Survival Analysis with Stata 2003

Conference Paper · March 2003

CITATIONS
0

READS
1,079

1 author:

Robert A Yaffee
New York University
47 PUBLICATIONS **96** CITATIONS

SEE PROFILE

Survival Analysis with STATA

Robert A. Yaffee, Ph.D.
Academic Computing Services
ITS
p. 212-998-3402
yaffee@nyu.edu
Office: 75 Third Avenue
Level C-3
2003

Outline

- 1. Outline
- 2. The problem of survival analysis
  - 2.1 Parametric modeling
  - 2.2 Semiparametric modeling
  - 2.3 The link between the two approaches
- 3. Basic Theory of Survival analysis
  - 3.1 The survivorship and hazard functions
    - the Survival function
    - the Cumulative hazard
    - the Hazard rate
  - 3.4 Censoring
    - 3.4.1 Right censoring
    - 3.4.2 Interval censoring
    - 3.4.3 Left censoring
- 4. Formatting and summarizing
  - survival data
- 5. Nonparametric models: Life Tables
- 6. Nelson-Aalen Cumulative Hazard rates
- 7. Semi-Parametric Models: The Cox Model
  - Derivation of the model
  - Fitting the model
  - Interpretation of coefficients
  - Assumptions
  - Tests of assumptions
- Recapitulation

Preparing survival data

- In this lecture we present methods for describing and summarizing data, as well as nonparametric methods for estimating survival functions.
  - 1. (st) Setting your data
    - 1.1 The purpose of the **stset** command
    - 1.2 The syntax of the **stset** command
    - 1.3 List some of your data
    - 1.4 **stdes**
    - 1.5 **stvary**
    - 1.6 Example: Hip fracture data
      - From Hosmer and Lemeshow

Describing the Survival Data

- The Kaplan-Meier product-limit estimator of the survivor curve
  - 2.1 The **sts graph** command
  - 2.2 The **sts list** command
  - 2.3 The **stsum** command
- 2.2 The Nelson-Aalen estimator of the cumulative hazard
- 2.3 Comparing survival experience
  - 2.3.1 The log-rank test
  - 2.3.2 The Wilcoxon test
  - 2.3.3 Other tests

The Problem of Survival Analysis

- We are studying time till an event
- The event may be the death of a patient or the failure of a system
- These are sometimes called event history studies or failure time models
- If we model the survival time without assuming statistical distributions pertain, this is called nonparameteric survival analysis.
  - In this case we use life tables analysis
- If we model the survival time process in a regression model and assume that a distribution applies to the error structure, we call this parametric survival analysis.

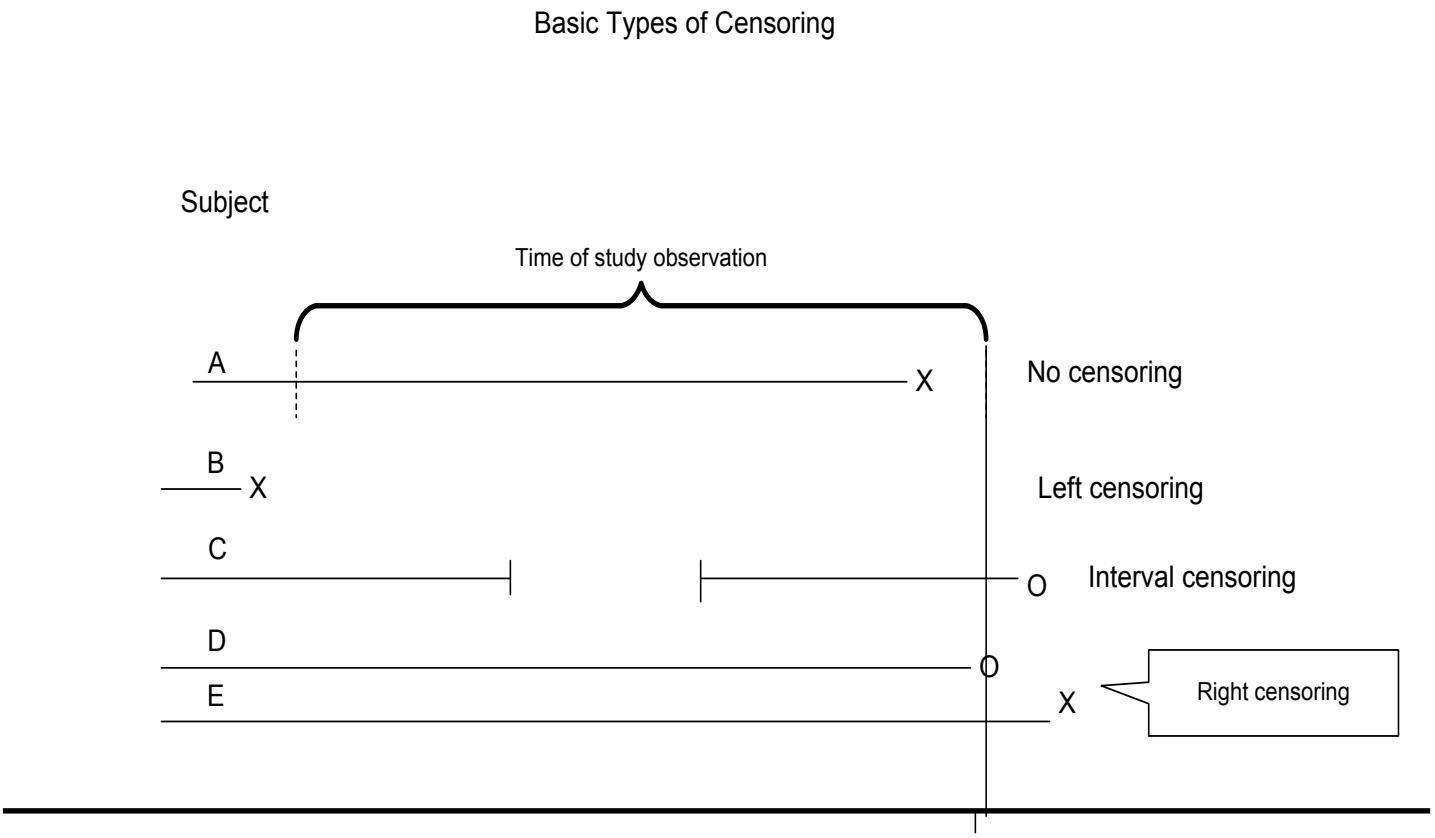
Censoring defined

1) Definition: Censoring occurs when cases are lost
2) What are the types:
    1) Left censoring: When the patient experiences the event in question before the beginning of the study observation period.
    2) Interval censoring: When the patient is followed for awhile and then goes on a trip for awhile and then returns to continue being studied.
    3) Right censoring:
        1) single censoring: does not experience event during the study observation period
        2) A patient is lost to follow-up within the study period.
        3) Experiences the event after the observation period
        4) multiple censoring: May experience event multiple times after study observation ends, when the event in question is not death.

Censored data

1) Definition: Data where the event beyond a particular temporal point was unobserved. The data within a particular range are reported at a particular limit of that range.
2) How it controls for the dropout
    1) The likelihood formula contains a probability factor that has an exponent of 1 when the event occurred and 0 when it was censored.
3) How we investigate it: We try to determine whether censoring is random or informative.

Censoring Depicted



Subjects D and E are right censored

Subject lost to follow-up not shown

Censoring and Truncation

- Truncation: Complete ignorance about the event of interest
- Left Truncation: Delayed entry
  - This could happen when the researchers do not administer the baseline interview before the patient dies

Survival Analysis Preprocessing

- The stset command
  - This command identifies the survival time variable as well as the censoring variable.
  - It sets up stata variables that indicate the entry, exit, and censoring time.

```
stset studytime, failure(died)
```

stset command

stset studytime, failure(died)

```
Notes:
      1.  (/m# option or -set memory-) 10.00 MB allocated to data
      2.  (/v# option or -set maxvar-) 5000 maximum variables

. use "D:\Admin\acs\lectures\Stata\Lectures\Survival\cancer.dta", clear
(Patient Survival in Drug Trial)

. stset studytime, failure(died)

     failure event:  died != 0 & died < .
obs. time interval:  (0, studytime]
 exit on or before:  failure

--------------------------------------------------------------------------
       48  total obs.
        0  exclusions
--------------------------------------------------------------------------
       48  obs. remaining, representing
       31  failures in single record/single failure data
      744  total analysis time at risk, at risk from t =         0
                            earliest observed entry t =         0
                                 last observed exit t =        39
```

Summary description of survival data set
stdes

- This command describes summary information about the data set. It provides summary statistics about the number of subjects, records, time at risk, failure events, etc.

