Análise de Sobrevivência - Modelo de Regressão de Cox

José Luiz Padilha

Julho de 2024

Exemplo 1: Sobrevida de Pacientes com Leucemia Aguda

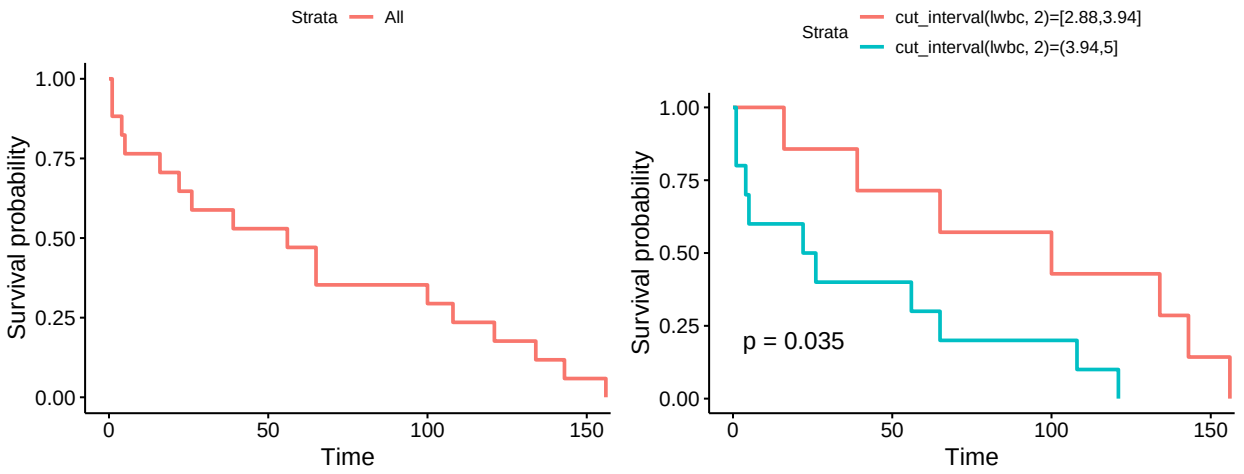
Retomemos os dados de leucemia aguda. Temos disponíveis os tempos de sobrevivência (em semanas) de 17 pacientes e suas contagens de glóbulos brancos (WBC) e seus correspondentes logaritmos, na base 10.

```
pacman::p_load(survival, survminer, tidyverse)
temp <- c(65, 156, 100, 134, 16, 108, 121, 4, 39, 143, 56, 26, 22, 1, 1, 5, 65)
cens <- rep(1, 17)
lwbc <- c(3.36, 2.88, 3.63, 3.41, 3.78, 4.02, 4.00, 4.23, 3.73, 3.85, 3.97, 4.51, 4.54,
          5.00, 5.00, 4.72, 5.00)
dados <- as.data.frame(cbind(temp, cens, lwbc))
```

Análise exploratória

Na análise exploratória nós dividimos a variável contínua WBC em dois grupos com igual amplitude e comparamos as curvas por meio do teste logrank.

```
ekm <- survfit(Surv(temp, cens) ~ 1, data = dados)
ekm2 <- survfit(Surv(temp, cens) ~ cut_interval(lwbc, 2), data = dados)
splots <- list()
splots[[1]] <- ggsurvplot(ekm, conf.int = FALSE)
splots[[2]] <- ggsurvplot(ekm2, pval = TRUE, conf.int = FALSE) +
  guides(colour = guide_legend(nrow = 2))
arrange_ggsurvplots(splots)
```



Ajuste do modelo de regressão de Cox

Na sequência ajustamos o modelo de regressão de Cox.

```
fit <- coxph(Surv(temp, cens) ~ lwbc, data = dados)
fit
```

```
## Call:
## coxph(formula = Surv(temp, cens) ~ lwbc, data = dados)
##
##        coef exp(coef) se(coef)     z       p
## lwbc 1.4672    4.3373   0.4937 2.972 0.00296
##
## Likelihood ratio test=9.36  on 1 df, p=0.002217
## n= 17, number of events= 17
```

A covariável `lwbc` é significativa. Podemos dizer que parte da variação observada nos tempos de sobrevivência pode ser explicada pela contagem de glóbulos brancos. A mesma conclusão é obtida pelo teste da razão de verossimilhanças.

Diagnóstico do modelo

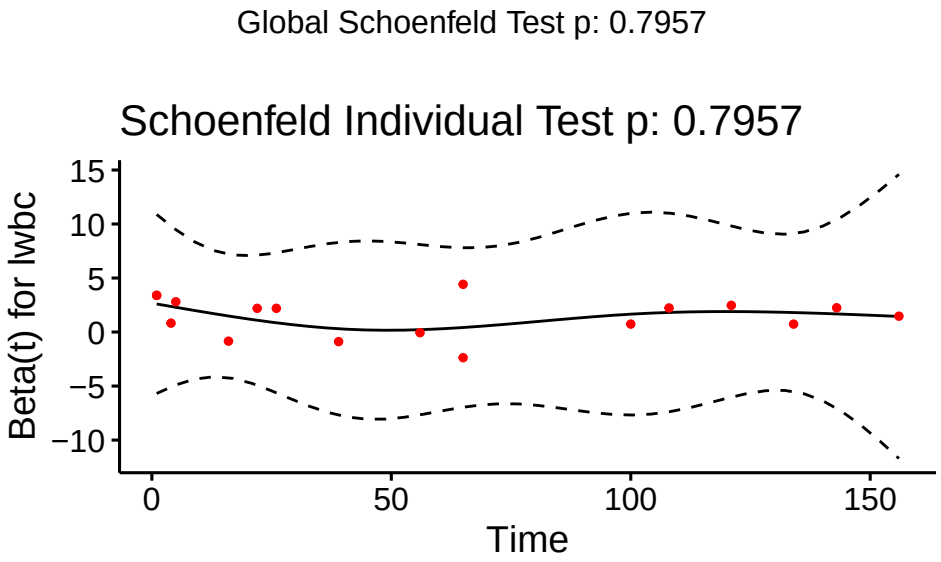
A suposição de razão de taxas de falhas proporcionais é avaliada por meio da função `cox.zph`, que aponta que o pressuposto básico do modelo é atendido.

```
zph <- cox.zph(fit, transform = "identity")
zph
```

```
##        chisq df   p
## lwbc   0.067  1 0.8
## GLOBAL 0.067  1 0.8
```

Alternativamente, avaliamos os resíduos de Schoenfeld.

```
ggcoxzph(zph)
```



Interpretação do modelo ajustado

O coeficiente de regressão ajustado para o modelo de Cox é positivo, o que indica que o risco do evento aumenta para maiores valores de contagem de glóbulos branco.

```
summary(fit)
```

```
## Call:
## coxph(formula = Surv(temp, cens) ~ lwbc, data = dados)
##
##   n= 17, number of events= 17
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## lwbc 1.4672    4.3373   0.4937 2.972  0.00296 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## lwbc     4.337     0.2306     1.648     11.41
##
## Concordance= 0.761  (se = 0.067 )
## Likelihood ratio test= 9.36  on 1 df,   p=0.002
## Wald test            = 8.83  on 1 df,   p=0.003
## Score (logrank) test = 9.54  on 1 df,   p=0.002
```

Tomando a exponencial dos coeficientes ajustados temos a razão de taxas de falhas. Tal efeito, referido em inglês como *hazard ratio* (HR), foi estimado em 4.34, $IC = (1.65;\ 11.41)$. Compare este resultado com aquele obtidos para o modelo paramétrico. Havíamos concluído que o modelo exponencial era adequado para representar os tempos de sobrevida. O modelo exponencial pertence à classe dos modelos de tempo de vida acelerados e à de taxas de falha proporcionais.

```
summary(survreg(Surv(temp, cens) ~ lwbc, dist = 'exponential', data = dados))
```

```
##
## Call:
## survreg(formula = Surv(temp, cens) ~ lwbc, data = dados, dist = "exponential")
##              Value Std. Error     z       p
## (Intercept)  8.477      1.711  4.95 7.3e-07
## lwbc        -1.109      0.414 -2.68  0.0073
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -83.9   Loglik(intercept only)= -87.3
##  Chisq= 6.83 on 1 degrees of freedom, p= 0.009
## Number of Newton-Raphson Iterations: 5
## n= 17
```

A interpretação para a exponencial do coeficiente estimado no modelo paramétrico se dá em termos de razão de tempos medianos. Assim, a cada aumento de uma unidade no logaritmo de WBC, o tempo mediano de vida dos pacientes fica reduzido para um terço ($e^{-1.109} = 0.3299$). Note que o sinal do coeficiente é invertido em relação ao modelo de Cox, mas as interpretações estão na mesma direção!

