Universidade Federal do Paraná Setor de Ciências Exatas Departamento de Estatística Programa de Especialização em *Data Science* e *Big Data*

Dalton Luiz Schneckenberg Filho

Modelo de Regressão de Cox

Curitiba 2020

Dalton Luiz Schneckenberg Filho

Modelo de Regressão de Cox

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. José Luiz Padilha

Modelo de Regressão de Cox - Uma Aplicação a Dados de Pacientes com a Doença de Chagas

Cox Regression Model - An Application to Data from Patients with Chagas Disease

Dalton Luiz S. Filho¹

¹Departamento de Estatística, Universidade Federal do Paraná Av. Cel. Francisco H. dos Santos, 100 - Jardim das Américas, Curitiba - PR, 81530-000, Brasil*

A análise de sobrevência é uma das áreas da estátistica que mais cresceu nas últimas duas décadas. A razão deste crescimento é o desenvolvimento e aprimoramento de técnicas estatísticas combinadas com computadores cada vez mais velozes. Este trabalho tem por objetivo utilizar conceitos do modelo semi-paramétrico de sobrevivência, em especial o modelo de Cox, para modelar o tempo de vida de doentes portadores da doença de chagas. O modelo de Cox estima a função de risco e muitas vezes é interessante estimar o efeito das covariáveis. Desta forma Cox propôs o modelo semi-paramétrico de riscos proporcionais que permite a análise de dados resultantes de estudos do tempo de vida, onde a variável resposta é o tempo de ocorrência de um evento de interesse ajustado por covariáveis.

Palavras-chave: Modelos semi-paramétricos. Modelo de Cox. Análise de sobrevivência

Survival analysis is one of the fastest growing areas of statistics in the past two decades. The reason for this growth is the development and improvement of statistical techniques combined with increasingly fast computers. This work aims to use concepts of the semi-parametric model of survival, especially the Cox model, to model the life span of patients with Chagas disease. Cox's model estimates the risk function and it is often interesting to estimate the effect of covariates. In this way, Cox proposed the semi-parametric model of proportional risks that allows the analysis of data resulting from life time studies, where the response variable is the time of occurrence of an event of interest adjusted by covariates.

Keywords: Semi-parametric models. Cox model. Survival analysis

1. Introdução

A doença de Chagas (ou Tripanossomíase americana) é a infecção causada pelo protozoário Trypanosoma cruzi. Apresenta uma fase aguda (doença de Chagas aguda – DCA) que pode ser sintomática ou não, e uma fase crônica, que pode se manifestar nas formas indeterminada, cardíaca, digestiva ou cardiodigestiva. Em função das ações de controle de vetores realizadas a partir da década de 1970, o Brasil recebeu em 2006 a certificação Internacional da interrupção da transmissão vetorial pelo Triatoma infestans, espécie exótica e responsável pela maior parte da transmissão vetorial no passado. Porém, estima-se que existam aproximadamente 12 milhões de portadores da doença crônica nas Américas, e que haja no Brasil, atualmente, pelo menos um milhão de pessoas infectadas por T. cruzi.

De 2008 a 2017 foram registrados no Brasil um total de 2172 casos agudos da doença de chagas, sendo a região norte (Pará) onde foi verificada a maior incidência. A taxa média de mortalidade registrada no Brasil de 2015 a 2017 foi de 2,19 para cada 100.000 habitantes.

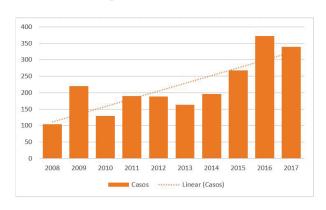


Figura 1: Casos da doença de chagas registrados no Brasil de 2008 a 2017.

^{*}daltonluiz26@gmail.com

2. Conceituação e definições

2.1. Publico alvo

Os dados utilizados no estudo tem como origem a base de informações do acompanhamento de doentes de chagas coletada no período entre os anos de 1999 e 2019 no Hospital das Clínicas da Universidade Federal de Minas Gerais.

2.2. Objetivo

O problema que pretendemos estudar aqui é a influência das variáveis de controle no risco de morte de pacientes com a doença de chagas.

3. Metodologia

3.1. Tratamento dos dados

Iniciou-se o trabalho no banco de dados com 619 linhas, 112 colunas e 272 óbitos Após a análise descritiva e tratamento dos dados restaram 615 linhas e 28 colunas, informações mais relevantes ao estudo proposto.

No conjuto de dados foram observados 355 pacientes do sexo masculino e 260 pacientes do sexo feminino apresentando a relação de 60/40 respectivamente Neste conjunto a idade média observada foi de 51 anos correspondente a 14 por cento dos pacientes, a idade minima e máxima encontrada foram 15 e 83 respectivamente.

3.2. Imputação de Dados Faltantes

Verificou-se que o conjunto de dados trazia inicialmente um grande número de dados faltantes.

A partir deste ponto o ajuste de modelos contendo todas as variáveis torna-se inviável caso não haja um tratamento nos dados ausentes uma vez que na maioria das técnicas de modelagem não são ajustados caso uma linha tenha pelo menos um dado ausente.

A exclusão dessas linhas não é viável e pode viesar o modelo, pois observações que apresentam valores ausentes podem ter uma relação distinta com a variável resposta. Neste caso optamos por preencher os valores ausentes através do método de imputação implementado pelo pacote MICE.

