Relatório PIBIC 2018: Estudo de Modelos para Dados Discretos em Análise de Sobrevivência

Daniel Lima Viegas[*]
Juliana Betini Fachini-Gomes[†]

1 Introdução

A Análise de sobrevivência é a área da Estatística aplicada nas áreas da saúde, das ciências sociais e econômicas chegando até às engenharias. Por possuir tantas aplicações, essa área também é chamada de Análise de Confiabilidade.

Nesse tipo de análise o objeto de estudo é o tempo até a ocorrência de um evento de interesse. Sua grande diferença com relação às outras áreas da Estatística é a presença de censura no estudo, ou seja, dados cujo a observação de tempo não se concluiu por algum motivo.

O tempo juntamente com essa censura, formam a variável resposta. Considerando o tempo como uma variável aleatória, é possível aplicar certos tipos de gráficos e técnicas com o propósito de descobrir uma possível distribuição para esses dados e assim realizar possíveis inferências e tirar as devidas conclusões.

Na grande maioria dos trabalhos nessa área mensuram a variável tempo de forma contínua, pela grande quantidade de distribuições de probabilidade as quais os dados podem se ajustar. Quando a variável é medida de forma discreta, é necessário utilizar de alguns artifícios para o estudo. Um desses artifícios é o uso da discretização de uma distribuição de probabilidade. Esse método causa uma adaptação nos métodos tradicionais da análise de sobrevivência. Segundo Colosimo e Giolo (2006), uma distribuição de probabilidade muito utilizada como alternativa de distribuições mais flexíveis é a Log-Logística, pois é aplicável em muitas situações práticas.

2 Revisão de Literatura

2.1 Análise de Sobrevivência

Em determinados tipos de estudo é desejável estudar o tempo até a ocorrência de determinado evento de interesse, na Estatística, o nome que se dá a esse tipo de estudo é Análise de Sobrevivência. O evento de interesse pode ter nomes diferentes para diferentes áreas, em geral o termo usado pela maioria é falha, na medicina esse evento pode ser a morte do paciente, a cura ou a manifestação de uma doença. Na engenharia, em geral, esse termo se refere a falha de um equipamento e nas ciências econômicas esse termo pode se referir a inadimplência de um determinadoo cliente.

O tempo nesse estudo pode ser medido de diversas formas como dias, meses, anos e até intervalos de tempo pré-determinados. Por ser um tipo de estudo que é observado ao longo do tempo, o acompanhamento de determinadas observações pode ser interrompido por diversos

[*]e-mail: daniel.limaviegas454@gmail.com
[†]e-mail: jfachini@unb.br

motivos. Dentre esses motivos podem estar a desistência de um paciente em participar do estudo por motivos pessoais ou o defeito de um produto por outro motivo que não o desejado.

Diferentemente de outras áreas da Estatística, na Análise de Sobrevivência esses dados incompletos também são utilizados e se caracterizam como censuras. Essa censura forma a variável resposta juntamente com as observações completas do tempo. Dentre essas censuras existem três tipos principais, à direita, à esquerda e a intervalar. A censura à direita acontece quando o tempo registrado no estudo é maior que o tempo de início do estudo.

Dentro da censura à direita, ainda é possível realizar a divisão dessa censura em três tipos. A censura de tipo I ocorre quando existe um tempo limite para a ocorrência desse evento, caso esse tempo seja atingido, todos as observações que não manifestaram o evento são marcadas como censura. Na censura de tipo II, o número de falhas é fixado no começo do estudo e ao atingir esse número de falhas, as outras observações são marcadas como censuras. A censura de tipo III é a censura aleatória e engloba as duas censuras anteriores, esta censura se caracteriza por possuir censuras que não se sabe o motivo dela ter acontecido. Nesse trabalho será utilizada a censura à direita aleatória.

A utilização desse tipo de censura no estudo se mostra importante, pois apesar de não ter apresentado falha, há a informação de que a observação ainda poderia apresentar tal falha caso o estudo tivesse continuado. A ausência dessa censura pode causar um viés às estimativas e ainda não mostrar a verdadeira distribuição dos dados.