Exemplo 2: Sobrevida de Pacientes com Câncer de Encéfalo

Retornemos os dados de sobrevida (em meses) de pacientes com câncer de encéfalo. A amostra é composta por 397 pacientes. As variáveis consideradas foram: idade (em anos), sexo (masculino ou feminino) e tipo de tratamento (radioterapia ou outros). Foram observadas 216 falhas e 181 censuras.

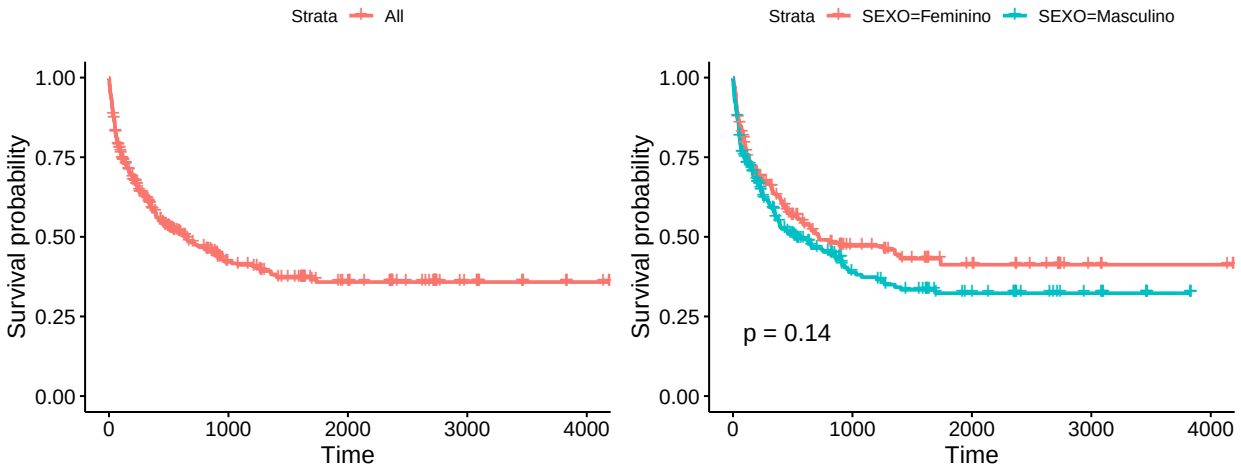
```
encefalo <- read.table("enc.txt", h = T)
head(encefalo)
```

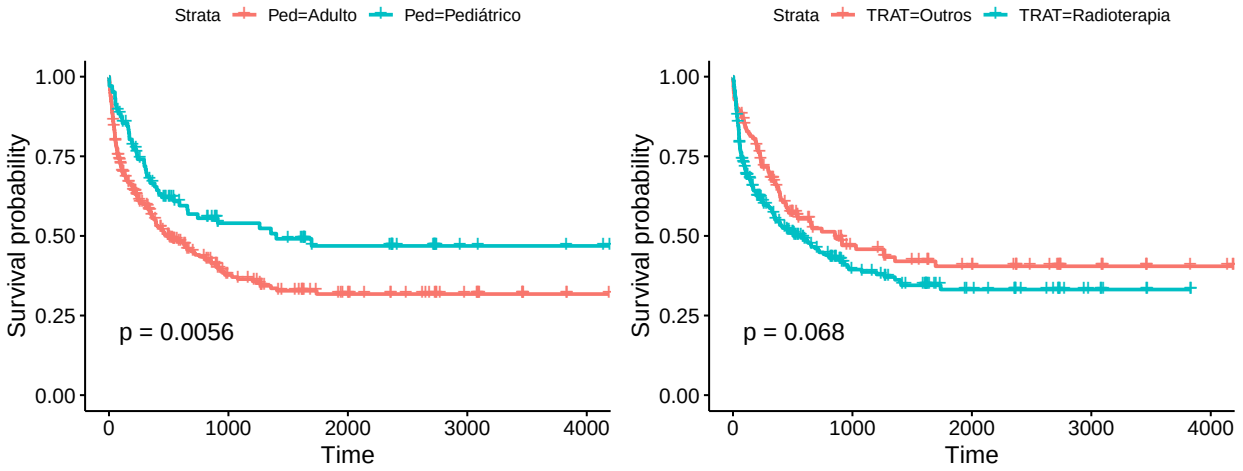
```
##   Sexo Idade Serv TRI Estadio AED Controle cens tempo    Temp_Mês        Ped
## 1    2     4    9   1       2   2        1    1  4137       137,9 Pediátrico
## 2    1    47    9   2      99   2        1    7   113 3,766666667     Adulto
## 3    2    56    9   2      99   2        1    4   688 22,93333333     Adulto
## 4    2     6    9   2       3   2        1    4  1401        46,7 Pediátrico
## 5    2    47    9   3      99   2        1    6  2486 82,86666667     Adulto
## 6    2    20    9   2      99   2        1    4  1739 57,96666667     Adulto
##        SEXO          TRAT cens.1
## 1  Feminino       Outros      0
## 2 Masculino Radioterapia      0
## 3  Feminino Radioterapia      1
## 4  Feminino Radioterapia      1
## 5  Feminino       Outros      0
## 6  Feminino Radioterapia      1
```

Análise exploratória

Para a construção das curvas de sobrevivência pelo estimador de Kaplan-Meier, a idade foi considerada numa versão binária. A seguir são mostrados os gráficos de sobrevivência estimada, marginalmente e por grupo.

```
ekm <- survfit(Surv(tempo, cens.1) ~ 1, data = encefalo)
ekm1 <- survfit(Surv(tempo, cens.1) ~ SEXO, data = encefalo)
ekm2 <- survfit(Surv(tempo, cens.1) ~ Ped, data = encefalo)
ekm3 <- survfit(Surv(tempo, cens.1) ~ TRAT, data = encefalo)
splots <- list()
splots[[1]] <- ggsurvplot(ekm, conf.int = FALSE)
splots[[2]] <- ggsurvplot(ekm1, pval = TRUE, conf.int = FALSE)
splots[[3]] <- ggsurvplot(ekm2, pval = TRUE, conf.int = FALSE)
splots[[4]] <- ggsurvplot(ekm3, pval = TRUE, conf.int = FALSE)
arrange_ggsurvplots(splots)
```





O teste logrank não aponta diferença entre os sexos. Uma diferença estatisticamente significativa foi encontrada entre os grupos de idade. Uma diferença marginal foi obtida para a variável tratamento.

Ajuste do modelo de Cox

O modelo de Cox incluindo as três variáveis é ajustado a seguir.

```
cox0 <- coxph(Surv(tempo, cens.1) ~ Idade + SEXO + TRAT, data = encefalo)
summary(cox0)
```

```
## Call:
## coxph(formula = Surv(tempo, cens.1) ~ Idade + SEXO + TRAT, data = encefalo)
##
##   n= 397, number of events= 216
##
##                       coef exp(coef) se(coef)     z Pr(>|z|)
## Idade             0.016898  1.017042 0.003485 4.849 1.24e-06 ***
## SEXOMasculino     0.186009  1.204433 0.143342 1.298    0.194
## TRATRadioterapia 0.047264  1.048399 0.150470 0.314    0.753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## Idade                1.017     0.9832    1.0101     1.024
## SEXOMasculino        1.204     0.8303    0.9094     1.595
## TRATRadioterapia     1.048     0.9538    0.7806     1.408
##
## Concordance= 0.609  (se = 0.019 )
## Likelihood ratio test= 28.64  on 3 df,   p=3e-06
## Wald test            = 28.19  on 3 df,   p=3e-06
## Score (logrank) test = 28.82  on 3 df,   p=2e-06
```

Em concordância com resultados anteriores, apenas a variável idade foi significativa. Reajustamos o modelo incluindo apenas o efeito de idade.

```
cox1 <- coxph(Surv(tempo, cens.1) ~ Idade, data = encefalo)
summary(cox1)
```

```
## Call:
## coxph(formula = Surv(tempo, cens.1) ~ Idade, data = encefalo)
```

```
##
##   n= 397, number of events= 216
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## Idade 0.017268  1.017418 0.003354 5.148 2.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## Idade     1.017     0.9829     1.011     1.024
##
## Concordance= 0.606  (se = 0.02 )
## Likelihood ratio test= 26.72  on 1 df,   p=2e-07
## Wald test            = 26.5  on 1 df,   p=3e-07
## Score (logrank) test = 27.08  on 1 df,   p=2e-07
```

O teste de razão de verossimilhanças comparando os dois modelos é mostrado a seguir.

```
anova(cox0, cox1)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(tempo, cens.1)
##  Model 1: ~ Idade + SEXO + TRAT
##  Model 2: ~ Idade
##    loglik  Chisq Df P(>|Chi|)
## 1 -1167.7
## 2 -1168.7 1.9188  2    0.3831
```

O valor-p aponta que podemos ficar com o modelo mais simples, que inclui apenas idade como preditor.

