Automatização na análise do tempo de sobrevivência em processos públicos

Aluno: Fernando Cesar Moreira Valle

Orientador: Prof. Dr Eduardo Monteiro de Castro Gomes



Universidade de Brasília Instituto de Ciências Exatas - IE Departamento de Estatística - EST

13 de julho de 2020

Sumário

- 1 Introdução e Justificativa
- Objetivos
 - Objetivo Geral
 - Objetivos Específicos
- Metodologia
 - Função de sobrevivência
 - Taxa de risco
 - Função taxa de falha acumulada
 - Tempo médio de vida
 - Vida média residual
 - Modelo de Regressão de Cox
 - Análise de Adequação ao modelo ajustado
- 4 Análise dos dados
 - Análise exploratória
 - Dashboard
- Cronograma
- 6 Referências

Sumário

- 1 Introdução e Justificativa
- Objetivos
 - Objetivo Geral
 - Objetivos Específicos
- Metodologia
 - Função de sobrevivência
 - Taxa de risco
 - Função taxa de falha acumulada
 - Tempo médio de vida
 - Vida média residual
 - Modelo de Regressão de Cox
 - Análise de Adequação ao modelo ajustado
- Análise dos dados
 - Análise exploratória
 - Dashboard
- Cronograma
- 6 Referências

Introdução

- Querendo sempre melhorar a qualidade do atendimento ao público, a 5^a Vara da Justiça Federal buscou incentivar a redução no tempo de processamento das ações judiciais, mas conjuntamente considerando o desafio de lidar diariamente com grandes volumes de demandas nos diversos setores internos.
- Preocupada com os prazos limites para resolução dos pleitos, a instituição buscou criar um conjunto de visualizações em Dashboard por meio do software R que fossem capazes de informar os servidores e advogados responsáveis sobre as análises documentais do tempo de vida médio já existente dos processos conjuntamente com suas análises descritivas.

Justificativa

- Com o propósito de informar a instituição sobre o tempo sobressalente ou faltante das atividades processuais, delimitou-se um conjunto de procedimentos visuais e analíticos que possuem como intuito evitar a quebra dos prazos limites designados as ações.
- Para usufruir de modelos em análise de sobrevivência no banco de dados, conciliou-se um grupo de censuras à direita (pleitos que não tiveram seu status de processamento encerrado até o dia 02/12/2019) com a intenção de gerar um conjunto de gráficos que evidenciem a probabilidade de conclusão dos pleitos.

Sumário

- Introdução e Justificativa
- Objetivos
 - Objetivo Geral
 - Objetivos Específicos
- Metodologia
 - Função de sobrevivência
 - Taxa de risco
 - Função taxa de falha acumulada
 - Tempo médio de vida
 - Vida média residual
 - Modelo de Regressão de Cox
 - Análise de Adequação ao modelo ajustado
- Análise dos dados
 - Análise exploratória
 - Dashboard
- Cronogram:
- 6 Referências

Objetivos

Objetivo Geral

Criar um sistema de Dashboard, por meio do software R, que contenha gráficos e tabelas que sejam capazes de atualizar e informar servidores e advogados sobre os prazos remanescentes para elaboração e estruturação dos processos descritos como de interesse. Dessa maneira, com auxilio do pacote *Shiny* e *ShinyDashboard* criou-se uma página web referente a instituição da 5ª vara da justiça federal com um layout bootstrap de fácil compreensão e interatividade para o usuário final.

Objetivos

Objetivos Específicos

- Estruturar um breve manual de uso afim de inteirar novos usuários sobre as características descritivas de cada processo analisado, conjuntamente evidenciando os prazos delimitados para as conclusões dos pleitos;
- Evidenciar medidas de como prosseguir com a utilização do software e dos códigos estruturados, disponibilizando documentação e concedendo arquivos via Github;

Objetivos

Objetivos Específicos

- Realizar um conjunto de visualizações simples e diretas que sejam capazes de informar o usuário a cerca dos tempo restante ou tardios para conclusão dos processos;
- Gerar gráficos que mostrem o desenvolvimento dos pleitos dentro da instituição de maneira interativa e eficiente;
- Realizar um estudo descritivo e analítico na área de análise de sobrevivência utilizando o banco de dados da Justiça Federal.

