Modelos de regressão paramétrica II

Vicente Garibay Cancho

O modelo log-logistica

A distribuição log-logística acomoda um modelo TFA, mas não um modelo RP.

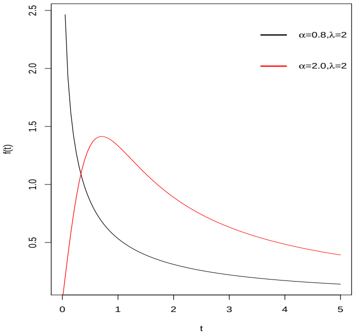
A função de risco

$$h(t) = \frac{\alpha\lambda t^{\alpha-1}}{1+\lambda t^{\alpha}},$$

com $\alpha > 0$ e $\lambda > 0$.

Formas da função de risco

- $\alpha \leq 1$: é decrescente;
- $\alpha > 1$: é unimodal.



O modelo log-logistica

A função de sobrevivência é dada por

$$S(t) = (1 + \lambda t^{\alpha})^{-1},$$

e correspondente f.d.p é dada por

$$f(t) = \frac{\alpha \lambda t^{\alpha - 1}}{(1 + \lambda t^{\alpha})^{2}}.$$

O $p-$ésimo quantil, denotado por t_p

$$t_p = \lambda^{-1/\alpha} \left[\frac{p}{1-p} \right]^{1/\alpha}, \; 0 < p < 1.$$

A mediana

$$t_{0,5} = \lambda^{-1/\alpha}.$$

O modelo logistica

Se T é uma v.a. que segue uma distribuição log-logístico com parâmetros $\alpha, \lambda > 0$, então a v.a. $Y = \log(T)$ segue uma distribuição logística com f.d.p. dada por

$$f(y) = \frac{1}{\sigma} \exp\left\{\frac{y-\mu}{\sigma}\right\}\left(1 + \exp\{\frac{y-\mu}{\sigma}\}\right)^{-2}, \; y \in R,$$

com $\mu \in R$ e $\sigma > 0$ os parâmetros de locação e escala, respectivamente.
Os parâmetros da distribuição log-logistica e logistica relacionadas pelas mesmas funções apresentadas no modelo Weibull, is é,

$$\alpha = 1/\sigma, \;\; e\lambda = \exp\{-\mu/\sigma\}.$$

O modelo logistica

As funções de sobrevivência e de risco são respectivamente,

$$S(y) = \left(1 + \exp\{\frac{y-\mu}{\sigma}\}\right)^{-1}$$

e

$$h(y) = \frac{1}{\sigma}\exp\left\{\frac{y-\mu}{\sigma}\right\}\left(1 + \exp\{\frac{y-\mu}{\sigma}\}\right)^{-1}, \; y \in R,$$

Como no modelo valor extremo, a distribuição logística tem a seguinte representação estocástica

$$y = \mu + \sigma\varepsilon,$$

onde ε é uma v.a. com distribuição logítico padrão ($\mu = 0$ e $\sigma = 1$), com f.d.p dada por

$$f(z) = \exp\{z\}\left(1 + \exp\{z\}\right)^{-2}, \; z \in R,$$

Modelos de chances proporcionais "proportional odds (PO)"

Em um modelo de sobrevivência de PO, a razão de chances é constante ao longo do tempo

- Chances de sobrevivência: chances de sobreviver além do tempo t
$$\frac{S(t)}{1-S(t)} = \frac{P(T>t)}{P(T\leq t)}.$$
- Chances de falhas: chances de obter o evento no tempo t
$$\frac{F(t)}{1-F(t)} = \frac{1-S(t)}{S(t)} = \frac{P(T\leq t)}{P(T>t)}.$$

Modelo PO

A chance de falha do modelo log-logística é dada por

$$\frac{1 - S(t)}{S(t)} = \frac{\lambda t^{\alpha}(1 + \lambda t^{\alpha})^{-1}}{(1 + \lambda t^{\alpha})^{-1}} = \lambda t^{\alpha}.$$

A razão de chance (odds ratio ,OR) para dois diferentes grupos (1 e 2) com α fixo, é

$$OR(1 \; vs \; 2) = \frac{\frac{1 - S_1(t)}{S_1(t)}}{\frac{1 - S_2(t)}{S_2(t)}} = \frac{\lambda_1 t^{\alpha}}{\lambda_2 t^{\alpha}} = \frac{\lambda_1}{\lambda_2}$$

Daí o modelo log-logístico é um modelo de PO.