Diagnóstico do modelo

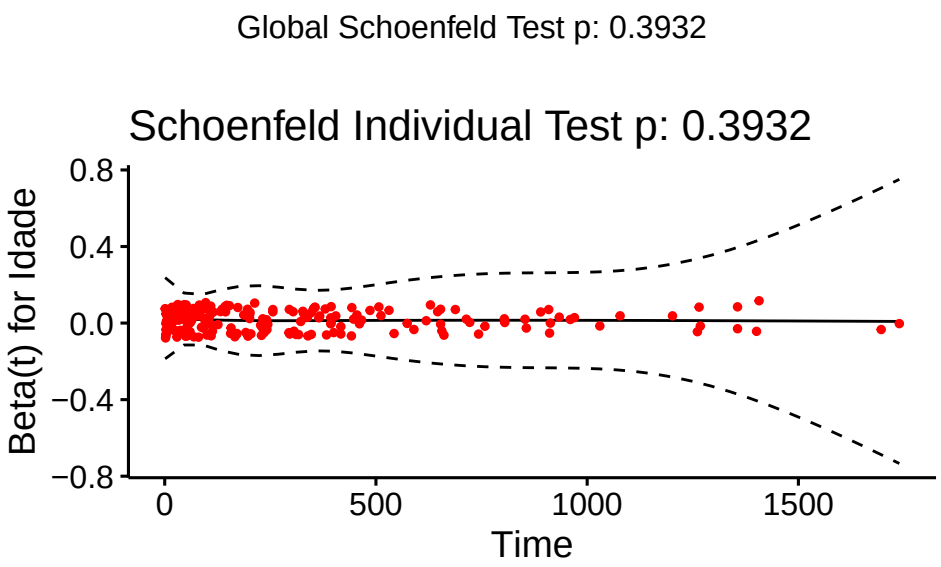
A verificação de proporcionalidade das taxas de falha é testada a seguir.

```
zph <- cox.zph(cox1, transform = "identity")
zph
```

```
##        chisq df    p
## Idade  0.729  1 0.39
## GLOBAL 0.729  1 0.39
```

O teste não rejeita a hipótese nula de taxas proporcionais. Alternativamente, avaliamos os resíduos de Schoenfeld.

```
ggcoxzph(zph)
```



Interpretação do modelo ajustado

```
cox1
```

```
## Call:
## coxph(formula = Surv(tempo, cens.1) ~ Idade, data = encefalo)
##
##           coef exp(coef) se(coef)     z        p
## Idade 0.017268  1.017418 0.003354 5.148 2.63e-07
##
## Likelihood ratio test=26.72  on 1 df, p=2.349e-07
## n= 397, number of events= 216
```

Temos as seguintes interpretações:

- O aumento em 1 ano de idade está relacionado com um aumento de 1.7% no risco (razão de taxas de falha) de o paciente ir a óbito ($e^{0.0173} = 1.0174$).
- Para uma diferença de idade de 10 anos (indivíduos com 40 e 30 anos, por exemplo) o aumento no risco de óbito é de 18.8% ($e^{10 \times 0.0173} = 1.1888$).

Na modelagem paramétrica havíamos concluído que o modelo log-normal era adequado. O ajuste deste modelo é mostrado para comparação.

```
fit_logno <- survreg(Surv(tempo, cens.1) ~ Idade, dis = "lognormal", data = encefalo)
summary(fit_logno)
```

```
##
## Call:
## survreg(formula = Surv(tempo, cens.1) ~ Idade, data = encefalo,
##     dist = "lognormal")
##               Value Std. Error    z       p
## (Intercept)  7.6594     0.2715 28.2 < 2e-16
## Idade       -0.0321     0.0063 -5.1 3.4e-07
## Log(scale)   0.8853     0.0521 17.0 < 2e-16
##
## Scale= 2.42
##
```

```
## Log Normal distribution
## Loglik(model)= -1655.3   Loglik(intercept only)= -1668.2
##  Chisq= 25.66 on 1 degrees of freedom, p= 4.1e-07
## Number of Newton-Raphson Iterations: 3
## n= 397
```

A razão de tempos medianos entre dois indivíduos com diferença de um ano é dada por $e^{-0.0321} = 0.968$.

- O coeficiente negativo no modelo log-normal indica que pacientes mais jovens apresentam sobrevida superior àquela de pacientes mais velhos.
- O coeficiente positivo no modelo de Cox indica que o risco de morte é maior para os pacientes mais velhos.

Exemplo 3: Dados de Aleitamento Materno

Retornemos aos dados de aleitamento materno. A variável resposta de interesse foi estabelecida como o tempo máximo de aleitamento materno (em meses), ou seja, o tempo contado a partir do nascimento até o desmame completo da criança. O quadro a seguir apresenta uma descrição das 11 covariáveis estudadas para as 150 mães avaliadas.

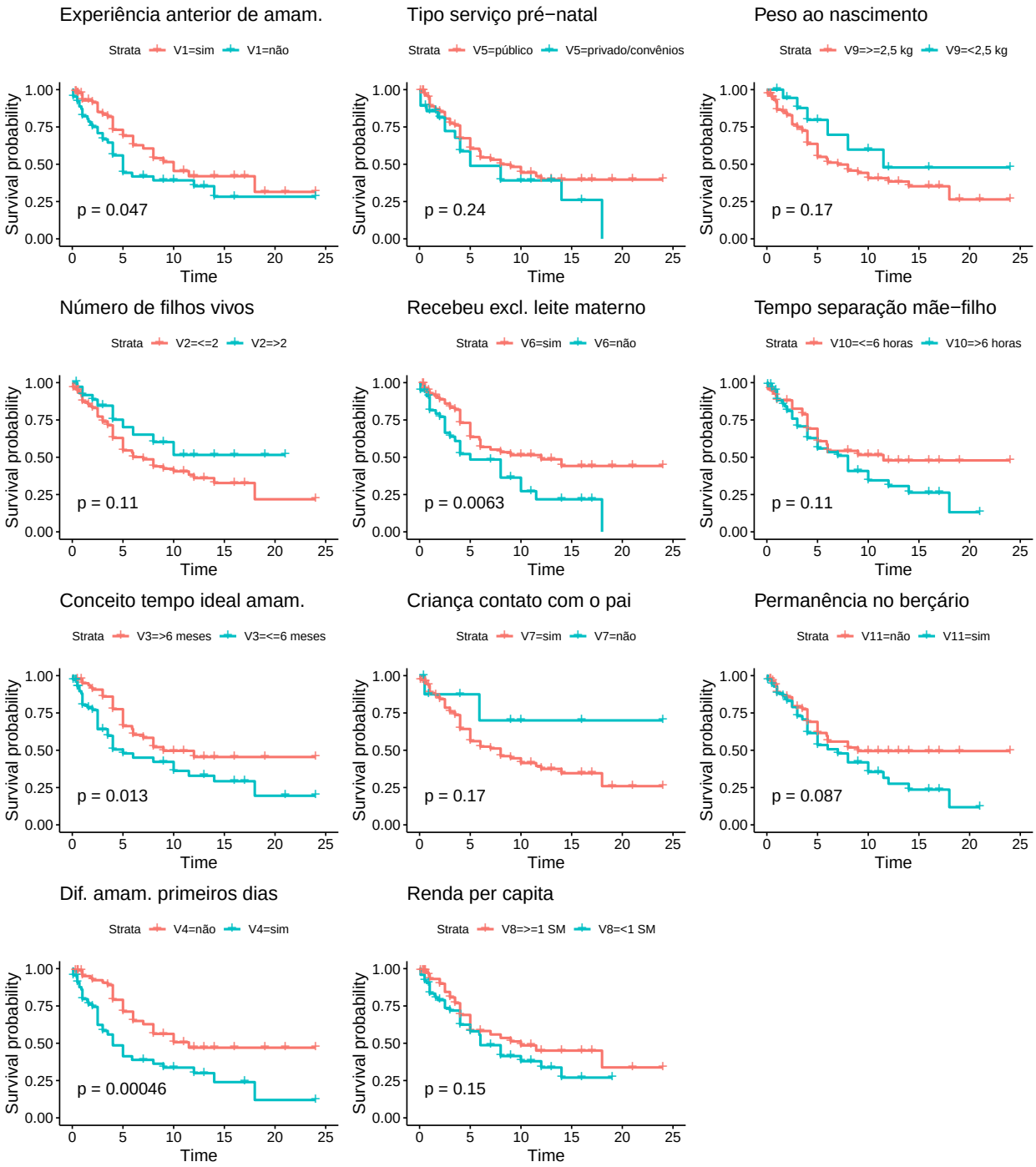
Código	Descrição	Categorias
V1	Experiência anterior de amamentação	0 se sim e 1 se não
V2	Número de filhos vivos	0 se dois ou menos e 1 se mais de dois
V3	Conceito materno sobre o tempo ideal de amamentação	0 se $>$ 6 meses e 1 se $\leq$ 6 meses
V4	Dificuldades para amamentar nos primeiros dias pós-parto	0 se não e 1 se sim
V5	Tipo de serviço em que realizou o pré-natal	0 se público e 1 se privado/convênios
V6	Recebeu exclusivamente leite materno na maternidade	0 se sim e 1 se não
V7	A criança teve contato com o pai	0 se sim e 1 se não
V8	Renda per capita (em SM/mês)	0 se $\geq$ 1 SM e 0 se $<$ 1 SM
V9	Peso ao nascimento	0 se $\geq$ 2,5kg e 1 se $<$ 2,5kg
V10	Tempo de separação mãe-filho pós-parto	0 se $\leq$ 6 horas e 1 se $>$ 6 horas
V11	Permanência no berçário	0 se não e 1 se sim

