Análise de sobrevivência e confiabilidade

Modelos paramétricos

Prof. Paulo Cerqueira Jr
Programa de Pós-Graduação em Matemática e Estatística - PPGME
Instituto de Ciências Exatas e Naturais - ICEN

https://github.com/paulocerqueirajr

Introdução

Introdução

- O método de máxima verossimilhança somente deve ser aplicado após ter sido definido um modelo probabilístico adequado para os dados.
- Se um modelo for usado inadequadamente, toda a análise estatística fica comprometida e consequentemente as respostas às perguntas de interesse ficam distorcidas.
- A ideia empírica consiste em ajustar os modelos probabilísticos típicos para dados de tempos de vida e com base na comparação entre valores estimados e observados, decidir qual deles "melhor" se ajusta aos dados amostrais.
- A escolha da "melhor" distribuição pode ser feita por meio de técnicas gráficas ou testes de hipóteses com modelos encaixados.

Introdução

- Método Gráfico.
  - Modelo candidato $\times$ Kaplan-Meier.
  - Linearização do modelo.
- Teste de hipóteses: utilizar o TRV para comparar o modelo proposto com a gama generalizada.
- Critérios informação.
  - Critérios de informação de Akaike (AIC) e Bayesiano(BIC).

Métodos gráficos

- Comparação direta da $S(t)$ estimada do modelo proposto com o Kaplan-Meier (no mesmo gráfico) e `selecionar o modelo cuja curva melhor se aproximar da curva`.
- Duas formas:
  - $S_{KM}(t) \times S_{MP}(t)$;
  - $S_{KM}(t) \times \text{tempo}$ e $S_{MP}(t) \times \text{tempo}$, no mesmo gráfico.
- Linearização da $S(t)$ para comparação com uma reta.

Métodos gráficos

> Exponencial:

$$S(t) = \exp\{-\alpha t\} \to -\log[S(t)] = \alpha t.$$

> Weibull:

$$S(t) = \exp\{-(\alpha t)^\gamma\} \to -\log[S(t)] = (\alpha t)^\gamma \to \log[-\log[S(t)]] = \gamma \log(\alpha) + \gamma \log(t).$$

> Lognormal:

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

Comparação para modelos encaixados

Teste da razão de verossimilhanças

- Como foi dito anteriormente, as técnicas gráficas são extremamente úteis na seleção de modelos.
- Entretanto, as conclusões a partir delas podem diferir para diferentes analistas.
- Outra forma de discriminar modelos é através de testes de hipóteses e este caso, não envolve qualquer componente subjetivo na sua interpretação.
- As hipóteses:

$$H_0 : \text{O modelo é adequado} \times H_1 : \text{O modelo não é adequado}$$

Comparação para modelos encaixados

Teste da razão de verossimilhanças

- Neste teste, usa-se a razão de verossimilhanças em modelos encaixados (Cox e Hinkley, 1977).
- Ou seja, devemos usar um modelo generalizado tal que os de interesse sejam casos particulares.
- O teste:
  - Calcular $\log L(\hat{\theta}_G)$;
  - Calcular $\log L(\hat{\theta}_M)$.

Comparação para modelos encaixados

Teste da razão de verossimilhanças

- A estatística da razão de verossimilhanças

$$TRV = -2 \log \left[\frac{L(\hat{\theta}_M)}{L(\hat{\theta}_G)} \right] = 2 \log \Big[\log L(\hat{\theta}_G) - \log L(\hat{\theta}_M) \Big],$$

sob H_0, tem aproximadamente uma distribuição qui-quadrado com graus de liberdade igual a diferença do número de parâmetros dos modelos sendo comparados.

Critérios de informação

- Após encontrar as estimativas dos parâmetros de interesse, os critérios de informação são usados para selecionar o melhor modelo ajustado aos dados em questão.

> Critério de informação de Akaike (AIC):

$$\mathrm{AIC} = -2\ell(\theta \mid D) + 2k,$$

em que k é o número de parâmetros.

