstico para o modelo ajustado

Figura 5 – Gráficos de diagnóstico referentes ao modelo com resposta normal inversa ajustado aos dados sobre biomassa foliar em Tílias-de-Folhas-Pequenas (*Tilia cordata*).



Fonte: Elaborado pelos autores.

No gráfico de resíduos de desvio contra os valores ajustados (Figura 7), observa-se que os pontos apresentam certa dispersão aleatória em torno da linha zero, o que sugere que a suposição de linearidade na escala da função de ligação foi, em geral, atendida. No entanto, nota-se a presença de um leve padrão em forma de funil, caracterizado por maior variabilidade dos resíduos para valores ajustados menores e redução dessa variabilidade à medida que os valores ajustados aumentam. Esse comportamento indica indícios de heterocedasticidade, o que pode comprometer a homogeneidade da variância dos resíduos, principalmente para as menores predições. Além disso, algumas observações ultrapassam os limites de ± 2, sugerindo possíveis valores discrepantes que podem merecer uma análise mais detalhada.

Por outro lado, o gráfico de valores ajustados contra o preditor linear (Figura 7) apresenta uma tendência linear crescente bem definida, como esperado, já que o preditor linear está diretamente relacionado à função de ligação da média da variável resposta. Esse comportamento confirma que o modelo foi capaz de capturar adequadamente a estrutura linear dos dados na escala transformada, sem evidência de distorções ou inadequações na especificação da relação entre as covariáveis e a biomassa foliar.

Tabela 6 – Coeficientes estimados do modelo com distribuição normal inversa sem a observação #24

Variável	Estimativa	Erro Padrão	t	p-valor
Intercepto	-0,0225	0,0153	-1,4700	0,1424
OriginNatural	0,1312	0,0236	5,5510	<0,0001
OriginPlanted	-0,0365	0,0196	-1,8710	0,0621
DBH	0,0918	0,0047	19,3600	<0,0001
Age	-0,0036	0,0013	-2,7860	0,0056

A Tabela 6 apresenta as estimativas do modelo ajustado sem a obervação #24. Os resultados mostraram que não houveram diferenças discrepantes com as estimativas do modelo ajustado com todas as observações (Tabela 5) e o modelo sem a observação #24, o que não a configura como um ponto de alavancagem das estimativas do modelo final.

4.1.3 Interpretação dos Parâmetros do Modelo Ajustado

O modelo de regressão normal inversa com função de ligação raiz quadrada estimado para a biomassa foliar (*Tilia cordata*) apresenta a seguinte interpretação prática:

$$\hat{\mu}_i = \left(-0,0455 + 0,1455 \cdot \text{Natural}_i - 0,0301 \cdot \text{Planted}_i + 0,1049 \cdot \text{DBH}_i - 0,0061 \cdot \text{Age}_i\right)^2$$

- ⊗ O modelo ajustado com função de ligação raiz quadrada revela importantes relações entre as variáveis preditoras e a biomassa foliar. O termo intercepto de -0,0455, quando transformado para a escala original (kg), resulta em aproximadamente $((-0,0455)^2 = 0,0021kg)$, representando a biomassa basal teórica para árvores da categoria *Coppice* com diâmetro (DBH) e idade zero. Este valor residual é biologicamente coerente, indicando que o modelo captura adequadamente o comportamento esperado na ausência dos fatores analisados.
- ⊗ As estimativas para a origem da árvore mostram diferenças significativas: árvores de origem natural apresentam incremento positivo de 0,1455 unidades na escala $kg^{1/2}$ (equivalente a 0,0212 kg na escala original) em relação à categoria de referência (Coppice), enquanto as árvores plantadas mostram redução de -0,0301 unidades $kg^{1/2}$ (0,0009 kg). Esses resultados sugerem que a procedência das árvores influencia significativamente a produção de biomassa foliar, possivelmente devido a diferenças adaptativas ou condições de crescimento.
- ⊗ A estimativa para DBH (0,1049) demonstra forte efeito positivo, onde cada centímetro

