Introdução à Análise de Sobrevida

1. Introdução
2. O Tempo
3. Funções de Sobrevida
4. Estimação Não-Paramétrica
5. Modelo de Cox
6. Análise de Resíduos

Métodos Avançados de Análise de Sobrevida

1. Covariável Tempo-dependente
2. Múltiplos Eventos
3. Fragilidade

Cronograma – Introdução

Dia	Data	Horário	Tema
qua	4 de mar	09:00 – 12:00	Introdução
qua	11 de mar	–	Viagem
qua	18 de mar	09:00 – 12:00	Funções de Sobrevida
qua	25 de mar	09:00 – 12:00	Estimação Não-Paramétrica
qua	1 de abr	09:00 – 12:00	Estimação Não-Paramétrica
qua	8 de abr	09:00 – 12:00	Modelo de Cox
sex	10 de abr	–	Semana Santa
qua	15 de abr	09:00 – 12:00	Análise de Resíduos
ter	21 de abr	–	Tiradentes
qua	22 de abr	–	Enforcado
qui	23 de abr	–	São Jorge
qua	29 de abr	09:00 – 12:00	Dúvidas/Avaliação

Bibliografia

- Kleinbaum, D., & Klein, M. *Survival analysis : a self-learning text.* Springer, 1997.
- Therneau, T. M., & Grambsch, P. M. *Modeling survival data: extending the Cox model.* Springer, 2000.
- Carvalho, M. S., Andreozzi, V. L., Codeço, C, T., Barbosa, M. T. S. & Shimakura, S. E.. *Análise de Sobrevida: teoria e aplicações em saúde.*

Agradecimentos

- À Fiocruz, que viabilizou escrever, testar e publicar o livro
- Às instituições e seus pesquisadores que cederam, mais do que seus dados, seus problemas, idéias, perguntas:
  - Departamento de Informação e Informática do SUS - Datasus;
  - Escola Nacional de Saúde Pública – Fundação Oswaldo Cruz;
  - Hospital Geral de Betin;
  - Hospital Universitário Clementino Fraga Filho – Universidade Federal do Rio de Janeiro;
  - Hospital Universitário Gaffrée e Guinle – Universidade Federal do Estado do Rio de Janeiro;
  - Instituto de Pesquisa Clínica Evandro Chagas – Fundação Oswaldo Cruz;
  - Instituto de Saúde Coletiva – Universidade Federal da Bahia;
  - Instituto Nacional do Câncer.

Material do curso

- Notas de aula e dados para exercícios na página do livro (http://dengue.procc.fiocruz.br/~sobrevida/)
- R software (www.r-project.org)
- Tutorial online do R
  - http://www.leg.ufpr.br/Rtutorial/
  - http://www.leg.ufpr.br/~paulojus/embrapa/Rembrapa/

Refrescando a memória

Supondo que TODOS conhecem modelos de regressão...

- o que é parâmetro?
- o que é estimativa?
- o que é distribuição – normal, binomial, Poisson?
- o que é um intervalo de confiança?
- o que é um p-valor?
- como se quantifica o efeito de variável?
- o que significa a expressão "controlando por idade e sexo"?
- quando se usa regressão logística?
- quando se usa regressão de Poisson?

Sobrevida

- Em que tipo de desenho de estudo se aplica a *Análise de Sobrevida*?
- Que perguntas podemos responder com os modelos de sobrevida (ou sobrevivência)?
- Definir taxa de incidência ou força de morbidade ou risco instantâneo

Sobrevida ou sobrevivência

- A análise de sobrevida, também chamada de análise de sobrevivência, será utilizada quando o tempo for o objeto de interesse, seja este interpretado como **o tempo até a ocorrência de um evento** ou o **risco de ocorrência de um evento por unidade de tempo**.
- As perguntas passíveis de resposta neste tipo de abordagem são:
  - Qual o efeito de um determinado anticancerígeno sobre o tempo de sobrevida?
  - Quais os fatores associados ao tempo de duração da amamentação?
  - Quais os fatores preditivos para reinternação hospitalar, considerando o tempo entre internações?
  - Qual o efeito da unidade assistencial na sobrevida após um infarto agudo do miocárdio?

Programa

O Tempo

Tempo até...

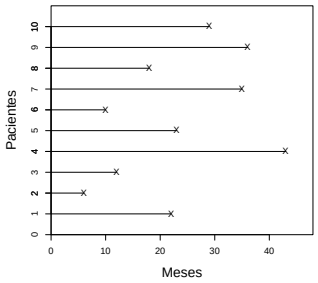
- óbito
- transplante
- doença
- cura

Medir o tempo

Tabela: Tempo de sobrevida (em meses) de 10 pacientes em diálise.

Paciente (i)	Tempo (T_i)
1	22
2	6
3	12
4	43
5	23
6	10
7	35
8	18
9	36
10	29

Representar o tempo



Cada linha representa a trajetória de um paciente e o símbolo **X** indica a ocorrência do evento ou falha.

Informação incompleta

- óbito por outras causas – morte do paciente por causas externas;
- término do estudo;
- perda de contato – mudança de residência;
- recusa em continuar participando do estudo;
- mudança de procedimento;
- abandono devido a efeitos adversos de tratamento (!!!);
- desconhecimento da data de início – em pacientes HIV+ com data de infecção desconhecida;
- dados truncados – prevalentes.

Censura e **truncamento**

Mecanismos de censura

Censura à direita

- É a mais comum.
- Sabe-se que o tempo entre o início do estudo e o evento é maior do que o tempo observado.
- Nesse caso aproveita-se a informação do tempo durante o qual a pessoa esteve sob observação sem que ocorresse o evento.
- Desprezar essa informação faria com que o risco fosse superestimado, pois o tempo até a evento é desconhecido, mas o paciente estava em risco de sofrer o evento pelo menos até o último momento observado.

Dados com censura à direita

Exemplo

- Visando estudar o tempo entre o diagnóstico de Aids e o óbito, 193 pacientes foram acompanhados em um ambulatório especializado de 1986 a 2000. Durante esse período, foram observados 92 óbitos. Sabemos que até a data de término do estudo, em dezembro de 2000, 101 permaneciam vivos. Não há informações mais recentes. Dizemos, então, que ocorreram 92 eventos e 101 censuras (à direita).

http://dengue.procc.fiocruz.br/~sobrevida/dados/aids.html

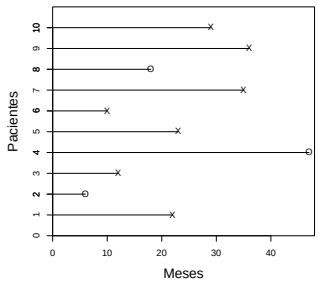
Dados com censura à direita

Dados de 10 pacientes

Paciente (i)	Tempo (T_i)	Censura
1	22	1
2	6	0
3	12	1
4	43	0
5	23	1
6	10	1
7	35	1
8	18	0
9	36	1
10	29	1

Dados com censura à direita

Graficamente

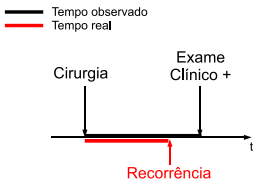


X indica ocorrência do evento e **O** corresponde à presença de censura.

