Lecture 6: Survival Analysis

Dankmar Böhning

Southampton Statistical Sciences Research Institute
University of Southampton

S\textsuperscript{3}RI, 2 - 4 March 2015
Introduction

Basic definitions

The hazard
A couple of questions and...

- What makes survival data so special that their analysis needs a special treatment, even as long as a one-term course?
- Why isn’t it simply covered as a sub-topic in, let’s say, regression analysis?
...a clarification

- Survival data subsume more than only times from birth to death for some individuals.
- Analysis of duration data, that is the time from a well-defined starting point until the event of interest occurs.
Examples

- how long patients survived after diagnosis or treatment
- the length of unemployment spells
- how long a marriage lasts
- how long PhD students need to finish writing their theses
- and more...
Features

- Survival data result from a dynamic process and we want to capture these dynamics in the analysis properly.
- The observation scheme for duration data can be rather complex, leading to data that are somehow cut.
The basic functions

In the following we will assume that time is running continuously, and we therefore will describe duration by a continuous random variable, denoted by $T$.

- $T \geq 0$
- $f(t) \Rightarrow$ density function
- $F(t) \Rightarrow$ cumulative density function (cdf)
- $S(t) \Rightarrow$ survival function
Recall that...

- The density function $f(t)$ describes how the total probability of 1 is distributed over the domain of $T$.
- The function $f(t)$ itself is not a probability and can take values bigger than 1. But still one can derive basic properties from looking at the density.
- For regions where the density has large values the area under the curve over an interval of given length will be larger as compared to an interval of same length where the density is lower.
- Regions over which the density is high are regions where we expect to observe more data points than in regions with low densities.
Recall that...

- The cdf $F(t)$ is defined as $F(t) := P(T \leq t)$ which can be computed from the density as

$$F(t) = \int_{0}^{t} f(s) \, ds$$

- A cdf is an increasing function, even strictly increasing if the density $f(t) > 0$ everywhere.
- $F(0) = 0$ and $\lim_{t \to \infty} F(t) = 1$.
- There is a one-to-one link between $f(t)$ and $F(t)$ as $F'(t) = f(t)$. Knowing one of the functions means, at least in principle, knowing the other (you may have to take the derivative or perhaps solve an ugly integral).
Recall that...

Instead of looking at the cdf, which gives the probability of surviving at most $t$ time units, one prefers to look at survival beyond a given point in time. This is described by the survival function $S(t)$:

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

Consequently, $S(t)$ starts at 1 for $t = 0$ and then declines to 0 for $t \to \infty$.

It should be obvious that knowing any one of $f(t)$, $F(t)$ and $S(t)$ allows to derive the other two functions.
To summarize

$$\Pr(a \leq T \leq b)$$

All the three functions introduced so far allowed to describe, in one way or another, how the survival times are distributed over the potential range.
The dynamic process

- Density, cdf and survival function look at the marginal distribution

- Conditioning on the survival experience so far, we have

\[
Pr(t < T \leq t + \Delta t \mid T > t)
\]

- Defining the Hazard Rate

\[
h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T \leq t + \Delta t \mid T > t)}{\Delta t}
\]
The hazard in more details

The basic information in the hazard is, first of all, its qualitative behavior.
Some useful identities

- $h(t) = \frac{f(t)}{S(t)} \Rightarrow f(t) = h(t)S(t)$
- $h(t) = \left[-\log S(t)\right]'$
- $S(t) = \exp\left\{-\int_0^t h(s)ds\right\}$
- Define the cumulative hazard $H(t)$

$$H(t) = \int_0^t h(s)ds \Rightarrow S(t) = \exp\{-H(t)\} \text{ or } \log S(t) = -H(t)$$
By using the definition of conditional probabilities

$$\Pr(t < T \leq t + \Delta t \mid T > t) = \frac{\Pr([t < T \leq t + \Delta t] \cap [T > t])}{\Pr(T > t)}$$

$$= \frac{\Pr(t < T \leq t + \Delta t \mid T > t)}{\Pr(T > t)}$$

It may be helpful to sketch this relation graphically.

