UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

TESE DE DOUTORADO

Estimação de razão de azares por meio de regressão quantílica para dados com censura à direita: uma abordagem computacional

Marina Bessel

Orientador: Prof. Dr. Álvaro Vigo

Porto Alegre, Setembro de 2017

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

TESE DE DOUTORADO

Estimação de razão de azares por meio de regressão quantílica para dados com censura à direita: uma abordagem computacional

Marina Bessel

Orientador: Prof. Dr. Álvaro Vigo

A apresentação desta tese é exigência do Programa de Pós-graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul, para obtenção do título de Doutor.

Porto Alegre, Brasil.
2017

CIP - Catalogação na Publicação

Bessel, Marina
Estimação de razão de azares por meio de regressão quantílica para dados com censura à direita: uma abordagem computacional / Marina Bessel. -- 2017.
36 f.
Orientador: Álvaro Vigo.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Medicina, Programa de Pós-Graduação em Epidemiologia, Porto Alegre, BR-RS, 2017.

1. regressão quantilica. 2. censura. 3. análise de sobrevivência. I. Vigo, Álvaro, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

BANCA EXAMINADORA

Profª. Drª. Suzi Camey, Programa de Pós-Graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul - UFRGS.

Prof. Dr. Marcio Valk, Departamento de Estatística, Instituto de Matemática e Estatística, Universidade Federal do Rio Grande do Sul - UFRGS.

Dr. Maicon Falavigna, Hospital Moinhos de Vento.

“Amo o Senhor, porque ele ouve a minha voz e as minhas súplicas. Porque ele inclinou para mim os Seus ouvidos, invocá-lo-ei enquanto eu viver. Volta, a minha alma, ao teu sossego, pois o Senhor tem sido generoso para comigo. Andarei na presença do Senhor na terra dos viventes.”

(Salmo 16:1-9)

AGRADECIMENTOS

À Deus, por estar sempre comigo em todos os momentos mas, principalmente, nos que eu mais preciso.

Ao meu marido Zenir, pelo carinho, paciência, apoio e incentivo. É o responsável por tudo o que sou, e por todas as minhas conquistas.

Às minhas irmãs Magali e Maglaine, meus sobrinhos Nathaniel e Thaise, minha mãe Cleci e em especial meu querido pai Balduino (*in memorian*).

Ao meu "filho" Peppe que me traz sempre momentos de alegria e relaxamento.

Aos membros da banca por disponibilizarem um pouco do seu tempo para contribuirem com este trabalho. Ao Prof. Álvaro, além das orientações, a paciência e ajuda na construção deste trabalho que tenho muito orgulho.

Ao projeto ELSA-Brasil, minha base como pesquisadora.

Ao CNPq pela bolsa de estudos proporcionada.

Ao Hospital Moinhos de Vento por incentivar a formação dos seus colaboradores. Em especial à Dra. Eliana e ao grupo do estudo POP-Brasil.

SUMÁRIO

RESUMO

O modelo de riscos proporcionais de Cox é um dos métodos mais utilizados na pesquisa clínica e epidemiológica para a análise de dados censurados, em grande parte por não exigir o conhecimento da distribuição de probabilidades do tempo. A principal suposição do modelo é a proporcionalidade de riscos ao longo do tempo, que pode ser restritiva em algumas situações práticas, como relações não lineares nas covariáveis ou efeitos de tratamentos que declinam no tempo. O modelo impõe uma estrutura global na função de sobrevivência e estima um único "efeito" médio, impossibilitando assim a estimação de "efeitos" das covariáveis localmente. A análise de dados censurados pode ser ainda mais complexa nos casos em que as censuras ocorrem somente em determinados períodos do tempo. Uma abordagem recente é o uso de modelos de regressão quantílica para dados de sobrevivência. São métodos robustos e flexíveis, no sentido em que permitem descrever a relação dos preditores em diferentes quantis da distribuição do tempo de sobrevivência. Pode ser vantajosa particularmente quando não estão atendidas as suposições de proporcionalidade de riscos e de linearidade. A grande maioria dos trabalhos sobre regressão de sobrevivência quantílica aborda aspectos da estimação dos parâmetros do modelo. No contexto epidemiológico, no entanto, frequentemente o objetivo é estimar o efeito (ou associação) de uma determinada exposição sobre o tempo até a ocorrência do evento. Este trabalho apresenta uma revisão das abordagens para estimação dos coeficientes do modelo de regressão quantílica para dados com censura à direita e uma abordagem computacional para estimar a função de risco (*hazard rate*) e razões de azares (*hazard ratio)* utilizando regressão quantílica. Os resultados das simulações mostraram que as estimativas de razão de azares diminuem na direção do valor de referência ao logo do tempo de acompanhamento.

ABSTRACT

The Cox proportional hazards model is one of the most widely used methods in clinical and epidemiological research for the analysis of censored data, largely because it does not require knowledge of survival time density. The main assumption of the model is proportionality of risks over time, which may be restrictive in some practical situations, such as nonlinear relationships in covariates or effects of treatments that decline over time. The model imposes a global structure on the survival function and estimates a single mean "effect", making it impossible to estimate the "effects" of the covariates locally. Analysis of censored data may be even more complex in cases where censoring occurs only in certain periods of time. A recent approach is the use of quantile regression models for survival data. They are methods robust and flexible, in the sense they allows to describe the relationship of the predictors in different quantiles of the distribution of survival time. It may be advantageous particularly where the proportionality assumptions of risk and linearity are not met. The vast majority of the work on quantile survival regression addresses aspects of estimation of parameters. In the epidemiological context, however, the objective is often to estimate the effect (or association) of a given exposure to the occurrence of the event over time. This work presents a review of the approaches to estimate the coefficients of the quantile regression model for right censored data as well as a computational approach to estimate the hazard rate and hazard ratio using quantile regression. The results of the simulation study show that the hazard ratio estimates decreases towards the reference value as the follow-up time increase.

1. APRESENTAÇÃO

Este trabalho consiste na tese de doutorado intitulada “Estimação de razão de azares por meio de regressão quantílica para dados com censura à direita: uma abordagem computacional”, apresentada ao Programa de Pós-Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul, em 21 de Setembro de 2017. O trabalho é apresentado em três partes, na ordem que segue:

1. Introdução, Revisão da Literatura e Objetivos
2. Artigo(s)
3. Conclusões e Considerações Finais.

2. INTRODUÇÃO

Em estudos epidemiológicos o modelo de riscos proporcionais de Cox tem sido o método frequentemente utilizado para a análise de dados de sobrevivência, pela flexibilidade de não exigir o conhecimento da distribuição de probabilidades do tempo de sobrevivência e inclusão de covariáveis. A suposição do modelo é de riscos proporcionais ao longo do tempo, ou seja, se, por exemplo, no início do estudo o indivíduo *i* tem um risco do evento igual a duas vezes o risco do indivíduo *j*, então esta razão de riscos será a mesma ao longo do tempo. Sob estas condições, uma limitação do modelo é a impossibilidade de captar mudanças no efeito das covariáveis ao longo do tempo, impondo uma estrutura global na função de sobrevivência(Cox, 1972; Allison, 2010).

A regressão quantílica oferece uma abordagem alternativa para a análise de sobrevivência. Sem fazer suposições globais sobre a forma da relação funcional das covariáveis e o tempo de sobrevivência, permite avaliar o efeito das covariáveis em diferentes quantis do tempo (Powell, 1986; Portnoy, 2003).

Este trabalho apresenta uma revisão das abordagens para estimação dos coeficientes do modelo de regressão quantílica para dados com censura a direita e, como objetivo principal, apresenta uma abordagem computacional com base na metodologia de Fizenberger & Wilke (2006) para estimar a função de risco (*hazard rate*) e razões de azares (*hazard ratio*)(Fitzenberger and Wilke, 2006).

3. REVISÃO DA LITERATURA

3.1 Conceitos básicos

A análise de sobrevivência, originalmente utilizada para avaliar o tempo até a ocorrência de óbitos, atualmente pode ser aplicada também a eventos como nascimentos, aposentadorias, doenças, entre outros. Alguns métodos, como Kaplan-Meier e o teste *logrank*, podem ser usados para estimar e comparar curvas de sobrevivência, no entanto, modelos de regressão são utilizados na maioria das aplicações. Modelos para dados de sobrevivência podem ser úteis para explorar a relação entre a experiência de sobrevivência e as variáveis explanatórias registradas no momento de entrada do indivíduo no estudo ou com variáveis medidas em diferentes ocasiões ao longo do tempo. Uma das principais características destes modelos é capacidade de incorporar a censura nas análises, fenômeno frequentemente presente em dados longitudinais. A censura à direita é o tipo mais comum, e é caracterizada por não ser possível mensurar o tempo até o evento para todos os indivíduos. Definições dos tipos de censura e de diferentes métodos de análise de dados de sobrevivência estão extensamente descrito na literatura (Bustamante-Teixeira et al. 2002; Allison 2010; Jr et al. 2011; Klein and Moeschberger 2013; Collett 2015).

O tempo de sobrevivência T é uma variável aleatória contínua não negativa, descrita pela função densidade $f(t)$, a qual pode ser interpretada como a probabilidade de ocorrência do evento de interesse em um intervalo instantâneo do tempo. A função de distribuição acumulada $F(t) = P(T \leq t)$ é a probabilidade de que um indivíduo experimente o evento antes do tempo t. Duas medidas centrais destes métodos são a função de sobrevivência $S(t)$ e a função de risco (*hazard function*) $h(t)$. A função $\mathrm{S(t)}$ representa a probabilidade de que, a partir do tempo de origem, o tempo de sobrevivência de um indivíduo seja maior do que t unidades de tempo, ou seja, $S(t) = 1 - F(t)$. A função $h(t)$, força de mortalidade condicional ou função de incidência, representa a taxa de ocorrência do evento no intervalo $(t, t + \Delta t)$, condicional ao fato de o indivíduo não ter experimentado o evento até o tempo t, ou seja,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t \mid T > t)}{\Delta t}.$$

Embora muitas vezes $h(t)$ seja chamada de função de risco, ela assume valores no intervalo $(0, +\infty)$ e, portanto, não é uma probabilidade. Outra função comumentemente utilizada é a função de risco (taxa) acumulado, definida como

$$H(t) = \int_0^t h(u)du$$

a qual mede a taxa de ocorrência do evento até o tempo t.

Estas funções descrevem o tempo de sobrevivência e estão relacionadas entre si, como mostrado a seguir:

$$F(t) = P(T \leq t) = \int_0^t f(u)du$$

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

$$h(t) = -\frac{d \ln S(t)}{dt} = \frac{f(t)}{S(t)}$$

$$H(t) = -\ln S(t)$$

$$S(t) = e^{-H(t)}.$$

Diferentes abordagens estão disponíveis para a estimação destas funções, via procedimentos paramétricos ou não paramétricos, sendo que na abordagem de estimação paramétrica é necessário o conhecimento da distribuição de probabilidades do tempo de sobrevivência $f(t)$. Estas funções estão descritas em detalhes na literatura (Allison, 2010; Jr et al., 2011; Klein and Moeschberger, 2013; Collett, 2015).

Na maioria das aplicações, diversos preditores são considerados na análise, sendo necessário o uso de modelagem. Entre as principais abordagens, estão os modelos paramétricos (modelo exponencial, Weibull, etc.) e o modelo de riscos proporcionais de Cox. Também foi proposto o uso do modelo de regressão quantílica para dados censurados para estimar a função taxa de falha(Fitzenberger and Wilke, 2006). Na pesquisa clínica e epidemiológica a distribuição de probabilidade do tempo de sobrevivência geralmente não é conhecida, sendo utilizado com mais frequência o modelo de Cox, descrito resumidamente na seção 3.2. Como alternativa, a abordagem de regressão quantílica é descrita na seção 3.3.

3.2 Modelo de riscos proporcionais de Cox

Em estudos epidemiológicos o modelo de riscos proporcionais de Cox tem sido o método mais utilizado para a análise de dados de sobrevivência. Em grande parte, seu uso se deve às propriedades do modelo, como não exigir o conhecimento da distribuição de probabilidades do tempo de sobrevivência, flexibilidade para o ajuste multivariável ou, quando necessário, para especificar preditores com possibilidade de poder analisar o efeito de covariáveis neste tempo. Em algumas situações essa distribuição é conhecida e modelos paramétricos também podem ser usado (Cox, 1972; Allison, 2010).

A principal suposição do modelo de Cox é a proporcionalidade de riscos ao longo do tempo, ou seja, o risco tem a mesma função para todos os indivíduos e as variáveis preditoras tem o mesmo efeito ao longo do tempo (Allison, 2010). Por exemplo, se aos 50 anos o risco de infarto em homens é duas vezes o risco de mulheres, essa razão de riscos deverá ser o mesma aos 60 anos, aos 70 anos e em outras idades. No entanto, esta é uma suposição forte, e muitas vezes restritiva, cuja veracidade deve ser verificada. Um exemplo clássico é o efeito de tratamento que tende a diminuir a medida que aumenta o tempo de uso do medicamento (Grambsch and Therneau, 1994).

Outras situações e exemplos práticos são apresentados por diversos autores, que chamam atenção para os cuidados no momento de interpretar os resultados quando são usados mecanismos para corrigir a falta de proporcionalidade, principalmente no uso de variáveis tempo-dependente. Estudos envolvendo câncer de mama, pulmão, entre outros, mostram que alguns preditores tendem a apresentar maior efeito quando avaliado em períodos curtos de tempo em relação a períodos maiores. Como consequência, a razão de azares pode ser subestimada ou superestimada (Xu and O'Quigley, 2000; Prentice et al., 2005; Peng and Huang, 2007; Schemper et al., 2009). O desenvolvimento de resistência a uma terapia com o passar do tempo é outro exemplo de riscos não proporcionais (Box-Steffensmeier et al., 2003).

Diferentes mecanismos podem conduzir à violação da proporcionalidade dos riscos, como, por exemplo, aumento ou diminuição da razão entre as funções de risco (*increasing or decreasing relative hazards*), funções de risco cruzadas (*crossing hazards*), funções de risco

divergentes (*diverging hazards*) ou funções de risco não monótonas (*non-monotonic hazards*) (Ng'andu, 1997)

Algumas abordagens foram propostas para avaliar a validade da suposição de riscos proporcionais, incluindo métodos gráficos e testes de hipóteses, tais como estimativas do modelo de Cox estratificado pelo tempo; modelo com variáveis dependentes do tempo, no qual são incluídos termos de interação entre o tempo e preditores ou o teste de correlação linear entre o rank dos tempos em que ocorrem os eventos, e a estimativa do resíduo parcial de Schoenfeld (Harrell Jr., Frank E. 2015).

Por meio de simulações e comparações empíricas, Ng'andu (1997) analisou a performance de cinco testes estatísticos, considerando a presença ou ausência de censuras. Um resumo dos principais testes (*Time-dependent covariate, Linear correlation test, Weighted residual test, Score process test e Omnibus test*) usados para verificar a suposição de riscos proporcionais do modelo de Cox, bem como resultados de simulações com diferentes cenários violando esta suposição: aumento do risco relativo, decréscimo do risco relativo, riscos cruzados, riscos divergentes e riscos não monotônicos foram apresentados. Para cada situação é apresentado o desempenho dos testes em relação ao poder de detecção da violação da proporcionalidade dos riscos, de acordo com o nível de significância. Para o nível de significância de 5% e riscos cruzados, *Omnibus test* apresentou pior desempenho, com poder igual a 55,3%, enquanto que o maior poder (79,4%) foi associado ao teste com variável tempo dependente (*Time-dependent test*). De maneira geral, para cenários simulados onde não há proporcionalidade de riscos, os testes *Time-dependent covariate, Linear correlation test e Weighted residual test* mostraram ter bom poder para detectar esta violação, seja para variáveis continuas ou dicotômicas, bem como foram pouco afetados pela censura (Ng'andu, 1997).

Uma abordagem gráfica baseada do diagrama de dispersão do resíduo parcial associado a cada preditor do modelo versus o tempo para avaliação da proporcionalidade dos riscos também foi proposta (Schoenfeld, 1982). Outras abordagens gráficas foram descritas por Harrell Jr (Harrell Jr., Frank E., 2015).

Para preditores quantitativos também é importante avaliar a suposição de linearidade, haja vista a especificação de forma funcional incorreta pode ser confundida com a violação da

suposição de riscos proporcionais(Keele, 2010). A categorização de preditores quantitativos, particularmente pela utilização de percentis da sua distribuição (tercis, quartis, etc.), nem sempre é recomendada. Os principais problemas que podem surgir com a categorização em quartis, por exemplo, são: a) necessidade de realizar múltiplos testes de hipóteses para comparações entre quartis, aumentando a chance de falsos positivos; b) a heterogeneidade do risco entre os grupos de indivíduos pode resultar em estimativas incorretas ou perda de poder; e, c) dificuldade de fazer comparações entre estudos, devido aos pontos de corte diferentes (Bennette and Vickers, 2012; Rothman et al., 2012).