Summary statistics about the total, mean, median, minimum and maximum of number of subjects, records, entry time, exit time, subjects with gap, time at risk and number of failure events.

stdes

```
. stdes

        failure _d:  died
  analysis time _t:  studytime
```

Category	total	per subject: mean	min	median	max
no. of subjects	48				
no. of records	48	1	1	1	1
(first) entry time		0	0	0	0
(final) exit time		15.5	1	12.5	39
subjects with gap	0				
time on gap if gap	0				
time at risk	744	15.5	1	12.5	39
failures	31	.6458333	0	1	1

```
.
```

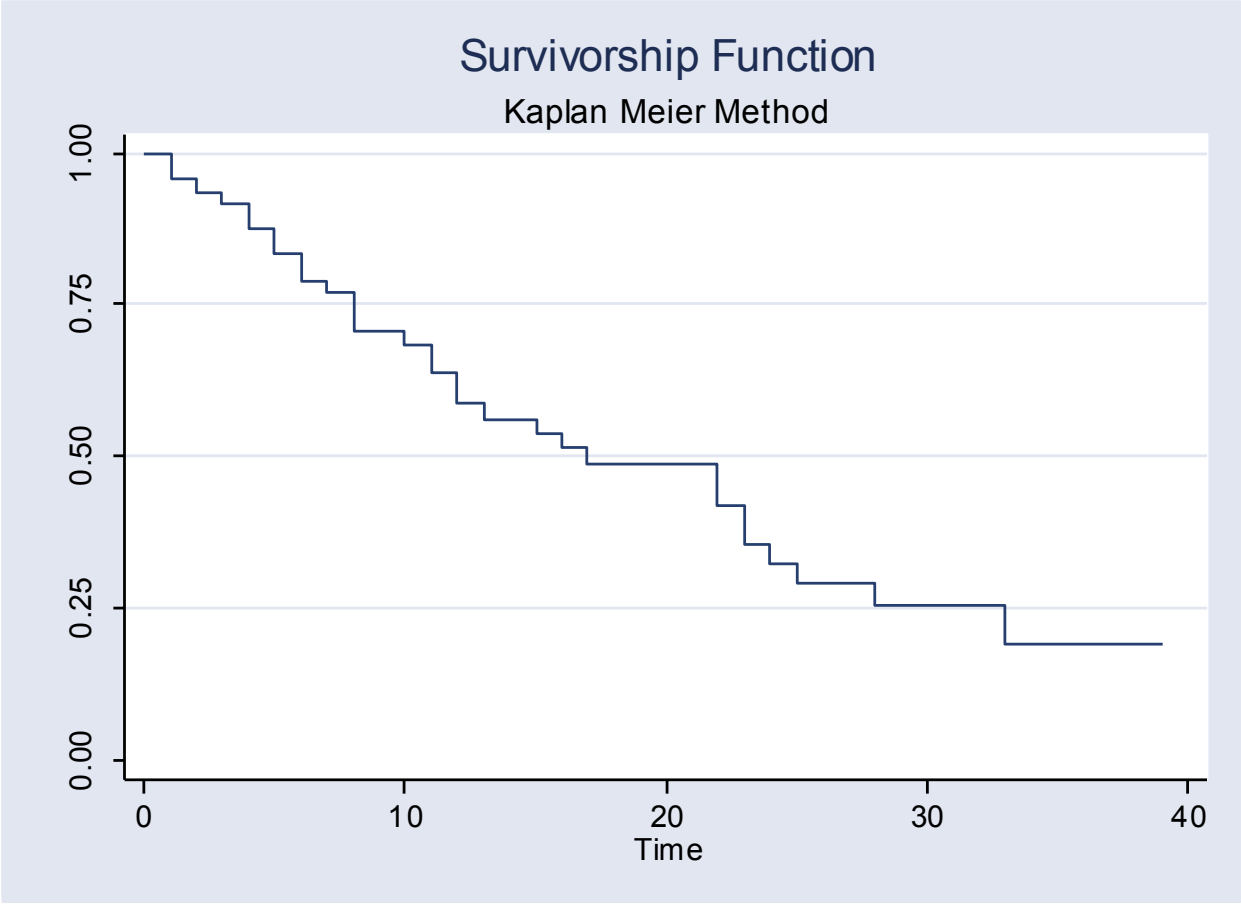
Describing the Survival Data
stsum stvary

```
. use cancer
(Patient Survival in Drug Trial)

. stset studytime, failure(died)

     failure event:  died != 0 & died < .
obs. time interval:  (0, studytime]
 exit on or before:  failure

------------------------------------------------------------------------------
       48  total obs.
        0  exclusions
------------------------------------------------------------------------------
       48  obs. remaining, representing
       31  failures in single record/single failure data
      744  total analysis time at risk, at risk from t =         0
                            earliest observed entry t =         0
                                 last observed exit t =        39

. stsum

         failure _d:  died
   analysis time _t:  studytime

                |                incidence      no. of |------ Survival time -----|
                |  time at risk     rate       subjects       25%       50%       75%
----------------+-------------------------------------------------------------------
          total |          744   .0416667          48         8        17        33

. stvary

         failure _d:  died
   analysis time _t:  studytime

                |              subjects for whom the variable is
       variable |  constant    varying      never missing  always missing  sometimes missing
----------------+-------------------------------------------------------------------------
           drug |        48          0             48             0             0
            age |        48          0             48             0             0

.
```

Graphing the data

Survival Probability of data set

sts graph, studytime is the stata command



As the study proceeds, this probability declines.

Basic Survival Analysis Theory

- We are interested in the Survivorship function S(t)
- The Survivorship function is a function of the probability of surviving plotted against time.
- We use the cancer.dta provided with STATA 7
- We graph the survivorship function

Computation of S(t)

1) Suppose the study time is divided into periods, the number of which is designated by the letter, *t*.
2) The survivorship probability is computed by multiplying a proportion of people surviving for each period of the study.
3) If we subtract the conditional probability of the failure event for each period from one, we obtain that quantity.
4) The product of these quantities constitutes the survivorship function.

Survival Function

- The survival probability is equal to the product of 1 minus the conditional probability of the event of interest.

$$S(t) = \prod_{t=1}^{T} (1 - h_i(t))$$

where

$S(t) =$ *estimated survivorship function at time* t

$h(t) =$ *conditional prob of event at time* t

Survival Function in Discrete Time

- The number in the risk set is used as the denominator.
- For the numerator, the number dying in period *t* is subtracted from the number in the risk set. The product of these ratios over the study time=

$$S(t) = \prod_{t(i) \le T} \frac{n_t - d_t}{n_t}$$

Survival Function and censoring

$$S(t) = \prod_{t(i) \le T} \frac{n_t - d_t}{(n_t - (c_t / 2))}$$

where c_t = *number censored in interval* t

The Survivorship Function

is the complement of the cumulative density function

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

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

F(t)=cumulative distribution of waiting time

The nature of the data

- The data are non-normal in distribution.
- They are right skewed.
- There may be varying degrees of censoring in the data.
- We have to use a nonparametric test to determine whether the survival curves are statistically different from one another.
- The early developers of tests include Mantel, Peto and Peto, Gehan, Breslow, and Prentice (Hosmer and Lemeshow, 1999).

The Structure of the Test

Table Testing Equality (homogeneity) of Survival Functions at Survival Time				
	Drug			
Event	*drug 1*	*drug 2*	*drug3*	***Total***
Die	d_1	d_2	d_3	d_i
Not die	N_1-d_1	N_2-d_2	N_3-d_3	N_i-d_i
At risk	N_1	N_2	N_3	n_i

Expected Value in the Table

$$e_i = \frac{n_i d_i}{n_i}$$

Tests for Equality across Strata

If $t_1<t_2<t_3<\ldots<t_k$ are the event times and $s=s_1,s_2,\ldots,s_c$ strata, then in this example c=3.

Then the test has the form:

$$Q_j = \frac{\sum_{i=1}^{k} (w_i(d_{ij} - \hat{e}_i))^2}{\sum_{i=1}^{k} w_i^2 {}_i v_i}$$

where

$$v_i = variance\, of\ d_i$$

$$w_i = weight \begin{cases} \log - rank, w_i = 1 \\ Gehan,\ w_i = n_i \\ Tarone - Ware, w_i = \sqrt{n_i} \end{cases}$$

Variance of d_i

$$Var_{jl} = \sum_{i=1}^{k} \frac{w_i^2 (n_i n_{il} \delta_{ji} - n_{ij} n_{il}) d_j s_i}{n_i^2 (n_i - 1)}$$

where

$i = event\ times$

$j = stratum$

$\delta_{jl} = 1\ if\ \ j = l, and\ 0\ otherwise$

$n_{ij} = size\ of\ \ risk\ \ set\ of\ \ jth\ stratum$

$$n_i = \sum_{j=1}^{c} n_{ij} \qquad s_i = n_i - d_i$$

$$d_j = \sum_{j=1}^{C} d_{ij}$$

The Weights w_i

- The Mantel Haenszel test or the Log-Rank test, developed by Peto and Peto in 1972, uses $w_i=1$.

- Gehan(1965) and Breslow(1970) generalized this test to allow for censoring. The weights $w_i=n_i$ the number of subjects at risk at each interval.