[Diagram of the relation]
An example
Survival Analysis: Non-Parametric Estimation

Dankmar Böhning

Southampton Statistical Sciences Research Institute
University of Southampton

S$^3$RI, 2 - 4 March 2015
General Concepts

Non-Parametric Estimation (no censoring)

Non-Parametric Estimation (including censoring)
Few remarks before starting

- Each subject has a beginning and an end anywhere along the time line of the complete study.
- In many clinical trials, subjects may enter or begin the study and reach end-point at vastly differing points.
- Each subject is characterized by
  1. Survival time
  2. Status at the end of the survival time (event occurrence or censored)
  3. The study group they are in.
Survival Analysis: Non-Parametric Estimation

General Concepts

Censoring

- The total survival time for that subject cannot be accurately determined.
  - A subject drops out, is lost to follow-up, or required data are not available
  - The study ends before the subject had the event of interest occur, i.e., they survived at least until the end of the study,
- There is no knowledge of what happened thereafter.
Censoring

- Right censoring: the period of observation expires, or an individual is removed from the study, before the event occurs.
- Left censoring: the initial time at risk is unknown.
- Interval censoring: both right and left censored
Estimation

- Random variable $T$ with cdf $F(t)$
- $S(t) = 1 - F(t)$
- With no censored observations:
  \[ \hat{S}(t) = 1 - \hat{F}(t) \]
- To estimate $F(t)$ at each time $t$:
  - data $t_1, \ldots, t_n$
  - parameter of interest $\theta = F(t) = \Pr(T \leq t)$
  - $\hat{\theta} = \frac{\# obs. \leq t}{n} = \frac{\sum_{i=1}^{n} I(0, t_i)(t)}{n}$
Confidence intervals

- Confidence interval for $F(t)$:
  \[
  \hat{\theta} \pm z_{\alpha/2} \sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}
  \]

- Confidence interval for $S(t)$:
  \[
  1 - \hat{\theta} \pm z_{\alpha/2} \sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}
  \]
Estimation

- To estimate the proportions $\theta_i$
  - $n_i =$ # of individuals at risk at the beginning of the $i$-th interval
  - $d_i =$ # of individuals experiencing the event
  
  $$\hat{\theta}_i = \frac{n_i - d_i}{n_i}$$

- Kaplan Meier estimator

  $$\hat{S}(t) = \prod_{i: t_i \leq t} \frac{n_i - d_i}{n_i}$$

- It reduces to $1 - \hat{F}(t)$ with no censored observations
## Example

<table>
<thead>
<tr>
<th>Subject</th>
<th>Group</th>
<th>Survival time in the interval</th>
<th># surviving at risk</th>
<th>Event</th>
<th># surviving after event</th>
<th>Cumulative survival rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>5</td>
<td>(1\times\frac{5}{6})</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>(1\times\frac{5}{6}\times\frac{4}{5})</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>(1\times\frac{5}{6}\times\frac{4}{5}\times\frac{3}{4})</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>(1\times\frac{5}{6}\times\frac{4}{5}\times\frac{3}{4}\times\frac{2}{3})</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>4.5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>(1\times\frac{5}{6}\times\frac{4}{5}\times\frac{3}{4}\times\frac{2}{3}\times\frac{1}{2})</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0.5</td>
<td>6</td>
<td>1</td>
<td>5</td>
<td>(1\times\frac{5}{6})</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>0.75</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>(1\times\frac{5}{6}\times\frac{4}{5})</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>(1\times\frac{5}{6}\times\frac{4}{5}\times\frac{3}{4})</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>1.5</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>(1\times\frac{5}{6}\times\frac{4}{5}\times\frac{3}{4}\times\frac{1}{2})</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>3.5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>(1\times\frac{5}{6}\times\frac{4}{5}\times\frac{3}{4}\times\frac{1}{2})</td>
</tr>
</tbody>
</table>
Example

Kaplan-Meier survival estimates

- Group 1
- Group 2

Survival analysis time
Understanding KM analysis

- The lengths of the horizontal lines represent the survival duration for that interval.
- The interval is terminated by the occurrence of the event of interest.
- The vertical distances between horizontal lines illustrate the change in the cumulative probability.
- The KM curve is a step-wise estimator, not a smooth function.
- What about estimate of point survival?
- Which is the effect of censoring?
Comparison of KM estimates

- It is simple to visualize the difference between two survival curves.
- The difference must be quantified in order to assess statistical significance.