Sumário

- Introdução e Justificativa
- Objetivos
 - Objetivo Geral
 - Objetivos Específicos
- Metodologia
 - Função de sobrevivência
 - Taxa de risco
 - Função taxa de falha acumulada
 - Tempo médio de vida
 - Vida média residual
 - Modelo de Regressão de Cox
 - Análise de Adequação ao modelo ajustado
- Análise dos dados
 - Análise exploratória
 - Dashboard
- CronogramaReferências

- Os dados obtidos por meio da parceria com a 5ª vara da justiça federal correspondem a 55 (cinquenta e cinco) diferentes tipos de classes processuais com diferentes tempos de circulação entre etapas, sendo estas administradas nos seis setores internos ao órgão (Secretaria, Gabinete, Central de mandados, Requerido, Requerente e Perito);
- Inicialmente os dados necessitaram de limpeza e manipulação para posteriormente realizar-se as delimitações de censuras à direita em relação aos pleitos que não obtiveram seu status de processamento encerrado até o dia 02/12/2019;

Após realização de censura nos dados, considerou-se as análises referentes a função de sobrevivência, taxa de risco, função taxa de falha acumulada, tempo médio de vida e vida média residual. Além disso, também levou-se em consideração os modelos de regressão de Cox e as análises referentes a adequação de modelos, para assim obter as medidas descritivas de interesse com seus respectivos intervalos de confiança.

Função de sobrevivência

- Esta é uma das principais funções probabilísticas usadas para descrever estudos na área de análise de sobrevivência S(t);
- É estimada com base nos dados obtidos em amostras e calculada com base no número de observações que não falharam ou sofreram o evento de interesse até o período de tempo t;

$$S(t) = P(T \ge t) = \int_{t}^{\infty} f(x) dx$$

- Pode ser definido como a probabilidade de um objeto em estudo não falhar até um determinado período de tempo $t_{(j)}$, ou seja, a probabilidade da observação analisada não ser censurada até o tempo $t_{(j)}$;
- Uma das técnicas amplamente utilizadas na estimação de $S_{(t)}$ é o Estimador de Kaplan-Meier, sendo este um método não paramétrico de estimação;

Estimador de Kaplan-Meier

• O estimador não paramétrico de Kaplan-Meier proposto por Kaplan e Meier em 1958 se defini como a medida mais popular para aferição da Função de Sobrevivência, estruturando assim, sua estimativa de $\hat{S}_{KM}(t)$ como sendo:

$$\hat{\mathcal{S}}_{\mathcal{KM}}(t) = \prod_{j:t_(j) \leq t} \left(1 - rac{d_j}{n_j}
ight)$$

• Devido ao fato de $\hat{S}_{KM}(t)$ está suscetível a variações amostrais presentes em sua variância, utilizou-se do intervalo de confiança Log descrito abaixo:

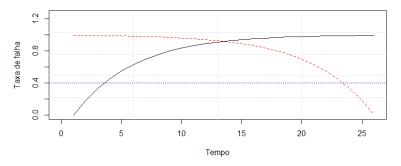
$$\left[e^{-Z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{KM(t)})}}\hat{S}_{KM}(t);e^{Z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{KM(t)})}}\hat{S}_{KM}(t)\right]$$

Taxa de risco

• É a probabilidade do objeto em estudo falhar entre os intervalos de tempo $[t_1, t_2)$ na função de sobrevivência sendo que este não falhou num período de tempo anterior a t_1 , dividida pela probabilidade da função de sobrevivência em t_1 vezes o comprimento do intervalo. Assim, a taxa de falha é expressa por:

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}$$

A taxa de risco de um objeto pode vir a assumir três formulações diferentes entre si, sendo estas: Função crescente (indicando um aumento na taxa de falha ao longo do tempo), Função constante (evidencia que a taxa não se altera com o passar do tempo) e Função decrescente (demonstra uma redução da taxa de falha à medida que o tempo passa).