Adequabilidade do modelo log-logístico

O log da chance de falha pode-se escrever como

$$\log\left(\frac{1-S(t)}{S(t)}\right) = \log\lambda + \alpha log t,$$

a qual é uma função linear em relação $\log(t)$.

- Esta propriedade permite uma avaliação gráfica da adequação de um modelo log-logístico plotando

$$\log\left(\frac{1-\widehat{S(t)}}{\widehat{S(t)}}\right) \quad vs \quad \log(t)$$

onde $\widehat{S(t)}$ é a estimativa K-M da função de sobrevivência.

Dados de Leucemia

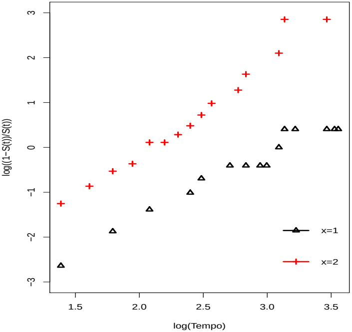
Os dados correspondem ao um estudo de tempos de remissão (em semanas) de pacientes de leucemia

	Tratamento		Placebo	
	time	log WBC	time	log WBC
1	6.00	2.31	1.00	2.80
2	6.00	4.06	1.00	5.00
3	6.00	3.28	2.00	4.91
4	7.00	4.43	2.00	4.48
5	10.00	2.96	3.00	4.01
6	13.00	2.88	4.00	4.36
7	16.00	3.60	4.00	2.42
8	22.00	2.32	5.00	3.49
9	23.00	2.57	5.00	3.97
10	6.00+	3.20	8.00	3.52
11	9.00+	2.80	8.00	3.05
12	10.00+	2.70	8.00	2.32
13	11.00+	2.60	8.00	3.26
14	17.00+	2.16	11.00	3.49
15	19.00+	2.05	11.00	2.12
16	20.00+	2.01	12.00	1.50
17	25.00+	1.78	12.00	3.06
18	32.00+	2.20	15.00	2.30
19	32.00+	2.53	17.00	2.95
20	34.00+	1.47	22.00	2.73
21	35.00+	1.45	23.00	1.97

Dados de Leucemia

A covariável $\log(WBC)$ foi categorizado

- $x = 1$, se $\log WBC < 2$,
- $x = 2$ se $\log WBC \geq 2$



$\Longrightarrow$ linhas aproximadamente retas indica o modelo log-logístico

O modelo TFA log-logística

O modelo TFA log-logística com um único preditor x

$$S(t) = \frac{1}{1 + \lambda t^{\alpha}} = \frac{1}{1 + (\lambda^{\frac{1}{\alpha}} t)^{\alpha}}$$

resolvendo em t tem-se

$$t = \left[\frac{1}{S(t)} - 1\right]^{\frac{1}{\alpha}} \times \frac{1}{\lambda^{1/\alpha}}.$$

Parametrizamos o fator da direita como

$$\frac{1}{\lambda^{1/\alpha}} = \exp\{\beta_0 + \beta_1 x\}.$$

Dai tem-se

$$t = \left[\frac{1}{S(t)} - 1\right]^{\frac{1}{\alpha}} \times \exp\{\beta_0 + \beta_1 x\}.$$

O modelo TFA log-logística: Fator de aceleração

Em termos de qualquer probabilidade fixa $S(t_q) = q$ obtemos

$$t_q = \left[q^{-1} - 1\right]^{\frac{1}{\alpha}} \times \exp\{\beta_0 + \beta_1 x\}.$$

O tempo mediano

$$t_{0,5} = [2 - 1]^{\frac{1}{\alpha}} \times \exp\{\beta_0 + \beta_1 x\} = \exp\{\beta_0 + \beta_1 x\}$$

O fator de aceleração γ para $S(t_q) = q$

$$\gamma(x = 2 \text{ } vs \text{ } x = 1) = \frac{\left[q^{-1} - 1\right]^{\frac{1}{\alpha}} \exp\{\beta_0 + 2\beta_1\}}{[q^{-1} - 1]^{\frac{1}{\alpha}} \exp\{\beta_0 + \beta_1\}} = \exp\{\beta_1\}$$

Codigo em R e Saida

```
fit=survreg(Surv(t, delta)~x1, dist="loglogistic")
summary(fit)
Call: survreg(formula = Surv(t, delta) ~ x1, dist = "loglogistic")
```

