Survival Analysis in R and Stata

Dr Cameron Hurst
cphurst@gmail.com

DAMASAC and CEU, Khon Kaen University

25th August, 2558

What I will cover....

In R and Stata

- Reading in data and 'setting up' survival outcome variables
- Kaplan-Meier curves
- Basic summary statistics
- Classical tests: the Log-Rank test
- Modeling survival outcomes using Cox proportional hazards regression
  - Fitting the model
  - βs and Hazard ratios (and their CIs)
  - Checking proportionality assumption

Conventions

Note:.....

Things to note will occur in a green box

Pitfalls:.....

Common pitfalls and mistakes in a red box

R SYNTAX:....

Most (important) R 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.

Stata SYNTAX:....

Most (important) Stata syntax will be in blue boxes and also be in `courier` font.

Motivating example

- Recall the Worchester 500 dataset I identified in the Intro to Survival Analysis session
- I will use this dataset (the WHAS500 data) throughout all of my examples

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)	

Fortunately you don't have to worry about the painful aspect of dealing with "Date" data.....they are already

Data preparation: R

To read data into R is done in the usual way...

Reading in data

```
library(survival)

#Read in data in R
setwd("f:/mydirectory")
tmp <- read.csv("WHAS500.csv")
hold.data <-data.frame(tmp)

#Specify a SURVIVAL analysis data object
#We will use the 'length of follow-up' outcome
WHAS500.surv<-Surv(time=hold.data$lenfol,
event=hold.data$fstat)
```

Note for survival analysis object in R (and Stata), we need to specify BOTH the survival time, AND the censoring variable.

Generating Kaplan-Meier curves in R

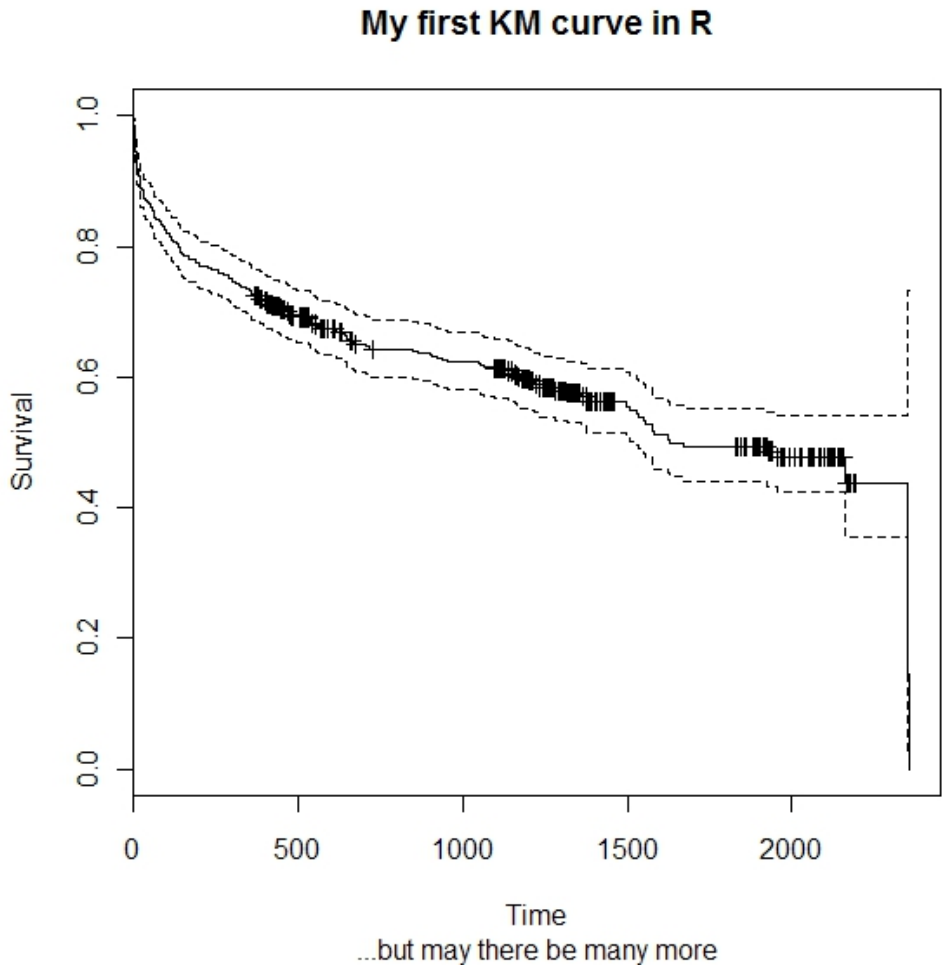
Let's start by generating the estimate of the survival curve using the Kaplan-Meier method (we won't consider any of the predictors yet)

