Introdução à Análise de Sobrevivência

Marilia Sá Carvalho
Dayse Pereira Campos
Raquel de V.C. de Oliveira

Fundação Oswaldo Cruz, Brasil

Introdução à Análise de Sobrevivência

- Introdução – Capítulo 1
- O Tempo – Capítulo 2
- Funções de Sobrevivência – Capítulo 3
- Estimação Não-Paramétrica – Capítulo 4
- Estimação Paramétrica – Capítulo 5
- Modelo de Cox – Capítulo 6
- Análise de Resíduos – Capítulo 7
- Covariável Tempo-Dependente – Capítulo 8

Métodos Avançados de Análise de Sobrevivência

1. Modelos com efeitos não lineares – Capítulo 9
2. Múltiplos Eventos – Capítulo 10
3. Eventos Competitivos – Capítulo 11
4. Fragilidade – Capítulo 12

Cronograma – Introdução

Dia	Tema
1° dia	Introdução, O tempo, Funções de Sobrevivência
2° dia	Estimação Não-Paramétrica, Cox
3° dia	Cox, Resíduos
4° dia	Resíduos, Tempo dependente

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 Sobrevivência: teoria e aplicações em saúde, 2a edição*.

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://sobrevida.fiocruz.br/`
- R software: `www.r-project.org`
- Tutorial online do R
  - `http://www.leg.ufpr.br/Rtutorial/`
  - `http://www.leg.ufpr.br/~paulojus/embrapa/Rembrapa/`

Sobrevivência

- Em que tipo de desenho de estudo se aplica a *Análise de Sobrevivência*?
  - Coorte – observacional ou de intervenção (ensaio clínico) – pressupõe o acompanhamento dos indivíduos ao longo do tempo
- Que perguntas podemos responder com os modelos de sobrevivência (ou sobrevida)?
- Definir taxa de incidência ou força de morbidade ou risco instantâneo

Sobrevivência

- A análise de sobrevivência, também chamada de análise de sobrevida, 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 sobrevivência?
  - 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 sobrevivência após um infarto agudo do miocárdio?
- Considerando a possível perda de seguimento (censura)

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?
- quando se usa regressão logística?
- quando se usa regressão de Poisson?
- o que é um intervalo de confiança?
- o que é um p-valor?
- o que é efeito de variável?
- o que significa a expressão ”controlando por idade e sexo”?

Refrescando a memória

- Modelo logístico: o efeito de um fator de exposição sobre o risco de ocorrência de um desfecho é uma probabilidade condicional de experiência do desfecho, dada a exposição – $Pr(D|E)$
- Taxa ou força de incidência ou força de morbidade ou risco instantâneo – $\lambda(t)$ – risco em expostos sobre não expostos em cada momento no tempo.

Outline

O Tempo

Tempo até ...

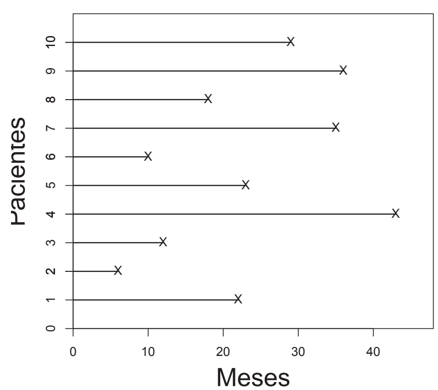
- óbito
- transplante
- doença
- cura

Medir o tempo

Tabela : Tempo de sobrevivência (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.

Causas de 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 – esquema de tratamento;
- abandono devido a efeitos adversos de tratamento;
- desconhecimento da data de início – em pacientes HIV+ com data de infecção desconhecida;
- use de dados prevalentes – óbitos antes do início do estudo.

Censura e truncamento

Mecanismos de censura

Censura à direita

- É a mais comum.
- Não se observa o desfecho.
- 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:

- 92 óbitos observados
- Ao término do estudo (dez/2000), 101 permaneciam vivos
- não há informação após essa data
- 92 eventos e 101 censuras (à direita)

http://sobrevida.fiocruz.br/

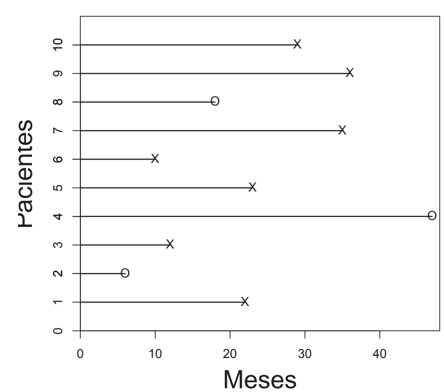
Dados com censura à direita

Dados de 10 pacientes Notação Clássica: T_i, δ_i

Paciente (i)	Tempo (T_i)	Status (δ_i)
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

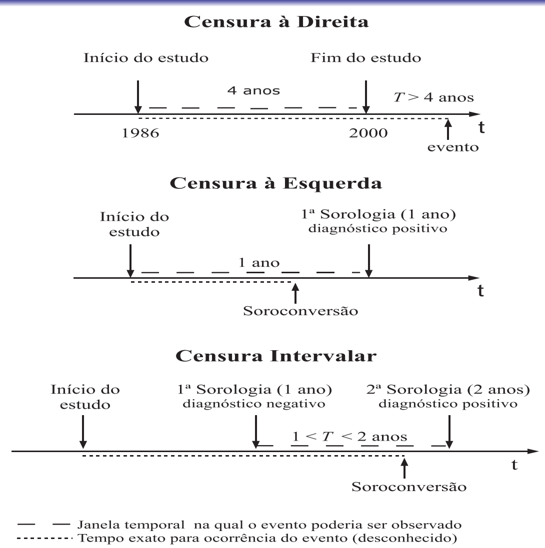
- Acontece quando não conhecemos o momento da ocorrência do evento, mas sabemos que ele ocorreu antes do tempo observado.
- Considere um estudo comunitário para investigar o fatores associados à soroconversão para leptospirose, após a entrada na comunidade onde é possível a transmissão. Caso o exame seja positivo, só podemos afirmar que a transmissão ocorreu entre a data da mudança para o local e a coleta do sangue.

Mecanismos de censura

Censura intervalar

- Ocorrência do evento entre tempos conhecidos
- No exemplo anterior seria a soroconversão entre dois exames (anuais).
- O tempo até a recorrência é **maior** do que a data do exame negativo e **menor** o primeiro exame positivo.

Mecanismos de censura



Informativa???

A censura ainda pode ser classificada em:

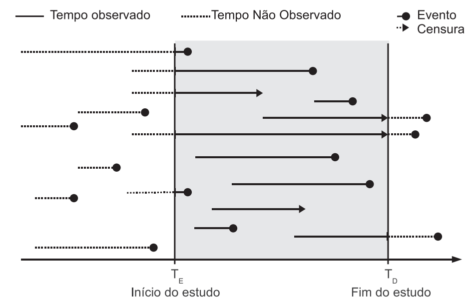
- Informativa: perda do indivíduo em decorrência de causa associada ao evento estudado.
- NÃO Informativa: quando não há razão para suspeitar que o motivo da perda de informação esteja relacionado ao desfecho.
- Avaliar a censura: comparação de censurados e não censurados segundo características.
- Evitar censura informativa – busca ativa!

Truncamento

- Indivíduos não são incluídos por motivo relacionado à ocorrência do evento estudado
- O estudo só inclui quem apresentou o evento na janela temporal (T_E, T_D), T_E – no momento do início do estudo; T_D – momento do desfecho.

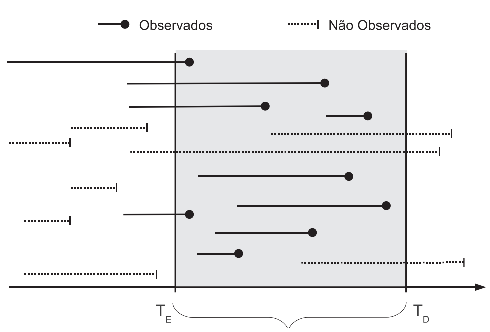
Truncamento à esquerda

- Indivíduos já experimentaram o evento antes do início do estudo
- Comum no uso de dados prevalentes, bases de dados secundários
- Como indivíduos com maior sobrevivência tem mais chance de entrar no estudo, o risco é subestimado



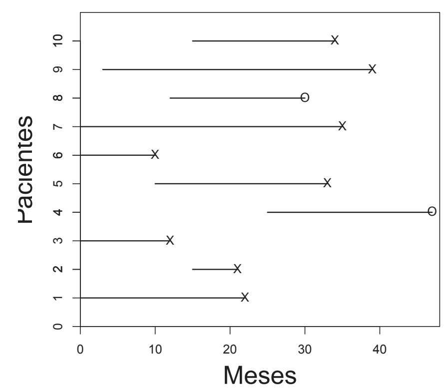
Truncamento à direita

- O critério de seleção inclui somente os que sofreram o evento, logo o risco é superestimado
- Não é problema em doenças com curta duração
- Comum em estudos que partem do óbito
- Não há censura à direita



Coorte aberta

Momento de entrada dos pacientes na coorte varia



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

Processo de contagem

A formulação do processo de contagem permite provar resultados importantes na análise de sobrevivência, acomodando censuras, truncamento, eventos múltiplos.
O par (T_i, δ_i) é substituído por $(N_i(t), Y_i(t))$, onde:

$N_i(t) =$ número (0, 1, 2,...) de eventos observados em $[0, t]$
evento único (óbito)$N_i(t) = 1$, eventos recorrentes (ex. doença oportunista) $N_i(t) = 0, 1, 2, 3 \cdots$

$Y_i(t) = 1$, 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.

Entender quem está em risco a cada momento é essencial na construção do banco de dados.

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 sobrevivência pode ser pensada como um processo de contagem onde $N(t)$ é o número de eventos observados até o tempo t e $\Delta N_i(t)$ é a diferença entre a contagem de eventos até o instante t e a contagem no momento imediatamente anterior a t.

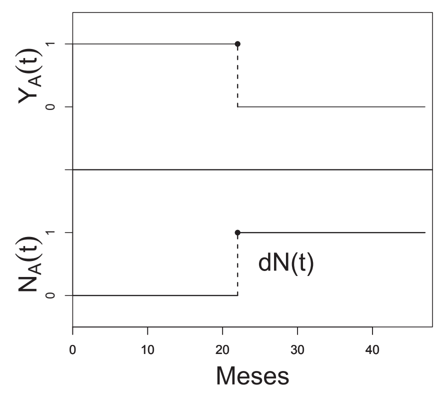
Registro do tempo

Tempo de observação de pacientes de uma coorte aberta.

Paciente	Tempo* inicial (I)	Tempo* final (F)	Tempo* T (final - inicial)	Status δ
1	0	22	22	1
2	15	21	6	0
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

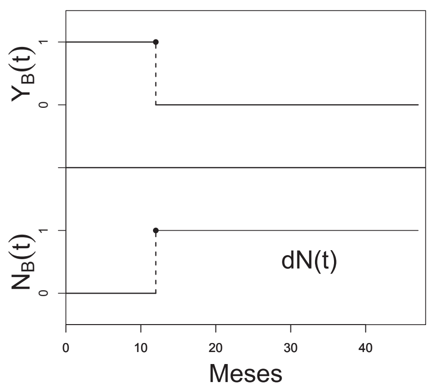
*Registrar as datas de entrada e do evento para cada paciente

Graficamente



Paciente 1: Diagnosticado no mês zero, acompanhado até o mês 22. A ocorrência do evento é assinalada pelo sinal •

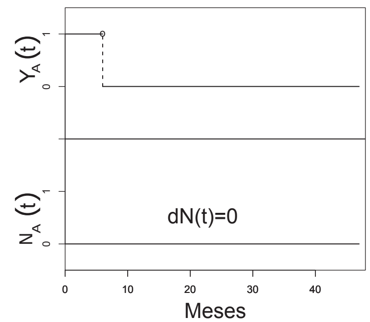
Graficamente



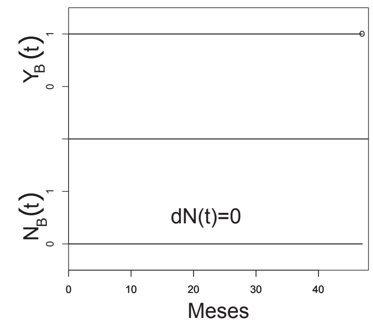
Paciente 3: Diagnosticado no mês zero, acompanhado até o mês 12.