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, discreta, não negativa. Pode-se descrever a função de sobrevivência como:

$$\begin{aligned} S(t) &= P(T > t) \\ &= \sum_{j=t+1}^{\infty} P(T = k), \qquad t = 0, 1, 2, 3, ... \end{aligned} \tag{1}$$

Segundo Colosimo e Giolo (2006), uma das funções mais importantes para a análise de sobrevivência é a função de risco ou função de taxa de falha. Esta função é definida pela probabilidade de que o evento ocorra dentro de um intervalo, dado que não ocorreu antes do limite inferior desse intervalo. Essa função pode ser definida como:

$$h(t) = \frac{-d \log S(t)}{dt}, \qquad t = 0, 1, 2, 3, ...$$

2.2 Estimação da Função de Sobrevivência por Kaplan Meier

O estimador a ser usado nesse 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:

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

.

Onde, 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 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 da forma de escada, uma vez que a estimação entre o tempo $t_{(j)}$ e $t_{(j+1)}$ é constante.

2.3 Distribuição Log-Logística

Para os casos onde T é uma variável aleatória contínua seguindo uma distribuição Log-Logística, sua função densidade de probabilidade é descrita como:

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

Com isso, é possível definir a função de sobrevivência para a distribuição Log-Logística como:

$$S(t) = \frac{1}{1 + \left(\frac{t}{\mu}\right)^{\lambda}}, \qquad t > 0, \mu > 0, \lambda > 0 \tag{3}$$

Segundo Santos (2017), dada uma variável aleatória contínua T, é possível encontrar sua função de probabilidade discretizada a partir de sua função de distribuição de probabilidade e função de Sobrevivência, através de:

$$\begin{aligned} p(t) &= P(T = t) \\ &= P(t \leq T < t + 1) \\ &= P(T < t + 1) - P(T \leq t) \\ &= F_T(t + 1) - F_T(t) \\ &= [1 - S_T(t + 1)] - [1 - S_T(t)] \\ &= S_T(t) - S_T(t + 1) \end{aligned} \tag{4}$$

Dado que T é uma variável com distribuição Log-Logística, a função de probabilidade discretizada dessa variável pode ser descrita como:

$$p(t) = \frac{1}{1 + \left(\frac{t}{\mu}\right)^{\lambda}} - \frac{1}{1 + \left(\frac{t+1}{\mu}\right)^{\lambda}}, \tag{5}$$

em que, μ é o parâmetro de escala e λ é o parâmetro de forma.

Com relação ao comportamento da função de risco, quando λ é menor ou igual a 1, esta é monótona decrescente, enquanto para valores maiores que 1 a função cresce até um valor

máximo e após isso tem um comportamento decrescente, ou seja, assumindo função de risco unimodal.

3 Modelo de Regressão Log-Logístico discreto

Uma das técnicas mais utilizadas para tentar explicar a variabilidade de uma variável segundo um conjunto de covariáveis é o modelo de regressão. Na análise de sobrevivência, utiliza-se de covariáveis para tentar explicar o tempo até a ocorrência do evento de interesse, com base na correlação entre a variável tempo e cada uma das covariáveis. Segundo Santos (2017), dado um vetor de covariáveis $\boldsymbol{x}^T = (1, x_1, ..., x_p)$ utiliza-se uma função de ligação $g(.)$ que conecte a variável resposta ao vetor $\boldsymbol{x}^T$.

Definindo o preditor linear como $\boldsymbol{x}^T\boldsymbol{\beta}$, em que, $\boldsymbol{\beta}$ é o vetor com p+1 coeficientes de regressão. Considerando uma variável aleatória T com distribuição log-logística discreta definida na seção 2.3, pode-se utilizar o parâmetro de escala μ, com $\mu > 0$ como a função de ligação, ou seja, $\mu = g(\eta) = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$. Com isso, é possível descrever o modelo de regressão Log-Logístico discreto como:

$$p(t|x) = \frac{1}{1 + \left(\frac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\lambda}} - \frac{1}{1 + \left(\frac{t+1}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\lambda}}, \qquad t > 0, \lambda > 0, -\infty < \boldsymbol{\beta} < \infty, \tag{6}$$

em que, λ é o parâmetro de forma e a $\exp(\boldsymbol{x}^T\boldsymbol{\beta})$ é o parâmetro de escala

A partir dessa função, é possível descrever também a função sobrevivência da seguinte forma:

$$S(t|x) = \frac{1}{1 + \left(\frac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta}}\right)^{\lambda}}, \qquad t > 0, \lambda > 0, -\infty < \boldsymbol{\beta} < \infty \tag{7}$$

Como a função de risco pode ser descrita como uma função da função de sobrevivência, esta possui a seguinte forma:

$$h(t|x) = 1 - \frac{1 + \left[\frac{t}{\exp(\boldsymbol{x}^T\beta)}\right]^{\lambda}}{1 + \left[\frac{t+1}{\exp(\boldsymbol{x}^T\beta)}\right]^{\lambda}}, \qquad t > 0, \lambda > 0, -\infty < \boldsymbol{\beta} < \infty \tag{8}$$

Onde, λ é o parâmetro de forma e $\boldsymbol{\beta}$ é o vetor de coeficientes do modelo.