Methods
- log-rank test ⇒ Most sensitive to consistent difference
- Wilcoxon test ⇒ Most sensitive to early differences
- hazard ratio ⇒ gives relative event rate in the groups
Log-Rank test: Example

<table>
<thead>
<tr>
<th>Time</th>
<th>Group 1 Event</th>
<th>Group 2 Event</th>
<th>Group 1 At Risk</th>
<th>Group 2 At Risk</th>
<th>Group 1 Expected</th>
<th>Group 2 Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td>0.55</td>
<td>0.45</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>4</td>
<td>1.20</td>
<td>0.80</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1.43</td>
<td>0.57</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0.80</td>
<td>0.20</td>
</tr>
<tr>
<td>3.5</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0.75</td>
<td>0.25</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4.5</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

The logrank test statistic is constructed by computing the observed and expected number of events in one of the groups at each observed event time and then adding these to obtain an overall summary across all time points where there is an event.

\[ \chi^2 = 3.07; \ p - value = 0.0798 \]
What to avoid

- Compare mean survival ⇒ Censoring makes this meaningless
- Overinterpret the tail of a survival curve ⇒ There are generally few subjects in tails
- Compare proportions surviving at a fixed time ⇒ Fine for description, not for hypothesis testing
Cox Proportional Hazards Regression for Survival Data

Dankmar Böhning

Southampton Statistical Sciences Research Institute
University of Southampton

S$^3$RI, 2 – 4 March 2015
Some simple distributions

The Cox PH model

Model diagnostics
Survival distributions

- Survival analysis focuses on the distribution of survival times.
- Although there are well known methods for estimating unconditional survival distributions, most interesting survival modeling examines the relationship between survival and one or more predictors.
- In principle, every distribution on $\mathbb{R}^+$ can serve to characterize survival data.
  - Constant hazard
  - Gompertz distribution
  - Weibull distribution
Cox Proportional Hazards Regression for Survival Data

Some simple distributions

Survival distributions

Modeling of survival data usually employs the hazard function

\[ h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t + \Delta t \mid T > t)}{\Delta t} \]

- Constant hazard: \( h(t) = \lambda \Rightarrow S(t) = e^{-\lambda t} \)
- Gompertz: \( h(t) = ae^{bt}, a > 0, b > 0 \Rightarrow S(t) = e^{a \left[ 1 - e^{bt} \right]} \)
- Weibull: \( h(t) = \lambda at^{a-1} \Rightarrow S(t) = e^{-\lambda t^a} \)
A parametric model based on the exponential distribution may be written as

\[ \log h_i(t) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \]

log-baseline hazard
The constant \( \beta_0 \) represents a kind of log-baseline hazard
The Cox model

The Cox model leaves the baseline hazard function
\[ \beta_0(t) = \log h_0(t) \] unspecified

\[ \log h_i(t) = \beta_0(t) + \beta_1x_{i1} + \cdots + \beta_px_{ip} \]

The model is semiparametric, because while the baseline hazard can take any form, the covariates enter the model linearly.

- The baseline hazard does not depend on covariates, but only on time
- The covariates are time-constant
- Proportional hazard assumption follows
The hazard ratio

For two observations $i$ and $j$, the hazard ratio

\[
\frac{h_i(t)}{h_j(t)} = \frac{h_0(t) \exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{h_0(t) \exp(\beta_1 x_{j1} + \cdots + \beta_p x_{jp})}
\]

\[
= \frac{\exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{\exp(\beta_1 x_{j1} + \cdots + \beta_p x_{jp})}
\]

\[
= \exp \left( \sum_{l=1}^{p} \beta_l (x_{il} - x_{jl}) \right)
\]

is independent of time $t$. Consequently, the Cox model is a proportional hazards model.
The hazard ratio: an example

- Only one covariate: Treatment
  - $x_i = 1 \Rightarrow$ Placebo
  - $x_j = 0 \Rightarrow$ Treatment
- Hazard ratio is then $\exp(\beta_1)$
- We expect that hazard is larger in the placebo group, i.e. the hazard ratio is expected greater than 1.
Time-constant covariates

