Lista nº 4 – Modelos de Regressão Paramétricos

1. (Colosimo, E. A. e Giolo, S. R. [2] – Exercício 3, pág. 112) Os dados mostrados as seguir representam o tempo até a ruptura de um tipo de isolante elétrico sujeito a uma tensão de estresse de 35 kV. O teste consistiu em deixar 25 destes isolantes funcionando até que 15 deles falhassem (censura do tipo II), obtendo-se os seguintes resultados, em minutos:

0,19	0,78	0,96	1,31	2,78	3,16	4,67	4,85
6,50	7,35	8,27	12,07	32,52	33,91	36,71	

   A partir desses dados amostrais, utilize **métodos não paramétricos** para obter as seguintes observações:

   a. Uma estimativa para o tempo mediano de vida deste tipo de isolante elétrico funcionando a 35 kV.
   b. Uma estimativa (pontual e intervalar) para a fração de defeitos esperada nos dois primeiros minutos de funcionamento.
   c. Uma estimativa (pontual) para o tempo médio de vida destes isolantes funcionando a 35 kV (limitados em 40 minutos).
   d. O tempo necessário para 20% dos isolantes estarem fora de operação.

   Identifique um **modelo paramétrico** para explicar esses dados e, em seguida, obtenha novamente as informações solicitadas anteriormente:

   e. Uma estimativa (pontual e intervalar) para o tempo mediano de vida deste tipo de isolante elétrico funcionando a 35 kV.
   f. Uma estimativa (pontual e intervalar) para a fração de defeitos esperada nos dois primeiros minutos de funcionamento.
   g. Uma estimativa (pontual e intervalar) para o tempo médio de vida destes isolantes funcionando a 35 kV (limitados em 40 minutos.
   h. O tempo necessário para 20% dos isolantes estarem fora de operação.

2. *Geração de amostra com dados censurados*.
   a. Implemente o seguinte procedimento para simular uma amostra com dados censurados para o modelo paramétrico do exercício (1):
      i. Defina o nível de censura [mantenha a taxa do exercício (1)]
      ii. Defina o tamanho do conjunto de dados (*n*) [mantenha o tamanho amostral do exercício (1)].
      iii. Defina os valores dos parâmetros reais que vai utilizar para simular os dados proveniente do modelo de sobrevida. [Considere como valores reais dos parâmetros as estimativas de máxima verossimilhança obtidas para os dados originais].
      iv. Sugestão de código em R. O aluno é estimulado a modificá-lo ou aprimorá-lo:

```
perc <- 0.05 # nível de censura
y # dados do modelo simulados sem censura
aa <- sort(y, decreasing = TRUE)
cutoff <- aa[ceiling(perc*n)]   # pto de corte
  # indicadora de censura à direita (tipo II)
cc <- matrix(1, n, 1)*(y >= cutoff)
y[cc == 1] <- cutoff
```

   v. O vetor `y` final é composto por tempos de sobrevida incluindo a censura. O vetor `cc` é indicador de censura.

3. *Bootstrap paramétrico*. Considere que já foi ajustado um modelo (por exemplo, exponencial, weibull, log-normal, etc.) a um conjunto de dados com *n* observações, obtendo-se a estimativa de máxima verossimilhança dos parâmetros.
   a. Implemente o seguinte procedimento para construção de um intervalo de confiança bootstrap:
      i. Gere B = 1000 amostras de tamanho *n* provenientes do modelo ajustado considerando como parâmetro real a estimativa de máxima verossimilhança obtida para os dados originais.
      ii. Para cada amostra bootstrap, ajuste o modelo em estudo e estime o parâmetro por máxima verossimilhança.
      iii. Guarde as 1000 estimativas do parâmetro e ordene-as. A partir dos percentis 2,5% e 97,5%, constroem-se, por exemplo, um intervalo de confiança bootstrap de 95% para o parâmetro de interesse.
   b. Determine o intervalo de confiança bootstrap para para o tempo médio de vida dos isolantes funcionando a 35 kV [exercício (1)].
   c. Compare com o intervalo paramétrico obtido no item (g) do exercício (1). Comente sobre as diferenças (ou não diferenças) encontradas.

