Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Eduardo Moreira Araújo 202043700
Johnata Alves Moura da Silva 180020340
Júlia Borges Nunes Lira 211039063

Modelos de sobrevivência e sua relação com os modelos lineares generalizados: uma abordagem teórica e aplicada sobre estimação, inferência estatística e interpretação de dados

Brasília, DF
30 de Novembro de 2023

Conteúdo

1 Introdução

Os modelos de sobrevivência são fundamentais em estatística quando o objetivo é analisar o tempo até a ocorrência de determinado evento de interesse. O fenômeno sob estudo pode envolver o tempo até a falha de um equipamento, o tempo até a morte de um paciente ou o tempo até a implementação de uma política pública, por exemplo. Nesse sentido, a análise de sobrevivência incorpora técnicas estatísticas poderosas e que podem ser aplicadas em diversas áreas como a médica, indústria, entre outras. Uma característica importante dessa classe de modelos é a possibilidade de trabalharmos com dados censurados, nos quais uma parte dos elementos ou indivíduos do estudo não experimentam o evento de interesse.

Na modelagem paramétrica nos baseamos em distribuições de probabilidade capazes em explicar a variável tempo e a probabilidade de sobrevivência ou falha. Nem todos os modelos propostos fazem parte da família exponencial, pré-requisito para serem considerados da classe dos modelos lineares generalizados. Além disso, a presença de dados censurados também muda a configuração da modelagem dos dados. A grande questão que surge é: podemos aproveitar a base de conhecimento dos MLGs e aplicar nos modelos de sobrevivência? A resposta é positiva e esse processo nos permite utilizar a estrutura dos MLGs para realizar estimação, inferência e interpretar resultados a partir de um modelo de regressão de Poisson. O pressuposto principal é que a distribuição escolhida para o tempo de sobrevivência deve fazer parte dos modelos de riscos proporcionais, no qual separamos a função de risco em um termo que depende apenas do tempo e outro que depende apenas de um conjunto de covariáveis por meio de um preditor linear.

O presente trabalho se utiliza da distribuição Weibull para modelar o tempo de sobrevivência. Essa é uma distribuição bastante utilizada na área por sua flexibilidade em modelar funções de risco. Além disso, tal distribuição faz parte dos modelos de riscos proporcionais, o que nos levará a aplicar um modelo de regressão pela abordagem dos MLGs a partir da distribuição Poisson (membro da família exponencial). Embora não seja pré-requisito, a distribuição Weibull com o parâmetro de escala fixado faz parte da família exponencial.

A estrutura do relatório está dividida em três grandes partes: Na primeira desenvolvemos a **Metodologia**, na qual discutimos teoricamente conceitos inicias de análise de sobrevivência, sua relação com os modelos lineares generalizados, a caracterização da distribuição Weibull e os métodos de estimação e inferência. Na segunda parte apresentamos os **Resultados** com uma análise descritiva dos dados, comparação entre as funções *glm* e *survreg* do software R, ajuste de um modelo de regressão, análise de resíduos e pontos de influência. Por fim, resumimos os resultados e apresentamos sugestões e limitações do estudo na parte de **Conclusão**.

A manipulação, visualização e análise dos dados foram realizadas por meio do

software R, através da interface RStudio na versão 4.3.1. Os pacotes utilizados foram: *tidyverse*, *ggplot2*, *survival* e *AdequacyModel*. Em anexo apresentamos o código completo. A aplicação foi realizada a partir de um banco de dados disponibilizado pela Fiocruz em um estudo acerca do efeito da adesão ao tratamento antirretroviral na falha terapêutica (virológica, imunológica, clínica ou óbito), realizado com pacientes assistidos no Ipec/Fiocruz em 2009.

2 Metodologia

2.1 Modelos de Sobrevivência

2.1.1 Conceitos e Fórmulas

"A análise de sobrevivência, eventualmente chamada de análise de sobrevida, será utilizada quando o tempo for o objeto de interesse, seja esse interpretado como **o tempo até a ocorrência de um evento** ou o **risco de ocorrência de um evento por unidade de tempo.**"(CARVALHO et al., 2011)

Nos dados de sobrevivência é comum a perda da informação temporal completa, no entanto, esses dados não são descartados, visto que ainda fornecem informações sobre o tempo em que os indivíduos estiveram expostos ao risco e omiti-los pode acarretar em conclusões viesadas na análise estatística. A essas observações parciais é dado o nome de **censura**.

Segundo Colosimo e Giolo (2021), conjuntos de dados de sobrevivência são caracterizados pelos tempos de falha e de censura, que correspondem à variável resposta. Geralmente, também são medidas covariáveis para cada indivíduo. O **tempo de falha** é constituído pelo **tempo inicial**, estabelecido no início do estudo, pela **escala de medida** e pelo **evento de interesse**, ou seja, a falha, que pode ocorrer por um único motivo ou vários, nesse caso denominados *riscos competitivos* (PRENTICE et al., 1978).

A informação da variável resposta associada a cada indivíduo é representada pela indicadora abaixo:

$$\delta_i = \begin{cases} 1 \text{ caso } t_i \text{ seja tempo de falha} \\ 0 \text{ caso } t_i \text{ seja tempo de censura} \end{cases}$$

Ademais, considerando uma única v.a contínua do tempo de falha (ou de sobrevivência) não-negativa, representada por T, podemos defini-la por algumas funções, conforme Lawless (2011):

Seja f(t) a densidade de probabilidade (f.d.p) de T e F(t) a função de distribuição acumulada (f.d.a) sendo

$$F(t) = Pr(T \leq t) = \int_0^t f(x)\, dx \tag{2.1.1}$$

A probabilidade de um indivíduo sobreviver até o tempo t é dada pela função de sobrevivência

$$S(t) = Pr(T \geq t) = \int_t^\infty f(x)\, dx \tag{2.1.2}$$

S(t) é uma função monótona decrescente e contínua, com $S(0) = 1$ e $S(\infty) = \lim_{t\to\infty} S(t) = 0$. Outra função muito importante é a de taxa de falha (ou de risco), que especifica a taxa instantânea de falha no tempo t dado que o indivíduo sobreviveu até o tempo t, definida como

$$h(t) = \lim_{\Delta_t \to 0} \frac{Pr(t \leq T < t | T \geq t)}{\Delta_t} = \frac{f(t)}{S(t)} \tag{2.1.3}$$

Também é útil definir a acumulada da função de risco

$$H(t) = Pr(T \leq t) = \int_0^t h(x)\, dx \tag{2.1.4}$$

É possível, ainda, relacionar as funções apresentadas a partir de suas propriedades.

A fim de estimar a função de sobrevivência na presença de censura, o procedimento mais indicado é por Kaplan-Meier, visto que é EMV de S(t). Supondo um estudo com n indivíduos, k falhas distintas nos tempos $t_1 < t_2 < ... < t_k$ distintos e ordenados de falha, e d_j o número de falhas em $t_j, j = 1, \ldots, k$, e n_j o número de indivíduos sob risco em t_j. O estimador é definido por:

$$\hat{S(t)} = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right) \tag{2.1.5}$$

Na presença de variáveis regressoras categóricas, o objetivo da análise torna-se comparar o efeito das categorias no tempo de sobrevivência. Para isso, é necessário verificar a suposição de **riscos proporcionais**, por meio do comportamento das curvas de sobrevivência de cada grupo, quando a razão das funções de risco dos grupos a serem comparados é aproximadamente constante (curvas não se cruzam).

É sabido que a v.a T apresenta, frequentemente, forte assimetria e que a distribuição Normal não é adequada. São distribuições comuns para modelar o tempo de sobrevivência: Exponencial, Weibull, Log-normal, Log-logística, etc

Dessa forma, existem várias formas que o gráfico da função de taxa de falha de T pode assumir, sendo necessário escolher o modelo mais adequado, o que é possível por exemplo, graficamente, a partir do gráfico do tempo total em teste (curva TTT) ou também pelo gráfico da função de risco acumulada estimada, $\hat{H}(t)$, mais indicado quando o número de censuras é grande.

2.1.2 Relação com os Modelos Lineares Generalizados

A regressão linear clássica é um dos principais modelos estatísticos para se trabalhar com dados em termos descritivos e preditivos. A sua concepção geral é relacionar uma variável resposta Y com uma ou mais variáveis que podem ser consideradas explicativas para o próprio fenômeno sob estudo. Nesse sentido, temos a seguinte configuração:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + \epsilon$$

No modelo acima, cada X_i representa uma possível variável que ajuda a explicar a variável resposta Y e os valores de β_i nos permitem quantificar tal associação. Além disso, temos uma componente aleatória ϵ para captar a incerteza do processo, uma vez que estamos lidando com fenômenos não determinísticos. Essa componente aleatória é caracterizada como os erros do modelo e é uma estrutura chave para a modelagem por regressão linear. Para a viabilidade do modelo, os erros devem atender certos pressupostos: independência, homocedasticidade, normalidade, entre outros.

Se os erros do modelo de regressão linear seguem uma distribuição Normal, então, a variável resposta também é modelada por uma distribuição Normal. Nessa perspectiva, surge o seguinte questionamento: como modelar certos fenômenos que, não necessariamente, são descritos por uma distribuição de probabilidade contínua como a Normal? Como lidar com dados de contagem, binários ou de múltiplas categorias, por exemplo? É nesse cenário que os modelos lineares generalizados (MLGs) surgem como uma ferramenta poderosa e mais ampla.

Os modelos lineares generalizados formam um conjunto maior de modelos de regressão, o que nos permite modelar dados de diferentes tipos, sejam eles contínuos, discretos, binários, de contagem, categóricos, etc. Nesse sentido, podemos trabalhar com o modelo de regressão linear com erros normais citado acima ou modelos não lineares como o logístico, exponencial, poisson, entre outros. No caso dos modelos não lineares, há um processo de linearização na modelagem dos dados (KUTNER; NACHTSHEIM; NETER, 2013).