(a) Crescente - Preto; Constante - Azul e Decrescente - Vermelha

Função taxa de falha acumulada

• É um procedimento útil na estimação não-paramétrica e na seleção de modelos mais apropriados para ajustar um determinado conjunto de dados. A função H(t) fornece o risco acumulado do indivíduo no tempo t. Caso esta seja definida como uma variável aleatória contínua, é descrita por:

$$H(t) = \int_0^t \lambda(u) du, \ t \ge 0$$

• Uma das técnicas amplamente usufruídas na estimação de H(t) é o estimador de Kaplan-Meie, sendo este um método não paramétricos de estimação.

Estimador de Kaplan-Meier

• Considerando-se que a função de sobrevivência de uma variável contínua pode vir a ser expressa em relação a função taxa de falha acumulada por meio da equação: S(t) = exp - H(t). Considera-se que o estimador $\hat{H}_{KM}(t)$ para a função de risco acumulado pode ser obtida por meio de:

$$\hat{H}_{\mathit{KM}}(t) = \mathit{log}\left[\hat{S}_{\mathit{km}}(t)
ight]$$

Com seu respectivo intervalo de confiança Log:

$$\left[\hat{H}_{KM}(t) - Z_{1-(\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{KM}(t))}; \hat{H}_{KM}(t) + Z_{1-(\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{KM}(t))}\right]$$

Tempo médio de vida

• É a representação da área gerada abaixo do gráfico da Função de sobrevivência, ou seja, o tempo médio que o objeto sobrevive sem presenciar a censura.

$$E(T^r) = \sum_{j:t(j) \le t}^{\infty} t^r p(t)$$
, para todo $r \ge 0$

Sendo p(t) obtido por meio da relação p(t) = S(t-1) - S(t) se t for maior que zero.

Vida média residual

• É a representação do tempo médio de vida restante para os elementos de interesse no estudo. A vida média residual no tempo t é a área sob a curva de sobrevivência à direita do ponto t, dividido pelo valor da Função de sobrevivência neste ponto mais o valor da Distribuição de probabilidade também nesse ponto .

$$V(t) = E[T - t | T \ge t] = \frac{1}{p(t) + S(t)} \sum_{k=t}^{\infty} S(k), \forall t = 0, 1, 2, ...$$

Considerando que p(t) é obtido por meio da relação p(t) = S(t-1) - S(t) se t for maior que zero.

Modelo de Regressão de Cox

- O modelo de regressão de Cox, proposto por Cox em 1972, é sem dúvida um dos mais populares na análise de sobrevivência.
- Possibilita que a análise dos tempos de vida até a ocorrência da censura seja realizada considerando-se as covariáveis de interesse no estudo.
- Cox, assim como outros autores, proprõe a modelagem dos dados de sobrevivência, na presença de covariáveis, por meio da função de risco.

$$\alpha_i(t|x_i) = \alpha_0(t) \exp\{\beta' x_i\}$$

- O modelo de Cox, definido acima, é conhecido como sendo semiparamétrico por considerar que as covariáveis agem multiplicativamente no risco pela relação $g(x,\beta) = exp\{\beta'x_i\}$ e por acatar $\alpha_0(t)$ arbitrário, ou seja, por não assumir nenhuma estrutura paramétrica em relação à $\alpha_0(t)$.
- Os pré-requisitos básicos para o uso dos riscos proporcionais de Cox são, portanto, que as taxas de falhas atuem de maneira proporcionais.

Suposição de riscos proporcionais no modelo de Cox

- O modelo de Cox é utilizado em situações que a suposição de riscos proporcionais é legitima, isto é, para situações em que as linhas das funções de riscos se cruzam. Para verificar tais suposições utilizam-se os passos:
 - Realizar a divisão dos dados em j estratos distintos segundo as j categorias de alguma covariável em interesse.
 - ② Estimar $\hat{H}_{0j}(t_i)$ para cada estrato j adquirindo as curvas de $log \ \hat{H}_{0j}(t)$ contra t, ou log(t).

Se as hipóteses forem válidas, as curvas de $log \hat{H}_{0j}(t)$ contra t, ou log(t), devem possuir diferenças constantes no tempo, ou seja, devem ser aproximadamente paralelas.

