

## Modelo de Regressão de Cox

### O modelo de Cox com taxas de falhas proporcionais

MAE 514 - Introdução à Análise de Sobrevida e Aplicações

IME-USP

## Introdução

Modelo de Cox(1972,1975)

- Não pressupõe um modelo paramétrico para os tempos de falha
- Generaliza o teste log-rank
- Modelo de regressão (ajuste de várias covariáveis simultaneamente)  
→ Detecção de diferenças não observáveis em análises univariadas
- Em sua forma mais simples, é um modelo com taxas de falhas proporcionais

## Taxas de falhas proporcionais

Taxas de falhas proporcionais

$$\alpha(t | \mathbf{X}) = \alpha_0(t) \exp\{\mathbf{X}'\beta\}$$

- $\mathbf{X}$ : vetor com  $p$  covariáveis
- $\beta$ : vetor com  $p$  parâmetros (desconhecidos)
- $\alpha_0(t)$ : função de taxa de falha basal (referência)

No modelo de Cox:

$$\alpha_0(t) \text{ é arbitrária}$$

Note que

$$\alpha_0(t) \geq 0, \quad \forall t \geq 0$$

## Interpretação de parâmetros - *Dummy*

$$\alpha(t | \mathbf{X}) = \alpha_0(t) \exp\{\mathbf{X}'\beta\}$$

Suponha  $p = 1$  com

$$X = \begin{cases} 0, & \text{se grupo A} \\ 1, & \text{se grupo B} \end{cases}$$

Então,

$$\frac{\alpha(t | X = 1)}{\alpha(t | X = 0)} = e^\beta$$

### Interpretação

Por exemplo, se  $e^\beta = 0,80$ , a taxa com que as falhas ocorrem para o grupo B é 80% da taxa com que as falhas ocorrem para o grupo A.

$X$  quantitativa (ex.: idade)

$$\frac{\alpha(t | X + 1)}{\alpha(t | X)} = e^\beta$$

- Taxas de falhas proporcionais: não depende do tempo

### Interpretação

Para um acréscimo de 1 unidade (mês, dia, ano...) na idade, a taxa de falha é multiplicada por  $e^\beta$ . Se  $e^\beta = 1,20$ , então, ao aumentar a idade em uma unidade, a taxa com que as falhas ocorrem aumenta em 20%.

- $e^\beta$  muito pequeno ou muito grande  $\Rightarrow$  escala de  $X$

## Estimação no modelo de Cox

$$\alpha(t | \mathbf{X}) = \alpha_0(t) e^{\mathbf{X}'\beta}$$

- $\beta$ : Parâmetro de interesse primário
- $\alpha_0(\cdot)$ : Parâmetro de perturbação

### Verossimilhança Parcial

$$L_1(\beta) = \prod_{i=1}^n \left( \frac{e^{\mathbf{X}_i'\beta}}{\sum_{\ell \in \mathcal{R}_i} e^{\mathbf{X}_\ell'\beta}} \right)^{\delta_i} \quad (1)$$

- $\mathcal{R}_i$ : índices das u.e. em risco na  $i$ -ésima falha

Maximização de  $L_1(\cdot) \Rightarrow \hat{\beta}$  (E.M.V.P.)

## Teste de Hipóteses

$$H_0 : \beta = \beta_0$$

- Não é possível obter a distribuição exata do EMVP

## Distribuição assintótica do EMVP

$$[i_o(\beta_0)]^{-1/2}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} \mathcal{N}_p(\mathbf{0}; \mathbf{I}_p)$$

- $i_o(\cdot)$ : matriz de informação *observada*
- $\mathbf{I}_p$ : matriz identidade de dimensão  $p \times p$

Três testes assintóticos são usualmente utilizados:

- 1 Teste de Wald
- 2 Teste de Rao ou Escore
- 3 Teste da Razão de Verossimilhanças (parciais)