Análise exploratória

Na sequência são mostradas as curvas de sobrevivência obtidas pelo estimador de Kaplan-Meier, juntamente com o valor-p do teste logrank.

```
# Leitura dos dados
dados <- read.table("desmame.txt", header = TRUE, dec = ",")
# Definição dos fatores
dados[, 4:14] <- lapply(dados[, 4:14], factor)
levels(dados$V1) <- c("sim", "não")
levels(dados$V2) <- c("<=2", ">2")
levels(dados$V3) <- c(">6 meses", "<=6 meses")
levels(dados$V4) <- c("não", "sim")
levels(dados$V5) <- c("público", "privado/convênios")
levels(dados$V6) <- c("sim", "não")
levels(dados$V7) <- c("sim", "não")
levels(dados$V8) <- c(">=1 SM", "<1 SM")
levels(dados$V9) <- c(">=2,5 kg", "<2,5 kg")
levels(dados$V10) <- c("<=6 horas", ">6 horas")
levels(dados$V11) <- c("não", "sim")
# Estimador de Kaplan-Meier
ekm_V1 <- survfit(Surv(tempo, cens) ~ V1, data = dados)
ekm_V2 <- survfit(Surv(tempo, cens) ~ V2, data = dados)
ekm_V3 <- survfit(Surv(tempo, cens) ~ V3, data = dados)
ekm_V4 <- survfit(Surv(tempo, cens) ~ V4, data = dados)
```

```r
ekm_V5 <- survfit(Surv(tempo, cens) ~ V5, data = dados)
ekm_V6 <- survfit(Surv(tempo, cens) ~ V6, data = dados)
ekm_V7 <- survfit(Surv(tempo, cens) ~ V7, data = dados)
ekm_V8 <- survfit(Surv(tempo, cens) ~ V8, data = dados)
ekm_V9 <- survfit(Surv(tempo, cens) ~ V9, data = dados)
ekm_V10 <- survfit(Surv(tempo, cens) ~ V10, data = dados)
ekm_V11 <- survfit(Surv(tempo, cens) ~ V11, data = dados)
# Lista de gráficos
splots <- list()
splots[[1]] <- ggsurvplot(ekm_V1, pval = TRUE, title = "Experiência anterior de amam.")
splots[[2]] <- ggsurvplot(ekm_V2, pval = TRUE, title = "Número de filhos vivos")
splots[[3]] <- ggsurvplot(ekm_V3, pval = TRUE, title = "Conceito tempo ideal amam.")
splots[[4]] <- ggsurvplot(ekm_V4, pval = TRUE, title = "Dif. amam. primeiros dias")
splots[[5]] <- ggsurvplot(ekm_V5, pval = TRUE, title = "Tipo serviço pré-natal")
splots[[6]] <- ggsurvplot(ekm_V6, pval = TRUE, title = "Recebeu excl. leite materno")
splots[[7]] <- ggsurvplot(ekm_V7, pval = TRUE, title = "Criança contato com o pai")
splots[[8]] <- ggsurvplot(ekm_V8, pval = TRUE, title = "Renda per capita")
splots[[9]] <- ggsurvplot(ekm_V9, pval = TRUE, title = "Peso ao nascimento")
splots[[10]] <- ggsurvplot(ekm_V10, pval = TRUE, title = "Tempo separação mãe-filho")
splots[[11]] <- ggsurvplot(ekm_V11, pval = TRUE, title = "Permanência no berçário")
# Junta os ggsurvplots
arrange_ggsurvplots(splots, print = TRUE, ncol = 3, nrow = 4)
```



O teste logrank aponta como significativas ao nível $\alpha = 5\%$ as variáveis `V1`, `V3`, `V4` e `V6`. Estas variáveis foram estatisticamente significativas quando incluídas na modelagem paramétrica anterior. Naquela oportunidade, o modelo log-normal foi considerado adequado para descrever o tempo até o desmame. Discutiremos agora um modelo alternativo.

Ajuste do modelo de Cox

Diferentes critérios de seleção de covariáveis podem ser consideradas na construção do modelo. Dentre as opções temos procedimentos automáticos, como a seleção *stepwise* implementada na função `step`, que

seleciona um modelo com base no critério AIC. Nesta ilustração, usaremos o argumento `method = "breslow"` no ajuste do modelo. O modelo final encontrado é o mesmo se usarmos a opção padrão (aproximação de Efron para tempos empatados).

```
fit.step <- step(coxph(Surv(tempo, cens) ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 +
                         V10 + V11, data = dados, x = T, method = "breslow"),
                 trace = FALSE)
fit.step
```

```
## Call:
## coxph(formula = Surv(tempo, cens) ~ V3 + V4 + V6 + V8, data = dados,
##     x = T, method = "breslow")
##
##               coef exp(coef) se(coef)     z        p
## V3<=6 meses 0.5635    1.7568   0.2590 2.176 0.029554
## V4sim       0.9224    2.5153   0.2567 3.593 0.000327
## V6não       0.5463    1.7269   0.2581 2.117 0.034273
## V8<1 SM     0.5687    1.7659   0.2572 2.211 0.027055
##
## Likelihood ratio test=26.14  on 4 df, p=2.971e-05
## n= 150, number of events= 65
```

Estimativas dos parâmetros

```
summary(fit.step)
```

```
## Call:
## coxph(formula = Surv(tempo, cens) ~ V3 + V4 + V6 + V8, data = dados,
##     x = T, method = "breslow")
##
##   n= 150, number of events= 65
##
##               coef exp(coef) se(coef)     z Pr(>|z|)
## V3<=6 meses 0.5635    1.7568   0.2590 2.176 0.029554 *
## V4sim       0.9224    2.5153   0.2567 3.593 0.000327 ***
## V6não       0.5463    1.7269   0.2581 2.117 0.034273 *
## V8<1 SM     0.5687    1.7659   0.2572 2.211 0.027055 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## V3<=6 meses     1.757     0.5692     1.058     2.918
## V4sim           2.515     0.3976     1.521     4.160
## V6não           1.727     0.5791     1.041     2.864
## V8<1 SM         1.766     0.5663     1.067     2.924
##
## Concordance= 0.708  (se = 0.034 )
## Likelihood ratio test= 26.14  on 4 df,   p=3e-05
## Wald test            = 25.44  on 4 df,   p=4e-05
## Score (logrank) test = 27.02  on 4 df,   p=2e-05
```

