Análise de Sobrevida
Teoria e Aplicações em Saúde

Caderno de Respostas

Marilia Sá Carvalho
Valeska Lima Andreozzi
Claudia Torres Codeço
Maria Tereza Serrano Barbosa
Silvia Emiko Shimakura

6

Modelos de regressão paramétricos

Exercícios

Exercício 6.1: O banco de dados *leite2.txt* contém dados de tempo de aleitamento de crianças de 4 comunidades. No ajuste não-paramétrico a esses dados, observamos que pertencer a uma comunidade não teve efeito no período de aleitamento. Confirme este achado, ajustando um modelo paramétrico a esses dados. Tente o modelo exponencial e o Weibull.

```
> require(survival)
```

```
[1] TRUE
```

```
> leite2 <- read.table("leite2.txt", header = T, sep = "")
> y <- Surv(leite2$tempo, leite2$status)
> modeloE1 <- survreg(y ~ factor(grupo), data = leite2, dist = "exponential")
> modeloW1 <- survreg(y ~ factor(grupo), data = leite2, dist = "weib")
> summary(modeloE1)
```

```
Call:
survreg(formula = y ~ factor(grupo), data = leite2, dist = "exponential")
               Value Std. Error     z        p
(Intercept)    1.609      0.258 6.233 4.57e-10
factor(grupo)2 0.113      0.365 0.310 7.56e-01
factor(grupo)3 0.410      0.365 1.123 2.62e-01
factor(grupo)4 0.052      0.365 0.142 8.87e-01

Scale fixed at 1
```

```
Exponential distribution
Loglik(model)= -165.2   Loglik(intercept only)= -166
        Chisq= 1.58 on 3 degrees of freedom, p= 0.66
Number of Newton-Raphson Iterations: 5
n= 60


> summary(modeloW1)


Call:
survreg(formula = y ~ factor(grupo), data = leite2, dist = "weib")
                Value Std. Error      z        p
(Intercept)     1.651      0.218  7.568 3.78e-14
factor(grupo)2  0.105      0.306  0.344 7.31e-01
factor(grupo)3  0.426      0.307  1.391 1.64e-01
factor(grupo)4  0.126      0.310  0.408 6.83e-01
Log(scale)     -0.175      0.103 -1.703 8.86e-02

Scale= 0.84

Weibull distribution
Loglik(model)= -163.8   Loglik(intercept only)= -164.9
        Chisq= 2.19 on 3 degrees of freedom, p= 0.53
Number of Newton-Raphson Iterations: 7
n= 60
```

Resposta: Note que realmente tanto sob o modelo exponencial quanto o Weibull a variável comunidade (aqui chamada 'grupo') não foi significativa. A estimativa do parâmetro de escala é marginalmente significativo (p-valor=0,089), ou seja, existe alguma indicação de que entre o modelo exponencial e o modelo Weibull o segundo pode ser mais adequado para estes dados.

Exercício 6.2: Um estudo foi realizado para estimar o efeito do transplante de medula óssea na sobrevida de pacientes com leucemia. As covariáveis analisadas foram: idade, fase da doença, ter ou não desenvolvido doença do enxerto crônica e ter ou não desenvolvido doença do enxerto aguda (para mais detalhes acerca desse estudo, refira-se ao Apêndice C.5). Ao se ajustar um modelo exponencial aos dados, obteve-se a seguinte saída do R:

```
                Value Std. Error      z        p
(Intercept)    7.13536     0.4992 14.293 2.44e-46
idade         -0.00179     0.0146 -0.122 9.03e-01
fase interm   -0.79363     0.3651 -2.174 2.97e-02
fase avançada -1.29759     0.4995 -2.598 9.39e-03
```

```
doençacronica  0.92521     0.3335  2.775 5.53e-03
doençaaguda   -1.43654     0.3158 -4.549 5.40e-06

Scale fixed at 1

Exponential distribution
Loglik(model)= -348.3   Loglik(intercept only)= -374.2
        Chisq= 51.96 on 5 degrees of freedom, p= 5.5e-10
Number of Newton-Raphson Iterations: 5
```