Graficamente

Trajetória de dois pacientes censurados



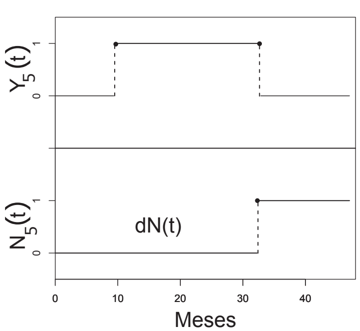
censura aos 6 meses

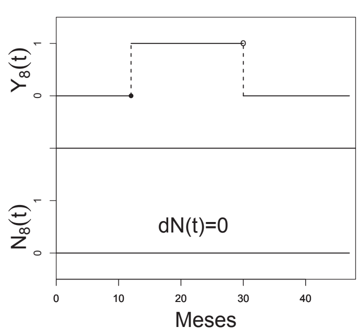


censura ao término do estudo

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: mudança de esquema ARV
- Evento múltiplos: sucessivos infartos do miocárdio
- Dados prevalentes: hemodiálise

Organização dos dados

id	tempo (T)	status (δ)	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

Tabela : Forma Clássica

Organização dos dados

id	inicio (I)	fim (F)	status (δ)	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

Tabela : Forma em Contagem

Tempo de Sobrevivência no R

- O R aceita os dois formatos de registro do tempo de sobrevivência.
- 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 sobrevivência – formato clássico

```
> 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
```

O objeto sobrevivência – formato contagem

```
id ini  fim tempo status
1 1243 2095   852      1
2 2800 2923   123      1
3 1250 2395  1145      1
4 1915 4670  2755      0
5 2653 4770  2117      0
6    3  332   329      0
7   36   96    60      1
8    1  152   151      1
```

```
> Surv(ipec$ini,ipec$fim,ipec$status)
```

```
 [1] (1243,2095 ] (2800,2923 ] (1250,2395 ] (1915,4670+]
 [5] (2653,4770+] (   3, 332+] (  36,  96 ] (   1, 152 ]
```


Tempo de Sobrevivência no R

- Surv(tempo,status) – clássico
- Surv(inicio,fim,status) – contagem

Resumo

Neste capítulo, foram apresentadas as diferentes abordagens – clássica e processo de contagem – para se estudar o tempo até a ocorrência de um evento, identificando-se:

- o tempo quando ocorre o evento;
- a população em risco em cada tempo;
- a censura não informativa e informativa;
- a censura à esquerda, à direita e intervalar;
- o truncamento à esquerda e à direita.

Outline

Funções de sobrevivência

- Introdução
- Função de Densidade de Probabilidade
- Função de sobrevivência
- Função de Risco (instantâneo)
- Comportamento da função de risco
- Função de Risco Acumulado
- Relação entre as funções
- Função de Verossimilhança

Introdução

- 50 pacientes, 4 anos de acompanhamento, 32 óbitos
- Taxa média de mortalidade: $32/50 = 0,64 = 64\%$ ou 16 óbitos por 100 pessoas/ano
- Mas... essa taxa não é homogênea no tempo.
- A análise de sobrevivência responde a:
  - Qual o **risco** de um paciente diagnosticado com Aids vir a falecer **em até** três anos após o diagnóstico?
  - Qual a **probabilidade** de um paciente sobreviver por **mais de** dois anos após o diagnóstico de Aids?
  - Qual seria o **número esperado** de óbitos em uma coorte de pacientes acompanhada por cinco anos?
  - Qual o **tempo mediano** de sobrevivência?

Função – densidade de probabilidade

- T – tempo de sobrevivência (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.

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

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

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

$$\hat{f}_x(t) = \frac{N_x(t)}{(\text{n}^{\underline{o}} \text{ total de ocorrências}) \times \Delta_x}$$

Tempos de sobrevivência – Aids, 32 pacientes

3	18	29	54	60	84	110	112	116	123	134
145	151	151	158	173	194	214	329	331	371	408
490	514	541	555	688	780	801	858	887	998	

Estimativa de probabilidade sem censura

Intervalo	$R_x(t)$	$N_x(t)$	Δ_x	$\hat{f}_x(t)$
(0,3]	32	1	3	0,010
(3,18]	31	1	15	0,002
(18,29]	30	1	11	0,003
(29,54]	29	1	25	0,001
(54,60]	28	1	6	0,005
(60,84]	27	1	24	0,001
(84,110]	26	1	26	0,001
(110,112]	25	1	2	0,016
(112,116]	24	1	4	0,008
(116,123]	23	1	7	0,004
(123,134]	22	1	11	0,003
(134,145]	21	1	11	0,003
(145,151]	20	2	6	0,010
(151,158]	18	1	7	0,004
(158,173]	17	1	15	0,002
⋮				

Função de sobrevivência

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 $Pr(T > 365)$?

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

$$S(t) = Pr(T > t)$$

Função de sobrevivência

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.

$$F(t) = Pr(T \leq t)$$

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

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

Estimando a sobrevivência – 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.

Cálculo da Função de sobrevivência – Aids

Intervalo	$R_x(t)$ (risco)	$N_x(t)$ (eventos)	$\hat{f}_x(t)$ (densidade)	$\hat{S}_x(t)$ (sobrevivência)
(0,3]	32	1	0,010	1,000
(3,18]	31	1	0,002	0,969
(18,29]	30	1	0,003	0,938
(29,54]	29	1	0,001	0,906
(54,60]	28	1	0,005	0,875
(60,84]	27	1	0,001	0,844
(84,110]	26	1	0,001	0,813
(110,112]	25	1	0,016	0,781
(112,116]	24	1	0,008	0,750
(116,123]	23	1	0,004	0,719
(123,134]	22	1	0,003	0,688
(134,145]	21	1	0,003	0,656
(145,151]	20	2	0,010	0,625
(151,158]	18	1	0,004	0,563
(158,173]	17	1	0,002	0,531
.				

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

$$\lambda(t) = \lim_{\epsilon \to 0} \frac{Pr((t < T < t+\epsilon) | T \geq t)}{\epsilon}$$

- $\lambda(t)$ 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 sobrevivência

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

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

Sobrevivência e risco são inversamente proporcionais: quando o risco aumenta, a probabilidade de sobrevivência 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 $\lambda(t)$.
- Comum nas tábuas de vida – demografia.

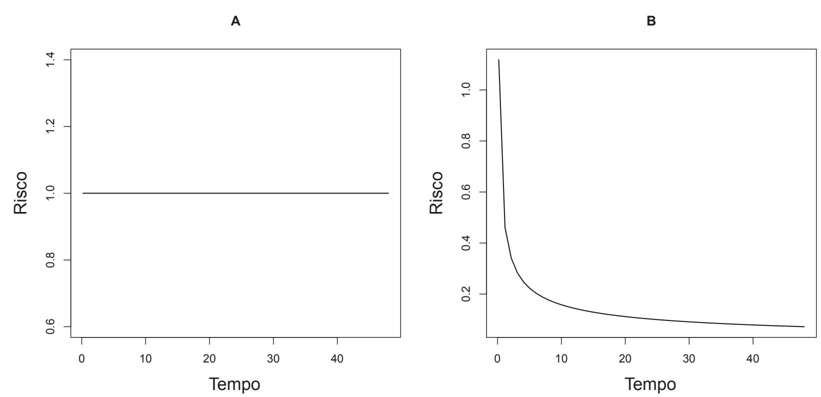
Planilha tempo.ods

Estimando risco

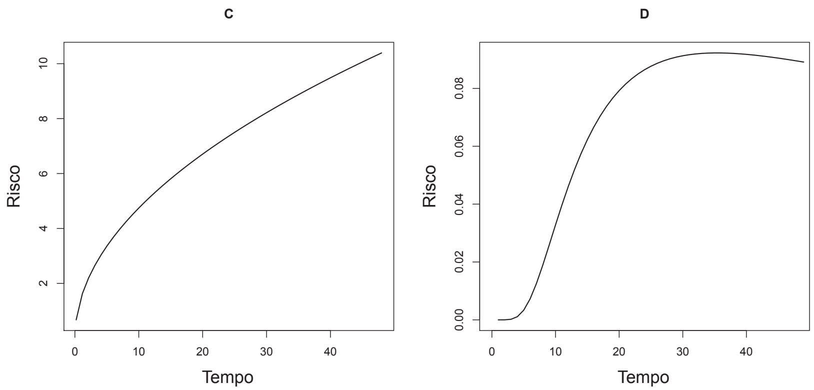
Intervalo	$R_x(t)$	$N_x(t)$	Δ_x	$\hat{f}_x(t)$	$\hat{S}_x(t)$	$\hat{\lambda}_x(t)$
(0,3]	32	1	3	0,010	1,000	$\frac{1}{32\times3} = 0,010$
(3,18]	31	1	15	0,002	0,969	$\frac{1}{31\times15} = 0,002$
(18,29]	30	1	11	0,003	0,938	$\frac{1}{30\times11} = 0,003$
(29,54]	29	1	25	0,001	0,906	$\frac{1}{29\times25} = 0,001$
(54,60]	28	1	6	0,005	0,875	$\frac{1}{28\times6} = 0,006$
(60,84]	27	1	24	0,001	0,844	$\frac{1}{27\times24} = 0,002$
(84,110]	26	1	26	0,001	0,813	$\frac{1}{26\times26} = 0,001$
(110,112]	25	1	2	0,016	0,781	$\frac{1}{25\times2} = 0,020$
(112,116]	24	1	4	0,008	0,750	$\frac{1}{24\times4} = 0,010$
(116,123]	23	1	7	0,004	0,719	$\frac{1}{23\times7} = 0,006$
(123,134]	22	1	11	0,003	0,688	$\frac{1}{22\times11} = 0,004$
(134,145]	21	1	11	0,003	0,656	$\frac{1}{21\times11} = 0,004$
(145,151]	20	2	6	0,010	0,625	$\frac{2}{20\times6} = 0,017$
(151,158]	18	1	7	0,004	0,563	$\frac{1}{18\times7} = 0,008$
(158,173]	17	1	15	0,002	0,531	$\frac{1}{17\times15} = 0,004$

.

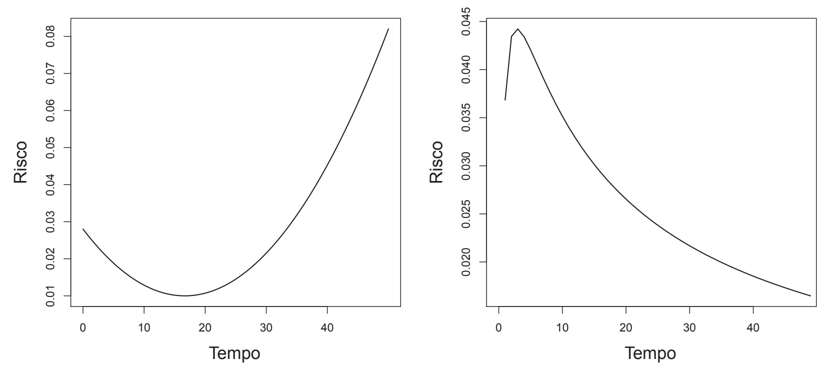
Comportamento da Função de Risco



Comportamento da Função de Risco



Comportamento da Função de Risco



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

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

- 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 tempo.ods

Relação entre as Funções

- Qual a probabilidade de sobreviver por mais de t unidades de tempo?
- Qual o risco de sofrer o evento no tempo t se sabemos que o paciente sobreviveu até aquele momento?
- Qual o risco de sofrer o evento até um determinado tempo t?

Relação entre as funções básicas de sobrevivência

$$S(t) = 1 - F(t)$$

$$S(t) = exp(-\Lambda(t))$$

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

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

$$\lambda(t) = \frac{f(t)}{1-F(t)}$$

$$\Lambda(t) = -\ln(S(t))$$

Função de verossimilhança

- A função de verossimilhança avalia o quanto os dados apoiam, concordam ou suportam cada valor possível do parâmetro a ser estimado.
- Exemplo: amostra para estimar prevalência de hipertensão.
  - 10% dos participantes são hipertensos
  - verossimilhança da proporção de hipertensos na população ser 90% é baixíssima
  - quanto mais próximo de 10%, maior a verossimilhança $\Rightarrow$ Máxima Verossimilhança
- Pressupostos do método de Máxima Verossimilhança:
  - Observações independentes
  - Tempos de sobrevivência independentes
  - Censuras independentes

Função de verossimilhança na sobrevivência

- Sem censura: $L \propto \prod_i f(t_i)$
- Com censura à direita: $L \propto \prod_{i \in O} f(t_i) \prod_{i \in D} S(t_i)$
- Com censura à esquerda: $L \propto \prod_{i \in O} f(t_i) \prod_{i \in D} S(t_i) \prod_{i \in E} [1 - S(t_i)]$
- Com censura intervalar: $L \propto \prod_{i \in O} f(t_i) \prod_{i \in D} S(t_i) \prod_{i \in E} [1 - S(t_i)] \prod_{i \in I} [S(t_i^-) - S(t_i^+)]$
- Com truncamento: probabilidade condicional – do indivíduo ser incluído no estudo.