Alguns autores apresentam estratégias de análise quando se detecta a não proporcionalidade. Tradicionalmente e bem discutida na literatura é a incorporação da interação do tempo (ou ainda *log* do tempo) com a covariável(Cox, 1972; Stablein DM et al., 1981; Gore et al., 1984; Verweij and van Houwelingen, 1995; Abrahamowicz et al., 1996). Outra alternativa avaliada por autores como Hess e Gray é o uso de splines, ferramenta bastante completa pois permite avaliar tanto a proporcionalide quanto a linearidade da covariável tempo dependente(Gray, 1992, 1994; Hess, 1994; Herndon and Harrell, 1995). Başar (2006) intoduz uma discussão bastante interessante de como sumarizar o efeito de uma covariável quando a hipótese de proporcioalidade é rejeitada. Além dos métodos como inclusão da interação tempo e covariável e modelos segmentados (*piecewise model*), o autor destaca o uso de splines como uma alternativa que pode apresentar bons resultados(BaşAr 2007).

A interpretação da razão de azares como uma estimativa de efeito causal pode ser equivocada, mesmo quando o modelo foi especificado corretamente, na ausência de confundimento não observado e de erros de medidas. Uma razão é a estimativa única da razão de azares média ao longo do período de acompanhamento. Portanto, as conclusões do estudo podem depender da duração do período acompanhamento, pois a estimativa média ignora a distribuição dos eventos no referido período(Hernán, 2010).

A análise de dados censurados pode ser ainda mais complexa nos casos em que as censuras ocorrem somente em determinados períodos do tempo (Leng and Tong, 2013). Uma abordagem recente é o uso de modelos de regressão quantílica para dados de sobrevivência. São métodos robustos e flexíveis, no sentido em que permitem entender como é a relação dos

preditores em diferentes quantis da distribuição do tempo. Pode ser vantajosa particularmente quando não estão atendidas as suposições de proporcionalidade de riscos e de linearidade (Rodriguez, Robert N., 2013).

3.3 Métodos de regressão quantílica para dados de sobrevivência

Um dos primeiros trabalhos envolvendo dados censurados foi o modelo de tobitos (*tobit model*)(Tobin 1958). O modelo foi posteriormente estendido para os quantis da variável resposta e chamado de modelo de regressão quantílica censurado (*censored regression quantile*), estimando os parâmetros pelo método dos mínimos desvios absolutos (*LAD - Least Absolute Deviations*). Posteriormente, Powell (1986), propôs uma abordagem relacionada a modelos lineares, mas limitando as observações a censura fixa. No entanto, esta abordagem não pode ser geralizada para situações em que o tempo de censura não pode ser observado para todos os indivíduos(Powell, 1984, 1986).

Modelos de regressão quantílica para dados com censura à direita tem se tornado uma ferramenta poderosa na análise de sobrevivência(Leng and Tong 2013), e tema de grande interesse na literatura. Esse modelo estima o quantil da distribuição do tempo de sobrevivência, condicional aos valores dos preditores, relaxa a restrição de proporcionalidade dos riscos ao longo do tempo e é uma escolha natural para modelagem de dados com heterogeneidade (Bilias et al., 2000; Wey et al., 2013; SAS, 2016).

Considerando T uma variável aleatória não negativa que representa o tempo de sobrevivência, o modelo de sobrevivência quantílico para $log(T)$ é escrito como

$$Q_{log(T)}(\tau|\boldsymbol{x}) = \boldsymbol{x}'\beta(\tau), \quad (1)$$

em que $Q_{log\,T}(\tau|\boldsymbol{x})$ é o τ-ésimo quantil de $log(T)$ condicional no valor do vetor de preditores $\boldsymbol{x}$, e β é o vetor de parâmetros de regressão desconhecidos. Como a função quantílica é invariante sob uma transformação monótona, o modelo também pode ser escrito como

$$Q_T(\tau|\boldsymbol{x}) = exp\{\boldsymbol{x}'\boldsymbol{\beta}(\tau)\}.$$

Esse modelo pode ser útil para avaliar o efeito dos preditores no tempo de sobrevivência. Para estimar a função de sobrevivência para um dado valor $\boldsymbol{x}$, pode-se utilizar sua relação entre a função quantílica

$$F\big(Q_T(\tau|\boldsymbol{x})\big) = P\big(T \leq Q_T(\tau|\boldsymbol{x})\big) = \tau,$$

ou seja, a função de distribuição acumulada $F_T(t|\boldsymbol{x})$ mapeia $Q_T(\tau|\boldsymbol{x})$ em τ e, assim, a função de sobrevivência $S_T(t|\boldsymbol{x})$ mapeia $Q_T(\tau|\boldsymbol{x})$ em $1-\tau$.

Utilizando-se de recursos como simulação de Monte Carlo, Buckinsky & Hahn (1998) apresentam uma solução para um problema de programação linear, podendo o estimador ser utilizado nos casos em que o ponto de censura é uma função desconhecida dos regressores. Este estimador possui boas propriedades para pequenas amostras, principalmente em relação ao viés induzido por grande quantidade de censura(Buchinsky and Hahn, 1998).

Métodos computacionais utilizando técnicas de reamostragem foram propostos para melhorar a aproximação da distribuição dos estimadores dos parâmetros do modelo de regressão quantílica censurado(Bilias et al., 2000).

Uma formulação da regressão quantílica foi utilizada por Koenker & Geling (2001) como forma de reanalisar as conclusões do estudo experimental sobre a mortalidade (específica por idade) da mosca da fruta mediterrânea. Estudos anteriores sobre a mosca da fruta utilizaram métodos de tabela de vida, que são úteis para avaliar o efeito de gênero e variáveis discretas na sobrevivência e mortalidade, mas são menos adequados para avaliar o papel de variáveis quantitativas. Por outro lado, modelos tradicionais de análise de sobrevivência (paramétricos ou semiparamétricos) são flexíveis para incorporar preditores categóricos e quantitativos, mas impõem condições exigentes sobre como eles influenciam o tempo de sobrevivência. Por exemplo, o modelo de tempo de vida acelerado (*AFT model*) e o modelo de Cox assumem que os preditores afetam somente a locação do tempo de sobrevivência transformado, mas não sua forma. Isto é particularmente restritivo no contexto da mosca da fruta, em que o interesse maior está na cauda superior do seu tempo de sobrevivência. Os resultados da reanálise trouxeram refinamentos importantes nas conclusões do estudo original, com o aparente declínio na mortalidade em idades mais avançadas e um efeito cruzado nas funções de sobrevivência entre os sexos. Como argumento para o uso da regressão quantílica, os autores afirmam que o método

oferece uma abordagem de análise mais poderosa e natural ao analisar a variabilidade causada pela heterocedasticidade nos dados ou heterogeneidade na população. Além do exemplo, o artigo também apresenta excelente revisão dos métodos(Koenker R,2001).

Contribuições para a estimação dos parâmetros do modelo foram sugeridas, visando incorporar aspectos específicos dos dados, como altas taxas de censura e de dimensionalidade (Chernozhukov and Hong, 2002), bem como diferentes abordagens para modelar o quantil do tempo de sobrevivência.

Em 2002, Machado & Portugal utilizaram a abordagem da regressão quantílica em dados de duração de desemprego obtidos de *Displaced Worker Survey* (DWS-1988) com o objetivo de mostrar o impacto de negligenciar a heterogeneidade nos estimadores da regressão quantílica, bem como, destacar o ganho de informação com o uso desta abordagem. A pesquisa mostrou que algumas variáveis (por exemplo, aviso prévio, salário anterior) tem maior efeito em tempos de duração de desemprego menores, outras desaparecem ao longo do período de desemprego, enquanto outras permanecem constantes ao longo do tempo (Machado, José A.F. and Portugal,2002).

Portnoy (2003) propôs uma generalização do estimador Kaplan-Meier para estimar a função quantílica do tempo de sobrevivência condicional nos preditores, utilizando uma abordagem baseada em um procedimento de estimação por reponderação recursiva, em que a massa de cada observação censurada é redistribuída para as observações não censuradas. O objetivo principal foi comparar esta nova abordagem com a abordagem tradicional do modelo de Cox e também mostrar como a regressão quantílica poderia complementar ao modelo de Cox, ou ainda oferecer uma melhor análise dos dados. A comparação com o modelo de Cox usando dados reais, mostrou que existem diferenças entre as abordagens, principalmente em quantis específicos, e que a análise usando regressão quantílica tende a ser mais sensível (Portnoy, 2003).

Em 2004, Guimarães et al. utilizaram a abordagem do modelo de regressão quantílica para dados censurados para avaliar o tempo de duração de desemprego entre a população americana. Alternativa a abordagem utilizada até recentemente, a regressão quantílica apresenta como ponto forte nesta análise, principalmente, a possibilidade de avaliar períodos curtos de

desemprego e períodos longos, ou até mesmo compará-los em relação a determinadas características(Guimarães et al., 2004).

Utilizando dados de desemprego na Alemanha, Fritzenberger & Wilke (2006) discutem as vantagens e desvantagens do uso da regressão quantílica em relação aos métodos tradicionais. Os autores apresentam o mesmo estimador proposto por Machado & Portugal (2002) e Guimarães et al.(2004), com pequenas modificações para resolver problemas de suavização (Fitzenberger and Wilke, 2006).

Peng & Huang (2008) propuseram um estimador visto como uma extensão do estimador Nelson-Aalen, com desempenho similar ao método de Portnoy (2003), porém, com algoritmos mais simples. O estimador assume que as censuras são condicionalmente independentes e que existe linearidade em todos os quantis (Peng and Huang, 2008).

Com foco no pacote *quantreg* do software R, Koenker (2008) descreve os métodos implementados neste pacote: Powell (1986) com estimador para censura fixa, Portnoy (2003) e Peng and Huang (2008) para censura aleatória. Através de simulações com quatro mecanismos de geração dos dados, dois para o tempo do evento e dois para censuras, o autor compara os estimadores Portnoy, Peng-Hung, Powell e máxima verossimilhança Gaussiana. Para censura constante o estimador de Powell demonstrou-se ligeiramente melhor, enquanto que para censura variável sua performance foi pior em relação aos demais(Koenker, 2008).

Wang & Wang (2009) apresentaram a abordagem de regressão quantílica ponderada localmente, que é menos restritiva do que os demais métodos, os quais geralmente exigem independência condicional nos tempos de sobrevivência e a censura, ou linearidade em todos os quantis(Wang and Wang, 2009). O método assume independência entre o tempo de sobrevivência e as censuras, condicional aos valores dos preditores, produzindo estimadores consistentes e assintoticamente normal.

O modelo de regressão quantílica com censura à direita foi generalizado para permitir a inclusão de preditores com relação não linear com a resposta, por meio de B-splines (Neocleous and Portnoy, 2009).

Wey et al.(2013) propuseram uma nova abordagem, utilizando particionamento recursivo (árvore de sobrevivência) para estimar os coeficientes da regressão, possibilitando maior flexibilidade para lidar com censuras que dependem das covariáveis(Wey et al. 2013).

Galvão et al. (2013) apresentaram duas abordagens para superar as limitações de métodos tradicionais na estimação do modelo de sobrevivência quantílico na presença de um número de parâmetros muito grande ou com covariáveis correlacionadas e heterogeneidade individual não observada. Estas duas abordagens tem como princípio os modelos de efeitos fixos da regressão quantílica com uma restrição de separação na probabilidade de censura. Simulações Monte Carlo foram conduzidas para avaliar os estimadores e indicaram boa performance em amostras finitas(Galvao et al., 2013).

A grande maioria dos trabalhos sobre regressão de sobrevivência quantílica aborda aspectos da estimação dos parâmetros do modelo. No contexto epidemiológico, no entanto, frequentemente o objetivo é estimar o efeito (ou associação) de uma determinada exposição sobre o tempo até a ocorrência do evento. Recentemente, Xue et al. (2016) utilizaram dados simulados e de estudos reais para discutir aspectos da interpretação das estimativas dos parâmetros da regressão quantílica no contexto da pesquisa clínica e de saúde pública, e também uma comparação com o modelo de Cox. Como uma das vantagens em relação ao modelo de Cox desenvolveram uma metodologia para validar o tempo até o evento predito pelo modelo(Xue et al. 2016).

Do nosso conhecimento somente um trabalho investigou a estimação da função de risco (*hazard rate*) a partir das estimativas do modelo de regressão quantílica(Fitzenberger and Wilke, 2006). O modelo de regressão quantílica para dados com censura à direita estima o quantil da distribuição do tempo de sobrevivência condicional aos valores dos preditores (ou seja, estima o valor do tempo de sobrevivência para cada quantil da distribuição, condicional aos preditores), mas é possível fazer inferências sobre a razão de azares a partir das estimativas do modelo.

Uma abordagem direta(Fitzenberger and Wilke 2006) é estimar uma densidade a partir dos quantis estimados $\hat{Q}_{T_i}(\tau|\boldsymbol{x}_i) = h^{-1}\left(\boldsymbol{x}_i'\boldsymbol{\beta}(\tau)\right)$. Uma estimativa aproximada da função de risco entre os quantis τ_1 e τ_2 pode ser obtida como

$$\hat{h}_i(t) = \frac{\tau_2 - \tau_1}{\left[h^{-1}\left(\boldsymbol{x}_i'\boldsymbol{\beta}(\tau_2)\right) - h^{-1}\left(\boldsymbol{x}_i'\boldsymbol{\beta}(\tau_1)\right)\right]\left[1 - \frac{1}{2}(\tau_1 + \tau_2)\right]}.$$

Contudo, os quantis τ_1 e τ_2 $(\tau_1 < \tau_2)$ devem ser escolhidos com uma distância suficientemente grande para que as estimativas condicionais estejam ordenadas corretamente, isto é, para satisfazer a relação

$$\hat{Q}_{T_i}(\tau_1|x_i) < \hat{Q}_{T_i}(\tau_2|x_i).$$

Um estimador similar foi proposto por Machado & Portugal (2002) e Guimarães et al.(2004), baseado na função esparsidade (inversa da função densidade), com as mesmas limitações em relação aos quantis. Estes autores também propuseram um método de estimação da função de risco por meio de simulações, o qual foi modificado por Fizenberger & Wilke (2006) e está descrito nas etapas abaixo:

Etapa 1: Gerar M variáveis aleatórias $\tau_m\,(m = 1,2,\ldots,M)$ de uma distribuição Uniforme em (τ_I, τ_S), de tal forma que os quantis extremos $0 \leq \tau < \tau_I$ e $\tau_S < \tau \leq 1$ não são considerados;

Etapa 2: Para cada τ_m, estimar o modelo de sobrevivência quantílico definido na equação (1), obtendo M vetores de estimativas dos parâmetros $\widehat{\boldsymbol{\beta}}(\tau_m)$;

Etapa 3: Para um dado valor do vetor de preditores $\boldsymbol{x}_0$, o vetor M-dimensional dos tempos de sobrevivência simulados são obtidos por

$$T_m^* \equiv \hat{Q}_{T_i}(\tau_m|\boldsymbol{x}_0) = exp\{\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}(\tau_m)\},$$

com $m = 1,2,\ldots,M$;

Etapa 4: Baseado na amostra $\{T_m^*;\ m = 1,2,\ldots,M\}$, estimar a densidade condicional $f^*(t|\boldsymbol{x}_0)$ e a função de distribuição condicional $F^*(t|\boldsymbol{x}_0)$;

Etapa 5: Estimar a função de risco condicional em $\boldsymbol{x}_0$ no intervalo (τ_I, τ_S) dada por

$$\hat{h}_0(t) = \frac{(\tau_S - \tau_I)\, f^*(t|\boldsymbol{x}_0)}{1 - \tau_I - (\tau_S - \tau_I)\, F^*(t|\boldsymbol{x}_0)}.$$

O método original proposto pelos autores utiliza um estimador kernel para a densidade condicional

$$f^*(t|\boldsymbol{x}_0) = \frac{1}{M\,h}\sum_{m=1}^{M} K\left(\frac{t-T^*}{h}\right)$$

em que h é a largura de banda e $K(.)$ é a função kernel. Assim, a função de distribuição do estimador é

$$F^*(t|\boldsymbol{x}_0) = \frac{1}{M}\sum_{m=1}^{M} K\left(\frac{t-T^*}{h}\right)$$

com $K(\upsilon) = \int_a^t K(v)dv$. Considerando que os tempos de sobrevivência são positivos, Machado & Portugal (2002) e Guimarães et al.(2004) sugerem utilizar $a = 0$, enquanto Fizenberger & Wilke (2006) sugeriram $a = -\infty$. Se os tempos são sempre positivos, uma alternativa melhor e mais simples seria a adoção de uma função kernel no logaritmo do tempo de sobrevivência, utilizando a transformação inversa para obter as estimativas da função densidade da variável original.

