

Linear Models in Stata: Survival analysis

Dr Cameron Hurst
cphurst@gmail.com

CEU, ACRO and DAMASAC, Khon Kaen University

25th June 2556



What we will cover....

- 1 Case study: The Worcester Heart Attack Study
- 2 Time-to-event outcomes
- 3 Descriptive statistics in survival analysis
 - Kaplan-Meier survival curves
 - Other descriptive statistics
- 4 Log-rank Test
- 5 Cox PH regression
 - A review of the Cox Ph model
 - Worked example of Cox PH regression: WHAS500
 - Disseminating the results
 - Assumptions of Cox PH regression

Conventions

Again, the conventions I will use:

Note:.....

Things to note will occur in a black box

Pitfalls:.....

Common mistakes and things to watch out for will occur in a red box

SYNTAX:.....

All Stata syntax will be in purple boxes and be in `courier` font. This will help you find it easily when you have to refer back to these notes.

Motivating example: Worcester Heart Attack Study

Background

- Before going into more detail about survival analysis, let's consider a real data set.
- Study considering factors and time trends associated with long-term survival following a myocardial infarction among residents of Worcester, Massachusetts
- In the first instance we will consider a trimmed down version of the dataset with several covariates (continuous and categorical) and a subset ($n=500$) of the large number of participants in this study.

Motivating example: Worcester Heart Attack Study

Variables in dataset

Covariates	Covariates (cont.)
Id	MI type (0=non-Q wave, 1=Q wave)
Age	Cohort year (1=1997, 2=1999, 3=2001)
Gender (0=male, 1=female)	Time and censoring variables
Initial heart rate	Admission date (a)
SBP	Discharge date (b)
DBP	Last follow-up date (c)
BMI	Length of stay (b-a)*
Family history of CVD(0=n,1=y)	Discharge status(0=alive, 1=dead)#
Atrial fibrillation(0=n,1=y)	Total length follow-up (c-a)*
Cardiogenic shock(0=n,1=y)	Status at last follow-up (0=alive, 1 dead)#
Congest. Heart compl(0=n,1=y)	
Complete heart block(0=n,1=y)	# censoring variables *analysis-time variables
MI order (0=first,1=recurrent)	

Time to event outcomes

- Unlike the other types of outcomes we model, time-to-event outcomes are a little different.
- Typically, they have two components:
 - (a) The quantity representing the time-to-event itself: **Survival time**
 - (b) Whether the event (e.g. death) actually occurred: **The censoring variable** (0=no[censored], 1=yes[died])
- The other types of outcomes we are used to (Binary outcomes, continuous outcomes, other types of categorical outcomes) only contain the first component.
- Hereafter, I will refer to Time-to-event outcomes as *Survival* outcomes (the most common time-to-event outcomes observed in clinical research)

Survival outcomes

To understand how these two parts of a survival outcome might be an advantage, let's consider an inappropriate analysis:

We sequentially enrol people who have experienced a myocardial infarction noting their treatment regimen and individual (patient) characteristics. The outcome of interest is whether the patient survived or not. The study ran for a two year period.

Here we have binary outcome (survival:yes/no) and a set of covariates → Binary logistic regression.

PROBLEMS???

Survival outcomes

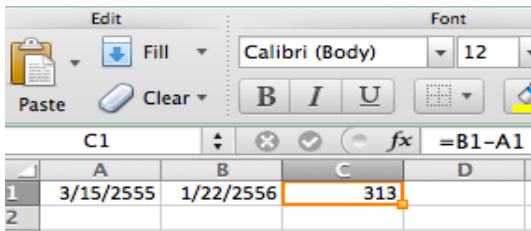
Answer: Time-at-risk. Those enrolled towards the end of the study have less 'exposure time' than those patients enrolled earlier. This would lead to biases in our results likely to render our analysis invalid (at best) and misleading (in the worst case)

The two components in time-to-event (survival) outcomes allow us to account for this time at risk.