4 Método de Máxima Verossimilhança para dados discretos

Para a estimação dos parâmetros da função de distribuição, e também para os 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.

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

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \left[p(t_i;\boldsymbol{\theta})\right]^{\delta_i} \left[S(t_i;\boldsymbol{\theta})\right]^{1-\delta_i}, \tag{9}$$

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

Para o modelo Log-Logístico, utilizando as equações 6 e 7, a função de máxima verossimilhança a ser maximizada possui a seguinte forma:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \left[\frac{1}{1+\left(\frac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\lambda}} - \frac{1}{1+\left(\frac{t+1}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\lambda}}\right]^{\delta_i} \left[\frac{1}{1+\left(\frac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta}}\right)^{\lambda}}\right]^{1-\delta_i}, \tag{10}$$

A partir da equação 10, é possível obter os parâmteros do modelo encontrando o ponto de máximo global na função. Isto pode ser feito ao resolver o sistema:

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

5 Análise de Dados

A fim de ilustrar o funcionamento do modelo de regressão Log-Logístico, serão utilizados dois bancos de dados com aplicações distintas. O primeiro banco de dados é o banco utilizado por Barreto et al. (1994). Este banco estuda o tempo desde a suplementação de vitamina A ou placebo, até o primeiro episódio de diarréia em crianças com idades entre 0 e 24 meses. Os dados foram obtidos ao acompanhar pouco mais de 1200 crianças e possui em sua estrutura 3 variáveis explicativas.

O segundo banco de dados foi cedido pela Universidade Estadual da Paraíba (UEPB). O banco estuda o tempo que um aluno leva para evadir do curso de Química nesta universidade com mais de 600 alunos presentes no conjunto. O banco possui informações sobre os alunos como covariáveis.

5.1 Suplementação de Vitamina em Episódios de Diarréia

No estudo de suplementação de vitamina, o evento de interesse era a ocorrência do primiero episódio de diarréia em crianças com idades entre 0 e 24 meses. Para o estudo, foi definido que um episódio da doença é quando a doença persiste por uma sequência de dias.

Além da covariável que indica o uso de vitamina A ou placebo, o banco também conta com a idade da criança e o sexo. Para o estudo, foram coletadas as informações de 1207 crianças.

5.1.1 Análise Descritiva

Para observar inicialmente o comportamento da variável tempo e levar em consideração as censuras ao longo do estudo, utiliza-se o estimador de Kaplan-Meier para a função de sobrevivência e assim poder estimar a curva de sobrevivência. Utiliza-se a função risco acumulado estimada para encontrar uma possível distribuição para realizar a modelagem através de seu gráfico.

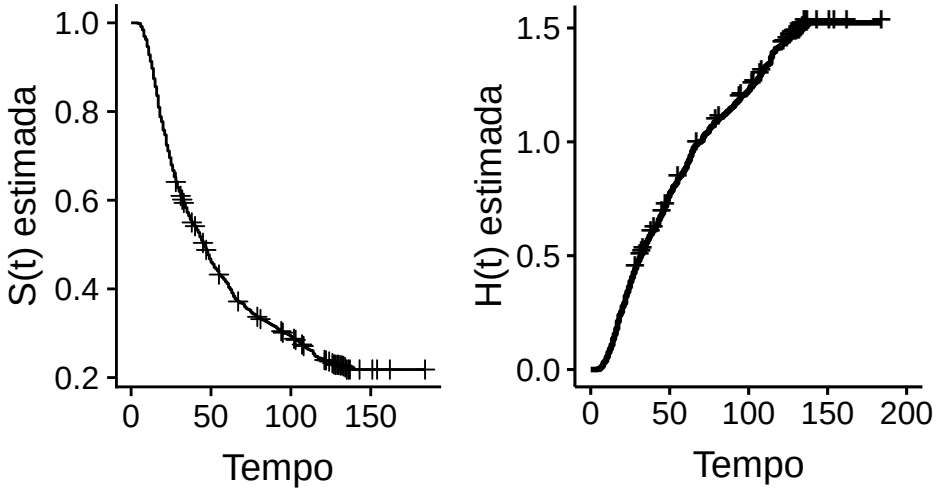


Figura 1: À esquerda, está a curva de sobrevivência e à direita, está a curva risco acumulado.