Outline

Estimação Não-Paramétrica

- Introdução
- Kaplan-Meier
- Nelson-Aalen
- Intervalos de confiança
- Tempo Mediano de sobrevivência
- Kaplan-Meier com estratificação
- Teste de Log-Rank
- Teste de Peto

Incorporando a censura

Sem suposições sobre a distribuição do tempo

Introdução

- Duas formas não paramétricas de estimação das funções de sobrevivência:
  - Kaplan-Meier – $S(t)$
  - Nelson-Aalen – $\Lambda(t)$
- COM 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.
- Estimador produto (ou estimador limite produto)

Kaplan-Meier

- Sejam $t_1 < t_2 < \cdots < t_m$ os m tempos onde ocorreram os eventos;
- $R(t_j)$ é o total de pessoas a risco no tempo t_j.
- $\Delta N(t_j)$ é o número de eventos ocorridos precisamente em t_j.
- Para os m tempos t_j em que ocorre um evento, a probabilidade de sobrevivência será estimada pelo número dos que sobreviveram até aquele tempo $(R(t_j) - \Delta N(t_j))$ sobre os que estavam em risco naquele tempo $(R(t_j))$.
- Como os eventos são independentes $S(t)$ é o produto das probabilidades de sobrevivência a cada tempo $t_j \le t$.

Kaplan-Meier

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

ou na forma de produtório:

$$\hat{S}_{KM}(t_j) = \prod_{j:t_j \le t} \frac{R(t_j) - \Delta N(t_j)}{R(t_j)}$$

Kaplan-Meier – o dado

- 21 pacientes com aids (n=21)
- 15 óbitos (m=15)
- 6 censuras (indicada pelo +)

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

Kaplan-Meier – gráfico

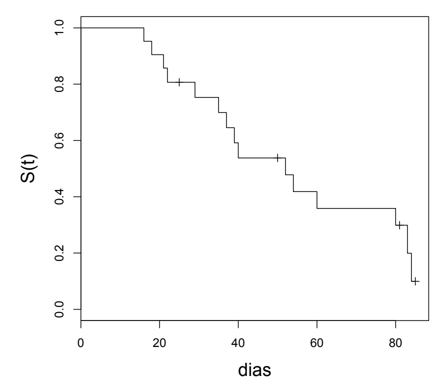


Figura : Função de sobrevivência dos pacientes com Aids. Os símbolos + localizam as censuras. É uma função em escada, que salta em cada tempo onde ocorre evento.

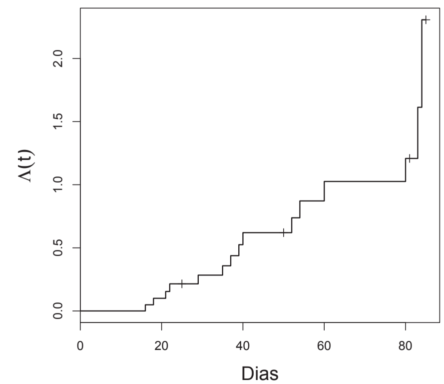
Da sobrevida ao risco

Função de Risco Acumulado

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

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

Gráfico da Função de Risco Acumulado



Estimador de –Aalen

Função de Risco Acumulado

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

Indicado para amostras muito pequenas
Equivalente ao K-M pra amostras grandes

planilha exerciciokm.ods

Estimativas de K-M e N-A

t_j	$R(t)$	$\Delta N(t)$	$S_{KM}(t)$	$\widehat{\Lambda}_{km}(t)$	$\widehat{\Lambda}_{na}(t)$
16	21	1	0,9524	0,0488	$\left(\frac{1}{21}\right) = 0,0476$
18	20	1	0,9048	0,1001	$\left(0,0476 + \frac{1}{20}\right) = 0,0976$
21	19	1	0,8571	0,1542	$\left(0,0976 + \frac{1}{19}\right) = 0,1503$
22	17	1	0,8067	0,2148	$\left(0,1502 + \frac{1}{17}\right) = 0,2091$
29	15	1	0,7529	0,2838	$\left(0,2091 + \frac{1}{15}\right) = 0,2757$
35	14	1	0,6992	0,3578	$\left(0,2757 + \frac{1}{14}\right) = 0,3472$
37	13	1	0,6454	0,4379	$\left(0,3472 + \frac{1}{13}\right) = 0,4241$
39	12	1	0,5916	0,5249	$\left(0,4241 + \frac{1}{12}\right) = 0,5074$
40	11	1	0,5378	0,6203	$\left(0,5074 + \frac{1}{11}\right) = 0,5983$
52	9	1	0,4781	0,7379	$\left(0,5983 + \frac{1}{9}\right) = 0,7094$
54	8	1	0,4183	0,8716	$\left(0,7094 + \frac{1}{8}\right) = 0,8344$
60	7	1	0,3585	1,0258	$\left(0,8344 + \frac{1}{7}\right) = 0,9773$
80	6	1	0,2988	1,2080	$\left(0,9773 + \frac{1}{6}\right) = 1,1440$
83	3	1	0,1992	1,6134	$\left(1,1439 + \frac{1}{3}\right) = 1,4773$
84	2	1	0,0996	2,3066	$\left(1,4773 + \frac{1}{2}\right) = 1,9773$

Intervalos de confiança

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

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

Intervalos de confiança

Assumindo erro α, o intervalo fica assim:

Limite inferior

$$\hat{S}_{KM}(t) - z_{\alpha/2}\sqrt{Var(\hat{S}_{KM}(t))}$$

Limite superior

$$\hat{S}_{KM}(t) + z_{\alpha/2}\sqrt{Var(\hat{S}_{KM}(t))}$$

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 ou igual a 1.

no R

- Criando o objeto sobrevida (tempo, status) (somente $t < 90$)

```
> tempo <- c(16, 18, 21, 21, 22, 25, 29, 35, 37,
  39, 40, 50, 52, 54, 60, 80, 80, 81, 83, 84, 85)
> status <- c(1,1,0,1,1,0,1,1,1,1,1,0,1,1,1,0,1,0,1,1,0)
# variável status=1 indica evento, 0 censura
> Surv(tempo,status)
```

```
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) ~ 1)
> summary(KM)
> plot(KM)
```

- Nelson-Aalen

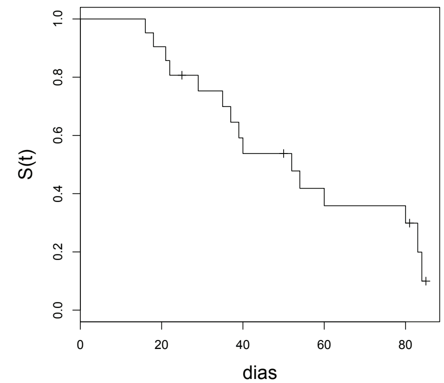
```
> sob.NA <- survfit(coxph(Surv(tempo,status)~1))
> sob.NA
> summary(sob.NA)
```

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.



Tempo Mediano de Sobrevivência

- Medida sumária mais comum
- Menor tempo para o qual metade dos indivíduos sofre o evento
- Com censura é tempo no qual o valor estimado da sobrevivência é $\leq 50\%$
- Sem censura é exatamente 50%

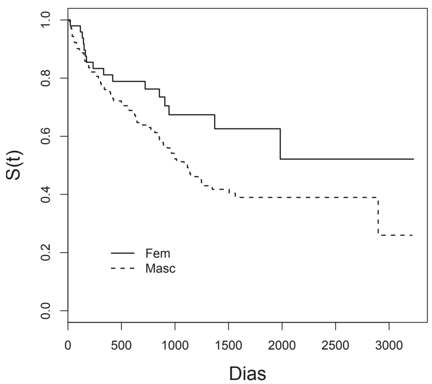
$$t_{med} = min(t_j | \hat{S}(t_j) \leq 0,5) \tag{1}$$

Kaplan-Meier com estratificação

- Descrever a sobrevivência segundo características: sexo, faixa etária, etc.
- A sobrevivência é estimada separadamente para cada estrato, utilizando Kaplan-Meier.
- no R

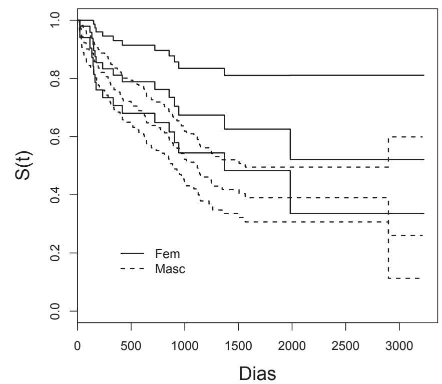
```
> ipec <- read.table("ipec.csv",header=T,sep=";")
> survaids <- survfit(Surv(tempo,status)~ sexo, data = ipec)
> survaids
Call: survfit(formula = resp ~ sexo, data = ipec)

         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.

Gráfico sobrevida estratificada



Com intervalo de confiança de 95%.

Testes

- Log-rank ou Mantel Haenszel
- Peto

Hipótese nula: não há diferença entre estratos

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

Teste Log-rank

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 ($k = 2$):

$$\text{Log-rank} = \frac{(O_1 - E_1)^2}{Var(O_1 - E_1)}$$

O_1 = total de eventos observados no estrato 1
E_1 = 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 $k = 2$):

$$Var(O_1 - E_1) = \sum_t \frac{R_1(t)R_2(t)\Delta N(t)[R(t) - \Delta N(t)]}{R(t)^2[R(t) - 1]}$$

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.

$$\text{Peto} = \frac{(O_1 - E_1)^2}{Var(O_1 - E_1)}$$

sendo que

$$O_1 - E_1 = \sum_{t_j} S(t_j)(O_1(t_j) - E_1(t_j))$$

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

no R – Log-rank

```
> 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 – Peto

```
> 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 *
```

Resumo

Neste capítulo foram apresentados:

- Método não paramétrico para estimação da função de sobrevivência – Kaplan-Meier;
- Método não paramétrico para estimação da função risco acumulado – Nelson-Aalen;
- Intervalos de confiança para as duas funções;
- Cálculo e interpretação do tempo de sobrevivência mediano;
- Intervalos de confiança para o tempo de sobrevivência mediano;
- Testes para comparação das curvas de sobrevivência entre diferentes estratos – log-rank e Peto.

Outline

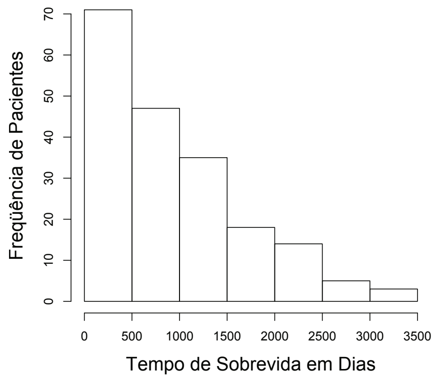
Modelagem Paramétrica

- Introdução
- Distribuições estatísticas para modelar as funções de sobrevivência
- Estimação
- Regressão paramétrica
- Seleção dos modelos
- Avaliação de ajuste do modelo

Introdução

- Os estimadores de Kaplan-Meier e Nelson-Aalen para as funções $S(t)$ e $\lambda(t)$ são obtidos a partir dos dados, supondo que a cada momento do tempo existe um processo diferente gerando as observações.
- Como cada intervalo de tempo é estimado de forma independente, a estimação não-paramétrica possui tantos parâmetros quantos intervalos de tempo.
- Na abordagem paramétrica o tempo segue uma distribuição de probabilidade conhecida.
- Para estimar o efeito de covariáveis –> modelagem

Distribuição do tempo da coorte de Aids



Tempo de vida acelerado

O tempo T obedece à:

$$ln(T) = \mu + \sigma W$$

sendo:
W –> distribuição de probabilidade que ajusta T
μ –> parâmetros de média $ln(T)$, também chamado locação
σ –> parâmetros de dispersão de $ln(T)$, escala

Distribuições

- Distribuições estatísticas para modelar as funções de sobrevivência:
  - Exponencial
  - Weibull
  - Log-normal
  - ...
- Funções assimétricas, contínuas, positivas

Distribuição Exponencial

Se a variável T possui uma distribuição exponencial,

- Densidade de probabilidade:

$$f(t) = \alpha \exp(-\alpha t), \qquad \alpha > 0$$

- Função de sobrevivência:

$$S(t) = \exp(-\alpha t)$$

- A função risco é constante para todo o tempo de observação t, ou seja:

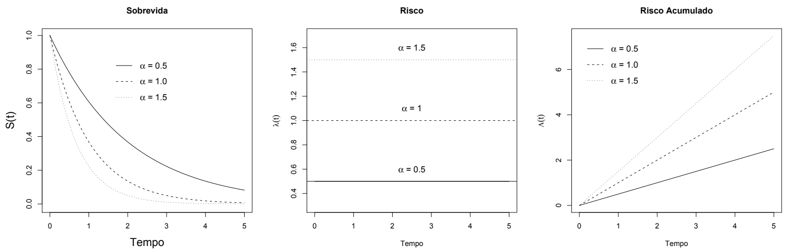
$$\lambda(t) = \frac{f(t)}{S(t)} = \alpha = \text{constante}$$

- A função de risco acumulado é uma função linear no tempo e é dada por:

$$\Lambda(t) = -\ln S(t) = \alpha t$$

Algumas exponenciais

Função de sobrevivência, de risco e de risco acumulado para a distribuição exponencial considerando diferentes valores de α



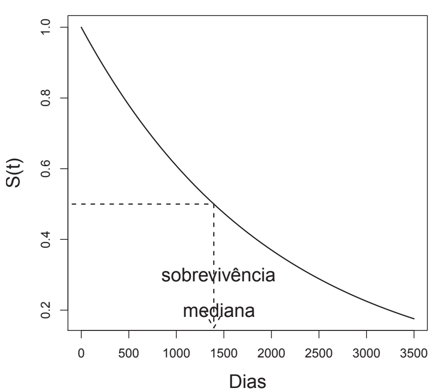
A distribuição exponencial é conhecida como distribuição exponencial padrão quando $\alpha = 1$.