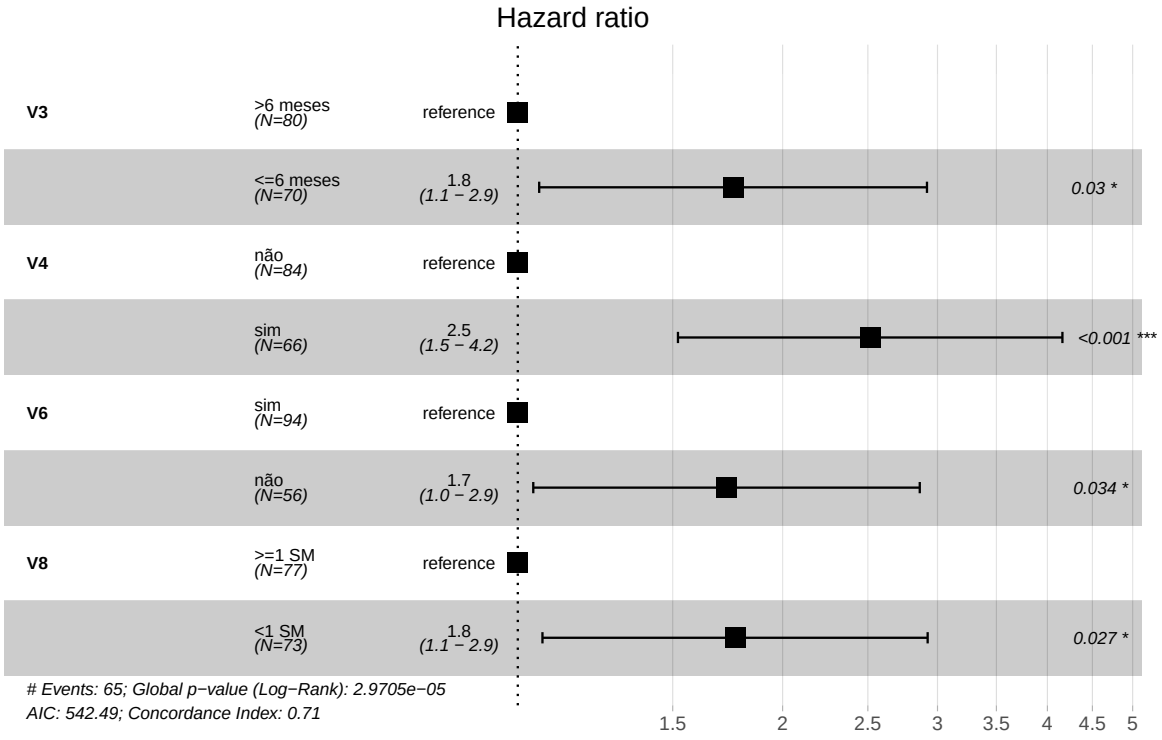
Temos obter as seguintes interpretações:

- A taxa estimada de desmame precoce de mães que acreditam que o tempo de amamentação ideal é ≤ 6 meses é 1.76 [1.06; 2.92] vezes a taxa das mães que acreditam que o tempo ideal de amamentação é >6 meses.

- A taxa estimada de desmame precoce em mães que apresentaram dificuldades de amamentar nos primeiros dias pós-parto é 2.52 [1.52; 4.16] vezes a taxa das mães que não apresentaram dificuldade.
- A taxa estimada de desmame precoce em crianças que não receberam exclusivamente leite materno na maternidade é 1.73 [1.04; 2.86] vezes a taxa de crianças que receberam exclusivamente leite materno.
- A taxa estimada de desmame precoce em famílias com renda per capita abaixo de 1 SM é 1.77 [1.07; 2.92] vezes a de famílias com renda per capita acima de 1 SM.

Visualizamos os resultados por meio do *forest plot*.

```
ggforest(fit.step, data = dados)
```



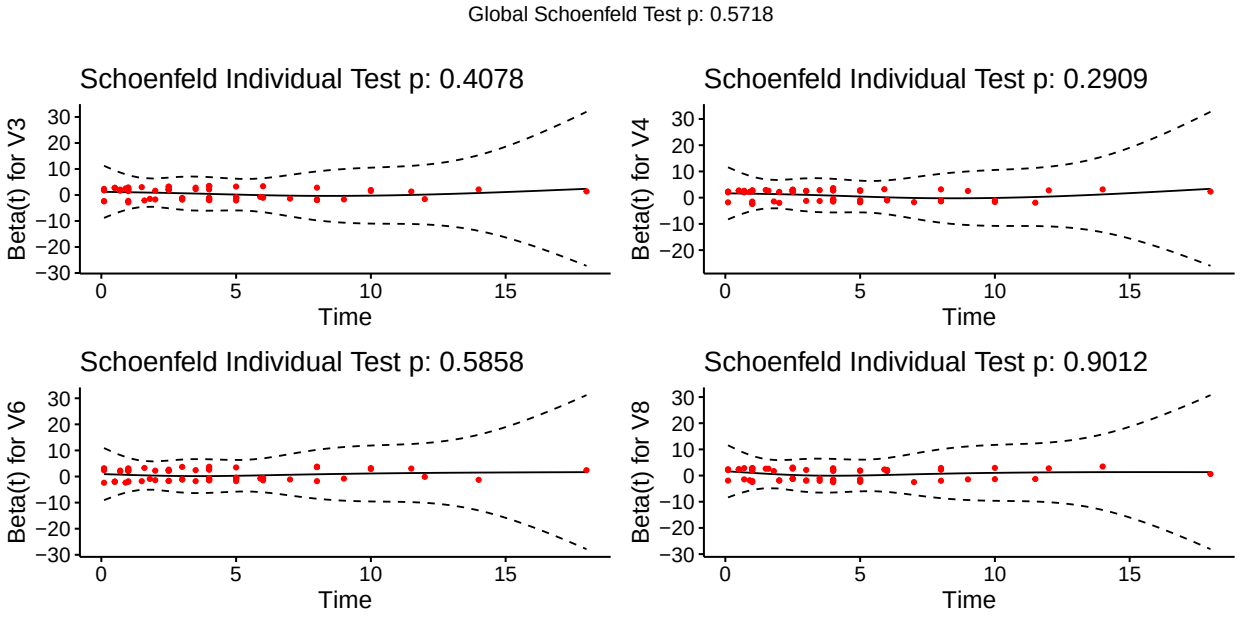
Diagnóstico do modelo

Verificamos se a suposição de taxas proporcionais é atendida por meio dos resíduos de Schoenfeld.

```
zph <- cox.zph(fit.step, transform = "identity")
zph
```

```
##         chisq df    p
## V3     0.6853  1 0.41
## V4     1.1152  1 0.29
## V6     0.2969  1 0.59
## V8     0.0154  1 0.90
## GLOBAL 2.9173  4 0.57
```

```
ggcoxzph(zph)
```



Verifica-se a ausência de tendência nos gráficos, o que pode ser confirmado também pelo resultado dos testes.

Exemplo 4: Análise de Dados de Câncer de Laringe (Klein e Moeschberger, 2003)

Os dados são provenientes de um estudo realizado com 90 pacientes do sexo masculino, diagnosticados com câncer de laringe, no período de 1970 a 1978 e acompanhados até 01/01/1983.

Foram registradas no diagnóstico as seguintes variáveis:

- Tempo de morte ou censura (meses)
- Idade (anos)
- Estágio (ordenado) da doença: I=tumor primário, II=envolvimento de nódulos, III=metástases e IV=combinações dos 3 estágios.