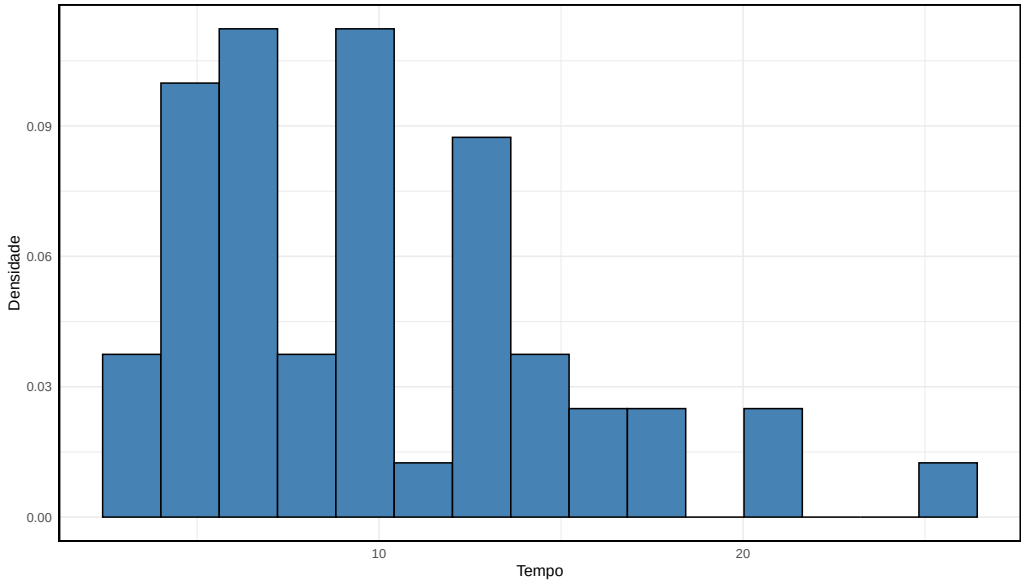
adicional no diâmetro do tronco aumenta em 0,1049 unidades a raiz quadrada da biomassa foliar. Este efeito, quando convertido para a escala original, mostra um crescimento não linear da biomassa em função do diâmetro, ressaltando a importância do desenvolvimento estrutural da árvore para a produção foliar.

⊗ Em contraste, o coeficiente negativo para idade (-0,0061) indica redução média anual de 0,0061 unidades $kg^{1/2}$ na biomassa foliar. Este resultado pode refletir processos de senescência foliar ou aumento da competição por recursos em árvores mais velhas. A transformação quadrática inerente à função de ligação garante que todas as estimativas de biomassa permaneçam positivas, adequando-se às características físicas da variável resposta.

4.2 Análise do Desempenho de Cinco Turbinas de Avião

Nesta aplicação, são utilizados dados extraídos do livro de Gilberto de Paula (2024), que também constam originalmente em Lawless (1982, p. 201). O experimento descrito tem como objetivo comparar o desempenho de cinco tipos de turbinas de alta velocidade utilizadas em motores de avião, sendo registrado, para cada uma delas, o tempo (em milhões de ciclos) até a perda de velocidade. Foram observados dez motores para cada tipo de turbina, totalizando 50 observações no conjunto de dados.

Figura 6 – Histograma dos tempos até a perda de velocidade.



Fonte: Elaborado pelos autores a partir de (PAULA, 2024).

A Figura 6 apresenta o histograma da variável tempo. A distribuição é nitidamente assimétrica à direita, com maior concentração de valores na região de menores tempos e uma cauda que se estende para valores mais altos. As principais medidas descritivas da variável tempo estão apresentadas na Tabela 7.

Tabela 7 – Medidas descritivas da variável tempo.