Standard Error of an Survival Function

Greenwood's formula

$$\hat{\sigma}(\hat{S}_i(t_i)) = \sqrt{\sum_{j=1}^{i} \frac{d_i}{n_j s_j}}$$

Examining the Survival Probability

Using the command, sts list, generates the survival table:

```
. sts list

        failure _d:  died
  analysis time _t:  studytime
```

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
1	48	2	0	0.9583	0.0288	0.8435	0.9894
2	46	1	0	0.9375	0.0349	0.8186	0.9794
3	45	1	0	0.9167	0.0399	0.7930	0.9679
4	44	2	0	0.8750	0.0477	0.7427	0.9418
5	42	2	0	0.8333	0.0538	0.6943	0.9129
6	40	2	1	0.7917	0.0586	0.6474	0.8820
7	37	1	0	0.7703	0.0608	0.6236	0.8656
8	36	3	1	0.7061	0.0661	0.5546	0.8143
9	32	0	1	0.7061	0.0661	0.5546	0.8143
10	31	1	1	0.6833	0.0678	0.5302	0.7957
11	29	2	1	0.6362	0.0708	0.4807	0.7564
12	26	2	0	0.5872	0.0733	0.4304	0.7145
13	24	1	0	0.5628	0.0742	0.4060	0.6931
15	23	1	1	0.5383	0.0749	0.3821	0.6712
16	21	1	0	0.5127	0.0756	0.3570	0.6483
17	20	1	1	0.4870	0.0761	0.3326	0.6249

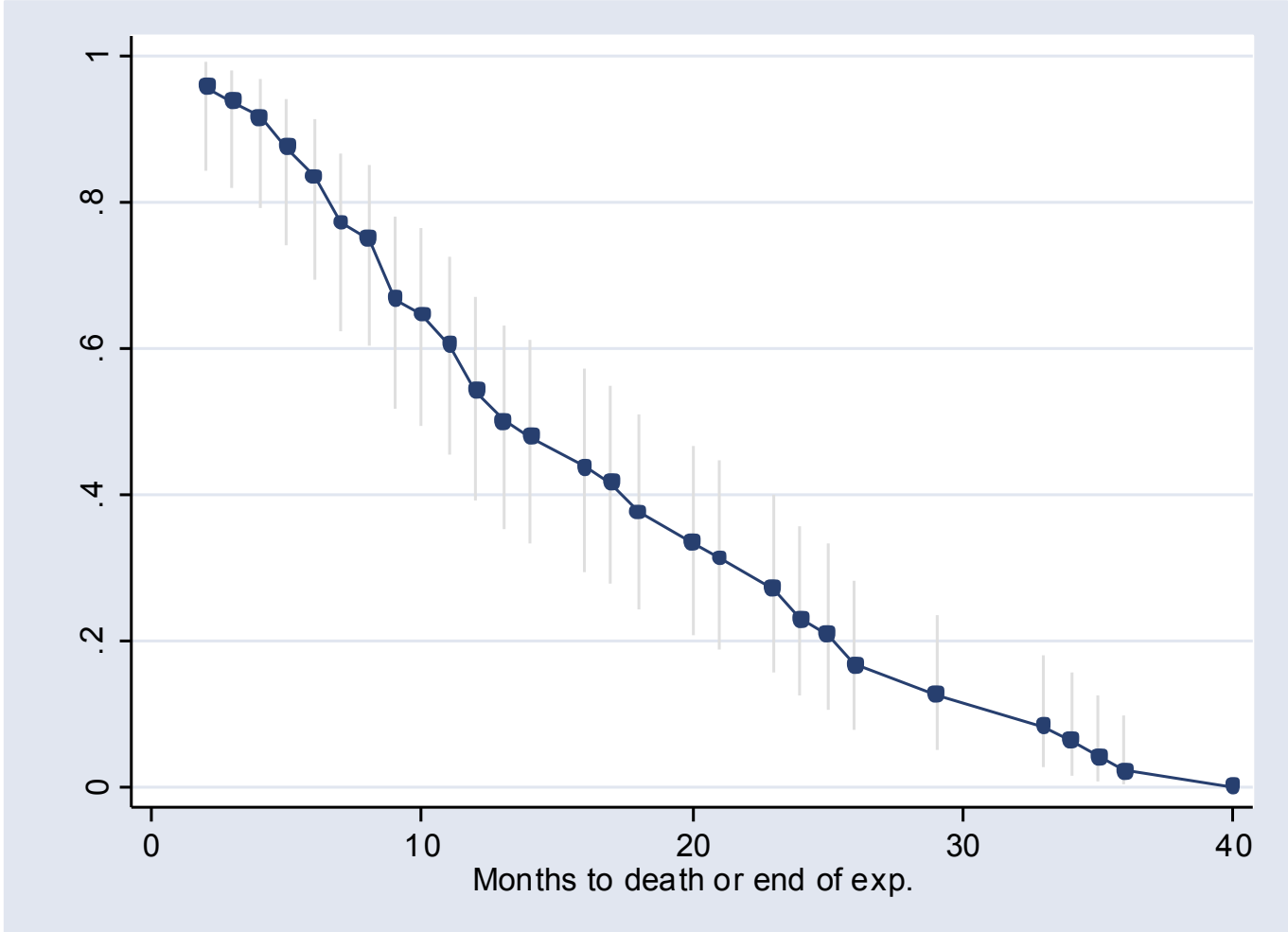
The Life Tables Analysis

```
. ltable studytime
```

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
1	2	48	2	0	0.9583	0.0288	0.8435	0.9894
2	3	46	1	0	0.9375	0.0349	0.8186	0.9794
3	4	45	1	0	0.9167	0.0399	0.7930	0.9679
4	5	44	2	0	0.8750	0.0477	0.7427	0.9418
5	6	42	2	0	0.8333	0.0538	0.6943	0.9129
6	7	40	3	0	0.7708	0.0607	0.6245	0.8660
7	8	37	1	0	0.7500	0.0625	0.6020	0.8495
8	9	36	4	0	0.6667	0.0680	0.5148	0.7807
9	10	32	1	0	0.6458	0.0690	0.4936	0.7628
10	11	31	2	0	0.6042	0.0706	0.4521	0.7262
11	12	29	3	0	0.5417	0.0719	0.3917	0.6696
12	13	26	2	0	0.5000	0.0722	0.3526	0.6307
13	14	24	1	0	0.4792	0.0721	0.3334	0.6110
15	16	23	2	0	0.4375	0.0716	0.2956	0.5707
16	17	21	1	0	0.4167	0.0712	0.2772	0.5503
17	18	20	2	0	0.3750	0.0699	0.2409	0.5087
19	20	18	2	0	0.3333	0.0680	0.2057	0.4661
20	21	16	1	0	0.3125	0.0669	0.1885	0.4445
22	23	15	2	0	0.2708	0.0641	0.1551	0.4003
23	24	13	2	0	0.2292	0.0607	0.1230	0.3549
24	25	11	1	0	0.2083	0.0586	0.1076	0.3317
25	26	10	2	0	0.1667	0.0538	0.0781	0.2840
28	29	8	2	0	0.1250	0.0477	0.0508	0.2344
32	33	6	2	0	0.0833	0.0399	0.0267	0.1821
33	34	4	1	0	0.0625	0.0349	0.0163	0.1545
34	35	3	1	0	0.0417	0.0288	0.0077	0.1257
35	36	2	1	0	0.0208	0.0206	0.0017	0.0958
39	40	1	1	0	0.0000	.	.	.

Graphing the survival probability

ltable studytime, graph



We need to develop tests that determine whether the survival rates are now statistically significantly different from one another

Stratifying the Survival Function

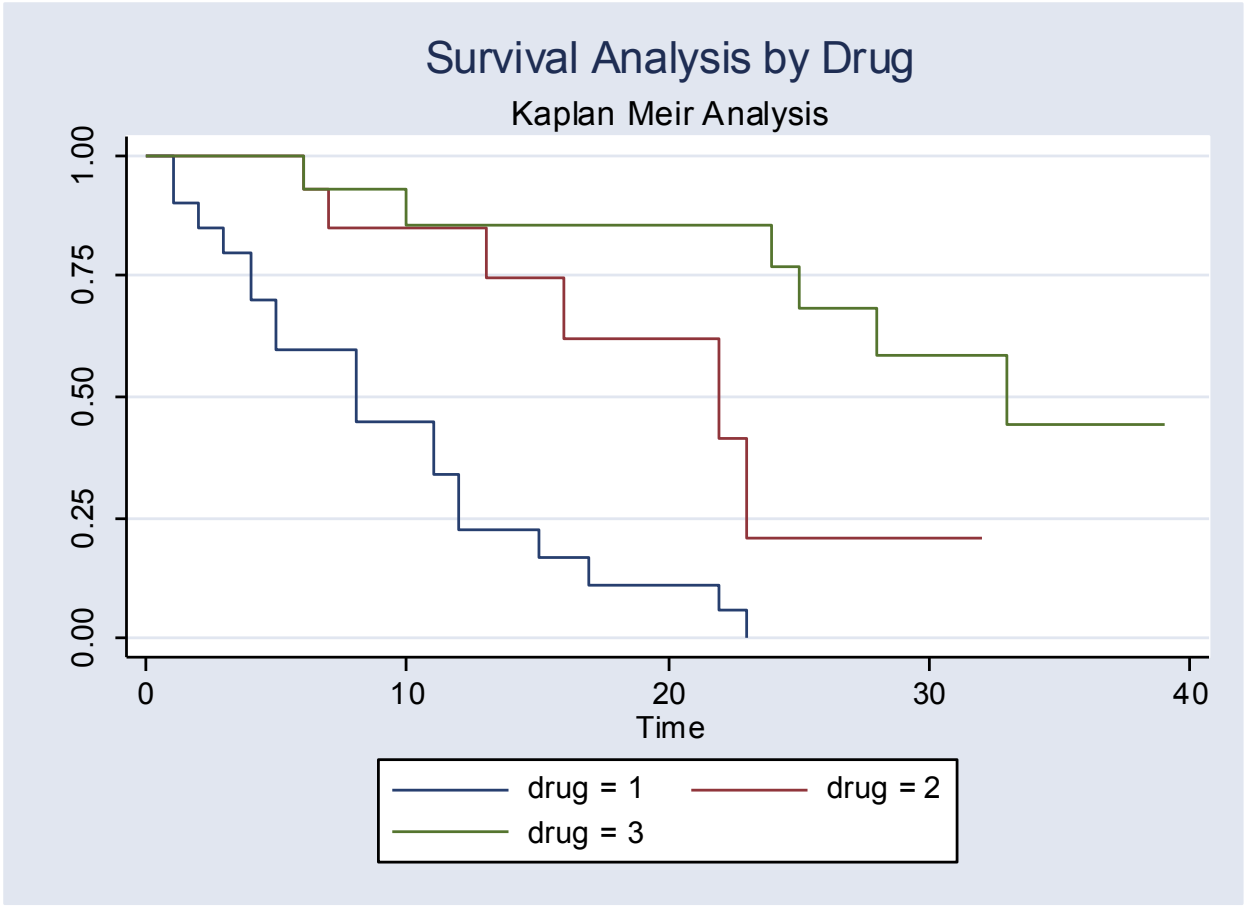
We test three drugs on the patients

If we were conducting a cancer clinical trial and were trying to slow down the impending death of terminally ill patients, we might test three different drugs.
The drugs in the three treatment arms of this clinical trial, we designate as drugs 1, 2, and 3.
We plot the survival functions of the three groups

