Universidade de Brasília
IE - Departamento de Estatística

Modelo de regressão Log-Logístico para dados grupados na presença de censura

Daniel Lima Viegas

Orientador: Prof.ª Juliana Betini Fachini Gomes

Brasília

2018

Daniel Lima Viegas

Modelo de regressão Log-Logístico para dados grupados na presença de censura

Orientadora:
Profª. Drª. **Juliana Betini Fachini Gomes**

Monografia apresentada para a obtenção do título de Bacharel em Estatística.

Brasília
2018

Agradecimentos

Agradeço primeiramente a Deus por me conceder saúde e disposição e por permitir a realização dos meus feitos.

À Profa. Dra. Juiliana Betini Fachini Gomes pelo apoio, paciência e compreensão. Uma das melhores professoras que tive na graduação. Agradeço pela paciência nesses um ano e meio de trabalho.

Aos meus pais por todo apoio que me deram durante toda a minha graduação. O suporte financeiro e principalmente emocional nos momentos que eu mais precisei, são coisas as quais eu serei eternamente grato.

Aos amigos que fiz na graduação, agradeço pela aprendizado nas matérias mais difíceis, a colaboração nos trabalhos e o companheirismo nos momentos complicados.

Aos amigos que fiz no Ipea, me ajudaram a superar diversos obstáculos no meu aprendizado profissional, além de ajudar nos problemas emocionais sofridos.

Com esses agradecimentos, percebo que sou uma pessoa abençoada e rodeada de pessoas dispostas a ajudar e a alegrar os dias.

Conteúdo

Resumo

Neste trabalho é proposto um modelo de regressão Log-Logístico para dados de sobrevivência grupados. Este tipo de modelagem é realizada quando o banco de dados em questão possui muitos empates nos tempos, ou seja, quando os tempos se repetem muitas vezes. Para o desenvolvimento do modelo foi utilizada a distribuição Log-Logística que tem se mostrado bastante flexível em situações práticas. Para a estimação dos parâmetros do modelo foi utilizado o algoritmo computacional BFGS. Por fim, dois bancos de dados reais foram utilizados para exemplificar o ajuste do modelo de regressão proposto.

Palavras-chave: Modelos de regressão; Distribuição Log-Logística; Dados de sobrevivência grupados;

1 Introdução

A análise de sobrevivência é um tópico importante utilizado em diversas áreas, como biologia, engenharia, medicina, entre outros. O principal objetivo desta análise é explicar ou predizer o tempo até a ocorrência do evento estudado, esse tempo é chamado de tempo de falha. A principal diferença desta técnica de modelagem para as demais é a capacidade de levar em consideração também os tempos em que não foi possível observar o evento de interesse, esse tipo de ocorrência é chamado de censura.

Dentre os tipos de censuras existentes, a mais genérica é a censura intervalar. Esse tipo de censura ocorre quando não é possível determinar o tempo de ocorrência, mas se tem o intervalo de tempo onde ele ocorreu. Por exemplo, no estudo sobre o tempo até uma lâmpada queimar, deixa-se a lâmpada ligada até que ela queime, em um dia, ela está funcionando, o pesquisador sai da área onde está acontecendo o experimento e quando retorna, a lâmpada está queimada. Neste caso, sabe-se que o intervalo de tempo onde a lâmpada queimou é entre o tempo em que o pesquisador saiu e o que ele voltou. O objetivo deste trabalho é estudar o comportamento de dados grupados, que é um caso particular da censura intervalar. O caso de dados grupados pode ser um caso particular da censura intervalar quando todos os indivíduos estão sendo avaliados no mesmo intervalo de tempo e também ocorre quando existem muitos tempos iguais dentro do banco de dados.

Em diversos estudos de sobrevivência, estuda-se o relacionamento de covariáveis e o tempo, tendo o objetivo de realizar as análises estatísticas e tentando encontrar o melhor uso dessas variáveis para a criação de um modelo de regressão para dados censurados.

O trabalho possui como objetivo, sugerir um modelo para dois bancos de diferentes áreas utilizando a metodologia de dados grupados. Um dos bancos foi utilizado no estudo de Barreto et al. (1994) na área da saúde e foi cedido pela Universidade Federal da Bahia. O outro banco de dados é um banco de dados na área da educação cedido pela Universidade Estadual da Paraíba. Para a realização das análises foi utilizado o *software* R.

2 Revisão de Literatura

A fim de dar fundamento teórico ao trabalho na área de análise de sobrevivência, serão apresentados, a seguir, conceitos e notações presentes nas literaturas do tema.

2.1 Variável resposta e censuras

Chama-se evento de interesse, aquilo que se deseja encontrar informações sobre a ocorrência. Na análise de sobrevivência, esse evento pode ser a morte de um indivíduo, a cura, um casamento, divórcio ou funcionamento de um dispositivo ou componente de uma máquina. Em análise de sobrevivência, a variável resposta é geralmente o tempo até a ocorrência de um evento de interesse, sendo esse tempo denominado tempo de falha.

A principal característica dos dados de sobrevivência é a presença de censuras, ou seja, observações que por algum motivo não consegue-se observar a ocorrência do evento de interesse do estudo. Existem três tipos principais de censura, a mais usual é a censura à direita. Esta censura acontece quando um indivíduo entra no estudo e por algum motivo ele não pode mais ser acompanhado e então não é possível registrar a ocorrência do evento de interesse. Em estudos médicos que analisam o tempo desde a obtenção da doença até a morte do paciente, por exemplo, este tipo de censura acontece quando o paciente é curado, morre por outra razão ou simplesmente não se pode mais observar tal paciente. Este tipo de censura tem três tipos de classificação:

- Censura do Tipo I: acontece quando a pesquisa tem um tempo pré-determinado. Ao final do estudo, as observações que não falharam são consideradas censuras. Nesse tipo de estudo, o percentual de censura é descrito como uma variável aleatória.
- Censura do Tipo II: É encontrada quando se obtém um determinado número de falhas dentro do experimento. Nesse tipo de experimento, o número de falhas deve ser determinado antes de começar o experimento, fazendo com que

o número de falhas seja constante. O número de falhas, claramente deve ser menor do que o tamanho da amostra.

- Censura Aleatória: engloba os outros dois tipos de censura. Acontece quando alguns componentes não podem mais ser acompanhados ou quando o motivo da observação falhar é diferente do que interessa. Esta censura ocorre sem intervenção do pesquisador.

Além dessa censura também existem censuras importantes como a censura à esquerda e a censura intervalar. A censura à esquerda ocorre quando o evento ocorre antes do começo do experimento. Por exemplo, deseja-se observar o tempo até uma criança aprender a ler, porém no começo do estudo, algumas crianças podem já ter aprendido a ler sem saber exatamente o tempo em que ela aprendeu, isto é caracterizado como censura à esquerda.

A censura intervalar pode ser dita como um caso genérico das outras censuras. Chama-se censura intervalar quando não se sabe o tempo em que ocorreu o evento de interesse ocorreu, porém sabe-se que ele não ocorreu antes de um determinado tempo. Por exemplo, em um estudo médico é necessário que hajam visitas regulares para a detecção de certas doenças, tal como câncer. Nesse tipo de experimento, sabe-se que a doença apareceu antes do tempo de uma consulta (V), mas também sabe-se que ela apareceu depois de uma consulta (U), ou seja, a doença se manifestou no intervalo [U, V). Quando $V = \infty$ tem-se a censura a direita, e quando a o tempo $U = 0$, essa censura se torna a esquerda. Daí vem o conhecimento de caso genérico da censura.

Uma das definições de dados grupados é um caso particular da censura intervalar, esta acontece quando os indivíduos são acompanhados durante o mesmo intervalo de tempo. Outra forma de definir dados grupados é quando existem muitos empates nos tempos de estudo dos indivíduos.

2.2 Funções

2.2.1 Função densidade de probabilidade

Dada uma variável aleatória contínua, não negativa, que represente o tempo de falha de uma observação. Chama-se função densidade, uma função *f*, que descreva a probabilidade de um indivíduo falhar em um intervalo de tempo, quando o limite desse intervalo tende a zero.

A partir desta função, é possível obter-se a função de distribuição acumulada, denominada função *F*. No caso contínuo, esta função é obtida a partir do cálculo da integral da função densidade sobre todo seu suporte. Ou seja, dada uma variável aleatória T:

Uma função densidade de probabilidade é uma função que satisfaz as seguintes condições:

1. $f(x) \geq 0$ para todo x,
2. $\int_{-\infty}^{+\infty} f(x)dx = 1$,
3. para quaisquer a, b com $-\infty < a < b < +\infty$, teremos $P(a \leq X \leq b) = \int_a^b f(x)dx$.