No contexto do modelo de sobrevivência quantílico não foram encontrados trabalhos que tenham explorado a estimação da razão de azares (*hazard ratio*), como estimativas de associação ou efeitos.

Os programas para regressão quantílica com censura não estimam diretamente a função de risco e nem a razão de azares. Implementada no pacote *quantreg* do software R, a função *crq* ajusta o modelo de regressão quantílica para dados censurados com três opções para o tratamento das censuras: método de censura fixa de Powell (1986), método de censura aleatória de Portnoy (2003) e também o método proposto por Peng & Huang (2008)(Koenker et al., 2013).

No programa SAS, o procedimento PROC QUANTLIFE disponibiliza tanto o estimador derivado da função de sobrevivência de Kaplan-Meier quanto do estimador de Nelson-Aalen, que não exigem o conhecimento da distribuição do tempo de sobrevivência e podem ser aplicados a dados heterogêneos(Rodriguez, Robert N., 2013; SAS, 2016).

4. OBJETIVOS

Objetivo Geral

Desenvolver um método computacional para estimar a função de risco (*hazard rate*) e a medida de razão de azares (*hazard ratio*) por meio de regressão quantílica para dados com censura à direita.

Objetivo Específico

Apresentar uma revisão sobre as abordagens para estimação dos coeficientes do modelo de regressão quantílica para dados com censura à direita e potencialidades de aplicações na pesquisa clínica e epidemiológica

REFERÊNCIAS BIBLIOGRÁFICAS

1. Abrahamowicz M, MacKenzie T, Esdaile JM. Time-Dependent Hazard Ratio: Modeling and Hypothesis Testing With Application in Lupus Nephritis. J. Am. Stat. Assoc. 1996;91(436):1432–9.

2. Allison PD. Survival Analysis Using SAS: A Practical Guide, Second Edition. SAS Institute; 2010.

3. BaşAr E. NON-PROPORTIONAL HAZARDS WITH APPLICATION TO KIDNEY TRANSPLANT DATA. Commun. Fac. Sci. Univ. Ank. Ser. A1Mathematics Stat. [Internet]. 2007 [cited 2017 Jul 8]; Available from: http://dergiler.ankara.edu.tr/dergiler/29/129/881.pdf

4. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Med. Res. Methodol. 2012 Feb 29;12(1):21.

5. Bilias Y, Chen S, Ying Z. Simple resampling methods for censored regression quantiles. J. Econom. 2000 Dezembro;99(2):373–86.

6. Box-Steffensmeier JM, Reiter D, Zorn C. Nonproportional Hazards and Event History Analysis in International Relations. J. Confl. Resolut. 2003;47(1):33–53.

7. Buchinsky M, Hahn J. An Alternative Estimator for the Censored Quantile Regression Model. Econometrica. 1998 May;66(3):653.

8. Bustamante-Teixeira MT, Faerstein E, Latorre M do R. Survival analysis techniques. Cad. Saúde Pública. 2002 Jun;18(3):579–94.

9. Chernozhukov V, Hong H. Three-Step Censored Quantile Regression and Extramarital Affairs. J. Am. Stat. Assoc. 2002;97(459):872–82.

10. Collett D. Modelling Survival Data in Medical Research, Third Edition. CRC Press; 2015.

11. Cox DR. Regression Models and Life-Tables. J. R. Stat. Soc. Ser. B Methodol. 1972;34(2):187–220.

12. Fitzenberger B, Wilke RA. Using quantile regression for duration analysis. Allg. Stat. Arch. 2006 Mar 1;90(1):105–20.

13. Galvao AF, Lamarche C, Lima LR. Estimation of Censored Quantile Regression for Panel Data With Fixed Effects. J. Am. Stat. Assoc. 2013;108(503):1075–89.

14. Gore SM, Pocock SJ, Kerr GR. Regression Models and Non-Proportional Hazards in the Analysis of Breast Cancer Survival. J. R. Stat. Soc. Ser. C Appl. Stat. 1984;33(2):176–95.

15. Grambsch PM, Therneau TM. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. Biometrika. 1994;81(3):515–26.

16. Gray RJ. Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis. J. Am. Stat. Assoc. 1992;87(420):942–51.

17. Gray RJ. Spline-Based Tests in Survival Analysis. Biometrics. 1994;50(3):640–52.

18. Guimarães, Machado, José A.F., Portugal, Pedro. Has long become longer or short become shorter? Evidence from a censored quantile regression analysis of the changes in the distribution of U.S. unemployment duration. 2004 Abril;

19. Harrell Jr., Frank E. Regression Modeling Strategies - With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis [Internet]. 2015 [cited 2017 Mar 29]. Available from: http://www.springer.com/br/book/9783319194240

20. Hernán MA. The Hazards of Hazard Ratios. Epidemiol. Camb. Mass. 2010 Jan;21(1):13–5.

21. Herndon JE, Harrell FE. The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariables. Stat. Med. 1995 Outubro;14(19):2119–29.

22. Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. Stat. Med. 1994 Maio;13(10):1045–62.

23. Jr DWH, Lemeshow S, May S. Applied Survival Analysis: Regression Modeling of Time to Event Data. John Wiley & Sons; 2011.

24. Keele L. Proportionally Difficult: Testing for Nonproportional Hazards in Cox Models. Polit. Anal. 2010 Mar 1;18(2):189–205.

25. Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. Springer Science & Business Media; 2013.

26. Koenker R. Censored Quantile Regression Redux. J. Stat. Softw. [Internet]. 2008 [cited 2017 Jul 8];27(6). Available from: http://www.jstatsoft.org/v27/i06/

27. Koenker R, code) SP (Contributions to CQ, code) PTN (Contributions to SQ, code) AZ (Contributions to dynrq code essentially identical to his dynlm, code) PG (Contributions to nlrq, advice) BDR (Initial (2001) R port from S (to my everlasting shame--how could I have been so slow to adopt R and for numerous other suggestions and useful. quantreg: Quantile Regression [Internet]. 2013 [cited 2014 Aug 2]. Available from: http://cran.r-project.org/web/packages/quantreg/index.html

28. Leng C, Tong X. A quantile regression estimator for censored data. Bernoulli. 2013 Feb;19(1):344–61.

29. Machado, José A.F., Portugal P. Exploring transition data through quantile regression methods: An application to US unemployment duration [Internet]. 2002 [cited 2014 Aug 31]. Available from: http://www.academia.edu/2956244/Exploring_transition_data_through_quantile_regression_methods_An_application_to_US_unemployment_duration

30. Neocleous T, Portnoy S. Partially linear censored quantile regression. Lifetime Data Anal. 2009 Sep;15(3):357–78.

31. Ng'andu NH. An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. Stat. Med. 1997 Mar 30;16(6):611–26.

32. Peng L, Huang Y. Survival analysis with temporal covariate effects. Biometrika. 2007 Aug 1;94(3):719–33.

33. Peng L, Huang Y. Survival Analysis With Quantile Regression Models. J. Am. Stat. Assoc. 2008;103(482):637–49.

34. Portnoy S. Censored Regression Quantiles. J. Am. Stat. Assoc. 2003;98(464):1001–12.

35. Powell JL. Least absolute deviations estimation for the censored regression model. J. Econom. 1984 Jul;25(3):303–25.

36. Powell JL. Censored regression quantiles. J. Econom. 1986 Jun;32(1):143–55.

37. Prentice RL, Langer R, Stefanick ML, Howard BV, Pettinger M, Anderson G, et al. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. Am. J. Epidemiol. 2005 Sep 1;162(5):404–14.

38. Rodriguez, Robert N. L Guixian. Using the QUANTLIFE Procedure for Survival Analysis. 2013;

39. Rothman KJ, Professor TLLA, Greenland S. Modern Epidemiology. Third, Mid-cycle revision edition. Philadelphia: LWW; 2012.

40. SAS. SAS Institute Inc. 2016. SAS/STAT® 14.2 User's Guide. Cary, NC: SAS Institute Inc. 2016;

41. Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratios by weighted Cox regression. Stat. Med. 2009 Agosto;28(19):2473–89.

42. Schoenfeld D. Partial Residuals for The Proportional Hazards Regression Model. Biometrika. 1982;69(1):239–41.

43. Stablein DM et al. Analysis of survival data with non-proportional hazard functions [Internet]. ResearchGate. 1981 [cited 2017 Mar 4]. Available from: https://www.researchgate.net/publication/15924893_Analysis_of_survival_data_with_non-proportional_hazard_functions

44. Tobin J. Estimation of Relationships for Limited Dependent Variables. Econometrica. 1958 Jan;26(1):24.

45. Verweij PJM, van Houwelingen HC. Time-Dependent Effects of Fixed Covariates in Cox Regression. Biometrics. 1995;51(4):1550–6.

46. Wang HJ, Wang L. Locally Weighted Censored Quantile Regression. J. Am. Stat. Assoc. 2009 Setembro;104(487):1117–28.

47. Wey A, Wang L, Rudser K. Censored quantile regression with recursive partitioning-based weights. Biostat. Oxf. Engl. 2013 Jan;15(1):170–81.

48. Xu R, O'Quigley J. Estimating average regression effect under non-proportional hazards. Biostatistics. 2000 Dec 1;1(4):423–39.

49. Xue X, Xie X, Strickler HD. A censored quantile regression approach for the analysis of time to event data. Stat. Methods Med. Res. 2016 May 10;

50. Reappraising Medfly Longevity: A Quantile Regression Survival Analysis. J. Am. Stat. Assoc. 2001;96(June):458–68.

ARTIGO 1

Regressão quantílica para dados com censura à direita como alternativa para o modelo de riscos proporcionais de Cox: uma revisão da literatura

Quantile regression for right censored data: a literature review

Marina Bessel, Doutoranda em Epidemiologia pela UFRGS;

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL (UFRGS)

A ser enviado para "*Cadernos de Saúde Pública*"

1. Introdução

Modelos para dados de sobrevivência podem ser úteis para explorar a relação entre a experiência de sobrevivência e as variáveis explanatórias registradas no momento de entrada do indivíduo no estudo ou com variáveis medidas em diferentes ocasiões ao longo do tempo. Uma das principais características destes modelos é capacidade de incorporar a censura nas análises, bastante frequente em dados longitudinais. A censura à direita é o tipo mais comum, caracterizada pelo fato do verdadeiro tempo de sobrevivência ser maior do que aquele observado (Bustamante-Teixeira et al., 2002; Allison, 2010).

O modelo de riscos proporcionais de Cox tem sido o método mais utilizado para a análise de dados de sobrevivência de estudos clínicos ou epidemiológicos, em grande parte por não exigir o conhecimento da distribuição de probabilidades do tempo de sobrevivência. A principal suposição deste modelo é a existência de proporcionalidade de riscos ao longo do tempo, ou seja, o risco tem a mesma função para todos os indivíduos e as variáveis preditoras tem o mesmo efeito ao longo do tempo. No entanto, proporcionalidade de riscos é uma suposição forte, e muitas vezes restritiva, como exemplo, o efeito de tratamento que tende a diminuir a medida que aumenta o tempo de uso do medicamento (Grambsch and Therneau, 1994).

Uma abordagem recente é o uso de modelos de regressão quantílica para dados de sobrevivência. São métodos robustos e flexíveis, no sentido de que permitem entender como é a relação dos preditores em diferentes quantis da distribuição do tempo. Pode ser vantajosa particularmente quando não estão atendidas as suposições de proporcionalidade de riscos e de linearidade(Rodriguez, Robert N., 2013).

Este artigo apresenta uma revisão das abordagens para estimação dos coeficientes do modelo de regressão quantílica para dados com censura a direita.

2. Revisão

James Tobin (Tobin, 1958) publicou o primeiro trabalho com dados censurados, sendo modificado pouco tempo depois para modelar os quantis da variável resposta, chamado de modelo de regressão quantílica censurado (*censored regression quantile*). Mais de duas décadas depois, foi proposta uma abordagem relacionando os modelos lineares limitando a pontos com censura fixa e conhecida (Powell, 1984, 1986).

Os modelos de regressão quantílica para dados com censura à direita para análise de sobrevivência estimam o quantil da distribuição do tempo de sobrevivência condicional aos valores dos preditores, e não exigem homogeneidade e proporcionalidade de riscos ao longo do tempo(Bilias et al., 2000; Wey et al., 2013; SAS, 2016).

O tempo de sobrevivência é representado pela variável aleatória não negativa T, e o modelo de sobrevivência quantílico para $log(T)$ é escrito como

$$Q_{log(T)}(\tau|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}(\tau), \quad (1)$$

em que $Q_{log\,T}(\tau|\boldsymbol{x})$ é o τ-ésimo quantil de $log(T)$ condicional no valor do vetor de preditores $\boldsymbol{x}$, e $\boldsymbol{\beta}$ é o vetor de parâmetros de regressão desconhecidos. Como a função quantílica é invariante sob uma transformação monótona, o modelo também pode ser escrito como

$$Q_T(\tau|\boldsymbol{x}) = exp\{\boldsymbol{x}'\boldsymbol{\beta}(\tau)\}.$$

Esse modelo pode ser útil para avaliar o efeito dos preditores no tempo de sobrevivência. Para estimar a função de sobrevivência para um dado valor $\boldsymbol{x}$, pode-se utilizar sua relação entre a função quantílica

$$F\big(Q_T(\tau|\boldsymbol{x})\big) = P\big(T \leq Q_T(\tau|\boldsymbol{x})\big) = \tau,$$

ou seja, a função de distribuição acumulada $F_T(t|\boldsymbol{x})$ mapeia $Q_T(\tau|\boldsymbol{x})$ em τ e, assim, a função de sobrevivência $S_T(t|\boldsymbol{x})$ mapeia $Q_T(\tau|\boldsymbol{x})$ em $1-\tau$.

Uma formulação da regressão quantílica foi utilizada por Koenker & Geling (2001) como forma de reanalisar as conclusões do estudo experimental sobre a mortalidade (específica por idade) da mosca da fruta Mediterrânea. Os resultados da reanálise trouxeram refinamentos importantes nas conclusões do estudo original, com o aparente declínio na mortalidade em idades mais avançadas e um efeito cruzado nas funções de sobrevivência entre os sexos(Koenker and Geling,2001).

Contribuições para a estimação dos parâmetros do modelo foram sugeridas, visando incorporar aspectos específicos dos dados, como altas taxas de censura e de dimensionalidade (Chernozhukov and Hong, 2002), e também diferentes abordagens para modelar o quantil do tempo de sobrevivência.

Em 2003 foi proposta uma generalização do estimador Kaplan-Meier para estimar a função quantílica do tempo de sobrevivência, condicional nos preditores, utilizando um procedimento de estimação por reponderação recursiva, em que a massa de cada observação censurada é redistribuída para as observações não censuradas. O objetivo principal foi comparar esta nova abordagem com o modelo de Cox e mostrar como a regressão quantílica poderia complementar as análises pelo modelo de Cox e as vantagens do modelo quantílico (Portnoy, 2003).

Dados sobre desemprego na Alemanha foram usados para discutir as vantagens e desvantagens do uso da regressão quantílica em relação aos métodos tradicionais. Os autores apresentam o mesmo estimador proposto por Machado & Portugal (2002) e Guimarães et al.(2004), com pequenas modificações para resolver problemas de suavização(Fitzenberger and Wilke, 2006).

Em 2008 foi proposto um estimador visto como uma extensão do estimador Nelson-Aalen, com desempenho similar ao método de Portnoy (2003), porém, com algoritmos mais simples. O estimador assume que as censuras são condicionalmente independentes e que existe linearidade em todos os quantis (Peng and Huang, 2008).

Outra abordagem, ponderada localmente, assume independência entre o tempo de sobrevivência e as censuras, condicional aos valores dos preditores, produzindo estimadores consistentes e assintoticamente normal. É menos restritiva do que os demais métodos, os quais exigem independência condicional nos tempos de sobrevivência e nas censuras, ou linearidade em todos os quantis (Wang and Wang, 2009).

O modelo de regressão quantílica com censura à direita foi generalizado para permitir a inclusão de preditores com relação não linear com a resposta, por meio de B-splines (Neocleous and Portnoy, 2009).

Em uma nova abordagem foi utilizado o método de particionamento recursivo (árvore de sobrevivência) para estimar os coeficientes da regressão, possibilitando maior flexibilidade para lidar com censuras que dependem das covariáveis(Wey et al.. 2013).

Para superar as limitações de métodos tradicionais na estimação do modelo de sobrevivência quantílico na presença de um número de parâmetros muito grande, ou com covariáveis correlacionadas e heterogeneidade individual não observada foram propostas duas abordagens, baseadas nos modelos de efeitos fixos da regressão quantílica com uma restrição de separação na probabilidade de censura. Simulações Monte Carlo foram conduzidas para avaliar os estimadores e indicaram boa performance em amostras finitas(Galvao et al., 2013).

