UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Eliane Ribeiro Carmes

ANÁLISE DE DADOS DE SOBREVIVÊNCIA NA PRESENÇA DE RISCOS COMPETITIVOS

Projeto de Pesquisa apresentado à disciplina Laboratório A do Curso de Graduação em Estatística da Universidade Federal do Paraná, como requisito para elaboração do Trabalho de Conclusão de Curso.

Orientadora: Profa. Dra. Suely Ruiz Giolo

CURITIBA
2014

Sumário

1 INTRODUÇÃO

A análise de sobrevivência, também denominada análise de sobrevida, é a área da Estatística que dispõe de técnicas e métodos que possibilitam a análise de dados em que há interesse no tempo de vida das unidades amostrais (indivíduos, equipamentos, etc.) na presença de censuras e de covariáveis.

Para dados dessa natureza, a variável resposta *T*, contínua e não-negativa, representa o tempo até que determinado evento de interesse ocorra, geralmente denominado tempo de falha. Por outro lado, censura é a observação parcial da resposta causada pela perda de acompanhamento por qualquer motivo que não o evento de interesse em estudo. Para os casos de censura à direita (KLEIN; MOESCHBERGER, 2003), a única informação que se tem disponível é que o tempo de falha das unidades amostrais classificadas como censuras é superior ao tempo de seguimento. Quanto às covariáveis, estas são características dos indivíduos em estudo as quais podem estar influenciando na variável resposta, acelerando ou retardando sua ocorrência (COLOSIMO; GIOLO, 2006).

As principais funções que caracterizam a distribuição da variável *T* são:

a) a função de sobrevivência, $S(t)$, que fornece a probabilidade de um indivíduo sobreviver ao tempo t ($t \geq 0$);
b) a função de distribuição, $F(t) = 1 - S(t)$, que fornece a probabilidade de um indivíduo não sobreviver ao tempo t;
c) a função taxa de falha ou de risco, $\lambda(t)$, que fornece a taxa instantânea de falha no tempo t dado que o indivíduo sobreviveu ao tempo imediatamente anterior a t e;
d) a função taxa de falha acumulada $\Lambda(t) = \int_0^t \lambda(u)du$.

Em diversas situações que envolvem dados de sobrevivência, é frequente que a causa que conduz à falha (isto é, ao evento de interesse) dos indivíduos em estudo seja uma dentre k causas possíveis ($k \geq 2$). Assim, a causa de falha registrada para cada indivíduo será aquela que ocorrer primeiro, dado que a ocorrência de uma delas impede a ocorrência de qualquer outra.

A presença de duas ou mais causas competindo pela falha, quando a ocorrência de uma impede a ocorrência de outra, caracteriza dados de sobrevivência com estrutura de riscos competitivos, com aplicação em diversas áreas do conhecimento. Na área da

Saúde, por exemplo, indivíduos com câncer estão em risco de morte pelo câncer bem como por outras causas. Nesse caso, a probabilidade de morte por câncer dependerá da taxa de mortalidade associada ao câncer e da taxa de mortalidade associada às demais causas. Já no campo das Ciências Políticas, a probabilidade de um político em cargo eletivo não se candidatar à reeleição dependerá da taxa de não candidatura associada a causas tais como: aposentadoria, interesses do partido político, interesses pessoais, dentre outras.

Na presença de riscos competitivos tem-se que:

a) os indivíduos estão em risco de falha por *k* causas diferentes;
b) a ocorrência de uma das *k* causas impede a ocorrência de qualquer outra.

Para dados de sobrevivência na presença de riscos competitivos é frequente o interesse na distribuição do tempo até a falha para uma causa específica *k (k = 1, ..., K)* na presença de todas as outras causas. Desse modo, diversas abordagens de análise para dados dessa natureza encontram-se propostas na literatura.