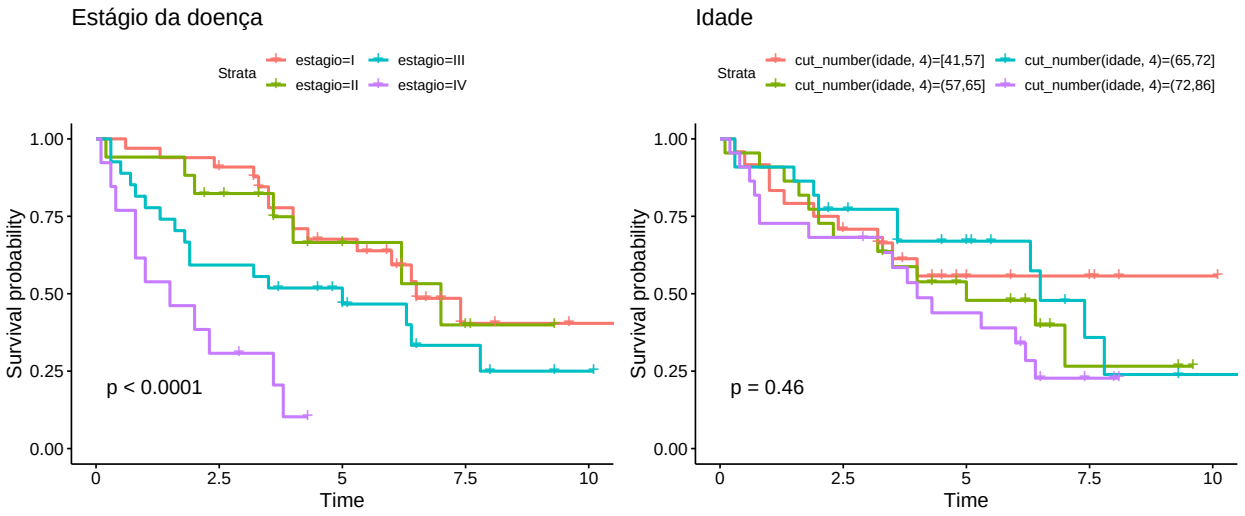
Análise exploratória

```
# Leitura dos dados
laringe <- read.table("laringe.txt", header = TRUE, sep = "")
laringe <- mutate(laringe, estagio = factor(estagio, labels = c("I", "II", "III", "IV")))
summary(laringe)
```

```
##        id            tempos            cens            idade       estagio
##  Min.   : 1.00   Min.   : 0.100   Min.   :0.0000   Min.   :41.00   I  :33
##  1st Qu.:23.25   1st Qu.: 2.000   1st Qu.:0.0000   1st Qu.:57.00   II :17
##  Median :45.50   Median : 4.000   Median :1.0000   Median :65.00   III:27
##  Mean   :45.50   Mean   : 4.198   Mean   :0.5556   Mean   :64.61   IV :13
##  3rd Qu.:67.75   3rd Qu.: 6.200   3rd Qu.:1.0000   3rd Qu.:72.00
##  Max.   :90.00   Max.   :10.700   Max.   :1.0000   Max.   :86.00
```

Para confecção das curvas de sobrevivência a variável idade foi categorizada em intervalos de aproximadamente igual número de observações.

```
# Estimador de Kaplan-Meier
ekm_V1 <- survfit(Surv(tempos, cens) ~ estagio, data = laringe)
ekm_V2 <- survfit(Surv(tempos, cens) ~ cut_number(idade, 4), data = laringe)
# Lista de gráficos
splots <- list()
splots[[1]] <- ggsurvplot(ekm_V1, pval = TRUE, title = "Estágio da doença") +
  guides(colour = guide_legend(nrow = 2))
splots[[2]] <- ggsurvplot(ekm_V2, pval = TRUE, title = "Idade") +
  guides(colour = guide_legend(nrow = 2))
# Junta os ggsurvplots
arrange_ggsurvplots(splots, print = TRUE, ncol = 2, nrow = 1)
```



O efeito de estágio da doença é mais perceptível que o efeito de idade. Note, contudo, que os valores-p acima do teste logrank consideram o efeito marginal. Nos modelos de regressão seremos capazes de fazer inferência incluindo os dois preditores conjuntamente.

Ajuste dos modelos

Na sequência ajustamos vários modelos de Cox, do mais simples (modelo nulo) até o mais complexo (envolvendo interação entre idade e estágio da doença).

```
fit1 <- coxph(Surv(tempos, cens) ~ 1, data = laringe, method = "breslow")
fit1
```

```
## Call:  coxph(formula = Surv(tempos, cens) ~ 1, data = laringe, method = "breslow")
##
## Null model
##   log likelihood= -197.2129
##   n= 90
```

```
fit2 <- coxph(Surv(tempos, cens) ~ estagio, data = laringe, x = T, method = "breslow")
fit2
```

```
## Call:
## coxph(formula = Surv(tempos, cens) ~ estagio, data = laringe,
##     x = T, method = "breslow")
##
##                coef exp(coef) se(coef)     z        p
## estagioII   0.06576   1.06797  0.45844 0.143   0.8859
## estagioIII 0.61206   1.84423  0.35520 1.723   0.0849
## estagioIV   1.72284   5.60040  0.41966 4.105 4.04e-05
##
## Likelihood ratio test=16.26  on 3 df, p=0.001001
## n= 90, number of events= 50
```

```
fit3 <- coxph(Surv(tempos, cens) ~ estagio + idade, data = laringe, x = T,
              method = "breslow")
fit3
```

```
## Call:
## coxph(formula = Surv(tempos, cens) ~ estagio + idade, data = laringe,
##     x = T, method = "breslow")
```

```
##
##               coef exp(coef) se(coef)     z        p
## estagioII  0.13856   1.14862  0.46231 0.300    0.764
## estagioIII 0.63835   1.89335  0.35608 1.793    0.073
## estagioIV  1.69306   5.43607  0.42221 4.010 6.07e-05
## idade      0.01890   1.01908  0.01425 1.326    0.185
##
## Likelihood ratio test=18.07  on 4 df, p=0.001197
## n= 90, number of events= 50
```

```
fit4 <- coxph(Surv(tempos, cens) ~ estagio*idade, data = laringe, x = T,
              method = "breslow")
fit4
```

```
## Call:
## coxph(formula = Surv(tempos, cens) ~ estagio * idade, data = laringe,
##     x = T, method = "breslow")
##
##                       coef exp(coef)  se(coef)      z      p
## estagioII        -7.946142  0.000354  3.678209 -2.160 0.0307
## estagioIII       -0.122500  0.884706  2.468331 -0.050 0.9604
## estagioIV         0.846986  2.332605  2.425717  0.349 0.7270
## idade            -0.002559  0.997444  0.026051 -0.098 0.9218
## estagioII:idade   0.120254  1.127783  0.052307  2.299 0.0215
## estagioIII:idade  0.011351  1.011416  0.037449  0.303 0.7618
## estagioIV:idade   0.013673  1.013767  0.035967  0.380 0.7038
##
## Likelihood ratio test=24.27  on 7 df, p=0.001021
## n= 90, number of events= 50
```

Podemos comparar os modelos por meio do teste da razão de verossimilhanças.

```
anova(fit2, fit3, fit4)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(tempos, cens)
##  Model 1: ~ estagio
##  Model 2: ~ estagio + idade
##  Model 3: ~ estagio * idade
##    loglik  Chisq Df P(>|Chi|)
## 1 -189.08
## 2 -188.18 1.8036  1    0.1793
## 3 -185.08 6.2039  3    0.1021
```

O teste aponta que o efeito da interação não é conjuntamente significativo ao nível de 5%. Note, contudo, que o efeito individual de interação para o segundo estágio é significativo para este nível. Prosseguiremos a análise com o modelo sem interação. A escolha entre qual modelo é mais adequado não deve ser dependente apenas do resultado de um teste estatístico. Aspectos como plausibilidade biológica e evidências da literatura devem ser considerados, inclusive para manter ou não o efeito (aqui não significativo) da idade.