Medida	Média	Mediana	Moda	Assimetria
Tempo (ciclos)	9,98	9,34	3,03	0,91

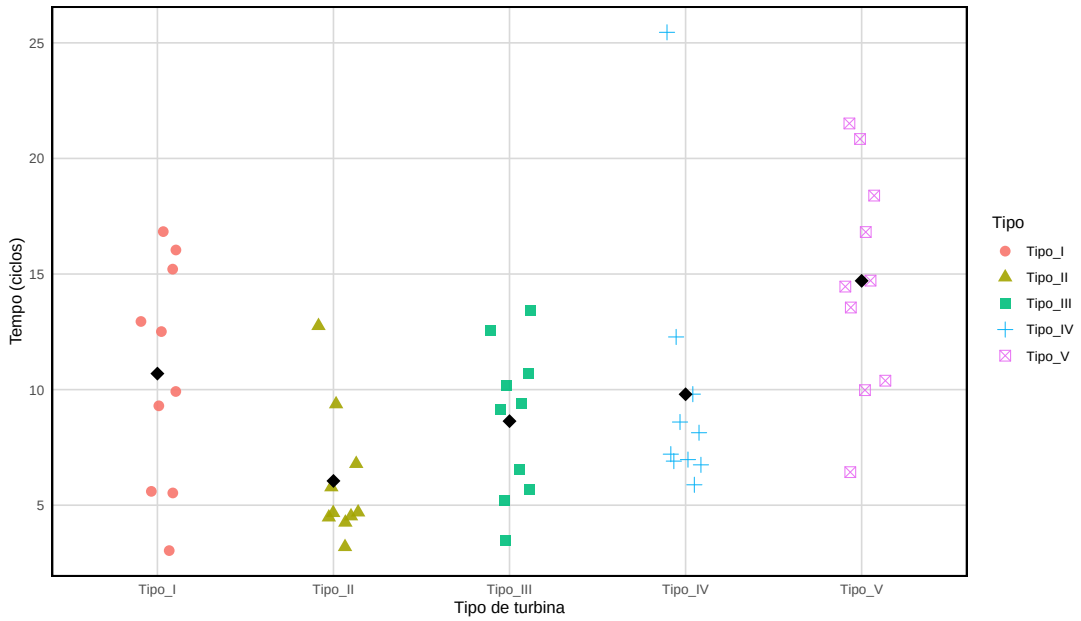
A tabela mostra que a média (9,98) é maior que a mediana (9,34), que por sua vez é maior que a moda (3,03), ou seja,

$$\text{Moda} < \text{Mediana} < \text{Média},$$

o que caracteriza uma distribuição assimétrica à direita. Essa conclusão é corroborada pelo coeficiente de assimetria positivo (0,91), indicando uma assimetria moderada a forte. Esse comportamento sugere a presença de valores extremos altos influenciando a média.

Para entender a relação entre o tempo e os diferentes tipos de turbina, considerou-se o gráfico de dispersão ilustrado na Figura 7. Nele, observa-se que o tempo até a falha varia consideravelmente entre os grupos. O Tipo V, por exemplo, apresenta os maiores tempos médios, enquanto o Tipo II tende a ter os menores. Dentro de cada grupo, a variabilidade também é notável, especialmente nos Tipos I, IV e V, que apresentam observações mais dispersas e alguns valores extremos.

Figura 7 – Gráfico de dispersão do tempo por tipo de turbina.



Fonte: Elaborado pelos autores.

4.2.1 Estimação e Seleção dos Modelos

Pela tendência da dispersão, podemos observar um leve crescimento exponencial ou linear entre os tempos e os tipos de turbinas. Com isso, optou-se por ajustar um modelo com distribuição Normal Inversa utilizando duas funções de ligação distintas: a identidade e a logarítmica. Além disso, nota-se no gráfico uma tendência de crescimento dos tempos médios ao longo dos tipos, sugerindo um comportamento levemente exponencial. Diante disso, a função de ligação logarítmica, $g(\mu) = \log(\mu)$, torna-se uma escolha apropriada, pois é capaz de linearizar relações exponenciais entre a média da variável resposta e a variável explicativa. Essa ligação também preserva a positividade da média e permite interpretar os efeitos das covariáveis de forma multiplicativa, sendo compatível com a modelagem utilizando a distribuição Normal Inversa.

Por outro lado, também foi considerado o ajuste com a função de ligação identidade, $g(\mu) = \mu$, cuja interpretação aditiva pode ser vantajosa em situações em que se deseja entender o impacto absoluto das covariáveis sobre a média da variável resposta. Assim, os modelos foram especificados como `glm(Tempo ~ Tipo, family = inverse.gaussian(link = "identity"))` para a ligação identidade, e `glm(Tempo ~ Tipo, family = inverse.gaussian(link = "log"))` para a ligação logarítmica, tendo como variável resposta o tempo até a falha e como

variável explicativa o tipo de turbina, considerado como fator com cinco níveis.

Abaixo, apresenta-se a tabela com os coeficientes estimados e os respectivos valores-p obtidos para cada um dos modelos ajustados.

