MAE514 - Introdução a Análise de Sobrevivência

Primeira Prova

1^o Semestre de 2021

Informações Importantes

- Esta prova está dividida em duas partes:

  - **PARTE 1:** deve ser feita em horário de aula e entregue no **dia 14/06**, via e-disciplinas exclusivamente;

  - **PARTE 2:** deve ser entregue no **dia 21/06**, via e-disciplinas exclusivamente.

- Ambas partes devem ser feitas **individualmente**, sem consulta a colegas de classe, amigos ou profissionais.
- Leia com atenção as instruções contidas em cada parte da prova.

Primeira Prova: PARTE 1 (5,0 pontos)

Instruções para a PARTE 1

- Esta parte da prova deverá ser realizada em horário de aula.
- A entrega deverá ser via e-disciplinas, exclusivamente, no **dia 14/06**. - A entrega deverá ser realizada até às **21h00** do dia 14/06.
- Haverá uma tolerância de algumas horas na entrega da prova, porém não serão aceitas de forma nenhuma provas entregues após a tolerância. Você poderá entregar a prova até 02h00 do dia 15/06 (2 horas da madrugada, não da tarde). Caso você tenha dificuldade de conexão, por favor entre em contato com a professora *antes do fim do prazo de entrega* da prova.
- A prova deverá ser feita **à mão** e escaneada (você pode tirar fotos também) para envio para o e-disciplinas. Não serão aceitas provas da parte I feita em latex, Libreoffice, Word ou outro editor de texto. Por favor, escreva de forma legível e verifique se a versão escaneada está legível antes da entrega.

QUESTÃO 1 (1,5 pontos)

Sejam $T_1, T_2, \ldots, T_p$ variáveis aleatórias e defina $T = \min(T_1, T_2, \ldots, T_p)$. Assuma que os dados serão coletados de forma que a informação disponível será

$$T = \min(T_1, T_2, \ldots, T_p) \text{ e } C = \begin{cases} 1, & \text{se } T = T_1 \\ 2, & \text{se } T = T_2 \\ \vdots & \\ p, & \text{se } T = T_p \end{cases}.$$

Isso significa que é conhecido o tempo de falha e a causa da falha. Observe que essa estrutura pode ser vista como uma generalização da estrutura de dados com censura à direita. Defina ainda

$$p_j = P(C = j), j = 1, 2, \ldots, p,$$

e

$$F(j,t) = P(C = j, T \leq t).$$

(a) Mostre que

$$S(t) = P(T > t) = \sum_{j=1}^{p} P(C = j, T > t) = \sum_{j=1}^{p} \bar{F}(j,t),$$

em que $\bar{F}(j,t) = (C = j, T > t)$. É importante observar que *não* está sendo feita a suposição de independência entre os tempos $T_1, T_2, \ldots T_p$ e que $F(j,t) + \bar{F}(j,t) = p_j$.

(b) Assuma que

$$\bar{F}(j,t) = \pi_j e^{-\lambda_j t}, j = 1, 2, \ldots, p,$$

com $\lambda_j > 0$, $\pi_j > 0$ e $\pi_1 + \pi_2 + \cdots + \pi_p = 1$. Calcule a função de taxa de falha de T. Em que situação a função de taxa de falha será constante em t?

(c) Ainda no contexto do item (b), calcule $P(T > t|C = m)$ e $P(C = m|T > t)$. Observe que $P(C = m|T > t)$ não é necessariamente constante em t. Em que situação será constante em t?

QUESTÃO 2 (2,0 pontos)

Em algumas aplicações práticas, um terceiro parâmetro é acrescentado na distribuição Weibull. A função de sobrevivência fica dada por

$$S(t) = \begin{cases} 1, & \text{se } t < G; \\ \exp\{-\lambda(t-G)^{\rho}\}, & \text{se } t \geq G. \end{cases}$$

(a) Suponha que se tenha $T_1^*, T_2^*, \ldots, T_n^*$ variáveis aleatórias independentes com distribuição Weibull com três parâmetros, sendo G um valor conhecido. As observações estão sujeitas à censura à direita, ou seja, observa-se de fato $T_i = \min\{T_i^*, C_i\}$, em que C_i é o tempo de censura, e a variável $\delta_i = I(T_i^* \leq C_i)$. Nessa situação, obtenha a função de verossimilhança dos parâmetros do modelo.

(b) Encontre o estimador de máxima verossimilhança de λ quando $\rho = 1$.

(c) Nas condições do item (b), com $\rho = 1$, obtenha a informação de Fisher (esperada) para os casos:

(*i*) $C_i = k, i = 1, 2, \ldots, n$, com $k > G$, ou seja, um esquema de censura Tipo I;

(*ii*) $C_i, i = 1, 2, \ldots, n$ são variáveis aleatórias independentes com função de sobrevivência dada por $S_C(c) = \exp\{-\theta(c - G)\}$ para $c \geq G$ e $S_C(c) = 1$ para $c < G$, G constante conhecida.