Mecanismos de censura

Censura à esquerda

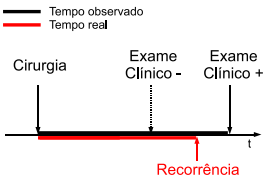
- Acontece quando não conhecemos o momento da ocorrência do evento, mas sabemos que a duração do evento é menor do que a observada.
- Considere um estudo para investigar o tempo de recorrência de um câncer após remoção cirúrgica. Três meses após a operação, pacientes são examinados e alguns apresentaram novos tumores. Para tais pacientes, sabemos que o tempo entre a cirurgia e a recorrência é menor que três meses, como indica o quadro abaixo



Mecanismos de censura

Censura intervalar

- Ocorrência do evento entre tempos conhecidos
- Aqui o paciente não apresenta recorrência na consulta após três meses da cirurgia, mas sim na consulta seguinte, ralizada 2 meses depois da anterior.
- O tempo até a recorrência é **maior** do que 3 meses e **menor** do que 5 meses.



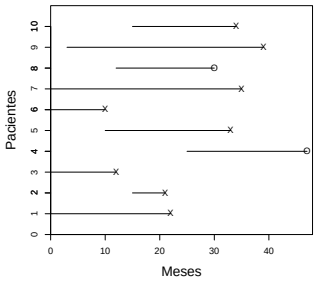
Informativa???

A censura ainda pode ser classificada em:

- Informativa: perda do indivíduo em decorrência de causa associada ao evento estudado. Por exemplo, abandono do tratamento devido a piora do paciente
- NÃO Informativa: quando não há razão para suspeitar que o motivo da perda de informação esteja relacionado ao desfecho

Coorte aberta

Momento de entrada dos pacientes na coorte varia



Trajetórias individuais de pacientes com censura e com diferentes tempos de entrada em observação.

Registro do tempo

Tempo de observação de pacientes de uma coorte aberta.

Paciente	Tempo* inicial (I)	Tempo* final (F)	Tempo* T (final - inicial)	Status (C)
1	0	22	22	1
2	15	21	6	1
3	0	12	12	1
4	25	47	22	0
5	10	33	23	1
6	0	10	10	1
7	0	35	35	1
8	12	30	18	0
9	3	39	36	1
10	15	34	19	1

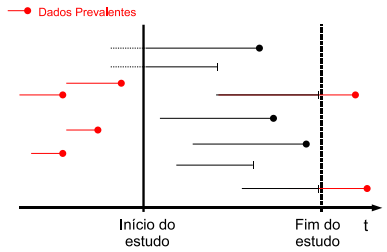
*Tempo calendário em meses

Truncamento

- À esquerda – quando a perda da informação está relacionada a indivíduos que foram excluídos do estudo porque já tinham experimentado o evento antes do início do estudo e não puderam ser observados (**dados prevalentes**).
- À direita – quando o critério de seleção inclui somente indivíduos que sofreram o evento. Não é problema em doenças com curta duração

Truncamento

Representação gráfica



Pacientes em vermelho não serão incluídos no estudo

Processo de contagem

O par (T_i, C_i) é substituído por $(N_i(t), Y_i(t))$, onde:

$N_i(t)$ é o número de eventos observados em $[0, t]$

$Y_i(t) = 1$, se o indivíduo i está sob observação e sujeito ao risco do evento no instante t

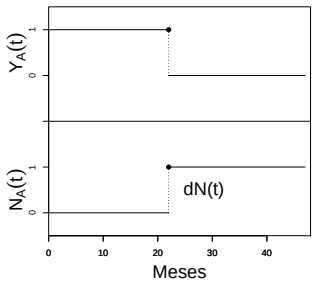
$Y_i(t) = 0$, se o indivíduo i não está em risco.

Processo de contagem

Formalmente:

- um processo de contagem é um processo estocástico $N(t)$ com $t > 0$, de tal forma que $N(0) = 0$ e $N(t) < \infty$;
- a trajetória de $N(t)$ é contínua à direita a partir de uma função escada com saltos de tamanho igual a um;
- a análise de sobrevida pode ser pensada como um processo de contagem onde $N(t)$ é o número de eventos observados até o tempo t e $dN_i(t)$ é a diferença entre a contagem de eventos até o instante t e a contagem no momento imediatamente anterior a t.

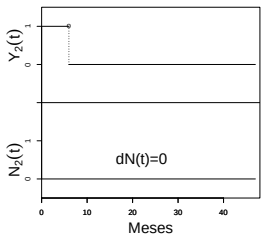
Graficamente



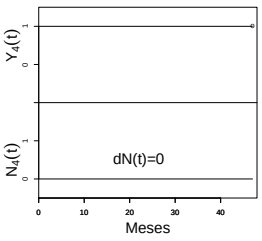
Paciente A: Diagnosticado no mês zero, acompanhado até o mês 22. A ocorrência do evento é assinalada pelo sinal
•

Graficamente

Trajetória de dois pacientes censurados



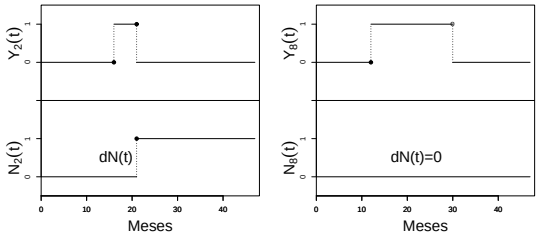
censura aos 6 meses



censura ao término do estudo

Graficamente

Trajetória de dois pacientes censurados que entraram na coorte ao longo do estudo



Qual o ganho?

O que se ganha com o processo de contagem?