Observe a saída do R e responda:

1. O modelo com covariáveis é melhor do que o modelo nulo (sem covariáveis)?

   Resposta: Não. Note que a log-verossimilhança do modelo com covariáveis é muito maior do que a do modelo nulo e o teste da Deviance entre o modelo com covariáveis e o modelo nulo resultou um p-valor praticamente nulo (p=5.5e-10). Portanto temos evidências altamente significativas contra o modelo nulo.

2. Que covariáveis estão associadas com a melhoria da sobrevida? Quais estão associadas com redução da sobrevida?

   Resposta: As covariáveis associadas com a redução da sobrevida são aquelas em que o efeito estimado é negativo: fase intermediária, fase avançada e doença aguda. A covariável idade tem efeito estimado negativo, porém este efeito é não significativo. A única covariável associada a um prognóstico favorável da sobrevida é doença crônica. Cabe observar que como o próprio nome indica, doença crônica necessariamente tem que evoluir durante um tempo razoável, ou seja, o paciente tem que sobreviver por um tempo razoável para apresentá-la.

3. Escreva a função de risco, $\lambda(t)$, estimada para esta coorte.

   Resposta:

$$\begin{aligned}\lambda(t) &= \exp(-(7.13536 - 0.00179 \times \text{idade} - 0.79363 \times \text{fase interm} \\ &- 1.29759 \times \text{fase avançada} + 0.92521 \times \text{doençacronica} \\ &- 1.43654 \times \text{doençaaguda}))\end{aligned}$$

4. Qual seria o risco de óbito de um paciente de 30 anos, em fase intermediária, com doença crônica?

   Resposta:

```
> lambdac <- exp(-(7.13536 - 0.00179 * 30 - 0.79363 + 0.92521))
> lambdac

[1] 0.0007367662
```

5. Qual seria o risco de óbito de um paciente de 30 anos, em fase intermediária, com doença aguda?

   Resposta:

```
> lambdaa <- exp(-(7.13536 - 0.00179 * 30 - 0.79363 - 1.43654))
> lambdaa

[1] 0.007816722
```

   O risco de óbito é 10,6 vezes maior para o paciente com doença aguda.

6. Um segundo modelo, mais simples, foi ajustado aos dados, contendo apenas a covariável fase. O logaritmo da função de verossimilhança deste modelo simples foi de -363.6. Compare este modelo com o mais completo acima e indique se o completo resultou em melhor ajuste.

   Resposta: *l1* e *l2* são as logverossimilhanças dos modelos completo e reduzido, respectivamente.

```
> l1 <- -348.3
> l2 <- -363.6
```

   Calculando a deviance *dev* e os graus de liberdade que é a diferença entre o número de parâmetros do modelo

```
> dev <- 2 * (l1 - l2)
> dev

[1] 30.6

> gl <- 6 - 4
```

   E, por último, calcula-se o p-valor da distribuição χ^2 sob a hipótese nula de que o modelo reduzido é melhor

```
> pvalor <- 1 - pchisq(dev, gl)
> pvalor

[1] 2.26618e-07
```

Rejeitamos o modelo mais simples com p-valor=0,00000023, ou seja, a redução no valor da verossimilhança dada pelo modelo mais completo foi significativa. Em outras palavras, doença crônica ou aguda é um fator prognóstico importante para o tempo de sobrevida.

Exercício 6.3: Em Aids, a terapia anti-retroviral evoluiu da monoterapia para a terapia combinada (2 componentes) e, por fim, para a terapia de alta potência (3 componentes). Espera-se que quanto mais componentes tiver mais efetiva seja a terapia em aumentar a sobrevida. Teste esta hipótese, ajustando um modelo exponencial aos dados da coorte de Aids (*ipec.csv*).