Interpretando risco exponencial

- média: $E(T) = \frac{1}{\alpha}$
- variância: $var(T) = \frac{1}{\alpha^2}$
- $T_{mediano} = ln(2)/\alpha$
- quanto maior o risco, menor o tempo médio de sobrevivência e menor a variabilidade deste em torno da média
- como a distribuição do tempo de sobrevivência T é assimétrica, usa-se mais o tempo mediano
- o modelo exponencial é matematicamente simples, mas a suposição de risco constante no tempo (sem memória) é pouco plausível
- aplicável quando o tempo é curto para supor risco constante (por ex., o risco de acidentes domésticos de crianças entre 2 e 5 anos pode ser considerado constante neste intervalo)

Exemplo – aids

Tempo médio de sobrevivência $= \frac{1}{\alpha} = \frac{1}{0{,}000497} = 2012$ dias; Tempo mediano de sobrevivência $= \frac{\ln(2)}{\alpha} = \frac{\ln(2)}{0{,}000497} = 1394$ dias.



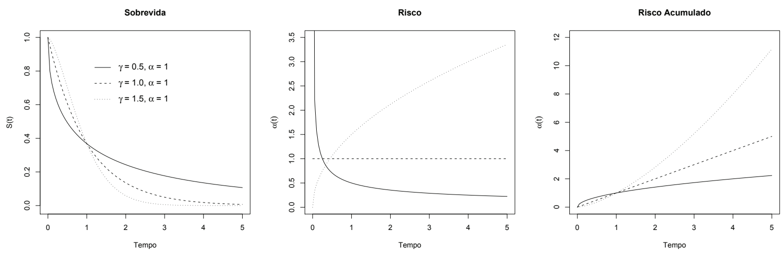
Distribuição Weibull

- permite variação do risco no tempo
- é uma generalização da distribuição exponencial:
- densidade –> $f(t) = \gamma\alpha^{\gamma} t^{\gamma-1} \exp(-(\alpha t)^{\gamma})$
- sobrevivência –> $S(t) = \exp(-(\alpha t)^{\gamma})$
- γ determina a forma da função de risco –> parãmetro de forma:

  $\gamma < 1$ função de risco decrescente
  $\gamma > 1$ função de risco crescente
  $\gamma = 1$ função de risco constante (equivalente ao modelo exponencial)
- a função de risco acumulado é: $\Lambda(t) = -\ln S(t) = (\alpha t)^{\gamma-1}$
- o parâmetro α determina a escala da distribuição
- Tempo mediano: $S(t) = 0,5 = exp(-(\alpha t)^{\gamma})$

Algumas Weibull

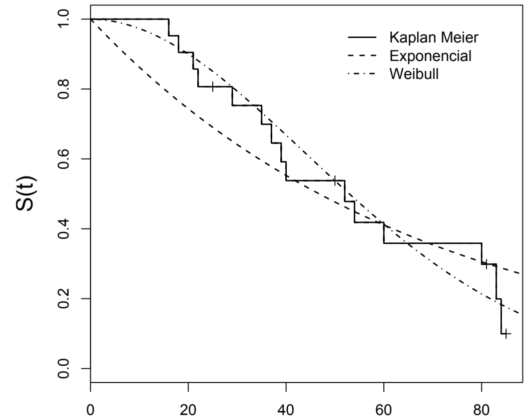
Função de sobrevivência, de risco e de risco acumulado com parâmetro escala $\alpha = 1$ e diferentes valores do parâmetro de forma γ



Exemplos: tumores, tempo de incubação do HIV

Comparando Não-paramétrico com paramétricos – Aids

$$N = 21$$



Modelo de Regressão Paramétrica

- Nos modelos paramétricos, a inclusão de covariáveis segue a forma utilizada em modelos lineares generalizados, podendo ser tanto contínuas – pressão sanguínea, idade, dosagens bioquímicas – como categóricas – gênero, tratamento, comportamentos.
- O objetivo de um modelo de regressão é o de estimar o efeito de covariáveis (ou variáveis independentes ou preditores), $x_1, x_2, \cdots, x_p$, sobre uma variável resposta (ou variável dependente), Y.
- Supondo uma distribuição da família exponencial para a variável resposta teremos um modelo linear generalizado.
- Ainda que a distribuição exponencial e a Weibull sejam parte desta família, os modelos de regressão paramétricos para tempo de sobrevivência não são parte dos GLM por causa de dados censurados.

Modelo de Regressão Paramétrica

- T –> tempo até o evento ou censura, variável resposta
- $\boldsymbol{x}$ –> vetor de covariáveis
- Função de risco: $\lambda(t|\boldsymbol{x}) = \lambda_0(t)g(\boldsymbol{x}\boldsymbol{\beta})$:
  - $\boldsymbol{\beta}$ –> coeficientes estimados
  - $g(.)$ –> função de ligação, positiva e contínua (exponencial, Weibull)
- Razão de riscos λ/λ_0 é função das covariáveis e não depende do tempo –> riscos proporcionais

Modelo de Regressão Paramétrica

- Assumimos que o parâmetro da distribuição depende de covariáveis segundo uma função
- Exemplo: $\alpha(\boldsymbol{x}) = \exp(\boldsymbol{x\beta})$
- Modelo Exponencial:

$$S(t|\boldsymbol{x}) = \exp(-\alpha(\boldsymbol{x})t) = \exp(-\exp(\boldsymbol{x\beta})t)$$

$$\lambda(t|\boldsymbol{x}) = \alpha(\boldsymbol{x}) = \exp(\boldsymbol{x\beta})$$

- Modelo Weibull:

$$S(t) = \exp(-(\alpha(\boldsymbol{x})t)^{\gamma}) = \exp(-(\exp(\boldsymbol{x\beta})t)^{\gamma})$$

$$\lambda(t) = \gamma\alpha(\boldsymbol{x})^{\gamma}t^{\gamma-1} = \gamma(\exp(\boldsymbol{x\beta}))^{\gamma}t^{\gamma-1}$$

Exemplo

Assumindo que o risco de morrer é constante ao longo do tempo, pode-se estimar o efeito da $idade$ na sobrevivência e no risco de 6.805 pacientes em diálise acompanhados durante um ano (1.603 morreram) através do modelo exponencial:

$$\lambda(t|idade) = \exp(\beta_0 + idade\beta_1)$$

Os parâmetros estimados são: $\beta_0 = -6,135$ e $\beta_1 = 0,037$, ou seja, para cada ano a mais de vida o risco aumenta de $exp(0,037) = 1,0377$.
Pode-se comparar o risco constante de morte no tempo, entre dois indivíduos submetidos à diálise, um com 30 anos e outro com 70, substituindo as estimativas dos parâmetros $\boldsymbol{\beta}$:

$$\frac{\lambda(t|x_1 = 70)}{\lambda(t|x_1 = 30)} = \frac{\exp(\beta_0 + 70\beta_1)}{\exp(\beta_0 + 30\beta_1)} = \frac{0,000713}{0,000162} = 4,39$$

Modelo Weibull

O tempo T segue uma distribuição de Weibull e o parâmetro de escala α depende das covariáveis.
Neste caso são estimados os parâmetros:

- β_0 – cuja exponencial representa o risco médio, quando todas as covariáveis são zero;
- β_1 – cuja exponencial é a parcela de variação no tempo de sobrevivência devida à idade do paciente;
- γ – a forma da função de risco ao longo do tempo.

Seleção do modelo

- Razão de Verossimilhança: $RV = 2(l_{maior} - l_{menor})$
- Teste de Wald – testa a hipótese nula H_0 de que o parâmetro β de cada covariável separadamente é igual a zero.

- Comparar um modelo com distribuição exponencial e outro com distribuição Weibull equivale a testar a hipótese nula de que o parâmetro de forma, γ, da distribuição Weibull é igual a 1. (compara-se o logaritmo da função de verossimilhança do modelo nulo exponencial com o modelo nulo Weibull)

Qualidade do ajuste do modelo

- *Deviance* –> $D = 2(l_{saturado} - l_{modelo})$
- D –> assintoticamente uma χ^2, com $n - p - 1$ graus de liberdade

Exemplo – exponencial

```
> survreg(formula=Surv(tempo,status)~1, data=dialise,
        dist='exponential')

Call:
survreg(formula=Surv(tempo, status)~1, data=dialise,
        dist="exponential")
Coefficients:
(Intercept)
   4.096059
Scale fixed at 1
Loglik(model)= -8169   Loglik(intercept only)= -8169
n= 6805
```


Exemplo – Weibull

```
> survreg(formula=Surv(tempo,status)~1, data=dialise,
        dist='weibull')

Call:
survreg(formula=Surv(tempo, status)~1, data=dialise,
        dist="weibull")
Coefficients:
(Intercept)
   4.388833
Scale= 1.257539
Loglik(model)= -8104.2   Loglik(intercept only)= -8104.2
n= 6805
```


Exemplo

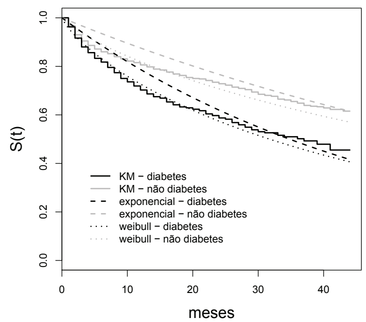
Comparando:
$D = 2(L_{weibull} - L_{exponencial}) = 2(-8104,2 - (-8169)) = 129,6$

Como D segue uma distribuição χ^2 com um grau de liberdade, $p = 0$, ou seja, rejeitamos a hipótese nula de que $\gamma = 1$.

Isto é, o modelo de Weibull, com $\gamma = 0,795$ é melhor do que o modelo exponencial.

Análise Gráfica

Comparar a curva do Kaplan-Meier com as estimadas parametricamente. Quanto mais próximo o modelo paramétrico estiver da curva do Kaplan-Meier, melhor.



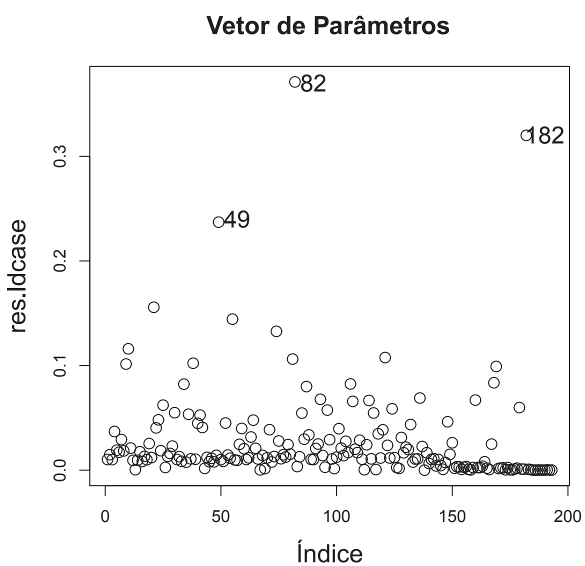
As três curvas em cinza referem-se aos paciente sem diabetes e as três curvas pretas aos pacientes com diabetes.