Análise de Adequação ao modelo ajustado

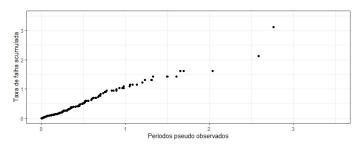
Resíduos de Cox-Snell

- Diferente da análise de resíduos efetuada em regressão linear, em análise de sobrevivência, não pode-se simplemente analisar os gráficos de resíduos devido a presença de censuras e ao fato dos próprios resíduos não seguirem uma distribuição normal.
- Para analisar a qualidade do ajuste efetuado ao modelo, em 1968, foi criado os resíduos de Cox-Snell.

$$e_i = \hat{H}_0(t_i) exp\Big(\sum_{k=1}^p x_{ip}\hat{\beta}_k\Big)$$

Resíduos de Cox-Snell

• Se o modelo estiver bem ajustado, os e_i 's podem ser visualizados como uma amostra censurada de uma distribuição exponencial padrão e, então, o gráfico de, por exemplo, $\hat{H}(e_i)$ contra e_i necessitaria ser algo próximo de uma reta.



(b) Resíduos de Cox-Snell

13 de julho de 2020

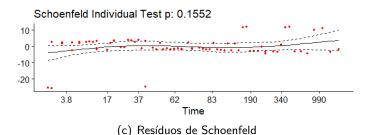
Resíduos de Schoenfeld

- O resíduo de Schoenfeld é a diferença entre os valores observados de covariáveis em um elemento contido no tempo (ti) e os valores esperados desde elemento também no período de tempo (ti) dado o grupo de risco $R_{(ti)}$.
- Um vetor de resíduos de Schoenfeld é adquirido em cada tempo observado de falha. Assim, se o elemento i é verificado falhar, o correspondente resíduo é obtido por meio de:

$$r_i = x_i - \frac{\sum_{j \in R_{(ti)}} x_j e^{\hat{\beta}_{xj}}}{\sum_{j \in R_{(ti)}} e^{\hat{\beta}_{xj}}}$$

Resíduos de Schoenfeld

• Considerando-se o plot de resíduos padronizados de Schoenfeld contra o tempo é verificável a ocorrência ou não de proporcionalidade, ou seja, se as suposições de riscos proporcionais forem satisfeitas não deverá existir nenhuma propensão sequencial no gráfico (Ho: p=0).



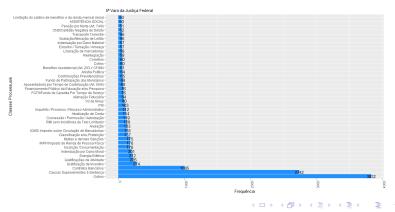
Sumário

- - Objetivo Geral
 - Objetivos Específicos
- - Função de sobrevivência
 - Taxa de risco
 - Função taxa de falha acumulada
 - Tempo médio de vida
 - Vida média residual
 - Modelo de Regressão de Cox
 - Análise de Adequação ao modelo ajustado
- Análise dos dados
 - Análise exploratória
 - Dashboard

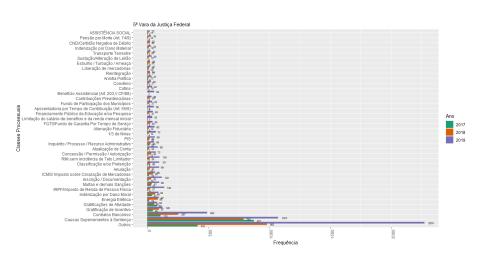
Análise exploratória

- O banco de dados consta inicialmente com 220.363 (duzentos e vinte mil trezentos e sessenta e três) observações separadas em duas classificações distintas, sendo estas Processual e Pje.
- Foram selecionados apenas os processos com classificações Pje, estes que contabilizam 78.260 (setenta e oito mil duzentos e sessenta) termos plausíveis de uso para averiguação das medidas descritivas em interesse.