Nos MLGs a natureza da variável resposta está relacionada com a distribuição de probabilidade do processo e dos próprios erros do modelo. Nesse sentido, a componente aleatória especifica a distribuição da variável aleatória $Yi|xi$, sendo esta Normal ou oriunda de uma outra distribuição. Uma característica essencial é que a distribuição de probabilidade em questão deve ser da família exponencial. Vejamos um resumo da estrutura do modelo, segundo Kutner, Nachtsheim e Neter (2013):

- Y_1,..., Y_n são n variáveis respostas independentes que seguem uma distribuição de probabilidade pertencente à família exponencial de distribuições de proba-

bilidade, com valor esperado $E[Yi] = \mu_i$. A família exponencial é composta por distribuições como a Normal, Gamma, Poisson, Binomial, entre outras.

- Um preditor linear baseado nas variáveis preditoras X_{i1},..., $X_{i,p-1}$ é utilizado, denotado por $X^T\beta$. Trata-se de uma componente sistemática que relacionada o parâmetro η_i ao vetor de covariáveis, por meio de uma combinação linear do vetor $\beta = (\beta_1, ..., \beta_{p-1})^T$, ou seja, $\eta_i = X^T\beta = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1}X_{p-1}$.
- Existe uma função de ligação que relaciona o preditor linear à média da variável resposta. Essa função precisa ser monótona e diferenciável, servindo como uma função que lineariza a relação determinística entre as variáveis. A relação pode ser expressa da seguinte forma: $g(\mu_i) = X^T\beta$.

Para exemplificar os MLGs, podemos pensar na distribuição de Poisson. Essa distribuição é indicada, por exemplo, na modelagem de dados de contagem por unidade de medida contínua como tempo ou área. A sua função densidade de probabilidade é dada por:

$$P(y;\lambda) = \frac{\exp^{-\lambda}\lambda^y}{y!} \quad com\ \lambda > 0\ e\ y \geq 0$$

A distribuição Poisson é membro da família exponencial com a seguinte esquematização: $\theta = \log\lambda$, $b(\theta) = \exp(\theta)$, $c(y,\phi) = -\log y!$, $a(\phi) = 1$ e $E[Y] = \mu = \lambda$.

$$P(y;\theta,\phi) = \exp\left[\left(\frac{y\theta - b(\theta)}{a(\phi)}\right) + c(y,\phi)\right] = \exp(y\log\lambda - \lambda - \log y!)$$

Existem muitas possibilidades para a função de ligação, mas a mais utilizada é a canônica considerando $\mu = \exp(\theta) = \exp(X^T\beta)$, logo, temos a função de ligação:

$$\log\mu = \log\lambda = X^T\beta$$

Após entendermos um pouco mais sobre os MLGs, uma dúvida surge: qual a relação entre os modelos de sobrevivência e os modelos lineares generalizados? Em suma, veremos que as análises para dados em sobrevivência, sob certas condições, podem utilizar o $framework$ dos MLGs em termos de estimação e inferência, considerando a própria distribuição Poisson mencionada acima.

Os modelos de sobrevivência paramétricos possuem como variável resposta o tempo decorrido até a ocorrência de determinado fenômeno ou evento de interesse. Algumas das principais distribuições utilizadas para modelar o tempo nesse campo não pertencem à família exponencial, logo, não fazem parte da classe dos MLGs. Como exemplo, podemos citar a distribuição Weibull com todos os parâmetros desconhecidos e a distribuição log-logística. Apesar disso, ainda assim é possível utilizar

a estrutura dos modelos lineares generalizados para esse tipo de dados que apresentam censura, a partir da suposição de riscos proporcionais.

Nos modelos de riscos proporcionais a função de risco pode ser dividida em dois elementos: o primeiro é $h_o(t)$ e depende apenas do tempo, ou seja, não está ligado ao preditor linear. Esse termo corresponde à própria função de risco no nível de referência das covariáveis (*baseline hazard*). O segundo termo depende apenas das variáveis explicativas e possui uma característica de efeito multiplicativo na função de risco (LINDSEY, 1997). Vejamos a fórmula:

$$h(t; x) = h_o(t) \exp^{X^T \beta}$$

Fazem parte dos modelos de riscos proporcionais distribuições como a Exponencial, Weibull, Weibull Generalizada e a Gombertz, por exemplo. No caso da distribuição Exponencial $h_o(t) = 1$, ou seja, a função de risco não depende do tempo. Na distribuição Weibull temos $h_o(t) = \gamma \alpha t^{\gamma - 1}$. Há uma relação entre os modelos de riscos proporcionais e a distribuição de Poisson, o que nos permite trabalhar com dados censurados como modelos lineares generalizados. Isso ocorre pois a verossimilhança dos modelos de riscos proporcionais é também proporcional em relação à verossimilhança de n variáveis aleatórias independentes com distribuição Poisson. Esse processo aleatório envolve a contagem do número de falhas, ou seja, temos: $D_j \sim Pois(\mu_j)$ com $\mu_j = (\alpha y_j{}^{\gamma}) \exp^{X_j{}^T \beta}$ (DOBSON; BARNETT, 2008).

Existe uma outra opção um pouco mais restritiva para trabalharmos com modelos de sobrevivência sobre a perspectiva dos MLGs. A ideia básica é utilizarmos uma classe de modelos denominada *accelerated failure time*. Nesse tempo de modelagem, considera-se o logaritmo do tempo como variável resposta e há a inclusão de um termo de erro que segue alguma distribuição de probabilidade (SAIKIA; BARMAN, 2017). A diferença é que aqui os efeitos das variáveis explicativas estão relacionadas diretamente com a função de sobrevivência, conforme vemos a seguir:

$$S(t \,|\, x) = s_0(\exp(-X^T \beta)t)$$

Aqui, $S(t \,|\, x)$ é a função de sobrevivência no tempo t e $s_0(\exp(-X^T \beta)t)$ é a função de sobrevivência no nível de referência das variáveis explicativas no tempo t. A função de risco obtida nesse tipo de modelo não é a mesma obtida pelos modelos de riscos proporcionais, exceto pela distribuição Weibull, única distribuição que possui ambas as propriedades (*accelerated failure time* e *proportional hazard*). Em suma, se $X^T \beta$ decresce, então o tempo até a falha acelera e a função sobrevivência decai mais rapidamente, do contrário, $X^T \beta$ cresce e o tempo até a falha desacelera e a função sobrevivência decai mais lentamente (DOBSON; BARNETT, 2008).

Os modelos AFT são também chamados de modelos de locação-escala e, quando utilizados o logaritmo do tempo de sobrevivência, temos a seguinte relação:

$$\log T = \mu + \beta_1 X_1 + ...\beta_p X_p + \sigma\epsilon$$

Nessa estrutura μ é o intercepto e σ o parâmetro de escala. Além disso, ϵ corresponde aos erros do modelo e segue uma distribuição de probabilidade especificada. Se T segue uma distribuição Weibull, então ϵ e, consequentemente $\log T$, segue uma distribuição do valor extremo. Se T segue uma distribuição Log-Normal, então, $\log T$ segue uma distribuição Normal. Se T segue uma distribuição Log-logística, então, $\log T$ segue uma distribuição Logística (SAIKIA; BARMAN, 2017).

Caso a distribuição de probabilidade do logaritmo dos tempos de sobrevivência faça parte da família exponencial, podemos utilizar a estrutura dos modelos lineares generalizados, sobretudo, no processo de estimação. Quando T é modelada por uma distribuição Log-Normal, por exemplo, a distribuição do logaritmo é a própria Normal, que faz parte da família exponencial. Nesse sentido, podemos estimar μ e σ da Normal e chegar até os parâmetros da distribuição Log-Normal a partir de uma reparametrização, em que $\alpha = \exp^{\mu}$ e $\gamma = 1/\sigma$. Essa relação entre os modelos AFT e os modelos lineares generalizados, entretanto, é mais complexa. Alguns autores indicam que essa relação vale para dados de sobrevivência com nenhuma ou pouca censura. Em sites especializados em estatística é possível verificar que alguns autores indicam que o uso da regressão Poisson discutido para modelos de riscos proporcionais também é válida para os modelos de tempo de falha acelerado. O foco deste trabalho está na classe dos modelos de riscos proporcionais, visto que o referencial teórico é mais robusto no material consultado.

2.2 Modelagem: Distribuição Weibull

2.2.1 Descrição do Modelo Weibull

A distribuição proposta por Weibull (1939) usada inicialmente para representar a distribuição da força/resistência de materiais. Posteriormente, Weibull (1951) também descreve algumas das aplicações da distribuição. Por ser uma transformação da exponencial, o parâmetro de forma γ traz uma certa flexibilidade para o modelo (JOHNSON; KOTZ, 1970).

Uma variável aleatória X tem distribuição Weilbull se:

$$Y = \left(\frac{X - \xi_0}{\alpha}\right)^{\gamma}$$

onde $\gamma > 0$, $\alpha > 0$ e Y tem distribuição exponencial com FDP dada por:

$$p_Y(y) = e^{-y}, \quad y > 0$$

E a FDP da v.a X é:

$$p_X(x) = \frac{\gamma}{\alpha}\left(\frac{X - \xi_0}{\alpha}\right)^{\gamma-1} e^{-\{(x-\xi_0)/\alpha\}^{\gamma}}, \quad x > \xi_0$$

Usualmente, pode-se fazer o parâmetro de localização 0, de forma que a função de densidade começa a crescer a partir do 0. Assim, temos que:

$$p_X(x) = \frac{\gamma}{\alpha}\left(\frac{x}{\alpha}\right)^{\gamma-1} e^{-\{x/\alpha\}^{\gamma}} = \frac{\gamma x^{\gamma-1}}{\alpha^{\gamma}} e^{-\{x/\alpha\}^{\gamma}}$$

De forma que γ é o parâmetro de forma e α é o parâmetro de escala. Logo, a distribuição exponencial é um caso particular da Weibull com $\gamma = 1$.