Análise de Resíduos

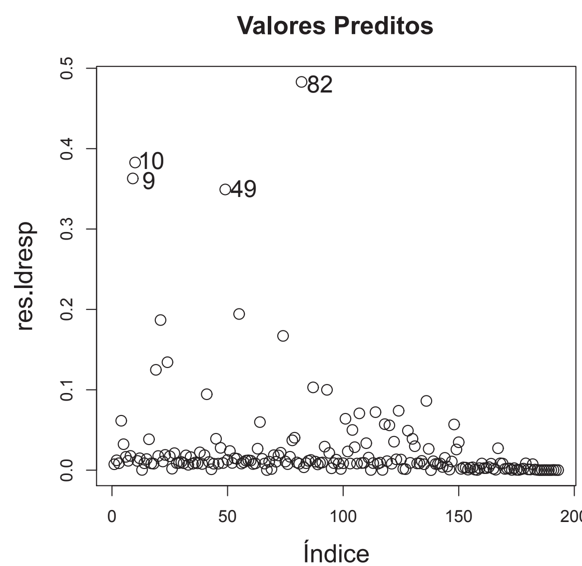
São três tipos de resíduos específicos dos modelos paramétricos (além dos que serão apresentados par o Modelo de Cox), que avaliam efeito de observações sobre:

- conjunto de parâmetros da regressão –> *ldcase*
- valores preditos (em unidades de DP) –> *ldresp*
- forma –> *ldshape*

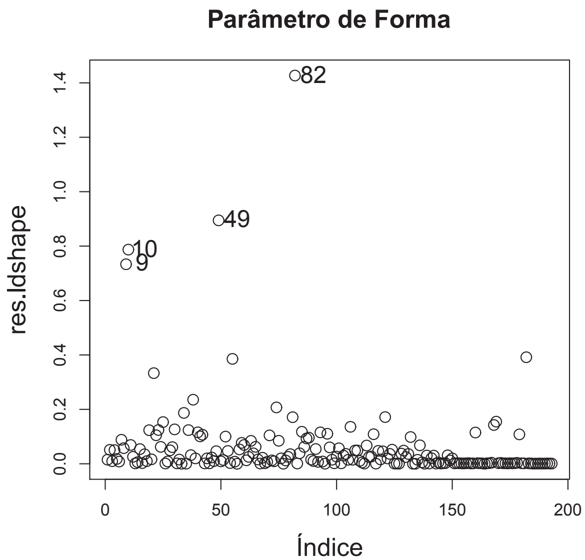
Análise de Resíduos – Vetor de Parâmetros



Análise de Resíduos – Valores Preditos



Análise de Resíduos – Parâmetro de Forma



Análise de Resíduos – Casos

```
> hiv[c(9,10,49,82,182),c(4,5,6,8,13)]
    tempo status sexo idade tratam
9    1563      1    M    44      0
10   1247      1    M    23      0
49   1344      0    M    30      0
82   1272      0    M    22      0
182    16      1    M    42      3
```


Reajustando o modelo

```
Call:
survreg(formula = Surv(tempo, status) ~ idade + sexo + tratam,
    data = hiv, dist = "weibull")
               Value Std. Error      z        p
(Intercept)  6.06842     0.5674 10.695 1.07e-26
idade        0.00951     0.0130  0.731 4.65e-01
sexoM       -0.23627     0.3277 -0.721 4.71e-01
tratam       1.48608     0.2273  6.538 6.25e-11
Log(scale)   0.14185     0.0862  1.647 9.97e-02

Scale= 1.15

Weibull distribution
Loglik(model)= -742   Loglik(intercept only)= -770.3
        Chisq= 56.64 on 3 degrees of freedom, p= 3.1e-12
Number of Newton-Raphson Iterations: 5
n= 193
```


Análise de Resíduos – Retirando Casos

```
Call:
survreg(formula = Surv(tempo, status) ~ idade + sexo + tratam,
    data = hiv, subset = -82, dist = "weibull")
              Value Std. Error      z        p
(Intercept)  5.7996     0.5760 10.069 7.60e-24
idade        0.0151     0.0133  1.137 2.55e-01
sexoM       -0.2603     0.3231 -0.806 4.20e-01
tratam       1.5490     0.2266  6.836 8.16e-12
Log(scale)   0.1281     0.0857  1.496 1.35e-01

Scale= 1.14

Weibull distribution
Loglik(model)= -739.2   Loglik(intercept only)= -769.7
        Chisq= 61.03 on 3 degrees of freedom, p= 3.5e-13
Number of Newton-Raphson Iterations: 5
n= 192
```

Outline

Modelo de riscos proporcionais de Cox (semi-paramétrico)

- Introdução
- Riscos proporcionais
- Modelo de Cox
- Cox estratificado
- Seleção dos modelos
- Qualidade do ajuste
- Tempos de vida empatados

Introdução

- O interesse é modelar o efeito de covariáveis sobre o tempo de sobrevivência (hazard)
- O modelo de regressão mais amplamente utilizado para dados de sobrevivência
- Ou seja, as covariáveis têm um efeito multiplicativo na função de risco
- Usando processo de contagem modela-se situações mais complexas –> Cox estendido (curso avançado).

Riscos Proporcionais

- Ajusta a função de risco $\lambda(t)$, considerando um risco basal $\lambda_0(t)$
- Inclui o vetor de covariáveis $\boldsymbol{x}$, de forma que:

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

A razão entre os riscos de ocorrência do evento de dois indivíduos i e j, com covariáveis $\boldsymbol{x}_k = (x_{k1}, x_{k2}, \cdots, x_{kp})$ e $\boldsymbol{x}_j = (x_{l1}, x_{l2}, \cdots, x_{lp})$ é:

$$\frac{\lambda_k(t|\boldsymbol{x}_k)}{\lambda_l(t|\boldsymbol{x}_l)} = \frac{\exp(\boldsymbol{x}_k\boldsymbol{\beta})}{\exp(\boldsymbol{x}_l\boldsymbol{\beta})}$$

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

Modelo de Riscos Proporcionais

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

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

- O risco acumulado basal é $\hat{\Lambda}_0(t) = \sum_{i:\, t_i \le t} \frac{\Delta N_i(t)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j \hat{\boldsymbol{\beta}})}$
- A sobrevivência basal é dada por $\hat{S}_0(t) = \exp[-\hat{\Lambda}_0(t)]$

Modelo de Cox

- Partindo do pressuposto de proporcionalidade, é possível estimar os efeitos das covariáveis sem qualquer suposição a respeito da distribuição do tempo de sobrevivência, 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.

Modelo de Cox - Pressupostos

- 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.
- Como o tempo é contínuo, não há empates na ocorrência do evento.

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 sobrevivência 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 sobrevivência.

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
- A função de Verossimilhança NÃO depende do risco basal

Verossimilhança Parcial

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

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

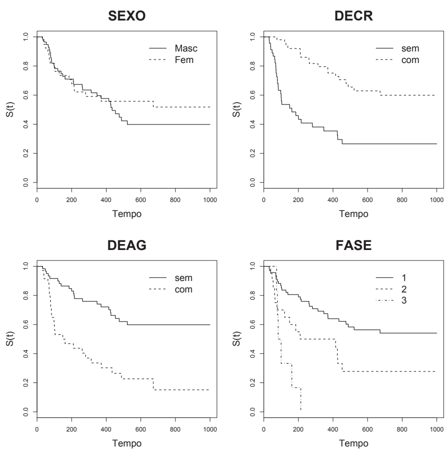
- $\Delta N_i(t)$ = diferença entre a contagem de eventos até o instante t e a contagem no momento imediatamente anterior a t.
- Numerador depende apenas da informação dos indivíduos que experimentam o evento
- Denominador utiliza informações a respeito de todos os indivíduos que ainda não experimentaram 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

```
> tmo <- read.table("tmoclas.dat", header=T, sep=",")
> tmo$sexo<-factor(tmo$sexo)
> m1 <- coxph(Surv(os,status)~idade+sexo,data=tmo,x=TRUE)
> summary(m1)

[...]
          coef exp(coef) se(coef)      z Pr(>|z|)
idade -0.02167   0.97857  0.01399 -1.548    0.122
sexo2 -0.37649   0.68626  0.32120 -1.172    0.241

      exp(coef) exp(-coef) lower .95 upper .95
idade    0.9786      1.022    0.9521     1.006
sexo2    0.6863      1.457    0.3657     1.288

Rsquare= 0.03   (max possible= 0.986 )
Likelihood ratio test= 2.92  on 2 df,   p=0.2320
Wald test            = 2.85  on 2 df,   p=0.2408
Score (logrank) test = 2.85  on 2 df,   p=0.2406
```


Cox estratificado

- O risco basal – $\lambda_0(t)$ – não é o mesmo para todos os indivíduos do estudo.
- $\lambda_{0_A}(t) \neq \lambda_{0_B}(t) \neq \lambda_{0_C}(t)$, definindo diferentes estratos
- É usado quando alguma covariável não atende à proporcionalidade
- A variável para a qual se estratifica NÃO terá o efeito estimado.

Selecionando modelos

- Teste de Wald

$$H_0 : \beta_j = 0$$
$$z = \widehat{\beta}_j / ep(\widehat{\beta}_j)$$

- Teste da Razão de Verossimilhança

$$\begin{aligned} H_0 &: Mod_{maior} = Mod_{menor} \\ RV &= 2(l_{maior} - l_{menor}) \\ RV &\sim \chi^2_{l-k} \end{aligned}$$

Selecionando Modelos

- Para modelos aninhados!
- Não se pode comparar modelos estratificados com não estratificados.
- A RV 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.
- Se existirem valores ausentes, modelos perdem a comparabilidade –> para retirar casos com variáveis com dados missing usar a função *complete.cases()*

Comparando modelos com função desvio

```
> anova(mod1,mod2,mod3,mod4)

Analysis of Deviance Table
 Cox model: response is  Surv(os, status)
 Model 1: ~ idade + sexo
 Model 2: ~ idade + sexo + fase
 Model 3: ~ idade + sexo + fase + deag
 Model 4: ~ idade + sexo + fase + deag + decr
   loglik  Chisq Df P(>|Chi|)
1 -201.94
2 -194.70 14.486  2 0.0007152
3 -188.15 13.109  1 0.0002939
4 -183.07 10.152  1 0.0014413
```


Qualidade do Ajuste

- O modelo se ajusta bem aos dados?
- Qual o poder explicativo de um modelo?
- Existem poucas estatísticas de ajuste:
  - Função Desvio (Deviance) $\propto R^2$
  - R^2
  - Probabilidade de concordância
  - Gráfico de Índice Prognóstico

Medida Global de Ajuste – R^2

- 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 (L(0)).

Medida Global de Ajuste – R^2

Modelo	l_{modelo}	R^2	% Variabilidade Explicada*
Nulo	-203,40	0,000	0,0%
Saturado	-1,39	0,986	100,0%
m1: *idade+sexo*	-201,940	0,030	3,0%
m2: m1+*fase*	-194,70	0,166	16,8%
m3: m2+*deag*	-188,15	0,272	27,6%
m4: m3+*decr*	-183,07	0,345	35,0%

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

Probabilidade de Concordância

- Medida global de ajuste quando o objetivo é obter um modelo preditivo
- Avalia o poder discriminatório e a acurácia preditiva do modelo
- Similar a interpretação da Área sob a curva (AUC) na curva ROC de um modelo logístico

Probabilidade de Concordância

Concordância(c)	Poder discriminatório
$0.3 < c < 0.4$	Baixo
$c = 0.5$	ao acaso
$0.6 \leq c < 0.7$	Comum
$0.7 \leq c < 0.8$	Muito bom
$0.8 \leq c < 0.9$	Excelente

No R

```
> summary(m4)

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

n=96, number of events=49

[...]

Concordance=0.768 (se=0.044)
Rsquare= 0.345   (max possible= 0.986 )
Likelihood ratio test= 40.96  on 6 df,   p=3.365e-07
Wald test            = 38.46  on 6 df,   p=9.113e-07
Score (logrank) test = 47.62  on 6 df,   p=1.405e-08
```

Índice Prognóstico – IP

Gráfico de sobrevivência 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 sobrevivência 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.

Índice Prognóstico – IP

- 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`)

$$
\begin{array}{lllll}
\beta_{idade} & \times 56 = & -0,005019 \times 56 = & -0,281064 \\
\beta_{sexo} & \times\ 0 = & -0,271984 \times\ 0 = & 0 \\
\beta_{fase2} & \times\ 1 = & 0,593973 \times\ 1 = & 0,593973 \\
\beta_{fase3} & \times\ 0 = & 0,938411 \times\ 0 = & 0 \\
\beta_{deag} & \times\ 1 = & 1,190381 \times\ 1 = & 1,190381 \\
\beta_{decr} & \times\ 0 = & -1,061750 \times\ 0 = & 0 \\
\hline
 & & \text{Total} \quad = & 1,50329
\end{array}
$$

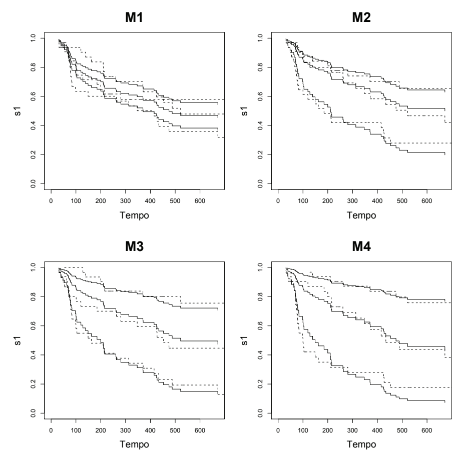
Índice Prognóstico – IP

- 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`)

$$
\begin{array}{lllll}
\beta_{idade} & \times 20 = & -0,005019 \times 20 = & -0,10038 \\
\beta_{sexo} & \times\ 1 = & -0,271984 \times\ 1 = & -0,271984 \\
\beta_{fase2} & \times\ 0 = & 0,593973 \times\ 0 = & 0 \\
\beta_{fase3} & \times\ 1 = & 0,938411 \times\ 1 = & 0,938411 \\
\beta_{deag} & \times\ 1 = & 1,190381 \times\ 1 = & 1,190381 \\
\beta_{decr} & \times\ 0 = & -1,061750 \times\ 0 = & 0 \\
\hline
 & & \text{Total} \quad = & 1,756428
\end{array}
$$

Índice Prognóstico – IP

Gráfico de sobrevivência estratificado por índice de prognóstico.