Estimativas dos parâmetros

```
summary(fit3)
```

```
## Call:
## coxph(formula = Surv(tempos, cens) ~ estagio + idade, data = laringe,
##     x = T, method = "breslow")
```

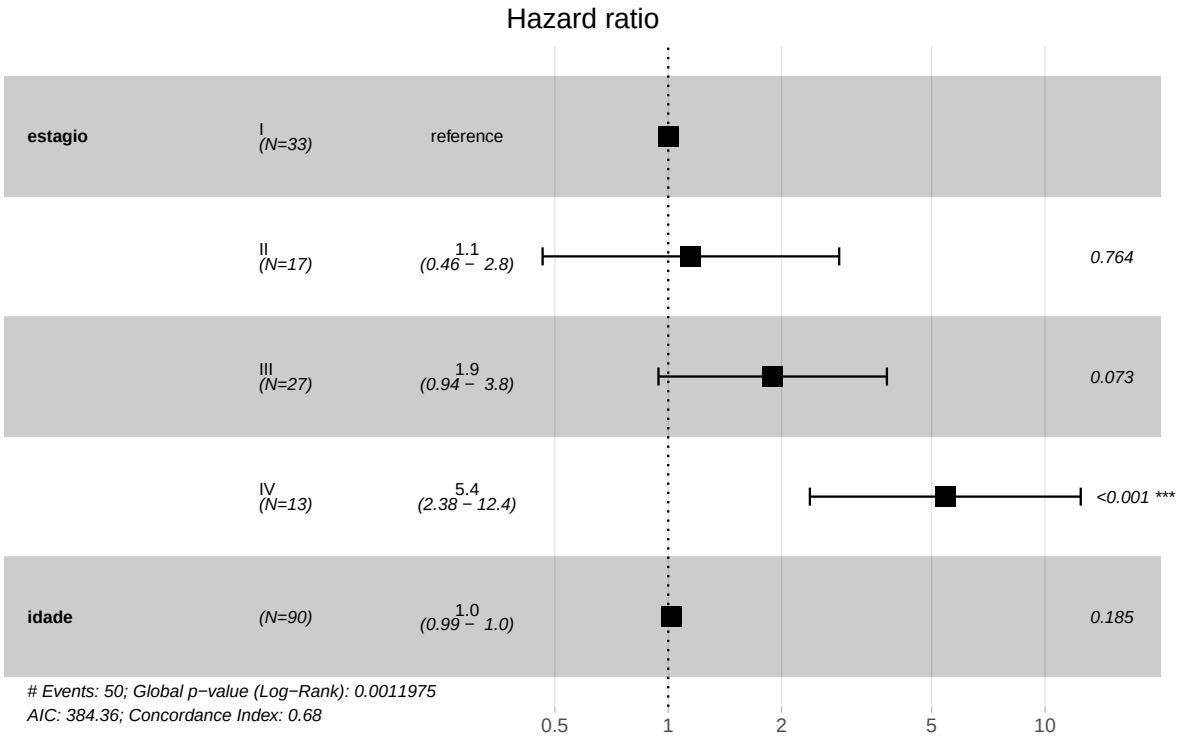
```
##
##   n= 90, number of events= 50
##
##               coef exp(coef) se(coef)     z Pr(>|z|)
## estagioII  0.13856   1.14862  0.46231 0.300    0.764
## estagioIII 0.63835   1.89335  0.35608 1.793    0.073 .
## estagioIV  1.69306   5.43607  0.42221 4.010 6.07e-05 ***
## idade      0.01890   1.01908  0.01425 1.326    0.185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## estagioII      1.149     0.8706    0.4642     2.842
## estagioIII     1.893     0.5282    0.9422     3.805
## estagioIV      5.436     0.1840    2.3763    12.436
## idade          1.019     0.9813    0.9910     1.048
##
## Concordance= 0.682  (se = 0.039 )
## Likelihood ratio test= 18.07  on 4 df,   p=0.001
## Wald test            = 20.82  on 4 df,   p=3e-04
## Score (logrank) test = 24.33  on 4 df,   p=7e-05
```

O efeito de idade não foi significativo. Para o estágio da doença temos as seguintes interpretações:

- A taxa de óbito de pacientes no estágio II não foi estatisticamente diferente da taxa dos pacientes no estágio I (referência).
- A taxa de óbito de pacientes no estágio III é estimada em 1.89 (IC=[0.94; 3.81]) vezes a taxa dos pacientes no estágio I.
- A taxa de óbito de pacientes no estágio IV é estimada em 5.44 (IC=[2.38; 12.44]) vezes a taxa dos pacientes no estágio I.

Visualizamos os resultados por meio do *forest plot*.

```
ggforest(fit3, data = laringe)
```



Podemos calcular as razões de taxas de falha (RTF) em comparações que não envolvam a categoria de referência no modelo (no caso o estágio I). Por exemplo, para dois pacientes na mesma idade e estágios IV e III, temos:

$$RTF = \frac{\widehat{\lambda}(t|\mathbf{x}_j)}{\widehat{\lambda}(t|\mathbf{x}_k)} = \frac{\exp\{\widehat{\beta}_3 + \widehat{\beta}_4 \times \text{idade}\}}{\exp\{\widehat{\beta}_2 + \widehat{\beta}_4 \times \text{idade}\}} = \exp\{\widehat{\beta}_3 - \widehat{\beta}_2\} = 2.87$$

A taxa de morte de pacientes no estágio IV da doença é de aproximadamente 3 vezes a taxa de morte de pacientes com a mesma idade no estágio III da doença.

Curvas de sobrevivência estimadas

A curva de sobrevivência estimada é dada por

$$\widehat{S}(t|\mathbf{x}) = [\widehat{S}_0(t)]^{\exp\{\widehat{\beta}_1 \text{I(Estágio II)} + \widehat{\beta}_2 \text{I(Estágio III)} + \widehat{\beta}_3 \text{I(Estágio IV)} + \widehat{\beta}_4 \text{Idade}\}}$$

No cálculo acima precisamos estimar $S_0(t)$. Para isso, fazemos uso da função `basehaz` que retorna estimativas da função taxa de falha acumulada e então usamos a relação entre esta e a função de sobrevivência.

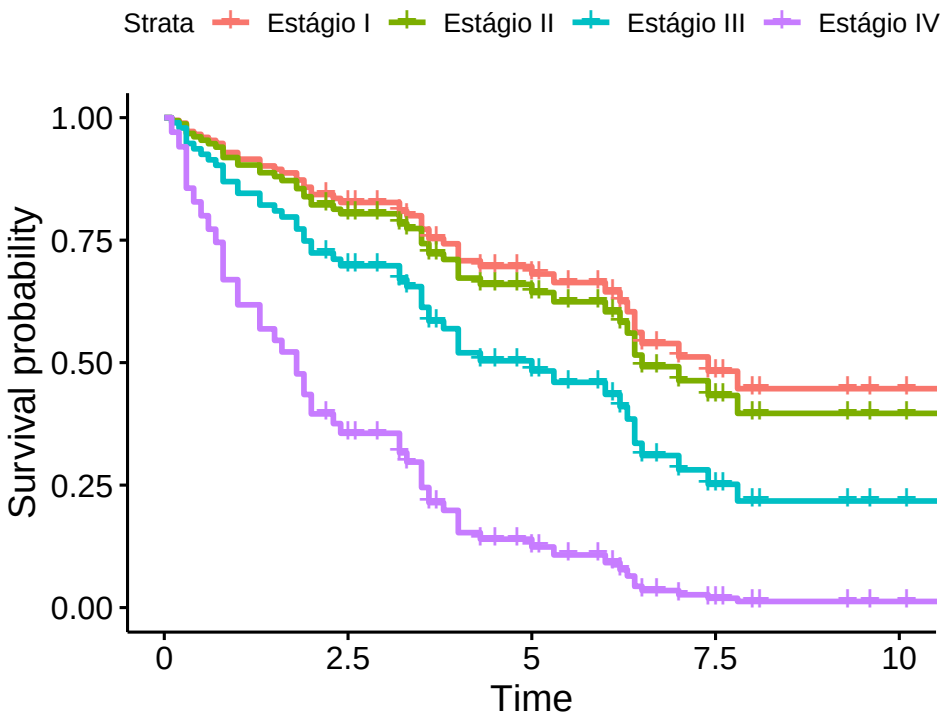
```
Lambda <- basehaz(fit3, centered = FALSE)
tempos <- Lambda$time
Lambda0 <- Lambda$hazard
S0 <- exp(-Lambda0)
head(round(cbind(tempos, S0, Lambda0), 4))
```

```
##      tempos     S0 Lambda0
## [1,]    0.1 0.9984  0.0016
## [2,]    0.2 0.9967  0.0033
## [3,]    0.3 0.9916  0.0084
```

```
## [4,]    0.4 0.9898  0.0102
## [5,]    0.5 0.9880  0.0121
## [6,]    0.6 0.9861  0.0140
```

A seguir são mostradas as curvas de sobrevivência estimadas para os quatro estágios, com a idade fixada em seu valor médio observado.

```
dados <- data.frame(estagio = c("I", "II","III","IV"), idade = mean(laringe$idade))
fit <- survfit(fit3, newdata = dados)
ggsurvplot(fit, data = dados, conf.int = FALSE,
           legend.labs = c("Estágio I", "Estágio II", "Estágio III", "Estágio IV"),
           ggtheme = theme_survminer())
```