A abordagem clássica, na presença de covariáveis, consiste em modelar a função taxa de falha causa-específica para as diferentes causas de falha sob a suposição de riscos proporcionais (LARSON, 1984; PRENTICE et al., 1978). Entretanto, alguns pesquisadores (GRAY, 1988; PEPE, 1991) observaram que, para uma particular causa de falha, uma determinada covariável pode apresentar efeitos diferentes sobre a função taxa de falha causa-específica e a sua correspondente função de incidência acumulada (FIA), também denominada subdistribuição e definida para a causa k por $F_k(t|\boldsymbol{x}) = P(T \leq t, \varepsilon = k|\boldsymbol{x})$. Desse modo, concluíram ser impossível testar, sob a formulação de funções taxa de falha causa-específica, o efeito de covariáveis sobre a função de incidência acumulada (FIA).

Essa limitação da abordagem clássica motivou esforços no sentido de se modelar diretamente as funções de incidência acumulada. Desses esforços resultou o modelo de regressão de Fine-Gray (FINE; GRAY, 1999) que, devido às suas similaridades com o modelo de regressão de Cox, é muito flexível e apresenta muitas das propriedades úteis desse modelo. A popularidade do modelo de Fine-Gray se deve, em parte, por este se encontrar implementado nos *softwares* R (pacote *cmprsk*) e Stata, e por fornecer, na prática, predições úteis e interpretações relativamente simples.

Após o modelo de Fine-Gray ter sido proposto, diversas extensões surgiram para o mesmo. Uma delas proposta por Scheike e Zhang (2008), acomoda situações em que não se faz necessário supor a proporcionalidade de riscos, o que implica ser possível

acomodar no modelo covariáveis com efeito variando no tempo (efeitos tempo-dependentes). Dentre outras extensões têm-se: os modelos que permitem a inclusão de covariáveis dependentes do tempo (BEYERSMANN; SCHUMACHER, 2008) e os que permitem a inclusão de um efeito aleatório ou termo de fragilidade (KATSAHIAN et al., 2006, SCHEIKE; SUN; ZHANG; JENSEN, 2010, KATSAHIAN; BOUDREAU, 2011, DIXON; DARLINGTON; DESMOND, 2011).

2 OBJETIVOS

2.1 Objetivo Geral

Estudar alguns modelos de regressão que, no contexto de dados de sobrevivência com riscos competitivos, modelam diretamente as funções de incidência acumulada.

2.2 Objetivos Específicos

1. Revisar a literatura no que diz respeito às abordagens de análise propostas para dados de sobrevivência no contexto de riscos competitivos;
2. Revisar em detalhes a formulação do modelo de regressão proposto por Fine e Gray (1999) e sua extensão proposta por Scheike e Zhang (2008);
3. Revisar os procedimentos de estimação e os métodos de diagnóstico e adequação propostos para os dois modelos mencionados;
4. Ajustar o modelo de Fine e Gray (FINE; GRAY, 1999) e sua extensão proposta por Scheike e Zhang (2008) a um conjunto de dados;
5. Comparar e discutir os resultados dos modelos ajustados.

3 CASUÍSTICA E MÉTODOS

3.1 Casuística

3.1.1 Conjunto de Dados

O conjunto de dados que se pretende utilizar para ajustar os dois modelos foco deste trabalho (o de Fine-Gray e sua extensão proposta por Sheike e Zhang) encontra-se disponível em http://www.uhnres.utoronto.ca/labs/hill/datasets/Pintilie/datasets/follic.txt. Este conjunto contém informações de 541 pacientes com linfoma de células foliculares (tipo I ou II) em estágio precoce, tratados com radioterapia ou radioterapia e quimioterapia. A idade dos pacientes (média = 57 anos e desvio padrão = 14) e os níveis de hemoglobina (média = 138 g/l e desvio padrão = 15) também estão disponíveis. O tempo mediano de seguimento dos pacientes foi de 5,5 anos.