BUT, it makes survival outcomes a little trickier to handle

Preparing survival outcomes

- If we are not provided with survival times we may have to calculate them using dates (e.g. Admission date, and Date of Death), using the differences between these dates.
- This is possible (but cumbersome) in Stata (I just use Excel)
- Note that you will have to use the US date format (mm/dd/yyyy), but can use Thai years (just be careful about leap years)



The screenshot shows the Microsoft Excel interface. The 'Edit' and 'Font' toolbars are visible at the top. The formula bar shows the formula `=B1-A1`. Below the formula bar, a table is displayed with columns A, B, C, and D. Row 1 contains the dates 3/15/2555 in column A and 1/22/2556 in column B. Cell C1 is highlighted with an orange border and contains the value 313, representing the difference between the two dates in Thai Buddhist Era (Buddhist Era 2555 corresponds to 2012 AD).

	A	B	C	D
1	3/15/2555	1/22/2556	313	
2				

stset: Setting up survival outcomes in stata

Recap:

Survival outcomes have two components:

- (a) **Continuous survival time**
- (b) **Binary censoring variable: [0=censored, 1=Died]**

We use the `stset` command in Stata set up our survival outcome:

SYNTAX: The `stset` command

```
*Open dataset(stata datafile)
use "c:\survWS\whas500.dta", clear

*Set up survival outcome
stset lenfol, failure(fstat)
```

In this example (from WHAS500 dataset), **lenfol** is the 'survival time' (Length of followup) and **fstat** is the 'censoring variable'.

Analysis of survival data

Like all of the other types of analysis, survival analysis can be used to generate:

- ① Descriptive statistics
- ② Bivariate methods (Specifically: Survival in terms of a single categorical covariate)
- ③ Multivariable models (e.g. Proportional Hazards [cox] regression)

The Kaplan-Meier survival curve

- Represents an estimator of the (true) survival curve
- aka the product-limit formula
- Purely empirical and non-parametric (makes no 'distributional' assumption about survival times)
- Accounts for censoring
- Generates the characteristic 'stair step' survival curves
- Does not account for confounding or effect modification by other covariates

Descriptive statistics: Kaplan-Meier survival curves

Once we have the data set up, analysis of survival data in Stata is quite simple. To generate a Kaplan-Meier (KM) survival curve:

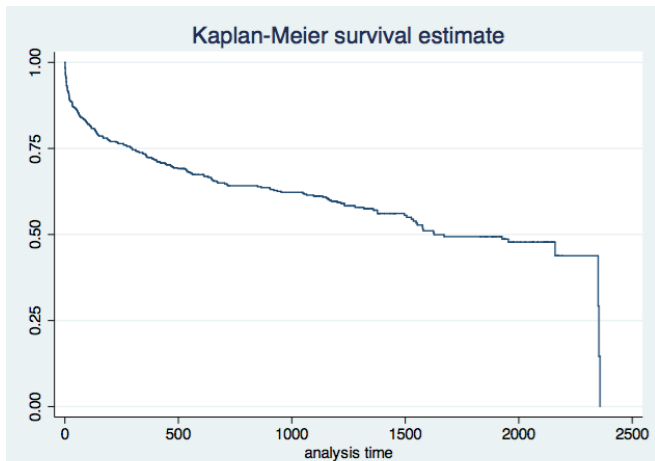
SYNTAX: KM survival curve(all individuals

```
*Generate a KM survival curve  
sts graph
```

The KM curve shows us what proportion of individuals are still alive (surviving) at any given time

KM curves

Figure : KM curve of all (WHAS500) individuals



Comparing KM curves of different groups

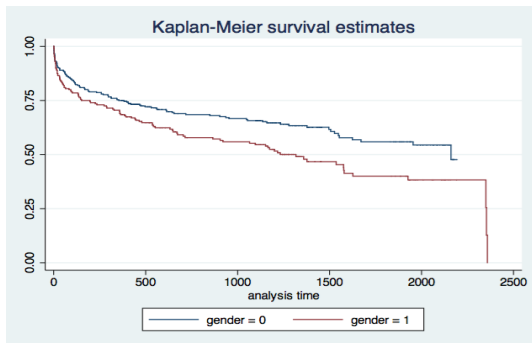
It is much more useful to compare the KM curves of two or more groups. Again, in Stata, this is very simple.