- Not changing over time (e.g. gender)
- Values are set at time $t = 0$
- Variables unlikely to change are often considered time-constant
- Other variables are sometimes treated as time independent
- Time-dependent covariates are allowed, but PH assumptions is not satisfied (an extended Cox model is needed)
Advantages

- Robustness
- Because of the model form, the estimated hazards are always non-negative
- We can estimate fixed effects and compute the hazard ratio even though the baseline hazard is left unspecified
Checking proportional hazards

- Test and graphical diagnostic for PH may be based on scaled Schoenfeld residuals
- Influential observations
- Nonlinearity
Survival Analysis in STATA

Dankmar Böhning

Southampton Statistical Sciences Research Institute
University of Southampton

$S^3$RI, 2 - 4 March 2015
Introduction

Coding

Kaplan-Meier

PH Cox model
Aim

Illustrate how to use Stata to

- prepare survival data for analysis
- estimate hazard and survival functions
Data manipulation

A manipulation of the data is needed to facilitate summary and analysis.

help st
Assumptions

- Continuous time survival data
- Single failure data, i.e. one record per unit
- No complications such as truncation and/or missing values
- Data do not need to be weighted
Data structure

Data have a very simple structure

- One row per unit (e.g. subject)
- The survival time and the censoring status must be included as variables (1 = failure, 0 = otherwise)
- Covariates (explanatory variables) could be included
Data description

```
. use "C:\Users\user\Documents\Didattica\Southampton\SC_Epidem\lung.dta", clear

Contains data from C:\Users\user\Documents\Didattica\Southampton\SC_Epidem\lung.dta
    obs: 228
    vars: 10  28 Jan 2013 11:16
    size: 3,648 (99.9% of memory free)

| variable name | storage type | display format | value label | variable label
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>inst</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>Institution code</td>
</tr>
<tr>
<td>time</td>
<td>int</td>
<td>%8.0g</td>
<td></td>
<td>Survival time in days</td>
</tr>
<tr>
<td>status</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>censoring status 0=censored, 1=dead</td>
</tr>
<tr>
<td>age</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>Age in years</td>
</tr>
<tr>
<td>sex</td>
<td>byte</td>
<td>%8.0g</td>
<td>Male=1 Female=2</td>
<td></td>
</tr>
<tr>
<td>phecog</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>ECOG performance score (0=good 5=dead)</td>
</tr>
<tr>
<td>ptkarno</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>Karnofsky performance score (bad=0-good=100) rated by physician</td>
</tr>
<tr>
<td>patkarno</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>Karnofsky performance score as rated by patient</td>
</tr>
<tr>
<td>mealcal</td>
<td>int</td>
<td>%8.0g</td>
<td></td>
<td>Calories consumed at meals</td>
</tr>
<tr>
<td>wtloss</td>
<td>byte</td>
<td>%8.0g</td>
<td></td>
<td>Weight loss in last six months</td>
</tr>
</tbody>
</table>
```
stset declares the data in memory to be *st* data

- **Main**
  - Time variable \(\Gamma\) survival time
  - Failure variable \(\Rightarrow\) censoring status

- **Options**
  - Origin time expression sets when a subject becomes at risk
  - Enter time expressions specifies when a subject first comes under observation
  - Exit time expression specifies the latest time under which the subject is both under observation and at risk.
stset in practice

![Image of stset dialog box]

- Time variable: `time`
- Failure event: `status`
- Options:
  - Multiple-record ID variable:
  - Do not show st setting information
  - Clear all settings
**stset in practice**

![Stata stset dialog box](image_url)
stset: example

. stset time, failure(status)
   
   failure event:  status != 0 & status < .
   obs. time interval:  (0, time]
   exit on or before:  failure

228  total obs.
   0  exclusions

228  obs. remaining, representing
165  failures in single record/single failure data
69593  total analysis time at risk, at risk from t = 0
      earliest observed entry t = 0
      last observed exit t = 1022
Using `stset`

New variables in the data, why? Which is your meaning? Should you use them?