A partir desta função, é possível obter-se a função de distribuição acumulada, denominada função *F*. No caso contínuo, esta função é obtida a partir do cálculo da integral da função densidade sobre todo seu suporte. Ou seja, dada uma variável aleatória T:

2.2.2 Função de sobrevivência

A função de sobrevivência é definida como a probabilidade de um indivíduo não falhar até um determinado tempo t, ou seja, é a probabilidade de uma observação viver além do tempo t. Dada uma variável aleatória T, contínua, não negativa. Pode-se descrever a função de sobrevivência como:

$$\begin{aligned} S(t) &= P(T > t) \\ &= \int_t^{\infty} f(v)dv. \end{aligned} \tag{1}$$

A função de sobrevivência também pode ser encontrada por meio da função de distribuição acumulada pela seguinte relação entre elas:

$$S(t) = 1 - F(t).$$

2.2.3 Função de Risco

Segundo Lawless (2003), esta funçao especifica a taxa de falha instantânea no tempo t dado que o indivíduo não falhou até esse tempo. Desta forma, é possível definir a função de risco como o limite da probabilidade de um indivíduo falhar no intervalo de tempo $[t, \Delta t)$, supondo que esse indivíduo não falhou até o tempo t, dividida pelo comprimento do intervalo infinitesimal Δt. Com isso, é possível visualizar matematicamente a definição dessa função como:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2}$$

Além disso, também é possível estabelecer uma relação entre esta função, a função de densidade e a função de sobrevivência:

$$h(t) = \frac{f(t)}{S(t)}. \tag{3}$$

2.2.4 Função de Risco Acumulado

A função de risco acumulado é uma função que não possui uma interpretação simples, porém possui importância dentro do campo da análise de sobrevivência.Esta função, denotada como $H(t)$, pode ser definida como o logaritmo da função de sobrevivência multiplicado por menos um, ou seja:

$$H(t) = -log(S(t)). \tag{4}$$

Esta função também pode ser obtida através da Função de Risco:

$$H(t) = \int_0^t h(u)du. \tag{5}$$

O gráfico desta função pode assumir algumas diferentes formas. Essas formas são utilizadas para determinar possíveis modelos probabilísticos que melhor se adequam aos dados. Com relação ao comportamento da função, o gráfico pode tomar as seguintes formas:

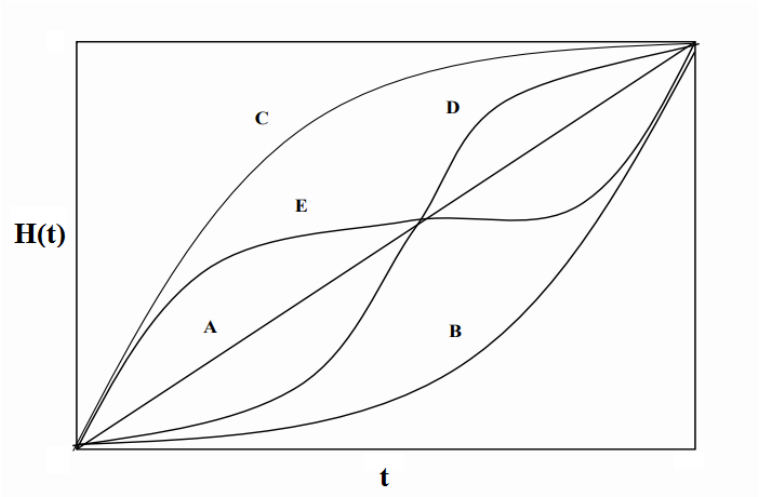

Figura 1: Retirado de Y. (2017), representa os comportamentos da Função Risco Acumulada

- A ⇒ Função de risco constante é adequada.
- B ou C ⇒ Função risco é monotonicamente crescente ou decrescente, respectivamente.
- D ⇒ Função risco tem forma de **U**.
- E ⇒ Função risco tem comportamento unimodal.

2.3 Estimadores da função de sobrevivência

2.3.1 Estimação na ausência de censura

A função de sobrevivência pode ser estimada amostralmente, como a proporção dos dados que não falharam até o tempo t. Esse estimador poderia ser escrito da seguinte forma:

$$\hat{S}(t) = \frac{n^o \textit{ de dados com tempo } > \textit{ t}}{n^o \textit{ total de individuos}}, \forall\ t\ \ \in t \geq 0$$

2.3.2 Kaplan-Meier

Os estimadores apresentados acima, não podem ser usados nesse tipo de estudo porque não existe nenhuma forma de se incluir censuras.

O estimador a ser usado neste trabalho será o estimador não-paramétrico de Kaplan-Meier. Esse estimador é muito popular em pesquisas que usam análise de sobrevivência. O estimador é escrito da seguinte forma:

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \frac{n_j - d_j}{n_j},$$

Em que, n_j representa o número de dados em risco de falha, d_j são os dados que falharam no tempo t_j, em que, $0 \leq t_{(1)} \leq \ldots \leq t_{(n)}$, são os tempos distintos e ordenados de falha. Esta técnica não utiliza covariáveis para a estimação, mas pode usar variável categóricas para verificar se as funções estimadas são diferentes.

A representação gráfica desse método se comporta em uma função na forma de escada, uma vez que a estimação entre o tempo $t_{(j)}$ e $t_{(j+1)}$ é constante.

2.4 Modelos de Probabilidade

2.4.1 Modelo Log-Logístico

Para os casos em que T é uma variável aleatória contínua seguindo uma distribuição Log-Logística com parâmetro de escala α e de forma γ, sua função densidade de probabilidade é descrita como

$$f(t) = \frac{\gamma \left(\frac{t}{\alpha}\right)^{\gamma-1}}{\alpha \left[1 + \left(\frac{t}{\alpha}\right)^{\gamma}\right]^2}, \qquad t > 0 \tag{6}$$

Em que α e γ são constantes, ambas maiores que 0. Com base na função densidade, é possível descrever a função de sobrevivência da variável aleatória T como:

$$S(t) = \frac{1}{1 + \left(\frac{t}{\alpha}\right)^{\gamma}} \tag{7}$$

E a partir desta função pode-se encontrar a função risco acumulado:

$$H(t) = -\log\left[\frac{1}{1 + \left(\frac{t}{\alpha}\right)^{\gamma}}\right] \tag{8}$$

A partir da equação (6) e (7) ainda é possível definir a função de risco do modelo:

$$h(t) = \frac{\gamma(t/\alpha)^{\gamma-1}}{\alpha[1 + (t/\alpha)^{\gamma}]}. \tag{9}$$

Com relação ao comportamento da função de risco, quando γ é menor que 1, esta é monótona decrescente, enquanto para valores maiores que 1 a função tem um comportamento monótono crescente até certo ponto e depois decresce, o que representa função de risco unimodal.

2.5 Método de Máxima Verossimilhança

Para a estimação dos parâmetros do modelo, existe uma grande variedade de formas para efetuar tal procedimento. Como a característica principal da análise de sobrevivência é a presença de censuras, o procedimento também deve incorporar tal característica. Por esse motivo, é descartado alguns métodos para a estimação.

Um método que consegue incorporar a censura é o método da máxima verossimilhança. Este método tem como objetivo encontrar o valor do parâmetro que maximiza a probabilidade da amostra observada ser encontrada. Este método mostra-se adequado por permitir a incorporação das censuras através da inclusão da função de sobrevivência para os tempos censurados, enquanto os tempos em que ocorreram falha, considera-se a função densidade de probabilidade.

Para os tipos de censura à direita, a função de máxima verossimilhança a ser maximizada pode ser descrita analíticamente e a menos de constantes, é dada por (Colosimo e Giolo, 2006):

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \left[f(t_i;\boldsymbol{\theta})\right]^{\delta_i} \left[S(t_i;\boldsymbol{\theta})\right]^{1-\delta_i}, \tag{10}$$

Em que δ_i é a variável indicadora de falha e $\boldsymbol{\theta}$ é o vetor de parâmetros que serão estimados.

Para encontrar o estimador de máxima verossimilhança, deriva-se a função de verossimilhança, ou o log da função, em relação aos parâmetros e iguala as funções resultantes a zero encontrando assim um valor para todos os parâmetros no vetor $\boldsymbol{\theta}$. Para a resolução do sistema são geralmente utilizados métodos computacionais de otimização numérica, tais como Newton-Raphson e BFGS devido a grande complexidade dos cálculos.