SYNTAX: Generating KM curves for multiple groups

```
*Generate a survival curves for males and females  
sts graph, by(gender)
```

Kaplan-Meier curves for each gender

Figure : Kaplan-Meier survival curves by sex



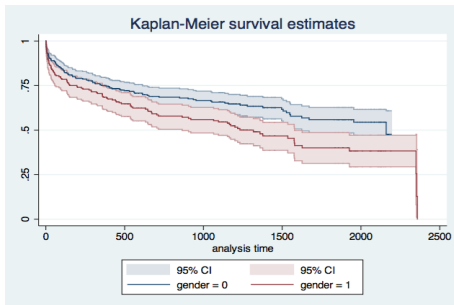
We can see that males[0] appear to have a more favorable survival experience than females[1]

Kaplan-Meier curves with Confidence intervals

SYNTAX: Generating KM curves with 95% CIs

```
*Generate a survival curves with CIs  
sts graph, by(gender) gwood
```

Figure : Kaplan-Meier curves with CIs



Median (and other quantile) survival time

SYNTAX: Generating medians and other quantiles

```
*Median(all individuals)
stci

*25 percentile (all individuals)
stci, p(25)

*Median (for each gender)
stci, by (gender)
```

Hint: Mean vs median survival time

As the distribution of survival times is often skewed, median (rather than mean) survival time is a better measure of the 'typical' survival time.

Median (and other quantile) survival time

Perusing the median survival times of the two genders...

```
. stci, by (gender)
```

```
      failure _d:  fstat  
      analysis time _t:  lenfol
```

gender	no. of subjects	50%	Std. Err.	[95% Conf. Interval]	
0	300	2160	.	1624	.
1	200	1317	177.0388	865	1579
total	500	1627	159.5555	1506	2353

As suggested by the KM curves, median survival time for males (300 individuals) seems better than females. 50% of males survived to day 2130, whereas, the median survival time for females was much lower (1317 days)

Question: Why is there no standard error for the males group??

Hint: Look at KM curve for males

The Log-Rank test

- Descriptive statistics are a nice way of perusing our sample, but more often we want to make statements (inferences) about the population (that align with our research objectives).
- For this we need inferential (hypothesis testing) methods.
- The first we will consider is the Log-Rank test. This test is:
 - Fully non-parametric (it makes no assumptions about the nature of the survival curve)
 - Bivariate: Only considers a single risk factor/treatment effect
 - Can only consider categorical predictors (i.e. "Groups")

The Log-Rank test in Stata

To perform a Log-Rank test in Stata:

SYNTAX: Testing equality of survival curves

```
*Log-rank test for gender  
sts test gender
```

This allows us to test the hypothesis:

H_0 : *Survival of males does not differ to that of females; vs*

H_A : *Survival of males differs from that of females*

The Log-Rank test in Stata

Running the syntax provided on the previous slide...

Log-rank test for equality of survivor functions

gender	Events observed	Events expected
0	111	130.73
1	104	84.27
Total	215	215.00

chi2(1) =	7.79
Pr>chi2 =	0.0053

We would conclude that there is a significant difference between the survival experience of males compared to females ($\chi_{LR} = 7.79$, $p = 0.0053$)

Limitations of the KM curves and Log-rank tests

- Both the Kaplan-Meier and Log-rank tests have the advantage of being non-parametric (distribution-free)
- But they have some distinct limitations:
 - They tell use very little about the magnitude of the association
 - They can't consider > 1 explanatory variable (at a time)
 - And this single explanatory variable must be categorical.
- We need a **modelling** approach for survival data
- There are a few, but we will consider that most commonly used: **Cox Proportional Hazards Regression**

Cox proportional hazards regression

- Most widely used methods for modelling time-to-event (survival) data
- Cox PH regression is **semi-parametric**; it makes no assumptions about the form of the survival distribution (or more specifically, the hazard distribution)
- However, nor is PH regression fully nonparametric as it makes assumptions about the functional form of the covariates (i.e. it is model based), and the central assumption of **proportionality** (more later)

The proportional hazards model

$$h_i(t) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}$$

where

$h_i(t)$ is hazard for subject i at time t

h_0 is the baseline hazard (common to all subjects)