Kaplan Meier curves

```
#Kaplan-Meier curve
my.survfit <- survfit(WHAS500.surv ~ 1)
#The '1' means constant(intercept) only

#Create plot
plot(my.survfit, xlab="Time", ylab="Survival")
title(main="My first KM curve in R")
```

The (overall) Kaplan Meier curve



Note that the black crosses represent censored values.

KM curves in terms of a categorical predictor

Now let's compare the survival curves of males and females

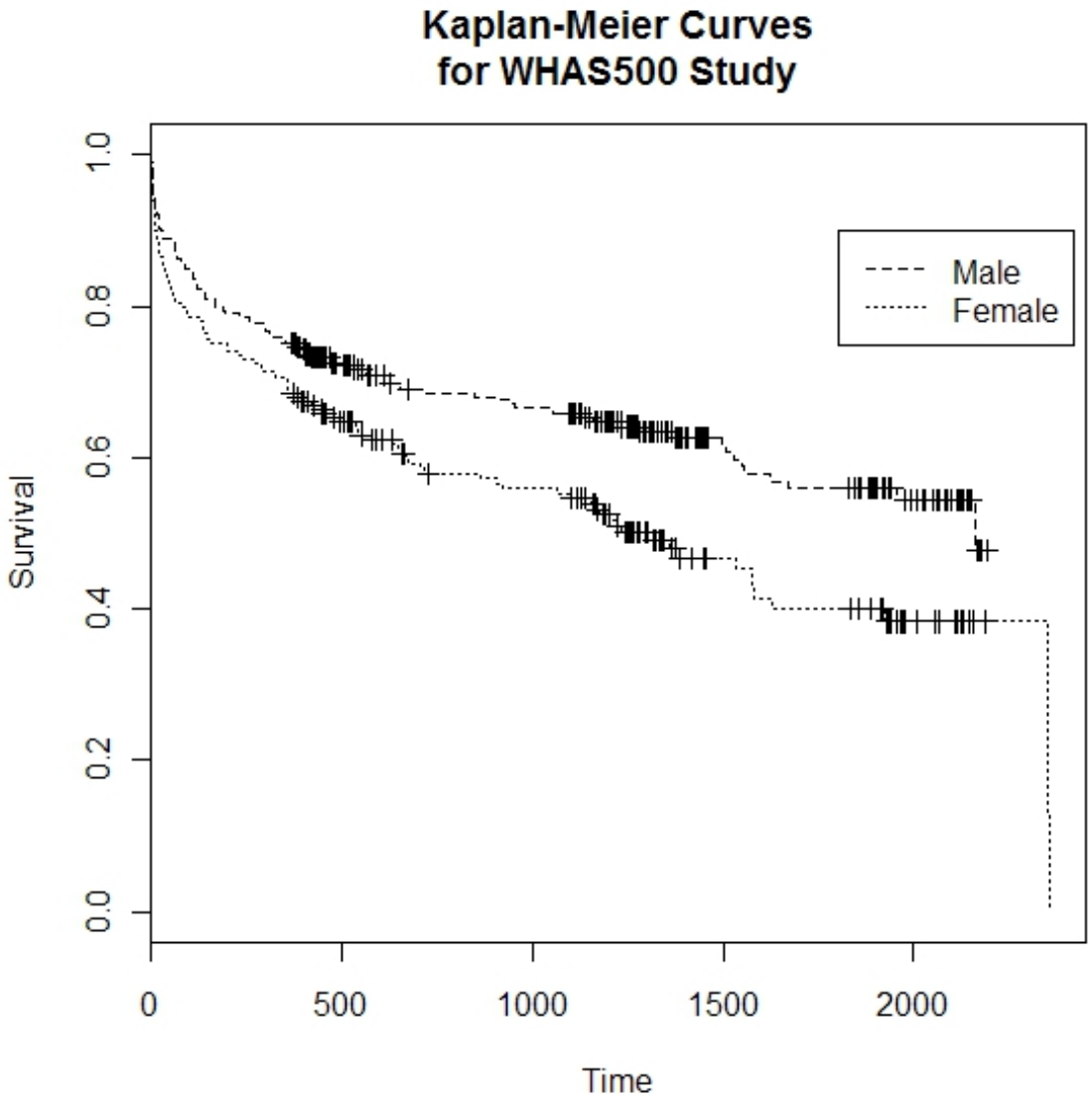
Kaplan Meier curves with categorical predcitors

```
#Kaplan-Meier curves by gender
my.survfit.gen <- survfit(WHAS500.surv~gender, data=hold.data)
#data = data frame holding covariates

plot(my.survfit.gen, xlab='Time', ylab='Survival')
#Optional extras
legend(1800, 0.9, c('Male', 'Female'))

#Arguments:  X coord, Y coord, "Names of classes"
```