Possibilidade de analisar:

- Mudança no valor de covariável
- Evento múltiplos
- Dados prevalentes

Organização dos dados

id	tempo (T)	censura (C)	sexo	idade
1	30	0	F	54
2	14	1	F	34
3	23	1	M	65
4	11	1	F	45
5	12	0	M	44

Organização dos dados

id	inicio (I)	fim (F)	censura (C)	sexo	idade
1	0	30	0	F	54
2	5	19	1	F	34
3	3	26	1	M	65
4	0	11	1	F	45
5	4	16	0	M	44

Tempo de Sobrevida no R

- O R aceita os dois formatos de registro do tempo de sobrevida.
- O comando *Surv()* tem como função combinar, em uma única variável, a informação referente ao tempo de sobrevivência de cada indivíduo e a informação a respeito do status do paciente.
  - Status = 1 (um), se ocorreu o evento
  - Status = 0 (zero) se o tempo foi censurado
- *require(survival)*
  - *Surv(tempo,status)*
  - *Surv(inicio,fim,status)*

O objeto sobrevida

```
> require(survival)
> ipec<-read.table("ipec.csv",header=T,sep=";")
> ipec[1:9,c("id","tempo","status")]
  id tempo status
1  1   852      1
2  2   123      1
3  3  1145      1
4  4  2755      0
5  5  2117      0
6  6   329      0
7  7    60      1
8  8   151      1
9  9  1563      1

> Surv(ipec$tempo,ipec$status)
[1]   852   123  1145  2755+ 2117+  329+   60   151  1563
```


Outline

Funções de sobrevida

- Densidade de Probabilidade
- Sobrevida
- Risco (instantâneo)
- Risco Acumulado

Função – densidade de probabilidade

- T – tempo de sobrevida (até a ocorrência de um evento);
- T é uma variável aleatória contínua e positiva;
- $f(t)$ é a sua função de densidade de probabilidade;
- a função $f(t)$ pode ser interpretada como a probabilidade de um indivíduo sofrer um evento em um intervalo instantâneo de tempo.

$$f(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t)}{\Delta t}$$

Estimativa de probabilidade sem censura

Se não houver censura, isto é, se todos os pacientes apresentarem o evento antes do fim do estudo, a função $f(t)$ pode ser estimada a partir da tabela de freqüência.

Nesta tabela, os valores observados de T são distribuídos em classes e para cada classe x, calcula-se $f_x(t)$:

$$\hat{f}_x(t) = \frac{\text{n}^o \text{ de ocorrências na classe } x}{(\text{n}^o \text{ total de ocorrências}) \times (\text{amplitude de } x)} \qquad (1)$$

Função de sobrevida

Qual é a probabilidade de um paciente com aids sobreviver 365 dias ou mais? Isto é, qual a probabilidade de T ser maior do que um determinado valor $t = 365$? Ou, mais formalmente, qual é $Pr(T \geq 365)$?

A função de sobrevida, $S(t)$, é a probabilidade de um indivíduo sobreviver por mais do que um determinado tempo t.

$$S(t) = Pr(T \geq t)$$

Função de sobrevida

Relembrando: a função de distribuição acumulada, $F(t)$, de uma variável aleatória é definida como a probabilidade de um evento ocorrer até o tempo t.

$$F(t) = Pr(T < t)$$

Logo, $S(t)$ é o complemento da função de distribuição acumulada $F(t)$:

$$S(t) = Pr(T \geq t) = 1 - Pr(T \leq t) = 1 - F(t)$$

Estimando a sobrevida – sem censura

$$\hat{S}_x(t_{inf}) = \frac{\text{n}^o \text{ pacientes com } T > t_{inf}}{\text{n}^o \text{ total de pacientes}}$$

em que t_{inf} é o limite inferior do intervalo de tempo considerado x.

Função de Risco

- Qual é o risco de um paciente com aids vir a óbito após sobreviver 365 dias?
- Esse risco de morrer aumenta ou diminui com o tempo?

$\lambda(t)$ –> probabilidade instantânea de um indivíduo sofrer o evento em um intervalo de tempo t e $(t + \epsilon)$ dado que ele sobreviveu até o tempo t.

Sendo ϵ infinitamente pequeno, $\lambda(t)$ expressa o risco instantâneo de ocorrência de um evento, dado que até então o evento não tenha ocorrido.

Função de Risco

$$\lambda(t) = \lim_{\epsilon \longrightarrow \infty} \frac{Pr((t < T < t + \epsilon)|T > t)}{\epsilon}$$

- $\lambda(t)$ também é denominada:
  - função ou taxa de incidência,
  - força de infecção,
  - taxa de falha,
  - força de mortalidade,
  - força de mortalidade condicional.
- Apesar do nome risco, $\lambda(t)$ é uma taxa ($tempo^{-1}$).
- Pode assumir qualquer valor positivo (não é probabilidade).

Função de Risco e de Sobrevida

$$\lambda(t) = \frac{f(t)}{S(t)}$$

$$\lambda(t) = -\frac{d \ln(S(t))}{dt}$$

Sobrevida e risco são inversamente proporcionais: quando o risco aumenta, a probabilidade de sobrevida diminui e vice-versa.

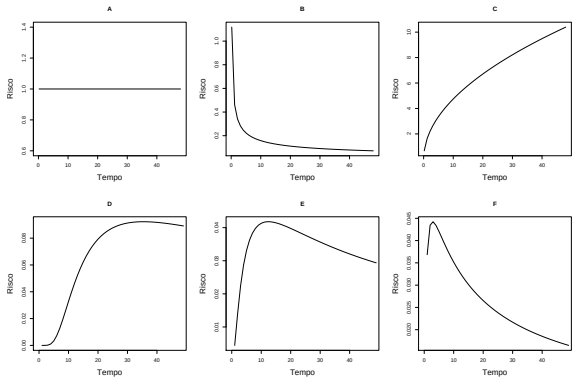
Estimando risco sem censura

$$\hat{\lambda}_x(t) = \frac{\text{n}^o \text{ ocorrências na classe } x}{R_x(t) \times (\text{amplitude de } x)}$$

Número de eventos observados no intervalo de classe x divididos pelo número de pacientes em risco no início do intervalo x e pela amplitude de x.

Uma maneira alternativa de estimar $\lambda(t)$ é utilizar as relações entre $S(t)$, $f(t)$ e $\lambda(t)$.

Comportamento do Risco



Função de risco com diversos formatos.

Função de risco acumulado

- Qual o risco de um paciente com aids vir a óbito no primeiro ano após o diagnóstico?
- Qual é o risco dele vir a óbito nos primeiros 2 anos?

$\Lambda(t)$ –> função de risco acumulado.
Mede o risco de ocorrência do evento até o tempo t.
É a soma (integral) de todos os riscos em todos os tempos até o tempo t.

$$\Lambda(t) = \int_0^t \lambda(u) d(u)$$

Também é uma taxa, logo não está restrita ao intervalo $[0; 1]$.

Estimando risco acumulado sem censura

$$\hat{\Lambda}_x(t) = \sum_{k=2}^{k=x-1} \hat{\lambda}_k(t) \times \text{amplitude de } k$$

- O risco acumulado até o tempo t é igual a:
  - o risco acumulado até o tempo $t-1$ mais
  - o risco instantâneo do período anterior vezes o intervalo de tempo até t.