Pode-se perceber pelos gráficos que as censuras começam a aparecer em tempos próximos de 50 e grande parte das censuras se concentram no final das curvas, próximo ao tempo 150. É possível perceber na curva risco acumulado na 1, um comportamento decrescente. Esse tipo de comportamento é o indicador de que a distribuição Log-Logística é uma das possíveis distribuições a serem estudadas.

Outra forma de se analisar descritivamente as variáveis em sobrevivência, é realizar o gráfico de sobrevivência facetado pelas covariáveis categóricas. Para a variável numérica idade, foi realizada uma categorização para a análise. Valores abaixo de 12 meses foram descritos como 0 e acima disso como 1.

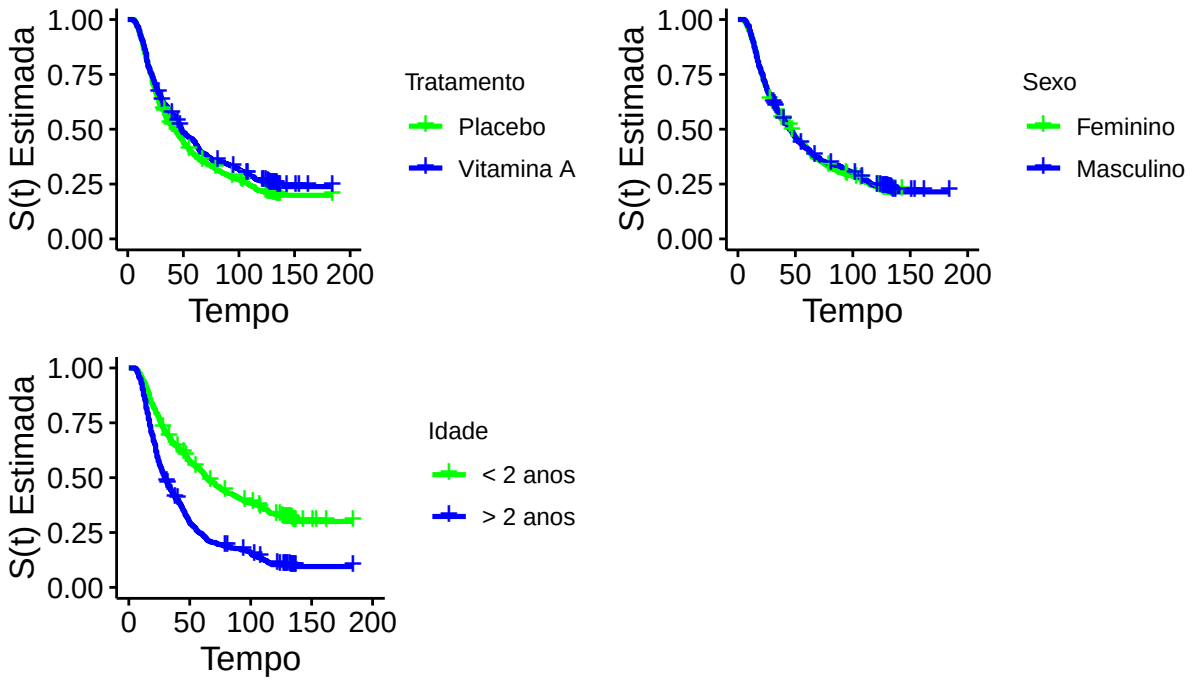


Figura 2: Curvas de sobrevivência facetadas pelas covariáveis

As variáveis que melhor explicam o tempo sem considerar nenhuma distribuição, são aquelas que possuem maior diferença nas curvas quando facetadas. Ao observar os gráficos na Figura 2 percebe-se que a variável de idade possui curvas muito diferentes, o que indica que a covariável explica bem o tempo. A variável tratamento também apresenta diferença, apesar de não ser tão perceptível. O sexo da criança não mostrou grande diferença, o que indica que esse fator não é muito explicativo para a variável resposta.

