

Survival Analysis

Ciro S. Costa

17 Nov, 2015

Análise de Sobrevivência

Análise que independe do tempo de monitoramento (período) ser igual ou não. Busca estudar o fato de atingir ou não um determinado evento de interesse (e quanto tempo demorou, caso atingido).

Tipicamente os dados se apresentam da forma:

qtd	indica a	
tempo	ocorrendia	
até	do evento	
evento	ou não	
ou nao		variáveis explicativas

time	status	x1	x2	...	xn
t0	1				
t1	0				
t2	1				
..	...				
tM	0				

Dados Censurados são dados para os quais o tempo até o evento crítico é não-conhecido. **NÃO DEVEM SER DESCARTADOS**, mas sim, incluídos no cálculo da função de sobrevivência.

Em tais modelos devemos levar em consideração a censura (o não-evento marcado). Suas causas são: - não-termino do monitoramento - missing-values - de fato o evento desejado não ter ocorrido no período

A análise pode ser realizada para apenas um ou múltiplos grupos: no primeiro caso, examina-se a curva de sobrevivência acumulado para o mesmo, apresentando então as probabilidades estimadas de sobrevivência após o final de cada período. Caso haja diversos grupos, curva de sobrevivência acumulada para cada grupo.

Procedimientos

Dois são os possíveis procedimentos a serem utilizados dependendo de nosso interesse: *Kaplan-Meier* e *Regressão de Cox*.

Kaplan-Meier

Procedimento para se criar tábuas de mortalidade a partir das quais avalia-se a probabilidade de um determinado indivíduo sobreviver a um determinado tempo. Como especificado acima, necessariamente precisamos de duas variáveis especiais: **time** e **status**. A primeira indica o tempo até o evento (ou não-evento). A segunda, a censura ou ocorrência.

No STATA iniciamos a análise indicando as variáveis relevantes à análise de sobrevivência:

```
* declara que os dados serão utilizados para
* análise de sobreviv.
. stset <tempo>, failure(<status>)
```

Conseguimos então gerar a tabela de sobrevida:

```
. ltable <tempo> <status>, survival
```

Com esta em mãos, somos capazes de responder questões relativas à probabilidade estimada de sobrevida após N tempo. Tal informação é obtida através da tabela de sobrevida.

A parte probabilística da tal pode ser feita à mão caso sejam disponibilizados apenas `deaths` e `lost`:

$$Pr(X > k) = \prod (1 - (\frac{death s_i}{beq Tot_i}))$$

```
. sts graph [if <cond>] , by(<varlist>)
```

Para verificar se há diferenças entre dois grupos realiza-se o teste de equidade das funções de sobrevivência segregando de acordo com a variável binárias desejada:

```
. sts test <varlist> [if <cond>]
```

Sabemos então se há diferença estatisticamente significativa (a um nível de 5% de significância) através do resultado de $\text{Pr}>\chi^2$: se for maior que 0.05, não há diferença. Caso contrário, há.

Vale notar que ao utilizar o procedimento *Kaplan Meier* não há a necessidade de se transformar as variáveis categóricas em dummies já que não se busca predição (o que será feito no caso da *Cox Regression*, adiante).

Outro ponto relevante ao procedimento é que ao realizar o teste verifica-se a presença ou ausência de significância entre *pelo menos* duas curvas em um conjunto de N curvas.

Regressão de Cox (Modelo de Riscos Proporcionais)

Como se trata de um modelo de análise de sobrevivência, iniciamos configurando-o:

```
. stset <time_var>, failure(<status_var>)
```

E então rodando a regressão:

```
. stcox <ls>, basesurv(survivor) basechaz(hazard)
. stepwise, pr(0.05): stcox <ls>
```

Com o resultado da regressão, verificamos se o modelo obtido é significativo:

- checamos se o efeito de uma ou mais variáveis são estatisticamente diferentes de zero por meio do teste χ^2 . Se `Prob > chi2` for menor que 0.05, rejeitamos a hipótese nula de que todas as variáveis apresentam parâmetros estatisticamente iguais a zero a um nível de significância de 5%, *i.e.*, pelo menos uma variável é significativa.
- checamos se todas as variáveis são significativas no modelo (note: como rodamos com `stepwise` já seriam descartadas aquelas não significantes, como mostraria em `P>|z|`).

Outro resultado do relevante do output é o *Hazard Ratio* (taxa de risco), que indica a mudança estimada no risco associada à alteração de **uma** unidade da variável preditora (similar aos modelos com *odds ratio*).

Exemplo 1: HR da variável categórica X é 0.3645, ou seja, *ceteris paribus*, o risco de ocorrer evento é 0.3645 vezes o risco de não-ocorrência de evento.

Exemplo 2: HR da variável quantitativa Y é 0.8989, ou seja, *ceteris paribus*, o aumento de uma unidade de Y faz com que o risco de evento seja multiplicado por 0.8989.

O gráfico das funções de sobrevivência exibe as probabilidades de sobreviver à ocorrência do evento após determinado tempo passado.

```
. stcurve, survival at1 (monitor=0) at2 (monitor=1)
```

Pode-se dizer que o *Hazard Rate* (e não *ratio*) mensura a taxa à qual o risco é acumulado com o passar do tempo. Quanto mais aumenta o risco, maior a inclinação da curva de risco acumulado no período. O gráfico de risco acumulado então retrata o crescimento dos riscos de ocorrência de evento.

```
. stcurve, cumhaz at1 (monitor=0) at2 (monitor=1)
```

Assim como no caso do *Kaplan Meier*, podemos verificar se há diferenças estatisticamente significantes entre duas curvas de sobrevida:

```
. sts test monitor
```