- Na presença de observações censuradas, o AIC pode não ser adequado. Dessa forma, Liang (2008), propôs uma versão modificada do AIC para dados censurados, e é expressa por

$$\mathrm{AIC}_{\mathrm{SUR}} = -2\ell(\theta \mid D) + 2k + \frac{2(k+2)(k+3)}{n-k-3},$$

em que k é o número de parâmetros, n é o tamanho amostral.

Critérios de informação

> Critério de infomação Bayesiano(BIC):

- O critério de informação bayesiano é formalmente definido por

$$BIC = -2\ell(\theta \mid D) + k \log(n).$$

- Para cada critério definido, `o menor valor indica a melhor qualidade de ajuste aos dados de interesse`.

Exemplo:

- Dados provenientes da UFPR.
- 20 pacientes com câncer de bexiga submetidos a um procedimento cirúrgico a laser.
- Resposta: tempo da cirurgia até a reincidência da doença (meses).
- Objetivo: mediano de vida destes pacientes.
- Dados (em meses): 17 falhas e 3 censuras.
- Modelos ajustados: `Exponencial`, `Weibull` e `Lognormal` e `Gama generalizado`.

```r
require(survival)
require(flexsurv)

#require(survminer)

tempos<-c(3,5,6,7,8,9,10,10,12,15,15,18,
          19,20,22,25,28,30,40,45)
cens<-c(1,1,1,1,1,1,1,0,1,1,0,1,1,1,1,1,
        1,1,1,0)
dados <- data.frame(tempos, cens)
ekm <- survfit(Surv(tempos, cens)~1)
```

```r
# Exponencial:
ajust1 <- survreg(Surv(tempos,cens)~1,
                  dist='exponential')
alpha  <- 1/exp(ajust1$coefficients[1])

# Weibull
ajust2  <- survreg(Surv(tempos,cens)~1,
                  dist='weibull')
alpha.w <- exp(ajust2$coefficients[1])
gama    <- 1/ajust2$scale
```

```r
# lognormal:
ajust3 <- survreg(Surv(tempos,cens)~1,
                 dist='lognormal')
mu    <- ajust3$coefficients[1]
sigma <- ajust3$scale

# gamma:
ajust4 <- flexsurvreg(Surv(tempos, cens) ~
                       dist = "gamma")
shape.g <- 1/ajust4$coefficients[1]
rate.g <- exp(ajust4$coefficients[2])
```

```r
# gamma Gen:
ajust5 <- flexsurvreg(Surv(tempos, cens) ~
                       dist = "gengamma.orig
shape.gg    <- exp(ajust5$coefficients[1])
scale.gg <- exp(ajust5$coefficients[2])
k.gg     <- exp(ajust5$coefficients[3])
```

Exemplo:

Função de sobrevivência estimada

> Exponencial:

$$S_E(t) = \exp\{-0.05t\}$$

> Weibull:

$$S_W(t) = \exp\left\{-0.05t^{1.54}\right\}$$

> Lognormal:

$$S_{LN}(t) = \Phi\left[-(\log(t) - 2.72)/0.76\right]$$

- `Modelo Gama` e `Gama Generalizado` obtidos numericamente.

Exemplo - Método gráfico

```r
#--- Função de sobrevivência:

tempo <- ekm$time
ste   <- exp(-alpha*tempo)
stw   <- pweibull(q = tempo, shape = gama, scale = alpha.w, lower.tail = FALSE)
stln  <- pnorm( -(log(tempo)-mu)/sigma )
stgamma <- pgamma(q = tempo, shape = shape.g, rate=rate.g,
                  lower.tail = FALSE)
stggamma <- pgengamma.orig(q = tempo, shape = shape.gg,
                           scale =scale.gg, k = k.gg,
                           lower.tail = FALSE)

cbind(tempo, ekm=ekm$surv, ste, stw, stln, stgamma, stggamma)
```