	Value	Std. Error	z	p
(Intercept)	4.094	0.586	6.98	2.92e-12
x1	-0.987	0.337	-2.93	3.40e-03
Log(scale)	-0.564	0.154	-3.67	2.41e-04

Scale= 0.569
Log logistic distribution
Loglik(model)= -111.2 Loglik(intercept only)= -115.4
Chisq= 8.28 on 1 degrees of freedom, p= 0.004
Number of Newton-Raphson Iterations: 4
n= 42

Codigo em R e Saida: Fator de aceleração

O fator de aceleração estimada comparando $x = 2$ (alta contagem) e $x = 1$ (contagem média) é agora:

$$\hat{\gamma} = \exp\{\hat{\beta}_1\} \qquad \begin{array}{l} > exp(fit\$coefficient[2]) \\ \text{x1} \\ 0.3728214 \end{array}$$

$$\Longrightarrow, \quad S_1(t) = S_2(0.37t)$$

Interpretação

O tempo de sobrevivência para o grupo com alta contagem (x = 2) é "Acelerada" por um factor de 0,37 em comparação com o grupo com contagem média (x = 1).

Comparação dos modelos PO e TFA log-logístico

- Modelo TFA: $\frac{1}{\lambda^{1/\alpha}} = \exp\{\beta_0 + \beta_1 x\}$.
- Modelo PO: $\lambda = \exp\{b_0 + b_1 x\}$.

Daí temos a seguinte relação

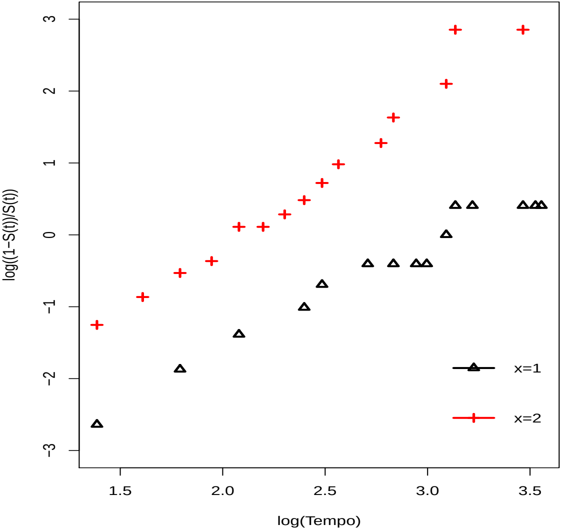
$$b_0 = -\alpha\beta_0, \ e \ b_1 = -\alpha\beta_1.$$

Nota: Se α é fixo: $TFA \Longleftrightarrow PO$
Assim, podemos calcular a estimativa da razão de chances (odds ratio, OR) com os coeficientes de regressão estimada do modelo TFA por:

$$\begin{aligned}\widehat{OR} &= \exp\{\hat{b_1}\}\\ &= \exp\{-\hat{\alpha}\hat{\beta}_1\} = 5,66.\end{aligned}$$

```
> beta1 <- fit$coefficient[2]
> alpha <- 1/fit$scale
> exp(-beta1*alpha)
5.662691
```

Avaliação gráfica



Avaliação gráfica

Proposição

1. Linhas retas $\Longrightarrow$ Log-logística.
2. Retas paralellas e log-logística $\Longrightarrow$ PO.
3. Log-logística e PO $\Longrightarrow$ TFA.

Prova: Considere dois grupos (1 e 2)

1. $\log\left(\frac{1-S(t)}{S(t)}\right) = \log(\lambda) + \alpha log(t)$.
2. Retas paralellas $\Longrightarrow$ α é a mesma para ambos grupos $\Longrightarrow$ $OR = \frac{t^\alpha \lambda_1}{t^\alpha \lambda_2} = \frac{\lambda_1}{\lambda_2}$
3. Para $S(t_q) = q$, o fator de aceleração γ

$$\gamma = \frac{\left[q^{-1} - 1\right]^{\frac{1}{\alpha}} \lambda_1}{\left[q^{-1} - 1\right]^{\frac{1}{\alpha}} \lambda_2} = \frac{\lambda_1}{\lambda_2}$$

Modelo log-Normal

Uma v.a T tem distribuição log-normal se sua f.d.p é dada por

$$f(t) = \frac{1}{\sqrt{2\pi} t\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{\log t - \mu}{\sigma} \right)^2 \right\}, \ t > 0,$$

onde $\mu \in R$ é média de log T, assim como $\sigma > 0$ é o desvio padrão, ou seja, log $T \sim N(\mu, \sigma^2)$.