5.1.2 Modelagem dos dados

Ao realizar a construção do modelo completo, obteu-se os seguintes parâmetros com seus respectivos p-valores:

	Parâmetros	P.valor
Intercepto	4.12	0.00
Idade	-0.68	0.00
Tratamento	0.13	0.04
Sexo	0.05	0.45

Tabela 1: Parâmetros do modelo completo

A partir da tabela, percebe-se que a idade e o tipo de tratamento são informações significativas para o modelo, enquanto a informação de sexo, não. Desta forma, utilizando o método *backward* para seleção de variáveis com o p-valor como critério de retirada.

Com isso, o modelo escolhido em seguida foi com as variáveis de idade e o tipo do tratamento:

	Parâmetros	P.valor
Intercepto	4.15	0.00
Idade	-0.68	0.00
Tratamento	0.13	0.04

Tabela 2: Parâmetros do modelo sem a variável sexo

Observando a Tabela 2, nota-se que todas as variáveis são significativas ao nível de 5% de significância. Pelo método *backward*, não tem nenhuma variável a ser retirada do modelo, sendo assim, este é o modelo escolhido no estudo.

Para verificar a qualidade do ajuste do modelo, utiliza-se o critério gráfico de Cox-Snell. Neste critério, compara-se a curva de sobrevivência do resíduo do modelo com a curva de sobrevivência de uma exponêncial com parâmetro 1.

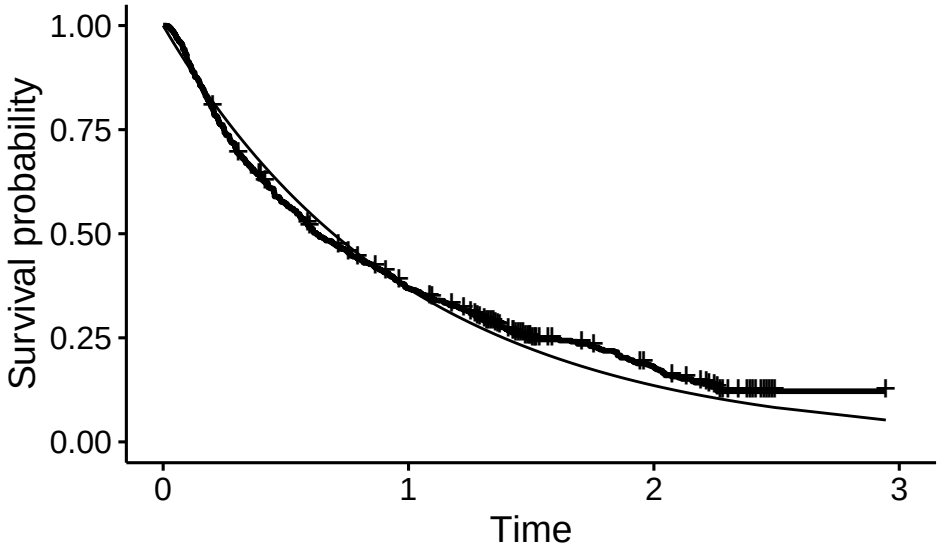


Figura 3: Resíduo de Cox-Snell.

Pela Figura 3, percebe-se que a curva estimada é próxima da curva da exponencial, o que mostra que o modelo ajustado possui boa qualidade, e por isso o modelo final para esse estudo é o modelo com as variáveis de idade e tipo de tratamento.

A interpretação do parâmetro de idade se dá por $e^{-0.6753061} = 0.5090006$, o que pode ser interpretado como o tempo mediano das crianças com idade menor que 2 anos, é aproximadamente metade do tempo mediano das quais são mais velhas. Com relação a variável do tipo de tratamento, sua interpretação $e^{0.1272767} = 1.1357312$, significa que o tempo mediano para pessoas que tomaram placebo é 1.13 vezes maior que as que foram suplementadas com vitamina A.

5.2 Evasão dos alunos de química

Para este estudo, o evento de interesse foi definido como a evasão do aluno, é dito que houve censura caso o aluno se forme, ou o estudo acabe antes da evasão. O banco conta com 671 alunos do curso de química e 6 possíveis variáveis explicativas, sendo essas: Sexo, Turno, Tipo da Escola, Por onde ingressou, Idade e a Origem do aluno. Todas as variáveis são

binárias, inclusive idade que está definida como “Maior que a idade mediana” e “Menor que a idade mediana”.