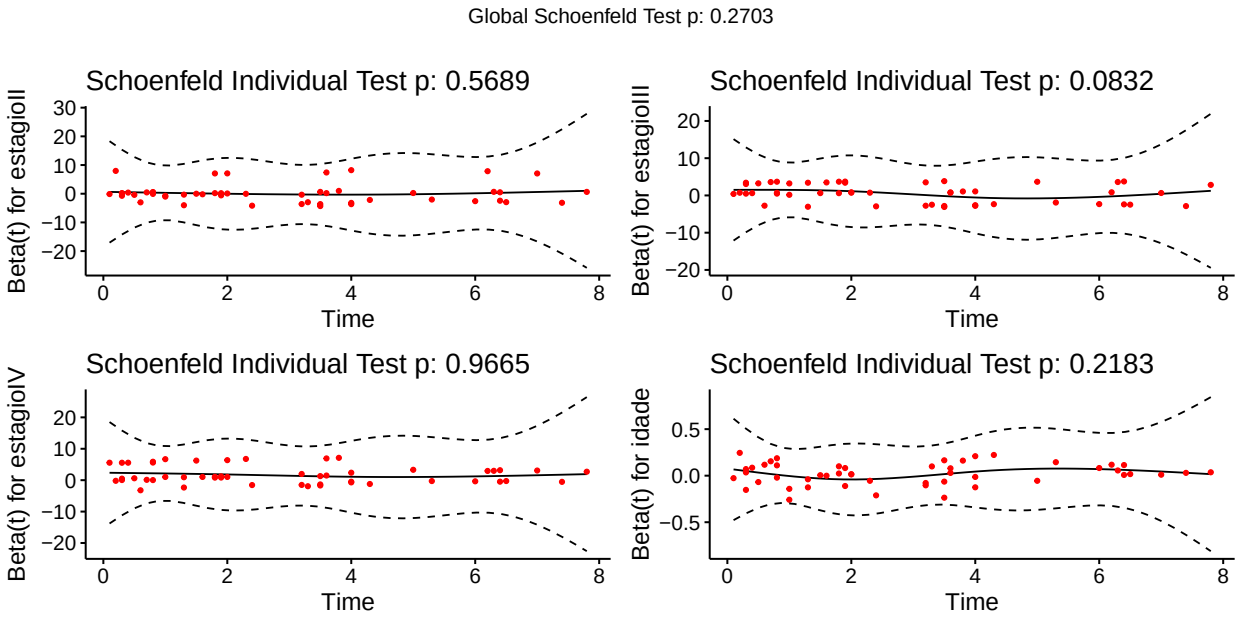
Diagnóstico do modelo

Os resultados a seguir, via testes e análise gráfica dos resíduos de Schoenfeld, indicam que a suposição de taxas proporcionais é atendida.

```
zph <- cox.zph(fit3, terms = FALSE, transform = "identity")
zph
```

```
##              chisq df     p
## estagioII  0.32458  1 0.569
## estagioIII 3.00146  1 0.083
## estagioIV  0.00176  1 0.967
## idade      1.51556  1 0.218
## GLOBAL     5.16980  4 0.270
```

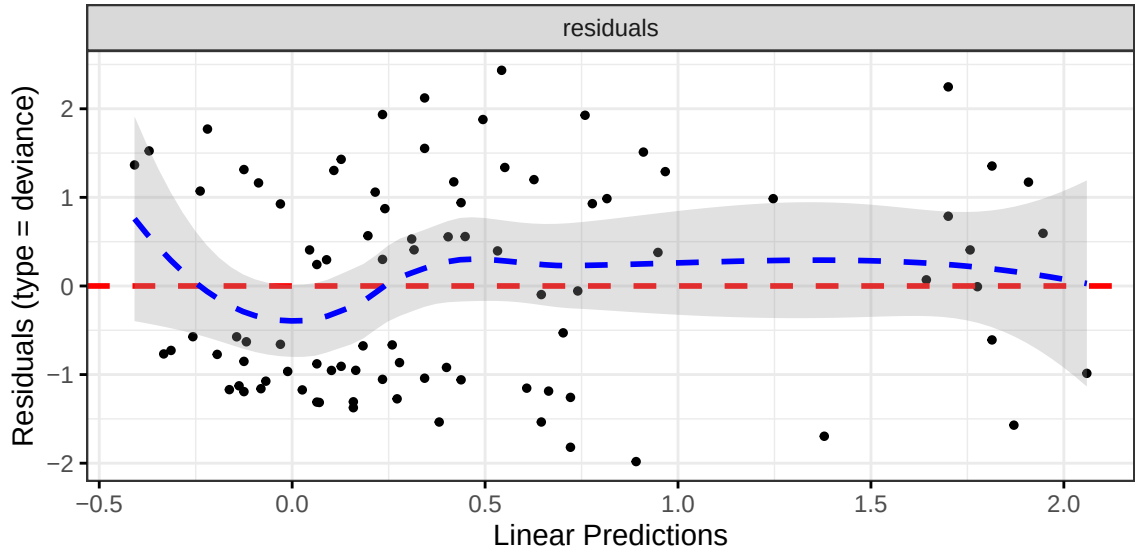
```
ggcoxzph(zph)
```



Por fim, ilustramos a inspeção de outros aspectos do ajuste do modelo de Cox. As funções `ggcoxdiagnostics` e `ggcoxfunctional` mostram gráficos para diferentes tipos de resíduos. É importante lembrar que a análise destes resíduos não é uma tarefa muito simples.

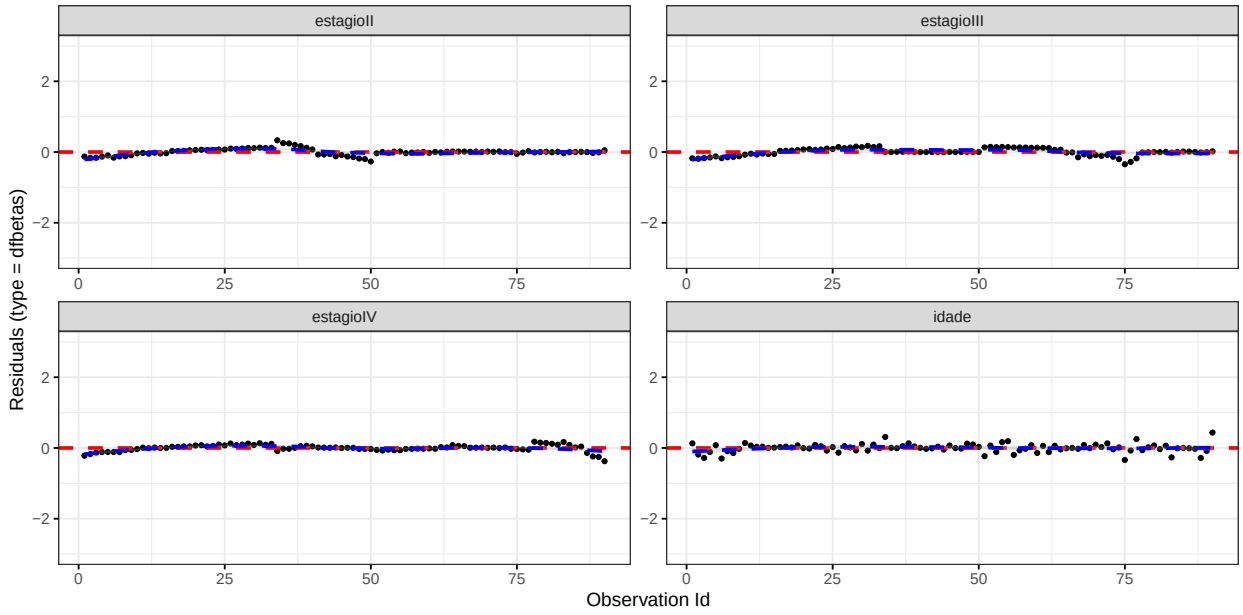
- Pontos atípicos (resíduos deviance)

```
ggcoxdiagnostics(fit3, type = "deviance")
```



- Pontos influentes (dfbeta padronizado)

```
ggcoxdiagnostics(fit3, type = "dfbetas", ox.scale = "observation.id", ylim = c(-3, 3))
```



- Forma funcional (resíduo martingal)

```
res.cox <- coxph(Surv(tempos, cens) ~ idade + I(log(idade)^2) + I(idade^2) + I(idade^3) +
                   I(sqrt(idade)), data = laringe)
ggcoxfunctional(res.cox, point.col = "blue", data = laringe, ylim = c(-2,1))
```