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

Tempos Empatados

- Tempo é contínuo
- Na prática –> DISCRETO
- Como a estimação só é feita quando ocorre evento, empate na censura não é problema
- Se censura e evento empatados –> considera-se que o evento ocorreu primeiro
- Empate de eventos –> estimação por Efron, Breslow, exata

Resumo

O modelo de Cox pode ser escrito como $\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp(\boldsymbol{x}\boldsymbol{\beta})$, sendo que:

- não se assume distribuição de probabilidade para o tempo T de sobrevivência;
- os coeficientes β são estimados por máxima verossimilhança parcial;
- modelos estratificados permitem a variação do risco basal $\lambda_0(t)$ entre os estratos;
- a avaliação da qualidade de ajuste dos modelos baseia-se na análise da função desvio, no R^2 e em análises gráficas (gráfico do índice prognóstico);
- modelos aninhados são selecionados através do teste da razão de verossimilhanças.

Outline

Análise de Resíduos

- Premissas e ajuste de modelo quanto à:
  - proporcionalidade do risco;
  - observações mal ajustadas pelo modelo (pontos aberrantes e influentes);
  - forma funcional das covariáveis.
- Tipos de resíduos:
  - Schoenfeld;
  - martingale;
  - *deviance*;
  - escore.

Introdução

- Proporcionalidade: a relação entre variável resposta e tempo é sempre a mesma, independente do momento de ocorrência do evento.
- Linearidade (log-linearidade, pois $\lambda(t) = \lambda_0(t)e^{\beta x}$): 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.
- O modelo estima efeito médio de covariáveis: pontos influentes (ou de alavanca) podem afetar a estimativa fortemente.

O resíduo obtido como a resposta observada menos a esperada não pode ser usado para os dados de sobrevida: a censura!!!

Riscos proporcionais – Schoenfeld

- O efeito de uma variável é sempre o mesmo durante todo o tempo observado?
- O resíduo de Schoenfeld é a diferença entre os valores observados de covariáveis de um indivíduo com tempo de ocorrência do evento t_j e os valores esperados em t_j dado o grupo de risco $R(t_j)$.
- Haverá tantos vetores de resíduos quanto covariáveis ajustadas no modelo, e que estes são definidos somente nos tempos de ocorrência do evento.

Riscos proporcionais – Schoenfeld

Para cada covariável x_i no tempo do evento t_i:

$$\begin{aligned} r_{ik} &= \delta_i(x_{ik} - a_{ik}) \\ a_{ik} &= \frac{\sum_{j \in R(t_j)} x_{ik} \exp(\boldsymbol{x}_i \widehat{\boldsymbol{\beta}})}{\sum_{j \in R(t_j)} \exp(\boldsymbol{x}_i \widehat{\boldsymbol{\beta}})} \end{aligned}$$

$R(t_j)$ é o conjunto de indivíduos em risco no tempo t_j
x_{ik} representa o valor da covariável k do indivíduo i pertencente ao grupo de risco

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:

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

- Se a premissa de proporcionalidade não é violada esperamos que o gráfico de $r_k^*(t_j)$ versus (t_j) (ou função de (t_j)) apresente uma reta com inclinação zero

Schoenfeld no *R*

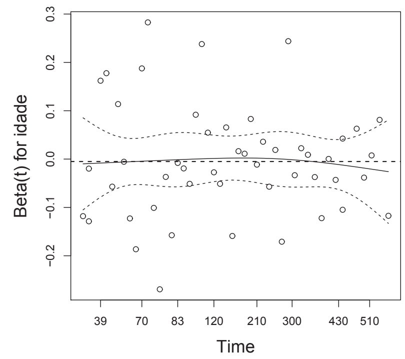
```
> residuo <- cox.zph(modelo)
> plot(residuo[1])
```

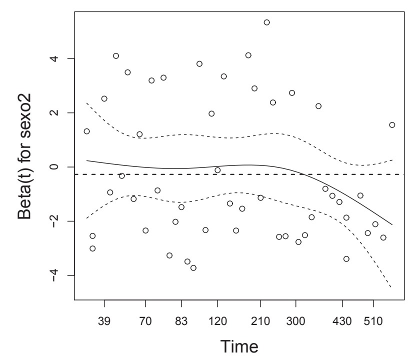
Atenção para a escala do tempo:

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

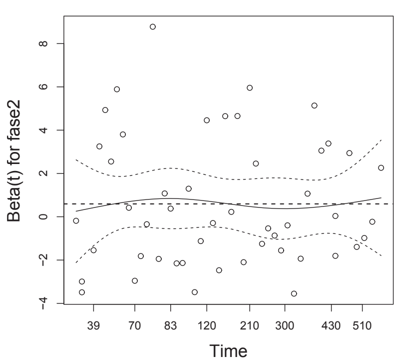
A linha curva é um `lowess`.

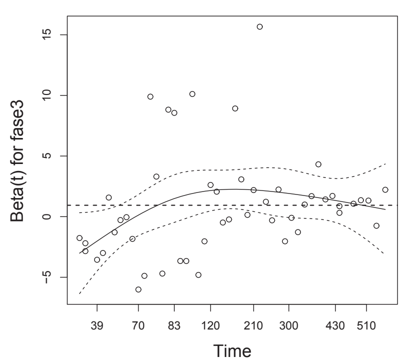
Gráficos de Schoenfeld - Exemplo TMO



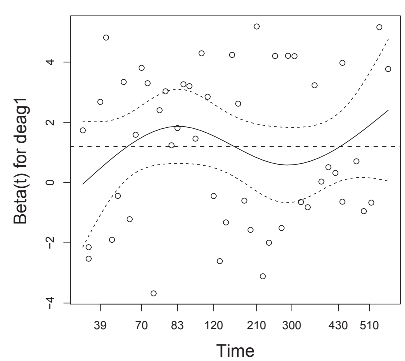


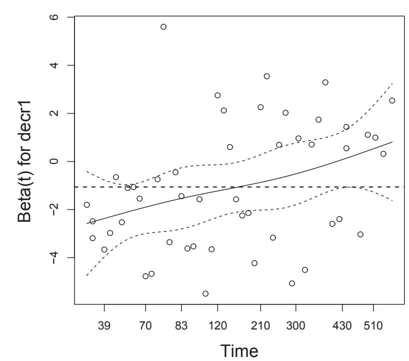
Gráficos de Schoenfeld - Exemplo TMO





Gráficos de Schoenfeld - Exemplo TMO





Schoenfeld - Exemplo TMO

- Testar H0 de que a correlação linear entre o resíduo de Schoenfeld e o tempo de sobrevida é nula
- Equivale a testar H0: inclinação igual a zero, ou H0: log do risco relativo é constante ao longo do tempo

```
m4.sch<-cox.zph(m4)
m4.sch

             rho    chisq      p
idade  -0.02226 2.92e-02 0.8644
sexo2  -0.18004 1.86e+00 0.1721
fase2  -0.00212 2.81e-04 0.9866
fase3   0.20766 2.91e+00 0.0881
deag1   0.05110 1.52e-01 0.6971
decr1   0.35133 7.22e+00 0.0072
GLOBAL       NA 1.35e+01 0.0362
```


Não proporcionalidade – soluções

- Frente a problema de proporcionalidade avaliar:
  - magnitude
  - pontos influentes
- 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.
Semelhante aos resíduos dos modelos de regressão linear em que:

- o valor esperado $= 0$ em torno do verdadeiro (e desconhecido) β
- não são simetricamente distribuídos em torno de zero, variando de $(-\infty, 1]$ e quando o tempo de sobrevivência é censurado o resíduo é negativo;
- 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

Martingale X resíduos de modelos lineares

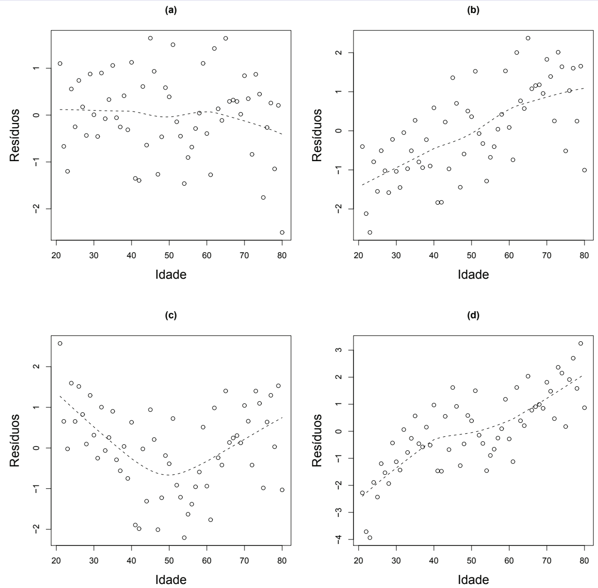
Diferentes dos resíduos da regressão linear porque:

- 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, nem log-normal, logo o qqplot não funciona;
- 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 – valores aberrantes
  - $M_i > 0 \Rightarrow$ observados $>$ esperados $\Rightarrow$ modelo subestima
  - $M_i < 0 \Rightarrow$ observados $<$ esperados $\Rightarrow$ modelo superestima
- M_i do modelo nulo (sem covariáveis) versus covariável com a superposição de uma curva de alisamento: para avaliar a forma funcional da covariável contínua a ser incluída no modelo
  - se linear – OK
  - se não linear – transformar variável, quebrar, suavizar

Gráficos Martingale – Resíduo modelo nulo X Idade



Martingale no *R*

A função para calcular o resíduo de Martingale é:

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

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

Ajuste forma funcional não linear – CURSO AVANÇADO

- 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

Pontos aberrantes: resíduos *deviance*

$$D_i = \text{sinal}(\widehat{M_i})\sqrt{-2(l_{i(modelo)} - l_{i(saturado)})}$$

- o sinal de $(\widehat{M_i})$ é o sinal do resíduo martingale.
- $l_{i(modelo)} - l_{i(saturado)}$: log da função de verossimilhança para cada observação i do modelo e do saturado.
- Resíduos são simetricamente distribuídos em torno do zero, portanto interpretação mais fácil.
- A soma não é necessariamente zero.
- Sem muita censura, os resíduos D_i parecerão uma amostra aleatória normal, e por isso são úteis na detecção de valores aberrantes.
- Três gráficos: resíduos deviance contra cada observação; contra preditos do modelo e gráfico quantil-quantil.

Pontos aberrantes: Martingale X *deviance*

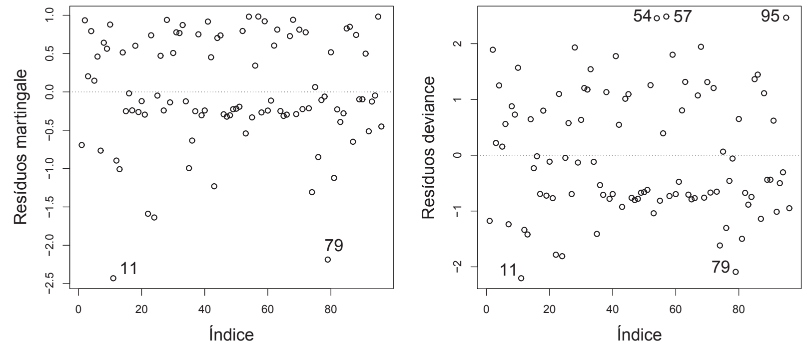
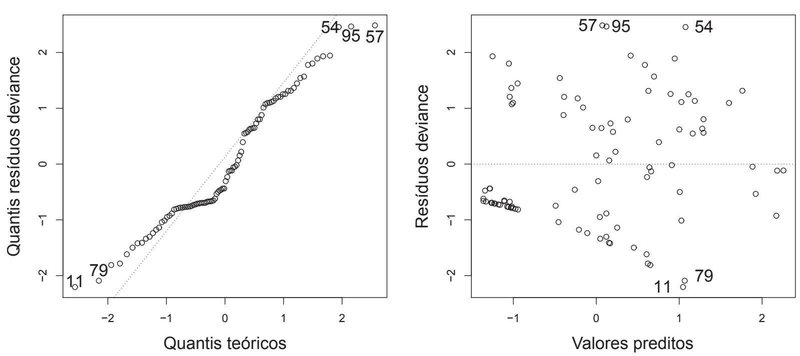


Figura : Resíduos martingale e *deviance*, identificando indivíduos mal ajustados pelo modelo `m4` (TMO). Observe que o resíduo *deviance* detecta indivíduos com grandes resíduos positivos, o que não foi possível com o resíduo martingale