A função de distribuição acumulada é dada por:

$$F_X(x) = 1 - e^{\{x/\alpha\}^{\gamma}}$$

e a função de sobrevivência é:

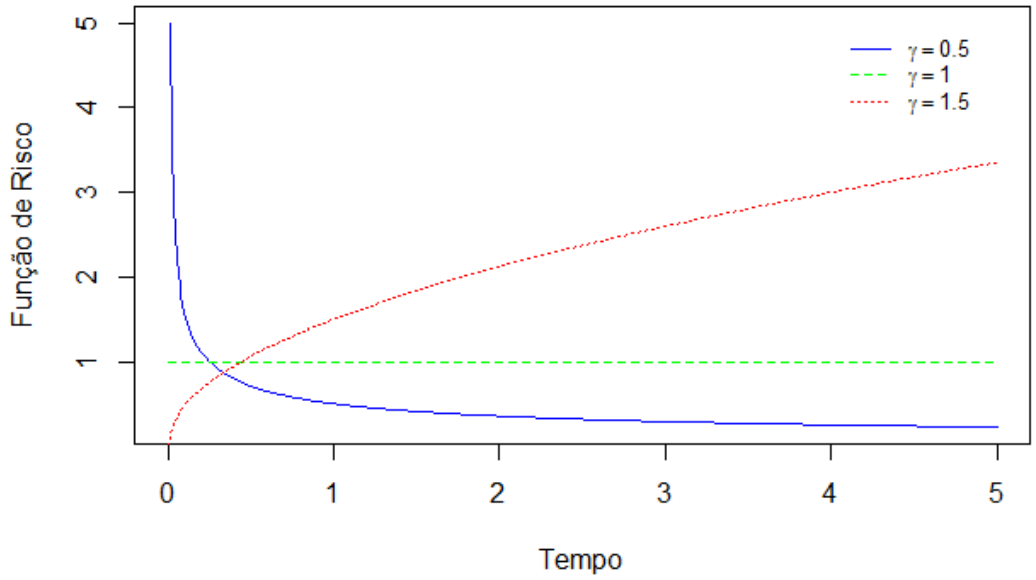
$$S_x(x) = 1 - F_X(x) = e^{\{x/\alpha\}^{\gamma}}$$

Logo, obtemos a função de risco como:

$$h_X(x) = \frac{p_X(x)}{S_x(x)} = \alpha^{-\gamma}\gamma x^{\gamma-1}$$

Assim, é possível observar que a função de risco é decrescente para $\gamma < 1$, constante para $\gamma = 1$ e crescente para $\gamma > 1$.

Figura 1: Função de risco para diferentes valores de γ



A função geradora de Momentos é dada por:

$$E[X^r] = \int_0^\infty x^r \frac{\gamma}{\alpha}\left(\frac{x}{\alpha}\right)^{\gamma-1} e^{-\{x/\alpha\}^\gamma} dx$$

fazendo a substituição $u = \left(\frac{x}{\alpha}\right)^\gamma$, de forma que $\frac{du}{dx} = \gamma\left(\frac{x}{\alpha}\right)^{\gamma-1}\alpha^{-1}$ e $x = \alpha u^{1/\gamma}$, assim temos que $u \to 0$ quando $x \to 0$ e $u \to \infty$ quando $x \to \infty$. Logo, juntando esses resultados, temos que:

$$E[X^r] = \int_0^\infty \alpha^r u^{r/\gamma} e^{-u} du = \alpha^r \int_0^\infty u^{\left(\frac{r}{\gamma}+1\right)-1} e^{-u} du = \alpha^n \Gamma\left(1 + \frac{r}{\gamma}\right)$$

E então, facilmente encontramos o valor esperado e variância da Weibull:

$$E[X] = \alpha\Gamma\left(1 + \frac{1}{\gamma}\right)$$

$$var[X] = \alpha^2\left[\Gamma\left(1 + \frac{2}{\gamma}\right) - \left(\Gamma\left(1 + \frac{1}{\gamma}\right)\right)^2\right]$$

2.2.2 Método de Estimação

Em um primeiro momento precisamos definir algumas notações importantes. Nesse sentido, n é o número de observações, x_j um vetor de variáveis explicativas, y_j o tempo de sobrevivência e δ_j uma indicação de censura em que $\delta_j = 1$ o tempo de sobrevivência é não censurado e $\delta_j = 0$ se o tempo é censurado. O método de estimação clássico em análise de sobrevivência é por máxima verossimilhança, dessa forma, vamos incorporar a informação sobre os dados não censurados a partir da função densidade da Weibull e a informação de observações censuradas é atribuída à função de sobrevivência. Portanto, obtemos a seguinte função de verossimilhança:

$$L = \prod_{j=1}^{n} f(y_j)^{\delta_j} S(y_j)^{1-\delta_j}$$

O próximo passo é obter a função log-verossimilhança e substituir a função densidade pela função de risco. Dessa forma, conseguimos o seguinte resultado:

$$l = \sum_{j=1}^{n} [\delta_j \log h(y_j) + \log S(y_j)]$$

A função log-verossimilhança dependerá dos parâmetros da distribuição de probabilidade e dos parâmetros da componente linear $X^T\beta$. Esses parâmetros podem ser estimados por diversos métodos iterativos como o Fisher Scoring ou Newton-Rapshon. Esse segundo método é um dos mais utilizados e a inversa da matriz de informação utilizada no processo de iteração nos retorna uma estimativa assintótica da matriz de variância e covariância dos parâmetros estimados (Dobson e Barnett, 2008).

Considerando que na análise exploratória dos dados observou-se que o tempo de sobrevivência pode ser modelado de forma satisfatória pela distribuição Weibull, então podemos estabelecer a função de log-verossimilhança na forma à seguir:

$$l = \sum_{j=1}^{n} [\delta_j \log(\gamma\alpha {y_j}^{\gamma-1} \exp^{{X_j}^T\beta}) - (\alpha {y_j}^{\gamma} \exp^{{X_j}^T\beta})]$$

Quando relembramos a log-verossimilhança da distribuição Poisson verificamos um resultado próximo, ou seja: $\sum_{i=1}^{n}(y_i \log\lambda - \lambda - \log y_i!)$. Em suma, a propriedade de riscos proporcionais faz com que a log-verossimilhança obtida no modelo Weibull seja proporcional à log-verossimilhança de n variáveis independentes com uma distribuição Poisson. Nesse sentido, temos uma contagem do número de observações com e sem censura, logo, resumimos da seguinte maneira: $D_j \sim Pois(\mu_j)$ com $\mu_j = (\alpha {y_j}^{\gamma}) \exp^{{X_j}^T\beta}$.

A função de ligação do modelo será a logarítmica. Além do preditor linear, temos a inclusão do tempo de sobrevivência e do parâmetro de forma γ, bem como

a incorporação do parâmetro de escala a partir do intercepto $\log \alpha + \beta_0$. Portanto, a relação obtida pode ser caracterizada da seguinte maneira:

$$\log \mu_j = \log \alpha + \gamma \log y_j + {X_j}^T \beta$$

É importante observar a presença do logaritmo dos tempos na função de ligação, uma vez que vamos chamar essa estrutura de *offset* e a utilizaremos na função *glm* do R. Como a distribuição Poisson faz parte da família exponencial, usamos como base o conhecimento dos MLGs em uma modelagem de sobrevivência. Esse modelo pode ser ajustado de forma iterativa a partir do método de verossimilhança:

$$\frac{\partial l}{\partial \gamma} = \frac{m}{\gamma} + \sum_{j=1}^{n} [\, (\delta_j - \mu_j) \log y_j \,] = 0$$

Nesse caso m é o número de observações não censuradas, ou seja, que apresentam tempo de falha e não de censura. Se começarmos com um chute inicial de $\gamma = 1$, o modelo pode ser ajustado por uma regressão de Poisson para produzir valores ajustados de μ que serão utilizados para estimar $\hat{\gamma}$. Nessa perspectiva, o estimador do parâmetro γ será dado por:

$$\hat{\gamma} = \frac{m}{\sum_{j=1}^{n} [\, (\mu_j - \delta_j) \log y_j \,]}$$

Em relação aos métodos iterativos de estimação citamos anteriormente os algoritmos Escore de Fisher (Fisher Scoring-FS) e de Newton-Raphson. Considerando que caímos em uma regressão de Poisson conforme citado acima, aproveitamos a estrutura de estimação dos MLGs. O primeiro algoritmo consiste em obter os estimadores a partir do logaritmo da função de log-verossimilhança, tendo como foco a atualização do vetor paramétrico β em que r representa o número de iterações, $U(\beta^{(r)})$ o vetor score avaliado no ponto $(\beta^{(r)})$ e $I^{-1}(\beta^{(r)})$ o inverso da matriz de informação.

$$\beta^{(r+1)} = \beta^{(r)} + I^{-1}(\beta^{(r)}) U(\beta^{(r)})$$

O algoritmo Fisher Scoring começa com um chute inicial $\beta^{(0)}$, passa pela obtenção das estimativas do preditor linear e atualização da média, obtenção da variável resposta ajustada e o cálculo atualizado de $\beta^{(r+1)}$. O processo é repetido até a convergência do algoritmo segundo algum critério preestabelecido. Por sua vez, o algoritmo de Newton-Raphson é utilizado para obter o estimador de máxima verossimilhança, entretanto, o cálculo envolve a matriz hessiana $H(\beta^{(r)})$ e o processo é repetido até que a convergência seja alcançada, na qual $\beta^{(r+1)} \approx \beta^{(r)}$.

$$\beta^{(r+1)} = \beta^{(r)} + H(\beta^{(r)}) U(\beta^{(r)})$$

2.2.3 Inferência Estatística

Sob o contexto de modelos lineares, há a necessidade de se avaliar a significância dos estimadores $\hat{\beta}_i$, i = 1,2...,p. Assim, tem-se o teste de hipóteses com $H_0 : B_j = 0$ e $H_1 : B_j \neq 0$, isto é, para o caso em que rejeita-se a hipótese nula, tem-se que o estimador é significantemente diferente de zero.