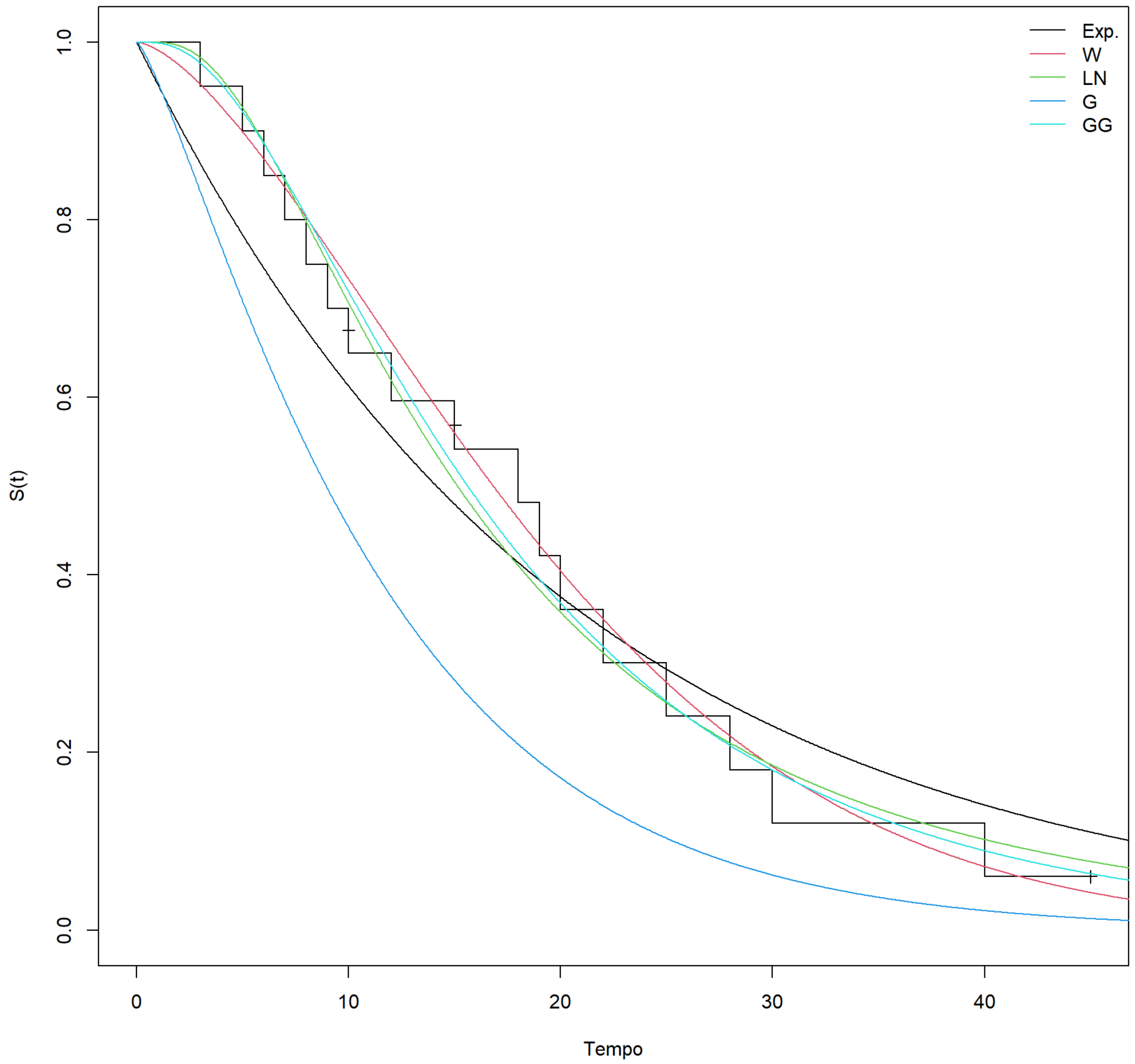
```
      tempo        ekm       ste        stw       stln    stgamma   stggamma
 [1,]     3 0.95000000 0.8633164 0.95274148 0.98283934 0.83180159 0.97656309
 [2,]     5 0.90000000 0.7827384 0.89897484 0.92624322 0.70824617 0.92159117
 [3,]     6 0.85000000 0.7453152 0.86839357 0.88685752 0.65037389 0.88550195
 [4,]     7 0.80000000 0.7096812 0.83609525 0.84337638 0.59583545 0.84591595
 [5,]     8 0.75000000 0.6757509 0.80253272 0.79781416 0.54482327 0.80423301
 [6,]     9 0.70000000 0.6434428 0.76809812 0.75169629 0.49737649 0.76158117
 [7,]    10 0.65000000 0.6126794 0.73313414 0.70611769 0.45343783 0.71883800
 [8,]    12 0.59583333 0.5554947 0.66278292 0.61931883 0.37557570 0.63555373
 [9,]    15 0.54166667 0.4795676 0.55966698 0.50475984 0.28122913 0.52148679
[10,]    18 0.48148148 0.4140186 0.46346069 0.41042396 0.20931642 0.42390649
[11,]    19 0.42129630 0.3942241 0.43345774 0.38317692 0.18948996 0.39512627
```