Planilha exerciciotempo.ods

Relação entre as funções básicas de sobrevida

$$\begin{aligned} S(t) &= 1 - F(t) \\ \lambda(t) &= -\frac{d\ln(S(t))}{dt} \\ \lambda(t) &= \frac{f(t)}{S(t)} \\ \lambda(t) &= \frac{f(t)}{1-F(t)} \\ \Lambda(t) &= -\ln(S(t)) \end{aligned}$$

Outline

Estimação Não-Paramétrica

- Estimadores de sobrevida e risco
- Kaplan-Meier e Nelson Aalen
- Intervalos de confiança
- Kaplan-Meier estratificado
- Testes de Log-Rank e Peto

Incorporando a censura

Sem suposições sobre a distribuição do tempo

Kaplan-Meier

- A probabilidade de sobrevida até o tempo t é estimada considerando que a sobrevivência até cada tempo é independente da sobrevivência até outros tempos.
- A probabilidade de chegar até o tempo t é o produto da probabilidade de chegar até cada um dos tempos anteriores.

Kaplan-Meier

- Seja $t_1 < t_2 < \cdots < t_m$ os tempos onde ocorreram os eventos;
- $Y_i(t) = 1$ se a pessoa i está em risco no tempo t e 0 caso contrário.
- $R(t_i)$ é o total de pessoas a risco no tempo t_i.
- A cada tempo t_i em que houver um evento, a probabilidade de sobrevivência será o número dos que sobreviveram até aquele tempo $(R(t_i) - N(t_i))$ sobre os que estavam em risco naquele tempo $(R(t_i))$.
- O estimador da distribuição $S(t)$ é o produto das probabilidades de sobrevivência a cada tempo $t_i \leq t$.

Kaplan-Meier

$$\hat{S}_{KM}(t) = \left(\frac{R(t_1) - N(t_1)}{R(t_1)}\right) \times \left(\frac{R(t_2) - N(t_2)}{R(t_2)}\right) \times \cdots \times \left(\frac{R(t_m) - N(t_m)}{R(t_m)}\right)$$

ou na forma de produtório:

$$\hat{S}_{KM}(t) = \prod_{t_i \leq t} \frac{R(t_i) - N(t_i)}{R(t_i)}$$

Da sobrevida ao risco

$$\hat{\Lambda}_{KM}(t) = -\ln \hat{S}_{KM}(t)$$

Logo.... pode-se estimar qualquer das funções.

Estimador de Nelson-Aalen

$$\hat{\Lambda}_{NA}(t) = \sum_{t_i \leq t} \frac{N(t_i)}{R(t_i)}$$

Melhor para amostras muito pequenas

planilha exerciciokm.ods

Intervalos de confiança

Variância do estimador Kaplan-Meier para a sobrevida
Estimador de Greenwood

$$Var(\hat{S}_{KM}(t)) = (\hat{S}_{KM}(t))^2 \sum_{t_i \leq t} \frac{N(t_i)}{R(t_i)(R(t_i) - N(t_i))}$$

Intervalos de confiança

Assumindo erro α, o intervalo fica assim:

$$\left[\hat{S}_{KM}(t) - z_{\alpha/2}\sqrt{Var(\hat{S}_{KM}(t))}; \hat{S}_{KM}(t) + z_{\alpha/2}\sqrt{Var(\hat{S}_{KM}(t))}\right]$$

Entretanto, este intervalo permite valores negativos e maiores do que 1, o que é incompatível com a definição de sobrevida.

Intervalos de confiança

Construindo intervalo simétrico para o risco $\ln \Lambda(t) = \ln(-\ln S(t))$, pode-se obter um intervalo assimétrico para $S(t)$, porém sempre positivo e menor do que 1 e igual a

$$[\exp(-\exp(l_s)); \exp(-\exp(l_i))]$$

onde

$$[l_i; l_s] = \left[\ln(\hat{\Lambda}_{KM}(t)) - z_{\alpha/2}dp; \ln(\hat{\Lambda}_{KM}(t)) + z_{\alpha/2}dp\right]$$

e o desvio padrão dp é:

$$dp = \sqrt{\frac{\sum_{t_i \leq t} \frac{N(t_i)}{R(t_i)(R(t_i)-N(t_i))}}{\left\{\sum_{t_i \leq t} \ln\left[\frac{R(t_i)-N(t_i)}{N(t_i)}\right]\right\}^2}}$$

no R

- Criando o objeto sobrevida (tempo, censura):

```
 > Surv(tempo,status)
# variável status=1 indica evento, 0 censura

16 18 21+ 21 22 25+ 29 35 37 39 40 50+ 52 54 60 80+ 80 81+ 83 84 85+
```

- Kaplan-Meier

```
> KM <- survfit(Surv(tempo,status), data = ipec90)
> summary(KM)
> plot(KM)
```

- Nelson-Aalen

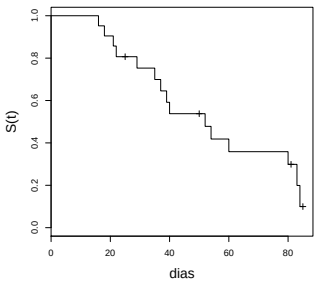
```
> sob.NA <- survfit(coxph(y~1, data = ipec90))
> sob.NA
> summary(sob.NA)
```


Saídas do R – summary(KM)

```
time n.risk n.event survival std.err lower95%CI upper95%CI
16     21      1    0.9524  0.0465   0.8655    1.000
18     20      1    0.9048  0.0641   0.7875    1.000
21     19      1    0.8571  0.0764   0.7198    1.000
22     17      1    0.8067  0.0869   0.6531    0.996
29     15      1    0.7529  0.0963   0.5859    0.968
35     14      1    0.6992  0.1034   0.5232    0.934
37     13      1    0.6454  0.1085   0.4642    0.897
39     12      1    0.5916  0.1120   0.4082    0.857
40     11      1    0.5378  0.1140   0.3550    0.815
52      9      1    0.4781  0.1160   0.2972    0.769
54      8      1    0.4183  0.1158   0.2431    0.720
60      7      1    0.3585  0.1137   0.1926    0.667
80      6      1    0.2988  0.1093   0.1459    0.612
83      3      1    0.1992  0.1092   0.0680    0.583
84      2      1    0.0996  0.0891   0.0172    0.575
```


Saídas do R – plot(KM)

Função de sobrevida dos pacientes com aids, utilizando o estimador produto Kaplan-Meier.
Os símbolos + localizam as censuras.