A partir do algoritmo utilizado para obter as estimativas por máxima-verossimilhança, a matriz de informação de Fisher também é obtida, de forma que:

$$I^{-1}(\hat{\beta}) = Cov(\hat{\beta}) = \phi^{-1}(X^T W(\hat{\beta})X)^{-1}$$

$$ep(\hat{\beta}) = \sqrt{\phi^{-1}}\nu_{j,j}$$

Assim, supondo que as condições de regularidade, espera-se encontrar o estimador $\hat{\beta}$ de β, tal que $U(\hat{\beta}) = 0$, e assim:

$$\mathbb{E}[U(\beta)] = 0 \; e \; Cov[U(\beta)] = \mathbb{E}[U^2(\beta)] = I(\beta)$$

Consequentemente $U(\hat{\beta})$ converge assintoticamente para uma distribuição normal com médias zero e variância $I(\beta)$. E, também, $U(\beta)^T I^{-1} U(\beta)$ converge assintoticamente para uma qui-quadrado com p graus de liberdade. Então, após esses valores serem estabelecidos , pode-se construir os testes e intervalos de confiança.

Assim, para assegurar a hipótese proposta anteriormente, pode-se usar 3 testes: Teste de Wald, Teste da Razão da Verossimilhança e Teste Escore.

- **Teste de Wald**: a estatística do teste, sob H_0 é dada por:

$$W_T^2 = \left(\frac{\hat{\beta}_j - \beta^{(0)}}{ep(\hat{\beta}_j)} \right)^2 \sim \chi_1^2$$

  E para a significância conjunta temos:

$$W_T^2 = (\hat{\beta} - \beta^(0))^T \hat{\phi}^{-1} (X^T W(\hat{\beta}X)^{-1} (\hat{\beta} - \beta^(0))$$

  que converge assintoticamente para uma distribuição χ_p^2 e $\hat{\phi}$ é o estimador consistente de ϕ. Por fim, o intervalo de confiança é dado por:

$$[\beta : W_T^2 \leq \chi_{p,1-\alpha}^2]$$

  onde $\chi_{p,1-\alpha}^2$ são ps quantis da distribuição χ^2 e $1 - \alpha$ o nível de confiança.

- **Teste Razão de Verossimilhanças**: a estatística sob H_0 é dada por:

$$\Lambda = -2l(\hat{\beta}^(0)) - l(\hat{\beta})$$

que também converge assintoticamente para para uma distribuição χ^2_p e o intervalo de confiança segue análogo:

$$[\beta : \Lambda \leq \chi^2_{p,1-\alpha}]$$

- **Teste Score**: por fim, este considera $U(\hat{\beta}^(0))$ uma vez que, se $\hat{\beta}$ é o EMV para β então $U(\hat{\beta}) = 0$ e a estatítica do teste é dada por:

$$S_T = U(\hat{\beta}^(0))^T I^{-1} U(\hat{\beta}^(0))$$

que também converge assintoticamente para para uma distribuição χ^2_p e o intervalo de confiança é análogo:

$$[\beta : S_T \leq \chi^2_{p,1-\alpha}]$$

O teste para a significância de ϕ é basicamente o teste de Wald:

$$W_T = \frac{\hat{\phi} - \phi^{(0)}}{\sqrt{Var(\hat{\phi})}} \sim N(0,1)$$

2.2.4 Qualidade de ajuste: Função Desvio

Uma das ferramentas mais importantes para avaliarmos a qualidade de modelos sob a perspectiva dos MLGs é a função desvio (Deviance). Nesse sentido, conseguimos identificar se a função de ligação se ajusta bem ao modelo e, consequentemente, aos dados (hipótese nula).

Nesse tipo de análise comparamos o modelo nulo com o modelo completo/saturado a partir da log-verossimilhança. O primeiro tipo de modelo é aquele que tem apenas um unico parâmetro na sua estrutura de regressão, ou seja, é formado pelo intercepto que representa a média geral e utilizamos o estimador de máxima verossimilhaça $\hat{\mu}$ para μ. O modelo completo é aquele que tem tantos parâmetros quanto observações e utilizamos o estimador de máxima verossimilhaça $\tilde{\mu}$ para μ. Portanto, a função desvio escalonada possui a seguinte fórmula:

$$D^*(y, \phi, \mu) = 2(loglik\, modelo\, completo - loglik\, modelo\, reduzido)$$

$$D^*(y, \phi, \mu) = 2(\, l(y, \phi, \tilde{\mu}) - l(y, \phi, \hat{\mu})\,)$$

3 Resultados

3.1 Banco de dados

O banco de dados utilizado para a aplicação prática do presente trabalho pode ser obtido em adesao.dat, disponibilizado gratuitamente pela Fiocruz como material do livro e site *Análise de Sobrevivência: teoria e aplicações em saúde*.

Conforme descrito no site, o banco traz informações sobre um estudo acerca do efeito da adesão ao tratamento antirretroviral na falha terapêutica (virológica, imunológica, clínica ou óbito), realizado com pacientes assistidos no Ipec/Fiocruz (CAMPOS et al., 2009). Além disso, a variável de interesse foi obtida a partir do Sistema de Controle Logístico de Medicamentos (SICLOM), desenvolvido para a dispensa de medicamentos antirretrovirais, e calculada como a razão entre o total de dias com atraso no contato com a farmácia para obter a medicação e os dias de acompanhamento entre a entrada no estudo e a falha terapêutica.

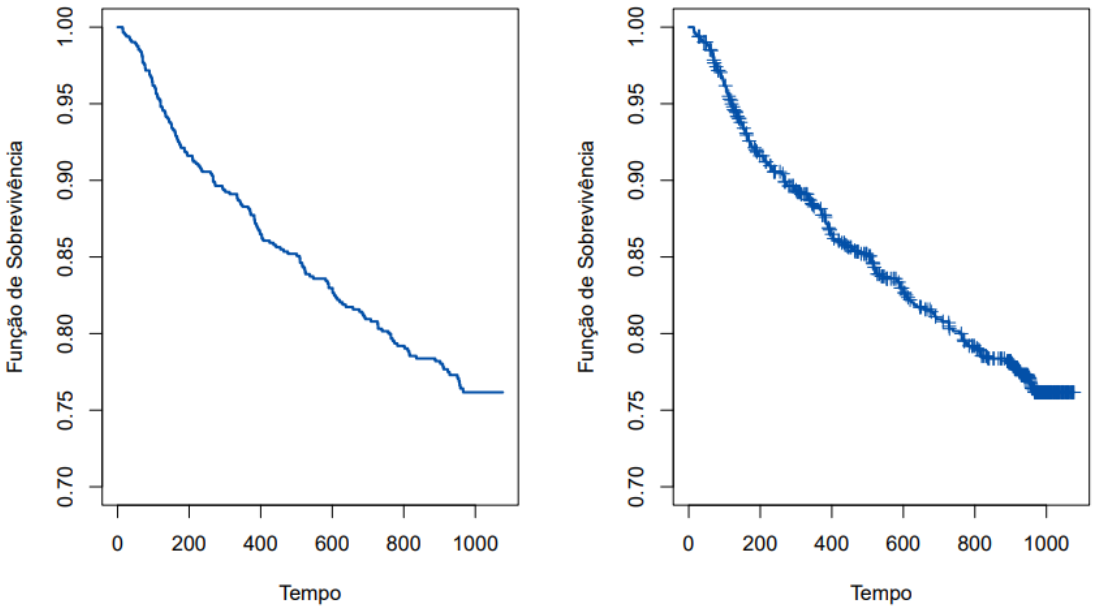
É composto por 711 indivíduos em 819 observações, organizado para análise de sobrevivência com uma linha para cada falha de cada paciente, porém sem covariáveis tempo-dependentes. As 8 variáveis presentes estão detalhadas a seguir:

Variável	**Descrição**
id	Identificação do paciente
ini	Data do início do acompanhamento da dispensação de medicamentos antirretrovirais (em dias)
fim	Data da falha terapêutica ou fim do estudo
tempo	ini - fim (em dias)
status	0 = censura, 1 = falha terapêutica
linha	Falha terapêutica ordenada, de 1 a 3
propatraso	Proporção de dias sem medicamento/total de dias de acompanhamento para cada período entre falhas ou censura
comprimdia	Número médio de comprimidos/dia previstos

3.2 Análise descritiva

Uma das primeiras etapas, quando trabalhamos com modelos de sobrevivência, consiste em realizar uma análise descritiva dos dados. Em um primeiro momento, podemos pensar em estimar a função de sobrevivência por Kaplan-Meier, conforme vemos abaixo:

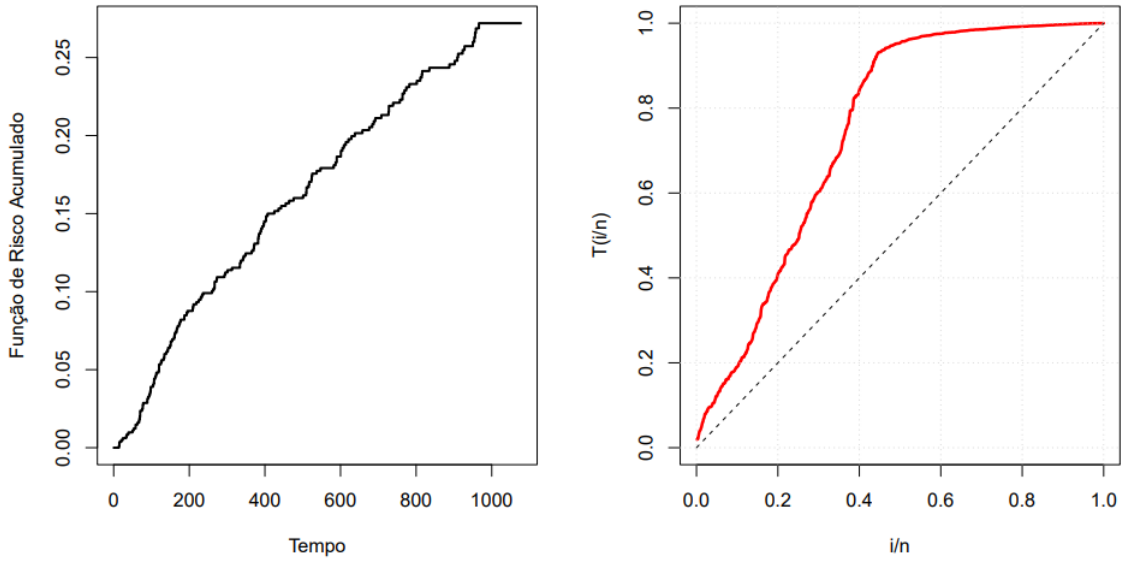
Figura 2: Função de Sobrevivência por Kaplain Meier



A partir do primeiro gráfico conseguimos visualizar o comportamento geral da função de sobrevivência. Nesse sentido, verificamos um decaimento da probabilidade de sobrevivência sem grandes oscilações até o patamar de aproximadamente 0,75 após o tempo de 900 dias. Essa queda para e a função de sobrevivência se estabiliza, ou seja, não decai para o nível zero pois existe um quantidade significativa de dados censurados. Esse fenômeno é perceptível quando olhamos para o segundo gráfico da Figura 2.

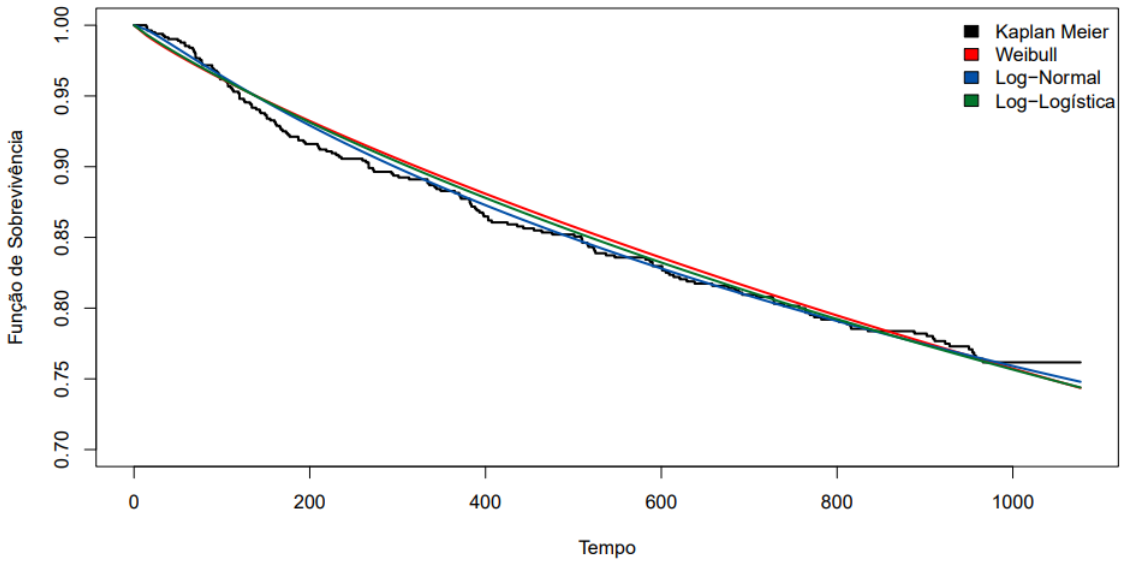
O próximo passo é verificar o comportamento da função de risco acumulado e o gráfico do Tempo Total em Teste (TTT). Nessa etapa, buscamos inferir sobre a real forma da função de risco para que possamos atribuir um modelo probabilístico ao estudo. Conforme vemos logo abaixo, o gráfico TTT apresenta uma curva totalmente voltada para cima, o que indica uma função de risco crescente. O fato negativo é que esse tipo de análise não possui tanto poder na presença de muitos dados censurados. Por sua vez, o comportamento da função de risco acumulada não é tão claro, logo, devemos testar mais de um modelo candidato. O modelo Weibull é bastante flexível para funções de riscos monótonas crescentes ou decrescentes. Por sua vez, modelos como log-normal e log-logistíco são frequentemente utilizados quando temos funções de riscos unimodais.

Figura 3: Função de Risco Acumulado e Gráfico do Tempo Total em Teste



Quando ajustamos os três modelos probabilísticos citados anteriormente verificamos resultados extremamente próximos. Vejamos a figura à seguir:

Figura 4: Função de Sobrevivência por Kaplan-Meier e Modelos probabilísticos



Na figura 4 podemos observar que a função de sobrevivência empírica estimada por Kaplain Meier possui um comportamento bem similiar às funções de sobrevivência nos modelos Weibull, log-normal e log-logístico. Uma outra forma de observar esse fato é com o cálculo de outras medidas para classificar e selecionar modelos tais como o Critério de Akaike (AIC), Critério de Akaike corrigido (AICc) e o Critério de informação Bayesiano (BIC).

Tabela 1: Critérios para classificação e seleção de modelos - AIC, AICc e BIC

Modelo	**AIC**	**AICc**	**BIC**
Log-Normal	3083,33	3094,93	3092,75
Log-Logístico	3092,09	3083,35	3101,51
Weibull	3094,92	3092,11	3104,34

A escolha do modelo é atribuída ao que apresentar os menores valores para as medidas acima. Apesar do modelo Weibull não ter sido o mais ajustado e com os melhores resultados, verificamos que a distância entre os três modelos é extremamente pequena e todos aparentam ser adequados à modelagem dos dados em questão. Portanto, escolhemos o modelo Weibull ($\alpha = 4459,835$ e $\gamma = 0,856$), uma vez que este é da classe dos modelos de riscos proporcionais e a abordagem por MLG pode ser utilizada.

Por fim, cabe ressaltar que temos uma variável categórica no banco de dados que será testada como variável explicativa. Na análise inicial, utilizamos o teste de Wilcoxon para a hipótese nula de igualdade entre as funções de sobrevivência nos seguintes tipos de falha terapêutica (virológica, imunológica, clínica ou óbito). Nesse cenário, a estatística do teste resultou em um valor de $\chi^2 = 2$ com 2 graus de liberdade e p-valor igual a $0,4$. Portanto, não rejeitamos a hipótese nula. Apesar disso, ainda vamos reavaliar tal variável na parte de regressão.

3.3 Comparativo: abordagem por MLG e Sobrevivência

Para verificar o processo de estimação discutido na metodologia, comparamos um ajuste inicial do modelo com todas as possíveis variáveis explicativas considerando duas abordagens. Na primeira, temos como variável resposta o próprio tempo de sobrevivência e a indicação de censura, ou seja, tratamos a análise da forma clássica dos modelos de sobrevivência utilizando a função *survreg* do pacote *Survival* do R.

```
survreg(Surv(tempo, censura) ~ comprimdia + propatraso +
    linha, dist = "weibull")
```

Na segunda abordagem utilizamos o fato do modelo Weibull fazer parte dos modelos de riscos proporcionais, o que nos permite utilizar a base de conhecimento dos MLGs. Nesse sentido, a log-verossimilhança é proporcional à log-verossimilhança de uma Poisson em que contamos o número de falhas terapêuticas.

```
glm(censura ~ offset(log(tempo)) + comprimdia + propatraso +
      linha, family = poisson(link="log")))
```

Em ambos os casos obtivemos a mesma interpretação. Apenas o intercepto e a variável propatraso (Proporção de dias sem medicamento/total de dias de acompanhamento para cada período entre falhas ou censura) foram considerados significativos. Quando olhamos para os valores absolutos das estimativas, erro padrão e a estatística de teste Z, percebemos que os resultados encontrados são extremamente próximos, conforme esperado. Portanto, trabalhar com os modelos de sobrevivência sob a perspectiva da estrutura dos MLGs funciona de forma eficiente.