## Teste de Wald

A estatística de teste é dada pela forma quadrática

$$W = (\hat{\beta} - \beta_0)' [i_o(\hat{\beta})] (\hat{\beta} - \beta_0)$$

Segue pela distribuição assintótica do EMVP que, sob  $H_0 : \beta = \beta_0$ ,

## Convergência de $W$

Sob  $H_0$ ,

$$W \xrightarrow{\mathcal{D}} \chi_p^2, \quad n \rightarrow \infty$$

- $p$ : dimensão do espaço paramétrico associado a  $\beta$

## Teste de Rao - Escore

$$R = [\mathbf{U}(\beta_0)]' [i_o(\beta_0)]^{-1} [\mathbf{U}(\beta_0)]$$

- $\mathbf{U}(\cdot)$ : função escore

Pode ser mostrado que, sob  $H_0 : \beta = \beta_0$ ,

## Convergência de $R$

Sob  $H_0$ ,

$$R \xrightarrow{\mathcal{D}} \chi_p^2, \quad n \rightarrow \infty$$

- $p$ : dimensão do espaço paramétrico associado a  $\beta$

### Teste da razão de verossimilhanças parciais

$$\Lambda = -2 \times \{\log[L_1(\beta_0)] - \log[L_1(\hat{\beta})]\}$$

- $L_1(\cdot)$ : verossimilhança parcial

### Convergência de $\Lambda$

Sob  $H_0$ ,

$$\Lambda \xrightarrow{D} \chi_p^2, \quad n \rightarrow \infty$$

- $p$ : diferença entre as dimensões dos espaços paramétricos irrestrito e restrito por  $H_0$

### Comentários

$$H_0 : \beta = \beta_0$$

### Estatísticas de testes

$$W = (\hat{\beta} - \beta_0)' [i_o(\hat{\beta})] (\hat{\beta} - \beta_0)$$

$$R = [\mathbf{U}(\beta_0)]' [i_o(\beta_0)]^{-1} [\mathbf{U}(\beta_0)]$$

$$\Lambda = -2 \times \{\log[L_1(\beta_0)] - \log[L_1(\hat{\beta})]\}$$

- Os testes são assintoticamente equivalentes
- O teste baseado em  $R$  não utiliza o EMVP
- $\Lambda$  é o teste mais poderoso entre os três (deve ser escolhido em caso de discrepância dos resultados)

### Exemplo: Avaliação de uma droga cancerígena

- $n = 100$  ratos
- $Z_i$ : tempo entre inoculação de células cancerosas e aparecimento de tumor ou censura
- $\delta_i$ : indicador de evento (aparecimento do tumor=1)
- $X_i$ : dummy indicando droga (1) ou placebo (0)

tempo	delta	tratamento
101	0	1
104	0	1
104	0	1
77	0	1
89	0	1
88	1	1
⋮	⋮	⋮

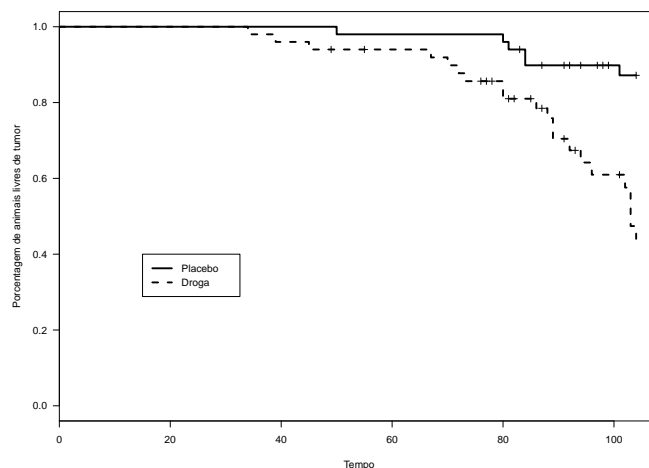
## Análise Descritiva

### Leitura dos dados e gráfico KM

```
> ratos <- read.table("ratos.dat",header=T)
> library(survival)
> ratos.km<-survfit(Surv(tempo,delta)~trat,ratos)
> plot(ratos.km, lty = c(1,2),lwd = c(3,3),las=1,
      xlab="Tempo", ylab="Porcentagem de animais livres
      de tumor")
> legend(15, 0.4, c("Placebo","Droga"),
      lty=c(1, 2),lwd=c(3, 3))
```