1. Ajuste um modelo com a variável tratamento apenas. O modelo com a variável tratamento é melhor do que o modelo sem covariáveis? Interprete o efeito dos tratamentos na sobrevida. A variável tratamento deve ser modelada como um fator, e não como numérica, lembrando que os efeitos dos tratamentos estão sendo estimados em relação à ausência de tratamento.

```
> ipec <- read.table("ipec.csv", header = T, sep = ";")
> mod.ipec <- survreg(Surv(tempo, status) ~ factor(tratam), data = ipec,
+     dist = "exp")
> summary(mod.ipec)

Call:
survreg(formula = Surv(tempo, status) ~ factor(tratam), data = ipec,
    dist = "exp")
                Value Std. Error     z         p
(Intercept)      6.14      0.177 34.73 2.91e-264
factor(tratam)1  1.59      0.226  7.07  1.58e-12
factor(tratam)2  2.68      0.445  6.01  1.80e-09
factor(tratam)3  3.01      1.016  2.97  3.00e-03

Scale fixed at 1

Exponential distribution
Loglik(model)= -742.9   Loglik(intercept only)= -774.6
        Chisq= 63.49 on 3 degrees of freedom, p= 1.1e-13
Number of Newton-Raphson Iterations: 6
n= 193
```

Resposta: Rejeitamos a hipótese nula de que o modelo nulo é melhor através da estatística de deviance igual a 63,49 que segue uma distribuição χ^2 com 3 graus de liberdade e p-valor menor que 0,001 (*p= 1.1e-13*). Conclusão, o modelo com a covariável tratamento é melhor.

Para calcular o risco temos que substituir os valores das variáveis *dummies* na expressão do risco do modelo exponencial que é $\lambda(t|\boldsymbol{x}) = \alpha(\boldsymbol{x}) = \exp(\boldsymbol{x}\boldsymbol{\beta})$. Como o R parametriza as distribuições de forma diferente, temos que **trocar o sinal dos coeficientes** para interpretarmos de acordo com o texto do capítulo.

Calculando o risco de um paciente sem nenhum tratamento:

```
> trat1 <- 0
> trat2 <- 0
> trat3 <- 0
> lambda0 <- exp(-(mod.ipec$coef[1] + mod.ipec$coef[2] * trat1 +
+     mod.ipec$coef[3] * trat2 + mod.ipec$coef[4] * trat3))
> lambda0

(Intercept)
0.002156625
```

Calculando o risco de um paciente com monoterapia (`tratam = 1`):

```
> trat1 <- 1
> trat2 <- 0
> trat3 <- 0
> lambda1 <- exp(-(mod.ipec$coef[1] + mod.ipec$coef[2] * trat1 +
+     mod.ipec$coef[3] * trat2 + mod.ipec$coef[4] * trat3))
> lambda1

 (Intercept)
0.0004381632
```

Calculando o risco de um paciente com terapia combinada (`tratam = 2`):

```
> trat1 <- 0
> trat2 <- 1
> trat3 <- 0
> lambda2 <- exp(-(mod.ipec$coef[1] + mod.ipec$coef[2] * trat1 +
+     mod.ipec$coef[3] * trat2 + mod.ipec$coef[4] * trat3))
> lambda2

 (Intercept)
0.0001485001
```

Calculando o risco de um paciente com terapia potente (`tratam = 3`):

```
> trat1 <- 0
> trat2 <- 0
> trat3 <- 1
> lambda3 <- exp(-(mod.ipec$coef[1] + mod.ipec$coef[2] * trat1 +
```

```
+     mod.ipec$coef[3] * trat2 + mod.ipec$coef[4] * trat3))
> lambda3

 (Intercept)
0.0001058985
```

Calculando os riscos relativos em relação ao paciente sem nenhum tratamento:

```
> lambda0/lambda1

(Intercept)
   4.921968

> lambda0/lambda2

(Intercept)
   14.52271

> lambda0/lambda3

(Intercept)
   20.36501
```

Todos os tratamentos são altamente significativos no aumento da sobrevida, mas a terapia potente aumenta mais a sobrevida do que a terapia combinada e esta tem um melhor efeito do que a monoterapia. Em termos do risco de óbito, a razão dos riscos de pacientes sem tratamento e com a terapia potente é aproximadamente 20,4, um valor extremamente alto.