$x_1, x_2 \dots, x_k$ are the subject-specific predictors; and

the exponential function, e , the inverse link function (i.e. a log link) implying a log-linear relationship between the hazard and the covariates

Recall our Worcester Heart Attack Study (n=500)

Covariates	Covariates (cont.)
Id	MI type (0=non-Q wave, 1=Q wave)
Age	Cohort year (1=1997, 2=1999, 3=2001)
Gender (0=male, 1=female)	Time and censoring variables
Initial heart rate	Admission date (a)
SBP	Discharge date (b)
DBP	Last follow-up date (c)
BMI	Length of stay (b-a)*
Family history of CVD(0=n,1=y)	Discharge status(0=alive, 1=dead)#
Atrial fibrillation(0=n,1=y)	Total length follow-up (c-a)*
Cardiogenic shock(0=n,1=y)	Status at last follow-up (0=alive, 1 dead)#
Congest. Heart compl(0=n,1=y)	
Complete heart block(0=n,1=y)	# censoring variables *analysis-time variables
MI order (0=first,1=recurrent)	

Modelling effect of MI type and BMI on survival

- First we will consider each covariate separately (i.e. in separate bivariate models).
- This will give us **crude** estimates of the hazard ratio for both risk factors
- Then we will include them in the same model (a multivariable model) to see:
 - Does taking both covariates into account (simultaneously) improve the model?
 - Does one covariate represent a confounder?
 - The risk factors will then be mutually adjusted → adjusted hazard ratios
- We should note that:
 - MI (0:non-Qwave; 1:Qwave) is categorical binary.
 - BMI is continuous and in this case is considered as time-independent

Cox regression in Stata

- As with most modelling in Stata, running a Cox PH Regression is very simple
- differs slightly as we don't have to specify the outcome (this is done is the `stset` command)
- We just have to specify the explanatory variables (risk factors, treatment effects etc)

SYNTAX: Cox regression for a continuous predictor

```
*Run Cox regression for BMI  
stcox bmi
```

Cox regression output: BMI

Figure : Stata output from Cox regression

```
Cox regression -- Breslow method for ties
```

No. of subjects =	500	Number of obs =	500
No. of failures =	215		
Time at risk =	441218		
Log likelihood =	-1203.497	LR chi2(1) =	48.16
		Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
bmi	.9064073	.0133633	-6.67	0.000	.8805906 .932981

Again, I will leave interpretation until the end

Cox regression: MI type

SYNTAX: Cox regression for a categorical predictor

```
*Myocardial infarction type  
stcox i.mitype
```

```
*Stata version <=10  
xi: stcox i.mitype
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
1.mitype	.5168868	.0864062	-3.95	0.000	.3724801	.7172786

Remember: non Q-wave MI type is the referent

Multivariable Cox regression

SYNTAX: MV Cox regression

```
*Multivariable Cox regression model
stcox bmi i.mitype
```

```
Log likelihood = -1196.7365      LR chi2(2)      =      61.69
                                Prob > chi2       =      0.0000
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
bmi	.9102218	.0133521	-6.41	0.000	.8844248	.9367713
1.mitype	.5550804	.0934396	-3.50	0.000	.399088	.7720459

Presenting the results

Similar to our logistic regression problem (yesterday), the results should be presented:

Table : Effects of BMI and Myocardial infarction type on survival

Effect	HR_{Crude}	$HR_{Adjusted}$	95%CI
BMI	0.91***	0.91***	0.88-0.94
MI type	0.52***	0.56***	0.40-0.77

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Interpreting results

- The overall (multivariable) model was highly significant ($\chi_{LR} = 61, 69, p < 0.001$)
- BMI was a significant predictor (of hazard) where (surprisingly) the chance of dying actually decreased by 9% every extra unit of BMI
- MI type was also a significant risk factor with those with Q-wave MIs having 44% less chance of dying, than those with Non Q-wave MIs (Referent)
- BMI confounds the MI effect (somewhat) with about a 7.5% change in the MI type effect ($HR_{Crude} = 0.52$ vs $HR_{Adjusted} = 0.56$) when BMI was adjusted for