Tabela 8 – Estimativas de máxima verossimilhança referentes aos modelos com resposta normal inversa com ligação identidade e logarítmica, ajustados aos dados de tempo até a falha em turbinas.

Efeito	Ligação Identidade		Ligação Logarítmica	
	Estimativa	EP	Estimativa	EP
Intercepto	10,693	1,695	2,3696	0,1585
Tipo II	-4,643	1,842	-0,5695	0,1983
Tipo III	-2,057	2,094	-0,2137	0,2131
Tipo IV	-0,895	2,254	-0,0874	0,2194
Tipo V	4,013	3,216	0,3187	0,2443
ϕ	0,02349		0,02349	
Desvio	1,0703		1,0703	
AIC	287,28		287,28	

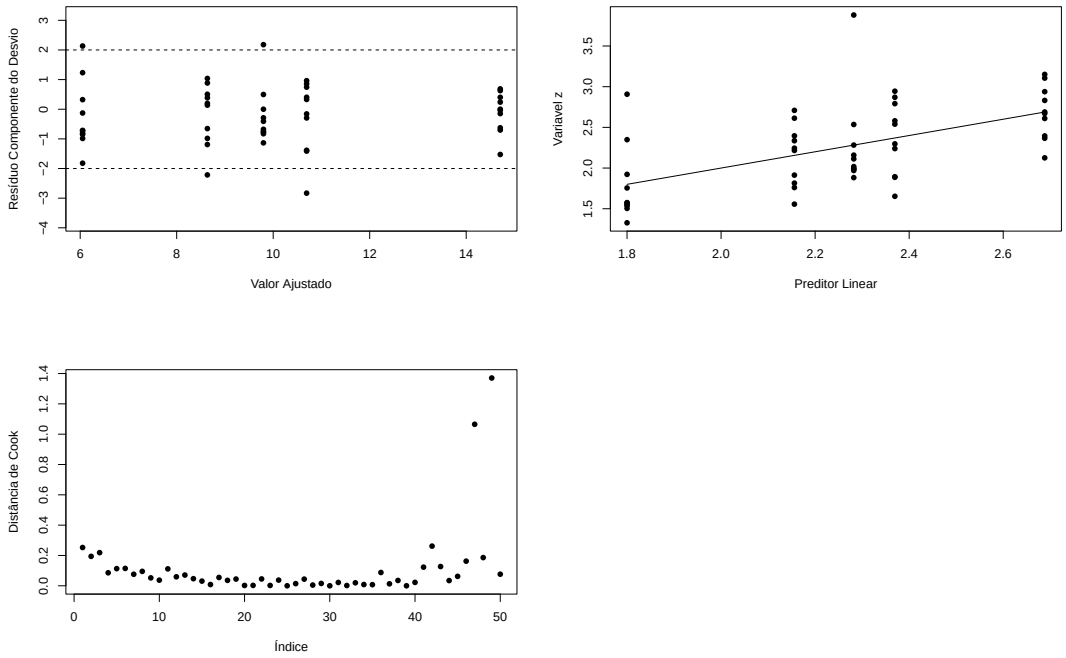
Nota: A categoria de referência para a variável *Tipo* é o Tipo I. EP: Erro Padrão.

Ambos os modelos com função de ligação identidade e logarítmica apresentam valores idênticos de AIC, desvio e parâmetro de dispersão, indicando desempenho estatístico equivalente. No entanto, considerando que os tempos são sempre positivos, a função logarítmica mostra-se mais apropriada. Portanto, opta-se pelo modelo com função logarítmica como o mais adequado para apresentar os dados.

4.2.2 Análise de Diagnóstico do Modelo

Para verificar o diagnóstico do modelo, foram analisados os principais gráficos de diagnóstico: resíduos componente do desvios versus valores ajustados, variável resposta transformada em função do preditor linear, e a distância de Cook (8)

Figura 8 – Gráficos de diagnóstico do modelo: resíduos componente do desvio vs. valores ajustados, variável transformada vs. preditor linear e distância de Cook.



Fonte: Elaborado pelos autores.

No primeiro painel da figura, observa-se que a maioria dos resíduos encontra-se dentro do intervalo entre -2 e 2. Distribuindo-se de forma aproximadamente simétrica ao redor do zero, sem padrões sistemático. Esse comportamento indica que não há sinais evidentes de heterocedasticidade.