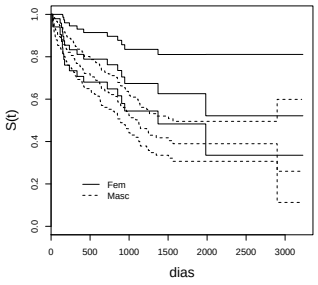


Kaplan-Meier estratificado

- A sobrevivência é estimada separadamente para cada estrato, utilizando Kaplan-Meier.
- no R

```
> ipec <- read.table("ipec.csv",header=T,sep=";")
> survaids <- survfit(Surv(tempo,status)~ sexo, data = ipec)
> survaids

Call: survfit(formula = resp ~ sexo, data = ipec)

         n events rmean se(rmean) median 0.95LCL 0.95UCL
sexo=F  49     16  2096       229    Inf    1371     Inf
sexo=M 144     74  1581       122   1116     887    1563
```


Gráfico sobrevida estratificada



Curvas de sobrevida de pacientes com aids, estratificado por sexo.
Estimação por Kaplan-Meier, com intervalo de confiança de 95%.

Testes

Hipótese nula: não há diferença entre estratos

$$H_0 : \lambda_1(t) = \lambda_2(t) = \cdots = \lambda_k(t)$$

Log-rank (ou Mantel-Haenszel)

Distribuição esperada de eventos igual em todos os estratos:

$$e_k(t) = N(t)\frac{R_k(t)}{R(t)}$$

Estatística de teste log-rank para dois estratos ($k = 2$):

$$\text{Log-rank} = \frac{(N_1 - E_1)^2}{Var(N_1 - E_1)}$$

com $N_1 =$ ao total de eventos observados no estrato 1 e $E_1 =$ ao total de eventos esperados no estrato 1.

Teste log-rank

A variância, que entra no cálculo como um fator de padronização, tem a fórmula (para $k = 2$):

$$Var(N_1 - E_1) = v_i$$

em que

$$v_i = \sum_{t_i} \frac{R_1(t_i)[R(t_i) - R_1(t_i)]N(t_i)[R(t_i) - N(t_i)]}{R(t_i)^2[R(t_i) - 1]}.$$

A estatística log-rank, sob a hipótese nula, segue uma distribuição χ^2 , com $k - 1$ graus de liberdade.

Teste de Peto

Dá maior peso às diferenças (ou semelhanças), no início da curva, onde se concentra a maior parte dos dados e por isso é mais informativa. Usa um ponderador $S(t)$ no estimador.

$$\text{Peto} = \frac{(N_1 - E_1)^2}{Var(N_1 - E_1)}$$

sendo que

$$N_1 - E_1 = \frac{\sum S(t_i)(N_1(t_i) - E_1(t_i))}{\sum S(t_i)}$$

$$Var(N_1 - E_1) = \frac{(\sum S(t_i)(N_1(t_i) - E_1(t_i)))^2}{\sum (S(t_i))^2 v_i}$$

Também a estatística Peto segue aproximadamente uma distribuição χ^2 com $k - 1$ graus de liberdade.

no R

```
> survdiff(Surv(tempo,status)~sexo, data=ipec,rho=0)

Call:
survdiff(formula = Surv(tempo, status) ~ sexo, data = ipec, rho = 0)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sexo=F  49       16     24.5      2.93      4.03
sexo=M 144       74     65.5      1.09      4.03

 Chisq= 4  on 1 degrees of freedom, p= 0.0447
```

O argumento *rho* determina o tipo de teste a ser realizado. Para log-rank, use *rho = 0* (*default*). Para o teste Peto, use *rho = 1*.

no R

```
> survdiff(Surv(tempo,status)~sexo, data=ipec,rho=1)

Call:
survdiff(formula = Surv(tempo, status) ~ sexo, data = ipec, rho = 1)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sexo=F  49     12.1     18.2     2.011      3.54
sexo=M 144     55.1     49.0     0.746      3.54

 Chisq= 3.5  on 1 degrees of freedom, p= 0.0598
```


Outline

Riscos Proporcionais

O modelo de regressão mais amplamente utilizado para dados de sobrevida ajusta a função de risco $\lambda(t)$, considerando um risco basal $\lambda_0(t)$ e incluindo o vetor de covariáveis $\boldsymbol{x}$, de forma que:

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp(x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p) = \lambda_0(t)\exp(\boldsymbol{x\beta})$$

Ou seja, as covariáveis têm um efeito multiplicativo na função de risco.

Riscos Proporcionais

A razão entre os riscos de ocorrência do evento de dois indivíduos i e j, com covariáveis $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ e $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \cdots, x_{jp})$ é:

$$\frac{\lambda_i(t|\boldsymbol{x}_i)}{\lambda_j(t|\boldsymbol{x}_j)} = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\exp(\boldsymbol{x}_j\boldsymbol{\beta})}$$

Observe que esta razão de riscos NÃO varia ao longo do tempo –> Modelo de Riscos Porporcionais

Riscos Proporcionais

O modelo RP também pode ser escrito em termos da função de risco acumulado ou da função de sobrevida:

$$\Lambda(t|\boldsymbol{x}) = \Lambda_0(t)\exp(\boldsymbol{x\beta})$$
$$S(t|\boldsymbol{x}) = [S_0(t)]^{\exp(\boldsymbol{x\beta})}$$

O risco acumulado basal é $\Lambda_0(t) = \sum_{i:\, t_i \le t} \frac{N_i(t)}{\sum_{j\in R(t_i)}\exp(\boldsymbol{x}_j\boldsymbol{\beta})}$ e a sobrevida basal é dada por $S_0(t) = \exp[-\Lambda_0(t)]$

Modelo de Cox

Partindo desta proporcionalidade, é possível estimar os efeitos das covariáveis sem qualquer suposição a respeito da distribuição do tempo de sobrevida, e por isso o modelo de Cox é dito semi-paramétrico.
Não se assume qualquer distribuição estatística para a função de risco basal, $\lambda_0(t)$, apenas que as covariáveis agem multiplicativamente sobre o risco e esta é a parte paramétrica do modelo.

Pressupostos do modelo de Cox

- As covariáveis agem multiplicativamente sobre o risco –> parte paramétrica do modelo.
- A razão de riscos é constante ao longo de tempo –> riscos proporcionais.
- Os tempos de ocorrência do evento são independentes.

Estimativa dos coeficientes

Para estimar os coeficientes da regressão paramétrica, a função de verossimilhança foi construída a partir da função de densidade de probabilidade calculada nos tempos de ocorrência do evento, multiplicada pela função de sobrevida calculada nos tempos de censura.

No Modelo de Cox o vetor de parâmetros $\boldsymbol{\beta}$ é estimado a partir de uma verossimilhança parcial.
De forma semelhante ao Kaplan Meier, considera-se apenas, a cada tempo t, a informação dos indivíduos sob risco, estimando os efeitos das covariáveis no tempo de sobrevida.