Analyzing stratified survival rates

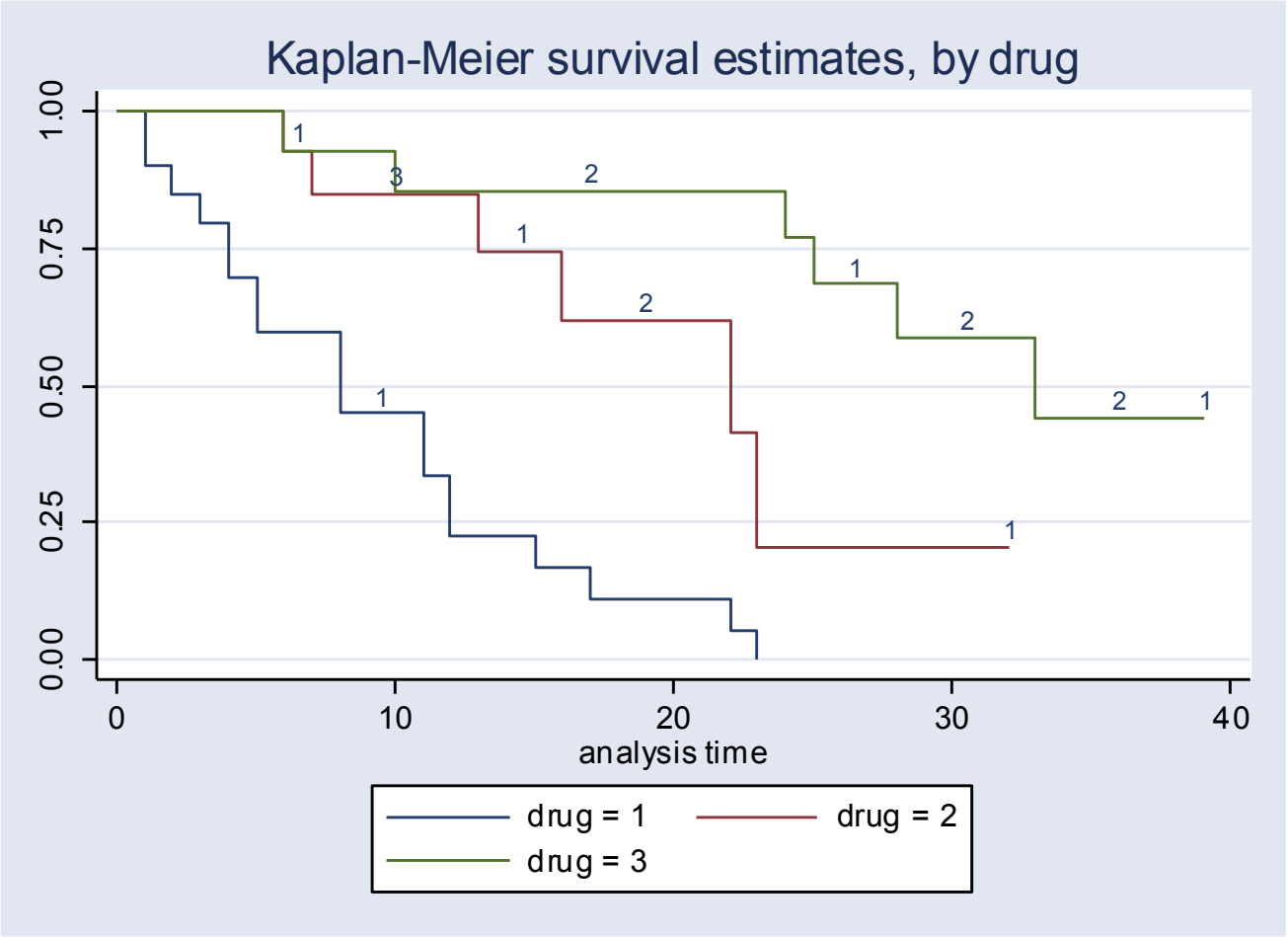
Stata command is

Sts graph, by(drug)



One can also identify the times of failure events in the survival estimates

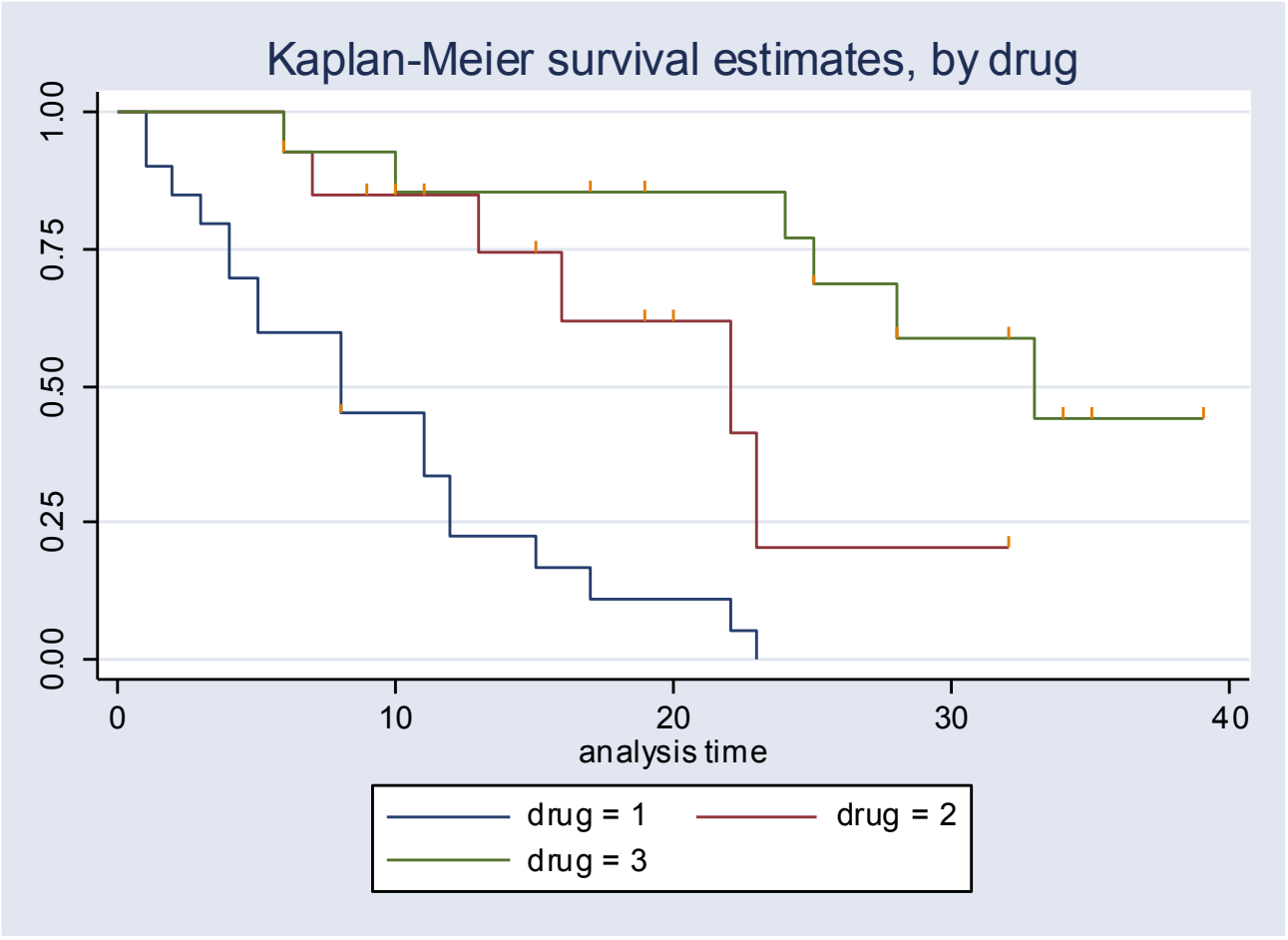
sts graph, by (drug) lost



Identifying the censored times

sts graph, by(drug) censored(single)

If there is multiple censoring, substitute multiple for single



Programming the Stratification Tests

sts test studytime, logrank strata(drug)

sts test studytime, wilcoxon

Logrank

```
. sts test studytime, logrank strata(drug)

         failure _d:  died
   analysis time _t:  studytime

Stratified log-rank test for equality of survivor functions
```

studytime	Events observed	Events expected(*)
1	2	0.20
2	1	0.16
3	1	0.21
4	2	0.68
5	2	0.96
6	2	0.21
7	1	0.15
8	3	2.93
9	0	0.15
10	1	0.30
11	2	2.12
12	2	2.63
13	1	0.28
15	1	1.85
16	1	0.45
17	1	2.05
19	0	0.59
20	0	0.45
22	2	3.18
23	2	4.68
24	1	0.25
25	1	0.72
28	1	1.00
32	0	1.78
33	1	0.75
34	0	0.75
35	0	0.75
39	0	0.75
Total	31	31.00

```
(*) sum over calculations within drug

                chi2(27) =      85.14
              Pr>chi2 =      0.0000
```