2.6 Dados Grupados

A análise de sobrevivência com censura intervalar, tem como caso particular os dados grupados. Essa particularidade acontece quando todos os indivíduos são

avaliados dentro do mesmo instante, se tornando assim o mesmo intervalo para todas as observações. A outra forma que define-se os dados grupados é quando acontece um número excessivo de empates, ou seja, vários indivíduos falham ao chegar em determinado tempo. Este tipo de experimento pode ser encontrado em estudos de medidas repetidas longitudinais.

Para esse tipo de estudo, os tempos são separados em intervalos arbitrários e disjuntos isso significa que os intervalos não precisam ter o mesmo tamanho e que não podem haver interseções entre os intervalos. As funções utilizadas na análise, tais como função de sobrevivência e de risco são as mesmas utilizadas quando o tempo é contínuo.

Os intervalos em questão são formados por dois valores arbitrários, a_j e $a_{j+1} \in \mathbb{N}$, em que $j = 1, ..., k$ e k é uma constante indicando o número de intervalos. O intervalo $I_j = [a_j, a_{j+1})$ deve ser tal que $I_j \cap I_{j+1} = \{\} \quad \forall\ j \in k$.

A função de verossimilhança se diferencia da forma utilizada na análise de sobrevivência devido a agregação dos valores em intervalos. Segundo Biazatti (2017), a função de verossimilhança para esse caso pode ser descrita como:

$$\begin{aligned} L(\boldsymbol{\theta}) = \prod_{j=1}^{k} \Bigg\{ & \prod_{i \in F_j} 1 - \frac{S(a_{j+1})}{S(a_j)} \\ & \times \prod_{i \in R_j} \frac{S(a_{j+1})}{S(a_j)} \\ & \times \prod_{i \in C_j} \left[\frac{S(a_{j+1})}{S(a_j)} \right]^{1/2} \Bigg\}, \end{aligned} \tag{11}$$

em que, $\boldsymbol{\theta}$ é o vetor de parâmetros, k é o número de intervalos no estudo, F_j é o grupo de individuos que falharam, R_j é o grupo de indivíduos que sobreviveram, ou não falharam, nesse intervalo e C_j são os individuos que foram censurados naquele intervalo. Os valores de a_j e a_{j+1} são os limites dos intervalos, sendo o primeiro o limite inferior e o segundo o limite superior.

3 Metodologia

3.1 Material

A fim de propor um modelo de regressão para dados grupados, irá ser utilizado um banco de dados cedido pelo Instituto de Saúde Coletiva da Universidade Federal da Bahia. Esses dados foram obtidos a partir de um estudo conduzido por Barreto et al.(1994). O banco de dados é formado por 1207 crianças com idade entre 6 e 48 meses no início do estudo, que receberam placebo ou vitamina A. Dentre as variáveis se encontra a variável tempo, que é o tempo entre a primeira dose de placebo ou vitamina A e a ocorrência de diarreia na criança, a variável idade(0 - $<$ 24 meses, 1 - $\geq$ 24 meses), tipo de tratamento(0 - Placebo, 1 - Vitamina A) e a variável sexo(0 - Feminino, 1 - Masculino). Para o agrupamento dos tempos, foram utilizados os mesmos 12 intervalos obtidos por Biazatti (2017).

Além deste, será utilizado um banco de dados cedido pela Universidade Estadual da Paraíba. Este é formado por informações de mais de 600 alunos do curso de química da UEPB e seu objeto de estudo é o tempo desde a entrada de um estudante até a evasão desse aluno. O banco tem o ano de entrada e o semestre de entrada de cada aluno e a partir daí o acompanhamento se segue até o acontecimento do evento de interesse, da censura ou do fim do tempo mínimo do aluno. Para a modelagem serão adotadas 6 covariáveis, sendo estas Sexo do aluno(0 - Masculino, 1 - Feminino), Turno(0 - Diurno, 1 - Noturno), Escola(0 - Privado, 1 - Publico), Ingresso(0 - Vestibular, 1 - ENEM), Idade(0 - $<$ 19 anos, $\geq$ 19 anos), Origem(0 - Outras Cidades, 1 - Campina Grande). Para a separação dos intervalos, foi decidido que cada semestre seria um intervalo, sendo que os semestres vão de 0 a 12.

3.2 Métodos

3.2.1 Modelo de Regressão Log-Logístico

Nos estudos em sobrevivência, é comum utilizar modelos de regressão para entender a efeito de covariáveis explicativas no tempo de falha em estudo. Nesse

caso, utiliza-se uma reparametrização de um parâmetro da distribuição escolhida através de uma função de ligação. Para este trabalho, foi escolhida a distribuição Log-Logística e o parâmetro que será reparametrizado, será o parâmetro de escala citado na subseção 2.4.1.

A função de ligação tem como objetivo conectar as covariáveis com a variável resposta através de um preditor linear definido como o produto da matriz transposta de covariáveis com o vetor de parâmetros, tal que para um conjunto de covariáveis o parâmetro de escala presente na distribuição pode ser descrito como:

$$\alpha = g(\boldsymbol{\eta}) = \exp(\boldsymbol{x}^T\boldsymbol{\beta}) \tag{12}$$

em que $\boldsymbol{x}^T = (1, x_1, ..., x_p)$ é o vetor de covariáveis e $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ é o vetor dos coeficientes associados às covariáveis.

Com isso, é possível reescrever as funções presentes em 2.4.1 para adaptá-las ao modelo de regressão Log-Logístico com covariáveis. As funções necessárias para a realização da modelagem será a função de sobrevivência e para calcular o resíduo de Cox-Snell será utilizada também a função risco acumulada.

Dessa forma, a função de sobrevivência do modelo de regressão Log-Logístico pode ser descrita como:

$$S(t) = \frac{1}{1 + \left(\dfrac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\gamma}}, \tag{13}$$

em que $\gamma > 0$ é o parâmetro de forma.

E a partir desta, é possível encontrar a função risco acumulada:

$$H(t) = -\log\left[\frac{1}{1 + \left(\dfrac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\gamma}}\right] \tag{14}$$

Para a estimação dos parâmetros do modelo é utilizada a função de verossimilhança. Utiliza-se o log desta função para a estimação dos parâmetros. Assim,

utilizando a função de verossimilhança definida na equação 11, tem-se:

$$\begin{aligned}\log(L(\boldsymbol{\theta})) = \sum_{j=1}^{k} \Bigg\{ & \sum_{i \in F_j} \log\left[1 - \frac{S(a_{j+1}|\boldsymbol{x}_i)}{S(a_j|\boldsymbol{x}_i)}\right] \\ & + \sum_{i \in R_j} \log\left[\frac{S(a_{j+1}|\boldsymbol{x}_i)}{S(a_j|\boldsymbol{x}_i)}\right] \\ & + \sum_{i \in C_j} 0.5 \times \log\left[\frac{S(a_{j+1}|\boldsymbol{x}_i)}{S(a_j|\boldsymbol{x}_i)}\right] \Bigg\}\end{aligned} \tag{15}$$

Finalmente, substituindo (13) em (15) tem-se que:

$$\begin{aligned}l(\boldsymbol{\theta}) = \sum_{j=1}^{k} \Bigg\{ & \sum_{i \in F_j} \log\left[1 - \frac{1 + \left(\frac{a_j}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\gamma}}{1 + \left(\frac{a_{j+1}}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\gamma}}\right] \\ & + \sum_{i \in R_j} \log\left[\frac{1 + \left(\frac{a_j}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\gamma}}{1 + \left(\frac{a_{j+1}}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\gamma}}\right] \\ & + \sum_{i \in C_j} 0.5 \times \log\left[\frac{1 + \left(\frac{a_j}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\gamma}}{1 + \left(\frac{a_{j+1}}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\gamma}}\right] \Bigg\}\end{aligned} \tag{16}$$

em que, $l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta}))$. A partir desse ponto, deve-se seguir o processo de otimização citado na subseção 2.5, ou seja, calcula-se a derivada da função log-verossimilhança e encontra o ponto onde todos as equações geradas são iguais a zero. Por se tratarem de diversos parâmetros, são utilizados procedimentos de otimização numérica para se estimar tais valores. Para estes procedimentos será utilizado o *software* R(R Core Team (2018)), utilizando o método BFGS da função *optim*.

3.3 Resíduos de Cox-Snell

Para avaliar a qualidade do ajuste do modelo, realiza-se o cálculo de medidas que possam mensurar essa qualidade. Por se tratar de modelos censurados, a análise de resíduos na análise de sobrevivência não são tão visualizáveis e por isso, existem diversas medidas de resíduos passíveis de serem utilizados. Esta subseção aborda os resíduos de Cox-Snell.