No contexto de riscos competitivos, as duas causas de falha consideradas neste estudo foram: (1) morte em recidiva da doença ou ausência de resposta ao tratamento e; (2) morte em remissão. Quanto à variável resposta, esta foi definida como o tempo decorrido desde o início do tratamento até a causa de falha que ocorrer primeiro. Para os pacientes sem registro de falha (censuras), o tempo considerado foi o decorrido desde o início do tratamento até a data final de seguimento de cada um deles (PINTILIE, 2006).

Dos 541 pacientes no estudo, foram observadas 348 falhas (óbitos), sendo 272 devidas à causa 1 (24 óbitos em pacientes que não responderam ao tratamento e 248 em pacientes em recidiva de doença) e 76 à causa 2 (óbitos em pacientes livres de doença, isto é, em remissão). O Quadro 1 apresenta uma descrição geral das informações disponíveis para o conjunto de dados descrito.

3.1.2 Recursos Computacionais

O *software* R, versão 3.1.1 (R CORE TEAM, 2014) será utilizado para ajustar os modelos aos dados descritos. Alguns pacotes a serem investigados com este propósito são: *cmprsk* (GRAY, 2014), *timereg* (SCHEIKE; MARTINUSSEN, 2006; SCHEIKE; ZHANG, 2011) e *timeROC* (BLANCHE, 2013).

Quadro 1 – Descrição das variáveis disponíveis no banco de dados de pacientes com linfoma de células foliculares

VARIÁVEL	DESCRIÇÃO
Stnum	identificação do paciente
COVARIÁVEIS	
Age	idade em anos
Hgb	hemoglobina em g/l (gramas por litro)
clinstg	estágio clínico: 1 = estágio I 2 = estágio II
Ch	quimioterapia: Y = sim em branco = não
RT	radioterapia: Y = sim em branco = não
DESFECHOS	
Resp	resposta após tratamento: CR = remissão completa NR = sem resposta
Relsite	sítio da recidiva: L = local D = distante B = local e distante em branco = sem recidiva
Survtime	tempo decorrido desde o diagnóstico até o óbito ou até o último seguimento
St	situação: 1 = óbito 0 = vivo
Dftime	tempo decorrido desde o diagnóstico até a falha que ocorrer primeiro: sem resposta, recidiva ou óbito
Dfcens	censura: 1 = falha 0 = censura

Fonte: Pintilie, M. Competing risks: a practical perspective. Chichester: John Wiley & Sons, Ltd, 2006, p. 17.

3.2 Métodos

Os modelos de regressão propostos para a análise de dados de sobrevivência com riscos competitivos, foco de estudo deste trabalho, são aqueles que modelam diretamente as funções de incidência acumulada.

Um destes modelos é o modelo de Fine-Gray (FINE; GRAY, 1999), que assume que a função de incidência acumulada para a causa *k* é modelada por:

$$F_k(t|\boldsymbol{x}) = P(T \leq t, \varepsilon = k \mid \boldsymbol{x}) = 1 - \exp\{-\Lambda_0(t)\exp(\boldsymbol{x}^T\boldsymbol{\beta}_k)\}, \qquad (1)$$

em que T denota a variável tempo, ϵ indica a causa de falha, $\boldsymbol{x} = (x_1, ..., x_p)$ corresponde a um vetor de covariáveis, $\boldsymbol{\beta}_k$ a um vetor de coeficientes de regressão associados à causa k e $\Lambda_0(t)$ a uma função não especificada e não decrescente tal que $\Lambda_0(0) = 0$.