Wilcoxon

```
. sts test studytime, wilcoxon

         failure _d:  died
   analysis time _t:  studytime
```

Wilcoxon (Breslow) test for equality of survivor functions

studytime	Events observed	Events expected	Sum of ranks
1	2	0.08	92
2	1	0.06	43
3	1	0.09	41
4	2	0.26	76
5	2	0.36	68
6	2	0.69	50
7	1	0.26	26
8	3	1.36	52
9	0	0.34	-14
10	1	0.74	1
11	2	1.32	7
12	2	1.03	14
13	1	0.56	4
15	1	1.20	-19
16	1	0.65	-1
17	1	1.40	-26
19	0	1.40	-46
20	0	0.70	-23
22	2	1.67	-20
23	2	1.97	-28
24	1	1.08	-17
25	1	2.36	-48
28	1	2.61	-52
32	0	2.61	-60
33	1	1.55	-27
34	0	1.55	-31
35	0	1.55	-31
39	0	1.55	-31
Total	31	31.00	0

```
chi2(27) =    114.56
 Pr>chi2 =    0.0000
```

Other tests

Tarone-Ware Test

This test is the same as the Wilcoxon test, with the exception that the weight function $wt=n^{1/2}$.

The STATA command is:

sts test studytime, tware

Peto-Peto Prentice Test

The only difference between the Wilcoxon test and this one is that the weight function is approximately equal to the K-M survival Function

$$wt \approx \hat{S}(t)$$

Stata command for the Peto-Peto Prentice(1978) test is:

Sts test studytime, peto

The hazard rate

- The hazard rate is the conditional probability of the death, failure, or event under study, provided the patient has survived up to an including that time period.
- Sometimes the hazard rate is called the intensity function, the failure rate, the inverse Mills ratio (Cleves et al., 2002).

- When it is applied to continuous data, it is sometimes referred to as the instantaneous failure rate (Cleves et al., 2002).

Formulation of the hazard rate

$$h(t) = \lim_{\Delta t \to 0} \left(\frac{\Pr(t + \Delta t > T > t \mid T > t}{\Delta t} \right)$$

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

The hazard rate is known as the conditional rate of failure. This is the rate of an event, given that a person has survived up to that time. It is given by the above formula.

It can vary from 0 to infinity. It can increase or decrease or remain constant over time. It can become the focal point of much survival analysis.

Rising hazard rates augur increasing peril. Falling hazard rates portend greater security.

Examples of hazard rates

- Cleaves, Gould and Guttierrez suggest that human mortality declines after birth and infancy, remains low for awhile, and increases with elder years. This is known as the bathtub hazard function.

- They also note that post-operative hazard rate declines with the time after operation (CGG, p.8).

The Cumulative Distribution of the density function

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

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

$$F(t) = \Pr(t \le T) = \int_{t=0}^{T} f(u)dt$$

$$f(t) = \frac{dF(t)}{dt} = F'(t)$$

The probability density function

- The probability density function is obtained by differentiating the cumulative failure distribution.

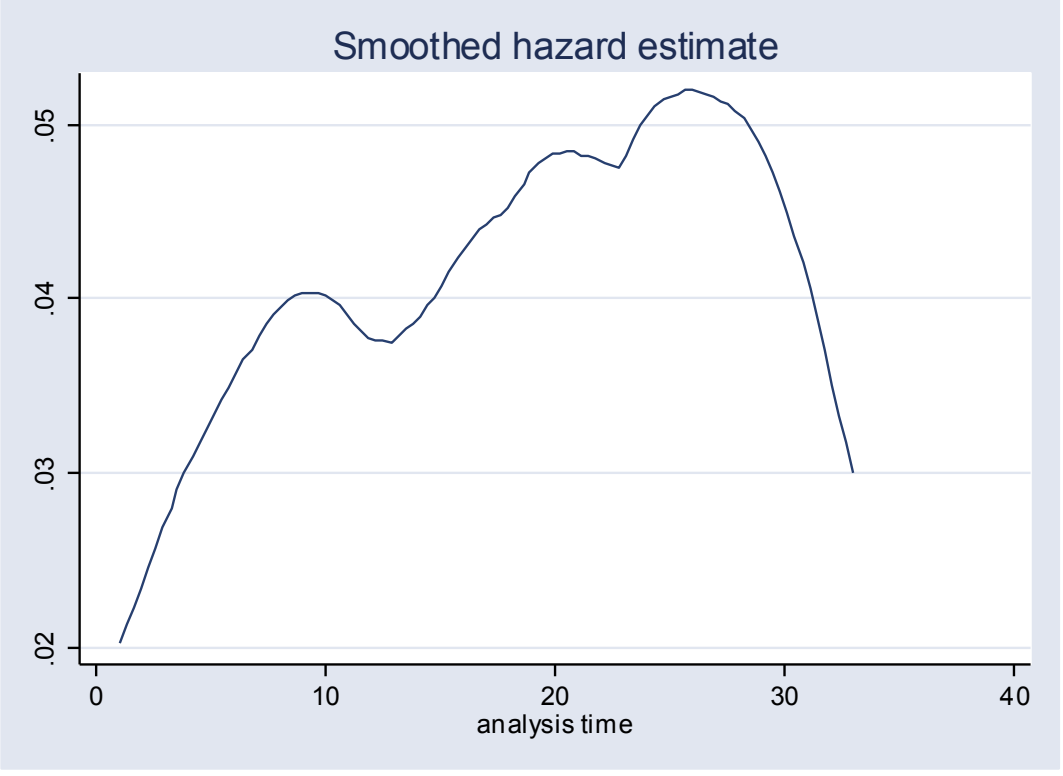
$$f(t) = \frac{dF(t)}{dt} = \frac{d(1 - S(t))}{dt} = -S'(t)$$

Programming the Survival Function

- The next few pages provide the preprocessing commands
- The Graphing Commands
- The testing commands for the survival function differences
- The menu options to use if you do not wish to use the commands