KM curves by groups



Generating summary statistics

I won't go into any great detail about generating summary stats (I will leave it as an exercise), but I will show you the basics:

Survival analysis summary statistics

```
# Basic summary statistics
print(my.survfit.gen)
```

See the Survival library for more survival analysis summary statistics including: Restricted mean (survival time), extended mean, quantiles etc.

(Classical) tests for comparing survival curves

- The `survdiff()` function in R provides a whole family of tests (the G-rho family defined by Harrington and Flemmington, 1982). When the rho parameter is set to zero, this simplifies down to the **Log-rank test**
- Again using Gender as the covariate of interest:

Log-rank test for difference between two survival curves

```
# Compare survival experience among groups
my.survdiff <- survdiff(WHAS500.surv ~
gender, data = hold.data, rho=0)
```

Results: Log-rank test

```
   N Observed Expected (O-E)^2/E (O-E)^2/V
gender=0 300 111 130.7 2.98 7.79
gender=1 200 104 84.3 4.62 7.79

chisq=7.8 on 1 degree of freedom, p=0.0053
```

So we can say there is a significant difference between the survival experience of males and females

Cox proportional hazards regression

A much more useful method for modelling survival data is Cox regression. Again considering gender:

Cox regression

```
# Cox regression
my.survfit.cox<-coxph(WHAS500.surv
~as.factor(gender), data= hold.data)
summary(my.survfit.cox)
```

Pitfall: Let R know variables types: categorical and continuous

A very common mistake in R (and all stats packages) is to forget to tell the pakage that a variable is categorical. If we have a variable coded 1, 2, 3 and 4 (representing four categories) and don't tell R, it will treat it as a continuous variable

OUTPUT FROM COX REGRESSION

```
n= 500

          coef exp(coef)se(coef) z    Pr(>|z|)
gender 0.3815 1.4645   0.1376 2.773  0.00556 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05

          exp(coef) exp(-coef) lower95 upper95
gender     1.464     0.6828     1.118   1.918

Rsquare= 0.015   (max possible= 0.993 )
Likelihood ratio test= 7.6  on 1 df,p=0.005843
Wald test            = 7.69  on 1 df,p=0.005555
```

Assessing the proportional hazards assumptions

Remember that the proportionality assumpition is central to Cox **Proportional Hazards** regression.