Segundo Colosimo e Giolo (2006), esses resíduos são obtidos através da função risco acumulada obtida pelo modelo ajustado:

$$\hat{e_i} = \hat{H}(ti|\boldsymbol{x_i}) \tag{17}$$

em que $\boldsymbol{x_i} = (1, x_{i1}, ..., x_{ip})$ é o vetor de covariáveis do i-ésimo indivíduo. Os resíduos $\hat{e_i}$ devem seguir distribuição exponencial padrão (Lawless (2003)). Cada tempo presente no banco de dados gera um resíduo, desta forma cada resíduo possui um indicador de censura. Para realizar a análise gráfica do modelo, calcula-se o estimador de Kaplan-Meier para a função de sobrevivência dos resíduos e cria o gráfico desta estimativa sobreposta pela curva de sobrevivência da distribuição exponencial padrão dos resíduos.

A qualidade do ajuste é determinada a partir da proximidade entre as duas curvas. Caso a curva gerada pelos resíduos aparente ter uma distribuição exponencial e sobrepor de forma satisfatória a curva gerada pela exponencial padrão, então diz-se que o modelo possui um bom ajuste.

4 Resultados e Discussão

4.1 Suplementação de Vitamina

4.1.1 Análise Descritiva

Para iniciar a análise dos resultados do banco de suplementação de Vitamina A, será realizada a análise exploratória do tempo e das covariáveis. Primeiro, é realizado um histograma da variável tempo incluindo todos os tempo de censura e de falha.

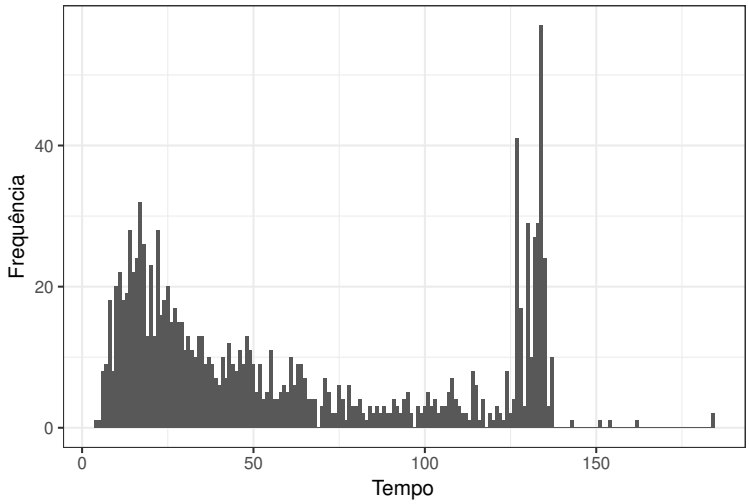

Figura 2: Histograma dos dados de Vitamina.

Pelo gráfico é possível notar que em grande parte dos tempos existem muitos empates nos tempos do estudo, ou seja, muitos tempos se repetem em diversas observações e essa é uma justificativa do porquê utilizar a metodologia de dados grupados.

Outro procedimento passível de ser realizado para a análise exploratória é a estimação da função de sobrevivência através de Kaplan-Meier. Na Figura 3 é possível notar que o decréscimo da função é mais rápido no começo da função e vai diminuindo conforme vai passando o tempo. Também é possível perceber uma grande concentração de censuras mais ao final do estudo.

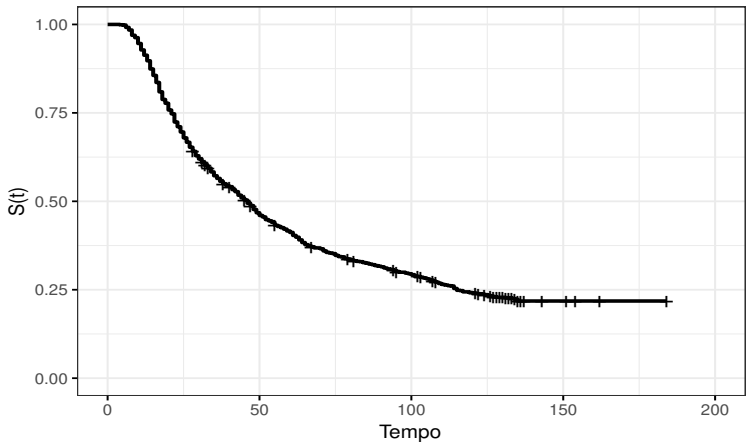

Figura 3: Função de Sobrevivência Estimada.

Para a realização do estudo utilizando a metodologia de dados grupados, o tempo de sobrevivência foi grupado em 12 intervalos, como dito na seção 3.1. A Tabela 1 apresenta os intervalos escolhidos, o número de falhas, o de censuras e o número de indivíduos sob risco.

Tabela 1: Descrição dos tempos de vida utilizando os intervalos grupados

Intervalo	Número de falhas	Número de censuras	Número sob risco
[4, 14)	124	0	1207
[14, 18)	106	0	1083
[18, 23)	103	0	977
[23, 29)	100	1	874
[29, 37)	92	3	773
[37, 48)	92	6	678
[48, 61)	90	1	580
[61, 73)	67	1	489
[73, 90)	46	3	421
[90, 108)	49	6	372
[108, 126)	10	11	317
[126, 185)	10	250	260

A Tabela 1 mostra um grande número de falhas em quase todos os intervalos com exceção das duas últimas, esse é mais um indicativo de que a metodologia para

dados grupados deve ser utilizada.

Após observar o comportamento da função de sobrevivência estimada considerando apenas o tempo, pode-se incluir covariáveis uma a uma e avaliar a diferença entre as curvas de sobrevivências geradas.

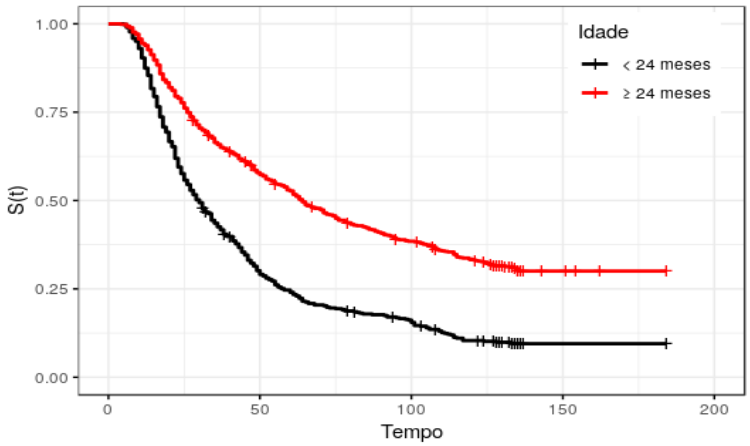

Figura 4: Função de Sobrevivência Estimada separada por idade.

Através da Figura 4, é perceptível que existe uma certa diferença entre os dois grupos. Crianças com idade maior ou igual 24 meses aparentam ter uma maior probabilidade de sobrevivência que crianças com idade menor que 24 meses.

Analisando a covariável sexo pela Figura 5, percebe-se que na maior parte do tempo, as curvas estão se sobrepondo o que indica que a variável não é muito significativa para o modelo, ou seja, o fato de ser menino ou menina não aparenta interferir significativamente na probabilidade de sobrevivência da criança.

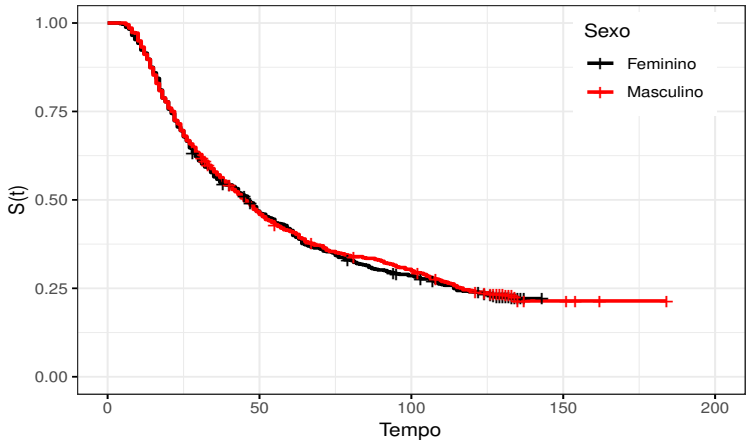

Figura 5: Função de Sobrevivência Estimada separada por sexo.

A Figura 6 mostra o efeito do tipo de tratamento na variável tempo. As curvas apresentam um certo distanciamento conforme o tempo passa. As crianças que foram suplementadas com Vitamina A aparentam ter probabilidade de sobrevivência maior que as que foram suplementadas com Placebo.