O segundo painel mostra a variável resposta transformada em função do preditor linear. Os pontos se distribuem de forma relativamente alinhada à reta traçada, sugerindo que a função de ligação escolhida é adequada à natureza dos dados.

O terceiro painel apresenta a distância de Cook, utilizada para identificar pontos influentes. Verifica-se que a grande maioria das observações possui baixa influencia sobre os parâmetros estimados, com duas observações sendo candidatos a serem possíveis pontos fluentes.

Dessa forma, os resultados apresentados na Figura **??** indicam que o modelo se ajusta bem aos dados, não havendo evidências claras de violação das suposições. Ainda assim, a investigação das observações 47 e 49 será importante para garantir a robustez das conclusões obtidas.

Tabela 9 – Comparação dos coeficientes estimados com diferentes exclusões de observações influentes.

Parâmetro	Completo	Sem 47	Sem 49	Sem 47 e 49
Intercepto	2,3659	2,3696	2,3696	2,3696
Tipo II	−0,7117	−0,7008	−0,5695	−0,7008
Tipo III	−0,2129	−0,2136	−0,2136	−0,2136
Tipo IV	−0,0793	−0,0874	−0,2830	−0,2830
Tipo V	0,3173	0,3187	0,3187	0,3187
Deviance residual	1,4989	0,9484	0,9347	0,8128
AIC	278,95	275,71	272,92	260,69

Com o objetivo de avaliar a robustez do modelo ajustado, realizou-se uma análise de sensibilidade por meio da exclusão das observações identificadas como influentes: os pontos de índice 47 e 49, que apresentaram distância de Cook superior a 1 e resíduos componente do desvio elevados. Foram ajustados três modelos alternativos: um com a exclusão da observação 47, outro com a exclusão da observação 49, e um terceiro com a exclusão simultânea de ambas. A Tabela 9 resume os principais resultados obtidos.

A exclusão da observação 47 resultou em uma queda considerável da deviance residual (de 1,50 para 0,95) e uma leve melhora no AIC. Os coeficientes estimados pouco se alteraram, mantendo-se próximas às estimativas do modelo completo, e o efeito do nível Tipo II permaneceu significativo.

Já a exclusão da observação 49 gerou um efeito mais visível sobre os coeficientes, com uma redução no módulo da estimativa associada a Tipo II ($-0{,}71$ para $-0{,}57$), ainda que o parâmetro continuasse estatisticamente significativo. A deviance residual e o AIC também diminuíram em relação ao modelo completo, sugerindo melhora no ajuste.

A exclusão simultânea das observações 47 e 49 resultou no melhor desempenho estatístico entre os modelos considerados, com a menor deviance residual (0,81) e o menor AIC (260,69). Os resíduos apresentaram menor dispersão, e o coeficiente de Tipo II manteve-se altamente significativo, com valor similar ao observado nos outros modelos. Os níveis Tipo IV e Tipo V passaram a se aproximar do limiar de significância, indicando que sua avaliação pode depender da presença ou ausência dessas observações influentes.

Apesar da melhora nos indicadores de ajuste com a exclusão das observações 47 e 49, optou-se por manter o modelo original com todas as observações. Isso se deve ao fato de que as conclusões principais, em especial a significância do efeito de Tipo II, permaneceram consistentes mesmo diante da exclusão de pontos influentes. Essa consistência reforça a robustez

e confiabilidade do modelo ajustado.

Com isso, após a análise de diagnóstico e a avaliação de sensibilidade em relação a observações influentes, optou-se por manter o modelo original com todas as observações, visto que os resultados foram robustos às exclusões e o ajuste apresentou qualidade satisfatória.

O modelo final ajustado assume a seguinte forma:

$$\log(\mu_i) = \beta_0 + \beta_1 \cdot \text{Tipo}_{\text{II}} + \beta_2 \cdot \text{Tipo}_{\text{III}} + \beta_3 \cdot \text{Tipo}_{\text{IV}} + \beta_4 \cdot \text{Tipo}_{\text{V}}$$

Onde μ_i representa a média da distribuição para a i-ésima observação, e foi utilizada a função de ligação logarítmica.