A média e variância

$$E(T) = \exp\{\mu + \sigma^2/2\}, \ \ e \ \ Var(T) = \exp\{2\mu + \sigma^2\}(\exp\{\sigma^2\} - 1).$$

O q-ésimo quantil

$$t_q = \exp\{\mu + \sigma z_q\}$$

onde z_q é q-ésimo quantil da distribuição normal padrão.

Modelo log-Normal

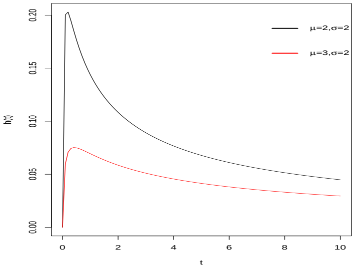
A função de sobrevivência não tem uma forma analática explicita e é representado por

$$S(t) = \Phi\left(-\frac{\log t - \mu}{\sigma}\right)$$

onde $\Phi(.)$ é a função de distribuição acumulada da distribuição normal padrão.

A correspondente função de risco é dado por:

$$h(t) = \frac{\exp\left\{-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right\}}{\sqrt{2\pi}\, t\sigma\Phi\left(-\frac{\log t - \mu}{\sigma}\right)}.$$



Adequabilidade do log-Normal

- $\Phi^{-1}(S(t))$ é linear em relação a $\log(t)$.

$$S(t) = \Phi\left(-\frac{\log t - \mu}{\sigma}\right), \Longrightarrow \Phi^{-1}(S(t)) = -\frac{\log t - \mu}{\sigma}$$

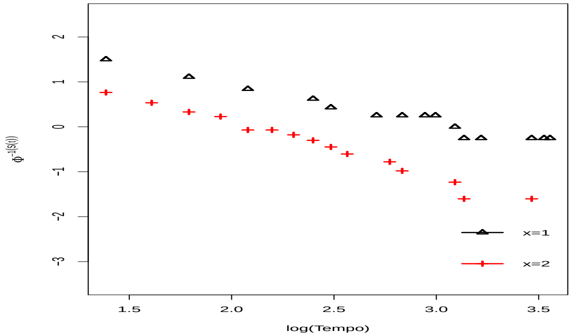
$$\Longrightarrow \log(-\log(S(t))) = \frac{\mu}{\sigma} - \frac{\log t}{\sigma}.$$

- Esta propriedade permite uma avaliação gráfica da adequação de um modelo Log-normal plotando

$$\Phi^{-1}(\widehat{S(t)}) \ \ vs \ \ \log(t)$$

onde $\widehat{S(t)}$ é a estimativa K-M da função de sobrevivência.

Gráfico da $\Phi^{-1}(\widehat{S(t)})$ *vs* $\log(t)$ para os dados de leucemia.



▷ Linea reta $\Longrightarrow$ log-normal.

O modelo TFA log-normal

O modelo TFA log-normal com un único preditor x

$$S(t) = \Phi\left(-\frac{\log t - \mu}{\sigma}\right)$$

resolvendo em t tem-se

$$t = \exp\left\{\Phi^{-1}\left(S(t)\right)\sigma + \mu\right\}.$$

Parametrizamos o parâmetro de locação, μ

$$\mu = \beta_0 + \beta_1 x.$$

Dai tem-se

$$t = \exp\left\{\Phi^{-1}\left(S(t)\right)\sigma\right\}\exp\{\beta_0 + \beta_1 x\}.$$

O modelo TFA log-normal: Fator de aceleração

Em termos de qualquer probabilidade fixa, $S(t_q) = q$ obtemos

$$t = \exp\left\{\Phi^{-1}(q)\,\sigma\right\}\exp\{\beta_0 + \beta_1 x\}.$$

O tempo mediano

$$t_{0,5} = \exp\left\{\Phi^{-1}(0,5)\,\sigma\right\}\exp\{\beta_0 + \beta_1 x\} = \exp\{\beta_0 + \beta_1 x\}$$