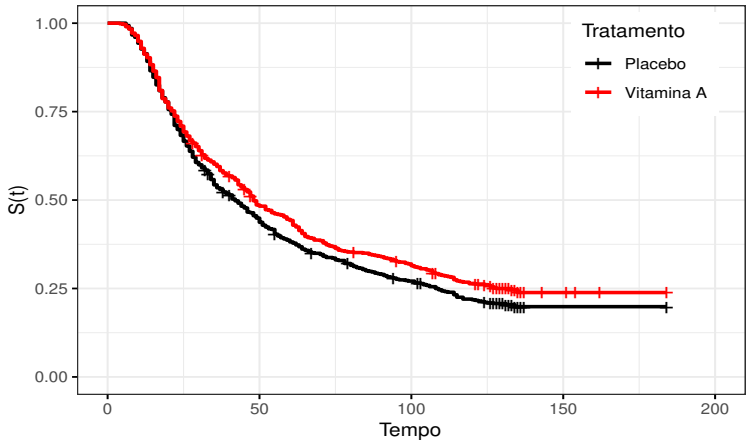

Figura 6: Função de Sobrevivência Estimada separada por tipo de tratamento.

O gráfico da função risco acumulada é utilizado para identificar um comportamento na função de risco e desta forma auxilia na suposição de um modelo adequado, como foi visto na seção 2.3.5.

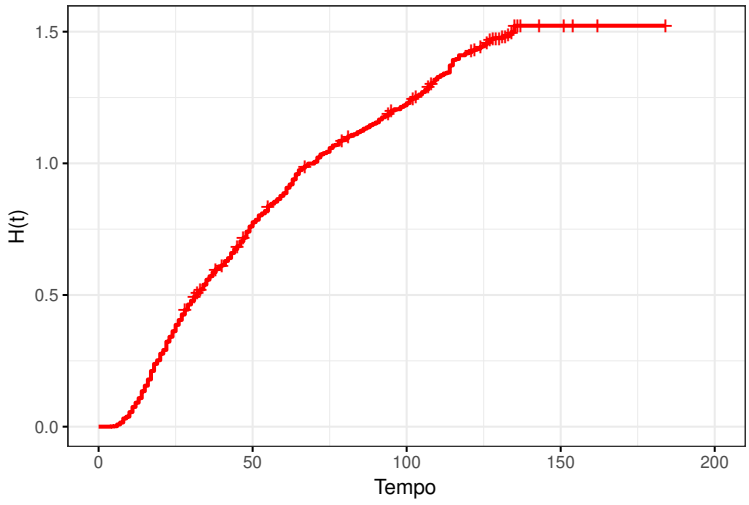

Figura 7: Função de Sobrevivência Estimada.

Observando a Figura 7, nota-se que o comportamento é muito similar ao (C) da seção 2.2.4 o que indica que a função de risco tem um comportamento decrescente. Desta forma, uma distribuição de probabilidade possível seria a Log-Logística.

Então com base na Figura 2 e na Tabela 1, conclui-se que a metodologia de dados grupados deveria ser utilizada. Com base na Figura 7, escolheu-se a distribuição Log-Logística para o modelo. Desta forma, o próximo passo será construir o modelo de regressão Log-Logístico para dados grupados.

4.1.2 Modelagem

Tendo como objetivo encontrar um modelo que explique bem a variável tempo de sobrevivência, nesta seção serão tratados a modelagem, a seleção de variáveis e a análise de resíduos.

Ao realizar a modelagem sem considerar nenhuma covariável e apenas estimar os parâmetros da distribuição Log-Logística, é possível realizar a comparação entre a função sobrevivência estimada pela função de máxima verossimilhança adapatada para o método de dados grupados e a função de sobrevivência estimada por Kaplan-Meier.

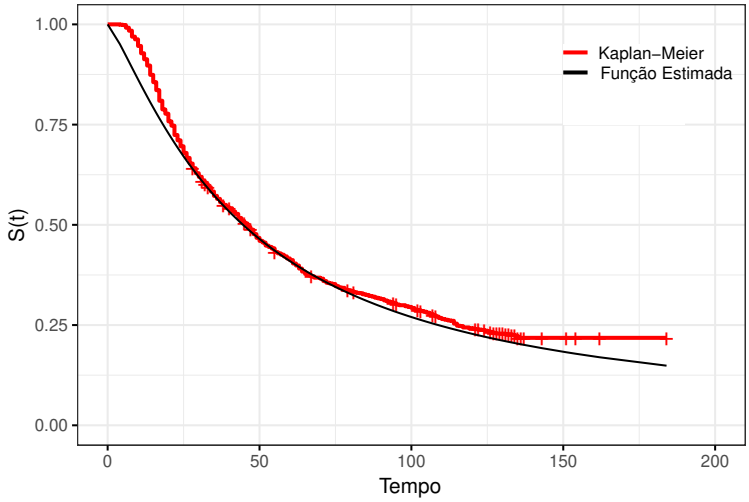

Figura 8: Função de Sobrevivência Estimada.

Ao que se pode observar a curva da distribuição estimada(curva suave em preto) se sobrepõe a boa parte da função estimada por Kaplan-Meier o que indica que a distribuição explica bem a variável resposta independente de covariáveis. Na Tabela 2 estão presentes os valores das estimativas de cada parâmetro, assim como seus erros padrões.

Tabela 2: Estimativas dos parâmetros da distribuição log-logística sem covariáveis.

Parâmetro	Estimativa	Erro Padrão
α	44,57	1,98
γ	1,23	0,045

Mais um indicativo para o uso do modelo em questão é o alto valor de α, o que indica uma grande heterogeneidade. Outra parte da modelagem que será realizada é a inclusão de covariáveis para verificar quais variáveis são significativas ao tentar explicar o tempo. Para isso, mais uma vez será realizada a estimação dos parâmetros através da função de máxima verossimilhança utilizando a metodologia em questão e para a inclusão das covariáveis, será utilizada a função de ligação $g(\eta) = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$. Assim, o parâmetro α pode ser escrito, utilizando todas as variáveis explicativas, da seguinte forma:

$$\alpha_i = exp(\beta_0 + \beta_1 Idade_i + \beta_2 Tratamento_i + \beta_3 Sexo_i) \tag{18}$$

Como pode ser observado na equação 18 as variáveis possuem o nome do que indicam. A Tabela 3 mostra os parâmetros estimados do modelo completo, assim como seus erros padrão e os p-valores com relação a significância do parâmetro.

Tabela 3: Estimativas dos parâmetros do modelo completo.

Parâmetro	Estimativa	Erro Padrão	p-valor
Intercepto	3,24	0,087	$< 0{,}001$
Idade	0,83	0,082	$< 0{,}001$
Tratamento	0,16	0,08	0,04
Sexo	0,05	0,07	0,48
γ	1,23	0.045	-

A tabela acima mostra que todas as covariáveis foram significantes menos a variável sexo. O Teste da Razão de Verossimilhança indica que o modelo retirando essa variável é adequado. A partir disto, é criada a Tabela 4 com as informações dos parâmetros utilizados nesse novo modelo.

Tabela 4: Estimativas dos parâmetros do modelo final.

Parâmetro	Estimativa	Erro Padrão	p-valor
Intercepto	3,28	0,075	$< 0{,}001$
Idade	0,83	0,082	$< 0{,}001$
Tratamento	0,16	0,08	0,04
γ	1,31	0.047	-

Pelos resultados, tem-se que todos as variáveis são significativas tanto a um nível de 5% de significância. Além disso, o Teste da Razão de Verossimilhança indica que não devem ser retiradas nenhuma variável do modelo e por isso, esse foi escolhido como candidato ao modelo final. Com o objetivo de verificar a qualidade do ajuste do modelo, realiza-se a análise gráfica de resíduos de Cox-Snell. A Figura 9 mostra os resíduos gerados pelo candidato a modelo final.

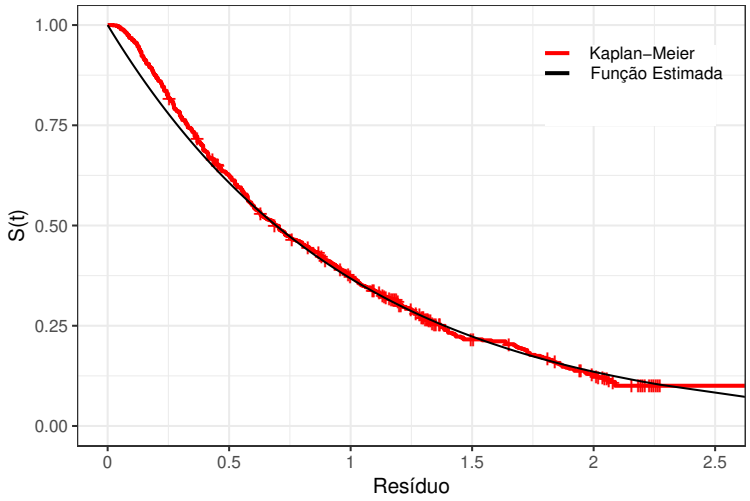

Figura 9: Resíduos de Cox-Snell.