Recentemente, Xue et al. (2016) utilizaram dados simulados e de estudos reais para discutir aspectos da interpretação das estimativas de parâmetros da regressão quantílica no contexto da pesquisa clínica e de saúde pública, e também uma comparação com o modelo de Cox. Como uma das vantagens em relação ao modelo de Cox desenvolveram uma metodologia para validar o tempo até o evento predito pelo modelo(Xue et al. 2016).

3. Discussão

Nos anos recentes foram desenvolvidas novas abordagens do modelo de regressão quantílica para a análise de dados censurados. Comparada aos métodos tradicionais, como o modelo de Cox e os modelos paramétricos, a regressão quantílica é menos restritiva, e possibilita estimar as medidas em diversos pontos ao longo do tempo.

Na pesquisa clínica e epidemiológica geralmente o principal interesse é estimar o impacto (efeitos ou associações) das covariáves sobre o tempo de sobrevivência. Neste sentido, uma única estimativa média pode não mostrar mudanças destas relações ao longo do tempo, perdendo-se parte da informação disponível nos dados. No contexto de um ensaio clínico randomizado para comparação de tratamentos, por exemplo, o efeito de tratamento pode diminuir ao longo do tempo, uma estimativa média do efeito de tratamento pode ser pouco informativa ou equivocada, haja vista que depende do tempo de acompanhamento dos indivíduose (Box-Steffensmeier et al., 2003; Hernán, 2010).

Os modelos paramétricos de análise de sobrevivência exigem o conhecimento da

distribuição de probabilidades para o tempo, que poucas vezes é conhecida na pesquisa clínica. No modelo de Cox, por sua vez, não exige o conhecimento da distribuição de probabilidades do tempo, pois os parâmetros são estimados considerando a ordem em que os eventos ocorrem, ignorando os reais valores observados para o tempo de sobrevivência. Também impõe proporcionalidade dos riscos ao longo do tempo, ou seja, se no início do estudo o indivíduo *i* tem um risco do evento igual a duas vezes o risco do indivíduo *j*, então esta razão de riscos será a mesma para todo o período de acompanhamento. Pode ser uma exigência restritiva e nem sempre verossímil na população da qual foram amostrados os dados(Hernán, 2010). Em ambas abordagens, nem sempre é trivial modelar relações não lineares com preditores quantitativos, e um modelo mal especificado gera estimativas viesadas para as associações ou confundimento residual.

A regressão quantílica para dados censurados pode ser uma poderosa alternativa de modelagens quando o modelo de Cox não for adequado, pois possibilida modelar a relação com as covariáveis em diferentes quantis do tempo de sobrevivência. Os métodos de estimação dos parâmetros do modelo quantílico têm sido constantemente aperfeiçoados e já são disponibilizados em diferentes programas de análise de dados. Cabe destacar as abordagens propostas por Portnoy (2003) e Peng & Huang (2008) no software SAS, e por Powell (1986), Portnoy (2003) e Peng & Huang (2008) no *software* R.

Uma limitação importante, no entanto, é que estes métodos estimam os parâmetros do modelo nos diferentes quantis da resposta, mas ainda não disponibilizam estimativas para as funções de sobrevivência e de risco, nem estimativas de efeito baseadas na razão de azares.

Por várias décadas o modelo de riscos proporcionais de Cox tem sido o método mais usado para analisar dados de sobrevivência. No entanto, a flexibilidade da regressão quantílica para acomodar relações não lineares e não impor proporcionalidade dos riscos ao longo do tempo, pode ser uma importante alternativa de modelagem para evitar inferências equivocadas baseadas em única estimativa média.

Referências

1. Allison PD. Survival Analysis Using SAS: A Practical Guide, Second Edition. SAS Institute; 2010.

2. Bilias Y, Chen S, Ying Z. Simple resampling methods for censored regression quantiles. J. Econom. 2000 Dezembro;99(2):373–86.

3. Box-Steffensmeier JM, Reiter D, Zorn C. Nonproportional Hazards and Event History Analysis in International Relations. J. Confl. Resolut. 2003;47(1):33–53.

4. Bustamante-Teixeira MT, Faerstein E, Latorre M do R. Survival analysis techniques. Cad. Saúde Pública. 2002 Jun;18(3):579–94.

5. Chernozhukov V, Hong H. Three-Step Censored Quantile Regression and Extramarital Affairs. J. Am. Stat. Assoc. 2002;97(459):872–82.

6. Fitzenberger B, Wilke RA. Using quantile regression for duration analysis. Allg. Stat. Arch. 2006 Mar 1;90(1):105–20.

7. Galvao AF, Lamarche C, Lima LR. Estimation of Censored Quantile Regression for Panel Data With Fixed Effects. J. Am. Stat. Assoc. 2013;108(503):1075–89.

8. Grambsch PM, Therneau TM. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. Biometrika. 1994;81(3):515–26.

9. Hernán MA. The Hazards of Hazard Ratios. Epidemiol. Camb. Mass. 2010 Jan;21(1):13–5.

10. Koenker R, Geling O. Reappraising Medfly Longevity: A Quantile Regression Survival Analysis. J. Am. Stat. Assoc. 2001;96(454):458–68.

11. Machado, José A.F., Portugal P. Exploring transition data through quantile regression methods: An application to US unemployment duration [Internet]. 2002 [cited 2014 Aug 31]. Available from: http://www.academia.edu/2956244/Exploring_transition_data_through_quantile_regression_methods_An_application_to_US_unemployment_duration

12. Neocleous T, Portnoy S. Partially linear censored quantile regression. Lifetime Data Anal. 2009 Sep;15(3):357–78.

13. Peng L, Huang Y. Survival Analysis With Quantile Regression Models. J. Am. Stat. Assoc. 2008;103(482):637–49.

14. Portnoy S. Censored Regression Quantiles. J. Am. Stat. Assoc. 2003;98(464):1001–12.

15. Powell JL. Least absolute deviations estimation for the censored regression model. J. Econom. 1984 Jul;25(3):303–25.

16. Powell JL. Censored regression quantiles. J. Econom. 1986 Jun;32(1):143–55.

17. Rodriguez, Robert N. L Guixian. Using the QUANTLIFE Procedure for Survival Analysis. 2013;

18. SAS. SAS Institute Inc. 2016. SAS/STAT® 14.2 User's Guide. Cary, NC: SAS Institute Inc. 2016;

19. Tobin J. Estimation of Relationships for Limited Dependent Variables. Econometrica. 1958 Jan;26(1):24.

20. Wang HJ, Wang L. Locally Weighted Censored Quantile Regression. J. Am. Stat. Assoc. 2009 Setembro;104(487):1117–28.

21. Wey A, Wang L, Rudser K. Censored quantile regression with recursive partitioning-based weights. Biostat. Oxf. Engl. 2013 Jan;15(1):170–81.

22. Xue X, Xie X, Strickler HD. A censored quantile regression approach for the analysis of time to event data. Stat. Methods Med. Res. 2016 May 10;

ARTIGO 2

Estimação de razão de azares por meio de regressão quantílica para dados com censura à direita: uma abordagem computacional

Estimation of hazard ratio by means of quantile regression for right censored data: a computational approach

Marina Bessel, Doutoranda em Epidemiologia pela UFRGS;

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL (UFRGS)

A ser enviado para “*Statistical Methods in Medical Research*”

1. Introdução

O modelo de riscos proporcionais de Cox (Cox,1972) ainda é o método mais utilizado na análise de dados de sobrevivência no contexto clínico e epidemiológico, por não exigir o conhecimento da distribuição de probabilidades do tempo. Apesar dessa flexibilidade, impõe proporcionalidade dos riscos ao longo do tempo, ou seja, se no início do estudo o indivíduo *i* tem um risco do evento igual a duas vezes o risco do indivíduo *j*, então esta razão de riscos será a mesma para todo o período de acompanhamento. Pode ser uma exigência restritiva e nem sempre verossímil na população da qual foram amostrados os dados(Hernán, 2010).

Diferentes mecanismos podem conduzir à violação da proporcionalidade dos riscos, como, por exemplo, aumento ou diminuição da razão entre as funções de risco (*increasing or decreasing relative hazards*), funções de risco cruzadas (*crossing hazards*), funções de risco divergentes (*diverging hazards*) ou funções de risco não monótonas (*non-monotonic hazards*)(Ng'andu,1997). A especificação incorreta do modelo, como violações na linearidade de preditores quantitativos e, talvez mais importante, efeitos de tratamentos que diminuem com o tempo também podem gerar riscos não proporcionais ao longo do tempo. Desse modo, é vital avaliar as suposições do modelo, em especial a suposição de riscos proporcionais, sob pena de obter estimativas viesadas da razão de azares. Em particular, no contexto em que o efeito de tratamento diminui ao longo do tempo, uma estimativa única de razão de azares pode ser inadequada, mesmo quando interpretada como estimativa média (Hernán, 2010; Harrell Jr., Frank E., 2015).

Entre as principais abordagens para avaliar a validade da suposição de riscos proporcionais podem ser citados os métodos gráficos (Schoenfeld,1982), testes de hipóteses baseados em resíduos ponderados (Grambsch and Therneau,1994), por meio de um modelo de Cox com variáveis tempo-dependentes, no qual são incluídos termos de interação entre o tempo e os preditores (Cox 1972; Stablein DM et al. 1981; Gore et al. 1984; Herndon and Harrell 1995); ou com o uso de funções splines, ferramenta bastante completa pois permite avaliar tanto a proporcionalide quanto a linearidade da covariável tempo dependente (Gray, 1992, 1994; Hess, 1994; Verweij and van Houwelingen, 1995; Abrahamowicz et al., 1996).

Uma abordagem recente é o uso de modelos de regressão quantílica para dados de sobrevivência. São métodos robustos e flexíveis, no sentido em que permitem entender como é a

relação dos preditores em diferentes quantis da distribuição do tempo. Pode ser vantajosa particularmente quando não estão atendidas as suposições de proporcionalidade de riscos e de linearidade(Rodriguez, Robert N., 2013; Xue et al., 2016)

O objetivo deste trabalho é investigar a utilização de regressão quantílica para dados de sobrevivência com censura à direita, propondo uma abordagem computacional para estimar razão de azares.

2. Métodos

O modelo de regressão quantílica para dados com censura à direita estima o quantil da distribuição do tempo de sobrevivência, condicional aos valores dos preditores. Ou seja, estima o valor do tempo de sobrevivência para cada quantil da distribuição, condicional aos preditores, mas é possível fazer inferências sobre a razão de azares a partir das estimativas do modelo.

Considerando T uma variável aleatória não negativa que representa o tempo de sobrevivência, o modelo quantílico para $log(T)$ é escrito como

$$Q_{log(T)}(\tau|\boldsymbol{x}) = \boldsymbol{x}'\beta(\tau), \quad (1)$$

em que $Q_{log\,T}(\tau|\boldsymbol{x})$ é o τ-ésimo quantil de $log(T)$ condicional no valor do vetor de preditores $\boldsymbol{x}$, e $\boldsymbol{\beta}$ é o vetor de parâmetros de regressão desconhecidos. A função quantílica é invariante sob uma transformação monótona, então o modelo pode ser escrito como

$$Q_T(\tau|\boldsymbol{x}) = exp\{\boldsymbol{x}'\boldsymbol{\beta}(\tau)\}.$$

A função de risco $h(t)$ é definida como $\frac{f(t)}{F(t)}$, em que $f(.)$ e $F(.)$ representam, respectivamente, a função densidade e a função de distribuição acumulada do tempo de sobrevivência. Se a função densidade $f(t)$ é desconhecida, a função de risco $h(t)$ não tem forma explícita. Assim, para estimar $h(t)$, as funções $f(t)$ e $F(t)$ precisam ser estimadas de forma não paramétrica. Para tanto, será utilizada a abordagem proposta por (Fitzenberger and Wilke, 2006), descrita por meio de um exemplo simulado, apresentada abaixo.

Geração dos dados

Utilizando o programa SAS, versão 9.4, o conjunto de dados do exemplo foi gerado considerando a matriz de correlação entre as variáveis Z_1 e Z_2

$$\rho = \begin{bmatrix} 1 & 0{,}7 \\ 0{,}7 & 1 \end{bmatrix}$$

a partir da qual foram geradas 5.000 observações independentes, seguindo os passos:

1) Foi gerada a matriz Z com dimensão (5.000×2), a partir da distribuição normal multivariada com vetor de médias $(0,0)$, variâncias $(1,1)$ e matriz de correlações ρ;

2) Foi definida a matriz $U(5.000 \times 2)$ por meio da transformação $U_j(Z_j) = \Phi(Z_j)$, $j = 1,2$, em que $\Phi(.)$ é a função de distribuição da normal padrão. Desse modo, as colunas de U são variáveis aleatórias com distribuição Uniforme em $[0,1]$, porém não são independentes;

3) Foi criada a matriz X (5.000×2), cujas colunas foram definidas por meio da transformação $X_j = F^{-1}(U_j)$, em que $F^{-1}(.)$ é a inversa da função de distribuição especificada: para X_1 foi usada a distribuição normal padrão e para X_2 foi usada a distribuição Uniforme em $[0,1]$;

4) Uma variável aleatória U, com distribuição uniforme em $[0,1]$ e independente de X_2 foi gerada para definir o preditor dicotômico GRUPO, por meio da dicotomização de X_2, tal que $GRUPO = 1$ se $X_2 \leq U$ e $GRUPO = 0$ em caso contrário;

5) O tempo de sobrevivência foi simulado, conforme descrito abaixo:

   a) Taxa em que ocorrem os eventos (*HazardRate* = 50);

   b) Taxa em que ocorrem as censuras (*CensorRate* = 75);

   c) Tempo máximo (*EndTime* = 150);

   d) Preditor linear, relacionando $GRUPO$ e X_1 com o tempo de sobrevivência: $LinPred = exp(0{,}2 * Grupo - 0{,}8 * x_1)$;

   e) O tempo do evento (*tEvent*) foi gerado utilizando a distribuição Weibull com parâmetros $\gamma = 3$ e $\delta = HazarRate * LinPred$;

   f) O tempo de censura (*c*) foi gerado utilizando a distribuição Weibull com parâmetros $\gamma = 2{,}5$ e $\delta = CensorRate$;

g) O tempo até ocorrência do evento ou censura (T) foi definido como $T = min(tEvent, c, EndTime)$, sendo posteriormente redefinido como o maior inteiro menor do que T, ou seja, $TEMPO = ceil(T)$;

h) A variável indicadora de censura foi definida como

$$CENSURA = \begin{cases} 1, & tEvent > c \; ou \; tEvent > 150 \\ 0, & caso \; contrário \end{cases}$$

6) Uma variável aleatória com distribuição uniforme em $[0,1]$ foi gerada para selecionar subamostras de tamanhos n=1000, n=2000 e n=5000 para os cenários das simulações.

Os dados simulados foram utilizados para estimar a razão de azares de ocorrência do evento, comparando o grupo 1 em relação ao grupo 0, ajustada pela variável x_1, por meio do modelo de Cox e do modelo de regressão quantílica para dados com censura à direita. O método proposto pode ser descrito nas seguintes etapas:

Etapa 1: Dividir o intervalo entre os quantis $\tau_I = 0{,}05$ e $\tau_S = 0{,}95$ do tempo de sobrevivência observado em M intervalos igualmente espaçados. Os quantis extremos $0 \leq \tau < \tau_I$ e $\tau_S < \tau \leq 1$ não são considerados devido às estimativas imprecisas nos extremos;

Etapa 2: Para cada τ_m; $m = 1,2,\cdots,M$, estimar o modelo de sobrevivência quantílico definido na equação (1), obtendo M vetores de estimativas dos parâmetros $\widehat{\boldsymbol{\beta}}(\tau_m)$;

Etapa 3: Para um dado valor do vetor de preditores $\boldsymbol{x}_j = (j, \bar{x}_1)$, o vetor M-dimensional dos tempos de sobrevivência simulados são obtidos por

$$T_m^* \equiv \hat{Q}_{T_i}(\tau_m|\boldsymbol{x}_j) = exp\{\boldsymbol{x}_j'\widehat{\boldsymbol{\beta}}(\tau_m)\},$$

com $m = 1,2,\ldots,M$, e $j = 0,1$ indica o correspondente grupo. Nesta etapa a média de x_1 foi usada para o ajuste multivariável;

Etapa 4: Baseado nas estimativas $\{T_m^*;\ m = 1,2,\ldots,M\}$, estimar a densidade condicional $f^*(t|\boldsymbol{x}_j)$ e a função de distribuição $F^*(t|\boldsymbol{x}_j)$. A densidade condicional $f^*(t|\boldsymbol{x}_j)$ foi estimada utilizando o método kernel de estimação de densidades do procedimento PROC KDE do programa SAS (SAS, 2016). Uma rotina computacional que utiliza o método de trapézio foi usada para integrar $f^*(t|\boldsymbol{x}_j)$ e obter os valores da função $F^*(t|\boldsymbol{x}_j)$ (Wicklin, 2011);

Etapa 5: Estimar a função de risco condicional em $\boldsymbol{x}_j$ no intervalo (τ_I, τ_S) dada por

$$\hat{h}_j(t) = \frac{(\tau_S - \tau_I)\, f^*\left(t \middle| \boldsymbol{x}_j\right)}{1 - \tau_I - (\tau_S - \tau_I)\, F^*\left(t \middle| \boldsymbol{x}_j\right)};$$

Etapa 6: Estimar a razão de azares do grupo 1 em relação ao grupo 0, usando

$$\widehat{HR}(t) = \frac{\hat{h}_1(t)}{\hat{h}_0(t)} = \frac{\dfrac{(\tau_S - \tau_I)\, f^*(t|\boldsymbol{x}_1)}{1 - \tau_I - (\tau_S - \tau_I)\, F^*(t|\boldsymbol{x}_1)}}{\dfrac{(\tau_S - \tau_I)\, f^*(t|\boldsymbol{x}_0)}{1 - \tau_I - (\tau_S - \tau_I)\, F^*(t|\boldsymbol{x}_0)}}$$

Etapa 7: Selecionar 500 amostras bootstrap da amostra original e, para cada uma delas, repetir as etapas (1) a (6) para estimar limites de confiança para $HR(t)$ baseados nos percentis 2,5% e 97,5% da distribuição empírica das respectivas estimativas das amostras bootstrap.