As estimativas dos coeficientes foram as seguintes:

Tabela 10 – Estimativas dos parâmetros do modelo final

Parâmetro	Estimativa	Erro padrão	p-valor
Intercepto	2,3659	0,1527	$< 0{,}001$
Tipo II	$-0{,}7117$	0,1815	$< 0{,}001$
Tipo III	$-0{,}2129$	0,1976	0,281
Tipo IV	$-0{,}0793$	0,2038	0,695
Tipo V	0,3173	0,2270	0,173

A interpretação dos coeficientes do modelo é feita no contexto da função de ligação logarítmica. Isso significa que as diferenças entre os grupos são interpretadas em termos de proporções da média do tempo.

O intercepto representa o logaritmo da média do tempo para o grupo de referência (Tipo I). Exponenciando esse valor, obtemos a estimativa da média de tempo para esse grupo.

O coeficiente associado ao nível Tipo II foi negativo e estatisticamente significativo, indicando que esse grupo apresenta, em média, tempos menores do que o grupo Tipo I. Mais precisamente, como $\hat{\beta}_1 = -0{,}7117$, temos:

$$\frac{\mu_{\text{Tipo II}}}{\mu_{\text{Tipo I}}} = \exp(-0{,}7117) \approx 0{,}490$$

Ou seja, o tempo médio no grupo Tipo II é cerca de 49% do tempo médio no grupo de referência, mantendo as demais variáveis constantes.

Para o grupo Tipo III, o coeficiente estimado foi de $-0{,}2129$, o que indica uma média de tempo aproximadamente 19% menor que no grupo de referência ($\exp(-0{,}2129) \approx 0{,}81$). No

entanto, essa diferença não foi estatisticamente significativa ($p = 0{,}281$), de modo que não se pode afirmar com segurança que há um efeito real desse tipo no tempo.

O grupo Tipo IV apresentou um coeficiente de $-0{,}0793$, implicando em um tempo médio cerca de 7,6% menor do que o do grupo Tipo I ($\exp(-0{,}0793) \approx 0{,}924$). Assim como no caso anterior, o efeito estimado foi pequeno e não significativo ($p = 0{,}695$), sugerindo ausência de evidência de diferença entre os grupos.

Já para o grupo Tipo V, o coeficiente estimado foi positivo, com valor de 0,3173. Isso sugere um tempo médio cerca de 37% maior em relação ao grupo de referência ($\exp(0{,}3173) \approx 1{,}373$). Apesar dessa diferença relativamente expressiva em termos práticos, o p-valor associado foi 0,173, indicando que a evidência estatística não é suficiente para confirmar essa diferença ao nível de significância de 5%.

A Figura 9 apresenta o gráfico de dispersão dos dados observados, com a média ajustada pelo modelo sobreposta a cada nível da variável Tipo. Como o modelo utiliza uma função de ligação logarítmica, os valores plotados correspondem a $\exp(\hat{\mu}_i)$, representando a média estimada diretamente na escala da variável resposta.