4. *Simulação*. Considere que se pretende avaliar a convergência assintótica da estatística da razão de verossimilhanças, utilizada no exercício (1), no contexto das seguintes hipóteses de interesse: H_0: modelo exponencial vs. H_1: modelo gama.
   a. Implemente o seguinte procedimento para avaliar a convergência assintótica da estatística de teste da razão de verossimilhanças:
      i. Gere B = 1.000 amostras de tamanho n [mesmo tamanho da amostra utilizada no exercício (1)], provenientes do modelo sob a hipótese nula.
      ii. Para cada uma dessas amostras, estime os parâmetros por máxima verossimilhança, sob as hipóteses nula e alternativa.
      iii. Em seguida, calcule o valor da estatística de teste da razão de verossimilhanças (RV).
      iv. Guarde os 1.000 valores da estatística de teste. Esses valores compõem a distribuição empírica dessa estatística de interesse.
      v. Faça um teste de hipóteses não paramétrico para avaliar se a distribuição dessa estatística segue uma qui-quadrado com 2 graus de liberdade, no contexto das hipóteses consideradas acima. [Por exemplo, `ks.test(RV, 'pchisq', 2)`].
      vi. Altere o valor de *n* para avaliar a convergência assintótica da estatística de teste.
   b. A partir de qual valor de *n* ocorre a convergência assintótica para uma qui-quadrado com 2 graus de liberdade, no contexto das hipóteses consideradas em (4)?

5. (Klein, J. P. e Moeschberger, M. L. [3] – Exercícios 12.3 [exceto item f], pág. 401 e 12.14, pág. 403). O conjunto de dados `hodg{KMsurv}` refere-se a estudo

de tempo até a morte ou recidiva (em dias) para 23 pacientes com linfoma não-Hodgkin (`NHL`), desses, 11 receberam transplante alogênico (`allo`) de irmão doador com compatibilidade de antígenos leucocitários humanos (*HLA-matched donor*) e 12 pacientes receberam transplante autólogo (`auto`), em que sua própria medula foi limpa e reimplantada após uma elevada dose de quimioterapia. Além desses, o conjunto contém dados de 20 pacientes com linfoma de Hodgkin (`HL`), 5 dos quais receberam transplante alogênico (`allo`) de irmão com compatibilidade (`HLA`) e 15, transplante autólogo (`auto`). Em razão de haver potencial para eficácias diferentes para os dois tipos de transplantes e para os dois tipos de linfomas, busca-se construir um modelo com efeitos principais para o tipo de transplante e para o tipo de linfoma, além de um termo de interação entre essas covariáveis.

a. Usando o modelo de regressão de Weibull, analise esses dados conduzindo um teste global de significância dos efeitos do tipo de transplante e do tipo de doença no tempo de sobrevida. Construa uma tabela Anova para resumir as estimativas dos coeficientes do risco e os resultados dos testes com 1 grau de liberdade para cada covariável do modelo. Obtenha a matriz de covariâncias de suas estimativas.
b. Usando o teste da razão das verossimilhanças, teste a hipótese de não haver interação entre o tipo de transplante-doença.
c. Encontre as estimativas pontuais e os intervalos de 95% de confiança para o risco relativo de morte para um paciente `NHL` de autotransplante comparado com um paciente `NHL` de transplante alogênico.
d. Teste a hipótese de que as taxas de morte são as mesmas para os grupos de pacientes `HL` e `NHL`, quando submetidos a transplantes alogênicos.
e. Teste a hipótese de que são as mesmas as taxas de morte para transplantes autólogos e alogênicos contra a hipótese de que eles são diferentes para pelo menos um dos grupos de doença.
   H_0: $\lambda(t|\texttt{allo}, \texttt{NHL}) = \lambda(t|\texttt{auto}, \texttt{NHL})$ e $\lambda(t|\texttt{allo}, \texttt{HL}) = \lambda(t|\texttt{auto}, \texttt{HL})$
f. Verifique o ajuste desse modelo usando os resíduos de Cox-Snell.
g. Repita (f), usando os resíduos *deviance*.
h. Repita (f) e (g), para o modelo de regressão log-logístico
i. Detalhes sobre o Banco de Dados `hodg`:

Dados provenientes de estudo realizado no *The Ohio State University Bone Marrow Transplant Unit*, em 43 pacientes com transplante de medula. Todos os pacientes tinham ou linfoma de Hodgkin (`HL`) ou linfoma não-Hodgkin (`NHL`), submetidos ou a um transplante alogênico (`allo`) ou a um transplante autólogo (`auto`). Detalhes desse estudo podem ser encontrados em: Avalos et al. Preparation for marrow transplantation in Hodgkin's and non-Hodgkin's Lymphoma using Bu/Cy. Bone Marrow Transplantation, n. 13, p. 133-138, 1993.

Variável	Descrição
gtype	tipo de enxerto (1 = alogênico, 2 = autólogo)
dtype	tipo de doença (1 = linfoma não-Hodgkin, 2 = linfoma de Hodgkin)
time	tempo até a morte ou recidiva, em dias
delta	indicador de morte/recidiva (0 = vivo, 1 = morto)

score	escala de desempenho de Karnofsky (escores de 0 = morto a 100% = normal, sem queixas de doenças)
wtime	tempo de espera até o transplante, em meses

6. (Carvalho, M. S. *et. al* [(1)] – Exercício 5.10, pág. 179) A Aids passou a ter tratamento apenas em 1991. Desde então a terapia antirretroviral evoluiu da monoterapia para a terapia combinada (2 ou mais componentes) e, por fim, para a terapia de alta potência (no mínimo 3 componentes, sendo um inibidor de protease). Espera-se que as terapias mais recentes sejam mais efetivas em aumentar a sobrevivência. Teste essa hipótese ajustando um modelo exponencial aos dados da coorte de Aids (banco `ipec.csv`).

    a. Ajuste um modelo apenas com a variável `tratamento`. O modelo com a variável tratamento é melhor do que o modelo sem covariáveis? Interprete o efeito dos tratamentos na sobrevivência (lembrando-se que os efeitos dos tratamentos estão sendo estimados em relação à ausência de tratamento).
    b. Faça uma análise gráfica do ajuste do modelo, comparando-o com a curva de Kaplan-Meier estratificada por tratamento. O que você tem a dizer sobre a adequação do modelo exponencial?
    c. Faça a análise de resíduos do modelo estimado. Existe algum ponto influente, sobre a estimativa dos parâmetros (`ldcase`), sobre os valores preditos (`ldcase`) ou sobre o parâmetro de forma?
    d. Caso considere algum ponto mais influente, retire-o e refaça a análise.
    e. Ajuste um outro modelo exponencial, adicionando variáveis de controle (sexo, idade e tipo de atendimento). Quais variáveis tiveram efeito significativo? Quais tiveram efeito protetor?
    f. Detalhes sobre o Banco de Dados `ipec.csv`:

    Dados provenientes de coortes hospitalares de pacientes portadores de HIV. A primeira coorte é constituída dos pacientes portadores de HIV atendidos entre 1986 e 2000 no Instituto de Pesquisa Clínica Evandro Chagas (Ipec/Fiocruz). Dessa coorte, obteve-se uma amostra de 193 indivíduos que foram diagnosticados como portadores de Aids (critério CDC 1993) durante o período de acompanhamento.