As etapas 6 e 7 não existiam no algoritmo original proposto por (Fitzenberger and Wilke, 2006). Foram considerados seis cenários para as simulações, de acordo com as combinações do tamanho de amostra (N=1000 e N=2000) sorteados da amostra de tamanho 5.000 e do número de quantis definidos no modelo (M=20 e M=30). No modelo de Cox a suposição de riscos proporcionais foi avaliada pelo teste disponível no procedimento PROC PHREG do programa SAS (Lin et al., 1993; SAS, 2016). A suposição de linearidade da variável x_1 foi avaliada pelo teste baseado nos quartis (Collett, 2015). As simulações foram realizadas em um computador da marca Dell, modelo Inspiron, processador Intel Core i7, memória de 8GB, sistema operacional 64bits.

3. Resultados

O tempo de simulação nos diferentes cenários variou entre 56,67 e 284,99 horas. A maior parte deste tempo de simulação é atribuída à etapa de estimação dos coeficientes de regressão do modelo, e cresceu com o aumento do número de quantis de M=20 para M=30. A Figura 1 mostra o histograma do tempo de sobrevivência simulado, por grupo, e o Quadro I mostra uma descrição da amostra variáveis x_1 e tempo, de acordo com os tamanho de amostra simulados $(n = 1.000, n = 2.000, n = 5.000)$, grupo e presença ou não de censura..

O Quadro II mostra as estimativas dos parâmetros do modelo de Cox e resultados dos testes para avaliar o atendimento da suposição de riscos proporcionais e de linearidade. Para todos os tamanhos de amostra, ao nível de significância de 5% a suposição de proporcionalidade dos riscos está atendida para a variável de grupo. , Para o preditor x_1, o pressuposto de proporcionalidade de riscos não é rejeitado para os tamanhos de amostra n=1000 e n=2000, mas é rejeitado para n=5000 (p=0,007). Os níveis descritivos amostrais associados ao teste da suposição de linearidade de x_1 para os tamanhos de amostra n=1000, n=2000 e n=5000 foram, respectivamente, p=0,0239, p=0,1155 e p =0,0018. As estimativas de razão de azares da comparação do Grupo 1 em relação ao Grupo 0 foram HR=0,59 (IC 95%: 0,50-0,70), HR=0,56 (IC 95%: 0,50-0,64) e HR=0,55 (IC95%: 0,51-0,59), respectivamente para os tamanhos de amostra n=1000, n=2000 e n=5000.

A Figura 2 mostra as estimativas dos coeficientes do modelo de regressão quantílica, em que se observa mudanças ao longo dos percentis do tempo, especialmente para o intercepto e para o coeficiente da variável grupo. O efeito de grupo, ajustado para a variável x_1, decresce suavemente ao longo do tempo de acompanhamento.

A Figura 3 mostra o comportamento das estimativas da função de risco $\hat{h}_0(t)$ e $\hat{h}_1(t)$ para os seis cenários simulados, e a Figura 4 mostra o comportamento das respectivas estimativas de razão de azares ao longo do tempo, e correspondentes limites de confiança de 95%. Valores selecionados destas estimativas são apresentados no Quadro III.

4. Discussão

Diferentemente do modelo de Cox, as simulações mostram que a estimativa de razão de azares (*HR-Hazard Ratio*) para a comparação dos grupos não é constante ao longo do tempo. No final do período de acompanhamento a estimativa de *HR* se aproxima do valor 1 e passa a ser não significativa, evidenciando uma diminuição do efeito de grupo. Este comportamento pode ser explicado pela diminuição do número de indivíduos em risco ao longo do tempo, devido à ocorrência de eventos e censuras, uma característica de praticamente todos os estudos de sobrevivência(Hernán, 2010).

Muitas vezes a estimativa de razão de azares é vista como objetivo principal de estudos clínicos ou epidemiológicos, mas sua interpretação como efeito causal pode ser arriscada mesmo

com um modelo corretamente especificado, na ausência de confundimento não observado e de erros de mensuração. No contexto de um ensaio clínico randomizado para comparação de tratamentos, por exemplo, se o efeito de tratamento diminui ao longo do tempo, então a razão de azares também varia no tempo. Sob estas circunstâncias, uma estimativa única de razão de azares representa uma média do possível efeito de tratamento, que depende do tempo de acompanhamento dos indivíduos, e pode ser pouco informativa ou equivocada(Box-Steffensmeier et al., 2003; Hernán, 2010).

A regressão quantílica aplicada para dados com censura a direita é uma técnica robusta e flexível, e é capaz de detectar mudanças no comportamento dos dados ao longo do tempo, muitas vezes negligenciadas por outros métodos. Não necessariamente usada de forma isolada, a regressão quantílica pode ser uma ferramenta associada a outras formas de análise quando estas têm suas suposições violadas.

As estimativas obtidas por meio da regressão quantílica podem ser úteis para fazer inferências mais verossímeis sobre o efeito de grupo, quando comparadas com a estimativa média de *HR* obtida pelo modelo de Cox. Além de mostrar uma redução do efeito de grupo ao longo do tempo, a regressão quantílica não é sensível à violação de suposição de linearidade da variável de confundimento (x_1) observada no modelo de Cox, no qual há potencialmente confundimento residual. No entanto, esta técnica não é muito explorada na área da saúde, possivelmente pela notação mais complexa, dificuldade em interpretar seu resultados ou ainda de validar suas estimativas(Xue et al. 2016).

Contudo, a abordagem para estimar a medida de *HR* é inédita, impossibilitando a comparação dos resultados. Algumas limitações do método computacional necessitam aperfeiçoamentos. Na etapa 3, os tempos de sobrevivência condicionais $T_m^* \equiv \hat{Q}_{T_i}(\tau_m|\boldsymbol{x}_j) = exp\{\boldsymbol{x}_j'\widehat{\boldsymbol{\beta}}(\tau_m)\}$ parecem instáveis nas caudas inferior e superior, possivelmente devido ao pequeno número de observações. Isto pode gerar instabilidade também na estimação da densidade do tempo de sobrevivência $f^*(t|\boldsymbol{x}_0)$ e da função de distribuição condicional $F^*(t|\boldsymbol{x}_0)$, na etapa 4. Também pode ser importante a avaliação e comparação de diferentes valores do parâmetro *bandwidth* na estimação da densidade $f^*(t|\boldsymbol{x}_0)$, bem como de outros métodos de estimação de densidade.

Foram usados dois cenários (M=20 e M=30) para o número de simulações de quantis do tempo de sobrevivência. O número excessivo de percentis pode aumentar o tempo de

processamento e, pelo mero acaso, também pode gerar inconsistências do tipo $T_s^* = \hat{Q}_{\tau_s}(\tau_s|\boldsymbol{x}_j) > \hat{Q}_{\tau_r}(\tau_r|\boldsymbol{x}_j)$, mesmo quando $\tau_s < \tau_r$.(Fitzenberger and Wilke, 2006). Considerando a amostra de tamanho 1000, foi observado um aumento de aproximadamente 17,5% no tempo de processamento comparando o uso de 30 percentis (M=30) em relação a relação a 20 percentis (M=20), porém não houve diferenças nas estimativas pontuais e na precisão.Além destes aspectos, o algoritmo pode ser estendido por meio da inclusão de um número maior de covariáveis.

Em conclusão, a abordagem proposta tem potencialidades para modelar dados de sobrevivência com censura à direita para contexto em que a forma funcional do preditor é complexa ou quando a suposição de riscos proporcionais não está atendida, permitindo assim estimar a medida de razão de azares não constante no tempo.

Referências

1. Abrahamowicz M, MacKenzie T, Esdaile JM. Time-Dependent Hazard Ratio: Modeling and Hypothesis Testing With Application in Lupus Nephritis. J. Am. Stat. Assoc. 1996;91(436):1432–9.

2. Box-Steffensmeier JM, Reiter D, Zorn C. Nonproportional Hazards and Event History Analysis in International Relations. J. Confl. Resolut. 2003;47(1):33–53.

3. Collett D. Modelling Survival Data in Medical Research, Third Edition. CRC Press; 2015.

4. Cox DR. Regression Models and Life-Tables. J. R. Stat. Soc. Ser. B Methodol. 1972;34(2):187–220.

5. Fitzenberger B, Wilke RA. Using quantile regression for duration analysis. Allg. Stat. Arch. 2006 Mar 1;90(1):105–20.

6. Gore SM, Pocock SJ, Kerr GR. Regression Models and Non-Proportional Hazards in the Analysis of Breast Cancer Survival. J. R. Stat. Soc. Ser. C Appl. Stat. 1984;33(2):176–95.

7. Grambsch PM, Therneau TM. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. Biometrika. 1994;81(3):515–26.

8. Gray RJ. Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis. J. Am. Stat. Assoc. 1992;87(420):942–51.

9. Gray RJ. Spline-Based Tests in Survival Analysis. Biometrics. 1994;50(3):640–52.

10. Harrell Jr., Frank E. Regression Modeling Strategies - With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis [Internet]. 2015 [cited 2017 Mar 29]. Available from: http://www.springer.com/br/book/9783319194240

11. Hernán MA. The Hazards of Hazard Ratios. Epidemiol. Camb. Mass. 2010 Jan;21(1):13–5.

12. Herndon JE, Harrell FE. The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariables. Stat. Med. 1995 Outubro;14(19):2119–29.

13. Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. Stat. Med. 1994 Maio;13(10):1045–62.

14. Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika. 1993 Sep 1;80(3):557–72.

15. Ng'andu NH. An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. Stat. Med. 1997 Mar 30;16(6):611–26.

16. Rodriguez, Robert N. L Guixian. Using the QUANTLIFE Procedure for Survival Analysis. 2013;

17. SAS. SAS Institute Inc. 2016. SAS/STAT® 14.2 User's Guide. Cary, NC: SAS Institute Inc. 2016;

18. Schoenfeld D. Partial Residuals for The Proportional Hazards Regression Model. Biometrika. 1982;69(1):239–41.

19. Stablein DM et al. Analysis of survival data with non-proportional hazard functions [Internet]. ResearchGate. 1981 [cited 2017 Mar 4]. Available from: https://www.researchgate.net/publication/15924893_Analysis_of_survival_data_with_non-proportional_hazard_functions

20. Verweij PJM, van Houwelingen HC. Time-Dependent Effects of Fixed Covariates in Cox Regression. Biometrics. 1995;51(4):1550–6.

21. Wicklin. The Area under a Density Estimate Curve: Nonparametric Estimates [Internet]. 2011 [cited 2017 Mar 7]. Available from: http://proc-x.com/2011/07/the-area-under-a-density-estimate-curve-nonparametric-estimates/

22. Xue X, Xie X, Strickler HD. A censored quantile regression approach for the analysis of time to event data. Stat. Methods Med. Res. 2016 May 10;

Quadros e Figuras

Quadro I - Descrição das variáveis x_1 e tempo, de acordo com os tamanhos de amostra simulados, grupo e presença ou não de censura.

Tamanho de amostra	Grupo	Censura	Frequência	Variável	Mínimo	Mediana	Máximo
n=1000	0	Não	367	x_1	-1,33	0,63	3,38
				Tempo	3,00	25,00	104,00
	1	Sim	253	x_1	-1,38	0,20	2,41
				Tempo	6,00	40,00	138,00
	0	Não	105	x_1	-2,50	-0,49	0,98
				Tempo	8,00	44,00	108,00
	1	Sim	275	x_1	-3,72	-0,90	0,77
				Tempo	14,00	56,00	134,00
n=2000	0	Não	745	x_1	-1,43	0,64	3,38
				Tempo	2,00	24,00	105,00
	1	Sim	468	x_1	-1,55	0,22	2,42
				Tempo	5,00	39,00	138,00
	0	Não	228	x_1	-2,50	-0,49	1,01
				Tempo	8,00	44,50	150,00
	1	Sim	559	x_1	-3,72	-0,88	1,29
				Tempo	7,00	55,00	150,00
n=5000	0	Não	1896	x_1	-1,73	0,65	3,38
				Tempo	2,00	25,00	113,00
	1	Sim	626	x_1	-2,66	-0,48	1,29
				Tempo	5,00	39,00	138,00
	0	Não	1111	x_1	-1,73	0,21	2,62
				Tempo	7,00	44,00	150,00
	1	Sim	1367	x_1	-3,72	-0,89	1,31
				Tempo	3,00	55,00	150,00

Quadro II – Estimativas dos parâmetros e respectivos testes de Wald para o modelo de Cox, teste da suposição de riscos proporcionais e da linearidade do preditor x_1, de acordo com os tamanhos de amostra simulados.

Tamanho de Amostra	Variável	$\hat{\beta}$	Erro padrão	p[1]	Riscos proporcionais (p)	Linearidade[2] (p)
n=1000	Grupo	-0,57	0,06	< 0,0001	0,543	-
	x_1	2,41	0,06	< 0,0001	0,130	0,0239
n=2000	Grupo	-0,57	0,09	< 0,0001	0,135	-
	x_1	2,31	0,08	< 0,0001	0,329	0,1155
n=5000	Grupo	-0,60	0,04	< 0,0001	0,290	-
	x_1	2,37	0,04	< 0,0001	0,007	0,0018

[1] Teste de Wald $(H_0: \beta = 0 \times H_1: \beta \neq 0)$

[2] Baseado nos quartis de x_1

Quadro III – Estimativas de *hazard ratio* (HR) da ocorrência do evento, para valores selecionados do tempo de sobrevivência estimado, considerando cenários de simulações definidos de acordo com o tamanho de amostra (N) e número de quantis da distribuição do tempo de sobrevivência (M).

Tempo estimado[1]	N=1000				N=2000				N=5000			
	M=20		M=30		M=20		M=30		M=20		M=30	
	HR[2]	95% CI[3]	HR[2]	95% CI[3]	HR[2]	95% CI[3]	HR[2]	95% CI[3]	HR[2]	95% CI[3]	HR[2]	95% CI[3]
25	0,66	0,51-0,80	0,65	0,48-0,82	0,58	0,49-0,67	0,55	0,46-0,66	0,56	0,51-,61	0,53	0,48-0,59
30	0,64	0,53-0,76	0,63	0,49-0,77	0,58	0,51-0,67	0,57	0,49-0,66	0,56	0,52-0,61	0,54	0,50-0,59
35	0,61	0,50-0,71	0,60	0,47-0,72	0,59	0,53-0,66	0,58	0,51-0,65	0,57	0,53-0,62	0,56	0,51-0,60
40	0,60	0,49-0,71	0,58	0,47-0,71	0,60	0,53-0,68	0,59	0,52-0,67	0,58	0,54-0,62	0,57	0,52-0,61
45	0,59	0,48-0,71	0,56	0,45-0,68	0,61	0,54-0,69	0,61	0,52-0,68	0,59	0,55-0,64	0,58	0,53-0,63
50	0,59	0,47-0,71	0,57	0,44-0,70	0,63	0,56-0,71	0,62	0,54-0,69	0,61	0,56-0,66	0,60	0,54-0,60
55	0,62	0,48-0,79	0,60	0,46-0,77	0,66	0,58-0,76	0,64	0,56-0,74	0,64	0,59-0,70	0,62	0,57-0,69
60	0,68	0,50-0,85	0,66	0,48-0,84	0,70	0,62-0,86	0,69	0,60-0,86	0,69	0,63-0,77	0,68	0,62-0,76
65	0,80	0,58-1,03	0,80	0,58-1,12	0,78	0,68-0,96	0,77	0,67-0,96	0,78	0,70-0,89	0,78	0,69-0,90
70	-	-	-	-	0,93	0,81-1,13	0,94	0,81-1,18	0,95	0,85-1,11	0,97	0,85-1,15

[1] Tempo de sobrevivência estimado pelo método de estimação de densidade, ajustado no valor médio da variável de confundimento x_1.