5.2.1 Análise Descritiva

Primeiramente observa-se o comportamento da variável tempo através do estimador de Kaplan-Meier, incorporando as censuras e verificando o comportamento da função risco acumulado:

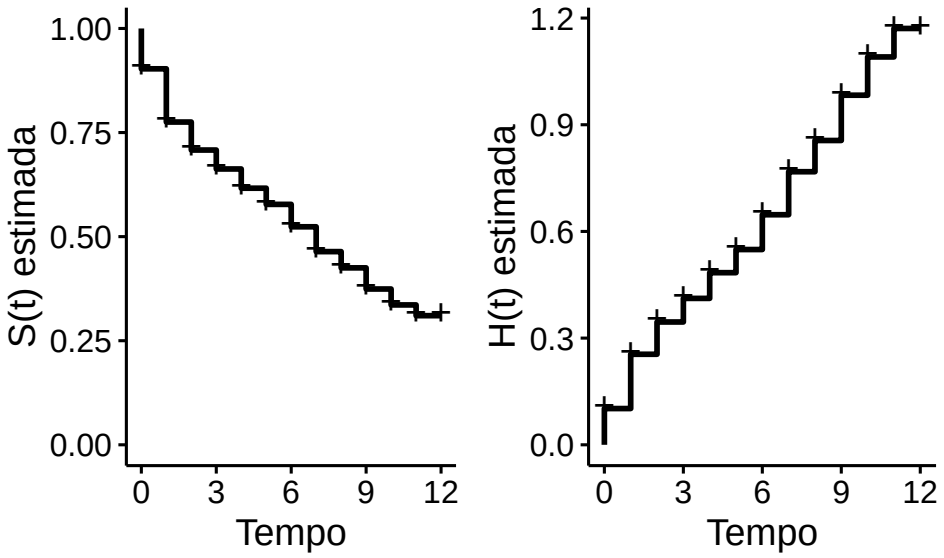
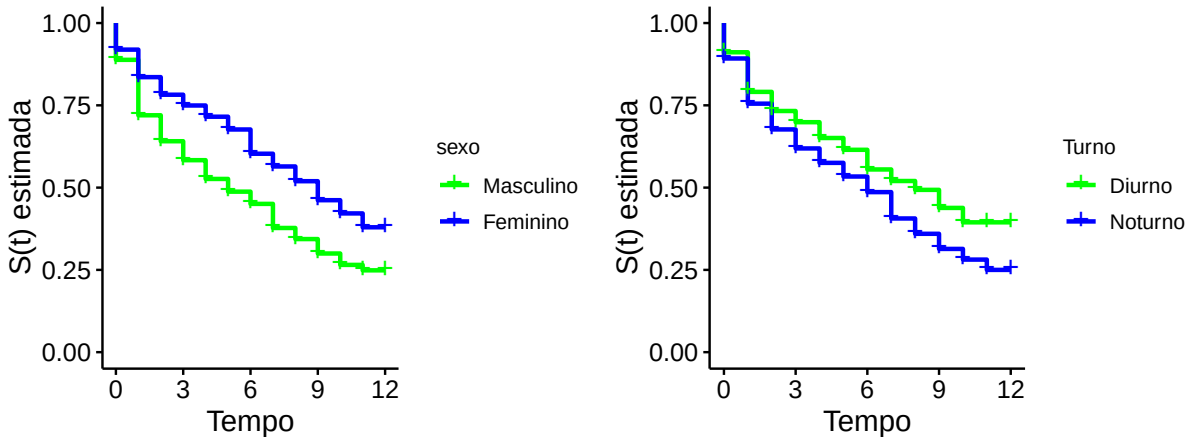


Figura 4: À esquerda, está a curva de sobrevivência e à direita, está a curva risco acumulado.

É interessante notar pelo gráfico que as probabilidades de sobrevivência vão diminuindo de forma muito mais rápida nos primeiros tempos e para os tempos finais a probabilidade vai convergindo para algo em torno dos 0,25. A função risco acumulada à direita mostra um comportamento decrescente o que mostra que a distribuição estudada se adequa bem aos dados.

Para verificar a relação entre o tempo e as variáveis explicativas, pode-se realizar o procedimento de Kaplan-Meier para cada uma das variáveis. Esse procedimento ajuda a entender se cada variável, individualmente, pode influenciar no tempo sem depender de distribuição nenhuma.