Tabela 2: Ajuste do modelo inicial pela função *survreg* do pacote *Survival* do R

Efeito	**Estimativa**	**Erro Padrão**	**Z**	**P-valor**
Intercepto	9,3021	0,2507	37,11	<0,001
comprimidos/dia	-0,0409	0,0227	-1,80	0,072
propatraso	-3,8712	0,4109	-9,42	<0,001
falha imunológica	-0,1113	0,2931	-0,38	0,704
clínica ou óbito	-0,8655	1,0494	-0,82	0,409

Tabela 3: Ajuste do modelo inicial pela função *glm* do R

Efeito	**Estimativa**	**Erro Padrão**	**Z**	**P-valor**
Intercepto	-9,20154	0,18010	-51,092	<0,001
comprimidos/dia	0,03932	0,02157	1,823	0,0683
propatraso	3,74499	0,34086	10,987	<0,001
falha imunológica	0,13906	0,27717	0,502	0,6159
clínica ou óbito	0,89436	1,00334	0,891	0,3727

A grande questão que surge é: por que os sinais das estimativas e da estatística Z são diferentes? A resposta está na característica da variável resposta. No ajuste pela função *survreg* estamos considerando o tempo de sobrevivência, dessa forma, se uma variável possui efeito negativo significa que a cada incremento de unidade o tempo de sobrevivência tende a decair. Tomando como exemplo, se a proporção de atraso aumenta então a probabilidade de sobrevivência diminui e isso reflete na função de sobrevivência. No ajuste pela função *glm* temos como variável resposta a contagem do número de falhas terapêuticas, dessa forma, se uma variável possui efeito positivo significa que a cada incremento de unidade o número médio de falhas tende a aumentar. Nesse caso, o tempo de sobrevivência também tende a decair, já que o número de falhas aumenta. Em suma, temos apenas uma leve mudança na forma que interpretamos os valores, embora o resultado e a interpretação final sejam exatamente iguais em ambos os cenários.

O valor estimado para o parâmetro γ foi de $0,957$ (transformação a partir da relação entre Weibull e Valor Extremo na função *survreg*). Utilizando a função *glm* e a fórmula de $\hat{\gamma}$ especificada na parte de estimação, a estimativa foi $\hat{\gamma} = 0,949$. Aqui ainda não estamos preocupados com o ajuste do modelo, já que não estamos no modelo final. A concepção desse tópico é apenas mostrar que as duas abordagens de modelagem convergem para resultados e interpretações semelhantes. Isso corrobora com a ideia de que conseguimos utilizar o conhecimento e a estrutura dos MLGs em certos casos de modelos de sobrevivência. Por fim, cabe ressaltar que a função *survreg* se utiliza por *default* do método de Newton-Raphson e a convergência ocorreu com 7 iterações. A função *glm* possui como *default* o método do Escore de Fisher e o número de iterações também foi de 7.

3.4 Ajuste do modelo e interpretação

Dados os desenvolvimentos anteriores, ajustamos então o modelo de regressão utilizando a função *glm()* com família Poisson e função de ligação logarítmica. Em um primeiro momento, utilizando o algoritmo de seleção de modelos *stepwise*, o modelo com menor AIC é apresentado na tabela 4, onde podemos observar que a covariável comprimidos/dia não foi significativa a 5%, o AIC desse modelo foi de 1122,5 e o Deviance residual 778,48.

Tabela 4: Ajuste do modelo GLM por seleção *stepwise*

Efeito	**Estimativa**	**Erro Padrão**	**Z**	**P-valor**
Intercepto	-9,1959	0,1799	-51,12	<0,001
comprimidos/dia	0,0393	0,0216	1,82	0,0681
propatraso	3,7773	0,3336	11,32	<0,001

Assim, como a estimativa do parâmetro da covariável comprimidos/dia não foi significativa, consideramos o modelo final apenas com propatraso de covariável:

Tabela 5: Modelo final

Efeito	**Estimativa**	**exp{estimativa}**	**I.C 95%**	**Erro Padrão**	**Z**	**P-valor**
Intercepto	-8,967	$1,276e^{-04}$	$[9,94e^{-05}$; $1,62e^{-04}]$	0,124	-72,29	<0,001
propatraso	3,828	45,9705	[23,76 ; 87,1]	0,331	11,56	<0,001

O modelo apresentado na tabela 5 obteve AIC igual a 1123,6 deviance igual a 781,63, sendo levemente maior que o modelo por stepwise. A estimativa de $\hat{\gamma}$ foi de 0,9358.

Interpretando os resultados na tabela 5 concluímos que para cada unidade aumentada na covariável propatraso (Proporção de dias sem medicamento/total de dias de acompanhamento para cada período entre falhas ou censura), o número médio de falhas terapêuticas aumenta em 46 vezes e consequentemente o tempo de sobrevivência diminui. Como estamos lidando com proporção na covariável a ideia de aumento de uma unidade inteira não é adequado, dessa forma, utilizamos como medida o valor de $0,01$. Dessa forma, para cada aumento de $0,01$ na covariável propatraso, o número médio de falhas terapêuticas aumenta em $e^{0,03828} = 1,04$ vezes e o decaimento da função de sobrevivência é acelerado.

Ainda do ajuste do modelos, faz-se necessário o teste de qualidade de ajuste. Este consiste em testar se a função de ligação utilizada ajusta corretamente os dados, de forma que buscamos não rejeitar a hipótese nula. O teste é basicamente aplicar um teste qui-quadrado sobre o deviance do modelo e foi obtido um valor p igual a 0.8903, indicando que a função de ligação *log()* se ajusta bem aos dados, considerando um nível de significância de 5%. Além disso, a estatística R^2 de Naglekerke foi de 0,1857 e o R^2 alternativo foi igual a 0,121, para modelos GLM consideramos valores

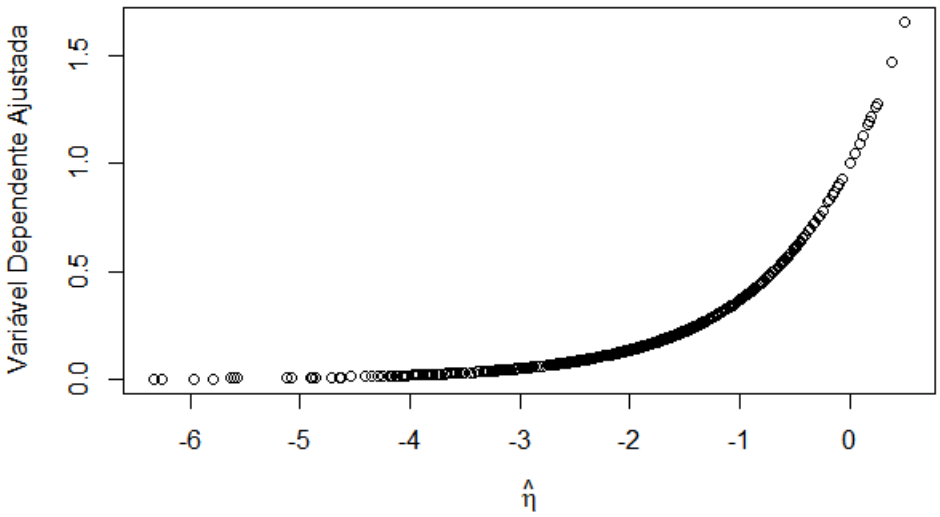
próximos de 0,5 como indicativo de um bom ajuste, o que não foi o caso por esse critério.

A análise do deviance (análogo à análise de variância ANOVA) é apresentada na tabela 6 e a partir dela podemos concluir que a variável propatraso reduz significativamente o deviance em relação ao modelo nulo. Além disso, se pegarmos o resíduo deviance do modelo ajustado pelo seu respectivo grau de liberdade temos o valor de $781,63/817 = 0,956$ que é menor que 1, o que indica um bom ajuste na modelagem dos dados.

Tabela 6: Deviance no Modelo final

	G.L	**Deviance**	**Resid. G.L**	**Resid. Deviance**	**P valor**
Nulo			818	889,12	
propatraso	1	107,50	817	781,63	<0,001

Figura 5: Análise visual da adequabilidade de ajuste do modelo

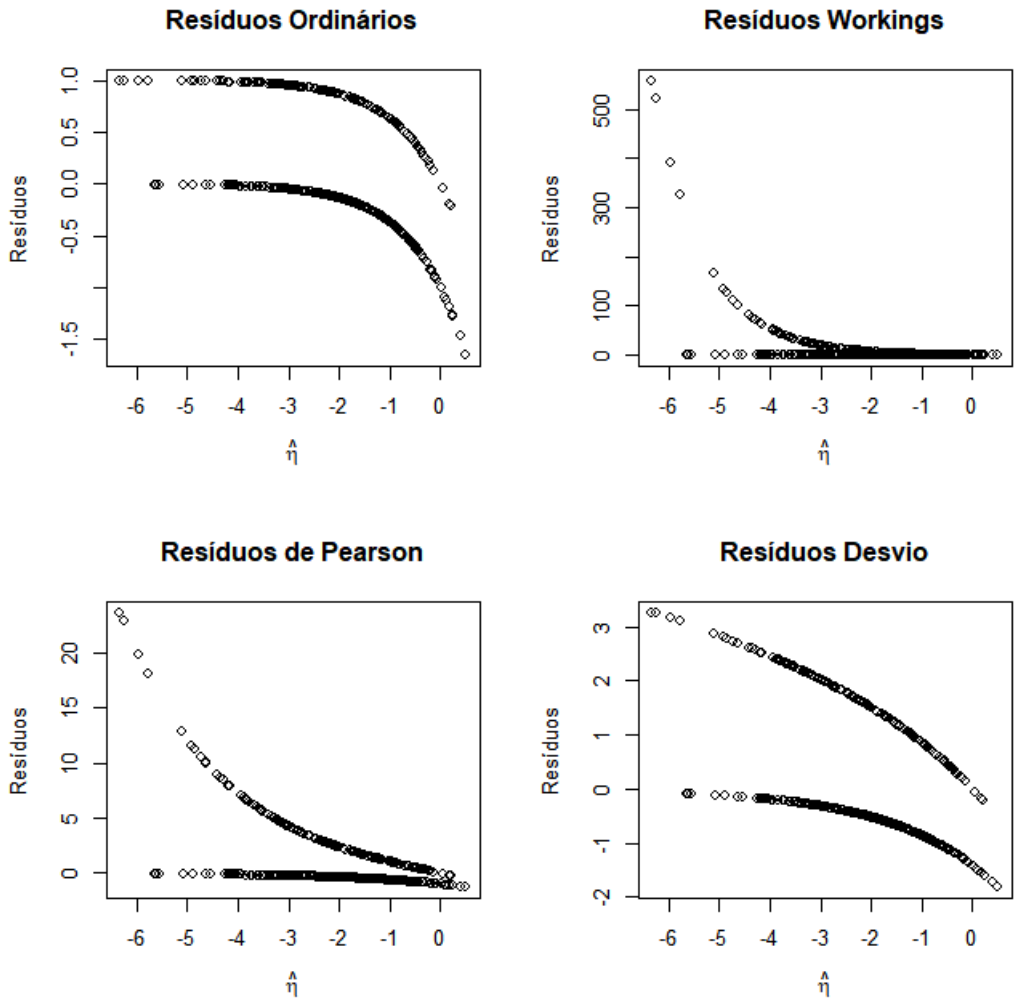


De forma contrária, o gráfico acima referente à variável dependente ajustada contra o preditor linear não sugere uma boa qualidade de ajuste da função de ligação, pois, os pontos não demonstram uma relação linear entre as duas quantidades.