- `inst`: Institution code
- `time`: Survival time in days
- `status`: Censoring status 0=censored, 1=dead
- `age`: Age in years
- `sex`: Male=1 Female=2
- `phecog`: ECOG performance score (0=good 5=dead)
- `phkarno`: Karnofsky performance score (bad=0-good=100)
- `patkarno`: Karnofsky performance score as rated by patient
- `mealcal`: Calories consumed at meals
- `wtloss`: Weight loss in last six months
- `_st`
- `_d`
- `_t`
- `_t0`
**Using `stset`**

- `_st` is a binary variable indicating cases included (1) or excluded (0) from the analysis.
- `_d` is a censoring indicator.
- `_t` is the survival time.
- `_t0` is the time at which units are observed to be at risk.
Using `stset`

```
. de _*
variable name  storage  display  value  variable label
            type      format   label
__st        byte   %8.0g
__d         byte   %8.0g
__t         int    %10.0g
__t0        byte   %10.0g

. sum _*
  Variable | Obs | Mean | Std. Dev. | Min | Max
-----------|-----|------|-----------|-----|-----
__st       | 228 | 1    | 0         | 1   | 1   
__d        | 228 | .7236842 | .4481588 | 0   | 1   
__t        | 228 | 305.2325 | 210.6455 | 5   | 1022
__t0       | 228 | 0    | 0         | 0   | 0   
```
Summary statistics

You must `stset` your data before using

- `stdescribe` produces a summary of the st data
- `stsum` summarizes survival-time data

```
. stdescribe
    . failure _d: status
        analysis time _t: time

<table>
<thead>
<tr>
<th>category</th>
<th>total</th>
<th>mean</th>
<th>min</th>
<th>median</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>no. of subjects</td>
<td>228</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>no. of records</td>
<td>228</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(first) entry time</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(final) exit time</td>
<td>305.2325</td>
<td>5</td>
<td>255.5</td>
<td>1022</td>
<td></td>
</tr>
<tr>
<td>subjects with gap</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time on gap if gap</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time at risk</td>
<td>69593</td>
<td>305.2325</td>
<td>5</td>
<td>255.5</td>
<td>1022</td>
</tr>
<tr>
<td>failures</td>
<td>165</td>
<td>.7236842</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

. stsum
    . failure _d: status
        analysis time _t: time

<table>
<thead>
<tr>
<th>time at risk</th>
<th>incidence rate</th>
<th>no. of subjects</th>
<th>25% Survival time</th>
<th>50% Survival time</th>
<th>75% Survival time</th>
</tr>
</thead>
<tbody>
<tr>
<td>69593</td>
<td>.0023709</td>
<td>228</td>
<td>170</td>
<td>310</td>
<td>550</td>
</tr>
</tbody>
</table>
```
Kaplan-Meier

- Simple single-spell type
- Right censoring
- No left censoring (truncation)
sts

Survival times are treated as observations on a continuous variable
- `sts list`
- `sts graph`
- `sts test`
- `sts generate`
sts list