Navigation icons

## Análise Descritiva - Kaplan-Meier



Navigation icons

## Análise Descritiva

### Teste log-rank

```
> survdiff(Surv(tempo,delta)~trat,ratos,rho=0)
```

Chisq= 15.2 on 1 degrees of freedom, p=9.75e-05

### Teste Gehan-Wilcoxon

```
> survdiff(Surv(tempo,delta)~trat,ratos,rho=1)
```

Chisq= 14 on 1 degrees of freedom, p= 0.000182

Evidências de que a droga altera o desenvolvimento de tumores em ratos

Navigation icons

## Ajuste do Modelo de Cox no R

### Função `coxph`

```
coxph(formula, data=parent.frame(), weights,
subset, na.action, init, control,
method=c("efron", "breslow", "exact"),
singular.ok=TRUE, robust=FALSE, model=FALSE,
x=FALSE, y=TRUE, ...)
```

Obs.:

- `formula`: `Surv(tempo,delta)~cov1+cov2+...`
- `data`: conjunto de dados (data-frame)
- `method`: Método para lidar com observações empatadas



## `coxph` aplicado aos dados dos ratos

### Comando R

```
> coxph(Surv(tempo,delta) ~ trat, ratos)
```

```
      coef  exp(coef) se(coef)      z      p
trat  1.64     5.14  0.465     3.52 0.00044
Likelihood ratio test=15.3 on 1 df, p=9e-05 n= 100
```

- Teste de Wald para efeito individual de droga  $p = 0,00044$
- Teste da razão de verossimilhanças  $p < 0,001$
- Taxa de falha relativa igual a 5,14:

**A taxa de ocorrência de tumores em ratos que recebem a droga é 5 vezes a taxa associada a ratos que recebem placebo**



## Mais informações

### Comando `summary`

```
> summary(coxph(Surv(tempo,delta)~trat,ratos))
```

```
      coef  exp(coef) se(coef)      z      p
trat  1.64     5.14  0.465     3.52 0.00044
      exp(coef) exp(-coef) lower .95 upper .95
trat  5.14         0.195     2.06     12.8
Rsquare= 0.142 (max possible= 0.901 )
Likelihood ratio test= 15.3 on 1 df, p=9e-05
Wald test = 12.4 on 1 df, p=0.000439
Score (logrank) test = 15.2 on 1 df, p=9.43e-05
```

- Intervalo de confiança para taxa de falha relativa
- Testes de Wald global (no caso é igual ao individual)
- Teste de Rao-Escore (no caso igual ao log-rank)



## Taxas de falhas proporcionais

$$\alpha(t | \mathbf{X}) = \alpha_0(t) \exp(\mathbf{X}\beta)$$

Lembrando que

$$S(t | \mathbf{X}) = \exp \left\{ - \int_0^t \alpha(s | \mathbf{X}) ds \right\}$$

vem que

$$S(t | \mathbf{X}) = \exp \left\{ - \int_0^t \alpha_0(s) \exp\{\mathbf{X}'\beta\} ds \right\} = [S_0(t)] e^{\mathbf{X}\beta}$$