[2] Comparando o Grupo 1 em relação ao Grupo 0.

[3] Limites de confiança baseados nos percentis 2,5% e 97,5% da distribuição da estimativa de HR utilizando 500 amostras bootstrap.

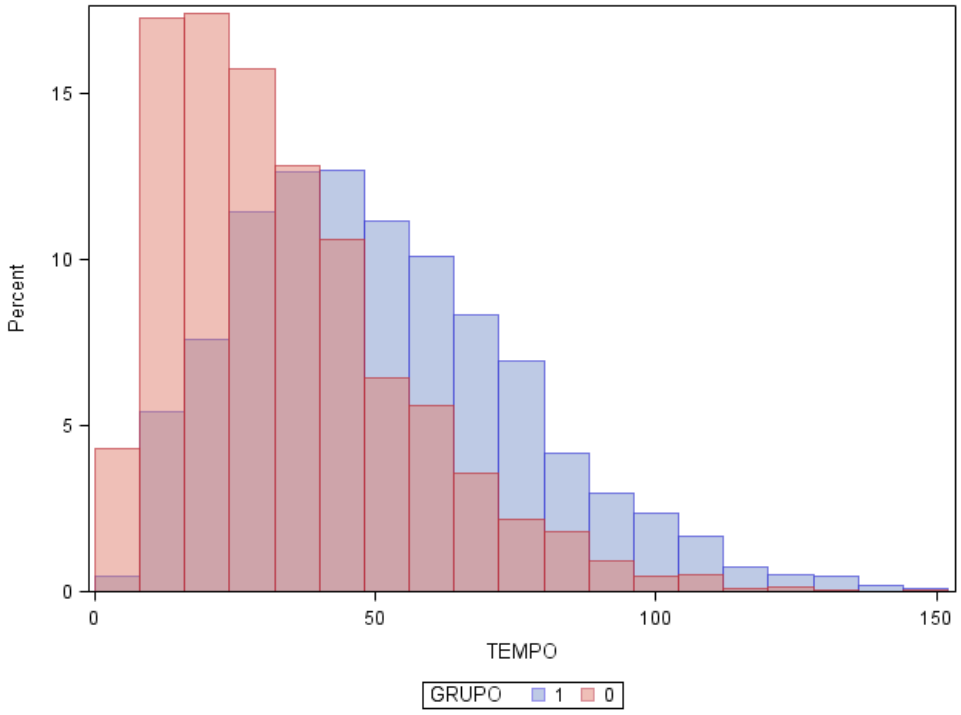


Figura 1 – Distribuição empírica do tempo de sobrevivência simulado, por grupo (n=5000).

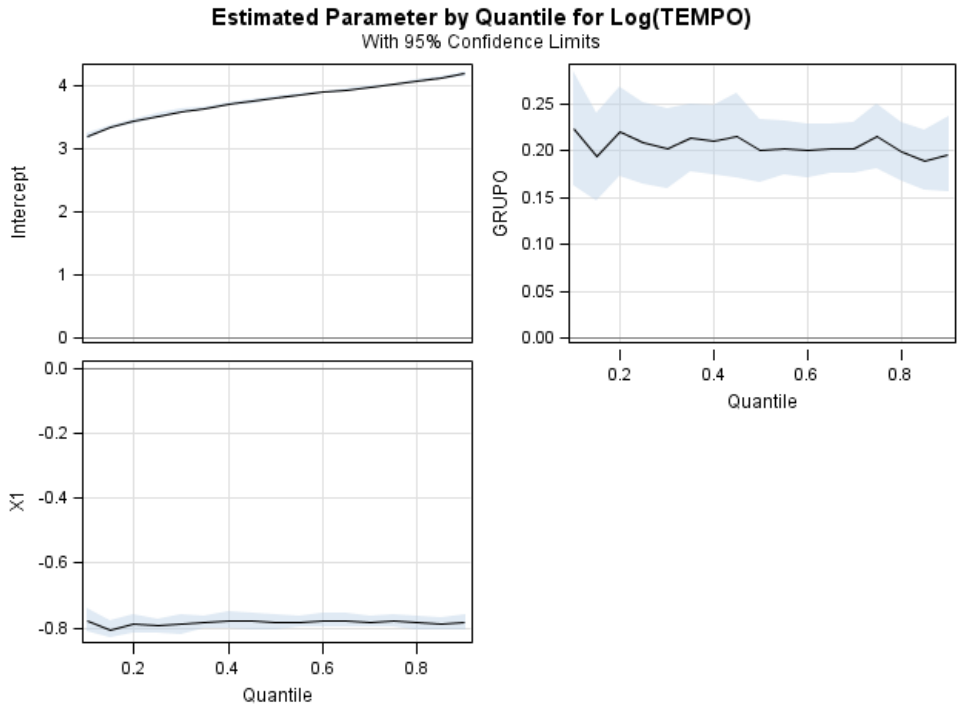
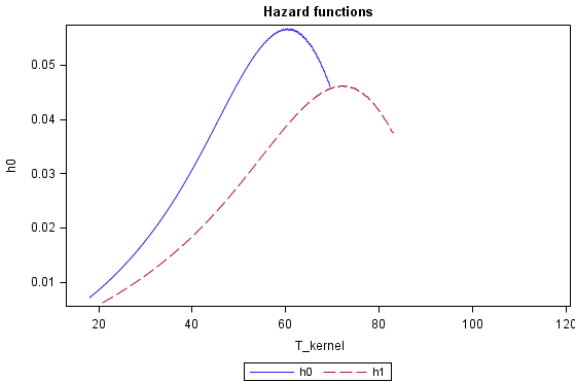
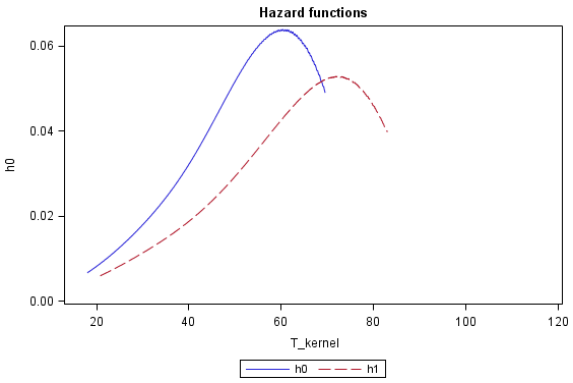


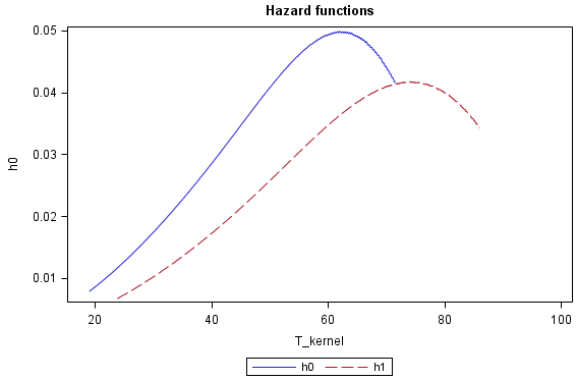
Figura 2 – Estimativas dos parâmetros do modelo de regressão quantílica com censura à direita, ao longo dos percentis do tempo de acompanhamento.



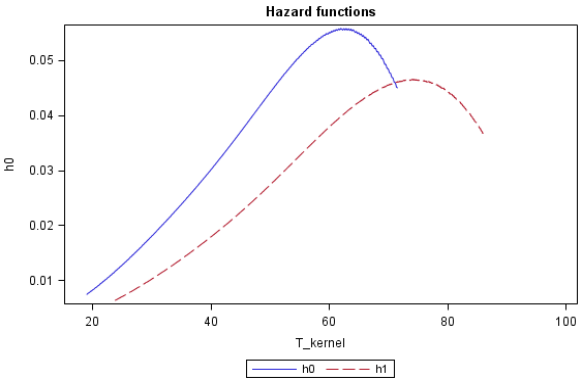
(a)



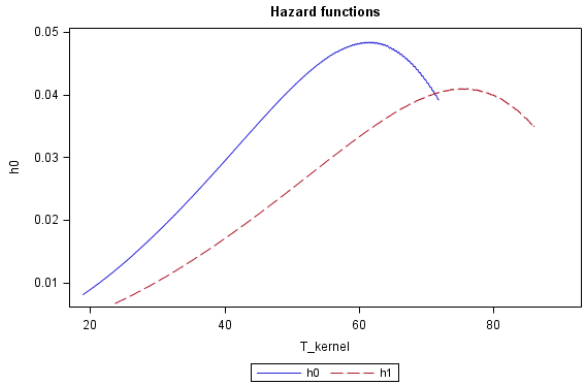
(b)



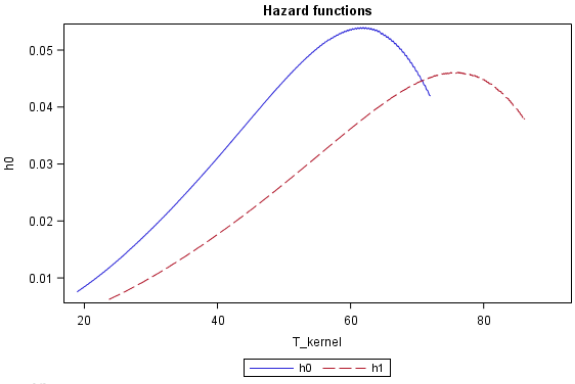
(c)



(d)

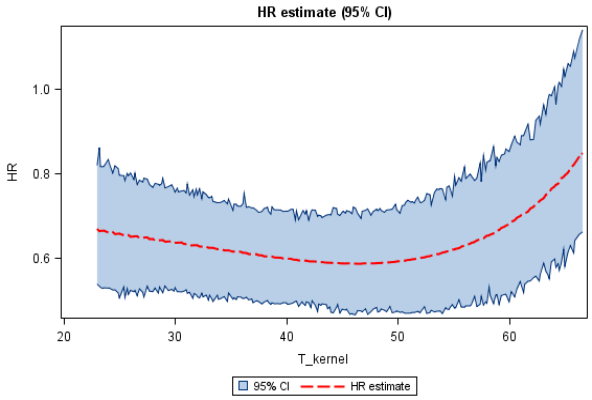


(e)

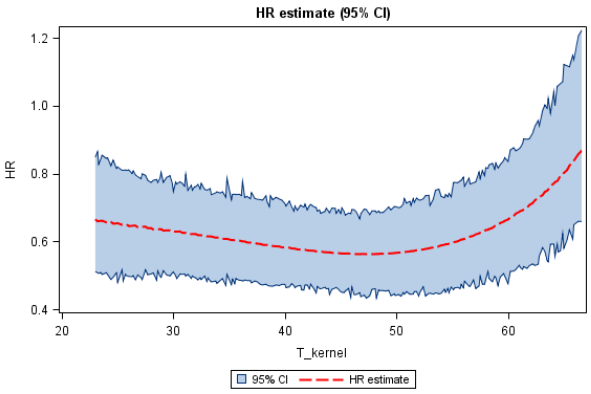


(f)

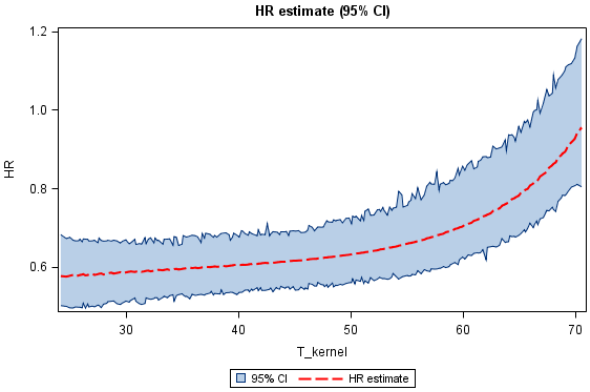
Figura 3 – Funções $\hat{h}_0(t)$ e $\hat{h}_1(t)$ estimadas pelo método proposto, para os diferentes cenários de simulações: (a) n=1000 e M=20, (b) n=1000 e M=30, (c) n=2000 e M=20, (d) n=2000 e M=30, (e) n=5000 e M=20; e, (f) n=5000 e M=30.



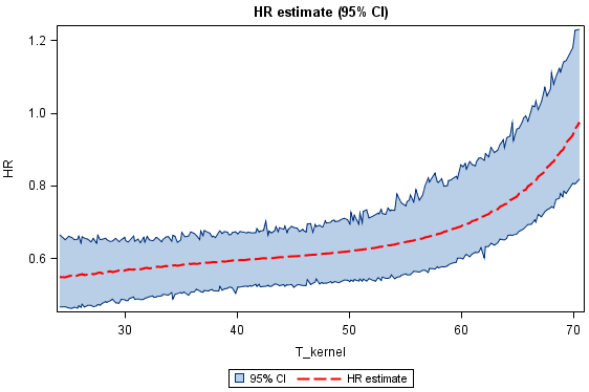
(a)



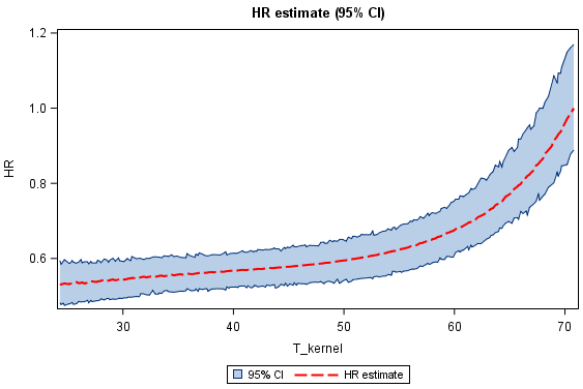
(b)



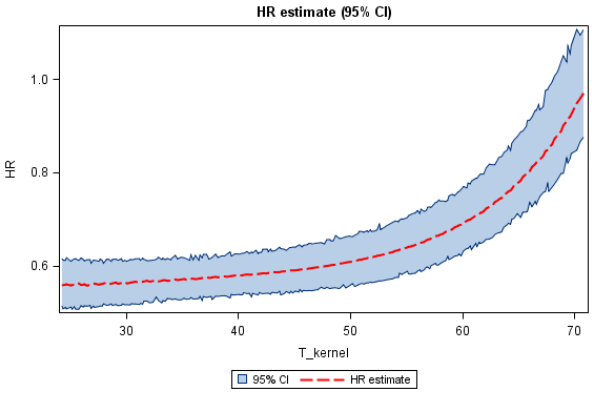
(c)



(d)



(e)



(f)

Figura 4 – Razão de azares (HR) e IC 95% estimadas pelo método proposto, para os diferentes cenários de simulações: (a) n=1000 e M=20, (b) n=1000 e M=30, (c) n=2000 e M=20, (d) n=2000 e M=30, (e) n=5000 e M=20; e, (f) n=5000 e M=30.

CONCLUSÕES E CONSIDERAÇÕES FINAIS

A grande maioria dos trabalhos sobre regressão de sobrevivência quantílica aborda apenas aspectos da estimação dos parâmetros do modelo, associados às covariáveis, nos diferentes quantis do tempo. É preciso ressaltar a importância de todas as técnicas desenvolvidas ao longo dos anos para estimar os parâmetros do modelo de regressão quantílica para que respondam questões de maior abrangência em análise de sobrevivência e para lidar com os diferentes tipos de censura. No contexto epidemiológico, no entanto, frequentemente o objetivo é estimar o efeito (ou associação) de uma determinada exposição sobre o tempo até a ocorrência do evento, e o modelo de riscos proporcionais de Cox ainda é o a abordagem dominante, mesmo havendo suposições (muitas vezes) restritivas que devem ser atendidas.

Quando não há certeza sobre a estrutura global para a função de sobrevivência, uma abordagem alternativa é o uso da regressão quantílica. É um método robusto a presença de outliers e capaz de modelar dados com heteroscedasticidade, permitindo ainda medir o efeito das covariáveis em diferentes quantis do tempo de sobrevivência. Apenas Fitzenberger and Wilke (2006) investigaram a estimação da função de risco (*hazard rate*) a partir das estimativas do modelo de regressão quantílica.

Os resultados do estudo de simulação mostraram que a suposição de risco constante ao longo do tempo do modelo de Cox pode ser muito restritiva ou irreal. Em situações práticas, isto ocorre nos casos em que o efeito de um tratamento diminui ao longo do tempo e, portanto, inferências baseadas em uma única estimativa de risco médio podem ser falsas, ao menos para partes do período de acompanhamento. Também pode ocorrer como consequência natural da distribuição dos eventos ou censuras ao longo do tempo, reduzindo o número de indivíduos ainda em observação, podendo gerar alterações na relação funcional das covariáveis em diferentes segmentos da distribuição do tempo. Assim, a modelagem dos quantis do tempo pode ser mais flexível para capturar essas mudanças da relação funcional.