Proportionaility $\Rightarrow$ relative risk (e.g. of an exposure) remains the same throughout the entire survival experience.

Two main methods for assessing this assumption:

1. Schoenfeld residuals plot (with a loess smooth curve fit): a flat and smooth straight line implies proportionality
2. **Test** of proportionality also using the Schonfeld residuals

Statistical tests for assessing assumptions

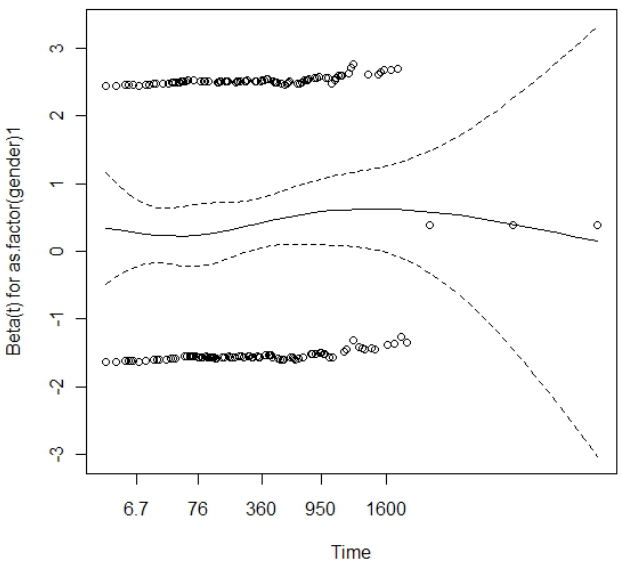
I don't like formal statistical tests of assumptions (e.g. Equal variances, Normality etc...) as they are rarely powered: a 'significance' doesn't always mean we have a problem, and a 'non-significance' doesn't always mean we are safe.

Assessing proportionaility of hazards in R

Assessing proportionality in R

```
# Assess the proportionalilty assumption
ph.assump<-cox.zph( my.survfit.cox)
```

Graphical method



'Test' method

	rho	chisq	p
gender	0.0547	0.642	0.423

Reasonably flat horizontal line (plot) and non-significant p-value $\Rightarrow$ proportionailty assumption (for gender) is valid

Data preparation in Stata

Like R, we need to tell Stata that we want to conduct a survial analysis. Specifically, we need to setup our survival outcome variable (which includes both survival time AND censorship status)

Data prepartion in Stata

```
use "F:\mydata\whas500.dta", clear

* Set up data for survival analysis
stset lenfol, failure(fstat)
```

Remember:

- `lenfol` is the amount of time followed (survival time)
- `fstat` is the censoring variable (experience the event, or not)

KM curves in Stata

Kaplan-meier curves in Stata are very easy:

Kaplan-Meier curves in Stata

```
* Generate basic KM curve
sts graph

* Now for each gender
sts graph, by(gender)
```

Cox regression in Stata

Let's start with a basic bivariate Cox regression (often called univariate in Survival analysis) :

Cox regression in Stata

```
*Fit gender effct and get HRs
stcox gender
```

```
Cox regression -- Breslow method for ties

No. of subjects =          500                     Number of obs    =       500
No. of failures =          215
Time at risk    =       441218

                                                   LR chi2(1)       =      7.59
Log likelihood  =   -1223.7851                     Prob > chi2      =    0.0059

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   1.464118   .2014392     2.77   0.006     1.118058     1.91729
------------------------------------------------------------------------------
```

We can see Gender is associated with survival, and females are 1.46 times more likely to die than males ($HR = 1.46$, $95\% CI$:1.12, 1.92, $p < 0.01$)

Cox regression in Stata: Multivariable model

Fitting a multi-variable model involves just including the extra covariates:

Cox regression in Stata

```
stcox gender bmi
```

```
Cox regression -- Breslow method for ties

No. of subjects =          500                 Number of obs    =        500
No. of failures =          215
Time at risk    =       441218
                                               LR chi2(2)       =      50.55
Log likelihood  =   -1202.3064                 Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   1.244664   .1759994     1.55   0.122     .9433863    1.642157
         bmi |    .911303   .0135847    -6.23   0.000     .8850627    .9383212
------------------------------------------------------------------------------
```

A quick interpretation

- FIRST, we see the overall model is significant ($\chi^2_{LRT} = 50.55$, $p < 0.001$)
- BMI had a major confounding effect on Gender (Gender is no longer significant, and the value of HR_{Gender} has changed considerably...certainly $\Delta OR > 10\%$)
- BMI itself is a significant risk factor ($HR_{BMI} = 0.91$; $95\% CI$: 0.89, 0.94; $p < 0.001$) and as we go up 1 unit in BMI, the chance of dying decreases by 9% (i.e.100% - 91%)

Global signifcance vs Local significance

As with ALL multivariable modeling, we MUST establish that the model is signifciant OVERALL, before going on to interpret the individual components of the model (i.e. the coeffcients)

Interaction effects

To determine whether there is an interaction (Is BMI an effect modifier??)

Cox regression in Stata

```
* Stata version 12+
stcox i.gender bmi i.gender*bmi

* Stata version <12
xi:  stcox i.gender bmi i.gender*bmi
```

Note that Stata (unlike R) seems to have problems with interactions involving continuous covariates

Assessing proportionality in Stata

Like R, Stata has a formal (test) and informal approach for assessing the proportionality assumption, but the graphical method in Stata is a little different (it uses the 'Log-minus-log' plot)

Using a Log-minus-log plot to assess proportionailty

We can see the log-minis-log plot as a mathematical transform to linearize the survival curve (as represented by the KM surve). That is, $ln(-ln(S(t)))$ should be a straight line

If we 'linearize' the survival curves for individual groups (e.g. Males and Females) then **Parallel log-minus-log curves implies proportionailty**