A função mice utiliza o procedimento chamado 'full conditional specification', no qual para cada variável com valores ausentes é construído um modelo de regressão condicional em todas as demais variáveis no

banco de dados. Foi aplicada a imputação simples utilizando as configurações padrões da função.

4. Modelagem do Problema

O modelo de regressão de cox permite a análise de dados provenientes de estudos de tempo de vida em que a resposta é o tempo até a ocorrência de um evento de interesse, ajustando por covariáveis.

O modelo de regressão de cox é bastante flexível devido à presença do componente não paramétrico, o que permite a análise de tempos de vida sem a necessidade de especificarmos uma distribuição de probabilidade para a variável resposta.

De forma genérica, condidere p covariáveis, de modo que x seja um vetor com os componentes x = (x1,...,xp)'. A expressão geral do modelo de regressão de cox considera:

$$\lambda(t|x) = \lambda_0(t) \ exp(x'\beta) \tag{1}$$

4.1. Ajuste do modelo

Foram ajustados os modelos contendo uma única variável e incluídas as variáveis significativas ao nivel de 0,10.

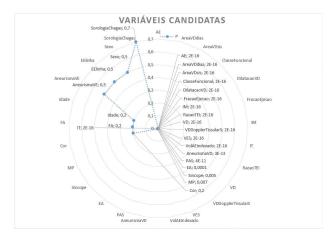


Figura 2: Conjunto de variáveis com pvalor < 0,10.

No passo dois as variáveis significativas do passo um foram ajustadas conjuntamente excluindo-se a cada iteração uma única variável, chegando assim ao conjunto final.

Evoluimos então para a análise dos resultados, foram selecionadas duas variáveis para avaliação.

A variável Classe funcional foi citada pela pesquisadora que forneceu os dados do estudo como importande do ponto de vista clinico, a variável está relacionada à capacidade funcional do coração (NYHA -

Dalton Luiz S. Filho 01-5

New York Heart Association) trata-se da classificação do nível de severidade do problema cardiaco do paciente, podendo variar do nível 1(doença cardiaca, sem limitações para atividade fisica) ao 4(doença cardiaca que acarreta em incapacidade para exercer qualquer atividade fisica).

Os resultados corroboraram a citação da pesquisadora.

Para os níveis 2, 3 e 4 verificamos o risco crescente em 24, 37 e 66% respectivamente em relação à referência, resultados realmente expressivos.

A segunda variável avaliada foi aneurismaVD, tratase de uma variável dicotomica(sim/não) que representa a presença de aneurisma no ventrículo direito do paciente, fato que se não verificado contribui com a redução de 38% no risco de morte.

O indice de concordância apresentado na análise foi de 0.82.

A figura 3 traz a análise completa dos dados para referência.

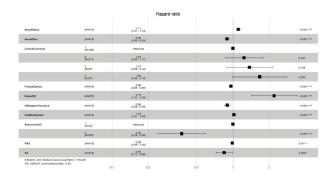


Figura 3: Análise de riscos, HR > 1 indica risco aumentado de morte

4.2. Validação do modelo

Embora o modelo tenha sido ajustado, ainda é necessária a verificação do atendimento da suposição dos riscos proporcionais para que o modelo de Cox seja considerado adequado.

De acordo com Colosimo e Giolo (2006), técnicas gráficas aplicadas aos diferentes resíduos obtidos na regressão são particularmente muito utilizadas para examinar a distribuição dos erros. Estas técnicas, como observados em Klein e Moeschberger (1997), devem ser utilizadas primariamente para rejeitar modelos inadequados e não para "provar" que um determinado modelo é correto.

Os métodos gráficos aplicados ao modelo em questão utilizaram o logaritmo da função de risco acumulada e os resíduos padronizados de Schoenfeld. Ambos devem ser construídos para cada variável do modelo.

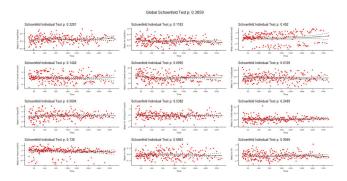


Figura 4: O Padrão aleatório verificado nos residuos Schoenfeld não rejeita o modelo

5. Conclusão

No presente trabalho foi estudado o modelo de cox e sua utilização na análise de sobrevivência para explicação dos efeitos das variáveis de controle na ocorrência do evento de interesse.

Agradecimentos

Agradeço ao Banco Bradesco pela confiança e investimento no desenvolvimento continuo dos funcionários. À minha esposa e filhos pelo apoio incondicional e compreenção nos períodos de ausência, à coordenação do Curso de Data Science Big Data da UFPR, seu corpo docente e administração que oportunizaram intenso avanço no conhecimento, e finalmente ao meu orientador José Luiz Padilha Silva, pela paciência e auxilio despendidos no desenvolvimento do presente trabalho.

Referências

- [1] Ministério da Saúde do Brasil, https://saude.gov.br/saude-de-a-z/doenca-dechagas
- [2] Análise de sobrevivência aplicada, *E*nrico Antônio Colosimo, Suely Ruiz Giolo. São Paulo: Blucher, 2006.
- [3] Multivariate Imputation by Chained Equations in R, Stef VAN BUUREN and Karin Groothuis-Oudshoorn, Journal of Statistical Software, 2011, https://www.jstatsoft.org/v45/i03/
- [4] Universidade Federal de Minas Gerais, $http://www.est.ufmg.br/enricoc/pdf/avancados_medicina/aula12$ $15_cox.pdf$