Nota-se que $F_k(t|\boldsymbol{x})$ fornece a probabilidade do indivíduo não sobreviver ao tempo *t* devido à k-ésima causa. Ainda, tem-se que a função que lineariza o modelo (1), denominada função de ligação, é a complemento log-log (cloglog), isto é,

$$\log(-\log(1 - F_k(t|\boldsymbol{x})) = \boldsymbol{x}^T\boldsymbol{\beta}_k + \log\big(\Lambda_0(t)\big).$$

Assim, denotando a função de ligação cloglog por $h(\cdot)$, tem-se o modelo de Fine-Gray expresso por:

$$\begin{aligned} h(1 - F_k(t|\boldsymbol{x})) &= \boldsymbol{x}^T\boldsymbol{\beta}_k + \log\big(\Lambda_0(t)\big) \\ &= \boldsymbol{x}^T\boldsymbol{\beta}_k + \eta\big(\Lambda_0(t)\big), \end{aligned}$$

com $\eta(\cdot)$ uma função de $\Lambda_0(t)$.

Se o principal interesse está em avaliar os efeitos das covariáveis sobre a função de incidência acumulada, outras funções de ligação conhecidas (por exemplo, a logito) podem ser consideradas a fim de tornar as interpretações mais simples. Desse modo, a função de incidência acumulada pode, de modo geral, ser expressa como uma função g de $\Lambda_0(t), \boldsymbol{\beta}_k$ e $\boldsymbol{x}$, isto é,

$$F_k(t|\boldsymbol{x}) = g(\Lambda_0(t), \boldsymbol{\beta}_k, \boldsymbol{x}),$$

de modo que:

$$h(F_k(t|\boldsymbol{x})) = \boldsymbol{x}^T\boldsymbol{\beta}_k + \eta\big(\Lambda_0(t)\big),$$

com $h(\cdot)$ a função de ligação considerada e $\eta(\cdot)$ uma função de $\Lambda_0(t)$.

Outro modelo alvo de estudo deste trabalho consiste de uma extensão do modelo de Fine-Gray proposta por Scheike e Zhang (2008) e implementada no pacote *timereg* do R por Scheike e Zhang (2011). Tal modelo é mais flexível que o modelo (1), tendo em vista que para seu uso não é necessário assumir proporcionalidade dos riscos, ou seja, pode-se acomodar nesse modelo covariáveis com efeitos tempo-dependentes. De modo geral, esse modelo é dado por:

$$h\{F_k(t \mid \boldsymbol{x}, \boldsymbol{z})\} = \boldsymbol{x}^T\boldsymbol{\beta}_k + \boldsymbol{z}^T\boldsymbol{\alpha}_k(t) + \eta\big(\Lambda_0(t)\big),$$

com $h(\cdot)$ uma função de ligação conhecida, $\boldsymbol{x}$ e $\boldsymbol{z}$ vetores de covariáveis, $\boldsymbol{\beta}_k$ coeficientes de regressão que não variam no tempo, $\boldsymbol{\alpha}_k(t)$ coeficientes de regressão que variam no tempo, ambos associados à causa k, e $\eta(\cdot)$ uma função de $\Lambda_0(t)$.

A estimação dos coeficientes de regressão associados aos modelos apresentados pode ser realizada por meio de uma abordagem baseada no ajuste direto de um modelo de regressão binomial. Essa abordagem, descrita por Scheike, Zhang e Gerds (2008), também será objeto de estudo neste trabalho, assim como os métodos de diagnóstico e adequação dos modelos (BLANCHE, 2013).

4 CRONOGRAMA DE ATIVIDADES

ATIVIDADES	OUT 2014	NOV 2014	JAN 2015	FEV 2015	MAR 2015	ABR 2015	MAI 2015	JUN 2015
1 Projeto de Pesquisa								
Definição do tema de estudo								
Definição do conjunto de dados e dos métodos estatísticos								
Elaboração e entrega do projeto de pesquisa ao orientador								
Apresentação do projeto de pesquisa								
2 Elaboração do Trabalho de Conclusão de Curso								
Revisão de literatura sobre o tema								
Análise dos dados e discussão dos resultados obtidos								
Redação do trabalho de conclusão de curso								
Leitura do trabalho pelo orientador e correções								
Entrega do trabalho redigido aos membros da banca								
3 Defesa do Trabalho de Conclusão de Curso								
Preparação e apresentação do TCC								
4 Elaboração da Versão Final do Trabalho de Conclusão de Curso								
Elaboração da versão final do TCC								
Entrega da versão final do TCC ao orientador								

REFERÊNCIAS

BEYERSMANN, J.; SCHUMACHER, M. Time-dependent covariates in the proportional subdistribution hazards model for competing risks. **Biostatistics**, Oxford, v. 9, p. 765-776, 2008.