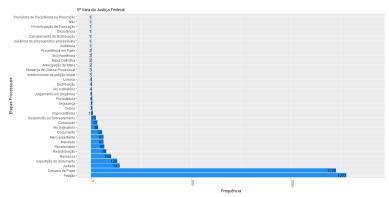
 Inicialmente plotou-se as frequências das classes processuais para saber como funcionaram as demandas da instituição entre os anos de 2014 a 2019.



- Observações com contagens inferiores a 50 unidades sofreram ajustes de nomenclatura para "Outros" afim de reduzir a poluição gráfica e otimizar a compreensão dos resultados propostos.
- Dentre as classes processuais que obtiveram maior demanda durante os oito anos de análise, destacam-se: Causas supervenientes à sentença (2.742 24.2%), Contratos bancários (1.005 8.9%) e Gratificação de incentivo (274 2.4%).
- Após análisar a frequência dos pleitos no gráfico acima, observouse que a maior incidência na classificação Pje ocorria entre os anos de 2017 a 2019.

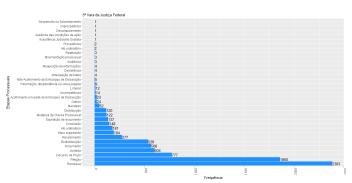


 Em relação a frequência de dados censurados existentes em cada uma das etapas, foi realizado um estudo com o objetivo de evidenciar o número de vezes que estes estágios não foram concluídos dentre as diversas classes processuais analisadas.

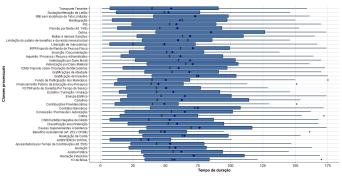


- Pode-se dizer que as etapas de "Decurso de prazo" (1277 observações) e "Petição" (1226 termos) sozinhas equivalem a 2.503 (duas mil quinhentas e três) frequências de estágios com status ainda em aberto até o dia de coleta das informações, ou seja, são essas as duas principais etapas onde a maior parte dos processos da instituição se encontram estagnados.
- Levando-se em consideração apenas as classes de pleitos que possuiram uma maior influência entre os processos nos anos de 2014 a 2019, observa-se que os dois termos de maior relevância em ordem crescente apresentam respectivamente os valores percentuais de 36.5% e 38% em relação ao banco como um todo.

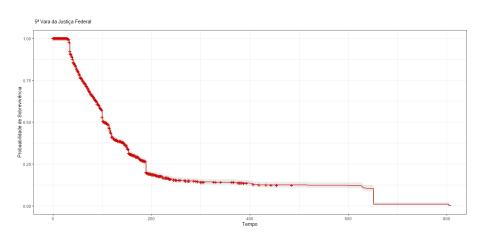
 Já em relação as etapas não censuradas, destaca-se uma maior ocorrência em ordem crescente nas etapas "Petição" e "Remessa" com respectivamente 1.860 e 2.383 casos distintos, ou seja, juntas contabilizam 4.243 observações ou 52.52% do valor total referenciado pelo sistema.



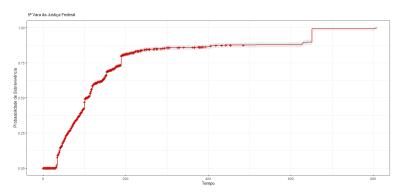
 Criou-se o Box-plot abaixo com o intuito de salientar as posições aproximadas do primeiro quartil, média, mediana e terceiro quartil de cada uma das variáveis presentes nas classes processuais em relação ao tempo.



- A função de sobrevivência serve para comparar os tempos de falha segundo variáveis qualitativas. No exemplo abaixo, pode-se comparar o comportamento do tempo até o arquivamento do processo de acordo com outras variáveis de interesse, como classe processual e etapa processual
- Observa-se que a função se inicia em um determinado momento no tempo, com 100% dos processos ainda com status em aberto, nos permitindo calcular qual a percentagem desses processos permanecem em aberto em relação a outros momentos ao longo do tempo, ou seja, serve para evidenciar o percentual de chance dos pleitos serem concluídos antes de seu arquivamento.