No gráfico, é possível perceber que a curva do resíduo se aproxima bem da curva da exponencial padrão. Com isso, percebe-se que o modelo explica bem o tempo e por isso esse modelo pode ser considerado como modelo final.

Pelos resultados dos parâmetros, percebe-se que o modelo apresenta taxa de falha unimodal devido a estimação do parâmetro γ ser 1,31. A estimativa do parâmetro de idade indica que crianças com idade maior que 24 meses possuem maior probabilidade de sobrevivência. Quanto a variável tratamento, quem apresenta maior

probabilidade são as que crianças que foram suplementadas com Vitamina A. Os resultados encontrados pelo modelo de regressão Log-Logístico refletem os resultados ilustrados pela análise exploratória, o que reforça a qualidade do modelo encontrado.

4.2 Evasão Curso de Química

4.2.1 Análise Descritiva

A fim de explorar inicialmente os dados, é realizado um histograma da variável tempo sem considerar nenhuma covariável nem a indicação de censura.

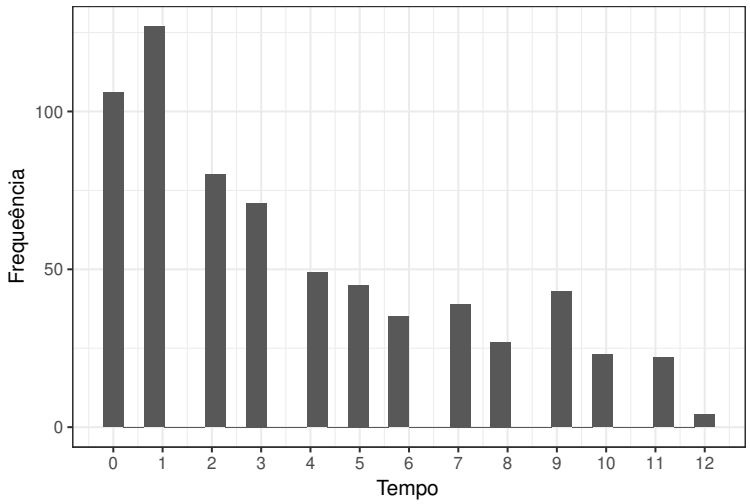

Figura 10: Histrograma dos tempos de evasão

Como é possível observar, devido ao fato de que os alunos só podem permanecer na universidade por 12 semestres existem muitos empates nos semestres. Além disso, o evento de interesse só é observado ao final dos semestres o que significa medir os dados a cada intervalo de um semestre e essa é a caracterização de dados grupados.

Para verificar a forma da estimativa da função de sobrevivência, é realizada a estimação por Kaplan-Meier e faz-se o gráfico dessa estimativa.

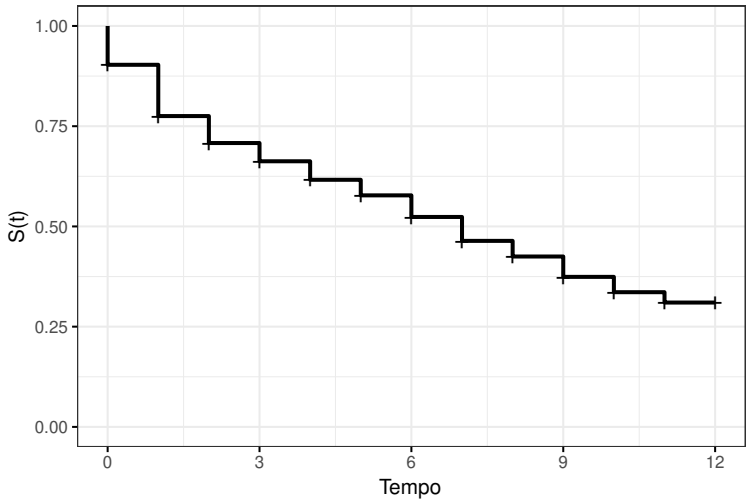

Figura 11: Histrograma dos tempos de evasão

A Figura 11 apresenta uma caída maior nos tempos iniciais, principalmente nos tempos 0 e 1. O gráfico possui um formato de escada devido ao tempo ser uma variável discreta e por isso as censuras só aparecem nos valores discretos. Como os tempos se tratam de semestres, o uso de um semestre como intervalo é uma forma natural de se pensar, assim são utilizados 13 intervalos. A Tabela 5 mostra os intervalos, o número de falhas, o de censuras e número de pessoas sob risco naquele intervalo.

Tabela 5: Descrição dos tempos de evasão utilizando os intervalos grupados.

Intervalo	Número de Falhas	Número de Censuras	Número sob risco
[0, 1)	65	41	671
[1, 2)	80	47	565
[2, 3)	38	42	438
[3, 4)	23	48	358
[4, 5)	20	29	287
[5, 6)	15	30	238
[6, 7)	18	17	193
[7, 8)	18	21	158
[8, 9)	10	17	119
[9, 10)	11	32	92
[10, 11)	5	18	49
[11, 12)	2	20	26
[12, 13)	0	4	4

A Tabela 5 mostra um grande número de empates na maioria dos intervalos, o que indica o uso da metodologia de dados grupados. Além disso, todas as observações são avaliadas nos mesmos intervalos, o que indica o caso particular da censura intervalar. Todos esses motivos apontam para o uso da metodologia de dados grupados.

Para observar o efeito das covariáveis, é realizado o gráfico de Kaplan-Meier para a variável tempo separada pelas variáveis explicativas.

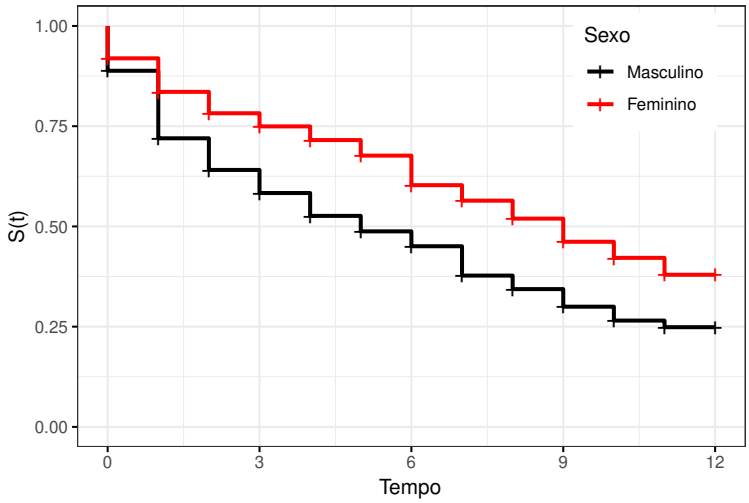

Figura 12: Função de sobrevivência por sexo

A Figura 12 mostra que existe um afastamento entre as curvas do sexo masculino e feminino o que pode indicar que a variável será significativa para o modelo. Essa figura ainda mostra que a probabilidade de sobrevivência das alunas do sexo feminino possuem uma probabilidade de sobrevivência menor que os alunos do sexo masculino.

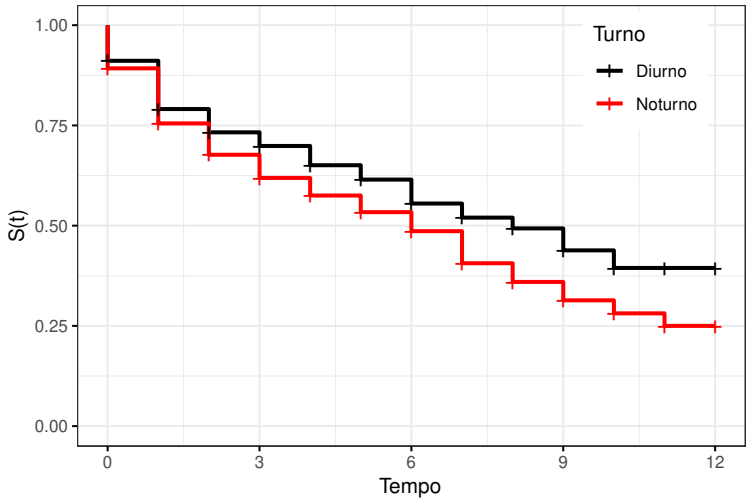

Figura 13: Função de sobrevivência por turno

Pela Figura 13 percebe-se que a diferença entre as curvas vai se acentuando conforme o tempo passa. Pelo distanciamento das curvas pode ser que a informação do turno do aluno seja significante, apesar do início próximo.