BLANCHE, P. **timeROC: time-dependent ROC curve and AUC for censored survival data**. R package version 0.2, 2013. Disponível em: <http://CRAN.R-project.org/package=timeROC>. Acesso em: 05 nov. 2014.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. São Paulo: Editora Blucher, 2006. 392 p.

DIXON, S. N.; DARLINGTON, G. A.; DESMOND, A. F. A competing risks model for correlated data based on the subdistribution hazard, **Lifetime Data Analysis**, Boston, v. 17, p. 473-495, 2011.

FINE, J. P.; GRAY, R. J. A proportional hazards model for the subdistribution of a competing risk. **Journal of the American Statistical Association**, New York, v. 94, n. 446, p. 496-509, 1999.

GRAY, R. J. A class of K-sample tests for comparing the cumulative incidence of a competing risk. **Annals of Statistics**, San Francisco, v. 16, p. 1141-1154, 1988.

GRAY, R. J. **cmprsk: subdistribution analysis of competing risks**. R package version 2.2-7. 2014. Disponível em: <http://CRAN.R-project.org/package=cmprsk>. Acesso em: 15 out. 2014.

KATSAHIAN, S.; RESCHERIGON, M.; CHEVRET, S.; PORCHER, R. Analysing multicenter competing risks data with a mixed proportional hazards model for the subdistribution. **Statistics in Medicine**, Chichester, v. 25, p. 4267-4278, 2006.

KATSAHIAN, S.; BOUDREAU, C., Estimating and testing for center effects in competing risks, **Statistics in Medicine**, Chichester, v. 30, p. 1608-1617, 2011.

LARSON, M. G. Covariate analysis of competing risks models with log-linear models, **Biometrics**, Washington, v. 40, p. 459-469, 1984.

KLEIN, J. P.; MOESCHBERGER, M. L. **Survival analysis: techniques for censored and truncated data**. 2. ed. New York: Springer, 2003. 536 p.

PEPE, M. S. Inference for events with dependents risks in multiple endpoint studies. **Journal of the American Statistical Association**, v. 86, p. 770-778, 1991.

PRENTICE, R. L.; KALBFLEISCH, J. D.; PETERSON, A. V.; FLOURNOY, N.; FAREWELL, V. T.; BRESLOW, N. E. The analysis of failure times in the presence of competing risks. **Biometrics**, Washington, v. 34, p. 541-554, 1978.

PINTILIE, M. **Competing risks: a practical perspective**. Chichester: John Wiley & Sons, Ltd, 2006. 224 p.

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna, Austria, 2014. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org/>.

SCHEIKE, T. H.; MARTINUSSEN, T. **Dynamic regression models for survival data**. New York: Springer, 2006. 470 p.

SCHEIKE, T. H.; ZHANG, M. J. A flexible competing risks regression modeling and goodness-of-fit. **Lifetime Data Analysis**, Boston, v. 14, p. 464-483, 2008.

SCHEIKE, T. H.; ZHANG, M. J.; GERDS, T. A. Predicting cumulative incidence probability by direct binomial regression. **Biometrika**, London, v. 95, p. 205-220, 2008.

SCHEIKE, T. H.; SUN, Y.; ZHANG, M. J.; JENSEN, T. K. A semiparametric random effects model for multivariate competing risks data. **Biometrika**, London, v. 97, p. 133-145, 2010.

SCHEIKE, T. H.; ZHANG, M. J. Analyzing competing risk data using the R timereg package. **Journal of Statistical Software**, Los Angeles, v. 38, n. 2, p. 1-15, 2011.