- Summarize survival-time data
- Describe survival-time data
- Report incidence-rate comparison
- Tabulate Mantel-Haenszel rate ratios
- Tabulate Mantel-Cox rate ratios
- Person-time, incidence rates, and SMR
- Tabulate failure rates and rate ratios
- Create survivor, hazard, and other variables
- List survivor and cumulative hazard functions
- Test equality of survivor functions
- Life tables for survival data
- CIs for means and percentiles of survival time
### sts list: example

```stata
. sts list

       failure _d: status
      analysis time _t: time

<table>
<thead>
<tr>
<th>Time</th>
<th>Beg.</th>
<th>Total</th>
<th>Fail</th>
<th>Net</th>
<th>Lost</th>
<th>Survivor Function</th>
<th>Std. Error</th>
<th>[95% conf. Int.]</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>228</td>
<td>228</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.9956</td>
<td>0.0044</td>
<td>0.9693</td>
</tr>
<tr>
<td>11</td>
<td>227</td>
<td>227</td>
<td>3</td>
<td>0</td>
<td></td>
<td>0.9825</td>
<td>0.0087</td>
<td>0.9539</td>
</tr>
<tr>
<td>12</td>
<td>224</td>
<td>224</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.9781</td>
<td>0.0097</td>
<td>0.9481</td>
</tr>
<tr>
<td>13</td>
<td>223</td>
<td>223</td>
<td>2</td>
<td>0</td>
<td></td>
<td>0.9693</td>
<td>0.0114</td>
<td>0.9367</td>
</tr>
<tr>
<td>15</td>
<td>221</td>
<td>221</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.9649</td>
<td>0.0122</td>
<td>0.9311</td>
</tr>
<tr>
<td>26</td>
<td>220</td>
<td>220</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.9605</td>
<td>0.0129</td>
<td>0.9278</td>
</tr>
<tr>
<td>30</td>
<td>219</td>
<td>219</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.9561</td>
<td>0.0136</td>
<td>0.9200</td>
</tr>
<tr>
<td>31</td>
<td>218</td>
<td>218</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.9518</td>
<td>0.0142</td>
<td>0.9146</td>
</tr>
<tr>
<td>33</td>
<td>217</td>
<td>217</td>
<td>2</td>
<td>0</td>
<td></td>
<td>0.9430</td>
<td>0.0154</td>
<td>0.9038</td>
</tr>
<tr>
<td>35</td>
<td>215</td>
<td>215</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.9386</td>
<td>0.0150</td>
<td>0.8985</td>
</tr>
<tr>
<td>59</td>
<td>214</td>
<td>214</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.9242</td>
<td>0.0164</td>
<td>0.8832</td>
</tr>
<tr>
<td>60</td>
<td>213</td>
<td>213</td>
<td>2</td>
<td>0</td>
<td></td>
<td>0.9254</td>
<td>0.0174</td>
<td>0.8828</td>
</tr>
<tr>
<td>61</td>
<td>211</td>
<td>211</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.9211</td>
<td>0.0179</td>
<td>0.8776</td>
</tr>
<tr>
<td>62</td>
<td>210</td>
<td>210</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.9167</td>
<td>0.0183</td>
<td>0.8725</td>
</tr>
<tr>
<td>65</td>
<td>209</td>
<td>209</td>
<td>2</td>
<td>0</td>
<td></td>
<td>0.9074</td>
<td>0.0192</td>
<td>0.8622</td>
</tr>
<tr>
<td>72</td>
<td>207</td>
<td>207</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.9025</td>
<td>0.0196</td>
<td>0.8572</td>
</tr>
<tr>
<td>79</td>
<td>206</td>
<td>206</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.8991</td>
<td>0.0200</td>
<td>0.8521</td>
</tr>
<tr>
<td>81</td>
<td>205</td>
<td>205</td>
<td>2</td>
<td>0</td>
<td></td>
<td>0.8904</td>
<td>0.0207</td>
<td>0.8420</td>
</tr>
<tr>
<td>88</td>
<td>203</td>
<td>203</td>
<td>2</td>
<td>0</td>
<td></td>
<td>0.8816</td>
<td>0.0214</td>
<td>0.8321</td>
</tr>
<tr>
<td>92</td>
<td>201</td>
<td>201</td>
<td>1</td>
<td>1</td>
<td></td>
<td>0.8772</td>
<td>0.0217</td>
<td>0.8271</td>
</tr>
<tr>
<td>93</td>
<td>199</td>
<td>199</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.8728</td>
<td>0.0221</td>
<td>0.8221</td>
</tr>
<tr>
<td>95</td>
<td>198</td>
<td>198</td>
<td>2</td>
<td>0</td>
<td></td>
<td>0.8640</td>
<td>0.0227</td>
<td>0.8172</td>
</tr>
<tr>
<td>105</td>
<td>196</td>
<td>196</td>
<td>1</td>
<td>1</td>
<td></td>
<td>0.8596</td>
<td>0.0230</td>
<td>0.8073</td>
</tr>
<tr>
<td>107</td>
<td>194</td>
<td>194</td>
<td>2</td>
<td>0</td>
<td></td>
<td>0.8507</td>
<td>0.0236</td>
<td>0.7974</td>
</tr>
<tr>
<td>110</td>
<td>192</td>
<td>192</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.8463</td>
<td>0.0239</td>
<td>0.7925</td>
</tr>
<tr>
<td>116</td>
<td>191</td>
<td>191</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.8418</td>
<td>0.0242</td>
<td>0.7876</td>
</tr>
<tr>
<td>118</td>
<td>190</td>
<td>190</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.8374</td>
<td>0.0245</td>
<td>0.7827</td>
</tr>
<tr>
<td>122</td>
<td>189</td>
<td>189</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.8330</td>
<td>0.0247</td>
<td>0.7778</td>
</tr>
<tr>
<td>131</td>
<td>188</td>
<td>188</td>
<td>1</td>
<td>0</td>
<td></td>
<td>0.8283</td>
<td>0.0250</td>
<td>0.7729</td>
</tr>
</tbody>
</table>
```
sts graph
sts graph: example
sts graph: example
sts graph: example