Verossimilhança parcial

- Considere m diferentes tempos até a ocorrência de um evento (sem empate), ordenados assim: $t_1 < t_2 < \ldots < t_m$.
- A verossimilhança individual, L_i, é a razão entre o risco $\lambda_i(t_i)$ do indivíduo i falhar em t_i e a soma dos riscos de ocorrência de evento de todos os indivíduos em risco:

$$L_i = \frac{\lambda_i(t_i)}{\sum_{j \in R(t_i)} \lambda_j(t_j)} = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j\boldsymbol{\beta})}$$

Verossimilhança parcial

- Sob o processo de contagem a verossimilhança individual é igual a

$$L_i = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\sum_{t \geq 0} Y_j(t) \exp(\boldsymbol{x}_j\boldsymbol{\beta})},$$

- com $Y_j(t)$ igual a 1 se o indivíduo j estiver em risco no tempo t e 0, caso contrário.

Verossimilhança Parcial

- A verossimilhança parcial $L(\boldsymbol{\beta})$ = produto das L_i

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{t \geq 0} \left\{ \frac{Y_i(t) \exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\sum_j Y_j(t) \exp(\boldsymbol{x}_j\boldsymbol{\beta})} \right\}^{dN_i(t)}$$

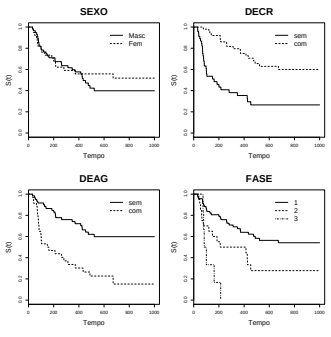
- $dN_i(t)$ = diferença entre a contagem de eventos até o instante t e a contagem no momento imediatamente anterior a t.
- Numerador depende apenas da informação dos indivíduos que experimentam o evento
- Denominador utiliza informações a respeito de todos os indivíduos que ainda não experimentaram o evento, incluindo aqueles que serão censurados mais tarde.

Exemplo TMO

- Avaliar os fatores prognósticos associados ao tempo de transplante de medula óssea TMO até o óbito nos pacientes com leucemia mielóide crônica tratados no INCA.
- covariáveis:
  - sexo,
  - idade,
  - fase da doença no momento do transplante (`fase`),
  - a ocorrência ou não de doença enxerto contra hospedeiro aguda (`deag`) ou crônica (`decr`).

Proporcionalidade

Curvas de KM para avaliar o pressuposto de proporcionalidade



No R

```
> tmocens <- read.table("tmoclas.dat", header=T, sep=",")
> mod1 <- coxph(Surv(os,status)~idade+factor(sexo),data=tmocens, x=TRUE)
> summary(mod1)

Call:
coxph(formula = Surv(os, status) ~ idade + factor(sexo), data = tmocens,
    x = TRUE)
  n= 96
                   coef exp(coef) se(coef)     z    p
idade          -0.0186     0.982   0.0141 -1.32 0.19
factor(sexo)2 -0.3299     0.719   0.3219 -1.02 0.31
              exp(coef) exp(-coef) lower .95 upper .95
idade             0.982       1.02     0.955      1.01
factor(sexo)2     0.719       1.39     0.383      1.35

Rsquare= 0.022   (max possible= 0.984 )
Likelihood ratio test= 2.16  on 2 df,   p=0.34
Wald test            = 2.11  on 2 df,   p=0.348
Score (logrank) test = 2.11  on 2 df,   p=0.348
```

Selecionando modelos

- Teste de Wald
- Análise da função desvio

Comparando quatro modelos

```
> anova(mod1,mod2,mod3,mod4,test='Chisq')
Analysis of Deviance Table

Model 1: Surv(os, status) ~ idade + factor(sexo)
Model 2: Surv(os, status) ~ idade + factor(sexo) + factor(fase)
Model 3: Surv(os, status) ~ idade + factor(sexo) + factor(fase) + deag
Model 4: Surv(os, status) ~ idade + factor(sexo) + factor(fase) + deag +
    decr
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        94     395.93
2        92     380.78  2    15.14 0.0005146
3        91     366.67  1    14.11 0.0001726
4        90     358.20  1     8.47 0.0036015
```


Selecionando Modelos

- A função desvio é assintoticamente semelhante à estatística de Wald quando o número de observações é grande.
- Para número de observações pequenos, a análise da função desvio é mais robusta.
- Outra ressalva a respeito de valores ausentes. Caso eles existam para algumas variáveis incluídas em alguns modelos, mesmo que aninhados, os modelos perdem a comparabilidade.

Medida Global de Ajuste

- R^2 – poder explicativo das covariáveis no tempo de ocorrência do evento em estudo.

$$\begin{aligned} R^2_{LR} &= 1 - \{L(0)/L(\hat{\boldsymbol{\beta}})\}^{2/n} \\ &= 1 - \exp(2\{l(0) - l(\hat{\boldsymbol{\beta}})\}/n) \end{aligned}$$

- Valor mínimo possível de R^2 é zero quando $L(0) = L(\hat{\boldsymbol{\beta}})$
- Valor máximo não é 1 (ou 100%), mas a razão entre as verossimilhanças do modelo saturado e do modelo nulo.

Medida Global de Ajuste

Modelo	ln(Verossimil.)	R^2	% Var. Explicada*
Nulo	-199,0424	0,000	0,0%
Saturado	-0,2670	0,984	100,0%
M1: Idade+Sexo	-197,9626	0,022	2,2%
M2: Mod1+Fase	-190,3905	0,165	16,8%
M3: Mod2+*deag*	-183,3364	0,279	28,4%
M4: Mod3+*decr*	-179,0992	0,340	34,6%

* $R^2_{modelo}/R^2_{saturado}$

Medida Global de Ajuste

Gráfico de sobrevida estratificado por índice de prognóstico (IP)

- IP é o preditor linear do modelo de Cox, $\boldsymbol{x\beta}$, calculado para cada indivíduo usando as covariáveis observadas e as estimativas dos coeficientes de regressão do modelo ajustado.
- Os indivíduos são estratificados em grupos de tamanhos aproximadamente iguais (grupos de alto, médio e baixo IP)
- Os valores médios de cada uma das covariáveis dentro de cada grupo são utilizados para obtenção de curvas de sobrevida sob o modelo ajustado.
- Espera-se, se o modelo for razoável, que o gráfico das curvas ajustadas pelo modelo em cada estrato sejam próximas das estimadas por Kaplan-Meier.