e assim

$$\log[-\log(S(t | \mathbf{X}))] = \log[A_0(t)] + \mathbf{X}\beta$$

$A_0(\cdot)$ : função de taxa de falha acumulada para placebo

## Verificação da suposição de taxa de falha proporcional

Para taxas de falhas proporcionais,

$$\log[-\log(S(t | \mathbf{X}))] = \log[A_0(t)] + \mathbf{X}\beta$$

Para o grupo placebo,

$$\log[-\log(S(t | \mathbf{X}))] = \log[A_0(t)]$$

Para o grupo que recebe a droga,

$$\log[-\log(S(t | \mathbf{X}))] = \log[A_0(t)] + \beta$$

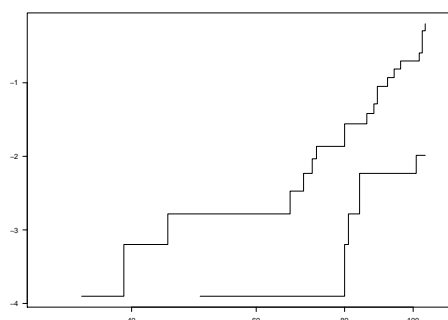
**Característica esperada**

**Curvas para placebo e droga paralelas**

## Averiguação para os dados dos Ratos

**Avaliação através do estimador Kaplan-Meier**

```
> plot(ratos.km, mark.time=F, fun='cloglog',
      xlim=c(30, 110), las=1)
```



## Dados de leucemia - Arquivo anderson.dat

### Variáveis

- Coluna 1: ID - Identificação
- Coluna 2: tempo - Tempo (semanas) de remissão dos sintomas
- Coluna 3: delta - indicador de censura
- Coluna 4: Sexo - 0 masculino, 1 feminino
- Coluna 5: logcgb - log da contagem de globulos brancos
- Coluna 6: Rx - tratamento: 0 novo, 1 padrão

### comando coxph

```
> anderson <- read.table("anderson.dat", header=T)
> coxph(Surv(tempo, delta) ~ sexo + logcgb + Rx, anderson)
```

### Saída

```
      coef  exp(coef)  se(coef)      z      p
sexo    0.416    1.52    0.460    0.905  3.7e-01
logcgb  1.706    5.51    0.345    4.947  7.5e-07
Rx       1.700    5.47    0.483    3.519  4.3e-04
likelihood ratio test=49.6 on 3 df, p=9.55e-11 n=
42, number of events= 30
```

- Testes de Wald individual: *Sexo* não é estat. significante
- Teste da razão de verossimilhanças ( $H_0$ : todos parâmetros são nulos): rejeita-se  $H_0$

## Modelo Final

### comando coxph

```
> summary(coxph(Surv(tempo, delta) ~ logcgb + Rx, anderson))
```

### Saída

```
      coef  exp(coef)  se(coef)      z  Pr(>|z|)
logcgb  1.717    5.566    0.344    4.99  5.96e-07 ***
Rx       1.534    4.635    0.441    3.47  0.000513 ***
      exp(coef)  exp(-coef)  lower .95  upper .95
logcgb  5.566    0.1797    2.837    10.92
Rx       4.634    0.2158    1.951    11.01
Rsquare= 0.687 (max possible= 0.988 )
Likelihood ratio test= 48.8 on 2 df, p=2.50e-11
Wald test = 34 on 2 df, p=4.21e-08
Score (logrank) test = 47.2 on 2 df, p=5.66e-11
```



	coef	exp(coef)	se(coef)	z	Pr(> z )
logcgb	1.717	5.566	0.344	4.99	5.96e-07
Rx	1.534	4.635	0.441	3.47	0.000513

- O teste de Wald individual indica que os dois fatores são estatisticamente significantes

Rsquare= 0.687 (max possible= 0.988 )  
Likelihood ratio test= 48.8 on 2 df, p=2.50e-11  
Wald test = 34 on 2 df, p=4.21e-08  
Score (logrank) test = 47.2 on 2 df, p=5.66e-11

- As três estatísticas globais levam à mesma conclusão

## Interpretação

	coef	exp(coef)	se(coef)	z	Pr(> z )
logcgb	1.717	5.566	0.344	4.99	5.96e-07
Rx	1.534	4.635	0.441	3.47	0.000513

- Para um mesmo valor de logcgb, a taxa com que os sintomas retornam para pacientes submetidos ao tratamento tradicional é aproximadamente 4.5 vezes a taxa de pacientes submetidos ao novo tratamento.
- Para um mesmo tratamento, um acréscimo de uma unidade no logcgb aproximadamente multiplica por seis a taxa de retorno dos sintomas.