2. 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?

```
> km <- survfit(Surv(tempo, status) ~ factor(tratam), data = ipec)
> plot(km, ylab = "S(t)", xlab = "dias", conf.int = F, col = 1:4,
+     mark.time = F)
> title("Tratamento em Aids")
```

Basta adicionar as curvas de sobrevida de acordo com o modelo exponencial.

a. Pacientes sem tratamento

```
> alpha0 <- exp(-6.14)
> sobre0 <- function(x) {
+     exp(-alpha0 * x)
+ }
> curve(sobre0, from = 0, to = 3500, lty = 2, add = T, col = 1)
```

b. Paciente em monoterapia

```
> alpha1 <- exp(-6.14 - 1.59)
> sobre1 <- function(x) {
+     exp(-alpha1 * x)
+ }
> curve(sobre1, from = 0, to = 3500, lty = 2, add = T, col = 2)
```

c. Paciente em terapia combinada

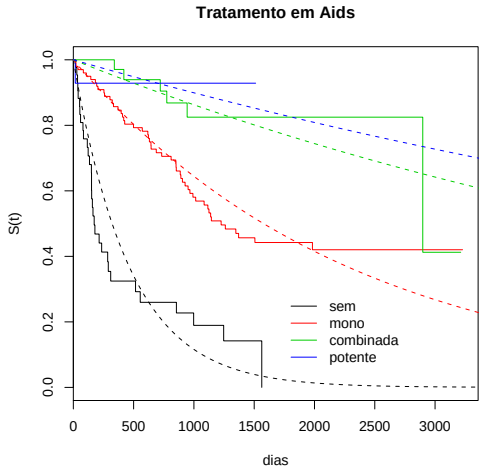
```
> alpha2 <- exp(-6.14 - 2.68)
> sobre2 <- function(x) {
+     exp(-alpha2 * x)
+ }
> curve(sobre2, from = 0, to = 3500, lty = 2, add = T, col = 3)
```

d. Paciente em terapia potente

```
> alpha3 <- exp(-6.14 - 3.01)
> sobre3 <- function(x) {
+     exp(-alpha3 * x)
+ }
> curve(sobre3, from = 0, to = 3500, lty = 2, add = T, col = 4)
> legend(1700, 0.3, c("sem", "mono", "combinada", "potente"), bty = "n",
+     col = 1:4, lty = 1)


> km <- survfit(Surv(tempo, status) ~ factor(tratam), data = ipec)
> plot(km, ylab = "S(t)", xlab = "dias", conf.int = F, col = 1:4,
+     mark.time = F)
> title("Tratamento em Aids")
> alpha0 <- exp(-6.14)
> sobre0 <- function(x) {
+     exp(-alpha0 * x)
+ }
> curve(sobre0, from = 0, to = 3500, lty = 2, add = T, col = 1)
> alpha1 <- exp(-6.14 - 1.59)
> sobre1 <- function(x) {
+     exp(-alpha1 * x)
+ }
> curve(sobre1, from = 0, to = 3500, lty = 2, add = T, col = 2)
> alpha2 <- exp(-6.14 - 2.68)
> sobre2 <- function(x) {
+     exp(-alpha2 * x)
+ }
> curve(sobre2, from = 0, to = 3500, lty = 2, add = T, col = 3)
> alpha3 <- exp(-6.14 - 3.01)
> sobre3 <- function(x) {
```

```
+     exp(-alpha3 * x)
+ }
> curve(sobre3, from = 0, to = 3500, lty = 2, add = T, col = 4)
> legend(1700, 0.3, c("sem", "mono", "combinada", "potente"), bty = "n",
+     col = 1:4, lty = 1)
```



O modelo exponencial se ajusta razoavelmente bem para os grupos em que foram observados mais óbitos.