Medida Global de Ajuste

- Assumindo modelo *mod4*
- Indivíduo 1: sexo masculino (`sexo = 0`) com 56 anos (`idade = 56`), na fase intermediária (`fase2 = 1` e `fase3 = 0`), com manifestação de doença do enxerto aguda (`deag=1`, `decr=0`)

$$
\begin{array}{lllll}
\beta_{idade} & \times 56 = & -0,0044 & \times 56 & = -0,2469 \\
\beta_{sexo} & \times 0 = & -0,2260 & \times 0 & = 0 \\
\beta_{fase2} & \times 1 = & 0,6413 & \times 1 & = 0,6413 \\
\beta_{fase3} & \times 0 = & 1,0279 & \times 0 & = 0 \\
\beta_{deag} & \times 1 = & 1,2530 & \times 1 & = 1,2530 \\
\beta_{decr} & \times 0 = & -0,9775 & \times 0 & = 0
\end{array}
$$

Soma $= 1,6474$

Medida Global de Ajuste

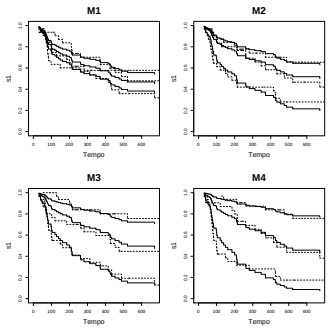
- Assumindo modelo *mod4*
- Indivíduo 2: sexo feminino (`sexo = 1`) com 20 anos (`idade = 20`), na fase avançada (`fase2 = 0` e `fase3 = 1`) com manifestação de doença do enxerto aguda (`deag=1`, `decr=0`)

$$
\begin{array}{lllll}
\beta_{idade} & \times 20 = & -0,0044 & \times 20 & = -0,0882 \\
\beta_{sexo} & \times 1 = & -0,2260 & \times 1 & = -0,2260 \\
\beta_{fase2} & \times 0 = & 0,6413 & \times 0 & = 0 \\
\beta_{fase3} & \times 1 = & 1,0279 & \times 1 & = 1,0279 \\
\beta_{deag} & \times 1 = & 1,2530 & \times 1 & = 1,2530 \\
\beta_{decr} & \times 0 = & -0,9775 & \times 0 & = 0
\end{array}
$$

Soma $= 1,9667$

Medida Global de Ajuste

Gráfico de sobrevida estratificado por índice de prognóstico.



Linha sólida representa o modelo ajustado e linha pontilhada a estimativa de Kaplan-Meier.

Outline

Objetivos

Analisar o ajuste e as premissas do modelo de Cox.

São três tipos de resíduos:

- Schoenfeld
- Martingale
- escore

Pressupostos

- proporcionalidade: a relação entre variável resposta e variável independe do tempo.
- linearidade (log-linearidade, pois a função de risco $\lambda(t)$ tem uma estrutura log-linear): a razão de riscos entre um indivíduo de 45 anos e um de 50 anos é idêntica àquela entre um indivíduo de 80 anos e um de 85 anos.
- Efeito de pontos influentes (ou de alavanca).
- O resíduo obtido como a resposta observada menos a esperada não pode ser usado para os dados de sobrevida: a censura!!!

Schoenfeld

$$\boldsymbol{r}_i(\boldsymbol{\beta}) = \boldsymbol{x}_i - \frac{\sum_{j \in R(t_i)} \boldsymbol{x}_j \exp(\boldsymbol{x}_j \boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j \boldsymbol{\beta})}$$

sendo j cada indivíduo e i $(i = 1, \cdots, m)$ o índice dos tempos observados de eventos.

O resíduo de Schoenfeld é a diferença entre os valores observados de covariáveis de um indivíduo com tempo de ocorrência do evento t_i e os valores esperados em t_i dado o grupo de risco $R(t_i)$. Haverá tantos vetores de resíduos quanto covariáveis ajustadas no modelo, e que estes são definidos somente nos tempos de ocorrência do evento.

Schoenfeld

Suponha um coeficiente β_k (k é cada covariável) que varia com o tempo t. β_k pode ser dividido em duas partes:

- uma média constante – $E[r_i(\beta_k)|R(t_i)]$, com variância $V(\beta_k)$
- e uma função $U(t)$ – que varia no tempo

O resíduo padronizado de Schoenfeld em t_i pode ser obtido por:

$$r_i^*(\beta_k) = \frac{r_i(\beta_k)}{V(\beta_k)}.$$

O valor esperado deste resíduo padronizado $r_i^*(\beta_k)$ para cada grupo em risco $R(t_i)$ é aproximadamente igual à parte de β_k que varia no tempo – a função $U(t)$ – GRÁFICO.

Schoenfeld no *R*

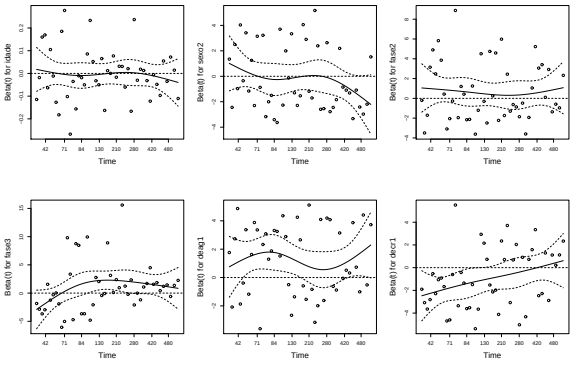
```
> residuo <- cox.zph(modelo)
> plot(residuo[1])
> abline(h=0,lty=2)
```

Atenção para a escala do tempo:

- Kaplan-Meier – nos tempos de falha
- Calendário – bom quando ajuste usnando processo de contagem, pode ficar pouco visível se concentra grande quantidade de eventos em um mesmo momento
- Rank – ordem dos eventos

A linha curva é um `lowess`.

Gráficos de Schoenfeld



Não proporcionalidade – soluções

- estratificar pela covariável tempo-dependente;
- particionar o eixo do tempo;
- outro tipo de modelo – tempo de vida acelerado;

Resíduos Martingale

É a diferença entre o número observado de eventos para um indivíduo e o esperado dado o modelo ajustado, o tempo de seguimento e o percurso observado de quaisquer covariáveis tempo-dependentes.