Cox regression assumptions

There are two main assumptions associated with Cox regression:

- ① If continuous covariate are associated with the hazard than it is log-linearly related i.e. Is the $\text{Log}(\text{Hazard})$ linearly related to covariates
 - ② Hazards remain proportional over the whole survival experience
- This second assumption is central to Cox **Proportional Hazards** regression
 - It is key to using a Maximum Likelihood like estimator to get the HR
 - All it means is that the Hazard ratio (relative risk) remains constant (for all survival times).
 - If one group has twice the chance of dying at day 1 relative to the referent, than this holds for day 50, 100, 150 and so on

Proportionality of hazard for MI type

Let's think about the implications of proportionality for our research problem.

- Research indicates that short-term mortality is substantially higher (about 2 fold) in Q-wave MIs compared to non Q-wave MI BUT
- This situation is reversed for long term mortality (mortality higher in non Q-wave)
- What does this mean for our proportionality assumption????

The relative hazard of Q-wave to non Q-wave is likely to change over the survival experience

We need to check if this is observed in our dataset.

Assessing the proportional hazard assumption

There are two main approaches to assessing the Proportional hazards assumption:

- ① The log minus log survival plot
- ② Schoenfeld residuals

Method 1: Log minus log survival plot

This is essentially the $\ln(-\ln(S(t)))$ for each level of the covariate (where $S(t)$ is the Kaplan-Meier estimate of the survival curve). Generally, if the log minus log survival curves are **parallel** for each level of the covariate (e.g. Q-wave and Non Q-wave) then the proportional hazard assumption is met.

Assessing proportionality in Stata

SYNTAX: Schoenfeld residuals and Log minus log plots

```
*Schoenfeld residuals test (Stata V12)  
stphtest, detail
```

```
*Schoenfeld residual test (Stata V10)  
xi: stcox bmi i.mitype, schoenfeld(sch*) scaledsch(sca*)  
stphtest, detail
```

```
*Log-log plots  
stphplot, by(mitype) adjust(bmi)
```

Results: Schoenfeld residuals (test)

Test of proportional-hazards assumption

Time: Time

	rho	chi2	df	Prob>chi2
bmi	0.08717	2.05	1	0.1520
mitype	-0.07112	1.10	1	0.2944
global test		2.96	2	0.2271

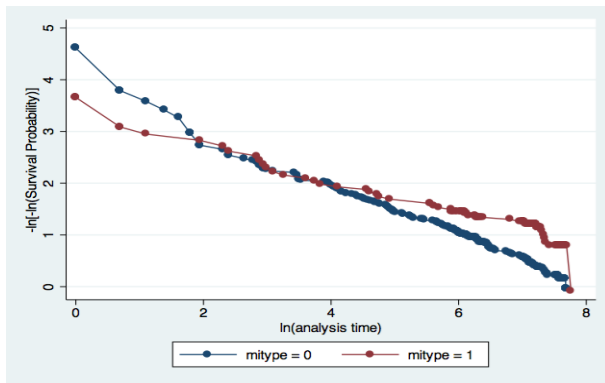
- Neither BMI or MI type seem to violate the proportionality assumption ($p_{BMI} = 0.152$ and $p_{MItype} = 0.294$ respectively)
- Nor was there an overall (global) departure from proportionality exhibited by the (overall) model ($p_{model} = 0.227$)

Note: Schoenfeld p-values

We are looking for a non-significant result for the schoenfeld residual test (i.e. NO significant departure from proportionality)

Results: Log minus log plot

Figure : Log minus log (Mltype)



Looking for parallel lines here. The cross-over is of some concern

THANK-YOU!!

Questions??

YOUR TURN

Exercises

Using the WHAS500 data:

- The analysis we conducted in this lecture was based on total follow up time (including after they were discharged from hospital).
- We are now interested in assessing the survival of the MI patients ONLY while they were in hospital (Hint: survival time = **los**, died in hospital (0=no, 1=yes) = **dstat**)....**You will have to run the `stset` command using these variables**
- Replicate the analysis we conducted in the lecture on this new survival outcome
- In the modelling step (Cox regression) feel free to consider other additional covariates