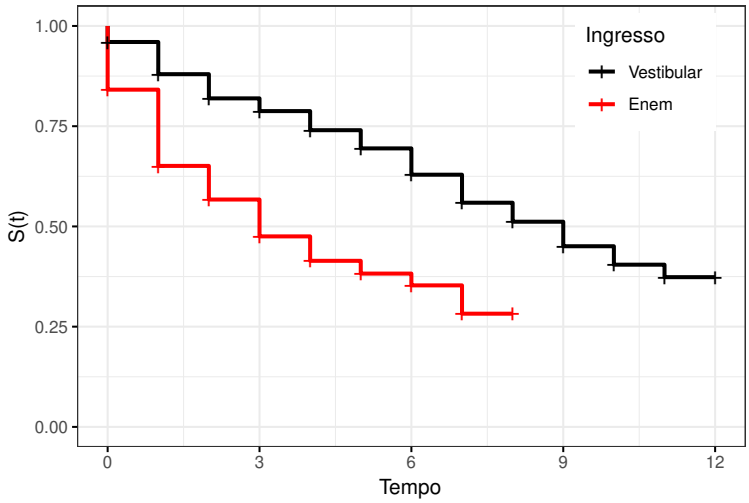

Figura 14: Função de sobrevivência por meio de ingresso

Percebe-se na Figura 14 uma grande diferença entre as curvas, o que pode indicar que a variável meio de ingresso explica bem a variável tempo de evasão.

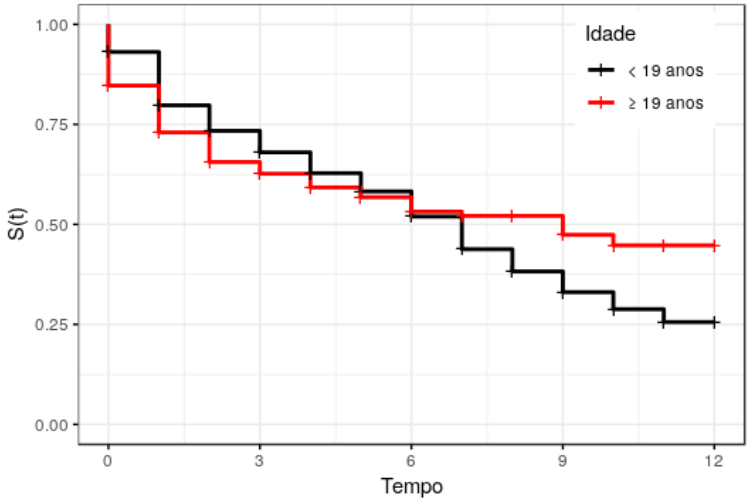

Figura 15: Função de sobrevivência por idade

A variável idade tem suas curvas bem próximas até o tempo 7 e após isso, as curvas se distanciam mais sendo difícil dizer se esta variável será ou não significativa, o que será mostrado ao realizar a modelagem.

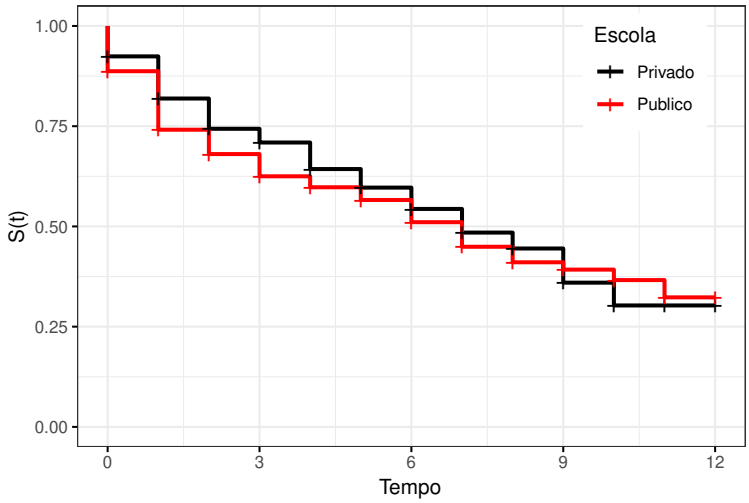

Figura 16: Função de sobrevivência por tipo de escola

O gráfico da variável escola (Figura 16) mostra que as curvas são extremamente próximas durante toda a sua extensão, podendo esta variável não ser significativa por não possuir diferença nas curvas.

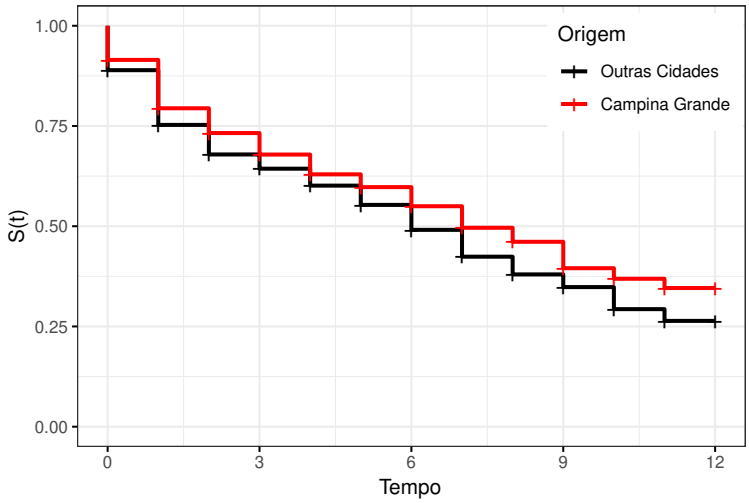

Figura 17: Função de sobrevivência por cidade de origem

Já a variável que indica a cidade de origem não aparenta explicar bem a variável tempo, devido a suas curvas serem muito próximas, como mostra a Figura 17.

Para a suposição da distribuição de probabilidade que será utilizada para a etapa de modelagem, é feito o gráfico da função risco acumulada com o objetivo de

observar seu comportamento e determinar a forma da função de risco, encontrando assim uma distribuição que seja adequada.

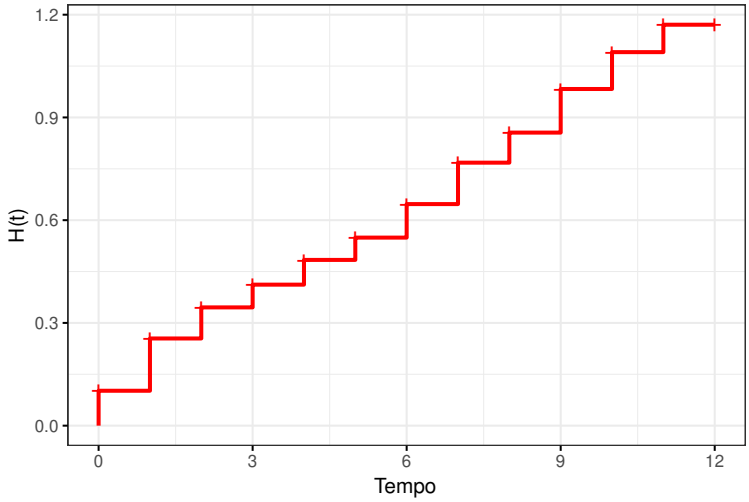

Figura 18: Função risco acumulada do tempo de evasão dos alunos.

Segundo a Figura 18, a função risco acumulada aparenta possuir um comportamento semelhante ao comportamento (C) da Figura 1 na seção 2.2.4, assim como na análise do banco anterior. Desta forma, também será utilizada a distribuição Log-Logística para o ajuste do modelo.

4.2.2 Modelagem

Nesta seção serão abordados os resultados do modelo final para o banco de dados do curso de química. Para a estimação do modelo, será utilizada a função de máxima verossimilhança para o caso de dados grupados e para a otimização da função será utilizada a função *optim()* do *software R*, assim como no modelo do banco anterior.

Realizando a modelagem sem nenhuma covariável, utilizando apenas os grupos dos tempos até a evasão dos alunos e a informação da censura pode-se verificar a explicação da variável apenas pela distribuição. A Figura 19 mostra o resultado desse modelo através da função de sobrevivência estimada por Kaplan-Meier(Curva Vermelha) e a função de sobrevivência da distribuição Log-Logística estimada pela verossimilhança modificada para incorporar as informações do grupo (Linha Preta).