O fator de aceleração γ para $S(t_q) = q$

$$\gamma(x = 2 \ vs \ x = 1) = \frac{\exp\left\{\Phi^{-1}(q)\,\sigma\right\}\exp\{\beta_0 + 2\beta_1\}}{\exp\left\{\Phi^{-1}(q)\,\sigma\right\}\exp\{\beta_0 + \beta_1\}} = \exp\{\beta_1\}$$

Ajuste dos dados de Leucemia do modelo TFA log-normal

```
fit.ln=survreg(Surv(t, delta)~x1, dist="lognormal")
summary(fit.ln)
Call: survreg(formula = Surv(t, delta)~ x1, dist = "lognormal")
```

	Value	Std. Error	z	p
(Intercept)	4.1851	0.598	6.9926	2.70e-12
x1	-1.0540	0.343	-3.0760	2.10e-03
Log(scale)	-0.0127	0.135	-0.0942	9.25e-01

```
Scale= 0.987
Log Normal distribution
Loglik(model)= -110.9509 Loglik(intercept only)= -115.4
Chisq= 8.88 on 1 degrees of freedom, p= 0.0029
Number of Newton-Raphson Iterations: 4
n= 42
```

Codigo em R e Saida: Fator de aceleração

O fator de aceleração estimada comparando $x = 2$ (alta contagem) e $x = 1$ (contagem média) é agora:

$$\hat{\gamma} = \exp\{-\hat{\beta}_1\} = 0,35 \qquad \begin{array}{l} > exp(fit.ln\$coefficient[2]) \\ \text{x1} \\ 0.3485551 \end{array}$$

$$\Longrightarrow, \quad S_1(t) = S_2(0.35t)$$

Interpretação

O tempo de sobrevivência para o grupo com alta contagem (x = 2) é "Acelerada" por um factor de 0,35 em comparação com o grupo com contagem média (x = 1). Então x elevado tem efeito negativo no tempo mediano.

Modelando o parâmetro de forma

Muitos modelos paramétricos contem um parâmetro de forma, que é geralmente considerado fixo

Exemplo

1. O modelo Weibull
   Relembre $h(t) = \alpha\lambda t^{\alpha-1}$,onde $\lambda = \exp\{b_0 + b_1 TRAT\}$ e α parâmetro de forma não é afetado por covariável.
2. Alternativamente o modelo Weibull $h(t) = \alpha\lambda t^{\alpha-1}$,onde $\lambda = \exp\{b_0 + b_1 TRAT\}$ e $\alpha = \exp\{\delta_0 + \delta_1 TRAT\}$.

- Se $\delta_1 \neq 0$ o valor de α difere para TRT.
- Se $\delta_1 \neq 0$ o modelo não é TFA e RP.

Critérios de seleção de modelos

- Ao selecionarmos modelos é preciso ter em mente que não existem modelos verdadeiros. Há apenas modelos aproximados da realidade.
- Deste modo, é necessário fazer a seleção do ?melhor? modelo, dentre aqueles que foram ajustados, para explicar o fenômeno sob estudo.
- Akaike (1974) utilizou a Informação de Kullback-Leibler para testar se um dado modelo é adequado.

Critério de informação de Akaike - AIC

O AIC é definido por

$$AIC = -2\ell(\hat{\boldsymbol{\theta}}) + 2p$$

onde $\ell(.)$ é a log-verossimilhança e p número de parâmetros do modelo.

Critérios de seleção de modelos

Critério de informação bayesiano - BIC

O BIC proposto por Schwarz (1978) é definido por

$$BIC = -2\ell(\hat{\boldsymbol{\theta}}) + p \log\ n$$

onde $\ell(.)$ é a log-verossimilhança e p número de parâmetros do modelo e n tamanho da amostra.

O melhor modelo é aquele AIC ou BIC com menor valor.

Comparação entre os modelos TFA log-logístico e log-normal

Na Tabela apresenta-se as estatística AIC e BIC para os dados de leucemina.

	AIC	BIC
log-logistico	226.42	306.42
log-normal	225.90	305.90

Ambos critérios indicam o modelo log-normal que ajusta melhor os dados de leucemia.

Exercícios

1. Ajuste os modelos de TFA e de PO log-logístico aos dados de leucemia com as duas covariáveis, interprete os parâmetros em termos do problema e faça um estudo de adequabilidade do modelo.
2. Ajuste o modelo de TFA log-normal aos dados de leucemia com as duas covariáveis interprete os parâmetros em termos do problema e faça um estudo de adequabilidade do modelo.