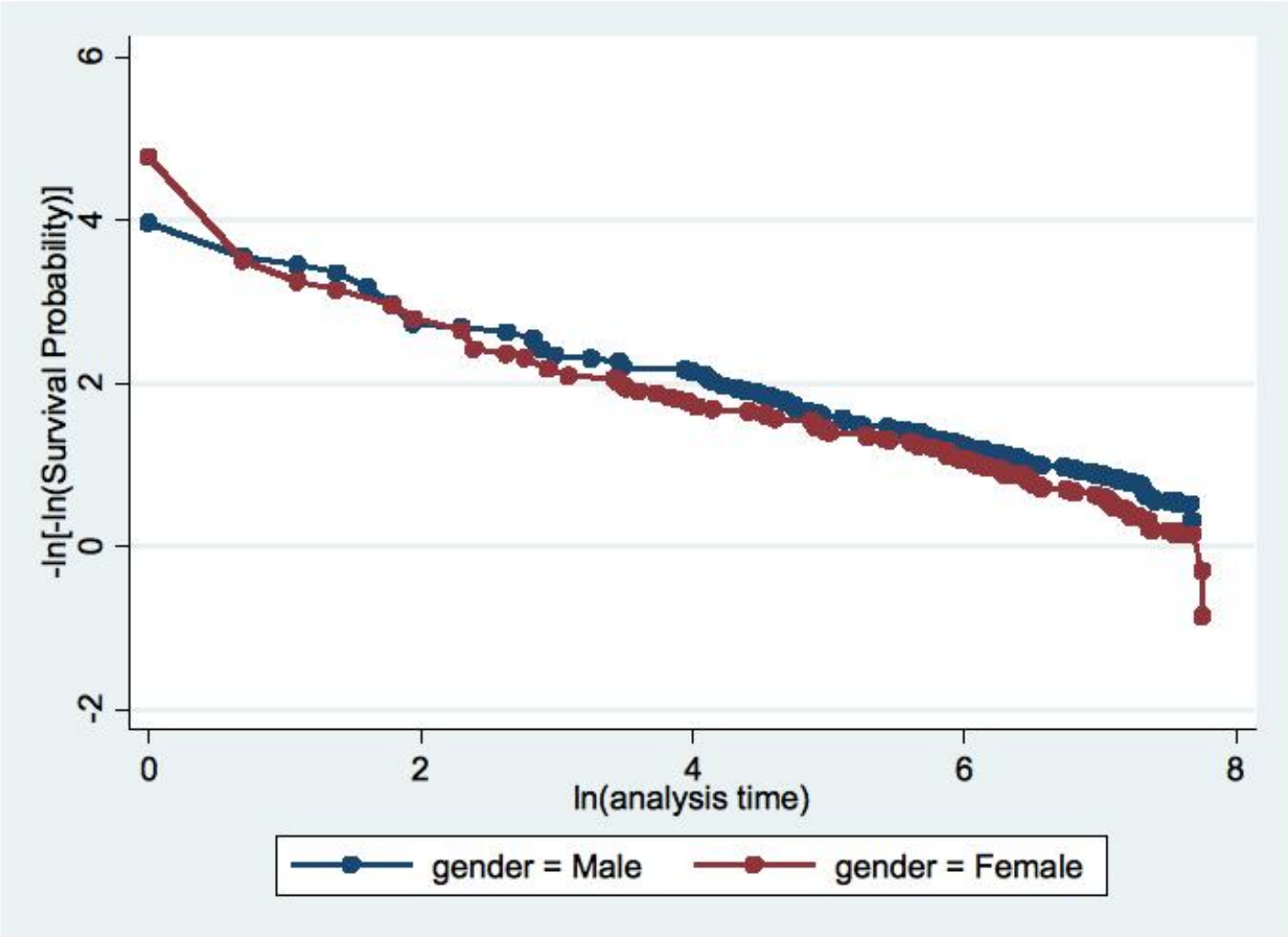
Log-minus-log plots for the Gender effect

Log-minus-log plots in Stata

```
* Generate a log-minus-log plot for gender
stphplot, by(gender) adjust(bmi)
```

Note that this is the log-minus-log plot from the multivariable model we just ran (that is, we are adjusting for BMI)

Log-minus-log plots for the Gender effect



Parallel??? Close enough

Schoenfeld residuals test in Stata

Now let's use the formal assumption test in Stata (very similar to that used in R)

Formal test of proportioanility in Stata

```
* Test the fit of Schoenfeld residuals
stphtest, detail
```

Now let's look at the proportionailty of all effects

:

```
. stphtest, detail

        Test of proportional-hazards assumption

        Time:  Time
                    |       rho     chi2       df   Prob>chi2
--------------------+---------------------------------------
_Igender_1          |   0.08295     1.57        1      0.2109
bmi                 |   0.10391     3.00        1      0.0835
--------------------+---------------------------------------
global test         |               3.77        2      0.1522
```

Overall (looks good), Gender (good), BMI (good)

Survival analysis in R and Stata

- Survival analysis is a wide and comprehensive area of biostatistics
- In many respects I have just touched the surface in terms of both
  - **Theory**: Different methods and models that can be used for Time-to-even data (in different situations); and
  - **Practice**: How we can use R and/or Stata to explore and analyze survival data
- I am concious that I spent very little time on how to read in survival data from 'date' variable types. **Why not?** It is detailed AND can be VERY FRUSTRATING.
- I suggest using Excel to convert your dates (calendar time) into survival (analysis time). If you have this type of data, just come and see me (even though it drives me crazy)

Any questions??????

Thank-you!!!!!!

QUESTIONS???