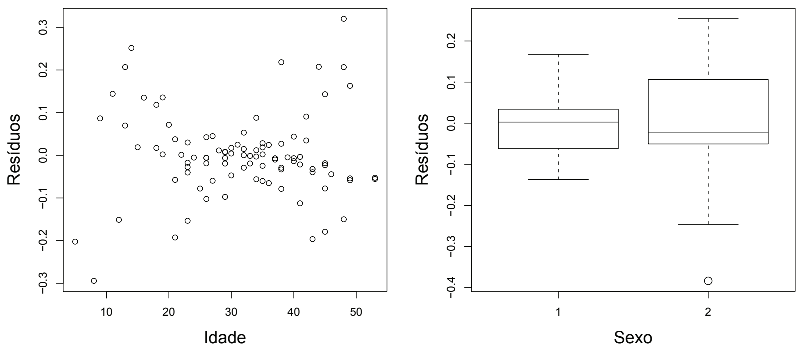
Pontos aberrantes: Martingale X *deviance*



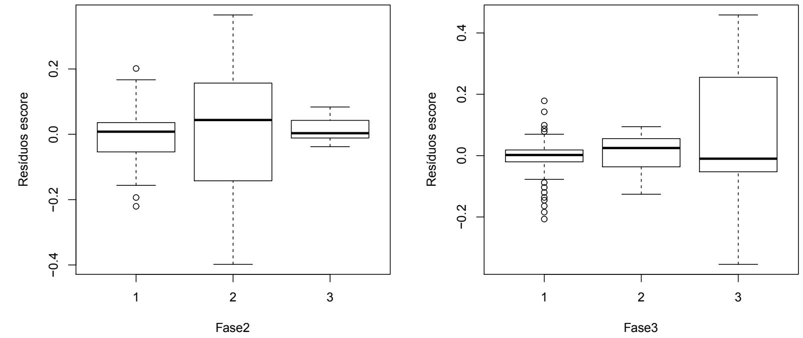
Resíduos escore – *dfbetas*

- 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 a $(x_i - \bar{x})\times$ 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

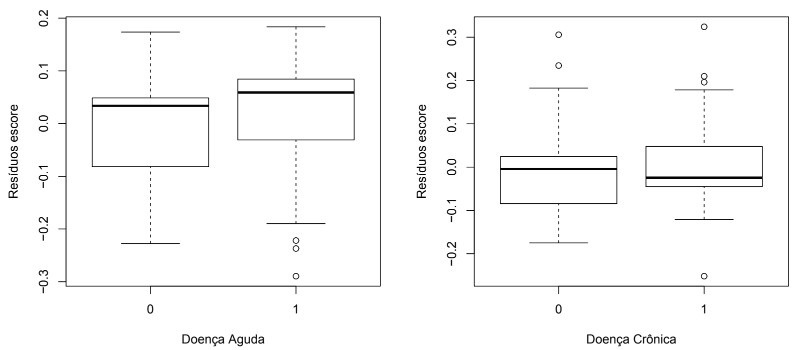
Resíduos escore - Exemplo modelo 4 TMO



Resíduos escore- Exemplo modelo 4 TMO



Resíduos escore



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 proporcionalidade global	teste de proporcionalidade global: função *cox.zph*
Avaliar proporcionalidade de cada variável	gráficos resíduo de Schoenfeld vs tempo
Identificar pontos aberrantes	resíduo martingale e resíduos *deviance*
Estudar forma funcional da variável	gráficos resíduo martingale do modelo nulo vs covariável
Identificar pontos influentes	gráficos resíduo escore vs covariável

Outline

Covariáveis Mudando no Tempo

- Introdução
- Estrutura do dado mudando no tempo
- Diagnóstico
- Dados prevalentes
- Intervalos de tempo descontínuos

Introdução

- Analisar a sobrevida quando as covariáveis mudam ao longo do tempo.
- Construir adequadamente o banco de dados na situação de covariáveis tempo-dependentes.
- O que muda? Tudo:
  - Idade
  - Terapia antiretroviral
  - Medicamento: crossover, efeitos colaterais
  - Hábitos: exercício, alimentação
  - Residência
  - Emprego
- Modelo de Cox Estendido – Processo de Contagem

Modelo de Cox Estendido

$$\lambda(t|\boldsymbol{x}(t)) \quad = \quad \lambda_0(t)\exp(\boldsymbol{x}(t)\boldsymbol{\beta})$$

Onde está a diferença?

Estrutura dos dados mudando no tempo

id	sexo	idade	status	inicio	fim	deag	decr	recplaq	fasegr
1	2	31	0	0	9	0	0	0	CP1
1	2	31	0	9	1000	0	0	1	CP1
2	2	38	0	0	28	0	0	0	CP1
2	2	38	1	28	39	1	0	0	CP1
3	1	23	0	0	27	0	0	0	CP1
3	1	23	0	27	36	0	0	1	CP1
3	1	23	0	36	268	1	0	1	CP1
3	1	23	1	268	434	1	1	1	CP1
4	2	5	0	0	24	0	0	0	CP1
4	2	5	1	24	69	1	0	0	CP1
5	2	15	0	0	22	0	0	0	CP1
5	2	15	0	22	83	1	0	0	CP1
5	2	15	0	83	446	1	0	1	CP1
5	2	15	1	446	672	1	1	1	CP1

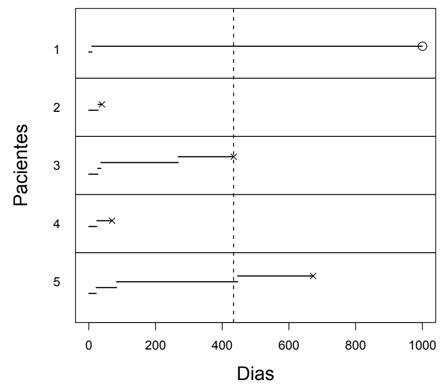
Estrutura dos dados

	Mudança			
	1^{a}	2^{a}	3^{a}	4^{a}
paciente 1	(0,9+]	(9,1000+]		
paciente 2	(0,28+]	(28,39]		
paciente 3	(0,27+]	(27,36+]	(36,268+]	(268,434]
paciente 4	(0,24+]	(24,69]		
paciente 5	(0,22+]	(22,83+]	(83,446+]	(446,672]

$+$ representa censura
($\rightarrow$ intervalo aberto, NÃO inclui o limite inferior
] $\rightarrow$ intervalo fechado, inclui o limite superior

Gráfico da estrutura dos dados de TMO

Quais pacientes estão em risco no tempo=434 dias (linha vertical)?



Estimação

- Não há superposição dos tempo
- A verossimilhança parcial utilizará no máximo uma observação de cada paciente em qualquer momento.
- A soma de indivíduos em risco será feita sobre um conjunto de observações independentes.

Exemplo – aids

Estudar o efeito da terapia anti-retroviral de alta potência (Haart) no tempo de sobrevida desde o diagnóstico de Aids até o óbito. Foi registrado a mudança de tratamento (haart = S ou N) ao longo do estudo. http://sobrevida.fiocruz.br/

reg	haart	ini	fim	sexo	escol	status	idade
4	N	942	1448	M	Gin	0	38
4	S	1448	1939	M	Gin	0	38
4	N	1939	1959	M	Gin	0	38
4	S	1959	3297	M	Gin	0	38
11	N	2162	2988	F	Prim	0	38
11	S	2988	3297	F	Prim	0	38
32	N	665	804	F	Prim	1	36
33	S	1498	1820	M	Univ	0	76
33	S	2400	3297	M	Univ	0	76
34	N	686	3200	M	Sec	0	33
35	N	769	1577	M	Sec	0	30
35	S	1577	1597	M	Sec	1	31
36	S	3255	3297	F	Prim	0	52

Tempo final da primeira linha do paciente 33 é diferente do tempo inicial da segunda linha. Por que?

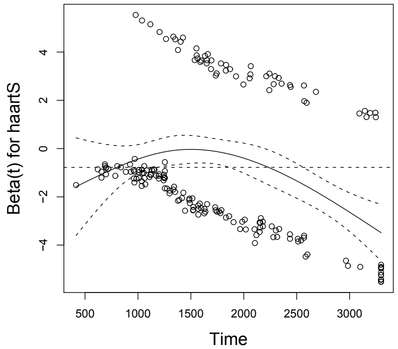
Exemplo

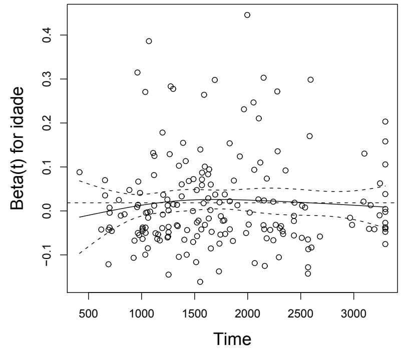
```
> muda.cox <- coxph(Surv(ini,fim,status)~haart+idade+
        escol+sexo,data=muda)
> muda.cox

Call:
coxph(formula = Surv(ini, fim, status) ~ haart +
                 idade + escol + sexo, data = muda)
                coef exp(coef) se(coef)      z       p
haartS     -0.7779     0.459  0.18508 -4.203 2.6e-05
idade       0.0185     1.019  0.00754  2.448 1.4e-02
escolAnalf -0.2342     0.791  0.76547 -0.306 7.6e-01
escolGin    0.5364     1.710  0.32688  1.641 1.0e-01
escolPrim   0.7438     2.104  0.31075  2.394 1.7e-02
escolSec    0.3265     1.386  0.33905  0.963 3.4e-01
sexoM       0.2253     1.253  0.16929  1.331 1.8e-01

Likelihood ratio test=35.1  on 7 df, p=1.08e-05  n= 1377
```


Uso dos resíduos

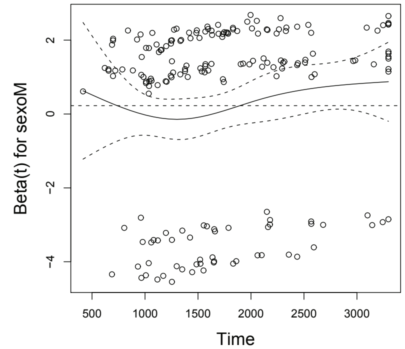
- Schoenfeld:
  - são calculados para os tempos de ocorrência do evento – definição e cálculo sem alteração para processo de contagem
  - valor da covariável utilizado nos cálculos corresponde ao tempo de evento
  - escala *default* é k-m, trocar para o tempo t (argumento *transform = "identity"*)
- Martingale:
  - podem ser calculados para cada registro – sem alteração
  - ou para cada indivíduo (argumento *collapse = id*)
- Escore: sem alteração.

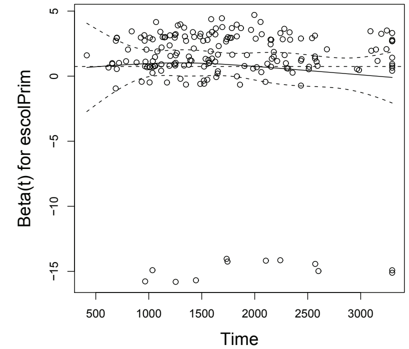
Resíduos Schoenfeld – aids



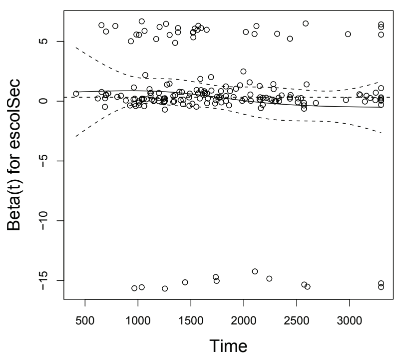


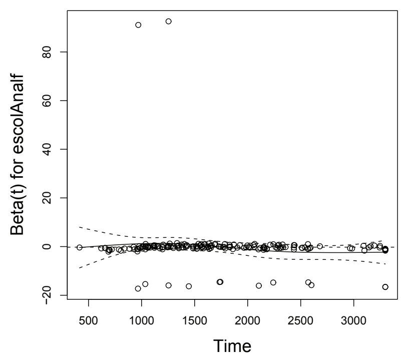
Resíduos Schoenfeld – aids



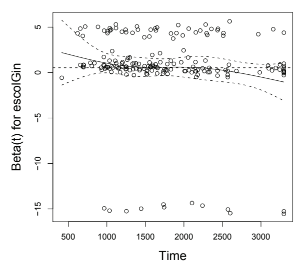


Resíduos Schoenfeld – aids





Resíduos Shoenfeld – aids



	rho	chisq	p
haartS	-0.26583	16.70605	0.000
idade	0.00627	0.00775	0.930
escolAnalf	-0.12455	2.86745	0.090
escolGin	-0.12721	3.03844	0.081
escolPrim	-0.07071	0.96321	0.326
escolSec	-0.10421	2.03111	0.154
sexoM	0.12002	2.94786	0.086
GLOBAL	NA	24.41845	0.000962