No nosso conhecimento este trabalho é o primeiro a usar o modelo de regressão quantílica com dados censurados à direita para estimar razão de azares e pode ser um início promissor para desenvolvimento de trabalhos futuros. Algumas limitações precisam ser mencionadas, as quais podem ser aspectos para continuidade da pesquisa e aprimoramento do método. Por exemplo, é necessário avaliar e comparar diferentes abordagens para a estimação não paramétrica da

densidade e da função de distribuição do tempo de sobrevivência e, por consequência, da função de risco acumulado. A abordagem também necessita ser comparada com outros métodos, como modelos paramétricos ou modelos com splines. A comparação empírica dos resultados aplicado em dados reais, analisados por meio de diferentes métodos, também pode produzir informações importantes. A extensão do método para outros tipos de censura é outro aspecto de possível desenvolvimento.

APÊNDICES – Rotinas Computacionais

APÊNDICE A - Geração dos dados da simulação

APÊNDICE B - Estimação da razão de azares para diferentes cenários

APÊNDICE C - Bootstrap para estimação do intervalo de confiança

APÊNDICE A - Geração dos dados da simulação

```
ods pdf file="C:\Doutorado\Projeto\Simulacao\2017MAR17\Dados\Simulacao_Dados_Cenario_170317.pdf";

options ps=58 ls=80 nocenter nodate nonumber formchar='|----|+|---+=|-/\<>*';

libname L1 'C:\Doutorado\Projeto\Simulacao\2017MAR17\Dados';


%macro Preditores(n=, seed1=, corr1=);
* Gera distribuicao dos preditores correlacionados;
    proc iml;
        call randseed(&seed1);

        * Especifica a matriz de correlacoes da distribuicao normal multivariada
          com medias 0 e variancias 1;
        sigma = {1.0 &corr1.,
                 &corr1. 1.0};
        *print sigma;


        * Gera nj observacoes da normal multivariada;
        Z = randnormal(&n, {0,0}, sigma);
        *print z;


        * Se F a f.d. da v.a. X, entao U=F(X)~U(0,1)
          a funcao 'cdf' aplica a f.d. da v.a.;
        U  = cdf("normal", Z); * As colunas de U sao v.a. U(0,1), mas nao sao ind.;


        * Se U~U(0,1), entao X=invF(U) ~ F;
        X1 = quantile("Normal", U[,1],0,1);       /* X1 ~ Normal   */
        X2 = quantile("Uniform", U[,2],0,1);      /* X2 ~ Uniform  */

        X = X1||X2;

        /* Se Z ~ MVN(0,Sigma), corr(X) geralmente é proxima de Sigma,
           em que X=(X1,X2,...,Xm) and X_i = F_i^{-1}(Phi(Z_i)) */
        varNames = ("X1":"X2");

        create XX from X[c=varNames];
        append from X;
        close XX;
    run;
    quit;

    data Xx;
        retain ID GRUPO X1 X2;
        set Xx;

        U = uniform(76849);
        GRUPO = (X2 <= U);

        ID = _N_;
    run;
%mend;

%Preditores(n=5000, seed1=31, corr1=0.7);
```

```
*Com censura;
data TEMPO;
    set Xx;

    call streaminit(10);
    HazardRate = 50;                    * Rate at which subject
experiences event;
    CensorRate = 75;                    * Rate at which subject drops out;
    EndTime = 150;                      * End of study period (days);

        *linpred = exp(0.3*Grupo - 1.0*X1 - 0.8*X2);
        linpred = exp(0.2*Grupo - 0.8*X1);
        tEvent = rand("WEIBULL", 3.0, HazardRate*linpred);
          * Time of event;
        rate = HazardRate*linpred;
        c = rand("WEIBULL", 2.5, CensorRate);
          * Time of censoring;
        T = min(tEvent, c, EndTime);
        CENSURA = (c < tEvent | tEvent > EndTime);

        TEMPO = ceil(T);
        SampleSelection = uniform(877887);
run;


proc sgplot data=TEMPO;
    histogram TEMPO / group=GRUPO transparency=0.5;
run;


proc means data=TEMPO n min median max;
    class CENSURA GRUPO;
    var X1 TEMPO;
run;


data CENARIO1;
    retain ID TEMPO CENSURA GRUPO X1 X2;
    set TEMPO;

    label TEMPO = "Tempo ate o evento (meses)";
    label CENSURA = "1=Censura, 0=Evento";
    keep ID TEMPO CENSURA GRUPO X1 SampleSelection;
run;


proc format;
    value grupof 1="Tratamento" 0="Placebo";
run;



proc freq data=CENARIO1;
    table GRUPO*CENSURA / nopercent nocol;
    format GRUPO grupof. CENSURA censf.;
run;
```

```
options ls=120;
proc means data=CENARIO1 min mean std max;
    class CENSURA;
    var X1 TEMPO;
run;


proc phreg data=CENARIO1;
    model TEMPO*CENSURA(1)= Grupo / rl ;
run;


proc phreg data=CENARIO1;
    model TEMPO*CENSURA(1)= Grupo X1 / rl ;
run;



proc quantlife data=CENARIO1 method=na log plot=quantplot;
    title1 "Quantile survival model (log Time)";
    model TEMPO*CENSURA(1)= Grupo / quantile=0.1 to 0.9 by 0.05;
run;


proc quantlife data=CENARIO1 method=na log plot=quantplot;
    title1 "Quantile survival model (log Time)";
    model TEMPO*CENSURA(1)= Grupo X1 / quantile=0.1 to 0.9 by 0.05;
    *baseline out=qlout covariates=PredAtX2 /*survival=qlsurv quantile=qly*/;
run;


* Ordena pelo SampleSelection;
proc sort data=CENARIO1; by SampleSelection; run;
data L1.Dados_Cenario_170317;
    set CENARIO1;
run;


ods pdf close;
```

APÊNDICE B - Estimação da razão de azares para diferentes cenários

```
options ps=58 ls=80 nocenter nodate nonumber formchar='|----|+|---+=|-/\<>*';
*options mprint symbolgen mlogic;


*******************************************************************************
     *****************
                                        Step 0

- Select (at random) a subset (n <= 5000) of the original sample
- Define simulation parameter M
- Define the filename suffix for simulation scenary

- Fit a Cox model for the original subsample only (not for bootstrap samples)
- Specify the number of bootstrap repetitions
*******************************************************************************
     *****************;
%global SubamostraN M Sscenaryname Bootrep NMissingValue basicpath folderdate;

%let SubamostraN = 1000;                    /* Select at random a subsample of
     size N           */
%let M = 30;                                /* # of time estimates defined by
     the percentiles of
                                            /*   time distribution (Taus) - See
     Step 1          */

%let Scenaryname  = N&SubamostraN._M&m.;     /* Filename suffix for the
     parameters estimates and */
                                            /* simularion results files
     */
%let Bootrep = 500;                         /* Number of bootstrap sample used
     to get 95% CI    */
%let NMissingValue = 300;                   /* Specify a minimum valid HR
     estimatives for each  */
                                            /* T_Kernel time in the bootstrap
     samples, in oreder*/
                                            /* to get the confidence interval
     */

%let basicpath = E:\Doutorado\Projeto\Simulacao;
%let folderdate = 2017MAR17;
data StartTime;
     Start_Datetime = datetime();
     put Start_Datetime = datetime18.;
     format Start_Datetime datetime18.;
     ONE = 1;
run;
%put _global_;


proc printto
     log="&basicpath.\&folderdate.\N&SubamostraN.\M&m.\Cenario_N&SubamostraN._M&
     m..log"; run;
proc printto
     print="&basicpath.\&folderdate.\N&SubamostraN.\M&m.\Cenario_N&SubamostraN._
     M&m..lst";run;
```

```
%let pathname = "&basicpath\&folderdate.\N&SubamostraN.\M&m.\";
libname Simul V9 "&basicpath.\&folderdate.\N&SubamostraN.\M&m.\";
libname Dados V9 "&basicpath.\&folderdate.\Dados\";
%include "&basicpath.\SASMacros\Bootstrap.sas";

ods pdf
     file="&basicpath.\&folderdate.\N&SubamostraN.\M&m.\Cenario_N&SubamostraN._M
     &m..pdf";


%macro Cenario;
     *Select (at random) a subset (n <= 5000) of the original sample;
     data Subset;
           if _n_ <= &SubamostraN.;
           set Dados.dados_cenario_170317;
     run;
%mend;
%Cenario;


* Cox model for the original (subset) sample;
proc phreg data=Subset;
    model TEMPO*CENSURA(1)= Grupo X1 / rl ;
     assess PH / resample;
run;




* Testing linearity assumption for X1;
%include "&basicpath.\SASMacros\LinTest_Macro_080217.sas";
%LinTest(sasproc=PHREG, datain=Subset, outcome=TEMPO, event=1, censor=CENSURA,
     testvar=X1, catvars=GRUPO, quantvars=);


*************************************************************************
                                        Step 1

 - Compute the M values for Taus in (0.05, 0.95)
*************************************************************************;
%global M Tau_Min Tau_Max byvalue;
%let Tau_Min = 0.05;
%let Tau_Max = 0.95;
%let byvalue = %sysevalf((&Tau_Max - &Tau_Min)/ (&M. - 1));
%put _global_;
```

```
*************************************************************************
                                  Etapa 2
   Para todo Taum, (m=1,2,...,M), estima o modelo de regressao quantilica de
   sobrevivencia Qlog(T)(Taum|x) = x'Beta
*************************************************************************;
ods output Quantlife.ParameterEstimates=Coefs;
proc quantlife data=subset method=na log;
     title1 "Quantile survival model (log Time)";
     model TEMPO*CENSURA(1)= Grupo X1 / quantile =&Tau_min to &Tau_Max by
     &byvalue;
run;
title1;

* Save betas;
data Simul.Original_Parameters_&scenaryname.;
     set Coefs;
run;


*************************************************************************
                                  Step 3
For each value of the predictor vector x_0, the M-Dimensional vector containing
the simulated survival times are obtained by means of

      T_m = Q_Ti(Tau_m|x_0) = exp{x'_0*Beta(Tau_m)}, for each m=1,2,...,M)
*************************************************************************;
* Means of the model predictor for multivariable adjustment;
proc means data=Subset mean noprint ;
     var GRUPO X1;
     output out=Means;
run;
data Means;
     set Means;
     if _STAT_ = "MEAN";
     keep GRUPO X1;
run;


%macro Ts_estimates;
     data Coefs;
           set Coefs;
           LoopControl = int(Quantile*10000);
     run;

     /*Estima T_m = Q_Ti(Tau_m|x_0) = exp{x'_0*Beta(Tau_m)}, para m=1,2,...,M)*/
     %do mm = 1 %to &M.;

           data Coefs&mm.(keep=/*Quantile*/ Parameter Estimate /*LoopControl
     M*/);
           set Coefs;
           if LoopControl = (int((&Tau_Min + (%sysevalf(&mm.) -
     1)*&byvalue)*10000));
           run;

     proc transpose data=Coefs&mm. out=Betas&mm.; run;
        data Betas&mm.;
                set Betas&mm.;

                rename COL1 = Intercept;
                rename COL2 = Grupo;
                rename COL3 = X1;
                drop _NAME_;
```

```
run;

proc iml;
      xvar={Intercept Grupo X1};
      xxvar={Grupo X1};
      use Betas&mm.;
            read all var xvar into Betas;
      close;
      *print Betas;

      Betas=Betas`;
      *print Betas;

      * Vetor com medias dos preditores;
      use Means;
            read all var xxvar into Means;
      close;
      *print Means;

 nparms = nrow(Betas);
      Um = j(1,1);
      *print Um nparms;

/* Define Categoria de "Referencia" para grupo, ajustado por X1 e X2 */
      Grupo0 = Means;
      Grupo0[1]  = 0;
      Grupo0 = Um || Grupo0;
      Grupo0 = Grupo0`;
      *print Grupo0;

 /* Define Categoria de "Exposicao" para grupo, ajustado por X1 e X2 */
      Grupo1 = Means;
      Grupo1[1]  = 1;
      Grupo1 = Um || Grupo1;
      Grupo1 = Grupo1`;
      *print Grupo1;


      * Etapa 3 - Calcula Tm = Quantil(Tauj | x0) = exp(Beta*Grupo);
   Ts = j(1,4,0);
      *print Ts;

   Ts[1,1] = %sysevalf(&mm.);
   *Ts[1,2] = %sysevalf(&Tau&mm.);
   Ts[1,2] = &Tau_Min + (%sysevalf(&mm.) - 1)*&byvalue;


      if Betas[1] > .Z & Betas[2] > .Z & Betas[3] > .Z
      then do;
                  T0 = exp(Betas`*Grupo0);
                  T1 = exp(Betas`*Grupo1);
                  *print T0 T1;
            end;
      else do; /* Missing values for time when they can't be estimate
due lack information for the respective percentiles*/
      T0 = .; T1 = .;
            end;

   Ts[1,3] = T0;
   Ts[1,4] = T1;
   *print "Valor Ts (M=&mm.)" Ts;
```

```
            varNames = ("TAUM"||"QUANTILE"||"T0"||"T1");
            CREATE T&mm. FROM Ts[colname=varnames];
             APPEND FROM Ts;
             CLOSE T&mm.;
                  quit;
            run;
     %end;
%mend;
%Ts_estimates;



* Dataset containing Ts estimates;
%macro Append;
     data T;
           set T1;
     run;
     %do mm=2 %to &M.;
          proc append
              BASE=T data=T&mm.;
          run;
     %end;
%mend;
%Append;


* Exclui datasets desnecessarios;
proc datasets lib=work nolist;
     delete Catmodel Ordinalmodel Lintest Coefs1-Coefs&M. Betas1-Betas&M. T1-
     T&M.;
run;
quit;


* Grava dados das Etapas 2 e 3;
data Simul.TemposEstimados_&scenaryname.;
     set T;
run;
```

```
*************************************************************************
                                  Step 4
The Kernel Density Estimatin method is applied in order to obtain estimatives of

                              f(t|x_0) and F(t|x_0)

associated to the estimated survival times {T_m, m=1,1,2,...,M}.
*************************************************************************;
%macro Integration;
     proc iml;
          use DENSITY;
               /*read all var {Value} into x;*/
               read all var {T} into x;
               read all var {Density} into y;
          close density;

          start TrapIntegral(x,y);
               N = nrow(x);
               dx    =   x[2:N] - x[1:N-1];
               meanY = ( y[2:N] + y[1:N-1] )/2;
               return( dx` * meanY );
          finish;

          /** compute the integral over all x **/
          Area = TrapIntegral(x,y);
          *print Area;


          /** Compute the integral for each X <= x **/
               *print x;
          N = nrow(x);
          Fcum = J(N,1,0)`;

          do k = 3 to N;
            idx = loc(x <= x[k]);
               *print idx;
               itx = TrapIntegral(x[idx],y[idx]);
               *print itx;
               Fcum[k] = itx;
          end;

          Fcum = Fcum`;
          *print X Fcum;

          &tvar._kernel = x || y || Fcum;
          *print &tvar._kernel;

     varNames = ("T_kernel"||"Density_&tvar."||"Fcum_&tvar.");
          CREATE &tvar._kernel FROM       &tvar._kernel[colname=varnames];
          APPEND FROM &tvar._kernel;
        CLOSE &tvar._kernel;
          quit;
     run;
     quit;
%mend;

%macro Step_4(tvar=, datain=);
     data T_&tvar.;
          set Work.&datain.;
          T_&tvar. = int(&tvar.*10)/10;
     run;
```

```
     proc kde data=T_&tvar.;
          univar T_&tvar. / method=SJPI plots=(density) out=density;
     run;

     * Quantile control;
     data Density;
          set Density;
          T = int(value*10)/10;
     run;

     * Computing (ft|x0) and F(t|x0);
    %Integration;
%mend;
%Step_4(tvar=T0, datain=T);
%Step_4(tvar=T1, datain=T);



*Etapa 5;
%macro Step_5;
     proc sort data=Work.T0_Kernel out=WORK._TABLE1_;
          by T_Kernel;
     run;
     proc sort data=Work.T1_Kernel out=WORK._TABLE2_;
          by T_Kernel;
     run;
     data T_Density ;
          merge WORK._TABLE1_ (in=TABLE1) WORK._TABLE2_ (in=TABLE2) ;
          by T_Kernel;
          if TABLE1 | TABLE2;
     run;

     data HR;
          set T_Density;

h0 = ((&Tau_Max. - &Tau_min.)*Density_T0)/(1 - &Tau_Min. - (&Tau_Max. -
     &Tau_min.)*Fcum_T0);
h1 = ((&Tau_Max. - &Tau_min.)*Density_T1)/(1 - &Tau_Min. - (&Tau_Max. -
     &Tau_min.)*Fcum_T1);

if not missing (Density_T0) & not missing (Density_T1) then
 HR = h1/h0;
     run;
    /*
     proc print data=HR;
          where not missing(HR);
          var T_Kernel HR;
     run;
    */
     data Simul.Resuls_&scenaryname.;
          set HR;
     run;
%mend;
%Step_5;

%bootstrap;


* Computing simulation time;
data EndTime;
     End_Datetime = datetime();
     format End_Datetime datetime18.;
```

```
     put End_Datetime = datetime18.;
     ONE = 1;
run;
data CountTime;
     merge StartTime EndTime;
    by ONE;
run;

data CountTime;
     set CountTime;
     Simulation_Time = (End_Datetime - Start_datetime)/(60*60);
     label Simulation_Time = "Time of simulation (hours)";
run;

title1 "Simulation Time (hours)";
proc print data=CountTime; run;
title1;


ods pdf close;
proc printto log=LOG; run;
proc printto print=PRINT; run;
```