3. Ajuste um outro modelo exponencial, adicionando variáveis de controle (sexo, idade e tipo de atendimento). Quais variáveis tiveram efeito significativo? Quais tiveram efeito protetor?

```
> mod2.ipec <- survreg(Surv(tempo, status) ~ factor(tratam) + sexo +
+     idade + factor(acompan), data = ipec, dist = "exp")
> summary(mod2.ipec)

Call:
survreg(formula = Surv(tempo, status) ~ factor(tratam) + sexo +
    idade + factor(acompan), data = ipec, dist = "exp")
                   Value Std. Error      z        p
(Intercept)      7.95467     0.6554 12.137 6.69e-34
factor(tratam)1  1.38695     0.2972  4.667 3.06e-06
factor(tratam)2  2.21397     0.4656  4.755 1.99e-06
factor(tratam)3  2.98559     1.0165  2.937 3.31e-03
sexoM           -0.07670     0.2833 -0.271 7.87e-01
idade           -0.00292     0.0120 -0.242 8.09e-01
factor(acompan)1 -1.70869     0.4064 -4.205 2.61e-05
factor(acompan)2 -2.23186     0.4664 -4.785 1.71e-06
```

```
Scale fixed at 1

Exponential distribution
Loglik(model)= -723.5   Loglik(intercept only)= -774.6
        Chisq= 102.25 on 7 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 6
n= 193
```

Resposta: Os fatores sexo e idade são não significativos. Dentre os fatores significativos somente tratamento teve um efeito protetor. O tipo de atendimento – internação hospitalar posterior ou imediata, comparadas a tratamento apenas ambulatorial – é significativo, e indica maior risco para pacientes que necessitam internação.

Exercício 6.4: Qual é o efeito da doença de base na sobrevida de pacientes em diálise, quando controlamos por idade? Usando a distribuição Weibull, ajuste o modelo:

```
> dialise <- read.table("dialise.csv", header = T, sep = ",")
> modelo1 <- survreg(Surv(tempo, status) ~ idade + cdiab + crim +
+     congenita, data = dialise)
> summary(modelo1)


Call:
survreg(formula = Surv(tempo, status) ~ idade + cdiab + crim +
    congenita, data = dialise)
              Value Std. Error       z        p
(Intercept)  6.7737    0.14999  45.161 0.00e+00
idade       -0.0428    0.00225 -19.017 1.24e-80
cdiab       -0.3605    0.07353  -4.903 9.44e-07
crim        -0.0384    0.08139  -0.472 6.37e-01
congenita    0.8855    0.27529   3.217 1.30e-03
Log(scale)   0.1951    0.02082   9.373 7.04e-21

Scale= 1.22

Weibull distribution
Loglik(model)= -7857.3   Loglik(intercept only)= -8104.2
        Chisq= 493.87 on 4 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 7
n= 6805
```

Resposta: A diabetes aumenta o risco e as doenças congênitas têm efeito protetor.

Existe evidência a favor da utilização de um modelo mais simples (exponencial)? Ou um modelo com menos variáveis? Remova as variáveis com p-valor menor do que 0,1 e compare o novo modelo com o modelo acima (uma dica: calcule a razão de verossimilhança entre os dois modelos).