Exemplo - Método gráfico

Estimador de Kaplan-Meier $\times$ modelo ajustado.

```
tempo <- seq(0.0001, 50, length.out=1000)
ste   <- exp(-alpha*tempo)
stw   <- pweibull(q = tempo, shape = gama,
                  scale = alpha.w,
                  lower.tail = FALSE)
stln  <- pnorm( -(log(tempo)-mu)/sigma )
stgamma <- pgamma(q = tempo, shape = shape.
                  rate=rate.g, lower.tail =
stggamma <- pgengamma.orig(q=tempo, shape=s
                           scale =scale.gg,
                           lower.tail = FAL
```

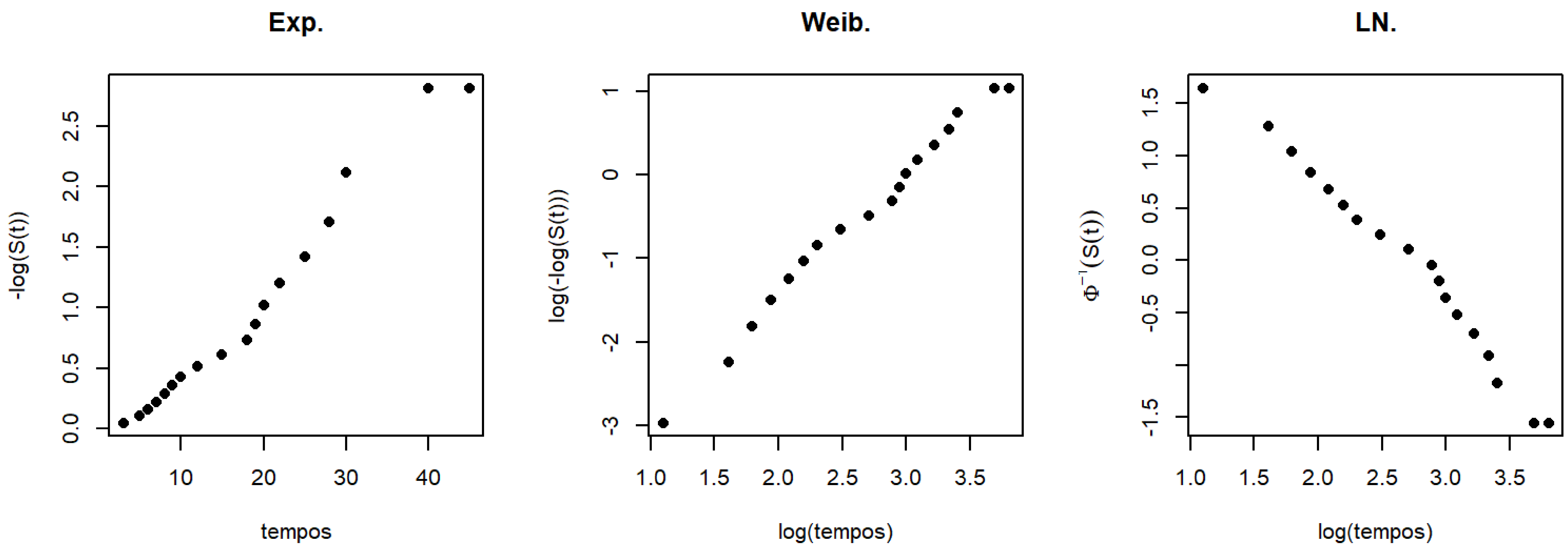
```
plot(ekm, xlab="Tempo",
  ylab="S(t)", mark.time=TRUE, conf.int=FAL
lines(tempo, ste, col=1)
lines(tempo, stw, col=2)
lines(tempo, stln, col=3)
lines(tempo, stgamma, col=4)
lines(tempo, stggamma, col=5)
legend("topright", c("Exp.", "W", "LN","G",
```