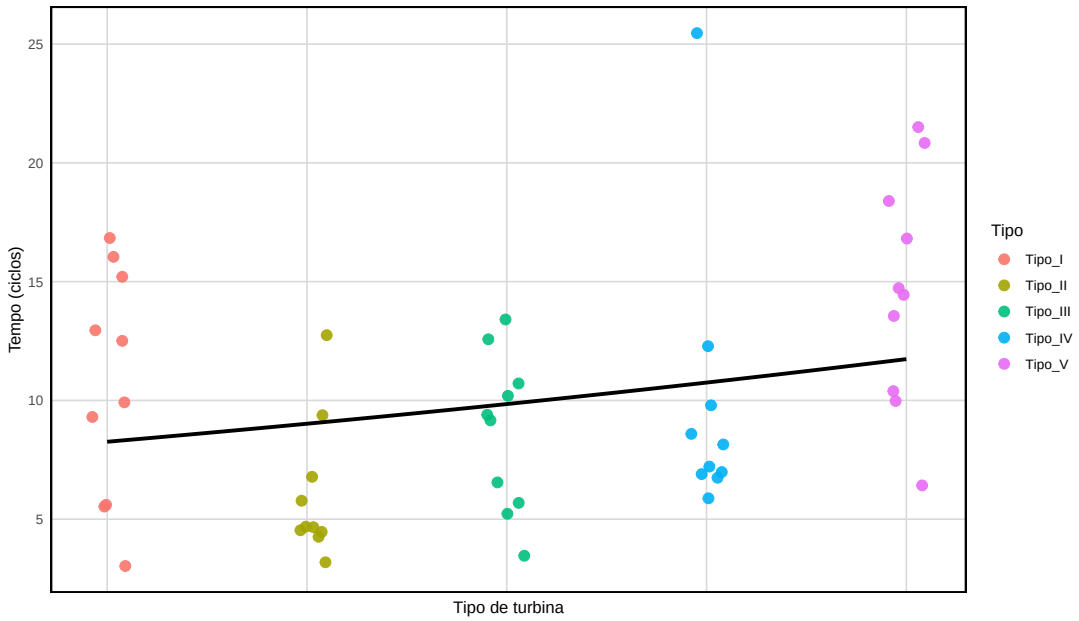


Figura 9 – Gráfico de dispersão do tempo por tipo, com curva ajustada pelo modelo final.

5 DISCUSSÃO

Na primeira aplicação sobre biomassa foliar em Tílias-de-Folhas-Pequenas, a análise gráfica apresentou indícios de uma relação exponencial entre a variável resposta (*Foliage*) e as variáveis preditoras. Nesse sentido, considerou-se inicialmente um ajuste por meio da função de ligação logarítmica. Além disso, também foram ajustados modelos com a função de ligação identidade e com função de ligação raiz quadrada. O ajuste por meio da função de ligação canônica não foi possível por conta de problemas de convergência no algorítimo de estimação, isso porque a variável resposta apresentava valores próximos de zero. O modelo ajustado com função de ligação identidade apresentou o menor AIC e o menor desvio, entretanto, ele foi adicionado na análise apenas para questões de comparação, uma vez que ele pode apresentar valores médios preditos menor que zero, ou seja, fora do espaço paramétrico da distribuição normal inversa.

Diante da dificuldade de estimação por meio das funções de ligação presentes na função `glm`, além da insatisfação com os ajuste dos modelos analisados, optou-se para a primeira aplicação o ajuste por meio da função de ligação raiz quadrada, a qual apresentou um ajuste considerado satisfatório em comparação com as demais funções de ligação. Para o uso da função raiz quadrada foi necessário a sua criação dentro do *Software* `R`, uma vez que ela não está presente de forma nativa na função `glm`.

O modelo ajustado evidencia que o diâmetro à altura do peito (DBH) é o fator com maior impacto absoluto na predição da biomassa foliar, seguido pela origem da árvore. A estrutura matemática adotada, com transformação raiz quadrada, mostrou-se adequada para capturar as relações não-lineares presentes nos dados, garantindo ao mesmo tempo a coerência biológica das predições (valores sempre positivos de biomassa). Estes resultados têm implicações práticas para o manejo florestal, sugerindo que estratégias que promovam o crescimento diamétrico podem ser mais eficazes para aumentar a produção de biomassa foliar nesta espécie.

Complementando com a análise dos tempos de falha em turbinas, foi possível observar que o modelo com resposta Normal Inversa se mostrou adequado para esse tipo de dado, caracterizado por ser positivo, contínuo e assimétrico à direita. A comparação entre diferentes funções de ligação revelou que a função logarítmica apresentou desempenho satisfatório, com menores desvios em relação à ligação identidade, embora ambas tenham fornecido ajustes plausíveis.

Apesar de alguns coeficientes não se mostrarem estatisticamente significativos ao

nível de 5%, a interpretação de todos os parâmetros contribuiu para uma visão mais completa sobre o comportamento dos diferentes tipos de turbinas em relação ao tempo até falha. A análise dos resíduos e da distância de Cook permitiu identificar observações influentes, cuja exclusão impactou as estimativas dos parâmetros, reforçando a importância do diagnóstico no ajuste de modelos com resposta Normal Inversa.