APÊNDICE C - Bootstrap para estimação do intervalo de confiança

```
%macro Bootstrap;
/*************************************************************************
     *
                Bootstrap Macro (Confidence Interval Estimation)

           - Select bootstrap samples
           - Repeat Steps 2 to 5
          - Merge datasets in order to obtain confidence interval based
          on the percentiles 2.5% and 97.5% of the empirical HR distribution
           - Graphs

           method=urs              Resample with replacement
           samprate=1              Each bootstrap sample has N observations
           outhits                 Frequency of sample unit selectec
           rep=&bootrep.           Number os bootstrap samples

*************************************************************************
     /

     %macro BootSamples;
        * Select bootstrap samples;
           proc surveyselect data=Subset
                 out=BootstrapSamples
                 seed=92395
                 method=urs
                 samprate=1
                 outhits
                 rep=&bootrep.;
           run;

           * Save bootstrap samples;
        data Simul.BootstrapSamples_&Scenaryname;
                 set BootstrapSamples;
           run;

           * Get parameter estimates for each bootstrap sample;
           ods listing select ParameterEstimates;
           ods select ParameterEstimates(persist);
           ods output ParameterEstimates(persist=proc)=BootCoefs;
           proc quantlife data=BootstrapSamples method=na log;
                 by Replicate;
                 title1 "Quantile survival model (log Time) - Bootstrap
     /samples";
                 model TEMPO*CENSURA(1)= Grupo X1 / quantile =&Tau_min to
     &Tau_Max by &byvalue;
           run;
           ods select all;
           *ods _all_ close;
           ods listing;
           title1;

                        * Save bootstrap samples;
        data Simul.BootCoefs_&Scenaryname;
                 set BootCoefs;
           run;
```

```
/*************************************************************************
                                        Step 3
        For each value of the predictor vector x_0, the M-Dimensional vector
     containing
           the simulated survival times are obtained by means of

              T_m = Q_Ti(Tau_m|x_0) = exp{x'_0*Beta(Tau_m)}, for each
     m=1,2,...,M)

     *************************************************************************/

           %do rep = 1 %to &Bootrep;
           * Means of the model predictor for multivariable adjustment;
                 proc means data=BootstrapSamples mean noprint ;
                       where Replicate = %sysevalf(&rep.);
                       var GRUPO X1;
                       output out=Means;
                 run;
                 data Means;
                       set Means;
                       if _STAT_ = "MEAN";
                       keep GRUPO X1;
                 run;

                 *options mprint symbolgen mlogic;
                 data Coefs;
                       set BootCoefs;
                       if Replicate = %sysevalf(&rep.);
                       LoopControl = int(Quantile*10000);
                 run;


                 %macro Ts_estimates;

/*Estima T_m = Q_Ti(Tau_m|x_0) = exp{x'_0*Beta(Tau_m)}, para m=1,2,...,M)*/
                       %do mm = 1 %to &M.;

                             data Coefs&mm.(keep=Parameter Estimate);
                                   set Coefs;
                                   if LoopControl = (int((&Tau_Min +
     (%sysevalf(&mm.) - 1)*&byvalue)*10000));
                             run;

                             proc transpose data=Coefs&mm. out=Betas&mm.; run;
                       data Betas&mm.;
                                   set Betas&mm.;

                                   rename COL1 = Intercept;
                                   rename COL2 = Grupo;
                                   rename COL3 = X1;
                                   drop _NAME_;
                             run;

                             proc iml;
                                   xvar={Intercept Grupo X1};
                                   xxvar={Grupo X1};
                                   use Betas&mm.;
                                         read all var xvar into Betas;
                                   close;
                                   *print Betas;

                                   Betas=Betas`;
```

```
                              *print Betas;

                              * Vetor com medias dos preditores;
                              use Means;
                                    read all var xxvar into Means;
                              close;
                              *print Means;

                        nparms = nrow(Betas);
                              Um = j(1,1);
                              *print Um nparms;

  /* Define Categoria de "Referencia" para grupo, ajustado por X1 e X2 */
                              Grupo0 = Means;
                              Grupo0[1]  = 0;
                              Grupo0 = Um || Grupo0;
                              Grupo0 = Grupo0`;
                              *print Grupo0;

   /* Define Categoria de "Exposicao" para grupo, ajustado por X1 e X2 */
                              Grupo1 = Means;
                              Grupo1[1]  = 1;
                              Grupo1 = Um || Grupo1;
                              Grupo1 = Grupo1`;
                              *print Grupo1;

       * Etapa 3 - Calcula Tm = Quantil(Tauj | x0) = exp(Beta*Grupo);
            Ts = j(1,4,0);
                              *print Ts;

            Ts[1,1] = %sysevalf(&mm.);
            *Ts[1,2] = %sysevalf(&Tau&mm.);
            Ts[1,2] = &Tau_Min + (%sysevalf(&mm.) - 1)*&byvalue;

             if Betas[1] > .Z & Betas[2] > .Z & Betas[3] > .Z
                              then do;
            T0 = exp(Betas`*Grupo0);
            T1 = exp(Betas`*Grupo1);
                                    *print T0 T1;
                                 end;
                              else do; /* Missing values for time when they
can't be estimate due lack information for the respective percentiles*/
                              T0 = .; T1 = .;
                                 end;

                        Ts[1,3] = T0;
                        Ts[1,4] = T1;
                        *print "Valor Ts (M=&mm.)" Ts;

                              varNames = ("TAUM"||"QUANTILE"||"T0"||"T1");
                              CREATE T&mm. FROM Ts[colname=varnames];
                           APPEND FROM Ts;
                        CLOSE T&mm.;
                              quit;
                           run;
                        %end;
             %mend;
             %Ts_estimates;
```

```
* Dataset containing Ts estimates;
%macro Append;
      data TBoot&rep.;
            set T1;
      run;
      %do mm=2 %to &M.;
      proc append
            BASE=TBoot&rep. data=T&mm.;
      run;
      %end;
%mend;
%Append;

* Exclui datasets desnecessarios;
proc datasets lib=work nolist;
      delete Coefs1-Coefs&M. Betas1-Betas&M. T1-T&M.;
run;
quit;

* Grava dados das Etapas 2 e 3;
data Simul.TemposEstimadosBoot&rep._&scenaryname.;
      set TBoot&rep.;
run;


**************************************************************************
                                    Step 4
          The Kernel Density Estimatin method is applied in order to obtain
     estimatives of

                              f(t|x_0), F(t|x_0) and h_0(t)=h(t|x_0)

                associated to the estimated survival times {T_m,
     m=1,1,2,...,M}, for each
                bootstrap sample (rep=1,2,3,...,Nrep).
**************************************************************************;
                %macro Step_4(tvar=, datain=);
                      data TBoot_&tvar.;
                            set Work.&datain.;
                            T_&tvar. = int(&tvar.*10)/10;
                      run;

                      proc kde data=TBoot_&tvar.;
                            univar T_&tvar. / method=SJPI out=density;
                      run;

                      * Quantile control;
                      data Density;
                            set Density;
                            T = int(value*10)/10;
                      run;

                      * Computing f(t|x0) and F(t|x0);
                %Integration;

                      data &tvar._Kernel_Boot&rep.;
                            set &tvar._kernel;
                      run;
                %mend;
                %Step_4(tvar=T0, datain=TBoot&rep.);
                %Step_4(tvar=T1, datain=TBoot&rep.);
```

```
*************************************************************************
                              Step 5
          Estimation of hazard function h_0(t)=h(t|x_0) and hazard ratio HR for
     each bootstrap sample (rep=1,2,3,...,Nrep).
*************************************************************************;
                %macro Step_5;
               * Check for repeated values of T_Kernel and keep the one with
     higher value
                     for the density f(t);
                     proc sort data = T0_Kernel_Boot&rep.;
                           by T_Kernel Density_T0;
                     run;;
                     data T0_Kernel_Boot&rep.;
                      set T0_Kernel_Boot&rep.;
                           by T_Kernel;
                           First_T_Kernel = First.T_Kernel;
                           Last_T_Kernel = Last.T_Kernel;
                           if (First_T_Kernel = 1 & Last_T_Kernel = 1) |
     (First_T_Kernel = 0 & Last_T_Kernel = 1) ;
                     run;
                     proc sort data = T1_Kernel_Boot&rep.;
                           by T_Kernel Density_T1;
                     run;;
                     data T1_Kernel_Boot&rep.;
                           set T1_Kernel_Boot&rep.;
                           by T_Kernel;
                           First_T_Kernel = First.T_Kernel;
                           Last_T_Kernel = Last.T_Kernel;
                           if (First_T_Kernel = 1 & Last_T_Kernel = 1) |
     (First_T_Kernel = 0 & Last_T_Kernel = 1) ;
                     run;

                     * Getting T_Density for each bootstrap sample;
                     proc sort data=Work.T0_Kernel_Boot&rep.
     out=WORK._TABLE1_;
                           by T_Kernel;
                     run;
                     proc sort data=Work.T1_Kernel_Boot&rep.
     out=WORK._TABLE2_;
                           by T_Kernel;
                     run;
                     data T_Density&rep. ;
                           merge WORK._TABLE1_ (in=TABLE1) WORK._TABLE2_
     (in=TABLE2) ;
                           by T_Kernel;
                           if TABLE1 | TABLE2;
                     run;

                     data HR_Boot&rep.;
                           set T_Density&rep.;

     h0 = ((&Tau_Max. - &Tau_min.)*Density_T0)/(1 - &Tau_Min. - (&Tau_Max. -
     &Tau_min.)*Fcum_T0);
     h1 = ((&Tau_Max. - &Tau_min.)*Density_T1)/(1 - &Tau_Min. - (&Tau_Max. -
     &Tau_min.)*Fcum_T1);

     if not missing (Density_T0) & not missing (Density_T1) then
     HR&rep. = h1/h0;
                     run;
```

```
            /*
                  proc print data=HR_Boot&rep.;
                        title1 "HR for Repetition = &rep.";
                        where not missing(HR&rep.);
                        var T_Kernel HR&rep.;
                  run;
             */
                  data Simul.ResulsBoot&rep._&scenaryname.;
                        set HR_Boot&rep.;
                  run;
             %mend;
             %Step_5;
      %end;

      data HR_Boot;
            merge HR_Boot1-HR_Boot&Bootrep.;
            by T_Kernel;
            keep T_Kernel HR1-HR&Bootrep.;
      run;

      data Simul.HR_Boot_&scenaryname.;
            set HR_Boot;
      run;

      proc datasets lib=work nolist;
            delete TBoot1-TBoot&Bootrep. T0_Kernel_Boot1-
T0_Kernel_Boot&Bootrep. T1_Kernel_Boot1-T1_Kernel_Boot&Bootrep.
                  T_Density1-T_Density&Bootrep. HR_Boot1-HR_Boot&Bootrep.;
      run;
      quit;
%mend;
%BootSamples;

%macro BootstrapCI;
      proc transpose data=HR_Boot out=HR_Boot_Transposed; run;

      *ods trace on;
      ods output Contents.DataSet.Attributes=Ncolumns;
      proc contents data=HR_Boot_Transposed; run;
      data HR_Boot_Transposed;
      length _NAME_ $ 14;
            set HR_Boot_Transposed;
      run;

      data Ncolumns;
            set Ncolumns;
            if Label1 = "Member Type";
            Ncolumns = input(cValue2, 8.) - 1;
            informat Ncolumns F8.0;
            format Ncolumns F8.0;
            keep Ncolumns;
      run;

      data _null_;
            set NColumns;
            call symput('NCols', NColumns);
      run;
      /*%put _global_;*/
```

```
%do col = 1 %to &NCols;
                proc univariate data=HR_Boot_Transposed(where=( _NAME_ ne
     "T_kernel")) pctldef=3 noprint;
                      var COL&col.;
                      output median=HR_Median pctlpts=2.5 97.5
     pctlpre=Percentile nmiss=NMissing out=CI;
                run;

                proc transpose data=CI out=CI_Transposed; run;
                data CI_Transposed;
                      set CI_Transposed;
                      rename COL1 = COL&col.;
                      keep _NAME_ COL1;
                run;

                data CI_Merged;
                      set HR_Boot_Transposed CI_Transposed;
                      if _NAME_ = "T_kernel" | _NAME_ = "HR_Median" | _NAME_ =
     "Percentile2_5" | _NAME_ = "Percentile97_5" | _NAME_ = "NMissing";
                      keep _NAME_ COL&col.;
                run;

                proc transpose data=CI_Merged out=HR_COL&col.; run;
                data HR_COL&col.(drop=_NAME_);
                      set HR_COL&col.;
                      rename Percentile2_5 = HR_Percentile2_5;
                      rename Percentile97_5 = HR_Percentile97_5;
                run;
          %end;
     %mend;
     %BootstrapCI;

     *Append;
     %macro HR_Append;
          data HR_CIs;
                set HR_Col1;
          run;
          %do col=2 %to &NCols;
          proc append
                BASE=HR_CIs data=HR_Col&col.;
          run;
                quit;
          %end;
          proc datasets lib=work nolist;
                delete HR_COL1-HR_COL%sysevalf(&NCols.);
          run;
          quit;
     %mend HR_Append;
     %HR_Append;


     proc sort data=HR_CIs out=WORK._TABLE1_;
          by T_Kernel;
     run;
     proc sort data=Work.HR out=WORK._TABLE2_;
          by T_Kernel;
     run;
     data HR_Results;
          merge WORK._TABLE1_ (in=TABLE1) WORK._TABLE2_ (in=TABLE2) ;
          by T_Kernel;
          if TABLE1 | TABLE2;
     run;
```

```
     data Simul.HR_Results_&scenaryname.;
           set HR_Results;
     run;


    * Lista resutlados;
     proc report data=HR_Results nofs headline headskip;
     where not missing (HR) & not missing (HR_Percentile2_5) & not missing
     (HR_Percentile97_5) & (NMissing < &NMissingValue.);
           title1 "HR estimate ((95% CI) for scenary &Scenaryname";
     column T_Kernel NMissing HR HR_Percentile2_5 HR_Percentile97_5;
     define T_Kernel          / display format=8.1 width=10 'Time';
     define NMissing          / display            width=10 'Missing/Bootstrap';
     define HR                / display format=8.2 width=10 'Hazard/ratio';
     define HR_Percentile2_5  / display format=8.2 width=10 'Percentile 2.5';
     define HR_Percentile97_5 / display format=8.2 width=10 'Percentile 97.5';
     run;
     title1;

*************************************************************************
                                   GRAPHS

        - HR Estimate versus time and confidence band based on the
          2.5th and 97.5th percentiles of empirical distribution for HR
          associated to the bootstrap samples

        - Hazard functions estimated for groups ho(t) and h1(t)
*************************************************************************;

     * Hazard ratio;
     ods listing gpath = &pathname;
     *goptions reset=all gsfmode=replace device=png gsfname=pasta hsize=6
     vsize=4;
     ODS GRAPHICS / RESET IMAGENAME = "HR_Estimate_&scenaryname." IMAGEFMT = png
     HEIGHT = 4in WIDTH = 6in;
     proc sgplot data=HR_Results;
           where not missing(HR) & (NMissing < &NMissingValue.);
           title "HR estimate (95% CI)";
         band x=T_Kernel lower=HR_Percentile2_5 upper=HR_Percentile97_5 /
     legendlabel="95% CI" outline fill;
              series x=T_Kernel y=HR / lineattrs=GraphPrediction (color=red)
     legendlabel="HR estimate";
               refline 0;
               *xaxis type=discrete;
     run;
     title;title1;

     * Hazard rates ho(t) and h1(t);
     *ods listing gpath = &pathname;
     *goptions reset=all gsfmode=replace device=png gsfname=pasta hsize=6
     vsize=4;
     ODS GRAPHICS / RESET IMAGENAME = "Hazard_functions_Estimates_&scenaryname."
     IMAGEFMT = png HEIGHT = 4in WIDTH = 6in;
     proc sgplot data=HR_Results;
           title "Hazard functions";
           series x=T_Kernel y=h0;
           series x=T_Kernel y=h1;
     run;
     title;
%mend;
```