Exemplo - Método gráfico

Linearização

```
par(mfrow=c(1,3))
st    <-ekm$surv
invst <-qnorm(st)

plot(ekm$time, -log(st),pch=16,xlab="tempos",ylab="-log(S(t))", main="Exp.")
plot(log(ekm$time),log(-log(st)),pch=16,xlab="log(tempos)",ylab="log(-log(S(t)))", main="Weib.")
plot(log(ekm$time),invst, pch=16,xlab="log(tempos)", ylab=expression(Phi^-1 * (S(t))), main="LN.")
```



Exemplo - TRV

- Comparação do modelo de interesse como o modelo Geral

```r
trv_ge <- round(-2*( ajust1$loglik[1]-ajust5$loglik),2)
p_trv_ge <- pchisq(trv_ge, df = 2, lower.tail = FALSE)
trv_gw <- round(-2*( ajust2$loglik[1]-ajust5$loglik),2)
p_trv_gw <- pchisq(trv_gw, df = 1, lower.tail = FALSE)
trv_gln <- round(-2*( ajust3$loglik[1]-ajust5$loglik),2)
p_trv_gln <- pchisq(trv_gln, df = 1, lower.tail = FALSE)
trv_gg <- round(-2*( ajust4$loglik-ajust5$loglik),2)
p_trv_gg <- pchisq(trv_gg, df = 1, lower.tail = FALSE)
```

Modelo	$\log L(\theta)$	**TRV**	**p-valor**
Gama gen.	-65.69	-	-
Exponencial	-68.27	5.16	0.075774
Weibull	-66.13	0.88	0.3482017
Log Normal	-65.74	0.09	0.7641772
Gama	-65.85	0.31	0.5776802