Variável	Descrição
id	identificação do paciente
ini	data do diagnóstico da aids (em dias)
fim	data do óbito (ou perda do paciente)
tempo	dias de sobrevivência do diagnóstico até o óbito
status	0 = censura, 1 = óbito
sexo	F = feminino, M = masculino
escola	0 = sem escolaridade, 1 = ensino fundamental, 2 = ensino médio, 3 = ensino superior
idade	idade na data de diagnóstico de Aids (20 a 68 anos)
risco	0 = homossexual masculino, 1 = usuário de drogas injetáveis, 2 = transfusão, 3 = contato sexual com HIV+, 5 = hetero com múltiplos parceiros, 6 = dois fatores de risco
acompan	acompanhamento: 0 = ambulatorial/hospital-dia, 1 = internação posterior, 2 = internação imediata

óbito	S = óbito, N = não óbito, I = ignorado
anotrat	ano do início do tratamento (1990 a 2000), 9 = sem tratamento
tratam	terapia antirretroviral: 0 = nenhum, 1 = mono, 2 = combinada, 3 = potente
doença	de apresentação: 1 = pcp, 2 = pcp pulmonar, 3 = pcp disseminada, 4 = toxoplasmose, 5 = sarcoma, 7 = outra doença, 8 = candidíase, 9 = duas doenças, 10 = herpes, 99 = definido por cd4
p ropcp	profilaxia para pneumocistis: 0 = sem profilaxia, 2 = primária, 3 = secundária, 4 = ambas

7. (Carvalho, M. S. *et. al* [1] – Exercício 5.11, pág. 180) Use a distribuição Weibull para ajustar um modelo para os pacientes em diálise (arquivo `dialise.csv`).

    a. Avalie qual é o efeito da doença de base na sobrevivência, controlado por idade.
    b. Existe evidência a favor da utilização de um modelo mais simples (exponencial)?
    c. Existe evidência a favor da utilização de um modelo com menos variáveis?
    d. Existe evidência de pontos influentes?
    e. Exclua do modelo as variáveis com p-valor > 0,1 e compare o novo modelo com o anterior, calculando a razão de verossimilhanças entre os dois.
    f. Detalhes sobre o Banco de Dados `dialise.csv`:

    Esses dados provém de uma coorte de 6.805 pacientes que foram submetidos a hemodiálise em 67 unidade de atendimento no Rio de Janeiro, no período de janeiro de 1998 a outubro de 2001. Os dados foram originados pelo sistema Apac (Autorização de Procedimentos de Alta Complexidade – DATASUS). Uma discussão detalhada do tema e da modelagem pode ser encontrada em Carvalho et al. (2003). A tabela a seguir contém a descrição das variáveis.

Variável	Descrição
unidade	número do centro de diálise
idade	idade ao iniciar a diálise
inicio	data do início da primeira diálise
fim	data da interrupção do acompanhamento
status	0 = censura, 1 = óbito
tempo	tempo de sobrevivência (meses) (`fim - inicio`)
sexo	0 = feminino, 1 = masculino
causa	`hip` = hipertensão, `dia` = diabetes, `ren` = renal, `con` = congênita, `out` = outras
grande	número de salas de diálise na unidade de tratamento 0 = uma ou duas salas e 1 = três ou mais salas
risco	0 = homossexual masculino, 1 = usuário de drogas injetáveis, 2 = transfusão, 3 = contato sexual com HIV+, 5 = hetero com múltiplos parceiros, 6 = dois fatores de risco
cdiab	1 = diabetes como causa da insuficiência renal e 0 = não
crim	1 = causas renais e 0 = não
congenita	1 = causas congênitas e 0 = não

Referências:

(1) CARVALHO, M. S. *et. al Análise de sobrevivência: teoria e aplicações em saúde*. Rio de Janeiro: Editora Fiocruz, 2011.
Bancos de dados: Disponíveis em http://sobrevida.fiocruz.br/
(2) COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. São Paulo: Edgard Blücher, 2006.
(3) KLEIN, J. P.; MOESCHBERGER, M. L. *Survival analysis: techniques for censored and truncated data*. New York, USA: Springer-Verlag, 1997.