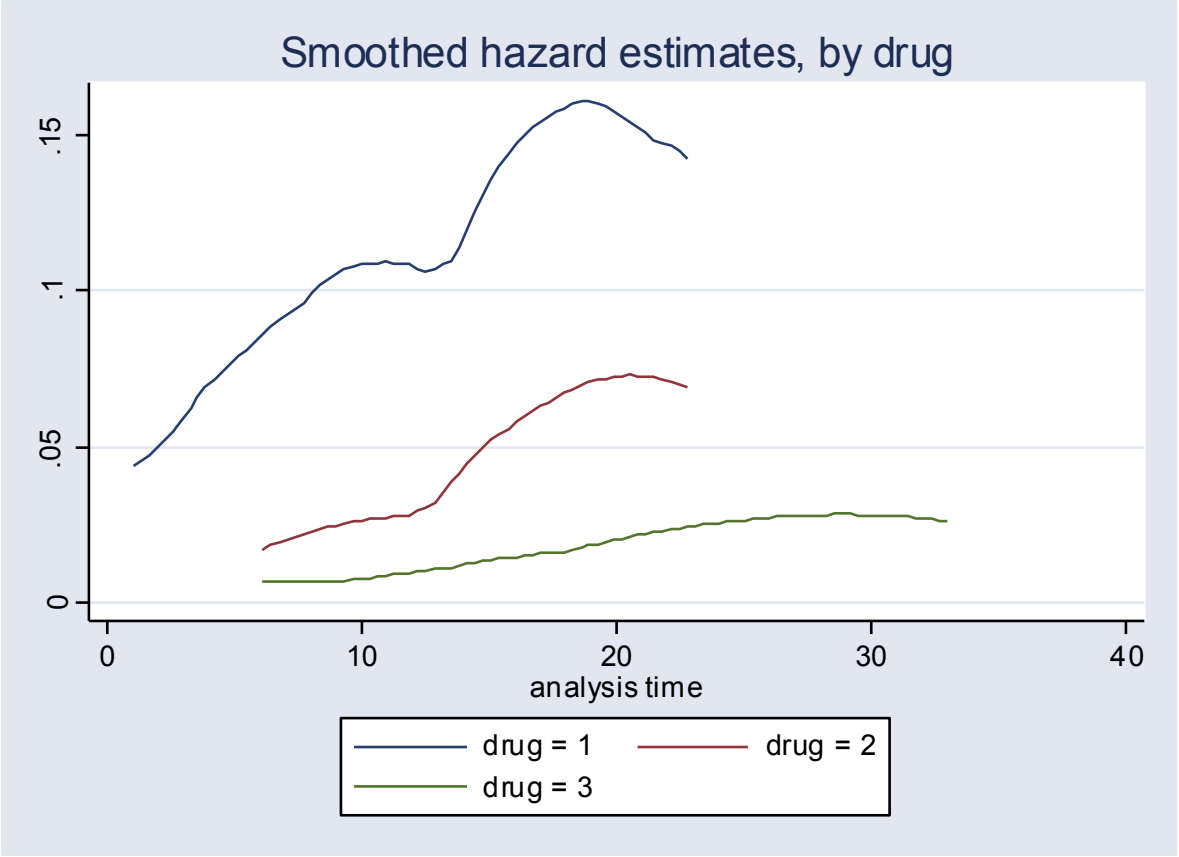
Graphing the hazard rate

sts graph, hazard



Graphing the respective hazard rates

- sts graph, by(drug) hazard

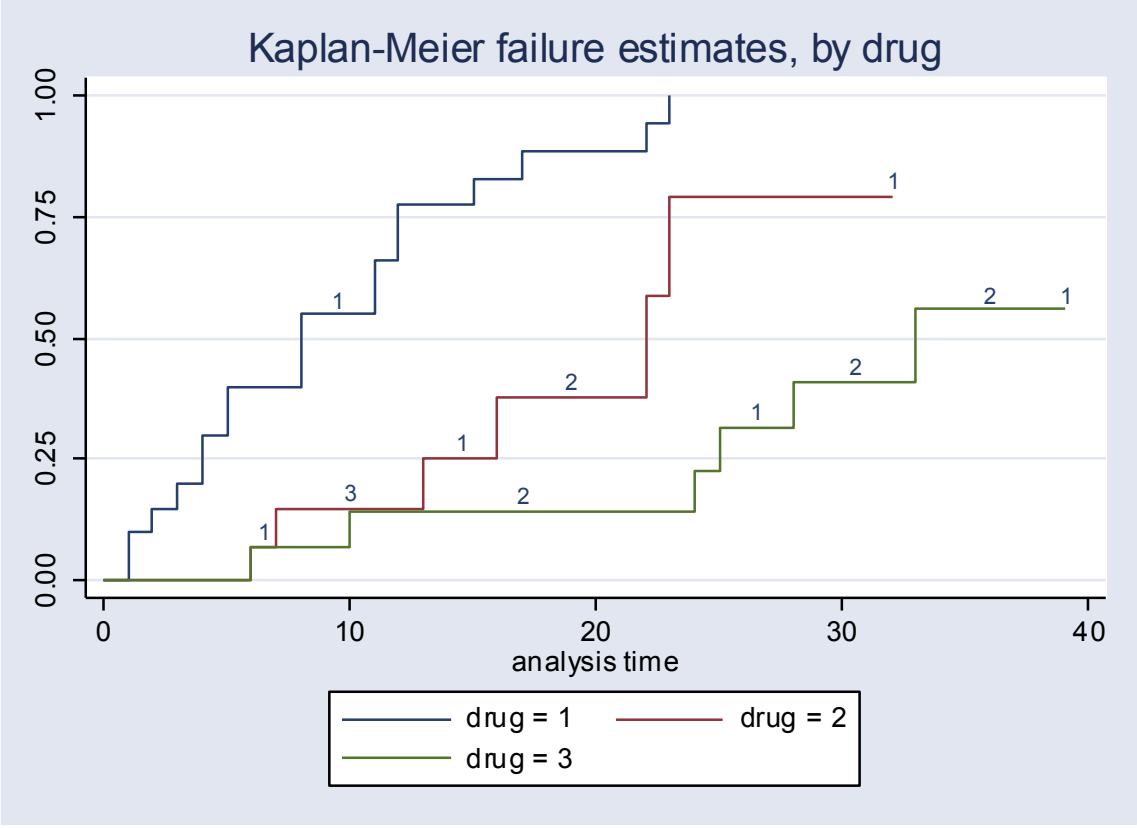


We will use the hazard rate as a dependent variable in the Cox models later.

Cumulative Probability of Failure

One can always graph F(t) with the following command:

sts graph, by (drug) failure lost



Nelson-Aalen Estimator

The Cumulative Hazard Function defined by Aalen in discrete time as

$$H(t) = \sum_{j \mid t_j \le t} \frac{d_j}{n_j}$$

dj= the number of failures at time j

nj = the number in the risk set at time j

Continuous Time version

$$H(t) = \int_0^t h(u)du$$

$$= \int_o^t \frac{f(u)}{S(u)}du$$

$$= \int_o^t \frac{1}{S(u)}\left\{\frac{dS(u)}{du}\right\}du$$

$$= -\ln\{S(t)\}$$

the Survival time as a function of the cumulative hazard function

$$H(t) = -\ln(\hat{S}(t))$$

$$\therefore \ln(\hat{S}(t)) = -H(t)$$

$$\therefore \hat{S}(t) = e^{-H(t)}$$

Let r be a function of the parameter vector.

$$H(t,x,\beta)=\int_0^t h(u,x,\beta)du$$

$$r(x,\beta)H_0(t)$$

$$if\ \ r(x,\beta)=e^{-(x,\beta)},$$

$$then:$$

$$S(t,x,\beta)=e^{-r(x,\beta)H_0(t)}$$

Listing data according to the Nelson-Aalen definitions

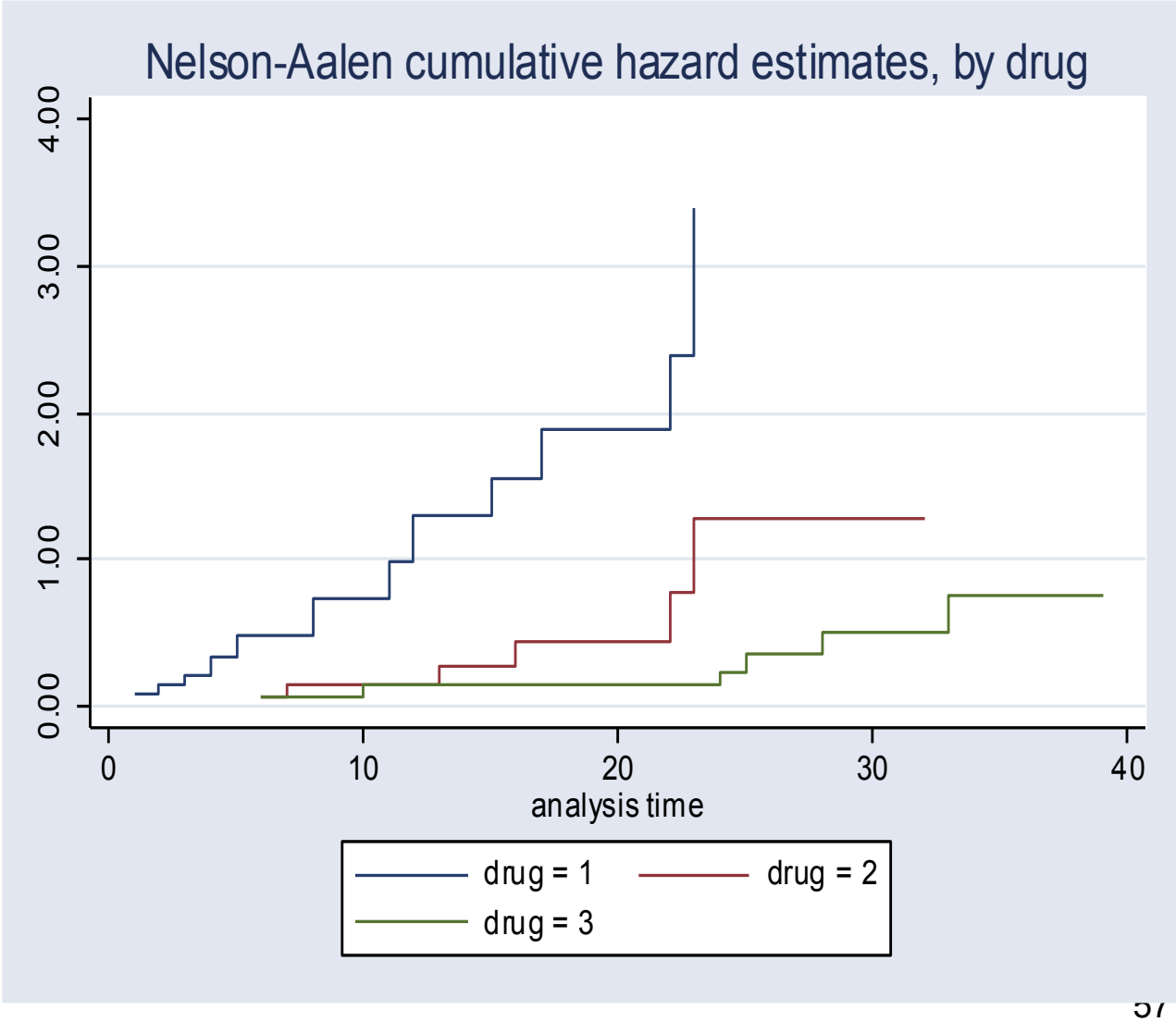
sts list, na

```
. sts list, na

        failure _d:  died
  analysis time _t:  studytime
```

Time	Beg. Total	Fail	Net Lost	Nelson-Aalen Cum. Haz.	Std. Error	[95% Conf.	Int.]
1	48	2	0	0.0417	0.0295	0.0104	0.1666
2	46	1	0	0.0634	0.0366	0.0204	0.1966
3	45	1	0	0.0856	0.0428	0.0321	0.2282
4	44	2	0	0.1311	0.0535	0.0589	0.2919
5	42	2	0	0.1787	0.0633	0.0893	0.3576
6	40	2	1	0.2287	0.0725	0.1229	0.4256
7	37	1	0	0.2557	0.0773	0.1414	0.4626
8	36	3	1	0.3391	0.0911	0.2003	0.5740
9	32	0	1	0.3391	0.0911	0.2003	0.5740
10	31	1	1	0.3713	0.0966	0.2230	0.6184
11	29	2	1	0.4403	0.1082	0.2719	0.7128
12	26	2	0	0.5172	0.1211	0.3268	0.8185
13	24	1	0	0.5589	0.1281	0.3566	0.8758
15	23	1	1	0.6024	0.1353	0.3879	0.9354
16	21	1	0	0.6500	0.1434	0.4218	1.0016
17	20	1	1	0.7000	0.1519	0.4575	1.0710
19	18	0	2	0.7000	0.1519	0.4575	1.0710