3.5 Análise dos resíduos

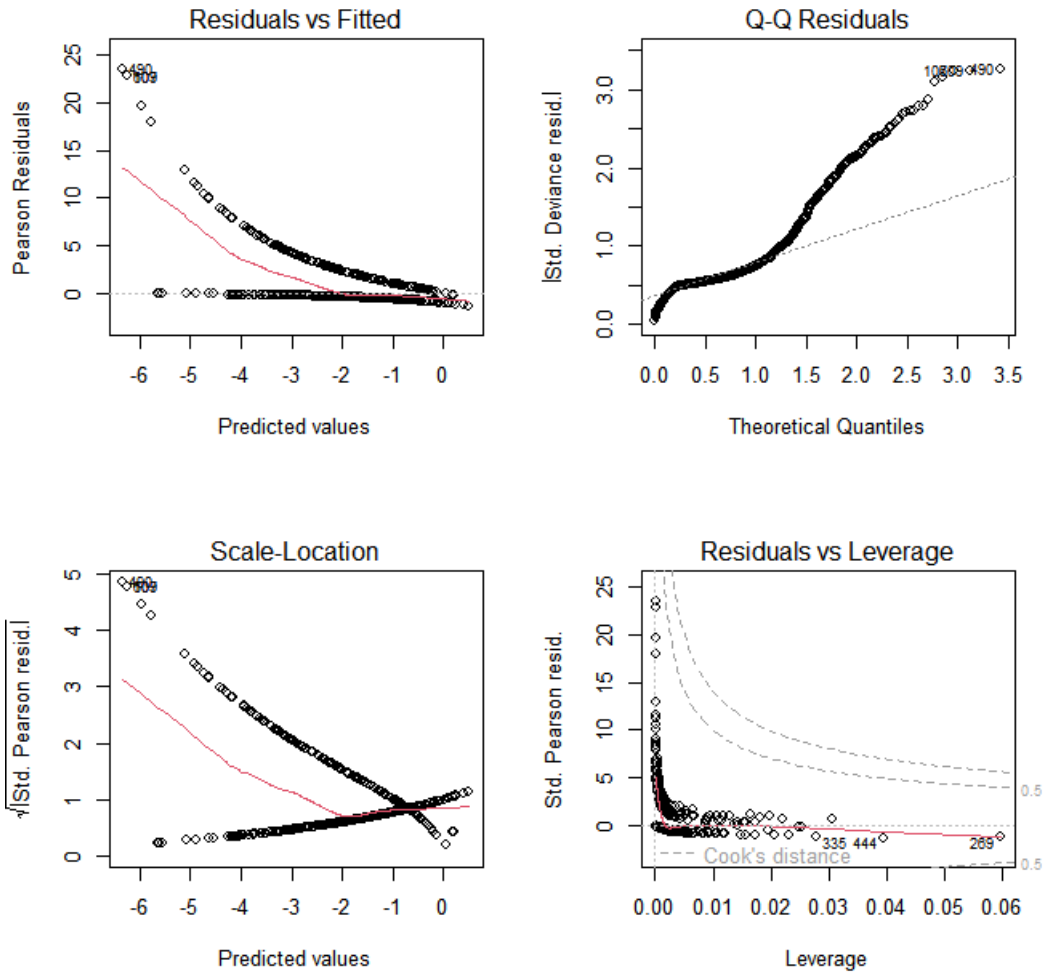
Após realizada a modelagem, se faz necessário diagnosticar e avaliar diferentes aspectos do modelo.

Figura 6: Comparação do grau de dispersão de diferentes tipos de resíduos em função do preditor linear



Os diferentes tipos de resíduos apresentados na Figura 6 mostram que os resíduos ordinários, ou seja, aqueles dados pela diferença entre os valores observados dos estimados, e os resíduos desvio apresentam tendência parecida, enquanto os resíduos Working e de Pearson também apresentam comportamentos parecidos entre eles. Porém, observa-se que em nenhum há pontos espalhados em torno de zero, além de tendências claras. Portanto, segue abaixo a análise residual padronizada a fim de solucionar o problema.

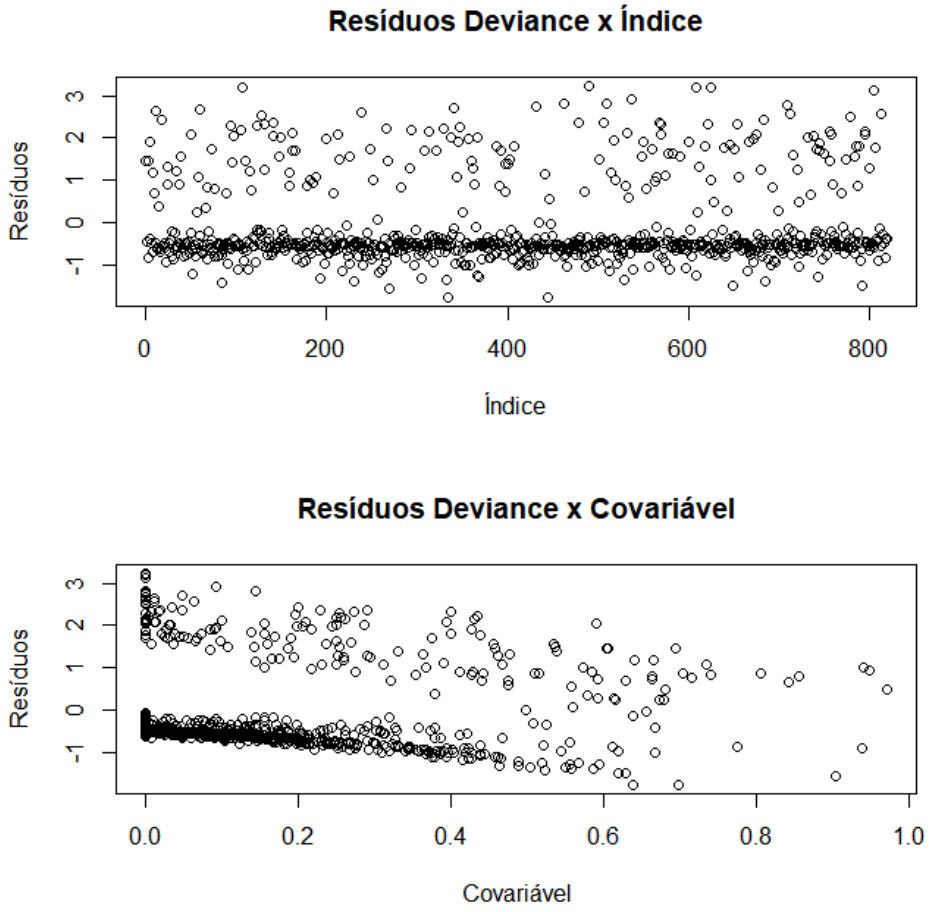
Figura 7: Diagnóstico básico dos resíduos do modelo GLM Poisson



As linhas vermelhas indicam a tendência dos dados, é percebido então que elas não seguem a linha pontilhada, o que sugere ausência de homogeneidade de variância dos resíduos. Além disso, nos 3 primeiros gráficos da Figura 7, há presença de observações, referentes às linhas do banco de dados, que aparecem identificados, sendo eles 490 e 609, indicando que essas observações estão distantes da média e possíveis *outliers* e/ou um potenciais pontos de influência. No entanto, em Residuals vs Leverage, o gráfico contendo os valores da distância de Cook não indica observações influentes, visto que estão todas abaixo de 0,8.

O gráfico dos resíduos de Pearson vs valores estimados, mostra claramente uma falta de espalhamento dos pontos em torno de zero, indicando mais uma vez que a variância dos resíduos não seja constante, assim como no gráfico da raiz quadrada dos valores absolutos (baseado nos resíduos de Pearson) demonstram comportamentos (tendências aparentes) no espalhamento dos pontos O segundo gráfico, especificamente, revela que os resíduos desvio padronizados não seguem distribuição Normal, fugindo muito à reta dos quantis teóricos.