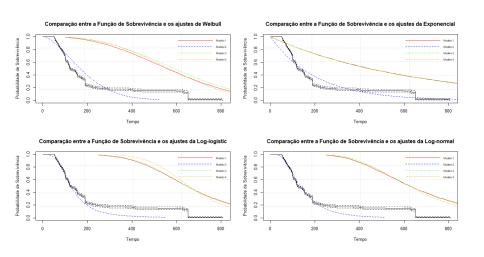


 Já a função da taxa de risco acumulada mostra a proporção de processos no todo que inicialmente encontram-se com status em aberto ou que não sofreram ainda o evento de interesse.

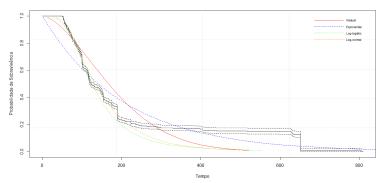


Utilizando-se do ajuste **AFT** por meio do comando Survreg, realizou-se a adequação do modelo de regressão a 4 (quatro) distribuições distintas de interesse (Weibull, Exponencial, Log-logistica e Log-normal) representados no gráfico abaixo. Dessa maneira, considerou-se a abrangência da aplicabildiade dos modelos acima citados para elaboração dos 4 (quatro) modelos paramétricos descritos:

- Modelo 1 apresenta em seus intereceptos apenas as relações entre as etapas processuais.
- Modelo 2 possui em seus intereceptos apenas as relações entre as classes processuais.
- **Modelo 3** dispõe em seus intereceptos uma relação de combinação por meio da soma entre as categorias de classes processuais e etapas processuais.
- Modelo 4 retém em seus intereceptos uma relação de combinação por meio da multiplicação entre as categorias de classes processuais e etapas processuais.



 Com base nos resultados obtidos selecionou-se o modelo 2 como melhor opção de ajuste aos dados apresentados, já que este possui uma maior aproximação de enquadramento em relação aos valores evidenciados pela função de sobrevivência em todas as distribuições.