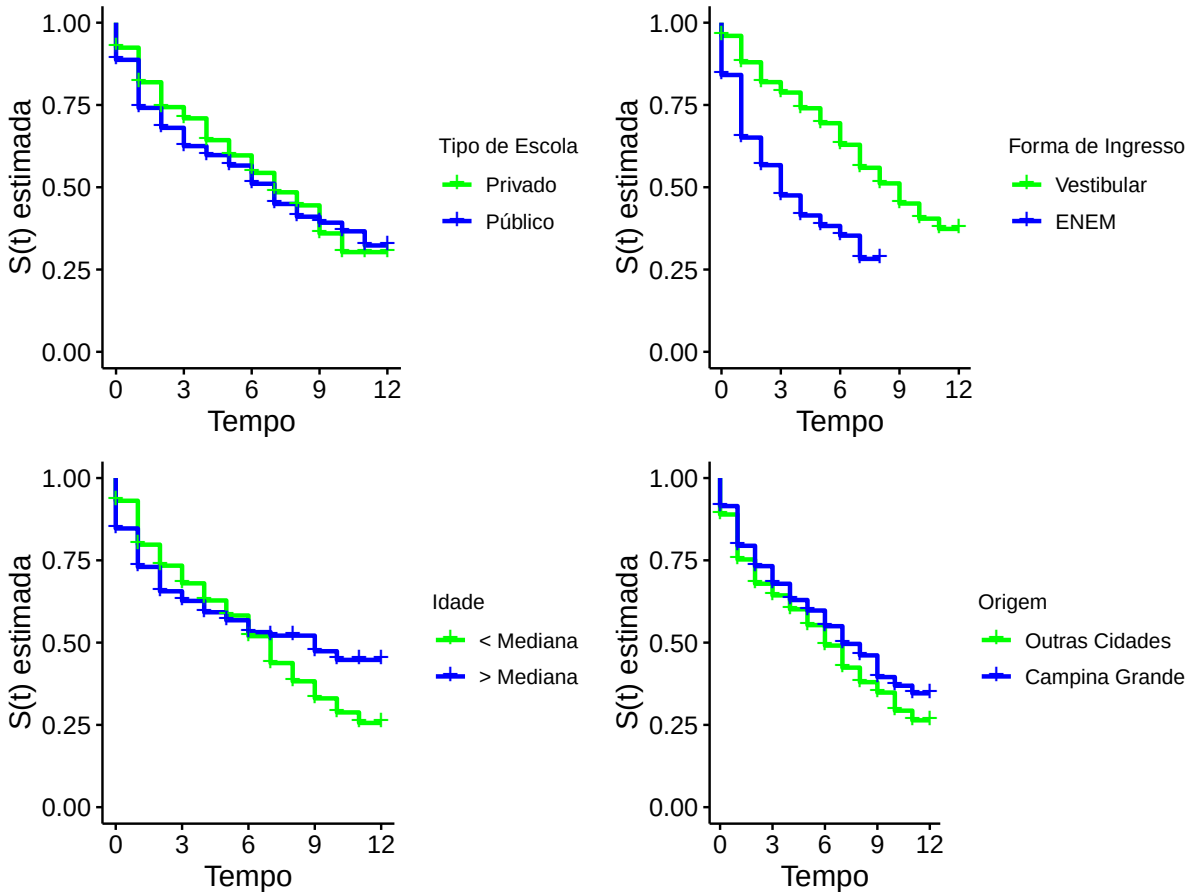


Figura 5: Curva de sobrevivência estimada para as covariáveis

Segundo os gráficos de cada covariável, nem todas aparentam ser bem explicativas para a resposta. Variáveis como tipo de escola e origem do aluno, possuem curvas muito parecidas entre suas categorias. Por outro lado, as variáveis de sexo e forma de ingresso se distanciam o suficiente para supor boa correlação com o tempo. As variáveis turno e idade são parecidas ao longo de parte da curva e depois começam a se diferenciar, não se sabe ao certo se elas poderiam explicar bem o tempo ou não.