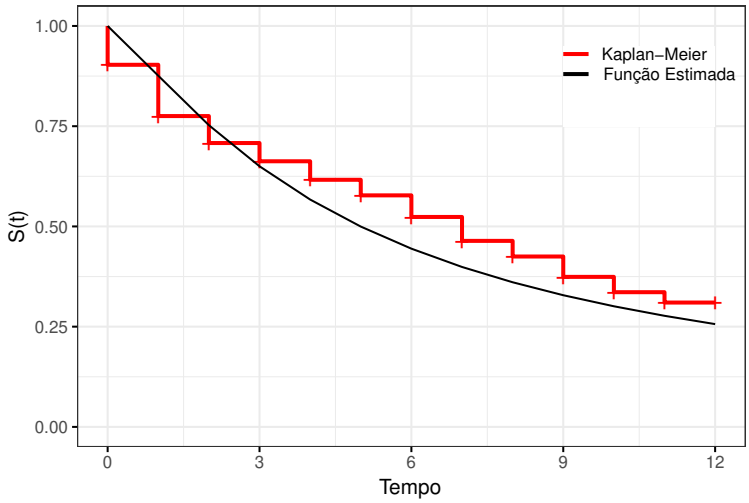

Figura 19: Função de Sobrevivência Estimada.

Como mostra a Figura 19, o modelo se aproxima bem no começo, subestima os valores após determinado tempo. Agora, são adicionadas covariáveis com o intuito de entender melhor se essas variáveis explicam o tempo de evasão dos alunos e quão bem elas explicam. A Tabela 6 mostra os parâmetros da Log-Logística estimada, assim como seus erros padrões.

Tabela 6: Estimativas dos parâmetros da distribuição log-logística sem covariáveis.

Parâmetro	Estimativa	Erro Padrão
α	4,99	0,33
γ	1,21	0,065

O preditor linear do modelo completo utilizado para ligar as variáveis explicativas a variável resposta pode ser descrito por:

$$\alpha_i = \exp(\beta_0 + \beta_1 Sexo_i + \beta_2 Turno_i + \beta_3 Escola_i + \beta_4 Ingresso_i + \beta_5 Idade_i + \beta_6 Origem_i) \tag{19}$$

Estima-se os parâmetros da equação (19) através do método de máxima verossimilhança. A Tabela 7 mostra o resultado da estimação dos parâmetros do modelo completo, seus erros padrão e o p-valor dessa estimativa.

Tabela 7: Parâmetros estimados do modelo completo.

Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	2,15	0,151	$< 0,001$
Sexo	0,41	0,109	$< 0,001$
Turno	-0,24	0,110	0,024
Escola	-0,01	0,111	0,91
Ingresso	-1,09	0,108	$< 0,001$
Idade	-0,18	0,121	0,123
Origem	0,12	0,110	0,284
γ	1,44	0,079	-

A Tabela 7 mostra que têm variáveis com p-valor muito baixo e outros muito altos. A um nível de significância de 10% o modelo possui 3 variáveis não significativas. Ao realizar um procedimento para a seleção de variáveis, foi definido que as variáveis Escola e Origem não são significativas a um nível de 10% e pelo teste da razão de verossimilhança foi determinado que o modelo candidato a modelo final está descrito na Tabela 8:

Tabela 8: Parâmetros estimados do modelo final.

Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	2,2	0,116	$< 0,001$
Sexo	0,42	0,109	$< 0,001$
Turno	-0,25	0,109	0,017
Ingresso	-1,07	0,109	$< 0,001$
Idade	-0,21	0,107	0,078
γ	1,44	0,079	-

Na Tabela 8, é possível ver que todas as variáveis têm o p-valor abaixo do nível de significância de 10% o que indica que todas essas informações são significativas a esse nível, o que aponta que o modelo é um bom candidato a modelo final.

Para decidir se esse modelo pode ser ou não o modelo final, é realizada a análise gráfica de Cox-Snell.

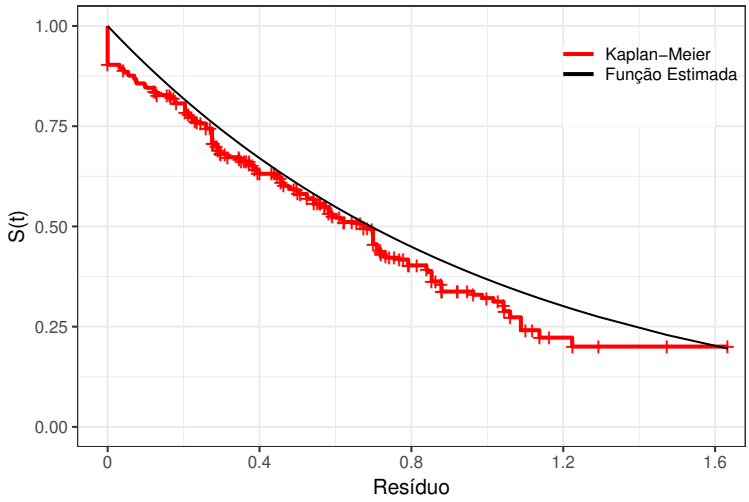

Figura 20: Gráfico do resíduo de Cox-Snell.

Como é possível perceber na Figura 20, a curva do resíduo se aproxima bem da curva da exponencial padrão, o que indica um bom ajustamento do modelo. Desta forma, é possível escolher esse modelo como modelo final.

O modelo mostra que as alunas do sexo feminino possuem maior probabilidade de sobrevivência com relação aos alunos do sexo masculino. Pela variável turno, os alunos que estudam no período diurno possuem maior probabilidade de continuar no curso. A variável ingresso indica que os alunos que ingressaram pelo Vestibular têm maior probabilidade de sobreviver no curso quando comparado com alunos ingressantes pelo Enem. Por fim, os alunos com idade maior ou igual a 19 têm uma probabilidade de sobrevivência menor que os alunos menores de 19 anos.

5 Conclusão

Neste trabalho foi proposto o uso do modelo de regressão Log-Logístico para dados de sobrevivência grupados. Para a aplicação do modelo foram escolhidos dois bancos de dados, o de Vitamina A e os dados do curso de química. Para a aplicação no primeiro banco encontrou-se que as variáveis que explicavam bem o modelo foram idade e tipo de tratamento, enquanto no segundo modelo encontrou-se que as variáveis significativas foram o sexo do aluno, o turno em que ele estuda, a forma de ingresso e a idade do mesmo.

Ambos os dados têm muitos empates nos tempos o que indica o uso da metodologia de dados grupados. Mesmo assim, o modelo de regressão Log-Logístico forneceu bons resultados para ambos. Outra especificidade do banco é a natureza dos tempos ser discreta, pois a metodologia de dados grupados ainda usa a função de sobrevivência do tempo de natureza contínua e mesmo assim, os resultados do modelo foram bons até mesmo em dados com poucos tempos diferentes no caso do banco de química.

O modelo de regressão Log-Logístico, aparenta ter sido uma boa escolha de distribuição tendo em vista o bom ajustamento dos resíduos de Cox-Snell. Como continuação do trabalho, seria interessante observar o ajuste levando em consideração outras distribuições com a função de risco decrescente.

Existe um indicativo de indivíduos curados nos dois bancos de dados, ou seja, que deixaram de ser sucetíveis ao evento de interesse. Com isso, talvez obtenha-se um melhor ajuste ao considerar o modelo de regressão Log-Logístico com fração de cura.

Referências

Barreto, M. L., Santos, L. M. P., Assis, A. M. O., Araújo, M. P. N., Franzena, G. G., Santos, P. A. B., & Fiaccone, R. L. (1994). Effect of vitamin a supplementation on diarrhoea acute lower-respiratory-tract infection in young children in brazil. *Lancet 344*, pages 228–231.

Biazatti, E. C. (2017). Modelo de regressão log weibull com fracão de cura para dados grupados. Master's thesis, Universidade de Brasília. Departamento de Estatística - Instituto de Ciências Exatas.

Colosimo, E, A. & Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada*. Editora Bucher.

dos Santos, D. F. (2017). Modelo de regressão log-logístico discreto com fracão de cura para dados de sobrevivência. Master's thesis, Universidade de Brasília. Departamento de Estatística - Instituto de Ciências Exatas.

Hashimoto, E. M. (2008). Modelo de regressão para dados com censura intervalar e dados de sobrevivência grupados. Master's thesis, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, (2nd edition ed.). John Wiley and sons.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Resende, V. S. (2017). Modelo de regressão log-beta burr iii para dados grupados. Master's thesis, Universidade de Brasília. Departamento de Estatística - Instituto de Ciências Exatas.

Y., N. E. (2017). *Um Curso de Análise de Sobrevivência*.