6 CONSIDERAÇÕES FINAIS

Os resultados obtidos nas análises demonstram a importância da seleção adequada da função de ligação em modelos de regressão com distribuição Normal Inversa, especialmente quando se trabalha com dados assimétricos e positivos, como os de biomassa foliar e tempo de falha de turbinas. Na primeira aplicação, a função de ligação raiz quadrada mostrou-se a mais adequada, garantindo predições biologicamente plausíveis (valores sempre positivos) e capturando eficientemente a relação não linear entre as variáveis. Apesar das dificuldades de convergência com a função de ligação canônica, a abordagem adotada permitiu um ajuste satisfatório, ressaltando a flexibilidade da distribuição Normal Inversa para modelar dados com assimetria positiva. Além disso, a análise de sensibilidade confirmou a robustez do modelo, com as estimativas dos parâmetros mantendo-se consistentes mesmo após a exclusão de possíveis pontos influentes.

Na segunda aplicação, a comparação entre as funções de ligação identidade e logarítmica evidenciou que ambas produziram resultados estatisticamente equivalentes em termos de AIC e desvio. No entanto, a função logarítmica destacou-se por sua capacidade de linearizar relações exponenciais e garantir predições positivas, sendo mais adequada para modelar tempos até falha. A análise de diagnóstico reforçou a qualidade do ajuste, com resíduos bem comportados e sem padrões sistemáticos de heterocedasticidade. A identificação e avaliação de observações influentes também foram cruciais para assegurar a confiabilidade das conclusões, demonstrando que o modelo é robusto mesmo na presença de valores extremos. Esses resultados têm implicações práticas para a engenharia de confiabilidade, fornecendo insights valiosos sobre o desempenho de diferentes tipos de turbinas.

Em síntese, este trabalho reforça a utilidade da distribuição Normal Inversa em modelos de regressão para dados assimétricos positivos, destacando a importância da escolha criteriosa da função de ligação e da realização de análises de diagnóstico rigorosas. As aplicações apresentadas ilustram como essa abordagem pode ser aplicada em contextos diversos, desde a ecologia florestal até a engenharia, fornecendo ferramentas estatísticas robustas para a tomada de decisões baseada em dados. Futuros estudos podem explorar o uso de outras funções de ligação ou a incorporação de efeitos aleatórios para capturar variações adicionais não consideradas nos modelos aqui propostos.

REFERÊNCIAS

AKRAM, M. N.; AMIN, M.; QASIM, M. A new liu-type estimator for the inverse gaussian regression model. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 90, n. 7, p. 1153–1172, 2020.

AMIN, M.; AMANULLAH, M.; ASLAM, M. Empirical evaluation of the inverse gaussian regression residuals for the assessment of influential points. **Journal of Chemometrics**, Wiley Online Library, v. 30, n. 7, p. 394–404, 2016.

AMIN, M.; ULLAH, M. A.; QASIM, M. Diagnostic techniques for the inverse Gaussian regression model. **Communications in Statistics - Theory and Methods**, v. 51, n. 8, p. 2552–2564, abr. 2022. ISSN 0361-0926, 1532-415X. Disponível em: https://www.tandfonline.com/doi/full/10.1080/03610926.2020.1777308.

CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. **Piracicaba: USP**, p. 31, 2008.

DEMÉTRIO, C. G. B. **Modelos lineares generalizados em experimentação agronômica**. [*S. l.*]: USP/ESALQ, 2001.

DUNN, P. K.; SMYTH, G. K. **Generalized linear models with examples in R**. New York, NY: Springer New York, 2018. (Springer Texts in Statistics Ser). ISBN 978-1-4419-0118-7.

FREITAS, R. B. d. **Modelo de regressão com resposta normal inversa**. 2025. Disponível em: <https://github.com/romulofreits/inverse_gaussian_model>. Acesso em: 20 jun. 2025.

MCCULLAGH, P.; NELDER, J. A. Binary data. In: **Generalized linear models**. [*S. l.*]: Springer, 1989. p. 21–31.

PAULA, G. A. **Modelos de Regressão com Apoio Computacional**. 2024. https://www.ime.usp.br/~giapaula/texto_2024.pdf. Instituto de Matemática e Estatística, Universidade de São Paulo. Disponível em: https://www.ime.usp.br/~giapaula/texto_2024.pdf.

SCHRÖDINGER, E. Zur theorie der fall-und steigversuche an teilchen mit brownscher bewegung. **Physikalische Zeitschrift**, v. 16, p. 289–295, 1915.

TWEEDIE, M. C. Statistical properties of inverse gaussian distributions. i. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 28, n. 2, p. 362–377, 1957.