QUESTÃO 3 (1,5 pontos)

Suponha que um pesquisador está interessado em comparar dois grupos, A e B, em termos do tempo de sobrevivência. Assuma um modelo de locação-escala, dado por

$$Y = \log T = \mu + \gamma x + \sigma\epsilon,$$

em que T denota o tempo de falha, x é uma variável binária indicadora de tratamento ($x = 0$ se tratamento A e $x = 1$ se tratamento B),(γ, σ) são parâmetros desconhecidos e ϵ é o erro aleatório. Assuma que a função de densidade de probabilidade do erro ϵ seja dada por

$$f_\epsilon(z) = \frac{e^z}{(1+e^z)^2}, -\infty < z < \infty.$$

(a) Obtenha a função de sobrevivência de T dado x.

(b) Defina a chance de que um indivíduo experiencie o evento de interesse antes de t por

$$(1 - S(t|x))\,/S(t|x).$$

Considere ainda que a chance de apresentar o evento de interesse (até t) para observações pertencentes ao grupo B é 20% **maior** do que a chance de apresentar o evento para observações do grupo A. Se o pesquisador obteve a informação de que o tempo mediano de sobrevivência estimado para o grupo A é igual a 35 meses, o que pode ser concluído sobre a função de sobrevivência estimada para o grupo B no instante $t = 35$ meses? Explique claramente como você encontrou os resultados.

Primeira Prova: PARTE 2 (5,0 pontos)

Instruções para a PARTE 2

- A entrega deverá ser via e-disciplinas, exclusivamente, no **dia 21/06**.
- A entrega deverá ser realizada até às **23h59** do dia 21/06. Não haverá tolerância para atrasos.
- Para a atividade 1, você poderá enviar um arquivo gravado ou enviar o link da gravação feita utilizando o *google meet*.
- Para a atividade 2, você deverá entregar um arquivo com a resolução da questão (você pode usar o editor de texto de desua preferência) contendo os códigos utilizados.

ATIVIDADE 1 (2,5 pontos)

Para esta atividade, você deverá gravar um vídeo no qual você deve explicar a construção do estimador de Kaplan-Meier da função de sobrevivência $S(t)$ para dados censurados à direita. Para auxiliar a explicação, você deverá usar um exemplo de um estudo com pacientes com câncer de mama (Liu, 2012), que foram acompanhadas por um período de 7 anos, até o óbito ou fim do estudo:

$5, 17, 20+, 24, 32, 35+, 40, 46, 47, 50, 59, 74+.$

As observações censuradas à direita estão sinalizadas com o símbolo "+". Não se esqueça de apresentar o estimador de Greenwood para a variância. Aborde também a obtenção de intervalos de confiança (ponto a ponto).

Esse vídeo pode ser no *google meet*, no formato em que as aulas estão sendo ministradas, mas pode ser também em qualquer outro formato que você se sentir mais confortável. Você pode utilizar qualquer recurso que tenha disponível, mas a avaliação será do domínio do conteúdo.

ATIVIDADE 2 (2,5 pontos)

Considere os dados provenientes do EBMT (*European Resgistry for Blood and Marrow Transplatation*), discutido em Putter, Fiocco Geskus (2007). Os dados são referentes a 2204 pacientes que receberam transplante de medula óssea entre 1995 e 1998 reportados ao EBMT. Os dados estão disponíveis no arquivo bmt3.csv e a descrição em bmt3.des. As variáveis nos dados são:

- Tempo, em dias, após o transplante até recuperação das plaquetas ou perda de acompanhamento
- Indicador de recuperação das plaquetas: 1, se recuperado; 0, se perda de acompanhamento
- Tempo livre de doença: tempo, em dias, após o transplante até óbito ou reincidência do câncer;
- Indicador de evento: 1, se óbito ou reincidência; 0 se censura
- Classificação da doença: leucemia linfoblástica aguda (ALL), leucemia mielóide aguda (AML) e leucemia mielóide crônica (CML)
- Idade, em categorias
- Correspondência de gênero doador-receptor
- Esgotamento de células T

Utilizando esses dados, responda os itens descritos a seguir:

(a) Faça uma análise descritiva dos dados considerando como desfecho ou variável resposta o tempo até recuperação das plaquetas.

(b) Faça testes de log-rank comparando as categorias das covariáveis no estudo.

(c) Faça uma análise descritiva dos dados considerando como desfecho ou variável resposta o tempo livre da doença.

(d) Faça testes de log-rank comparando as categorias das covariáveis no estudo. Considere também a inclusão de ponderações nesse caso. Discuta os resultados.

(e) Considerando o tempo livre da doença, ajuste um modelo Weibull aos dados. Apresente os resultados do modelo completo, com todas as covariáveis incluídas. Faça um processo de seleção de variáveis utilizando o teste da razão de verossimilhanças e apresente o resultado do modelo final obtido. Você precisa descrever claramente o processo de seleção das variáveis adotado, mas deve apresentar apenas as estimativas e resultados de dois modelos: modelo completo e modelo final. Você pode apresentar os resultados do modelo na parametrização de locação-escala.

(f) Interprete os parâmetros do modelo final obtido em (e).

(g) De forma semelhante ao item (e), ajuste um modelo log-logístico aos dados. Faça da mesma forma (porém utilizando a distribuição log-logística) e apresente os resultados do modelo completo e do modelo final.

(h) Interprete os parâmetros do modelo final obtido em (g).

Importante: Em todos os itens, os resultados apresentados devem ser interpretados. A redação será também avaliada. Não se esqueça de apresentar códigos dos programas utilizados.