Figura 8: Resíduos Deviance (Desvio)



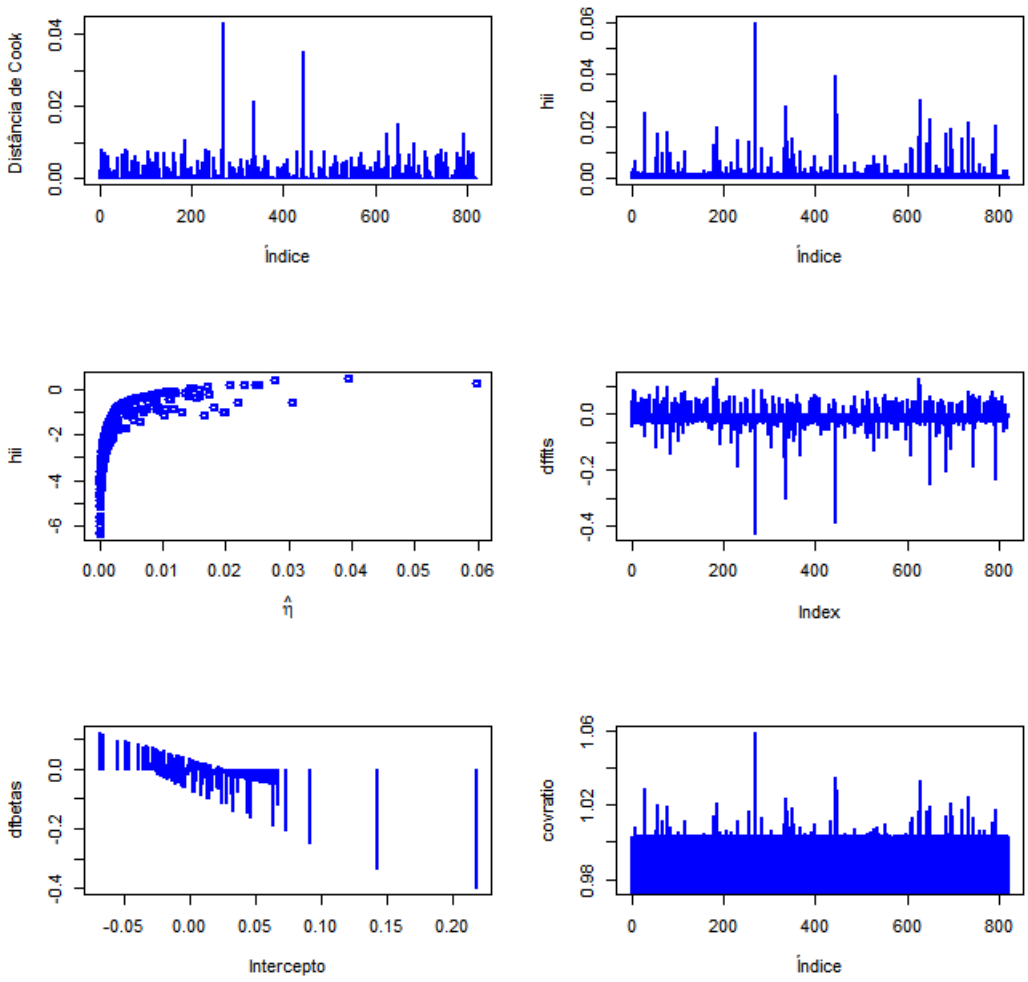
Na Figura 8 é observado em Resíduos Deviance vs Índice da observação considerável espalhamento em todo o gráfico, sendo a maior concentração abaixo da referência zero. Este, diferente do que foi visto nos gráficos anteriores, apresenta homocedasticidade e não há problemas de autocorrelação serial. Já em Resíduos Deviance vs Covariável, que no caso é propatraso, é possível verificar se existe relação entre os resíduos e uma variável incluída no modelo. Grande parte dos resíduos encontram-se à esquerda e abaixo de 0, para valores baixos da variável explicativa, apesar de haver um maior espalhamento à direita, porém com poucos pontos conforme a proporção de dias sem medicamento/total de dias de acompanhamento aumenta.

3.6 Análise dos pontos de influência

A seguir serão analisadas as medidas de influência considerando os seguintes critérios, sendo $p = 2$ e $n = 819$:

- $h_{ii} > \frac{3p}{n}$=0,00732
- Di ≥ 0.8
- $|\,1 - \text{COVRATIO}\,| \geq \frac{3p}{n-p} = 0,00734$
- $|\text{DFBETAS}| \geq \frac{2}{\sqrt{n}} = 0,0699$
- $|DFFITS_i| \geq 2\sqrt{\frac{p}{n}} = 0,0988$

Figura 9: Medidas de Influência

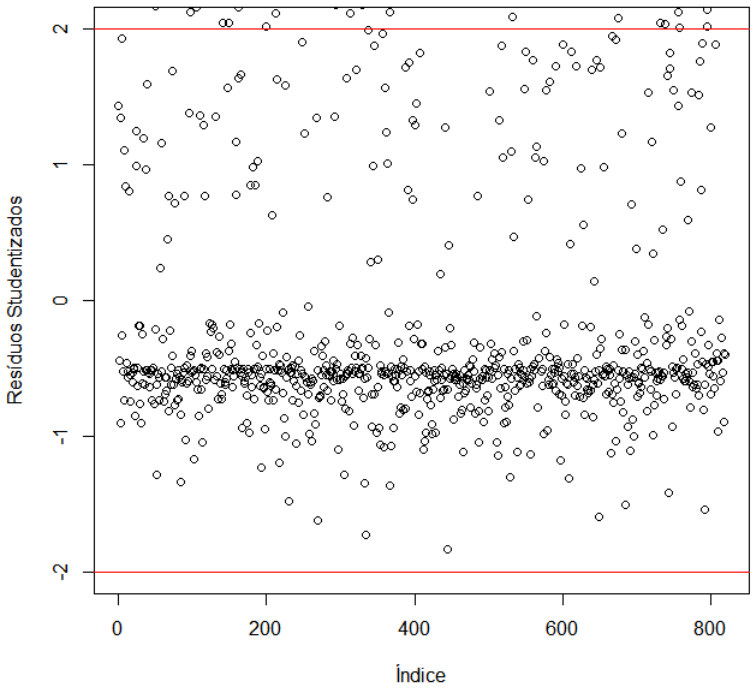


Dando seguimento com as análises, é de destacar que, segundo os critérios mencionados, apenas por distância de Cook nenhuma observação pode ser considerada

ponto de influência. Quanto ao restante das medidas, apresentam diversas observações que têm um impacto significativo nos resultados do modelo, por hii, por exemplo, são 44 e por COVRATIO são 52 observações.

É possível também visualizar por meio do gráfico dos resíduos studentizados, o qual espera-se ver um padrão aleatório entre ±2.

Figura 10: Resíduos Studentizados



Assim como visto nas medidas de influência, pelo gráfico acima observou-se variáveis que estão fora do intervalo para cima de 2, indicando pontos de alavancagem.

4 Conclusão

Com base nas análises realizadas, constatou-se que o comparativo das duas abordagens de modelagem para o processo de estimação (por análise de sobrevivência, com o tempo de sobrevivência como variável resposta, e MLG, considerando o número de falhas terapêuticas) revelou resultados consistentes e interpretações semelhantes. Os principais parâmetros, como o intercepto e a variável propatraso, foram significativos em ambos os cenários. Os resultados obtidos reforçam a viabilidade e a consistência das abordagens utilizadas, consolidando a compreensão de que a aplicação do conhecimento e estrutura dos MLGs em modelos de sobrevivência é uma prática eficaz e confiável.

Os resultados da análise de regressão utilizando o modelo Poisson com função de ligação logarítmica indicam que a covariável propatraso é significativa na explicação do número médio de falhas terapêuticas. A interpretação dos resultados revela que um aumento de 0,01 na covariável propatraso está associado a um aumento de aproximadamente 4% no número médio de falhas terapêuticas, indicando uma diminuição no tempo de sobrevivência. A análise da qualidade de ajuste sugere que a função de ligação logarítmica se ajusta bem aos dados.

Entretanto, a análise de resíduos e medidas de influência não refletiram bem a qualidade do ajuste evidenciada em 3.4, de modo que os resultados indicam a necessidade de uma revisão mais detalhada do modelo, com especial atenção para a heterogeneidade da variância dos resíduos, a presença de possíveis outliers e pontos de influência, bem como a relação entre os resíduos e as variáveis explicativas.

Referências

CAMPOS, D. P. et al. *Efeito do critério de diagnóstico da AIDS e da adesão ao tratamento anti-retroviral na progressão clínica em HIV/AIDS*. Tese (Doutorado), 2009.

CARVALHO, M. S. et al. *Análise de sobrevivência: teoria e aplicações em saúde*. [S.l.]: SciELO-Editora FIOCRUZ, 2011.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. [S.l.]: Editora Blucher, 2021.

DOBSON, A. J.; BARNETT, A. G. *An introduction to generalized linear models*. [S.l.]: CRC press, 2008.

JOHNSON, N.; KOTZ, S. *Continuous Univariate Distributions*. J. Wiley, 1970. (Continuous Univariate Distributions, v. 1). ISBN 9780471446262. Disponível em: <https://books.google.com.br/books?id=SQ7vAAAAMAAJ>.

KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J. *Applied Linear Statistical Models*. McGraw-Hill Irwin, 2013. (McGrwa-Hill international edition). ISBN 9780071122214. Disponível em: <https://books.google.com.br/books?id=0xqCAAAACAAJ>.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. [S.l.]: John Wiley & Sons, 2011.

LINDSEY, J. K. *Applying generalized linear models*. [S.l.]: Springer Science & Business Media, 1997.

PRENTICE, R. L. et al. The analysis of failure times in the presence of competing risks. *Biometrics*, JSTOR, p. 541–554, 1978.

SAIKIA, R.; BARMAN, M. P. A review on accelerated failure time models. *International Journal of Statistics and Systems*, v. 12, n. 2, p. 311–322, 2017.

WEIBULL, W. Astatistical theory ofthe strength of materials. *Proc. Royal 4cademy Engrg Science*, v. 15, 1939.

WEIBULL, W. A statistical distribution function of wide applicability. *Journal of applied mechanics*, 1951.

5 Anexo

O código em R completo utilizado no trabalho pode ser visualizado e baixado a partir do seguinte endereço eletrônico: Código em R.