We may graph the cumulative hazard by the Nelson-Aalen definition

- sts graph, by (drug) na



Cox proportional hazards regression models

- Cox's proportional hazards method.
  - 1. Introduction
    - 1.1 The Cox model theory
    - 1.2 Interpreting coefficients
    - 1.3 The effect of units on coefficients
    - 1.4 The baseline hazard and related functions
    - 1.5 The effect of units on the baseline functions
    - 1.6 Summary of **stcox** command

    - 2.1 Indicator variables
    - 4.2 Categorical variables
    - 4.3 Continuous variables
    - 4.4 Interactions
    - 4.5 Time-varying variables
    - 4.6 Testing the proportional-hazards assumption
    - 4.7 Residuals

  - 3. Stratified analysis
    - 3.1 Obtaining coefficient estimates
    - 3.2 Obtaining the baseline functions
    - 3.3 The calculation of results

Aliases

Proportional Hazards model

Proportional hazards regression model

Cox Proportional Hazards model

"The hazard functions are multiplicatively related and that their ratio is constant over the survival time (Hosmer and Lemeshow, 1999)."

Cox Regression

- The Cox model presumes that the ratio of the hazard rate to a baseline hazard rate is an exponential function of the parameter vector.

$$\frac{h(t)}{h_o(t)} = \exp(x'b)$$

We would like to ascertain what variables potentiate or diminish the hazard rate

If we make some assumptions we can set up a model that can answer these questions.

We have to assume that the proportional hazard remains constant.

$$\frac{h(t)}{h_0(t)} = \exp(X'B) = e^{b_1x_1 + b_2x_2 + \ldots + b_px_p}$$

We have to assume that the baseline is not important to our primary considerations in this model.

A relative risk model

$$hazard\ ratio(t, x_1, x_0) = \frac{h(t, x_1, \beta)}{h(t, x_0, \beta)}$$

$$= e^{\beta(x_1 - x_0)}$$

Hazard rate as an exponential function of the covariate vector

$$h(t, x) = h_0(t)\, e^{x'\beta}$$

We take the natural log of the equation

We can convert this model to a linear model by taking the natural log of the equation.

$$\ln(h(t)) = \ln(h_o(t)) + b_1x_1 + b_2x_2 + \ldots + b_px_p$$

The natural log of the baseline hazard rate can be considered a constant in the model. "This component expresses the hazard rate changes as a function of survival time, whereas the covariate vector expresses the natural log of the hazard rate as a function of the covariates (Hosmer and Lemewhow, 1999).".

When the hazard is logged, the coefficients are called the risk score.

Semi-Parametric model

- The baseline is not explicitly described

Derivation

$$h(t,x,\beta) = \frac{f(t,x,\beta)}{S(t,x,\beta)}$$

$$f(t,x,\beta) = h(t,x,\beta)S(t,x,\beta)$$

$$likelihood(\beta) = \prod_{i=1}^{n} \left\{ [h(t_i,x_i,\beta)S(t_i,x_i,\beta)]^c [S(t,x,\beta)]^{1-c} \right\}$$

$$= \prod_{i=1}^{n} \left\{ [h(t_i,x_i,\beta)]^c S(t_i,x_i,\beta)] \right\}$$

$$LogL(\beta) = \sum_{i=1}^{n} c[h_0(t_i)] + c_i x_i \beta + e^{x_i\beta} \ln(S_0(t_i))]$$

When the individual is censored, the c=1 and when the individual is not censored c=0. This may change with the package, in LIMDEP, it is the opposite.

Partial Likelihood

The partial likelihood concentrates not on the baseline, but on the parameter vector of interest.

Let $R(t_i)$=risk set at time t_i with subjects whose survival or censored time are ge current time(H and L, p.98)

For the time being, it ignores censoring when c=0.

$$likelihood(\beta) = \prod_{i=1}^{n} \frac{e^{x_i'\beta}}{\sum_{j \in R(t_i)} e^{x_j'\beta}}$$

We take the ln of the expression

$$LL(\beta) = \sum_{i=1}^{n} c_i \left[x_i \beta - \ln \left\{ \sum_{j \in R(t_i)} e^{x_j \beta} \right\} \right]$$

where x_i = *value of covariate with ordered survival times*

Solving for beta

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} c_i \left\{ x_i - \frac{\sum_{j \in R(t_i)} x_j e^{x_j \beta}}{\sum_{j \in R(t_i)} e^{x_j \beta}} \right\}$$

$$= \sum_{i=1}^{n} c_i \left\{ x_i - \sum_{j \in R(t_i)} w_{ij}(\beta) x_j \right\}$$

$$= \sum_{i=1}^{n} c_i \left\{ x_i - \bar{x}_{iw} \right\}.$$

$$where \quad w_{ij}(\beta) = \frac{\sum_{j \in R(t_i)} e^{x_j \beta}}{\sum_{j \in R(t_i)} e^{x_j \beta}}$$

and

$$\overline{x}_{wi} = \sum_{j \in R(ti)} w_{ij}(\beta) x_j$$

Deriving the Standard Errors

- We take the 2nd derivative of the log likelihood to obtain the information matrix.

$$\frac{\partial^2 L(\beta)}{\partial \beta^2} = -\sum_{i=1}^{m} \sum_{j \in R(ti)} w_{ij} (x_j - \bar{x}_{wi})^2$$

$$= -I(\beta)$$

The variances of the variables are in the inverse of the information matrix.

$$Var(\hat{\beta}) = I(\hat{\beta})^{-1}$$

SE(β)

$$SE(\beta) = \sqrt{Var(\beta)}$$

Programming the Proportional Hazards model with stcox

```
stcox age drug, schoenfeld(sch*) scaledsch(sca*) nohr

failure _d:  censor
analysis time _t:  survtime

Iteration 0:   log likelihood = -299.19502
Iteration 1:   log likelihood = -281.73399
Iteration 2:   log likelihood = -281.70404
Iteration 3:   log likelihood = -281.70404
Refining estimates:
Iteration 0:   log likelihood = -281.70404

Cox regression -- Breslow method for ties

No. of subjects =          100                 Number of obs   =               100
No. of failures =           80
Time at risk    =         1136
LR chi2(2)      =           34.98
Log likelihood  =   -281.70404                  Prob > chi2     =   0.0000


    _t   |      Coef.        Std. Err.       z     P>|z|     [95% Conf.Interval]

Age   |    .0915319   .0184879     4.95   0.000     .0552963 .1277675
Drug  |    .9413856   .2555104     3.68   0.000     .4405943 1.442177


.  stphtest, plot(age) yline(0)

.  stphtest, plot(drug) yline(0)
```

Interpretation

- If the nohr option is invoked, the coefficients are the log hazard ratios, not the hazard ratios.
- If the option nohr is not used the hazard ratio is the dependent variable.

Modeling the Baseline Rate

- *There is no b_o and hence, there is no intercept in this model.*

- *When the x_i=0, then the relative hazard, exp(x'b) =1.*

Correction for Ties

Breslow's partial likelihood (adjustment for ties)

$$L_p(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)}\beta}}{\left[\sum_{j \in R_{t_i}} e^{x_{(i)}\beta} \right]^{di}}$$

$$d_i = number\ of\ subjects\ with\ survival\ time\ t(i)$$

Fitting the Cox Regression Model

1. We can fit these models according to the residual reduction.
2. We can fit these models according to the log likelihood.
3. The higher the – log likelihood, the better the model.
4. The larger the LR chi-square the better the model.

Partial Likelihood Ratio Test

G is the difference between the covariate model and the null model (constant only).

$$G = 2\{L_p(\hat{\beta}) - L_p(0)\}.$$

where

$$L_p(0) = \sum_{i=1}^{m} \ln(n_i)$$

This is distributed as a chi square with m df.

Interpretation of the Coefficients

1. *This depends on whether the dependent variable has been logged or not.*
2. *If the dependent variable has been logged, then a unit increase in the independent variable is associated with β increase in the log hazard rate.*
3. *If the dependent variable is the hazard ratio, so that the nohr has not been invoked, then a unit increase in the covariate is associated a e^{β} increase in the hazard ratio.*

For Example

```
.    stcox age

         failure _d:  censor
   analysis time _t:  survtime

Iteration 0:   log likelihood = -299.19502
Iteration 1:   log likelihood = -288.53901
Iteration 2:   log likelihood = -288.51804
Refining estimates:
Iteration 0:   log likelihood = -288.51804

Cox regression -- Breslow method for ties

No. of subjects =          100                 Number of obs    =        100
No. of failures =           80
Time at risk    =         1136
                                               LR chi2(1)       =      21.35
Log likelihood  =   -288.51804                 Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.084809   .0189139     4.67   0.000     1.048365    1.12252
------------------------------------------------------------------------------

. stcox age, nohr

         failure _d:  censor
   analysis time _t:  survtime

Iteration 0:   log likelihood = -299.19502
Iteration 1:   log likelihood = -288.53901
Iteration 2:   log likelihood = -288.51804
Refining estimates:
Iteration 0:   log likelihood = -288.51804

Cox regression -- Breslow method for ties

No. of subjects =          100                 Number of obs    =        100
No. of failures =           80
Time at risk    =         1136
                                               LR chi2(1)       =      21.35
Log likelihood  =   -288.51804                 Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0814042   .0174352     4.67   0.000     .0472317    .1155766
------------------------------------------------------------------------------
```

Significance tests of Coefficients

Wald statistic $z = \dfrac{\hat{\beta}}{SE(\hat{\beta})}$

Confidence Intervals for the hazard ratios

$$for\ dichotomous\ variables:$$

$$\exp(\hat{\beta} \pm 1.96SE(\hat{\beta})].$$

$$Categorical\ variables\ are\ dummied.$$

For continuous variables with c units change

$$= (x + c)\beta - x\beta$$

$$= \exp(c\hat{\beta} \pm z_{1-\alpha} c * SE(\hat{\beta}))$$

Time Varying Covariates

- The tvc (x3 x4 x5) option may be added to the model to specify time varying covariates.