Smoothed hazard estimate

analysis time
sts graph: stratification
sts graph: stratification

Kaplan–Meier survival estimates

sex = 1  sex = 2
sts test

Survival Analysis - Stata

Kaplan-Meier
### sts test

```stata
. sts test sex

    failure _d: status
    analysis time _t: time

Log-rank test for equality of survivor functions

<table>
<thead>
<tr>
<th>sex</th>
<th>Events observed</th>
<th>Events expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>112</td>
<td>91.58</td>
</tr>
<tr>
<td>2</td>
<td>53</td>
<td>73.42</td>
</tr>
<tr>
<td>Total</td>
<td>165</td>
<td>165.00</td>
</tr>
</tbody>
</table>

    chi2(1) = 10.33
    Pr>chi2 = 0.0013

. sts test sex, wilcoxon

    failure _d: status
    analysis time _t: time

Wilcoxon (Breslow) test for equality of survivor functions

<table>
<thead>
<tr>
<th>sex</th>
<th>Events observed</th>
<th>Events expected</th>
<th>Sum of ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>112</td>
<td>91.58</td>
<td>3148</td>
</tr>
<tr>
<td>2</td>
<td>53</td>
<td>73.42</td>
<td>-3148</td>
</tr>
<tr>
<td>Total</td>
<td>165</td>
<td>165.00</td>
<td>0</td>
</tr>
</tbody>
</table>

    chi2(1) = 12.47
    Pr>chi2 = 0.0004
```
Survival Analysis - Stata

PH Cox model

stcox
stcox: options for model checking
stcox: example

```
. stcox sex age, sch(global*) sca(local*)

    failure _d:  status
    analysis time _t:  time

Iteration 0:  log likelihood =  -750.12202
Iteration 1:  log likelihood =  -743.09465
Iteration 2:  log likelihood =  -743.07965
Iteration 3:  log likelihood =  -743.07965
Refining estimates:
Iteration 0:  log likelihood =  -743.07965

Cox regression -- Breslow method for ties

No. of subjects =        228  Number of obs    =        228
No. of failures =        165  Time at risk    =    69593
Log likelihood =  -743.07965  LR chi2(2) =      14.08
                       Prob > chi2 =     0.0009

|   _t  |   Haz. Ratio | Std. Err. |      z |     P>|z| | [95% Conf. Interval] |
|--------|--------------|-----------|--------|--------|----------------------|
|   sex  |    0.5989574  |  0.1003026 |  -3.06  |      0.002  |     0.431372, 0.8316487 |
|  age   |    1.0171580  |  0.0093802 |   1.84  |      0.065  |     0.9989388, 1.03571  |
```
**stphplot: model checking**
**stphplot: model checking**

![Diagram showing Cox PH model with survival probability calculations](image)
estat phtest: model checking
estat phtest: model checking

<table>
<thead>
<tr>
<th>Time:</th>
<th>rho</th>
<th>chi2</th>
<th>df</th>
<th>Prob&gt;chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>0.12535</td>
<td>2.52</td>
<td>1</td>
<td>0.1125</td>
</tr>
<tr>
<td>age</td>
<td>-0.02090</td>
<td>0.07</td>
<td>1</td>
<td>0.7851</td>
</tr>
<tr>
<td>global test</td>
<td>2.65</td>
<td>2</td>
<td>0.2659</td>
<td></td>
</tr>
</tbody>
</table>