Critérios de informação

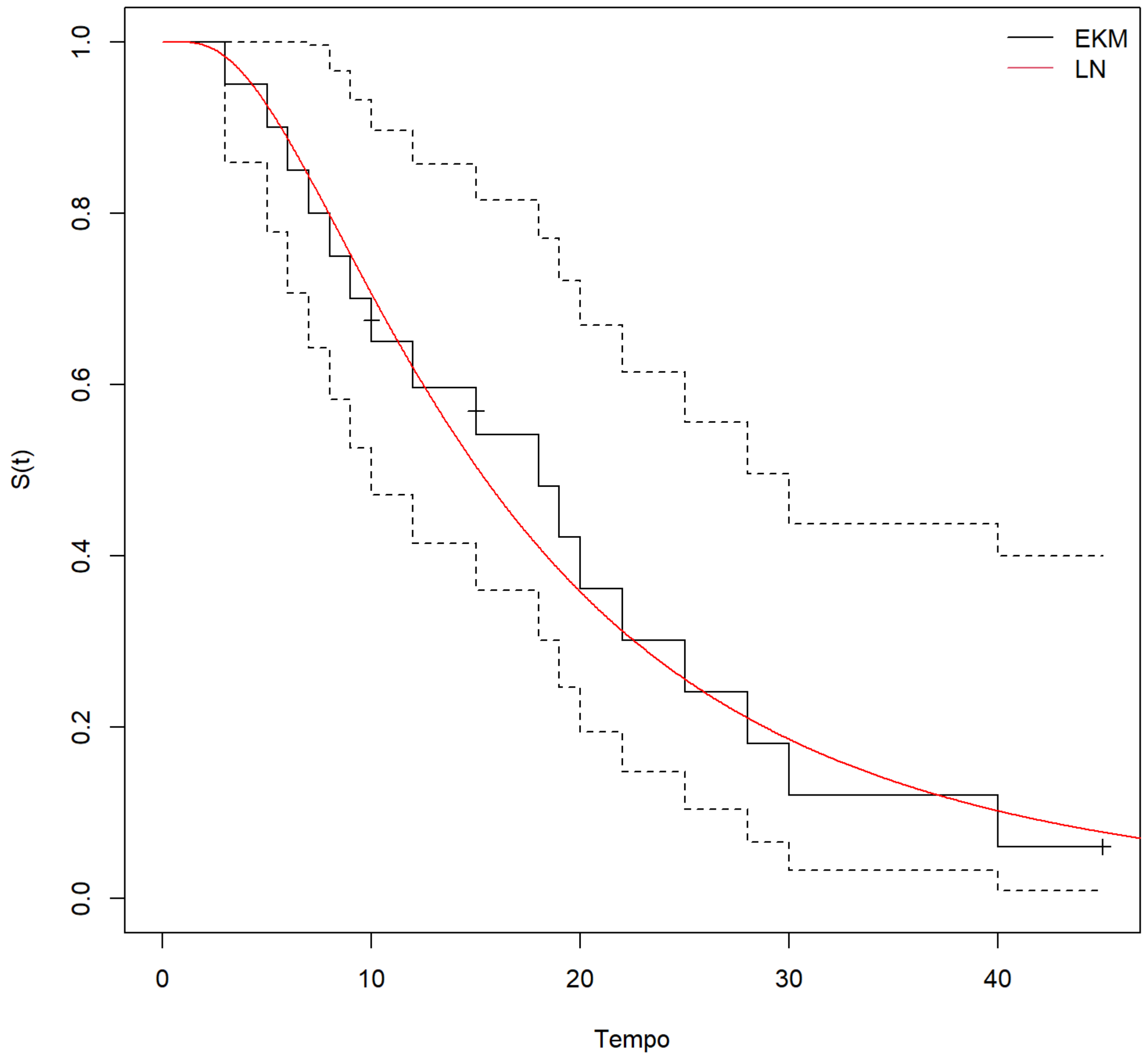
Menores valores de AIC, AIC_{Sur} e BIC indicam uma melhor qualidade de ajuste aos dados.

```r
aic_e <- round(AIC(ajust1, k=1),2)
aic_w <- round(AIC(ajust2),2)
aic_ln <- round(AIC(ajust3),2)
aic_g <- round(AIC(ajust4),2)
aic_gg <- round(ajust5$AIC,2)

aic_s_e <- round(AIC(ajust1, k=1)+((2*(1+2)*(1+3))/(dim(dados)[1])-1-3 ),2)
aic_s_w <- round(AIC(ajust2)+((2*(2+2)*(2+3))/(dim(dados)[1])-2-3 ),2)
aic_s_ln <- round(AIC(ajust3)+((2*(2+2)*(2+3))/(dim(dados)[1])-2-3 ),2)
aic_s_g <- round(AIC(ajust4)+((2*(2+2)*(2+3))/(dim(dados)[1])-2-3 ),2)
aic_s_gg <- round(ajust5$AIC+((2*(2+2)*(2+3))/(dim(dados)[1])-2-3 ),2)

bic_e  <- round(BIC(ajust1),2)
bic_w  <- round(BIC(ajust2),2)
bic_ln <- round(BIC(ajust3),2)
```

Modelo	AIC	AIC_{surv}	BIC
Exponencial	137.55	134.75	139.54
Weibull	136.27	133.27	136.27
Log Normal	135.48	132.48	137.47
Gama	135.7	132.7	137.69
Gama Gen.	137.39	134.39	140.37

Estimativas do modelo final

```r
plot(ekm, xlab="Tempo",
  ylab="S(t)", mark.time=TRUE, conf.int=TRU
lines(tempo, stln, col="red")
legend("topright", c("EKM", "LN"),
       col=1:2, lty=1, bty="n")
```



Estimativas sobre o tempo:

> Tempo mediano (EKM):

$$\hat{t}_{0,5;EKM} = 18 \text{ meses.}$$

> Tempo mediano (Log-normal):

$$\hat{t}_{0,5} = 15.1375069 \text{ meses.}$$

> Tempo medio (Log-normal):

$$\widehat{E(T)} = 20.2802566 \text{ meses.}$$