Resposta:

Ajustando um modelo exponencial

```
> modeloE <- survreg(Surv(tempo, status) ~ idade + cdiab + crim +
+     congenita, dist = "exp", data = dialise)
> summary(modeloE)


Call:
survreg(formula = Surv(tempo, status) ~ idade + cdiab + crim +
    congenita, data = dialise, dist = "exp")
             Value Std. Error       z        p
(Intercept)  6.1643    0.10897  56.568 0.00e+00
idade       -0.0365    0.00176 -20.676 5.66e-95
cdiab       -0.3092    0.06029  -5.127 2.94e-07
crim        -0.0313    0.06696  -0.467 6.41e-01
congenita    0.7550    0.22616   3.338 8.43e-04

Scale fixed at 1

Exponential distribution
Loglik(model)= -7905.3   Loglik(intercept only)= -8169
        Chisq= 527.4 on 4 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 6
n= 6805
```

Comparando os modelos weibull e exponencial, isto é, testando a hipótese que o o parâmetro de forma γ é igual a 1, através da estatística de deviance

```
> dev <- 2 * (modelo1$loglik[2] - modeloE$loglik[2])
> dev


[1] 95.95186


> gl <- 1
> pvalor <- 1 - pchisq(dev, gl)
> pvalor


[1] 0
```

Alternativamente pode-se testar a redução do modelo usando o comando *anova()*. Observe que na coluna *Deviance* temos o mesmo valor que calculamos anteriormente (*dev = 95.95*)

```
> anova(modeloE, modelo1)

                             Terms Resid. Df    -2*LL Test Df Deviance
1 idade + cdiab + crim + congenita      6800 15810.56      NA       NA
2 idade + cdiab + crim + congenita      6799 15714.61    =  1 95.95186
     P(>|Chi|)
1           NA
2 1.177112e-22
```

Ajustando um modelo sem a covariável causas renais (*crim*)

```
> modelo2 <- survreg(Surv(tempo, status) ~ idade + cdiab + congenita,
+     data = dialise)
> summary(modelo2)

Call:
survreg(formula = Surv(tempo, status) ~ idade + cdiab + congenita,
    data = dialise)
              Value Std. Error      z        p
(Intercept)  6.7623    0.14798  45.70 0.00e+00
idade       -0.0428    0.00225 -19.00 1.70e-80
cdiab       -0.3510    0.07061  -4.97 6.69e-07
congenita    0.8951    0.27454   3.26 1.11e-03
Log(scale)   0.1951    0.02082   9.37 7.05e-21

Scale= 1.22

Weibull distribution
Loglik(model)= -7857.4   Loglik(intercept only)= -8104.2
        Chisq= 493.64 on 3 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 7
n= 6805
```

Testando a hipótese nula de que o modelo reduzido é melhor, ou em outras palavras, que o coeficiente da covariável *crim* não é significativo.

```
> dev <- 2 * (modelo1$loglik[2] - modelo2$loglik[2])
> dev

[1] 0.2218873
```

```
> gl <- modelo1$df - modelo2$df
> gl

[1] 1

> pvalor <- 1 - pchisq(dev, gl)
> pvalor

[1] 0.6376056
```

Alternativamente pode-se testar a redução do modelo usando o comando anova

```
> anova(modelo2, modelo1)

                            Terms Resid. Df    -2*LL  Test Df  Deviance
1        idade + cdiab + congenita      6800 15714.83       NA        NA
2 idade + cdiab + crim + congenita      6799 15714.61 +crim  1 0.2218873
  P(>|Chi|)
1        NA
2 0.6376056
```

Segundo o teste apresentado no sumário do modelo1 (modelo Weibull) e no teste de razão de verossimilhança, a estimativa do parâmetro de escala é significativamente diferente de 1 e portanto a redução para um modelo com menos parâmetros (exponencial) não seria adequada.

Adotando-se o modelo Weibull, testou-se a redução do modelo por exclusão da variável ***crim***, e o modelo reduzido não pode ser rejeitado sendo possível a retirada de tal variável do modelo. Isto significa que, segundo o modelo Weibull, a variável ***crim*** não é um fator prognóstico importante para a sobrevida.