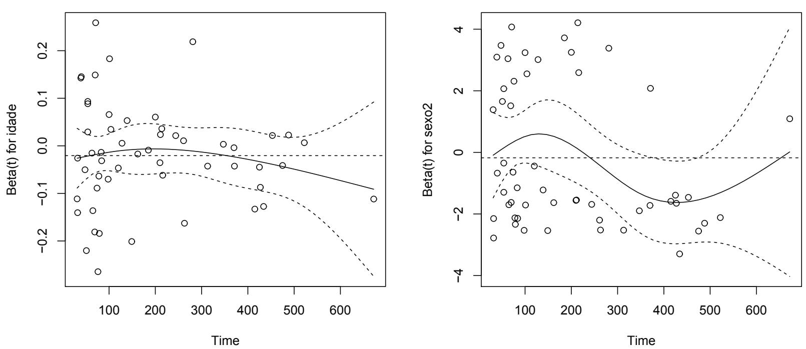
Exemplo - TMO

```
Call:
coxph(formula = Surv(inicio, fim, status) ~ idade + sexo +
      fasegr + deag + decr + recplaq, data = tmopc)

                coef exp(coef) se(coef)      z       p
idade        -0.0206     0.980   0.0140 -1.471 1.4e-01
sexo2        -0.1766     0.838   0.3093 -0.571 5.7e-01
fasegrOther   0.9266     2.526   0.3108  2.981 2.9e-03
deag1         1.0531     2.866   0.2917  3.610 3.1e-04
decr1         0.4370     1.548   0.3859  1.133 2.6e-01
recplaq0      1.9630     7.120   0.4671  4.203 2.6e-05

Likelihood ratio test=50.3  on 6 df, p=4.05e-09  n= 259,
number of events= 53
```

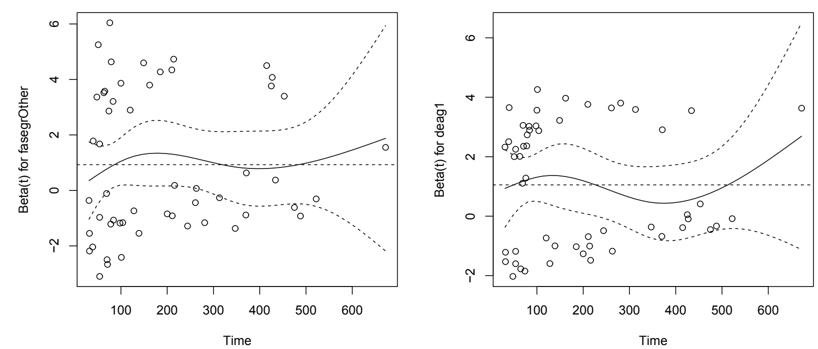

Diagnóstico – Schoenfeld – TMO

```
> tmo.sch <- cox.zph(tmo.cox)
> tmo.sch
                 rho   chisq      p
idade       -0.08369 0.41389 0.5200
sexo2       -0.25846 3.67790 0.0551
fasegrOther  0.04967 0.16535 0.6843
deag1       -0.03694 0.06742 0.7951
decr1        0.01235 0.00958 0.9220
recplaq0     0.00507 0.00177 0.9665
GLOBAL            NA 4.41245 0.6210
```

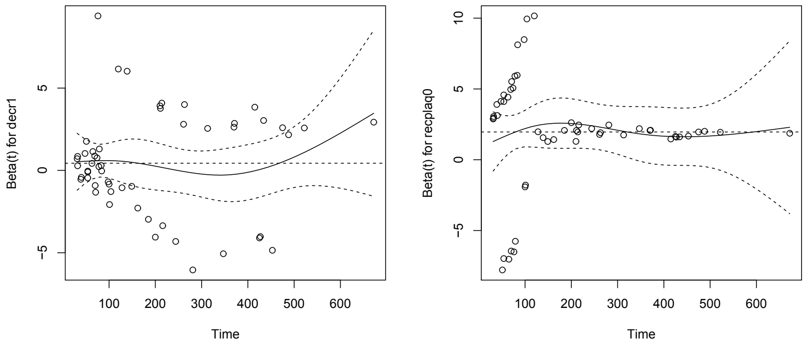

Schoenfeld – TMO



Schoenfeld – TMO



Schoenfeld – TMO

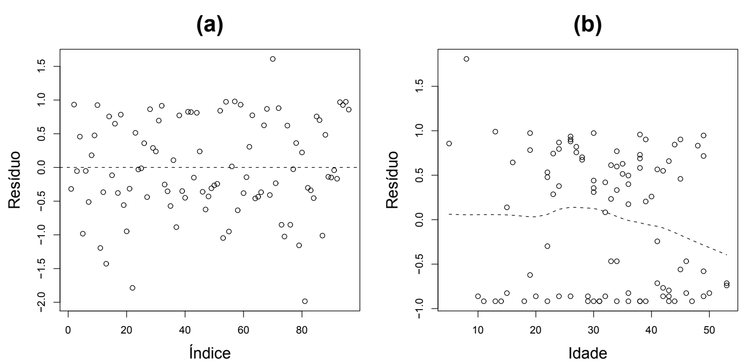


Resíduos Martingale

- Podem ser calculados para cada um dos intervalos de tempo nos quais não há mudança de covariável (cada linha)
- Ou para cada um dos n indivíduos (resíduo individual $=$ soma dos resíduos do indivíduo em cada intervalo de tempo)
- incluir argumento *`collapse=id`* para o obter resíduo individual

Resíduos martingale

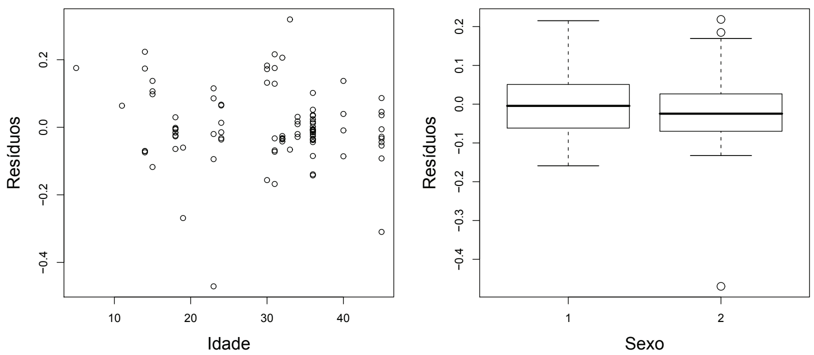
Resíduos de martingale para o modelo `tmo.cox` versus índice (a) e para o resíduo do modelo nulo versus idade (b) (covariável contínua).



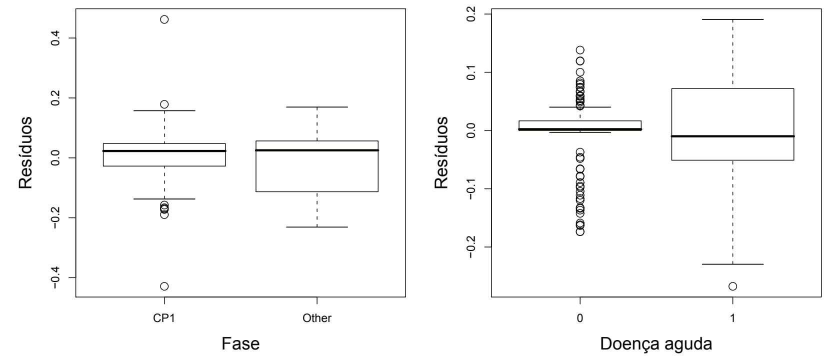
Resíduos escore

- Permite identificar pontos de alavanca por períodos de tempo (linha)
- Ou indivíduos alavanca
- *`collapse=id`* – para o indivíduo

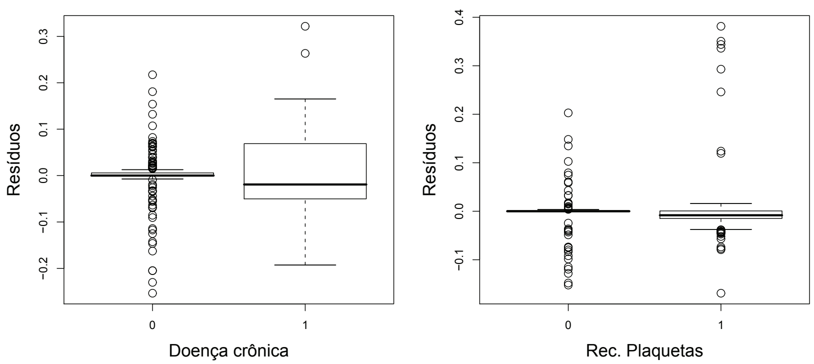
Resíduos escore – TMO



Resíduos escore – TMO



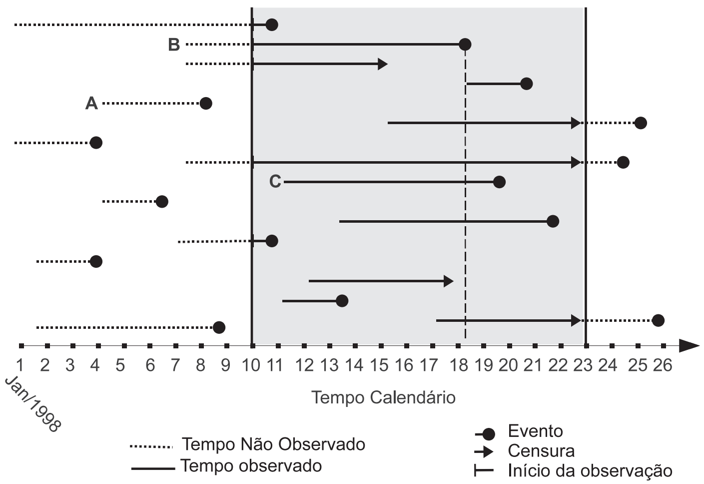
Resíduos escore – TMO



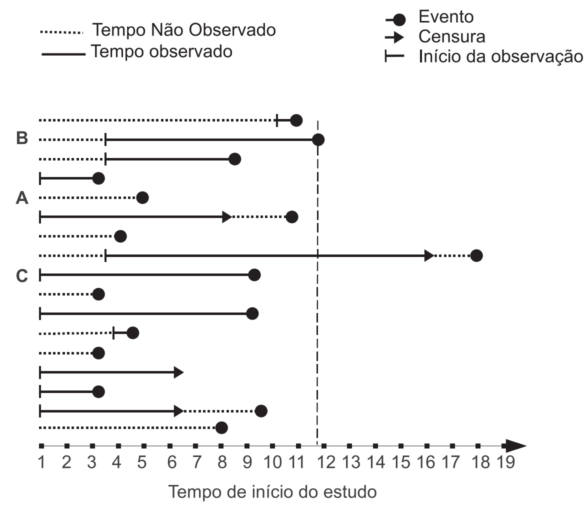
Dados prevalentes

- Identificar os valores corretos das covariáveis para cada paciente em cada intervalo de tempo (construir corretamente o banco de dados): vale para dados prevalentes ou truncados à esquerda
- Definir como data de referência t_0 a data mais antiga no banco de dado, ou
- Calcular o tempo de entrada na coorte de cada indivíduo: limite inferior do seu 1ºintervalo de tempo, o momento de entrada no estudo
- Cada indivíduo será analisado dentro de sua janela temporal, eliminando o viés potencial da introdução na coorte de sobreviventes com tempos mais longos
- E a forma de interpretar os efeitos é condicional – dado que o indivíduo sobreviveu até entrar em observação

Dados prevalentes



Dados prevalentes



Dados prevalentes

- A escolha da estratégia depende do objetivo:
  - Fatores de risco individuais – tempo de início do estudo
  - Fator presente em todos ao mesmo tempo – tempo calendário
- Tempo total de observação não deve ser usado pois não ajusta dados prevalentes.

Tempo descontínuo

- Podem ocorrer por: ausência de informação, afastamentos por viagem, interrupção, eventos múltiplos (próximo tópico)
- O mesmo mecanismo de registrar os intervalos de tempo (início,fim) permite tratar adequadamente dados de indivíduos com risco descontínuo ao longo do estudo

Comparando abordagens de tempo

- 6805 pacientes que iniciam hemodiálise, acompanhados por 44 meses
- analisando como se somente acompanhados a partir do 20^o – 5891
- Comparando modelos com dado completo e dado truncado, nas 3 formas de incluir o tempo.

Comparando abordagens de tempo

Tabela : HR para dados prevalentes por tempo calendário, tempo de diálise e tempo total de cada indivíduo

	Dados Completos		**Dados Truncados**		
Covariável	Tempo total	Calen-dário	Tempo total	Tempo em diálise	Calen-dário
Idade	1,04*	1,04*	1,04*	1,04*	1,04*
Causa:					
Congênita	0,49*	0,47*	0,51*	0,56*	0,53*
Diabetes	1,34*	1,38*	1,34*	1,32*	1,35*
Outras	1,04	1,07	1,04	1,02	1,05
Renal	1,04	1,04	1,07	1,09	1,09

$^*p < 0,05$

Resumo

A principal questão é montar o banco de dados após identificar adequadamente:

- o **tempo inicial** do acompanhamento de cada indivíduo ou que define mudança no valor da covariável;
- o **tempo final**, seja do acompanhamento ou por mudança no valor da covariável;
- o **status** em cada período entre mudanças de covariável e ao final do acompanhamento do indivíduo.