Dashboard

Tela referente a identificação de usuário



Tela referente a análise processual





Tela referente ao manual de uso

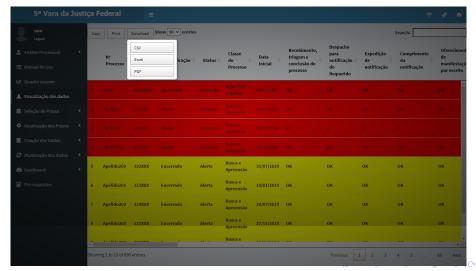




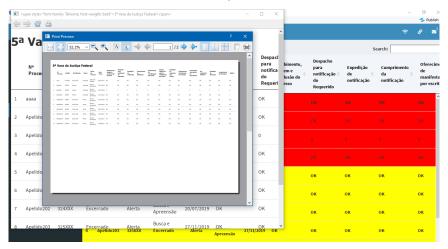
Tela referente ao quadro Resumo



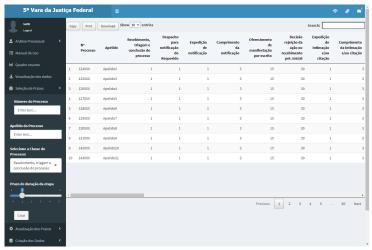
Tela referente a Visualização de dados



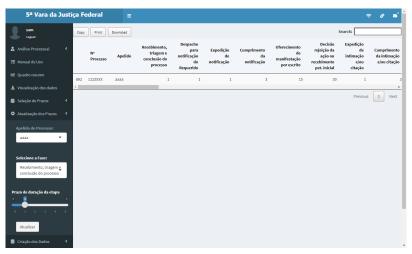
Tela referente a Impressão dos dados



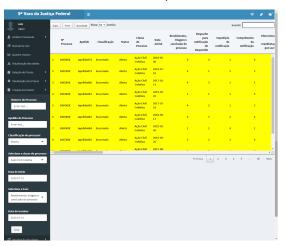
Tela referente a Seleção de prazos



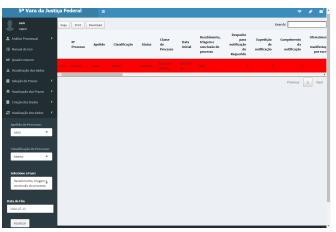
Tela referente a Atualização dos prazos



Tela referente a Criação de dados



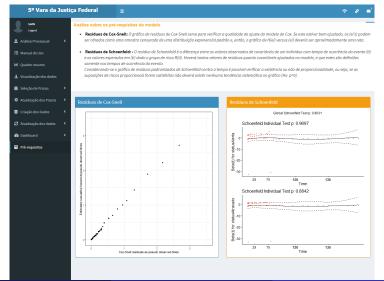
Tela referente a Atualização de dados



Tela referente ao Dashboard



Tela referente aos Pré-requisitos do modelo



Sumário

- Introdução e Justificativa
- Objetivos
 - Objetivo Geral
 - Objetivos Específicos
- Metodologia
 - Função de sobrevivência
 - Taxa de risco
 - Função taxa de falha acumulada
 - Tempo médio de vida
 - Vida média residual
 - Modelo de Regressão de Cox
 - Análise de Adequação ao modelo ajustado
- Análise dos dados
 - Análise exploratória
 - Dashboard
- 6 Cronograma
- 6 Referências



Cronograma

Cronograma - 2/2019

Atividades	Jul	Ago	Set	Out	Nov	Dez
Escolha do tema a ser abordado						
Estudo do tema						
Estudo de técnica						
Elaboração de relatório parcial						
Entrega e correção do relatório parcial						

Cronograma

Cronograma - 1/2020

Atividades	Fev	Mar	Abr	Maio	Jun	Jul	Ago	Set
Desenvolvimento dos modelos em estudo								
Análise dos Resultados								
Elaboração do relatório Final								
Entrega do relatório final ao prof. orientador								
Correção do relatório final								
Apresentação do relatório final								
Entrega do relatório final a banca								

Sumário

- Introdução e Justificativa
- Objetivos
 - Objetivo Geral
 - Objetivos Específicos
- Metodologia
 - Função de sobrevivência
 - Taxa de risco
 - Função taxa de falha acumulada
 - Tempo médio de vida
 - Vida média residual
 - Modelo de Regressão de Cox
 - Análise de Adequação ao modelo ajustado
- Análise dos dados
 - Análise exploratória
 - Dashboard
- Cronograma
- 6 Referências



Referências

Colonismo, E.A.; Giolo, S.R. Análise de Sobrevivência Aplicada. São Paulo: Edgard Blucher, ano 2006.

J.F. Lawless.

Estatistical Models and Methods for Lifetime Data.

John Wiley Sons, New York, and 1982.

Poder Judiciário Justiça Federal.

Corregedoria-Regional da Justiça Federal da 2ª Região. Manual de Rotinas e Procedimentos Internos. Brasília.

DF, ano 2009.

Cox, D.R.

Regression models and life tables.

Journal of Royal statistical society. Series V.39, P.1-38, 1972.

Referências

Cox, D.R.

Partial likehood..

Biometrika, V.62, N.2, P.269-276, MAR. 1975.

🍆 Giolo, S. R

Modelos de análise de sobrevivência para experimentos dose-resposta..

Campinas: Disseratação de Mestrado, 1994.

Giolo, S. R

Variáveis latentes em análise de sobrevivência e curvas de crescimento..

Piracicaba: Tese de Doutorado, 2003.

🍆 Hougaard, P.

Analysis of multivariate survival data...

New York:Springer Verlag, 2000.

Referências

陯 Kaplan, E. L.; Meier, P.

Non-parametric estimation from incomplete observations.. Jour-non-parametric statistical association, V.53,P.547-481, 1958.

Giolo, S. R.

Modelos de Riscos Proporcionais.

Paraná: Universidade Federal do Paraná, 2018.

🍆 Colonismo, E. A.; Giolo, S. R.

Análise de Sobrevivência Aplicada.

Edgar Blucher, 2006.

Nakano, E. Y.

Um curso de Análise de sobrevivência.

Brasília: Universidade de Brasília, 2018.