Resíduos Martingale

Semelhante aos resíduos dos modelos de regressão linear:

- o valor esperado $= 0$
- o somatório dos resíduos observados $= 0$
- os resíduos M_i são não correlacionados, mas as estimativas $\hat{M}_i$ são negativamente correlacionadas, ainda que fracamente

Resíduos Martingale

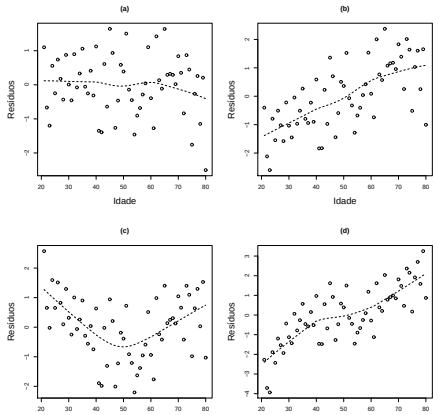
E diferentes dos resíduos da regressão linear:

- a soma de quadrados dos resíduos não auxilia na avaliação do ajuste global do modelo (o melhor modelo de Cox ajustado não tem a menor soma de quadrados de resíduos martingale);
- a distribuição dos resíduos não é aproximadamente normal;
- o gráfico de resíduos versus valores ajustados não funciona para resíduos martingale pois estes são negativamente correlacionados com os valores ajustados.

Gráficos Martingale

- M_i versus índice do indivíduo: permite revelar indivíduos mal ajustados pelo modelo;
- M_i do modelo nulo (sem covariáveis) versus covariável com a superposição de uma curva de alisamento: para avaliar a forma funcional da covariável a ser incluída no modelo.

Gráficos Martingale

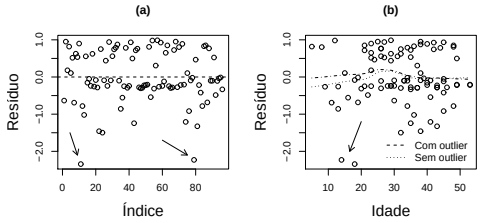


Martingale no *R*

A função para calcular o resíduo de Martingale é:

```
> res <- resid(modelo,type="martingale")
```

em que *modelo* é o objeto que recebeu o modelo de Cox.



(a) resíduo por indivíduo (b) covariável X modelo nulo

Ajuste forma funcional não linear

- Incluir uma função de alisamento: *smoothing splines*
- Vantagem sobre polinômios é ser não paramétrica
- São tratadas como covariáveis usuais, inclusive testes de hipótese para não-linearidade
- Permite estimar intervalos de confiança

Smoothing spline

- O objetivo é estimar β de tal forma que se obtenha, simultaneamente, o menor número possível de nós e a menor soma dos resíduos quadráticos para a covariável em questão
- Parâmetro θ indica afastamento da reta:
  - $\theta \to 0$, a solução converge para uma reta
  - $\theta \to 1$ a curva passa por todos os pontos
- Número de pontos pelos quais a curva passará são os graus de liberdade acrescentados ao modelo
- Escolha pelo critério de informação de Akaike (*Akaike Information Criteria* – AIC)

No *R*

```
> coxph(formula = Surv(os, status) ~ pspline(idade,df=0)+
  sexo+fase+decr+deag, data=tmocens, x=T)

Call: coxph(formula = Surv(os, status) ~ pspline(idade, df=0)
    + sexo + fase + decr + deag, data = tmocens, x = T)

                       coef    se(coef) se2    Chisq DF   p
pspline(idade,df=0),l  -0.0117 0.0157   0.0157  0.56 1.00 0.45000
pspline(idade,df=0),n                           12.27 4.63 0.02400
sexo2                  -0.2623 0.3445   0.3405  0.58 1.00 0.45000
fase2                   0.7428 0.3982   0.3946  3.48 1.00 0.06200
fase3                   1.2538 0.5657   0.5593  4.91 1.00 0.02700
decr1                  -1.1182 0.3476   0.3444 10.35 1.00 0.00130
deag1                   1.4174 0.3592   0.3537 15.57 1.00 0.00008

Iterations: 6 outer, 19 Newton-Raphson        Theta= 0.581
Degrees of freedom for terms= 5.6 1.0 2.0 1.0 1.0
Likelihood ratio test=55.3  on 10.5 df, p=4.38e-08
                n= 96
```


No *R*

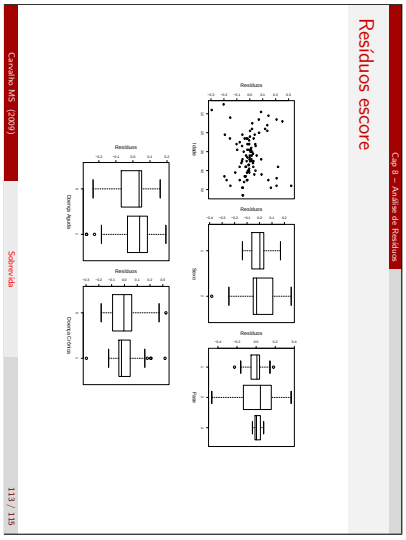
```
> anova(mod4,mod5,test='Chisq')

Analysis of Deviance Table
Model 1: Surv(os, status) ~
            idade + sexo + fase + deag + decr
Model 2: Surv(os, status) ~ pspline(idade, df = 0)
            + sexo + fase + decr + deag
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        90     358.20
2        74     342.74 16    15.46      0.49
```

Observar:

- os graus de liberdade são fracionários
- o componente linear da idade `pspline(idade,df=0),1` é não significativo
- o componente suavizado tem $p < 0,05$
- `summary(mod5)` – indica 18 nós para idade

Resíduos escore

- Verifica a influência de cada observação no ajuste do modelo
- Permite a estimação robusta da variância dos coeficientes de regressão (útil para dados em cluster)
- A influência de cada observação deve ser proporcional à $(x_i - \bar{x}) \times$ resíduo
- O gráfico do resíduo escore para cada covariável $\Delta\beta_k$ versus x mostra pontos de alavanca
- Vantagem – definidos para todos os tempos, mesmo onde não ocorre evento, melhorando a análise quando há muita censura



Resíduos escore no *R*

```
> res.esco <- resid(modelo,type="dfbetas")
> par(mfrow=c(1,2))
> plot(banco$var1,res.esco[,1],
  xlab='Var1', ylab='Resíduos')
> plot(banco$var2,res.esco[,2],
  xlab='Var2', ylab='Resíduos')
```

Observar que o objeto `res.esco` guarda em cada coluna as variáveis incluídas no modelo, na ordem em que foram colocadas. Para lembrar quais são, veja `modelo$call`

Sumário

Para	Fazer
Avaliar o pressuposto de proporcionalidade global	teste de proporcionalidade global fornecido pela função que calcula o resíduo de Schoenfeld
Avaliar o pressuposto de proporcionalidade de cada variável	gráficos do resíduo de Schoenfeld contra o tempo
Estudar a forma funcional da variável	gráficos do resíduo de martingale vs a covariável com a superposição de um alisamento da covariável
Linearizar a forma funcional da variável quando não linear	alisamento *spline* (`pspline()`) da covariável diretamente no modelo de Cox
Avaliar efeito de valores aberrantes	gráficos de resíduos escore