For example,

stcox x1 x2, nohr tvc(x2)

Indicates that of the two covariates, the second is time-varying.

Testing the Adequacy of the model

1. We save the Schoenfeld residuals of the model and the scaled Schoenfeld residuals.
2. For persons censored, the value of the residual is set to missing.

Schoenfeld residuals

$$r_s = c_i\left(x_{ik} - \hat{\bar{x}}_{w_i k}\right)$$

where

$$\hat{\bar{x}}_{w_i k} = \frac{\sum x_{jk} e^{x_j'\hat{\beta}}}{\sum_{j \in R(ti)} e^{x_j'\hat{\beta}}}$$

Rescaled Schoenfeld Residuals

- m = number of uncensored survival times

$$r_{rs_i} = mVar(\hat{\beta})r_{s_i}$$

Creating the Residuals

```
stcox age drug, schoenfeld(sch*) scaledsch(sca*)
    nohr
```

Testing the Assumptions

- The hazard rates must be chosen so that h(t,x,b)>0.
- $h_0(t)$ characterizes the baseline hazard function, and this holds when x=0.

- The baseline hazard is a function of time and not of the covariates.

$$\ln(h(t,x,\beta) = \ln(h_0(t)) + x'\beta$$

An Objective Test

- stphtest, detail

```
. stphtest, detail

      Test of proportional hazards assumption

      Time:  Time
```

	rho	chi2	df	Prob>chi2
age	0.01317	0.01	1	0.9126
drug	-0.05248	0.21	1	0.6457
global test		0.25	2	0.8824

After the rescaled Schoenfeld residuals have been generated, this test may be conducted.

The detail option shows the individual as well as the global test of the proportional hazards assumption. NS results implies the proportional hazards assumption.

A graphical test of the proportion hazards assumption

- A graph of the log hazard would reveal 2 lines over time, one for the baseline hazard (when x=0) and the other for when x=1.
- The difference between these two curves over time should be constant= β

If we plot the Schoenfeld residuals over the line y=0, the best fitting line should be parallel to y=0.

Graphical tests

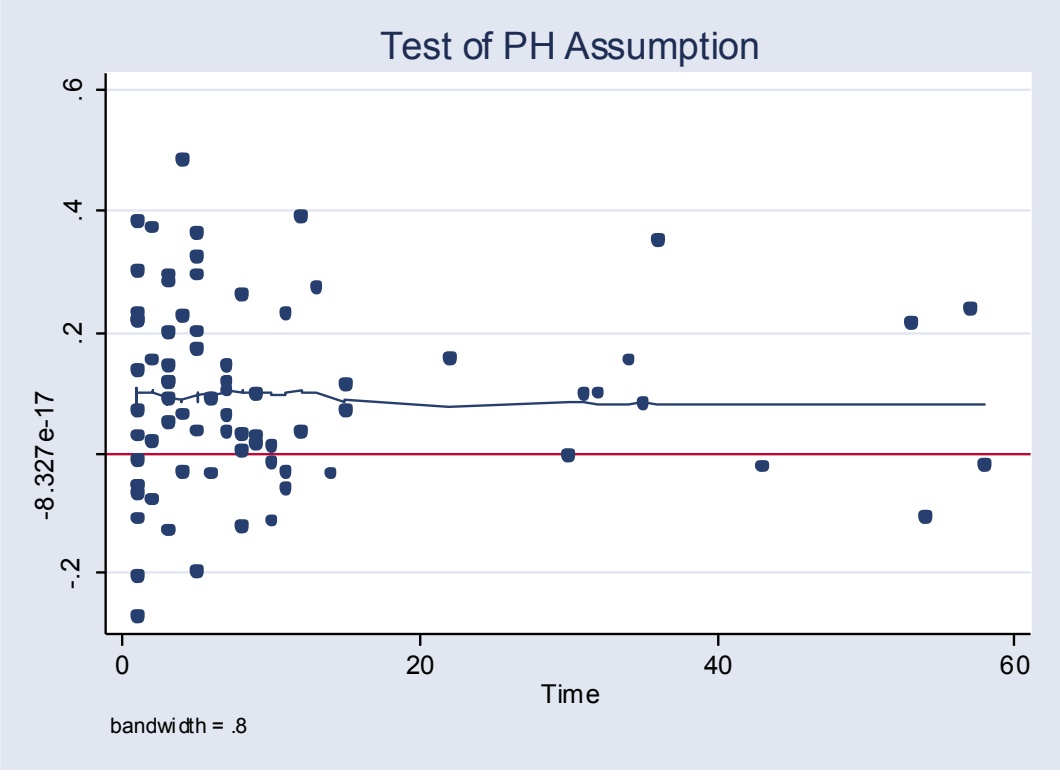
- Criteria of adequacy:

The residuals, particularly the rescaled residuals, plotted against time should show no trend(slope) and should be more or less constant over time.

Stphtest

- This tests the Schoenfeld residuals or the scaled Schoenfeld residuals against time.
- We hope to find that there is a level line that is close to 0. If there is, then the proportional hazards assumption holds.
- The stata command after creating the Schoenfeld residuals to test age is:
- **stphtest, plot(age) yline(0)**

Graph created to test ph assumption re age



The Model is time dependent

- Because this model is time dependent, it can handle time varying covariates
- If we have categorical predictors, we may wish to recode them as dummy variables.

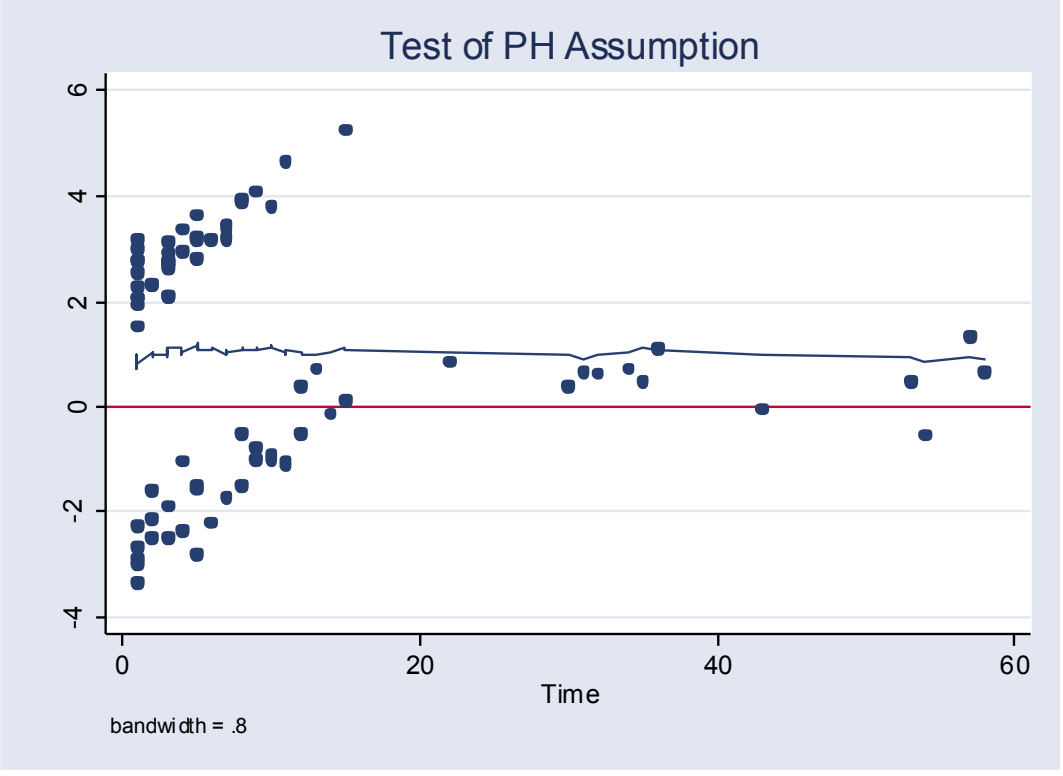
stphtest

- To test the drug use variable,
- The stata command is:

stphtest, plot(drug) yline(0)

This generates the following graph.

Test of Ph assumption with the Drug abuse variable



Other issues

- Time-Varying Covariates
- Interactions may be plotted
- Conditional Proportional Hazards models:
- Stratification of the model may be performed. Then the stphtest should be performed for each stratum.

References

Cleves, M., Gould, W.M., & Gutierrez, R.G. (2002). An Introduction to Suvival Analysis using Stata. College Station, Tex: Stata Press, pp.7, 34, .

Hoesmer, D. & Lemeshow, S. (1999). Applied Survival Analysis. New York: Wiley, pp. 58-65, 90.