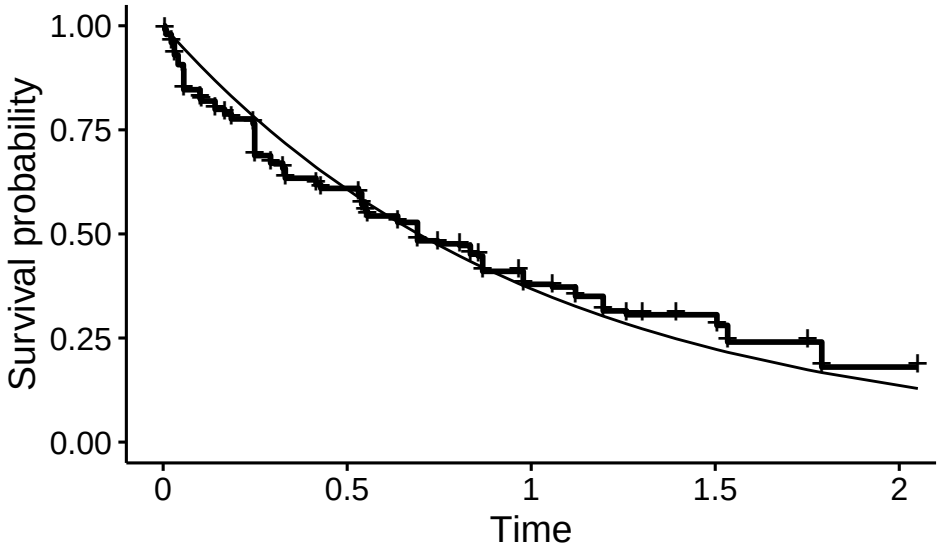
5.2.2 Modelagem dos dados

Para modelagem deste banco de dados, será realizada uma adaptação mais fraca do método *forward* de seleção de variáveis. O método começa com o modelo apenas com uma variável, desta forma, são gerados 6 modelos e é mantido o modelo que teve a variável mais significativa, após isso, é repetido o processo para as variáveis restantes, é escolhida a variável mais significativa e novamente repete-se o processo. O método para quando nenhuma das variáveis a serem adicionadas é significativa.

Seguindo essa ideia, após realizar o modelo para todas as variáveis uma a uma, foi escolhida a variável ingresso, após isso, foi realizado o modelo novamente e a variável mais significativa se mostrou ser a variável sexo. Ao tentar a inclusão de outra variável, nenhuma se mostrou significativa e por isso não foi mais adicionada nenhuma variável. Desta forma o modelo escolhido como final é:

	Parâmetro	P.valor
Intercepto	2.08	0.00
Sexo	0.27	0.00
Ingresso	-0.84	0.00

Tabela 3: Parâmetros do modelo completo



Como é possível notar pelo gráfico, a curva da distribuição exponencial é muito próxima à curva de sobrevivência do resíduo do modelo e por isso, é possível dizer que o modelo aparenta ter boa qualidade de ajuste. Com isso, o modelo final para esse estudo é o modelo que explica o tempo segundo as variáveis Sexo e Forma de Ingresso.

A interpretação da variável Sexo se dá por $e^{0.2713798} = 1.3117732$, isso significa que o tempo mediano dos alunos de sexo feminino é 1.311 vezes maior que o tempo mediano referente ao sexo masculino. Enquanto a variável indicadora de forma de ingresso pode ser interpretada por $e^{-0.8363476} = 0.4332902$ que significa que o tempo mediano das pessoas que ingressaram pelo ENEM é pouco maior que 2 vezes o tempo mediano dos ingressantes pelo vestibular.

6 Considerações Finais

Na análise de sobrevivência não existem muitas distribuições discretas para se realizar a modelagem, como uma alternativa para esse problema está a discretização de distribuições contínuas, tais como a Log-Logística.

Pelos resultados apresentados no texto, a distribuição Log-Logística discretizada é uma boa alternativa no que tange a uma alternativa para tempos discretos na análise de sobrevivência. Os modelos estimados em ambos os bancos mostraram bons ajustes.

No primeiro banco, o resultado seguiu esperado na análise exploratória com as variáveis de idade e tratamento sendo significativas. O mesmo aconteceu no segundo banco com as variáveis de sexo e turno. Essas variáveis em ambos os estudos aparentavam explicar bem o tempo independente da distribuição.

7 Referências

Barreto, M. L., L. M. P. Santos, A. M. O. Assis, M. P. N. Araújo, G. G. Franzena, P. A. B. Santos, e R. L. Fiaccone. 1994. "Effect of vitamin A supplementation on diarrhoea acute lower-respiratory-tract infection in young children in Brazil". *Lancet 344*, 228–31.

Colosimo, A., E, e S. R. Giolo. 2006. *Análise de Sobrevivência Aplicada*. São Paulo: Editora Bucher.

Santos, D. F. dos. 2017. "Modelo de Regressão Log-Logístico discreto com fracão de cura para dados de sobrevivência". Dissertação de mestrado, Brasília: Universidade de Brasília. Departamento de Estatística - Instituto de Ciências Exatas.